UNIVERSITÉ DE LORRAINE

École doctorale IAEM Lorraine

Facing Data Scarcity in Dialogues for Discourse Structure Discovery and Prediction

THÈSE

présentée et soutenue publiquement le 2023-08-24

pour l'obtention du

Doctorat de l'Université de Lorraine

(mention informatique)

par

Chuyuan (Lisa) Li

Composition du jury

Président :	Mathieu Constant	Université de Lorraine, France
Rapporteurs :	Benoît Crabbé	Université Paris Cité, France
	Junyi Jessy Li	University of Texas at Austin, USA
Examinateurs :	Chloé Clavel	Télécom-Paris, France
	Giuseppe Carenini	University of British Columbia, Canada
Encadrants :	Maxime Amblard	LORIA, Université de Lorraine, Nancy, France
	Chloé Braud	IRIT, Université Paul Sabatier, Toulouse, France

Laboratoire Lorrain de Recherche en Informatique et ses Applications — UMR 7503

Mis en page avec la classe thesul.

Abstract

A document is more than a random combination of sentences. It is, instead, a cohesive entity where sentences interact with each other to create a coherent structure and convey specific communicative goals. The field of discourse examines the sentence organization within a document, aiming to reveal its underlying structural information. Discourse analysis plays a crucial role in Natural Language Processing (NLP) and has demonstrated its usefulness in various downstream applications like summarization and question answering. Existing research efforts have focused on automatically extracting discourse structures through tasks such as discourse relation identification and discourse parsing. However, these data-driven methods have predominantly been applied to monologue scenarios, leading to limited availability and generalizability of discourse parsers for dialogues. In this thesis, we address this challenging problem: discourse analysis in dialogues, which presents unique difficulties due to the scarcity of suitable annotated data.

We approach discourse analysis along two research lines: "Discourse Feature Discovery" and "Discourse Structure Prediction". In the first research line, we conduct experiments to investigate linguistic markers, both lexical and non-lexical, in text classification tasks. We are particularly interested in the context of mental disorder identification since it reflects a realistic scenario. To address the issue of data sparsity, we propose techniques for enhancing data representation and feature engineering. Our results demonstrate that non-lexical and discourse-level (even though shallow) features are reliable indicators in developing more general and robust classifiers. In the second research line, our objective is to directly predict the discourse structure of a given document. We adopt the Segmented Discourse Representation Theory (SDRT) framework, which represents a document as a graph. The task of extracting this graph-like structure using machine learning techniques is commonly known as discourse parsing. Taking inspiration from recent studies that investigate the inner workings of Transformer-based models ("BERTology"), we leverage discourse information encoded in Pre-trained Language Models (PLMs) such as Bidirectional Encoder Representations from Transformers (BERT) and propose innovative extraction methods that require minimal supervision. Our discourse parsing approach involves two steps: first, we predict the discourse structure, and then we identify the relations within the structure. This two-stage process allows for a comprehensive analysis of the parser's performance at each stage. Using self-supervised learning strategies, our parser achieves encouraging results for the full parsing. We conduct extensive analyses to evaluate the parser's performance across different discourse structures and propose directions for future improvements.

Keywords: Discourse analysis, machine learning, dialogue, data scarcity, self-supervised learning

Résumé Court

Un document est plus qu'une combinaison aléatoire de phrases. Il s'agit plutôt d'une entité cohésive où les phrases interagissent les unes avec les autres pour créer une structure cohérente et transmettre des objectifs de communication spécifiques. Le domaine du discours examine l'organisation des phrases au sein d'un document, dans le but de révéler les informations structurelles sous-jacentes. L'analyse du discours joue un rôle crucial dans le Traitement Automatique des Langues (TAL) et a démontré son utilité dans diverses applications telles que le résumé et les systèmes de questions-réponses. Les efforts de recherche existants se sont concentrés sur l'extraction automatique des structures du discours à travers des tâches telles que l'identification des relations du discours et l'analyse du discours (*discourse parsing*). Cependant, ces méthodes axées sur les données ont été principalement appliquées à des scénarios de monologues, ce qui a conduit à une disponibilité et une généralisation limitées des analyseurs de discours pour les dialogues. Dans cette thèse, nous abordons ce problème difficile en raison de la rareté des données annotées : l'analyse du discours dans les dialogues.

Nous abordons l'analyse du discours selon deux axes de recherche : la « découverte des marqueurs du discours » et la « prédiction de la structure du discours ». Dans le premier axe de recherche, nous menons des expériences pour étudier les marqueurs linguistiques, à la fois lexicaux et non lexicaux, dans les tâches de classification de texte. Nous nous intéressons particulièrement au contexte de l'identification des troubles mentaux qui est un cas d'application. Pour résoudre le problème de la rareté des données, nous proposons des techniques d'amélioration de la représentation des données et de l'ingénierie des traits. Nos résultats démontrent que les marqueurs non lexicaux au niveau du discours (même s'ils sont surfaciques) sont des indicateurs fiables pour développer des classificateurs plus généraux et plus robustes. Dans un second axe, notre objectif est de prédire directement la structure du discours d'un document. Nous adoptons le cadre de la théorie *Segmented Discourse Reprsentation Theory* (SDRT), qui représente les relations rhétoriques présentes dans un document sous la forme d'un graphe. L'extraction de cette structure à l'aide de techniques d'apprentissage automatique est communément appelée *discourse parsing*. En nous inspirant des études récentes portant sur le fonctionnement interne des modèles basés sur les Transformers (« BERTology »), nous exploitons les informations discursives encodées dans les modèles de langage pré-entraînés (PLMs) tels que les *Bidirectional Encoder Representations from Transformers* (modèle BERT) et proposons des méthodes d'extraction innovantes qui qui minimisent la supervision. Notre approche de l'analyse du discours comporte deux étapes : tout d'abord, nous prédisons la structure du discours, puis nous identifions les relations au sein de la structure. Ce processus en deux étapes permet une analyse complète des performances de l'analyseur à chacune d'entre elles. En utilisant des stratégies d'apprentissage auto-supervisé, notre analyseur obtient des résultats encourageants dans l'analyse complète du discourse. Nous effectuons des analyses approfondies pour évaluer les performances de l'analyseur sur différentes structures de discours et proposons des pistes d'amélioration pour de futurs travaux.

Mots-clés: Analyse du discours, apprentissage automatique, dialogue, rareté des données, apprentissage auto-supervisé

Résumé Long

Un document n'est pas un ensemble de segments textuels aléatoires et indépendants, mais plutôt formé de séquences de phrases, ordonnées et liées entre elles, qui forment un ensemble cohérent et signifiant : cette organisation est appelée **structure du discours** (Hobbs, 1979). Dans cette thèse, nous nous sommes particulièrement intéressés à la compréhension des liens entre les clauses (extraits de texte qui ont une longueur inférieure ou égale à celle des phrases) : comment elles interagissent les unes avec les autres, quel est le type de relation qui décrit la connexion, et comment pouvons-nous automatiquement extraire cette structure d'un document.

En Traitement Automatique des Langues (ci-après TAL), l'analyse du discours est le traitement du langage au-delà des limites de la phrase. Elle se réfère à la récupération de la structure inhérente des documents, qui comprend différents niveaux d'analyse tels que la *structure thématique* : les signaux lexicaux et la distribution des mots indiquent les changements de sujet, la *structure référentielle* : les liens de coréférence entre les pronoms et les entités pour créer une cohérence locale, et la *structure de cohérence relationnelle* : deux extraits de texte sont liés par une relation rhétorique spécifique à l'aide de connecteurs explicites ou implicites (Stede, 2011).

Contrairement à l'analyse lexicale ou syntaxique, qui étudient les mots et l'interaction des mots dans une phrase individuelle, les éléments de base du discours sont des extraits de texte similaires à des clauses, connus sous le nom de *Unités de Discours* (*Discourse Units*, ci-après `DUs`). Les plus petites unités de discours sont les *Unités Discursives Élémentaires* (*Elementary Discourse Units*, ci-après `EDUs`). Nous considérons une `EDU` comme le plus petit porteur d'information, ou comme le dit Stede (2011), « une unité d'information complète et distincte, à laquelle le discours subséquent peut se connecter ». Normalement, une `EDU` reste dans la portée d'une phrase, de sorte qu'il n'y a pas d'`EDUs` inter-phrastiques. La combinaison des `EDUs` sont les *Unités Complexes de Discours* (*Complex Discourse Units*, ci-après `CDUs`). Comme première étape de l'analyse du discours, une segmentation de bonne qualité doit être effectuée de manière la plus neutre possible, pour ne pas influencer le processus d'analyse subséquent (Braud, 2015). Aussi simple que cela puisse paraître, la tâche de *Segmentation des Unités de Discours* n'est pas triviale. Ce n'est que récemment que la performance moyenne sur la tâche de segmentation pour différentes langues a finalement atteint des scores proches de 90 ($F_1 \approx 92\%$) (Zeldes et al., 2021).

Une fois les `EDUs` obtenues, l'étape cruciale suivante consiste en la construction d'une structure qui illustre les interactions entre ces unités, éventuellement enrichie de relations telles que Élaboration et Contraste. Dans la *Segmented Discourse Representation Theory* (ci-après SDRT) (Asher and Lascarides, 2003), un document est représenté comme un Graphe Orienté Acyclique (*Directed Acyclic Graph*, DAG), avec des sommets représentant les `EDUs` et des arêtes codant les relations discursives. Le principal cadre discursif que nous employons dans cette thèse est la théorie de la SDRT.

D'autres cadres discursifs ont des représentations structurales différentes. Certains d'entre eux utilisent des arbres, comme dans la *Rhetorical Structure Theory* (ci-après RST) (Mann, 1984) et le *Linguistic Discourse Model* (Polanyi and Scha, 1984; Polanyi, 1988). De plus, la RST donne également une importance aux deux `DUs` liées, connue sous le nom de « nucléarité ». Le *noyau* est l'unité discursive centrale et le *satellite* est celle qui fournit des informations auxiliaires. Il faut noter que tous les cadres discursifs ne montrent pas la structure complète d'un document : le modèle du *Penn Discourse Treebank* (PDTB) (Prasad et al., 2008a), par exemple, se concentre particulièrement sur la relation entre les segments discursifs. Il utilise des connecteurs discursifs (*donc*, *parce que*, *cependant*, etc.) pour révéler des relations discursives locales, qui ne couvrent pas nécessairement tous les `DUs` d'un document. On parle en général de *Chunking Discursif* (*Discourse Chunking*) ou d'*Analyse Discursive de Surface* (*Shallow Discourse Parsing*,

définition de la tâche partagée organisée lors de la conférence CoNLL 2015).

La représentation du discours sous forme de graphes ou d'arbres est très utile. Ces structures reflètent le flux d'information dans un document cohérent : où se trouve une nouvelle phrase et comment elle s'intègre dans le contexte actuel. De plus, des informations telles que les types de relation et la *nuclearité* reflètent l'importance relative des unités discursives. Ces informations sont bénéfiques pour de nombreuses applications en TAL, telles que la classification de textes (Ji and Smith, 2017; Ferracane et al., 2017), l'analyse de sentiment (Bhatia et al., 2015; Hogenboom et al., 2015; Nejat et al., 2017), la segmentation thématique (Jiang et al., 2021a), la traduction automatique (Marcu, 2000; Tu et al., 2013; Joty et al., 2017), le résumé (Louis et al., 2010; Hirao et al., 2013; Yoshida et al., 2014; Gerani et al., 2014; Xu et al., 2020), et la tâche de question-réponse (Verberne et al., 2007b; Jansen et al., 2014). En particulier, la représentation discursive de type dépendance a été étudiée intensivement ces dernières années pour des tâches liées au dialogue, telles que la compréhension du dialogue sous forme de réponse à des questions (Ma et al., 2021; Li et al., 2021b; He et al., 2021), et le résumé de dialogue (Feng et al., 2021b; Chen and Yang, 2021).

Les théories du discours telles que la RST (Mann, 1984), la SDRT (Asher and Lascarides, 2003), et le PDTB (Prasad et al., 2008a) ont conduit divers projets d'annotation à travers le monde, produisant des **corpus de discours** en plusieurs langues : l'anglais (Carlson et al., 2002a), le français (Péry-Woodley et al., 2011; Afantenos et al., 2012a), le basque (Iruskieta et al., 2013), le chinois (Cao et al., 2017, 2018), le russe (Shelmanov et al., 2019), etc. Parmi ceux-ci, le corpus de style RST, RST-DT (Carlson et al., 2002b), et le corpus de style SDRT, STAC (Asher et al., 2016), sont les plus couramment utilisés pour former et tester les parsers de discours pour les monologues et les dialogues, respectivement.

Malgré leur popularité, ces corpus sont relativement limités en taille : RST-DT est composé de seulement 385 articles de Wall Street Journal (environ $21,8k$ `DUs`), et STAC comprend 45 conversations de jeux (environ $10k$ `DUs`). Les autres ressources disponibles sont encore plus petites en taille. D'autres problèmes dans les corpus de discours incluent l'annotation non standardisée provenant de différentes théories du discours (Braud, 2015), l'utilisation de critères d'évaluation non comparables (Zeldes et al., 2021), et parfois la qualité de l'annotation qui est problématique. Il y a de bonnes raisons de croire que les performances en analyse du discours ont encore un long chemin à parcourir pour atteindre de bonnes performances (Morey et al., 2017; Zeldes et al., 2019).

Les approches traditionnelles d'analyse du discours se concentrent presque exclusivement sur les modèles supervisés, entraînés et testés dans le même domaine. Ces modèles peuvent être grossièrement catégorisés en approches basées sur les transitions ou basées sur la représentation graphique : la première se concentre sur l'optimisation globale de toute la structure, tandis que la seconde se concentre sur l'optimal locale. Les modèles état de l'art sur le corpus STAC (Asher et al., 2016) tels que *Deep Sequential* (Shi and Huang, 2019), *Structure-aware GNN* (Wang et al., 2021a), et *Structural-joint* (Chi and Rudnicky, 2022) atteignent les F_1 scores proche de 70% pour la prédiction de structure *nue* (sans relations), et seulement $\approx 55\%$ pour le parsing complet.

En raison du problème de la rareté des données et de la prévalence des techniques d'apprentissage par transfert, les chercheurs ont commencé à explorer différentes formes d'approches semi-supervisées et faiblement supervisées. Dans le travail pionnier de Liu and Lapata (2018), les auteurs ont produit des structures d'arbres latents à partir de tâches de résumé. Même si les arbres générés se sont avérés être superficiels et triviaux (Ferracane et al., 2019), leur approche

pour inférer la structure de l'arbre de discours à partir de mécanismes d'attention a inspiré de nombreuses études ultérieures, y compris notre propre recherche sur la prédiction de structure de discours *nue* dans le Chapitre 7.

Pour le paradigme de la supervision à distance, plusieurs études ont émergé qui exploitent les informations d'autres tâches telles que l'analyse de sentiment (Huber and Carenini, 2019), le résumé (Xiao et al., 2021), et la segmentation thématique (Jiang et al., 2021a). Ces études visent à inférer la structure du discours uniquement à partir des informations obtenues par des tâches auxiliaires, éliminant ainsi le besoin d'annotation humaine. Bien que ces approches offrent des idées innovantes et des résultats perspicaces, les performances de leurs modèles ont tendance à être relativement faibles. La plupart de ces modèles sont axés sur la prédiction de structure, avec peu ou pas de discussion sur la prédiction des relations. De plus, leur évaluation a été principalement effectuée dans le scénario de monologue, spécifiquement avec l'analyse de style RST. Une autre ligne de recherche explore le potentiel de l'apprentissage faiblement supervisé (Badene et al., 2019b,a), où l'idée est de faire un léger compromis entre qualité et quantité.

La récente montée en puissance des méthodes de transcription fiables et une augmentation de la communication en ligne ont conduit à une explosion impressionnante des données de dialogue. Par conséquent, le besoin de systèmes automatiques pour traiter les dialogues a considérablement augmenté. Par exemple, le résumé de réunions ou d'échanges avec des agents de service clientèle pourrait être utilisée pour améliorer les collaborations ou analyser les problèmes des clients (Li et al., 2019; Feng et al., 2021a); la compréhension de lecture automatisée sous forme de question-réponse pourrait améliorer les performances des agents de dialogue et aider à la construction de graphes de connaissance (He et al., 2021; Li et al., 2021b).

Les dialogues sont généralement moins structurés, entrecoupés d'un usage linguistique plus informel (Sacks et al., 1978), et ont des particularités structurelles telles que des structures en forme de losange (Asher et al., 2016). Ces caractéristiques font la richesse des dialogues, mais posent également des difficultés pour l'analyse. Par conséquent, les caractéristiques simples de niveau superficiel ne sont souvent pas suffisantes pour extraire des informations précieuses des conversations (Qin et al., 2017). Il est plutôt nécessaire de comprendre les relations sémantiques et pragmatiques qui structurent le dialogue, telles que l'utilisation de l'information discursive et de la structure de relation de cohérence.

Par conséquent, nous proposons dans cette thèse deux **questions de recherche** liées à l'analyse du discours dans les dialogues :

RQ1 Comment pouvons-nous utiliser efficacement le discours et les informations structurelles comme les marqueurs linguistiques pour les tâches de classification de texte pour le dialogue, surtout dans la détection de troubles mentaux ?

RQ2 Comment pouvons-nous générer des structures discursives avec des techniques d'apprentissage automatique en utilisant une supervision minimale pour obtenir la meilleure applicabilité dans des scénarios réels ?

Les deux questions de recherche sont abordées à travers plusieurs sous-projets.

Pour répondre à la RQ1, nous poursuivons la première direction de recherche intitulée « Découverte des Marqueurs Discoursifs », qui vise à étudier le discours dans un sens général qui ne se limite pas aux structures de type SDRT ou RST. Au départ, nous nous concentrons sur les tâches de classification de texte qui impliquent l'utilisation de marqueurs discursifs de base tels que les connecteurs discursifs et les actes de dialogue. Deux tâches sont menées dans le domaine du déficit cognitif : la première est la détection de la schizophrénie, qui a donné lieu à deux

publications (Amblard et al., 2020; Li et al., 2021a) et à plusieurs présentations, notamment lors de la Journée commune AFIA-THL / ATALA - la santé et le langage en France et lors de *Workshop on the Semantics and Pragmatics of Dialogue* (SemDial 2021); la seconde est la détection de la dépression avec une publication internationale et une présentation à la conférence SIGDial (Li et al., 2022).

Le contexte du déficit cognitif constitue une situation réaliste. Aujourd'hui, environ 1% des adultes dans le monde sont touchés par la schizophrénie. L'impact de la dépression est encore plus grand : environ 4% de la population mondiale, et un taux plus élevé chez les personnes âgées, selon les chiffres rapportés par l'Organisation Mondiale de la Santé. Ces maladies mentales présentent divers symptômes, parmi lesquels des troubles linguistiques tels que le *langage désorganisé* et la *pauvreté du vocabulaire* (Kuperberg, 2010a). Les praticiens du TAL peuvent s'intéresser à ces troubles du langage et considérer leurs analyses linguistique comme potentielle source de descriptions des symptômes et être utile à une meilleure compréhension de la maladie et de ses manifestations. Cela pourrait aider à la détection précoce de la maladie et éventuellement fournir une aide dans son traitement. Cependant, les modèles actuels pour la détection des troubles mentaux sont loin d'être idéaux. La majorité des recherches dans ce domaine s'appuient sur les données des réseaux sociaux (Benton et al., 2017; Mitchell et al., 2015; Birnbaum et al., 2017a; Guntuku et al., 2017), avec un accent particulier sur l'information lexicale. Cependant, comme souligné dans notre étude (Li et al., 2021a), ces approches ont des limites dans certaines langues et pourraient conduire à des résultats biaisés.

Notre objectif est de développer des modèles plus fiables et robustes, ce qui nous incite à explorer des marqueurs linguistiques qui dépendent moins de l'information lexicale et privilégient plutôt l'information structurelle. Les résultats de notre investigation sont présentés dans deux projets. Le Chapitre 4 présente le premier projet sur la détection linguistique de la schizophrénie. Nous observons que les caractéristiques lexicales, bien que très précises, présentent un fort biais. Par conséquent, nous explorons des caractéristiques délexicalisées telles que les arbres syntaxiques et des caractéristiques moins lexicalisées telles que les connecteurs discursifs.

Ensuite, dans le Chapitre 5, nous nous penchons sur la détection de la dépression dans les dialogues. En l'absence de structures discursives de référence, nous proposons d'incorporer l'information discursive dans le cadre de l'Apprentissage Multi-Tâches (*Multi-Task Learning*, MTL). Nous adoptons une approche simple mais efficace connue sous le nom de schéma *entièrement partagé* (*fully-shared*), où les couches cachées sont partagées entre toutes les tâches.

Pour répondre à la RQ2, nous établissons une deuxième direction de recherche intitulée « Prédiction de Structure du Discours ». Notre travail s'ancre dans la SDRT et utilise le corpus STAC pour sa mise en oeuvre pratique. Nous adoptons une approche en deux étapes pour aborder cette ligne de recherche. La première étape concerne la prédiction de la structure discursive nue, présentée dans le Chapitre 7. Il convient de noter que les structures nues ont été démontrées comme étant des caractéristiques précieuses pour certaines tâches, telles que la sélection de contenu (*content selection*) (Louis et al., 2010) et l'extraction de fils de discussion (*thread extraction*) (Jiang et al., 2020). Il s'agit d'un travail collaboratif mené avec des collègues de l'Université de Colombie-Britannique à Vancouver durant mon stage au sein du groupe TAL de l'UBC. Ce projet a abouti à une publication lors de la conférence EACL 2023 (Li et al., 2023) et à une présentation lors du 4ème *Workshop on Computational Approaches to Discourse* (CODI 2023).

Contrairement aux études précédentes qui s'appuient sur une supervision complète (Afantenos et al., 2015; Shi and Huang, 2019; Chi and Rudnicky, 2022), notre objectif est d'effectuer un parsing discursif avec moins de données nécessitant une annotation humaine, afin que notre

analyseur puisse être utilisé dans des cas plus généraux. Cependant, sélectionner les signaux de supervision à distance (par exemple de sentiment ou de résumé) ou faiblement supervisés (de règles heuristiques) n'est pas simple. Au vu des résultats prometteurs de l'information discursive capturée dans les modèles de langage pré-entraînés (*Pre-trained Language Models*, ci-après PLMs), comme introduit dans le Chapitre 6, nous choisissons finalement les PLMs comme source de supervision. Nous explorons divers PLMs et découvrons que le réseau encodeur du modèle BART (Lewis et al., 2020) est le plus performant. Nous proposons également des tâches de *fine-tuning* adaptées aux dialogues pour renforcer l'information discursive codée dans les matrices d'attention, sans nécessiter d'annotation supplémentaire.

Par la suite, dans le Chapitre 8, nous menons une étude sur la prédiction des relations discursives basée sur la structure extraite des PLMs. En nous inspirant des approches décrites dans Nishida and Matsumoto (2022), nous utilisons des stratégies de bootstrapping via le *self-training*. À l'aide de quelques documents annotés, nous formons d'abord un modèle source puis l'utilisons pour générer des étiquettes pseudo sur des données non annotées. Les instances étiquetées pseudo de haute confiance sont sélectionnées et combinées avec les documents originaux pour un nouveau cycle de construction du modèle.

À l'issue de ces expériences, nous disposons de multiples élements pour répondre à nos problématiques.

Dans le premier projet, nous proposons deux méthodes pour aborder le problème de la rareté des données dans la tâche de classification de textes pour la schizophrénie. La première méthode consiste à explorer différents niveaux d'ingénierie des *features*, y compris des marqueurs lexicaux (*Bag-Of-Words*), syntaxiques (*POS tagging*) et discursifs (*Backchannel* response, *Open Class Repairs*, connecteurs discursifs). La seconde méthode consiste à modéliser les dialogues en limitant l'analyse aux tours de parole des patients et en testant différentes fenêtres de contexte pour améliorer la représentation des données. Nous comparons plusieurs algorithmes de classification et constatons que le *Naive Bayes* fonctionne bien avec les décomptes lexicaux, tandis que les SVM et *Logistic Regression* sont mieux adaptés aux données rares et aux caractéristiques de haute dimension. L'analyse révèle que les patients ont tendance à discuter volontairement de leur maladie et de leur traitement, aboutissant à des sujets liés à la maladie, ce qui biaise fortement le lexique. Les modèles délexicalisés, qui mettent l'accent sur les informations morpho-syntaxiques et les caractéristiques discursives de haut niveau, sont plus généralisables. Nous découvrons également des résultats intéressants concernant les caractéristiques des patients schizophrènes, tels que leur utilisation de davantage de phrases verbales et adverbiales et moins d'expressions phatiques, ce qui est cohérent avec les études précédentes.

Le second projet examine la structure hiérarchique du discours dans les dialogues et son potentiel pour la détection de la dépression. Pour pallier le problème de rareté des données, nous nous inspirons du cadre d'apprentissage multi-tâches et apprenons conjointement des caractéristiques à partir de plusieurs tâches connexes. Nous considérons trois tâches auxiliaires : la classification des émotions, la classification des actes de dialogue et la classification des sujets, pour explorer comment une information superficielle sur la structure du dialogue peut améliorer les performances. Pour intégrer l'organisation du dialogue, nous proposons une architecture hiérarchique spécifique au dialogue, où deux tâches (classification des émotions et des actes de dialogue) sont réalisées au niveau du tour de parole, tandis que deux autres (détection de la dépression et classification des sujets) sont réalisées au niveau du document. Nous observons des améliorations significatives lors de l'ajout de chaque tâche séparément. Apprendre conjointement les quatre tâches entraîne une amélioration dans tous les indicateurs (F_1 +27%, Accuracy +11%). Nos études d'ablation montrent que la détection des émotions et de la dépression se

renforcent mutuellement. Les résultats positifs pour les marqueurs superficiels, tels que les actes de dialogue et les sujets, indiquent également leur pertinence pour la structure du dialogue.

Dans le troisième projet, nous proposons un cadre conçu en pipeline pour l'extraction automatique de la structure et des relations discursives. La partie extraction de la structure du discours innove en utilisant des méthodes semi-supervisées et non supervisées pour traiter les problèmes de rareté des données dans les dialogues et extraire des informations discursives à partir de modèles de langue pré-entraînés. Nous examinons la robustesse et la localité des structures discursives dans les PLMs en analysant les informations capturées à travers les têtes d'attention et diverses tâches de *fine-tuning*. Choisir la meilleure tête d'attention est un problème crucial lors de l'utilisation des PLMs pour extraire des informations discursives au niveau du document. Les résultats expérimentaux sur le corpus STAC montrent que les méthodes non supervisées et semi-supervisées surpassent un *baseline* assez puissant (F_1 $56,8\%$), offrant des gains substantiels sur l'ensemble de données complet (F_1 $59,3\%$) et des améliorations supplémentaires sur le sous-ensemble structuré en arbre (F_1 $68,1\%$). L'analyse qualitative des structures déduites montre que notre modèle prédit avec succès plus de 82% des arcs projectifs, certains s'étendant sur quatre `EDUs`. Ce résultat est encourageant et suggère que notre approche est capable d'extraire des structures discursives raisonnables avec une supervision minimale.

Le deuxième module – prédiction de la relation – est construit sur la partie extraction de structure et se concentre sur l'exploitation des PLMs par le *self-training*. Nous examinons diverses techniques de sélection des données pseudo-étiquetées, et constatons que la sélection des échantillons basée uniquement sur les scores de confiance n'est pas suffisante. Bien que le *self-training* puisse améliorer les performances du modèle, l'amélioration est modeste (environ 1%). Le défi principal de *self-training* réside dans la génération d'étiquettes pseudo précises et diversifiées. Pour surmonter cette limitation, nous étudions le potentiel d'une stratégie « *human-in-the-loop* » en fournissant une annotation correcte pour les exemples incertains ayant des scores de confiance faibles. Nos résultats suggèrent que l'effort humain peut être bénéfique, mais nécessite une quantité considérable d'annotation. Cependant, dans des situations pratiques, il peut être difficile d'obtenir une telle supervision étendue.

De plus, nous fournissons des résultats de parsing complet qui combinent la segmentation `EDU`, la prédiction de structure et la classification des relations, établissant ainsi la première référence pour un analyseur discursif complet pour les dialogues formés à l'aide d'une supervision faible. Les résultats empiriques montrent une progression graduelle, bien que modeste, qui ouvre la voie à un parsing discursif complet dans les dialogues.

Bien que nous ayons abordé les deux questions de recherche avec les projets précédents, il y a encore de la place pour des améliorations.

Pour modéliser l'interaction dans la classification de la langue de la schizophrénie, des réseaux de neurones pourraient être utilisés à la place des modèles probabilistes classiques. Une possibilité est d'utiliser le *adversarial learning* au sein d'un modèle neuronal. Dans le *adversarial learning*, un modèle adversaire est formé pour maximiser une fonction de perte opposée à celle du modèle original. En introduisant cette composante antagoniste, le modèle original est forcé d'apprendre des caractéristiques plus généralisables moins sujettes au biais (Zhang et al., 2018a). Nous pouvons nous inspirer des travaux qui s'attaquent au biais de genre, comme dans Bordia and Bowman (2019); Liu et al. (2020).

Afin de coder la structure du dialogue, nous pourrions également envisager des informations structurales plus profondes telles que le parsing discursif. Cependant, cette approche pose un défi direct en raison du manque de parseurs discursifs généraux et puissants, un problème que nous prévoyons d'aborder dans des travaux futurs. Une étape supplémentaire consistera à étudier la

généralisation de notre modèle à d'autres troubles de la santé mentale, tels que la détection de la démence.

Dans la dernière partie de la thèse, bien que nous montrions des résultats initiaux prometteurs sur la capacité à capter des structures discursives valides à partir de méthodes semi-supervisées et de *self-training*, la performance de nos méthodes proposées reste limitée, notamment par rapport aux systèmes entièrement supervisés sur les modèles de parsing intra-domaine.

Nous laissons plusieurs questions sans réponse qui pourront être approfondie dans nos futurs travaux.

La première concerne l'amélioration de la structure discursive extraite pour qu'elle s'aligne mieux avec les graphes de type SDRT. Une approche possible serait de ré-implémenter les méthodes de *Integer Linear Programming* présentées dans Perret et al. (2016) mais avec des PLMs comme structures de base. La deuxième question ouverte concerne l'approche conçue en pipeline employée pour le parsing discursif, qui est susceptible de propager d'erreurs. Dans ce cas, une simple proposition du modèle conjoint est d'augmenter les structures discursives avec des informations supplémentaires. Par exemple, si une relation de haute confiance est identifiée entre deux `EDUs` qui n'ont pas été reliées, nous pourrions effectuer un raffinement a posteriori et ajouter l'attachement manquant à la structure. Troisièmement, après avoir montré toutes les applications en synergie en aval dans le Chapitre 3, nos parseurs discursifs ont un potentiel significatif pour être appliqués à de nouveaux domaines et utilisés pour d'autres tâches. Nos approches semi-supervisées sont actuellement les ressources les plus efficaces pour produire des structures discursives pour des documents bruts.

Acknowledgements

I would not been doing a thesis without Benoit Crabbé and Jonathan Ginzburg, my Master's Professors at Paris Diderot, now known as Université Paris Cités. This work would not have been done without the support of my supervisors: Maxime Amblard and Chloé Braud. From them, I learned not only scientific proficiency but also invaluable life lessons, a balanced perspective on work and life, and resilience in the face of challenges, whether they be unfavorable outcomes, harsh reviews, or tricky situations. Their comprehensive insights have immensely enriched this thesis and eased my journey as a Ph.D. student over the past four years. I would like to express my gratitude to Maria Boritchev and Siyana Pavlova for proofreading the abstract.

A heartfelt acknowledgment to the jury members present at my defense: Junyi Jessy Li, Benoit Crabbé, Giuseppe Carenini, Chloé Clavel, and Mathieu Constant. I am extremely grateful that you took the time to weigh in on my work. Thank you for the constructive suggestions. I will always remember on a hot and humid day end of August, we sat in a stuffy room but still, engaged in several rounds of interesting discussions.

I owe a special thanks to Giuseppe Carenini, whose guidance was indispensable. His offer for an internship at UBC was a turning point for me. I am looking forward to our upcoming endeavors together! Likewise, many thanks to Patrick Huber, Xiao Wen, Xing Linzi, Raymond Li, and the NLP team at UBC. I also wish to thank my previous teachers at Paris Diderot, including Benoit Crabbé, Marie Candito, and the entire Linguistic Department. Their influence is undeniably a reason I'm working in NLP today.

I'd like to extend my heartfelt appreciation to the members of (mostly) Semagramme team at LORIA, in no particular order: Philippe de Groote, Sylvain Pogodalla, Michel Musiol, Bruno Guillaume, Karen Fort, Guy Perrier, Aurore Coince, Isabelle Herlich, Isabelle Blanchard, Adrien Coulet, Christophe Cerisara, Gael Guibon, Philippe Muller (IRIT), Nicholas Asher (IRIT), Damien Sileo (Inria), as well as to my peers: Pierre Lefebvre: an excellent engineer with whom we conducted beautiful projects: UFO and the SDRT-Anno platform. Maria Boritchev and Titouan Carette, Pierre Ludmann, Clément Beysson-Cabaret, William Babonnaud: former colleagues who welcomed me and helped me settle in Nancy. Siyana Pavlova, Khensa Amandi Daoudi, Marc Anderson: my lovely office-mates, enduring moves from the *frigo* to the *four* without a word of complaining. Vincent Tourneur, Gabriel Sauger, Valentin Richard, Marie Cousin, Amandine Bob Decker, Bertrand Remy, Amandine Lecomte, Samuel Buchel, Fanny Ducel, Hee-Soo Choi, Priyansh Trivedi, and Matthieu Amet: your infectious laughter resonates in the hall. It is nice to see the team growing and I am lucky to finish my manuscript with such a happy team. Kelvin Han and Anna Liednikova, friends from the SyNaLP team for the intriguing discussion on generation. Timothée Mickus, a conference-frequent friend and talented story-maker.

Numerous individuals have supported me throughout this rigorous four-year research journey: Salomé Valade and her parents, my dearest friends and family in France, always showering me with warmth and support. Ma Anji, my boyfriend, supported and encouraged me during my toughest moments. Feng Hailan's vivacity always brings light to my days. Liu Mo's lucid thinking helped soothe my nerves. During the hard COVID years, the band I found among fellow Chinese Ph.D. students at Du Kou's house was invaluable. Together, we enjoyed games, food, discussions, and celebrated traditions. I'm honored to be in the company of such brilliant minds. Here's to our future meetings, whether in China or any corner of the world!

Lastly, thank you, parents! I'm grateful for your unwavering support as I navigate this significant milestone in my life. Three days prior to my defense, my grandfather celebrated his 90^{th} birthday. While I couldn't be there in person, watching the videos of our big family made me feel close. Best wishes to you, grandpa!

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Contents

1.1 Discourse Structure & Parsing

A document is not a random and independent text spans, but instead sequences of ordered and related sentences which together make coherent and meaningful documents: this organization is called **discourse structure** (Hobbs, 1979). In this thesis, we are particularly interested in understanding the connection between clauses (text spans that are shorter than or equal to sentences in length): how they interact with each other, what is the relation type to describe the attachment, and how can we automatically extract the structure out of a document.

In Natural Language Processing (NLP), discourse analysis is language processing beyond the sentence boundary. It refers to the retrieval of the inherent structure of documents, which include different levels of analysis such as *topic structure*: lexical signals and word distribution indicate topic shifts, *referential structure*: coreference links between pronouns and entities in order to create local coherence, and *coherence-relational structure*: two text spans are linked together with specific semantic relation using explicit or implicit connectives (Stede, 2011).

Different from lexical or syntactic analysis, which study words and the interaction of words in individual sentence, the basic elements in discourse are clause-like text spans, known as *Discourse Units* (`DUs`). The smallest units of `DUs` are *Elementary Discourse Units* (short in `EDUs`), and the combination of `EDUs` are *Complex Discourse Units* (short in `CDUs`). Normally, a `EDU` stays within the range of a sentence, so that there are no inter-sentential `EDUs`. In the related literature, we do not find a consensus on the definition of `EDUs` (Section 2.1). Linguists hold their own opinions when defining the criteria with linguistic phenomena such as *ellipsis*, *relative clause*, and *prepositional phrases* (Mann, 1984; Polanyi and Scha, 1984; Asher, 1993; Tofiloski et al., 2009). We regard a `EDU` as the smallest piece of information carrier, or as put in Stede (2011), "a complete, distinct unit of information that the subsequent discourse may connect to". We show a concrete example from the Strategic Conversation Corpus (STAC) (Asher et al., 2016), a corpus of online conversations during the game Settlers of Catan and was annotated under the Segmented Discourse Rhetorical Theory (SDRT) (Asher and Lascarides, 2003):

(1) 167 *gwfs:* [so how do people know about the league?]$_1$

170 *lj:* [i did the trials]$_2$

174 *tk:* [i know about it from my gf]$_3$

175 *gwfs:* [yeah me too,]$_4$

176 *tk:* [did not do the trials]$_5$

178 *gwfs:* [i did them]$_{6a}$ [because a friend did]$_{6b}$

In example (1), each line is a speech turn, i.e., one entry, which contains a speech index (167, 170, etc.), a participant (*gwfs*, *lj*, *tk*), and a text span. STAC corpus contains sub-dialogues or threads that divide and merge as the dialogue proceeds. For readability, we only extract the speech turns in one thread, which explains the disjoint speech indices.

This dialogue consists of 6 speech turns and 7 *Discourse Units* (marked with subscript numbers), the first five are `EDUs` and the last one is a `CDU`. As the very first step in discourse analysis, a good quality segmentation should be performed in an objective and impartial manner to lay the ground for subsequent analysis such as link attachment and relation prediction (Braud, 2015). Simple may it looks, the task of *Discourse Unit Segmentation* is non-trivial. Only recently, the average performance on segmentation task for different languages has finally reached the low 90*s* (Zeldes et al., 2021). Thanks to the DISRPT shared tasks, we now have state-of-the-art `EDU` segmentors such as ToNy (Muller et al., 2019), DisCoDisCo (Gessler et al., 2021), and DisCut (Ezzabady et al., 2021) that work well in 11 languages[1].

Disposed of elementary discourse units, the next crucial step is to build a structure that illustrates the interactions among these units, eventually enriched with relations. We show a realization of SDRT-type discourse structure for example (1) in Figure 1.1. `EDUs` are ranged vertically to echo the order of their appearance[2]; they are linked with each other with typed edges, reflecting the discourse relations. Speaker *gwfs* first asks a question, other two participants both address it – thus creating two *question answer pairs* ("qap"). With the development of the conversation, we discover more relation types, such as *parallel* (170 - 175) when the *lj* and *gwfs* share a common theme, *elaboration* (174 - 176) when *tk* provides more information on her previous speech, and *contrast* (176 - 178) as participants *tk* and *gwfs* present opposite opinions, etc. With vertices representing `EDUs` and edges encoding discourse relations, a document[3] is thus represented as a Directed Acyclic Graph (DAG) – the standard discourse structure in the SDRT framework. In part III of this thesis (Chapter 7, 8), we adopt SDRT as our theoretical foundation for discourse analysis and we perform discourse parsing to automatically extract such graph structure from a given document.

Other discourse frameworks have different structure representation (Section 2.2). Some of them use trees, such as in Rhetorical Discourse Theory (RST) (Mann, 1984) and Linguistic Discourse Model (Polanyi and Scha, 1984; Polanyi, 1988). Additionally, RST also gives relative importance to the linked `DUs`, namely *nulcearity*. *Nucleus* is the core discourse unit and *satellite* is the one that provides auxiliary information. An example of an RST-style discourse tree is shown

[1]According to the results from DISRPT shared task on `EDU` segmentation: `https://sites.google.com/georgetown.edu/disrpt2021/results`, where automatic segmentors have been tested on 13 datasets in 11 languages.

[2]Note that theoretically, in SDRT, discourse units are represented in embedded boxes for hierarchical structures and are placed horizontally or vertically for different relation types (cf details in Section 2.2.2). In STAC corpus, boxes are removed and speech turns are shown in chronological order in a diagram for better visualization, as presented in `https://www.irit.fr/STAC/stac_game_graphs/readme.html`.

[3]In this thesis, we use the term "documents" to refer to written texts, including both monologues and dialogues.

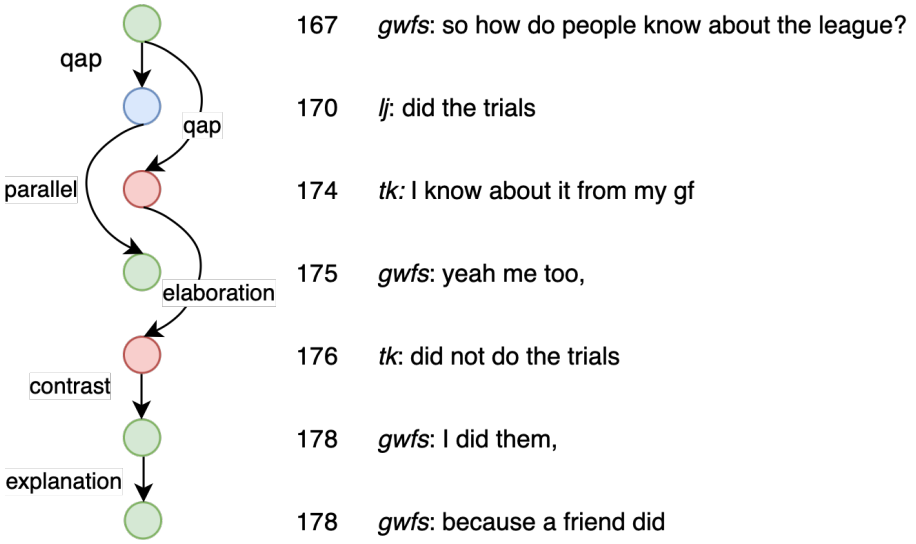


Figure 1.1: On the left: SDRT-style discourse structure of example (1). Circles are `EDUs`. Speakers are distinguished using various colors. Arrows indicate link attachment (from *head* to *dependent*) between `EDUs`. Attachments are typed with rhetorical relations. "qap": *QA pair*.

in Figure 1.2. Note that not all the discourse frameworks show the full structure of a document: the Penn Discourse Treebank's framework (PDTB) (Prasad et al., 2008a) for instance, has a particular focus on the relationship between discourse segments, they utilize connectives (*so*, *because*, *however*, etc.) to reveal *local* discourse relations, which not necessarily cover all the `DUs` in a document. We call discourse analysis in PDTB-style parsing *Shallow Discourse Parsing*.

Discourse represented in graph- or tree- structure is very useful. These structures reflect the information flow in a coherent document: where a new sentence is located and how it fits into the current context. Further, information such as the relation types and *nuclearity* reflect the relative importance of discourse units. This information is beneficial for many downstream applications in NLP (Section 3.3). We discover synergistic tasks such as text classification (Ji and Smith, 2017; Ferracane et al., 2017), sentiment analysis (Bhatia et al., 2015; Hogenboom

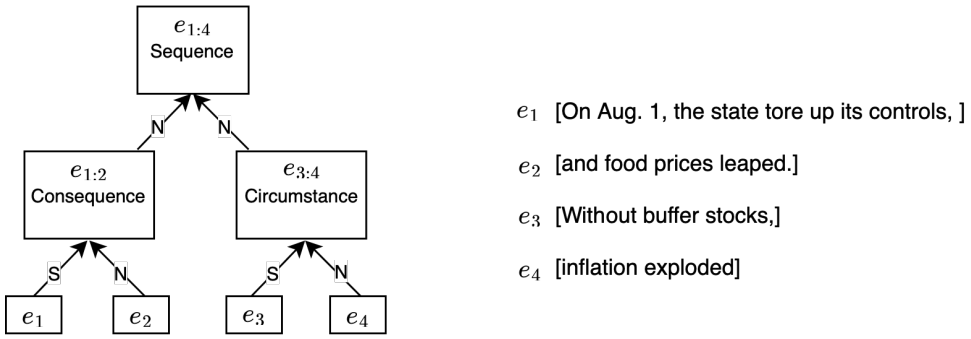


Figure 1.2: On the left: RST tree representation of a text example extracted from Wall Street Journal (*wsj_1146*). From bottom to top, adjacent `EDUs` are combined into intermediate `DUs`. "S": *satellite*; "N": *nucleus*.

et al., 2015; Nejat et al., 2017), topic segmentation (Jiang et al., 2021a), machine translation (Marcu, 2000; Tu et al., 2013; Joty et al., 2017), summarization (Louis et al., 2010; Hirao et al., 2013; Yoshida et al., 2014; Gerani et al., 2014; Xu et al., 2020), and question answering (Verberne et al., 2007b; Jansen et al., 2014). In particular, dependency-style discourse representation has been studied intensively in recent years for dialogue-related tasks such as dialogue comprehension in the form of question answering (Ma et al., 2021; Li et al., 2021b; He et al., 2021), and dialogue summarization (Feng et al., 2021b; Chen and Yang, 2021).

1.2 Resources & Existing Models

Discourse theories such as RST (Mann, 1984), SDRT (Asher and Lascarides, 2003), and PDTB's framework (Prasad et al., 2008a) have lead various annotation projects worldwide, leaving **discourse corpora** in multiple languages (Section 2.3): English (Carlson et al., 2002a), French (Péry-Woodley et al., 2011; Afantenos et al., 2012a), Basque (Iruskieta et al., 2013), Chinese (Cao et al., 2017, 2018), Russian (Shelmanov et al., 2019), etc[4]. Among these, the RST-style corpus RST-DT (Carlson et al., 2002b) and the SDRT-style corpus STAC (Asher et al., 2016) are the most commonly used for training and testing automatic discourse parsers in monologue and dialogue settings, respectively. Despite their popularity, these corpora are relatively limited in size: RST-DT consists of only 385 Wall Street Journal news articles (approximately $21.8k$ `DUs`), and STAC comprises 45 gaming conversations (approximately $10k$ `DUs`). Other resources available are even smaller in size. In comparison to research on syntax parsing, Universal Dependencies[5] (Nivre et al., 2016) offers a vast collection of over 200 treebanks spanning over 100 languages. For English alone, there are nine treebanks available, comprising more than $46k$ annotated sentences. The size of annotated discourse treebanks may hinder the development of general and high-functional discourse parsers, making them not easily applicable to downstream applications (Vargas et al., 2022). Other issues in discourse corpora include the unstandardized annotation guidelines originating from different discourse theories (Braud, 2015), the un-matchable evaluation criteria (Zeldes et al., 2021), and sometimes the questionable annotation quality (Section 2.3.5). There is good reason to believe that performance on discourse analysis and parsing has a substantial way to go (Morey et al., 2017; Zeldes et al., 2019). Through shared tasks like DISRPT, the discourse community shares the desire to foster collaboration and promote standardized data formats, consistent evaluation guidelines, and diverse discourse tasks. Through these collective efforts, we anticipate achieving greater transparency in comparing different systems and their performance in the field.

Traditional discourse parsing approaches are near-exclusively focusing on supervised models, trained and tested in the same domain (Section 3.2.1). These models can be roughly categorized into transition-based and graph-based approaches: the first one focuses on global optimization over the entire structure, while the second focuses on local optimal. State-of-the-art models on STAC (Asher et al., 2016) corpus such as *Deep Sequential* (Shi and Huang, 2019), *Structure-aware GNN* (Wang et al., 2021a), and *Structural-joint* (Chi and Rudnicky, 2022) reach the low $70s$ on *naked* structure prediction (without relations), and only middle-$50s$ on the full parsing.

Due to the data sparsity issue and the prevalence of transfer learning techniques, researchers started to explore different forms of semi-supervised and weakly-supervised approaches. In the pioneering work of Liu and Lapata (2018), authors produced latent tree structures from summarization task. Even though the generated trees are proven to be shallow and trivial (Ferracane

[4]For more languages, check the latest DISRPT github: `https://github.com/disrpt/sharedtask2023`.
[5]`https://universaldependencies.org`

et al., 2019), their approach of inferring discourse tree structure from attention mechanisms has inspired many subsequent studies, including our own research on *naked* discourse structure prediction (Chapter 7). In the paradigm of distant supervision (Section 3.2.2.1), several studies have emerged that leverage information from other tasks such as sentiment analysis (Huber and Carenini, 2019), summarization (Xiao et al., 2021), and topic segmentation (Jiang et al., 2021a). These studies aim to infer discourse structure solely based on the information obtained from auxiliary tasks, eliminating the need for human annotation. While these approaches offer novel ideas and insightful findings, their model performances tend to be relatively low. Additionally, most of these models are focused on structure prediction, with limited or no discussion on relation prediction. Furthermore, their evaluation has been primarily conducted in the monologue scenario, specifically with RST-style parsing. Another line of research explores the potential of weakly supervised learning (Section 3.2.3), where the idea is to make a slight trade-off between quality and quantity. For instance, Badene et al. (2019b,a) employed expert-composed heuristics within the Snorkel framework (Ratner et al., 2017) to capture `EDU` attachment on raw data. They demonstrated promising results on the STAC corpus, comparable to those of a locally supervised model (Perret et al., 2016). However, this approach has a drawback in terms of the complex rule-writing process, which requires experts and a large validation set for verification. Moreover, these rules can only address a limited number of relation attachments, resulting in biased outcomes.

In real-life scenarios, how can we make use of pre-trained discourse parsers on the target domain? Research by Liu and Chen (2021) shows that direct transfer results in poor performance which can be lower than simple baselines. The generalization issue in discourse parsing thus triggered studies in unsupervised domain adaptation (Section 3.2.4). Particularly, we advocate the work by Nishida and Matsumoto (2022) where authors apply several bootstrapping strategies, including self-training, co-training, and tri-training for domain adaptation. By using pseudo-labeled data to enrich the model during retraining, they increased the initial performances in the dialogue setting by 6 and 2 points for *naked* structure and full structure parsing, respectively. Inspired by their work, we build up our research for discourse relation prediction using self-training strategies (Chapter 8).

1.3 Focus & Contributions

A dialogue corresponds to exchanges between two or more people, in contrast to monologues which are usually authored by a single person. Dialogues are generally less structured, interspersed with more informal linguistic usage (Sacks et al., 1978), and have structural particularities such as *diamond-shaped* structures (Asher et al., 2016) (Section 2.4). These characteristics construct the richness in dialogues but also pose difficulties in analysis. Our focus in this thesis is **discourse in dialogues**.

The recent rise of reliable transcription methods and a spike in online communication led to an astonishing explosion of dialogue data. As a result, the need for automatic systems to process dialogues has increased dramatically. For example, summarization of meetings or exchanges with customer service agents could be used to enhance collaborations or analyze customers issues (Li et al., 2019; Feng et al., 2021a); machine reading comprehension in the form of question-answering could improve dialogue agents' performance and help knowledge graph construction (He et al., 2021; Li et al., 2021b). However, simple surface-level features are oftentimes not sufficient to extract valuable information from conversations (Qin et al., 2017). Instead, it is necessary to comprehend the semantic and pragmatic relationships that structure the dialogue, such as the

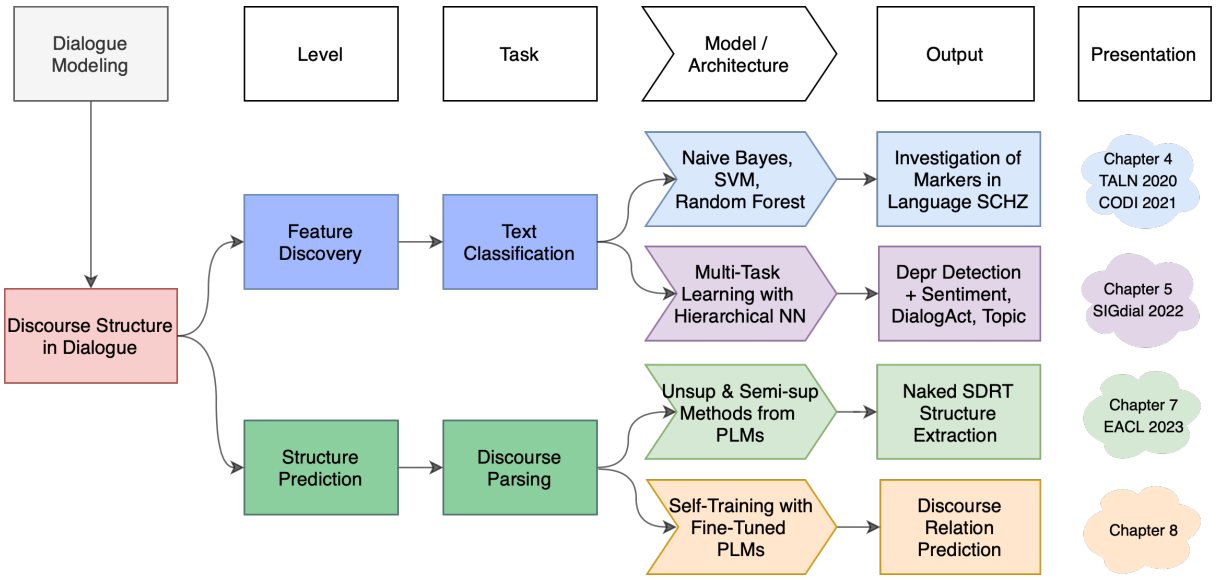


Figure 1.3: Thesis projects overview.

use of discourse information and coherence-relation structure. Consequently, we propose two **research questions** related to discourse analysis in dialogues:

RQ1 How can we effectively use discourse and structural information as linguistic features in text classification tasks for dialogue, such as mental disorder illness detection?

RQ2 How can we generate discourse structures with machine learning techniques using minimal supervision to achieve the greatest applicability in real-life scenarios?

Both research questions are approached with a few sub-projects. We illustrate the research objectives ("Level"), corresponding tasks ("Task"), employed methods ("Model / Architecture"), and outcomes ("Output") in Figure 1.3. Each project has typically one publication, and we intend to showcase them in their respective chapters ("Presentation").

To address RQ1, we pursue the first research line "Feature Discovery", which aims to investigate discourse in a general sense that is not limited to SDRT-style or RST-style structures. Initially, we focus on text classification tasks that involve the use of basic discourse markers like discourse connectives and dialogue acts. Two tasks are conducted in the cognitive impairment field: the first one is Schizophrenia detection (Chapter 4), which leads to two publications (Amblard et al., 2020; Li et al., 2021a) and a few communication talks including French national *Health and Language Seminar*[6] and *Semantics and Pragmatics of Dialogue Workshop* (SemDial 2021)[7]; the second one is depressive detection (Chapter 5) with one international publication and presentation at SIGDial conference (Li et al., 2022). The cognitive impairment setting makes for a realistic situation. Today, approximately 0.5% of adults worldwide are affected by Schizophrenia. Depression's impact is even larger: around 4% of the world population and a higher rate in elderly people, according to the numbers reported by the World Health Organization. These mental illnesses manifest varied symptoms, among which there are linguistic disorders such as the *disorganized language* and *poverty in vocabulary* (Kuperberg, 2010a). NLP

[6]Journée commune AFIA-THL / ATALA - la santé et le langage.
[7]`https://semdial2021.ling.uni-potsdam.de/programme/`.

practitioners can leverage language disorders as a potential source of symptoms for linguistic analysis to gain insights into the disease and its manifestations. This, in turn, could aid in the early-stage detection of the disease and eventually provide assistance in its treatment. Current models for mental disorder detection, however, are far from ideal. The majority of research in this field relies on social media data (Benton et al., 2017; Mitchell et al., 2015; Birnbaum et al., 2017a; Guntuku et al., 2017), with a particular emphasis on lexical information. However, as highlighted in our study (Li et al., 2021a), these approaches have limitations in certain languages and could lead to biased results. Our objective is to develop more reliable and robust models, which prompts us to explore linguistic features that rely less on lexical information and instead leverage structural information. The outcomes of our investigation are gradual and unfold across two projects. Chapter 4 presents the first project of the language detection of Schizophrenia. We observe that lexical features, although highly accurate, exhibit heavy bias. As a result, we explore delexicalized features such as syntactic trees and less-lexicalized features such as discourse connectives. Further, in our exploration of dialogue structure modeling, we introduce diverse context window sizes to investigate the influence of context length. This approach not only enables us to expand the training instances but also serves as a partial remedy for the limited annotated data. By replicating state-of-the-art results, we confirm some previous observations regarding specific linguistic features present in the language of Schizophrenia. Following that, in Chapter 5, we delve into the detection of depression in dialogues. In the absence of gold discourse structures, we propose to incorporate discourse information into the Multi-Task Learning (MTL) framework by utilizing *shallow* discourse features, such as dialogue acts, from another annotated resource. We adopt a simple yet effective approach known as the *fully-shared* scheme, where hidden layers are shared across all tasks. To enhance the modeling of dialogue structures, we introduce a hierarchical structure within the MTL framework. Our approach achieves the highest performance compared to existing studies, validating the advantages of incorporating multi-level structural-aware model architecture.

To address RQ2, we establish a second research line called "Structure Prediction". Our work is grounded in the Segmented Discourse Representation Theory and utilizes the STAC corpus for practical implementation. We embark on a two-step approach to tackle this research line. The first step is naked discourse structure prediction, presented in Chapter 7. It is worth noting that naked structures have been demonstrated to be valuable features for specific tasks, such as content selection (Louis et al., 2010) and thread extraction (Jiang et al., 2020). It is a collaborative effort with colleagues at the University of British Columbia in Vancouver during my internship at the UBC NLP group. This project results in a publication at the EACL 2023 conference (Li et al., 2023) and a presentation at the 4^{th} Workshop on Computational Approaches to Discourse (CODI 2023). Unlike previous studies that rely on full supervision (Afantenos et al., 2015; Shi and Huang, 2019; Chi and Rudnicky, 2022), our goal is to perform discourse parsing with less human-annotated data, so that our parser can be used in more general cases. However, selecting the appropriate distant or weak supervision signals is not an easy feat. In view of the promising findings of discourse information captured in pre-trained language models (PLMs), as introduced in Chapter 6, we ultimately choose PLMs as the source of supervision. We explore various PLMs and discover that the encoder network in BART model (Lewis et al., 2020) performs the best. We also propose fine-tuning tasks tailored to dialogues to enhance discourse information encoded in attention matrices, without requiring additional annotation. Subsequently, in Chapter 8, we carry out a study on discourse relation prediction based on the structure extracted from PLMs. Taking cues from the approaches outlined in Nishida and Matsumoto (2022), we employ bootstrapping strategies via self-training. Using a few annotated documents, we first train a source model and then use it to generate pseudo labels on unannotated data. High-confidence

pseudo-labeled instances are selected and combined with the original documents for a new round of model training. Through iterative self-training, we obtain a model that achieves optimal accuracy and class coverage. In addition, we provide complete parsing results that combine EDU segmentation, structure prediction, and relation classification, thereby establishing the first benchmark for a full discourse parser for dialogues trained using weak supervision. The empirical findings demonstrate a gradual advancement, albeit modest, that signifies the pioneering nature of this project in the expectation of full discourse parsing in dialogues.

1.4 Thesis Organization

The thesis is organized into three parts. **Part I**, titled "Discourse Analysis Foundations", focuses on discourse theories and state-of-the-art models for discourse parsing. It comprises two chapters.

In Chapter 2, we briefly introduce basic elements of discourse analysis, followed by a presentation of two influential discourse theories: Rhetorical Structure Theory (Mann, 1984) and Segmented Discourse Representation Theory (Asher and Lascarides, 2003). These theories have inspired the creation of many discourse corpora, including RST-DT (Carlson et al., 2002a) and STAC (Asher et al., 2016), the latter being the primary corpus for our experiments. We also explore other discourse corpora, including the PDTB framework (Prasad et al., 2008a) and datasets constructed under other frameworks. To address the concerns about the quality of discourse annotation, we conduct a detailed examination of the recent SDRT-style corpus Molweni (Li et al., 2020). Finally, we expand the discourse analysis discussion to the specificities of discourse in different language settings, including distinctions between spoken and written language, as well as between monologues and dialogues.

Chapter 3 explores the existing discourse parsing models, typically under the RST and the SDRT frameworks. In the past decade, supervised methods with graph-based or transition-based parsing paradigms have been commonly used. They are trained and tested in the same domain and contribute the state-of-the-art performances on corpora such as RST-DT and STAC. In recent years, transfer learning strategies have shown rapid development, but most semi-supervised and distantly-supervised methods have only been applied to monologues. We study these models in great detail by comparing their architecture and training process. Lastly, we focus on the practical applications of discourse in Natural Language Understanding (NLU) and Natural Language Generation (NLG) tasks. We conclude this chapter by discussing the current state of discourse usage in downstream applications and offering insights for the development of future discourse-aware models.

Part II, titled "Discourse Structure Discovery", is dedicated to addressing the first research question (RQ1). It showcases two text classification projects in two chapters.

Chapter 4 discusses the task of classification of the language of Schizophrenia. We start by analyzing the existing work in the field, along with the drawbacks of current models due to lexical biases and the limited size of the dataset. To address these challenges, we propose various strategies, including better data representation and dialogue structure modeling. The experiments are performed on a small French corpus derived from the SLAM project. Our results demonstrate the effectiveness of delexicalized and less-lexicalized features in building more robust models. Moreover, we conduct an extensive analysis of lexical, syntactic, discourse, and dialogue features in the context of the language of Schizophrenia.

Chapter 5 delves into the challenge of detecting depression in dialogues. In this chapter, we place a greater emphasis on enhancing dialogue structural modeling. We discover a larger, publicly available corpus DAIC-WOZ (DeVault et al., 2014) that allows for comparisons with existing

work. Our design is a hierarchical architecture that encodes interactions in a dialogue. We test this model under a Multi-Task Learning framework, allowing us to learn from disease-related information, such as sentiment and dialogue acts. Our model attains the highest performance compared to existing approaches, establishing a new state-of-the-art. Our analysis highlights the crucial role of integrating structural information and discourse-relevant signals.

Part III, titled "Discourse Structure Prediction", tackles the second research question (RQ2) and aims to construct a full discourse parsing pipeline with minimal supervision. This part contains three chapters.

In Chapter 6, we provide an introduction to Pre-trained Language Models (PLMs) and delve into the field of "BERTology", which focuses on studying the inner workings of Transformer-based models. This chapter serves as a foundation for the subsequent chapters, as we utilize PLMs as the backbone to learn and extract discourse information. We also explore various studies in the discourse field that employ probing tasks or self-supervised learning to extract discourse information, thereby establishing the context for our research.

Chapter 7 introduces our innovative approaches for extracting naked discourse structures from the attention matrices of PLMs. This step is crucial in the development of a complete discourse parser. We begin by providing a detailed presentation on the selection of PLMs, discourse tree inference methods, and semi-supervised and unsupervised strategies for identifying the most discourse-rich attention heads. Our experiments on the STAC corpus yield promising results, even with a small annotated dataset of only 50 documents. We then conduct a comprehensive analysis to investigate various factors that influence model performance, for instance, document length and the distance between `EDUs`. To assess the generalization of our approach, we also evaluate our model on the dialogue portion of the GUM corpus (Zeldes, 2017), albeit with less satisfactory results. It is worth mentioning that GUM uses a different annotation framework (RST) and it contains monologue-like conversations. In a deployment scenario, we utilize predicted `EDUs` instead of gold-standard ones for link prediction, aiming to evaluate the performance of our model under realistic conditions. Lastly, we explore methods to extend the tree structures into graphs and present modest improvements.

In Chapter 8, we present our experiments on relation prediction as a second step towards achieving full discourse parsing. Our approach involves utilizing a self-training strategy inspired by the work of Nishida and Matsumoto (2022). We employ a BERT-base model, fine-tuned with small-size annotated relation data. Despite its simplicity, leveraging PLMs as backbones has proven effective in capturing implicit relations Shi and Demberg (2019). We conduct an in-depth analysis to evaluate the impact of iterative self-training on relation prediction. Our findings reveal that this approach improves the performance of infrequent relations, although it necessitates careful tuning. We also explore the trade-off between the model's reliability and coverage, investigating how different strategies can balance these two aspects. In addition, we delve into the potential benefits of incorporating human feedback to further enhance the performance of our model.

Part I

Discourse Analysis Foundations

Part I of this thesis focuses on establishing the foundation for discourse analysis, which is divided into two chapters. In Chapter 2, we cover the fundamental elements and theories of discourse analysis, as well as annotation projects inspired by those theories. We discuss discourse in general, as well as in specific scenarios, such as different language devices (monologues versus dialogues, spoken versus written), to provide a general understanding of this topic. This chapter offers a step-by-step explanation of discourse processing for those unfamiliar with the concept, while for those who are, we hope that it serves as a summary and refresher.

In Chapter 3, we shift our focus to a specific task in discourse processing, i.e., discourse parsing. We provide an overview of the most commonly used methods for automatically extracting a discourse structure (SDRT-style graph or RST-style tree) from full documents. Due to the scarcity of annotated data, various Machine Learning methods have been proposed, mostly focusing on supervised learning strategies. Other strategies include transfer learning and weakly supervised learning to tackle insufficient learning examples. We conduct a thorough analysis of different strategies for discourse parsing and include pointers to our own experiments where applicable. For readers who are already familiar with these studies, they may utilize these references to proceed to our contributions in part III. Finally, at the end of the chapter, we expand the discourse parsing discussion to include its application in downstream tasks. We explore how discourse knowledge can be beneficial for NLP tasks such as summarization and sentiment analysis, and provide a comprehensive summary of current state-of-the-art discourse-aware models. We also provide suggestions on how to enhance the integration of discourse information in these tasks. Although we do not conduct any experiments on downstream tasks in this thesis, it is an intriguing topic for future research.

Chapter 2

Discourse Theories & Corpora

Contents

Discourse refers to the use of language by humans in a variety of contexts, such as essays, conversations, speeches, and more. It can take different forms, including spoken or written, monologues or dialogues, and can appear in various domains such as online technical forums or news articles. Regardless of these different expressions, the term *discourse* refers to the organization of language in a context. The objective of discourse analysis is to reveal the structural organization of language and understand how sentences interact with each other in order to give a plausible interpretation of communicative goals.

In this chapter, our focus is on the theoretical background of discourse analysis. We begin by presenting some key elements in discourse analysis in Section 2.1. **Discourse Units** (`DUs`) are the building blocks of discourse analysis, and refer to spans of texts that serve as the basic information carrier. The smallest discourse units are known as *Elementary Discourse Units* (`EDUs`), and the composition of `EDUs` creates intermediate discourse units, called *Complex Discourse Units* (`CDUs`) in some theories. **Discourse connectives** provide important clues for discourse relations. For instance, the word *but* shows strong evidence of the relation *Contrast* relation, and *because*

demonstrate an *Explanation* relation. Different formalisms interpret what kind of relationship should be established in discourse, with some being intention-based and stressing communicative goals, while others are semantic-based and use states and event descriptions. Following the presentation of the basic elements in discourse analysis, we delve into two major discourse theories in Section 2.2: Rhetorical Structure Theory (RST) (Mann, 1984; Mann and Thompson, 1987) and Segmented Discourse Representation Theory (SDRT) (Asher, 1993; Lascarides and Asher, 1993; Asher and Lascarides, 2003). These theories are widely recognized as the most influential in **full discourse analysis**, which involves constructing the global structure of a document, as opposed to **local discourse analysis** (also known as "chunking" or "chunk parsing") such as the Penn Discourse Treebank- (PDTB-) style analysis. In Section 2.3, we discuss several annotation projects influenced by various formalisms, such as RST and SDRT, along with other frameworks. These annotated corpora are essential for training automatic discourse parsers, providing valuable examples and patterns for machine learning. Additionally, we draw attention to potential annotation issues in a recently released corpus Molweni (Li et al., 2020). Finally, in Section 2.4, we conclude the chapter by comparing discourse across different language settings (e.g., monologues vs. dialogues) and reviewing recent attempts to adapt existing discourse theories to new scenarios (e.g., from written to spoken language).

2.1 Basic Elements in Discourse Analysis

We start by introducing the basic elements of discourse analysis, including *Elementary Discourse Units*, discourse connectives, and discourse relations. Connectives and discourse relations are useful for identifying important information in a text. For example, an *Elaboration* relation signifies a more detailed explanation of a given statement, while an explicit connective like *but* typically indicates a *Contrast* between two text spans, with more emphasis on the second. Some frameworks also provide information on the relative importance of elements in addition to discourse relations, which could be useful for downstream NLP tasks like text summarization.

2.1.1 Discourse Units

Given a two-sentence text, as shown in example (1), how many discourse units are there? This might seem to be an easy question at first sight, but has arose lots of discussion in the field of computational linguistics since the 80s (Grosz and Sidner, 1986; Polanyi, 1988; Hobbs, 1979; Mann and Thompson, 1988; Passonneau and Litman, 1997). Before addressing this question, it is important to establish a definition of what constitutes a discourse unit. These segments, largely known as *Elementary Discourse Units* (`EDUs`) or *Basic Discourse Units* (`BDUs`, in Polanyi (1988)), are the building blocks for discourse analysis. For written text, it is basically taken for granted that sentence boundary is also `EDU` boundary, which means that `EDUs` do not span across sentences (Stede, 2011). So, the question becomes whether a sentence should be further divided into smaller units.

(1) [But he added:]$_1$ ["Some people use the purchasers' index as a leading indicator,]$_2$ [some use it as a coincident indicator.]$_3$ [But the thing it's supposed to measure]$_4$ [– manufacturing strength –]$_5$ [is missed altogether last month."]$_6$

wsj_0627 in RST-DT (Carlson et al., 2002b)

Taken from the guidelines of Rhetorical Structure Theory (RST) (Mann and Thompson, 1988), annotators of RST identified 6 discourse units (bracketed units sub-scripted with numbers)

in example (1). However, in the Linguistic Discourse Model (LDM) (Polanyi and Scha, 1984; Polanyi, 1988), authors would probably argue that segments $4, 5, 6$ together form one basic unit so they would end up with 4 discourse units in total. This is because, in the LDM, proper discourse units are only those that can be *independently continued* in the subsequent discourse. So in this example, the interpolation segment "– manufacturing strength –" can not be considered as a separate segment.

Other commonly seen linguistic problems when defining the criteria for `EDUs` include the treatment of various kinds of ellipsis, relative clauses (example (2)), and prepositional phrases (example (3)).

(2) a. [The car that was red]$_1$ [narrowly won the race.]$_2$

b. [The red car,]$_1$ [which my friend had bought last week,]$_2$ [narrowly won the race.]$_3$

Example 4.14 and 4.15 in Discourse Processing (Stede, 2011)

For relative clauses, the restrictive relative clause ("that was red" in (2)a) serves **only** to identify a referent which often can be paraphrased with an adjectival modifier (identical to "the red car"), so that we would not want to separate this clause from the noun phrase "The car". In the case of non-restrictive clause ((2)b), on the other hand, "which my friend had bought last week" provides a new information piece and thus a new discourse unit, so we could treat it as a separate `EDU`. Note that not all restrictive/non-restrictive relative clauses are easy to distinguish, and not all the theories agree on treating non-restrictive relatives clauses as separate `EDUs` – as the case in Polanyi et al. (2004).

(3) [Tom is late]$_1$ [because of the rain.]$_2$

Researchers have different perspectives when it comes to prepositional phrases. According to Tofiloski et al. (2009), every `EDU` must contain a verb. Therefore, in example (3), the second part should not be separated. However, in the RST-DT annotation guideline (Carlson et al., 2002b), it is a classic case with two `EDUs` and an inner-sentential relation of causality.

In the Linguistic Discourse Model (Polanyi et al., 2004), *discourse segments* are defined as "syntactic constructions that encode a minimum unit of meaning and/or discourse function interpretable relative to a set of contexts". The "minimum unit of meaning" communicates information about no more than one event, event type, or state of affairs, and the "minimal functional unit" encodes information about how it relates structurally, semantically, internationally, or rhetorically to other units in the discourse or to the extra-linguistic context. Based on *discourse segments*, basic discourse units are then identified, which are "discourse segments of a type that can be independently continued". Operator segments that are heavily integrated into other nominal (such as the interpolation in Example (1)) or verbal constructions cannot be accessed for independent continuation, thus not a `EDU`. A more loose definition of `EDUs` consider that these small units convey a minimum unit of meaning, similar to "words" in syntactic analysis, as in RST (Mann and Thompson, 1988) which do not stress on the independent continuation property. The Penn Discourse Treebank (PDTB) model (Prasad et al., 2008a) does not incorporate the concept of discourse units. Instead, it employs the notion of *arguments*, which are text segments that express discourse relations. This distinction arises from PDTB's unique annotation process, which involves first identifying discourse relations and then identifying segments. Essentially, Stede (2011) proposed a general definition of `EDU` as follows:

Definition 1 (Elementary Discourse Unit (EDU)) *A span of text, usually a clause, but in general ranging from minimally a (nominalization) NP to maximally a sentence. It denotes a single event or type of event, serving as a complete, distinct unit of information that the subsequent discourse may connect to. An EDU may be structurally embedded in another.*

From the discussion above, we realize that it is very difficult to reach one precise definition of discourse segments. Different theories have different reasoning for linguistic phenomena and thus different criteria to segment discourse units. Most of them are not even clearly described (Braud, 2015). It is clear, however, that a discourse segment must serve a specific purpose in relation to the other parts of the text. These semantic and/or pragmatic functions determine the relationships that are established between these segments. As the initial step in discourse analysis, the segmentation of discourse units should be performed in an objective and impartial manner, in order to avoid any potential bias in subsequent processing.

2.1.2 Discourse Connectives

Having looked at elementary discourse units, we now present signals that help to identify the coherence relations, typically, the connectives. Words (or multiwords) such as *because*, *but*, *although*, and *in contrast* provide signals for new pieces of information, and indicate how they link with the previous ones: seeing *because*, readers expect an explanation to follow; reading *but*, readers anticipate a contrast and often times expect a more crucial information to come.

What do connectives link and how do they contribute to the interpretation of text? In Rouchota (1996), the author presented two frameworks to explain the semantic and pragmatic properties of connectives: coherence-based framework and relevance-theoretic framework. The former follows the line of research of Mann and Thompson (1988); Fraser (1990); Sanders et al. (1993); Knott and Dale (1994), stressing coherent text and how these "cue words" can make such discourse relations explicit. The latter – relevance-theoretic approach – focuses on communicative purpose and how connectives can encode procedural information, as supported by Blakemore (1987); Wilson and Sperber (2012). In this section, we mainly focus on the *coherence* aspect of connectives and discuss how connectives can help identify coherence relations.

We consider that discourse connectives form a closed set and can be of different morpho-syntactic categories. Different languages have different ways of relation expression, thus the size of connective inventories varies. PDTB (Prasad et al., 2008a) for instance, contains around 100 forms of English connectives and is classified into 3 sense hierarchies (i.e. relation). Note that the modified forms of connectives are treated as belonging to the same type as the unmodified forms. It annotates both explicit connectives – including subordinating conjunctions (e.g., "when", "although", "if"), coordinating conjunctions (e.g., "and", "but"), and adverbial connectives (e.g., "however", "therefore"), and implicit connectives which are identified between adjacent sentences that are not related by an explicit connective[1]. German connective lexicon DiMLex (Stede and Umbach, 1998) was constructed around 170 frequent connectives. For Chinese, since the morphological forms of connectives are more flexible[2], 282 Chinese connectives are annotated on the

[1]The annotation of implicit connectives is intended to capture discourse relations that are implicitly expressed between adjacent sentences. During annotation, annotators were asked to provide an explicit connective that can best describe the relation. In this thesis, we coarsely characterize discourse relations into two types: explicit connectives and implicit connectives. But this classification is not accurate. Note that there are cases where an implicit connective could not be provided. In PDTB 2.0 annotation guideline, "AltLex", "EntRel" and "NoRel" are used for these cases (Prasad et al., 2008a).

[2]Chinese connectives can contain more than one word and can be discontinuous. For example "不是...而是(is not... is)" or even paired connectives such as "因为...所以(because...so)".

Chinese Discourse Treebank (CDTB) (Xue et al., 2005; Li et al., 2014d). LexConn (Roze et al., 2012) is a French connective inventory that contains 328 forms of connectives. Subsequently, French Discourse Treebank (FDTB1) (Steinlin et al., 2015) was created based on LexConn and it gathered more than $10k$ connectives corresponding to 353 forms, much larger than that in English. Other connective lexicons include for instance Spanish (Alemany et al., 2002), Czech (Mírovskỳ et al., 2017), and Italian (Feltracco et al., 2016). Very recently, an effort towards multilingual lexicon resources for connectives has been put forward by Stede et al. (2019), where an online discourse database Connective-Lex displays the existing and newly-created lexicons in 13 different languages[3].

Apart from the forms of connectives, each connective can present in different positions of the sentence, depending on their syntactic role. Take English connectives as an example, discourse connectives could take different morpho-syntactic forms, such as coordinating conjunction (*and, but*), subordinating conjunction (*if, because*), discourse adverbials (*however, since, consequently*), nominal phrases (*the reason, as a result of*), or even some verbs (*cause*). Due to their syntactic nature, they can occur at different places in a sentence (Rouchota, 1996). Conjunctions, for instance, can only occur at the beginning of the clause they introduce: *but* is the prototypical example in this category, where it can only appear at the beginning of the sentence ((4)a) and not elsewhere ((4)b):

(4) a. John bought a house, *but* he is not happy.
b. John bought a house, he is, *but*, not happy.*

Other connectives, such as adverbials, have more flexible syntactic properties so that they may occur at the beginning, middle, or end of the sentence. A similar example by replacing *but* to *however*:

(5) a. John bought a house. *However,* he is not happy.
b. John bought a house. He is, *however,* not happy.
c. John bought a house. He is not happy, *however.*

In principle, connectives are very useful indices for identifying coherence relations. However, recognizing the form of a connective is not sufficient. For one point, some relations are always implicit, meaning that they are simply not lexicalized, such as *Frame* in ANNODIS corpus Afantenos et al. (2012a). This relation describes a relationship that links a detached adverbial at the head of a proposition, introducing a frame that localizes a situation temporally or spatially, and the segment to which this frame relates. An example in Braud (2015): "The next day, Mr Pitoun was found safe and sound.", where the relation between the two text spans can not be lexicalized.

For lexicalized relations, sometimes, explicit connectives can also be omitted. For instance, in the following example, the omitted discourse connective is *however*, and the discourse relation (or *sense* in the PDTB terminology) is *Comparison.Contrast*. We can manually add back the connective. However, it is worth noting that this action could, sometimes, modify the sense or remove the existing sense. In example (6), inferred connective *however* does not contain the idea of a temporal succession between the sentences.

(6) ["Kemper is the first firm to make a major statement with program trading."]$_{arg1}$ He added that ["having just one firm do this isn't going to mean a hill of beans."]$_{arg2}$

[3] `http://connective-lex.info`

wsj_1000 in PDTB (Prasad et al., 2008a)

For another point, the form of connectives can be ambiguous at two levels: first, they can be used in *discourse-usage* or *non-discourse-usage* settings. One word with the form of a connective is not always employed for discourse use. For instance, *and* is a conjunction connective when it links the propositions (in (7)b *and* implies *continuation* relation), but it not when it coordinates nominal words (in (7)a). Another example is the word *once*: it can be either a temporal discourse connective or simply a word meaning "formerly".

(7) a. Lithuania, Latvia *and* Estonia thus open themselves to the multiparty system.

b. The CGT transport federation have risen against "the lack of consultation" *and* consider that employees have "nothing positive to expect from this restructuring."

FDTB1 (Danlos et al., 2012), translation provided by Laali and Kosseim (2017)

Secondly, discourse connectives may be used to signal more than one discourse relation. For example, the word *since* can serve as either a temporal or causal connective. In the shared task CoNLL-2016 (using PDTB-2.0 dataset (Prasad et al., 2008a)), we find connectives with more than 8 senses, such as *then*, *as*, *when*, and *but*, making the sense classification very difficult. There exists a fruitful line of research on the disambiguation of discourse connectives: pioneered by Pitler and Nenkova (2009) where authors proposed to use syntactic features and connectives themselves, with very promising results on PDTB. Follow-up works such as Lin et al. (2014) further increased the results using contextual and lexico-syntactic information. There has been relatively little research on connective disambiguation in languages other than English, likely due to the lack of annotated corpora. Nonetheless, some work has been done in French (Laali and Kosseim, 2017) and in Arabic (Al-Saif and Markert, 2011).

Despite the ambiguity property in discourse connectives, there are works investigating plausible semantic sense applicable to a particular connective. Typically, we notice the work by Sileo et al. (2020) where authors used a pre-trained model to predict discourse markers with known semantic relations such as discourse relations and sentiment, and study the linkage between discourse markers and relations. They showed association patterns between discourse connectives and semantic categories in discourse corpora such as PDTB, STAC (Asher et al., 2016), GUM (Zeldes, 2017) and several corpora in Natural Language Inference (NLI) tasks, revealing inconspicuous but sensible discourse markers for discourse relations. For example, in example (8), the relation *Contradiction* can be expressed with any of the following markers: *in contrast*, *initially*, and *curiously*. At first glance, the association of *initially* and *curiously* with *Contradiction* might seem surprising, as one would typically link *initially* to *Explanation* or *Background* relations, and *curiously* to *Elaboration* or *Explanation* relations. However, upon closer examination of the context, they do seem like reasonable matches. A similar case goes for *seriously* as a marker of *Sarcasm*. This work provides a novel approach to semantic analysis, utilizing unsupervised methods on a large scale. Rather than relying on established discourse connective sets, the authors investigated a wider range of potential connections between discourse relations and connectives. This approach led to an expanded set of possible associations, although some of these require individual examination and may not be widely applicable[4].

[4]DiscSense is publicly available at `https://github.com/synapse-developpement/DiscSense`.

(8) a. You will seldom meet new people, *in contrast*, in medellin you will definitely meet people.

b. If I burn a fingertip I'll moan all night. *Initially,* it didn't look so bad.

c. He puncture is about the size of a large pea. *Curiously,* he can see almost no blood.

Discourse markers with *contradiction* relation, in Discovery (Sileo et al., 2019).

2.1.3 Discourse Relations

With discourse units and connectives as clues, we can then use specific relations to link these units. In full discourse parsing theories (such as RST and SDRT), the linkage between two `DUs` is created in a recursive manner. In a text, two `EDUs` are connected to one another, forming a larger discourse unit (an internal node, or *complex discourse unit* (`CDU`) as we call it in SDRT), which in turn is also subject to relation linking. Recall the previous example (1):

(1) [But he added:]$_1$ ["Some people use the purchasers' index as a leading indicator,]$_2$ [some use it as a coincident indicator.]$_3$ [But the thing it's supposed to measure]$_4$ [– manufacturing strength –]$_5$ [is missed altogether last month."]$_6$

wsj_0627 in RST-DT (Carlson et al., 2002b)

When we process RST-style relation linking for this example, *elaboration* is the relation that links $\texttt{EDU}_4$ and $\texttt{EDU}_5$ since the interpolation "manufacturing strength" provides precision on "the thing it's supposed to measure", making them together a larger `DU`. *Contrast* relation attaches $\texttt{EDU}_2$ and $\texttt{EDU}_3$ since the two `EDUs` provide different opinions on "what is the purchasers' index". Between the two larger `DUs` ($\texttt{EDU}_2$-$\texttt{EDU}_3$ and $\texttt{EDU}_4$-$\texttt{EDU}_5$), a *contrast* relation can also be established. The result of relation attachment is an RST-style labeled tree, as shown in Figure 2.1.

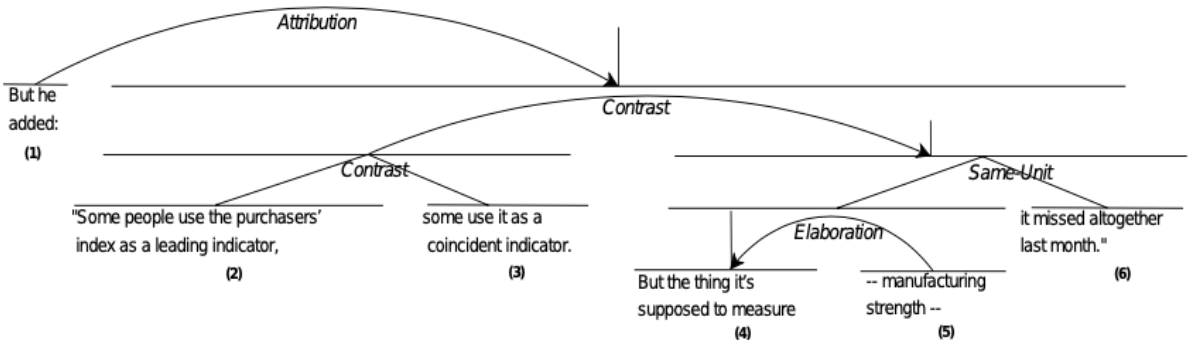


Figure 2.1: Discourse relation of a two-sentence example in RST, from Joty et al. (2015).

The concept of discourse structure is linked to the principle of coherence. Theoretical frameworks assume that all coherent discourse has a structure and aim to account for this coherence by describing the organization of discourse. Discourse relations are generally viewed as binary predicates, taking two discourse units. Depending on the theoretical frameworks, they are defined based on various criteria, which leads to different sets of relations, sometimes refined in extensions and in discourse corpora (such as RST and RST-DT corpus). The inventories generally include relations of temporal, causal, conditional, additive, and comparative types. Theoretical

frameworks and formalisms also differ in terms of the constraints imposed on the final structure of the document, and thus the way units are linked, possibly prohibiting certain configurations. Depending on the constraints, the final structure obtained is either a tree (as in RST) or a graph (as in SDRT).

In the three main theoretical frameworks, we find different sizes of relations. Initially, Mann and Thompson (1988) had suggested about 25 relations in the RST framework. When constructing the first large RST-style Discourse Treebank (RST-DT), Carlson et al. (2002a) used a much finer relation inventory: 53 mono-nuclear relations where one `DU` is more salient than the other one, and 25 multi-nuclear relations where two `DUs` are of equal importance. The relations are then grouped into 16 coarse-grained categories, see Carlson and Marcu (2001) for a detailed description of the relations. STAC (Asher et al., 2016) is the most commonly used corpus under the SDRT framework (Asher and Lascarides, 2003). It contains 16 relations. Different from the relations in monologues, dialogue-specific relations such as *question-answer-pair*, *comment*, and *acknowledgment* occupy a large portion of the STAC corpus. In the most popular local discourse analysis corpus PDTB (Webber et al., 2019), there are 3 sense hierarchies for discourse relations. The first level contains 4 coarse relations: *temporal*, *contingency*, *comparison*, and *expansion*. Level-2 provides precision on the sub-types of Level-1 relations, for instance, *concession* and *similarity* are sub-relations of *comparison*. Level-3 encodes the direction for asymmetric level-2 relations such as *concession*, *cause*, and *purpose*. Precisely, a sense relation R is symmetric if and only if $R(Arg_1, Arg_2)$ and $R(Arg_2, Arg_1)$ are semantically equivalent. If a relation is not symmetric, it is asymmetric. Although different discourse frameworks have varying relation types, there has been work trying to map discourse relations between frameworks. For instance, an interesting survey by Demberg et al. (2019) explored the compatibility of discourse relations in RST-DT and PDTB 2.0. The findings revealed that RST-DT and PDTB exhibit higher agreement on explicit relations (over 70%) than implicit relations (less than 50%). The ambiguity of connectives emerged as a significant source of disagreement for mapping, and some relations were inherently challenging to distinguish (such as *contrast* and *concession*), possibly due to different frequency of usage in RST-DT and PDTB.

An interesting follow-up question is: how to *infer* these semantico-pragmatic relations between discourse units?

Originated with Hobbs (1979), there is a line of research trying to model the reasoning process such as the Rhetorical Structure Theory, the Segmented Discourse Representation Theory, Discourse Lexicalized Tree Adjoining Grammer (DLTAG) (Webber et al., 2003), *etc.*. As presented in the previous section, explicit discourse markers such as discourse connectives can be of great help when recognizing discourse relations. Their role is similar to that of the "cue word" for topic segmentation and can be interpreted successfully once two discourse units are being set into correspondence with each other (Stede, 2011). On the other hand, with no overt linguistic signals, people need to reason about the most likely relation between two segments, possibly by inserting implicit connectives. This *inferential* process may be more challenging. In a comparison experiment conducted by Soria and Ferrari (1998), subjects were asked to identify the coherence relation in a text with and without connectives. The study tested three relations: *additive*, *cause*, and *contrast*, and revealed that recognition rates significantly decreased when connectives were absent: 73% → 64%, 89% → 60%, 83% → 43% for the three relations respectively. This is an intriguing discovery that suggests that while explicit linguistic signals aid in the comprehension and reasoning of discourse relations to a great extent, the context itself can provide sufficient information for readers to make accurate *inferences* even in the absence of connectives.

More recently, some shared tasks around discourse relation prediction have been proposed,

such as *Shallow Discourse Parsing* in CoNLL-2015 (Xue et al., 2015) and CoNLL-2016[5]. CoNLL-2016 used the Penn Discourse Treebank and Chinese Discourse Treebank as the shared task datasets to conduct shallow discourse parsing. The parsing task is referred to as "shallow" because the relations in a document are not connected to one another to form a connected structure. Starting from 2021, the CODI workshop has been organizing DISRPT shared tasks[6], including `EDU` segmentation, discourse connective identification, and discourse relation classification. These tasks are designed to accommodate various discourse frameworks and are applicable to multiple languages.

Lastly, we briefly present other indices that convey coherence relation. In the PDTB, relations can be expressed lexically by non-connective expressions. In this case, the label "AltLex" is assigned to indicate that adding an implicit connective to express an inferred relation results in redundancy. For example, in sentence (9), the phrase "mayhap this metaphorical connection" in bold indicates the relationship with the previous segment, and no additional connective is required.

(9) Ms. Bartlett's previous work, which earned her an international reputation in the non-horticultural art world, often took gardens as its nominal subject. AltLex **Mayhap this metaphorical connection made** the BPC Fine Arts Committee think she had a literal green thumb.

Example (7) in Prasad et al. (2008a)

Some verbs can also bring strong clues, such as "concede" for *Concession* relation and "cause" for *Cause* relation. Punctuation, such as dashes (–) or two colons (:), can succinctly express *Explanation* and *Cause* relations. The presence of numbers, such as money or percentages, or comparative lexicons ("stronger", "better", etc.) indicate comparative relationships; dates, days of the week, or months can show temporal relationships (Braud, 2015).

2.2 Different Views of Discourse Analysis

So far we have presented the notions of elementary discourse unit, coherence relation, and explicit signals for relation recognition – discourse connectives. These basic ingredients provide local information about relations between text spans. We are aware that in a text, the linear order of text spans is not arbitrary. Rather, it reflects an underlying logic. To examine the inner coherence in a larger context, we now move from atomic elements to hierarchical structures. Different theoretic frameworks have been proposed to study discourse at the document level. These frameworks aim to define the nature of structures that glue a document together. We present here two frameworks that led to the annotation of corpora at the discourse level: Rhetorical Structure Theory (RST) in Section 2.2.1 and Segmented Discourse Representation Theory (SDRT) in Section 2.2.2.

2.2.1 Rhetorical Structure Theory

The Rhetorical Structure Theory (thereafter RST) is a theory that describes a text by assigning a structure to it. It was first proposed by Mann and Thompson (1987, 1988) and enriched by the

[5] `https://www.cs.brandeis.edu/~clp/conll16st/intro.html`

[6] `https://sites.google.com/georgetown.edu/disrpt2021`, `https://sites.google.com/view/disrpt2023/`

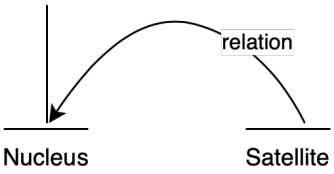


Figure 2.2: The generic RST schema.

work of Marcu (1997). It has been largely influenced by Grimes and Grimes (1975); McKeown (1985); Mann (1984). With the original inception for text planning (i.e. generation), RST was also extended in various applications in computational linguistics, cross-linguistic studies, dialogues, and multimedia settings (Taboada and Mann, 2006). RST was initially created for just one purpose: text organization (Mann, 1984; Mann and Thompson, 1987). In order to create a comprehensive theory of text organization while limiting immediate task complexity, Mann and Thompson developed two parts of RST which stand for two theoretic meanings: descriptive and constructive. A *Descriptive RST* gives almost all small published texts an RST analysis by showing what relations are essential in composing text and how they are linked together. RST analysis is informative about the phenomena of clause combining, conjunction, and related issues about forms and functions. It also has rich knowledge representation. On the other hand, a *Constructive RST* is the basis of an autonomous computational text planner. It goes beyond the descriptive theory by describing an approach for structure synthesis. It can mimic part of the generation of natural texts and produce appropriate structures. The building process is composed of multiple "Oracles" blocks, such as belief oracle, evidence proposing oracle and evidence supporting oracle.

RST Schemas: Descriptive RST is built upon the elementary analysis units, called *schemas*. A generic *schema* is diagrammed in Figure 2.2. It indicates how a particular unit of text is decomposed into multiple components: the two horizontal lines represent two text spans and are linked together by a curved line with *relation*; the vertical line points to one of the text spans which is called *nucleus* while other text spans are called *satellites*. Conceptually, a schema is an abstract pattern that depicts the constituency arrangement of text. They are loosely analogous to grammatical rules (Mann and Thompson, 1988).

In most cases, a schema takes two text spans (one satellite and one nucleus), as in Figure 2.3(a). The majority of both schemas and schema applications follow this pattern. Multiple satellites and one nucleus are also allowed in RST, such as the *Inform* schema in Figure 2.3(d), with the middle text span being the nucleus and two satellites set aside. Another schema type *Request*, containing *motivation* and *enablement* as relations, shares a similar pattern. Figure 2.3(b) is a multinuclear schema to represent a few equal-importance text spans. Here, *Contrast* schema has exactly two nuclei. *Sequence* (Figure 2.3(e)) and *Joint* (Figure 2.3(c)) schemas, on the other hand, have indefinitely many elements. We do not use narrowed curves in the latter case but simply attached curves for multi-nuclei. Depending on the relation types, a satellite may appear on the left or right side of the nucleus. Schema names are the same as the corresponding relation names. We use uppercase for the first letter in schema name and lowercase in relation name, for distinction. In order to initiate an instance of schema, the nucleus must be present, but all satellites are optional. Conventionally, the creation of schemas is not restricted to certain orders, while the analysis of a text takes the left-to-right order when applicable.

A short text is given with a full rhetorical RST structure in (Mann, 1984). The text is an

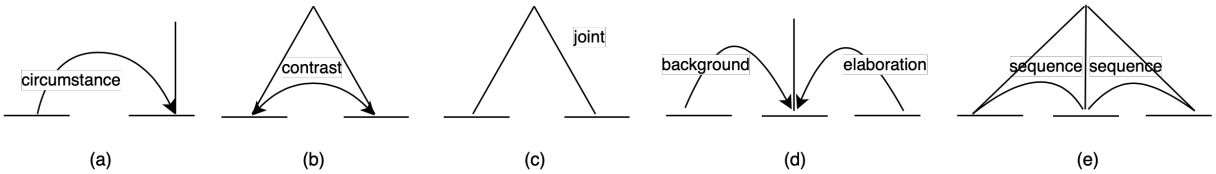


Figure 2.3: Five schema types in RST.

advocacy article from a political magazine. For readability, we only take the first part of the article (example (10)) and show its structure in Figure 2.4.

(10) [I don't believe that endorsing the Nuclear Freeze Initiative is the right step for California CC.]$_1$

[Tempting as it may be,]$_2$

[we shouldn't embrace every popular issue that comes along.]$_3$

[When we do so]$_4$

[we use precious, limited resources where other players with superior resources are already doing an adequate job.]$_5$

[Rather, I think we will be stronger and more effective]$_6$

[if we stick to those issues of governmental structure and process, broadly defined, that have formed the core of our agenda for years.]$_7$

[Open government, campaign finance reform, and fighting the influence of special interests and big money, these are our kinds of issues.]$_8$

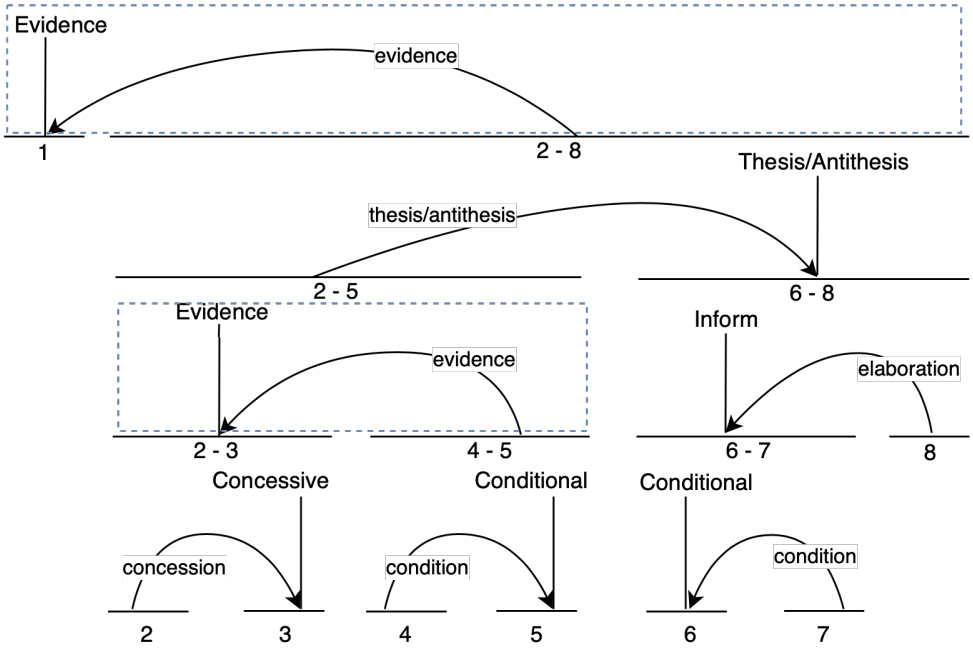


Figure 2.4: A RST diagram of a (partial) advocacy test, from Mann (1984).

The clauses are discourse units. In total, there are 7 applications of 5 different schemas: *Concessive, Conditional, Thesis/Antithesis, Evidence* and *Inform*. We will present the corresponding

relations in the next section (precisely for *evidence* relation, circled in dashed boxes). As we can see, the constituent tree is built in a bottom-up fashion. An analysis of such structure consists of recursive decomposition into intermediate units (such as [1, 2-8], [2-5, 6-8]) and finally, the smallest text spans (e.g.: [2, 3], [4, 5]).

RST Relations: Relation definition is at the heart of RST. A definition is given between two non-overlapping text spans (one *nucleus* (N) and one *satellite* (S)), and consists of four fields: (1) constraints on N; (2) constraints on S; (3) constraints on the combination of N and S; (4) the effect. The "effect" field shows how the application of such a relation could satisfy the writer's desire. It secures a valid analysis for a coherent text. When applying RST schemas, "effect" serves as a constraint against inappropriate use of relations. Mann and Thompson (1987) mentioned that descriptive RST is a functional account of a text since this analysis always tries to find out what the writer wants to convey in each part of the text. We retake the example (10) and show one example relation *evidence* in Table 2.1:

relation:	EVIDENCE
constraint on N:	Reader might not believe N to a degree satisfactory to Writer
constraints on S:	Reader believes S or will find it credible
constraints on N+S:	Reader's comprehending S increases reader's belief on N
effect:	Reader's belief of N is increased

Table 2.1: Definition fields in RST, example with *Evidence* relation. N = nucleus; S = satellite.

In RST, the application of schemas is recursive. In example (10), we apply twice the *evidence* relation: the first one is between spans [2-3] and [4-5] and the second one is between [1] and the rest of text spans [2-8] (circled in dashed boxes in Figure 2.4). In the first scenario, clauses [2-3] make a statement on "we should not embrace every popular issue that comes along" – which may not be believed by a reader at this point; the writer immediately proposes evidence (clauses [4-5]) with an argument on ineffective usage of "precious resources" to support her claim. As such, if the reader finds the evidence convincing, it will increase her belief in the initial statement. Similarly, in the second case, the writer uses the whole text span ([2-8]) as evidence to back up her first claim "endorsing NFI is not the right step".

Apart from *evidence*, *justify* satellite also intends to increase the reader's belief in the nucleus statement. Instead of providing a new piece of evidence, *justify* clarifies the statement and increases the readiness. Since the two relations share the same *effect*, they form a sub-group in the RST relation definition list (Mann and Thompson, 1988). We show the organization of relation groups in Table 2.2. Each group reassembles relations that share a number of characteristics and differ in other attributes. Take another example in the sub-group *Antithesis* & *Concession*: both relations aim to cause positive regard towards the nucleus. *Antithesis* reach the objective by using contrast, while *concession* does not. Meanwhile, we note that almost all relations are asymmetric. For instance, if span A serves as evidence of span B, then B is definitely not serving as evidence for A.

Another way of organizing RST relations is dichotomous division, such as semantic vs. pragmatic, ideational or non-ideational, etc. Mann and Thompson (1988) proposed a two-way distinction based on "subject-matter" and "presentational" aspects. In the "subject-matter" group, the relation satellite presents parts of the subject matter, through a causal effect (*volitional cause* or *non-volitional cause*), a new piece of information (*elaboration*, *solutionhood*), etc. Relations

Antithesis & Concession	Condition & Otherwise	Enablement & Motivation
Antithesis	Condition	Enablement
Concession	Otherwise	Motivation
Relations of Cause	Interpretation & Evaluation	Restatement & Summary
Volitional Cause	Interpretation	Restatement
Non-Volitional Cause	Evaluation	Summary
Volitional Result	Evidence & Justify	Other relations
Non-Volitional Result	Evidence	Sequence
Purpose	Justify	Contrast
Background	Circumstance	Elaboration
Solutionhood		

Table 2.2: Organization of relation definition in RST, from Mann and Thompson (1988).

in the "presentational" group, on the other hand, are used to facilitate the presentation process itself. The communicative goal is to increase readers' belief (*evidence* and *justify*), desire (*motivation*), or positive view (*antithesis* and *concession*) on the nucleus statement.

The set of relations in RST is not closed (Mann and Thompson, 1987). New relation types can be added; old ones can be subdivided or even manipulated to meet the needs of specific phenomena or new domains. The initial relation list in Mann's 1988 work has 23 relations (Table 2.2), and then enriched to 30 relations. Among these, *Joint* is a schema and not a relation.

Nuclearity: We have already mentioned *nucleus* and *satellite* in the schema section. They refer to the relative importance of two text spans in the RST. This characteristic is known as nuclearity. Nuclearity describes the hierarchical structure in a schema. In principle, combined text spans have different functions: one span is more prominent and essential than the others. It delivers the core information and appears at the higher level of schema. Such span is known as *nucleus* and the remaining ones are *satellites*. For schemas with multiple relations, there is a single *nucleus* that all other *satellites* are related to. Most of the relations in RST obey the single nucleus principle. However, multi-nuclear relations also exist: *Sequence* and *Contrast* are two typical relations with multi-nuclei, as shown in sentences (11)a and (11)b, respectively. Their schema types correspond to resp. (b) and (e) in Figure 2.3. The text spans have equal importance in sentence (11): each is a nucleus.

(11) a. [Animals heal]$_1$, [but trees compartmentalize.]$_2$

b. [1. Peel oranges,]$_1$ [2. and slice crosswise.]$_2$ [3. Arrange in a bowl]$_3$ [4. and sprinkle with rum and coconut.]$_4$ [5. Chill until ready to serve.]$_5$

Rhetorical representation theory website:
`https://www.sfu.ca/rst/01intro/definitions.html`

Interestingly, nuclearity seems not able to cover all text organization. In *enveloping structure* (i.e., letters and mails) for instance, the conventional openings and closing are not easily described with nuclearity. Similar in *parallel structures* such as recipes and product manuals, parallelism organization of text makes the nuclearity assignment less apparent.

Satellite before nucleus:	Antithesis, Background, Conditional, Concessive, Justify, Solutionhood
Nucleus before satellite:	Elaboration, Enablement, Evidence, Purpose, Restatement

Table 2.3: Canonical orders of *satellite* and *nucleus* in some relations.

Naturally, the concept of nuclearity is closely linked to relations. We have noted that most relations are **asymmetric** in RST. If A is the consequence of B, then B is not the consequence of A. These asymmetries form a pattern which is represented in the relation definition. Thus, the assignment of *nucleus* and *satellites* is generally a by-product of relation labeling.

As we indicated above, the schema and relation do not constrain the order of spans in the text. However, in practice, there are strong tendencies of frequent ordering of spans for certain relations and thus, the relative position of *nucleus* and *satellite*. For instance, *satellite* usually appears before *nucleus* in relation *Background*; while *nucleus* appears prior to *satellite* in *Elaboration* relation. In Mann and Thompson (1988), authors present the canonical ordering for some relations (Table 2.3).

RST Construction: The RST construction process is in fact the recursive application of schemas on the whole text to obtain a tree-shape structure. Mann (1984) presented 7 application conventions, from which we can resume four constraints (Braud, 2015): First of all, one schema should be instantiated to describe the entire text: this is the convention for **completeness**. The **connection** constraint requires that each text span must be connected to at least another span, either an elementary unit or an intermediate unit built with smaller spans. However, the schema does not constrain the order of the nucleus or satellite. Thirdly, one schema must contain a nucleus but allow multiple satellites. Only one relation type is allowed between a nucleus and a satellite. This constraint is called **uniqueness**. Lastly, as already shown in Figure 2.4, all schemas are constructed within **adjacent** text spans. The constraints, however more or less strict, later became the targets of criticism. Uniqueness and adjacency particularly pose problems (Taboada and Mann, 2006). For instance, adjacency is abandoned in Segmented Discourse Representation Theory (SDRT) in order to cover long-distance relations between text spans.

2.2.2 Segmented Discourse Representation Theory

The Segmented Discourse Representation Theory (SDRT) (Lascarides and Asher, 1993; Asher, 1993; Asher and Lascarides, 2003) is a dynamic representation theory of discourse extended from the Discourse Representation Theory (DRT) (Kamp, 1981; Kamp and Reyle, 2013). Different from the RST (Mann and Thompson, 1988) which focuses on an intention-based approach that models the communication goals, SDRT favors a semantic-based approach using states or event description (Amblard and Pogodalla, 2014), similar to the Linguistic Discourse Model (LDM) (Polanyi, 1988). As we know, discourse analysis is the analysis applied at the document level. It distinguishes from the sentence-level analysis since the semantic content of a sentence is not necessarily the same as that in a larger context. In other words, there exists the notion of **dynamics** which underlies the content of discourse. Precisely, imagine we have some elements and knowledge that are already established in a given world (i.e., context), such as a person, her name, her status, an ongoing event, etc. When a new context (i.e., new sentences) is introduced, new elements will, in turn, access and modify the old world while maintaining coherence; meanwhile, make the current world accessible for future new contexts. This process can continue infinitely.

We need to dynamically mimic the movement and the impact that comes along. SDRT is one of the frameworks that enable such dynamic modeling of discourse. To make the new states of the world understandable, some elements should remain accessible. This brings up the second crucial feature in discourse analysis – **coherence**. In SDRT, such coherence is obtained via a structure of rhetorically connected propositions. We say that discourse is coherent in case (a) every introduced proposition is rhetorically connected to another piece of information, and (b) all anaphoric expressions can be resolved (Asher and Lascarides, 2003).

In this section, before diving into the SDRT framework, we first lay the ground by briefly revising dynamic semantics and the predecessor of SDRT: Discourse Representation Theory (DRT). We then present the relations and structures in SDRT.

Dynamic Semantics: In semantics, the meaning of a sentence derives from the meaning of its parts and how they combine together. This principle is known as **compositionality**, sometimes called *Frege's Principle* (Pelletier, 1994), pioneered by Frege (1988). It has had a tremendous impact on modern linguistics ever since Montague Grammars became known (Montague, 1970, 1973). For instance, take the following example from Asher and Lascarides (2003):

(12) a. A man walked in.

This example is easy to process if we know the meaning of "man" and "walk". Once combining them, we understand that a male person does an action which is walking. The temporal description further tells us that this action has already been accomplished. Now, if we add another sentence right after (12)a, the discourse becomes:

(12) a. A man walked in.

b. He ordered a beer.

A reasonable reader would naturally consider the pronoun "he" refers to the "man" in the first sentence. Therefore, the individual who walked in is exactly the same individual who ordered a beer. However, this reasoning poses problems for static semantics such as in Montague's or Davidson's (Davidson, 2001). Since in static semantics, the meaning of a sentence is merely the set of models it satisfies; it can not bridge cross-sentence anaphora. While in dynamic semantics, the meaning of a sentence is the application results between a set of prior contexts being proceeded and a set of posterior contexts that represents the content of the discourse including that sentence. Therefore, dynamic semantics is able to solve anaphoric resolution.

The basic idea of dynamic semantics is to develop a notion of context and of contextual interpretation so that the context for new information (such as (12)b) can take into consideration the material from its previous sentences ((12)a). Generally speaking, it must provide a context where discourse referents can be stored and accessed. It achieves the such effect by making the assignment functions that map free variables to individuals of the content in discourse. We will give an illustration of how Discourse Representation Theory (DRT) (Kamp, 1981; Kamp and Reyle, 2013) manages to achieve it in the next section. Meanwhile, we note that similar works have been proposed in other dynamic semantic theories such as the File Semantics (Heim, 1982, 1983) and different ways of the combination of Montague Semantics and discourse dynamics (Groenendijk and Stokhof, 1990; Muskens, 1996; De Groote, 2006).

Discourse Representation Theory (DRT): It is a formalism introduced and developed by Kamp (1981) and Kamp and Reyle (2013). It provides a paradigm of a dynamic semantic theory. Compared with other dynamic alternatives, it is said to be the most explicit analysis of anaphoric phenomena (Asher and Lascarides, 2003).

DRT represents discourse with a Discourse Representation Structure (DRS). A DRS is represented as a box with a pair of sets $< U, C >$. The element U stands for discourse referents (such as constants and variables of individuals), and C represents a list of conditions (i.e., properties and relations that hold among referents). For instance, a DRS for the two-sentence discourse in (12) can be written in the following logic formula:

(13) $\exists x, y, z.man(x) \wedge walked\_in(x) \wedge ordered(y, z) \wedge male(y) \wedge beer(z) \wedge y = x$

The initial DRS box is empty. Because of the existential quantifier, (12)a updates the context with a new variable x – introduced on the top part of the box. Additionally, the formula keeps track of the properties referent x satisfies: two unary predicates $man(\cdot)$ and $walked\_in(\cdot)$, so both conditions are listed at the lower part of the box. As shown in (14):

(14)

x
man(x)
walked_in(x)

Sentence (12)b then adds two news referents: y and z and two properties: two unary predicates $beer(\cdot)$ and $male(\cdot)$, and a binary predicate $ordered(\cdot)$. The only issue in this new context is the value for variable y. In sentence-level discourse analysis, we do not have enough information to deduce the linkage between y and other referents.

(15)

y, z
ordered(y, z)
male(y)
beer(z)
y = ?

When we merge the two DRSs ((14) and (15)), we simply consider that all the referents and conditions join together. On the top, we have referents x, y, and z. In the second half of the box, we list all the properties. The '?' in the linkage in (14) is initiated with a discourse referent which is *accessible*[7] from the referents. Naturally, we link the pronoun y to the predicate $man(\cdot)$, which also satisfies the condition that y is male. We thus resolve the anaphoric issue (box (16)):

[7]We do not explain DRT accessibility explicitly here. We consider it the most basic structure, with no subordination or conditional. Thus the referents in the universe of DRS are all accessible to the conditions. For a precise definition and application rules, refer to Section 2.2 in Asher and Lascarides (2003).

(16)

x, y, z
man(x)
walked_in(x)
ordered(y, z)
male(y)
beer(z)
y = x

SDRT – A Refined DRT: We have shown with example (12) how dynamic semantics such as DRT could help resolve anaphora phenomena. However, this problem is not completely resolved in DRT. The constraints on anaphora in DRT are in fact very coarse-grained, which could lead to over-generation or under-generation issues. That is one of the main motivations for a more refined DRT. A second drawback of DRT is the analysis of temporal structures, which could result in a logical interpretation that is contradictive to reality. We will give examples on both points. To tackle these shortcomings, Asher (1993); Asher and Lascarides (2003) proposed to enrich the discourse structure with rhetorical structures, thus the creation of Segmented Discourse Representation Theory (SDRT).

Before presenting concrete examples, we first present an important constraint in SDRT: the **Right Frontier Constraint** (RFC), firstly proposed by Polanyi (1985). When adding a new sentence to existing discourse, we need to decide where this sentence should be attached. In the surface form, the new sentence is linearly attached to the previous one. However, considering the anaphora or temporal order, they can only be attached at certain positions. RFC assumes that the last sentence is a possible location for attachment, as well as any nodes (sentence) that subordinate it, visually seeming like a frontier at the right side of the discourse. Such a rule is called the right frontier constraint.

Now consider a classic example in SDRT (shown in (17)). In DRT, when there is no subordination such as "every", "each" or conditional structure "if ... then ..." in the discourse, the DRS form is a simple atomic box, where we have no constraints for accessibility. From this point of view, there is no blockage for the pronoun "it" in the last sentence (f) to referent the previous discourse. Therefore, it should be accessible to all the referents: "salmon", "cheese", and "competition", and attach itself to "salmon". However, in reality, we consider (17)f to be an odd continuation of (17)a-e. Interestingly, the continuation of (17)f to (17)a-d is odd as well. DRT does not predict these. It obviously over-generates anaphoric phenomena.

(17)
a. π_1: John had a great evening last night.
b. π_2: He had a fantastic meal.
c. π_3: He ate salmon.
d. π_4: He devoured lots of cheese.
e. π_5: He won a dancing competition.
f. $*$ It was a beautiful pink.

SDRT, on the other hand, can tackle this issue by utilizing rhetorical relations and the right frontier constraint. The rhetorical representation of example (17) is shown in Figure 2.5. Each

sentence is represented with a discourse marker π (π_1 to π_5). We colored the intermediate boxes (π_7, π_6 and π_0) for a clearer presentation. Figure 2.5 is a hierarchical structure: some relations induce subordination: create a deeper level such as *Elaboration*, and others cause coordination which horizontally extends the content, such as *Narration*. The discourse starts with π_1, then elaborated with π_2 "a fantastic meal"; π_2 is further refined by π_3 and π_4 describing the meal. π_3 and π_4 together form π_7. Finally, π_5 gives a new piece of information on "dancing competition" which moves up in the structure and links with π_2 as a *Narration*. Together, π_2 and π_5 form π_6 which extends π_1 "a great evening". By convention, we label the whole discourse π_0.

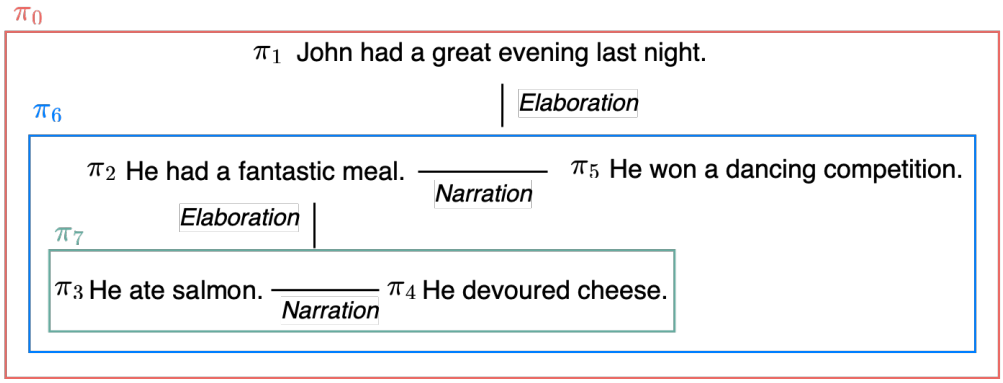


Figure 2.5: Rhetorical representation of example (17).

The right frontier constraint regulates that a new sentence can be attached to the previous discourse (called LAST in the terminology in Asher and Lascarides (2003)) and the nodes that subordinates it. The notion of *accessibility* constraints anaphora. In our case, sentence (17)f can be attached to π_5 (LAST) and above (π_6, π_1, π_0). Since π_3 is not included, we can not attach the anaphora "it" to the referent "salmon" – what we wish for to avoid the odd continuation. We see that the usage of rhetorical relations and right frontier constraint limit the over-generation issue, showing the superiority of SDRT over DRT in anaphora resolution.

A second argument for using rhetorical relations in SDRT is about **temporal anaphora**. Asher and Lascarides found that the append-based approach to dynamically construct the logical form is insufficient. Rather, one needs to rank decisions before appending them to the established structure. They gave a pair of discourses for illustration:

(18) a. Max fell. John helped him up.
b. Max fell. John pushed him.

In example (18), both discourses have the same tense forms. Under the DRT framework, one would expect that (18)a and b have the same temporal structure, i.e., the order of occurrence of two events. In reality, however, only (18)a matches the temporal order; (18)b exhibits the opposite of textual order: "John pushed Max" happens before "Max fell". Such reasoning can be enhanced using rhetorical relations. (18)a is obviously a *Narration*, describing the event that happens afterward, whereas (18)b is an *Explanation* which gives the cause of the previous discourse. This analysis is more complex than the append-based definition of discourse update. It calls for in-depth reasoning about which rhetorical connections to hold. Nevertheless, SDRT rhetorical-enhanced discourse update accounts for more general pragmatic phenomena bridging anaphora in both entity and temporal aspects.

SDRT Structure: In SDRT, the basic structure is called segmented discourse representation structure (SDRS). A well-formed SDRS contains the following vocabulary (Asher and Lascarides, 2003):

Definition 2 (SDRS Vocabulary Sets)

vocab-1. A set Ψ*: logical forms for atomic natural language clauses.*

vocab-2. A set of labels: $\{\pi_1, \pi_2, \ldots, \pi_k\}$*. Each* π *is a discourse unit;* π *can be elementary or intermediate discourse.*

vocab-3. A set of relation symbols Φ*:* $\{R_1, R_2, \ldots, R_n\}$*.*

With these vocabularies, we can define formally an SDRS:

Definition 3 (SDRS Structure) *SDRS as a tuple* $\langle \mathcal{A}, \mathcal{F}, LAST \rangle$*, where:*

- $\mathcal{A}$ *is a set of labels in vocab-2.*
- *LAST is a label in A. Intuitively, it is the last clause added in* $\mathcal{A}$*.*
- $\mathcal{F}$ *is a function that assigns each member of* $\mathcal{A}$ *a member of* Φ*.*

In the short version, we can also write an SDRS as a couple of $\langle \mathcal{A}, \mathcal{F} \rangle$ if no confusion of LAST. Again let us illustrate with the example (17). The following is a well-formed SDRS, where K_π represents the content of discourse π, i.e. the text.

(19) $\langle \mathcal{A}, \mathcal{F}, LAST \rangle$, where:

- $\mathcal{A} = \{\pi_0, \pi_1, \pi_2, \pi_3, \pi_4, \pi_5, \pi_6, \pi_7\}$
- $\mathcal{F}(\pi_1) = \mathrm{K}_{\pi_1}$
  $\mathcal{F}(\pi_2) = \mathrm{K}_{\pi_2}$
  $\mathcal{F}(\pi_3) = \mathrm{K}_{\pi_3}$
  $\mathcal{F}(\pi_4) = \mathrm{K}_{\pi_4}$
  $\mathcal{F}(\pi_5) = \mathrm{K}_{\pi_5}$
  $\mathcal{F}(\pi_0) = \mathrm{Elaboration}(\pi_1, \pi_6)$
  $\mathcal{F}(\pi_6) = \mathrm{Narration}(\pi_2, \pi_5) \wedge \mathrm{Elaboration}(\pi_2, \pi_7)$
  $\mathcal{F}(\pi_7) = \mathrm{Narration}(\pi_3, \pi_4)$
- $\mathrm{LAST} = \pi_5$

SDRT Relations: Given the formal definition of SDRS, we still have one important aspect to discuss: the rhetorical relations in vocabulary Φ. In SDRT, the choice of relation types must be justified on the basis of **truth condition** of semantic interpretation. In other words, they count R as a distinct relation if and only if R affects the truth condition of the elements it connects, but do not consider the subtle differences in intentions or goals during communication. Compared to RST, this principle narrows down the vocabulary of relations. For instance, *Contrast*, *Antithesis*, *Concession* are all eligible relations to convey the meaning of contrast in RST, while in SDRT, these relations are grouped into one type: *Contrast*. On such a basis, one can define the truth-conditional effects of relations. We will not enter into syntactic details of the relation definition part, but present briefly several levels of relation in SDRT. These relation levels are extracted from Appendix D, page 459 in Asher and Lascarides (2003):

- Content-level relations involve events and individuals. It contains most common relations such as *Alternation*, *Background*, *Consequence*, *Continuation*, *Elaboration*, and *Explanation*.
- Text-structuring-level relations such as *Parallel* and *Contrast*. These two relations require that the contents of the discourse they linked (i.e. K_{π_1} and K_{π_2}) have the same propositional structure, i.e., they always express propositions.
- Cognitive-level relations have semantics that is specified in terms of the intentions and beliefs of the dialogue agents, including *Acknowledgement*, *Indirect Question Answer Pair* (IQAP), *Plan Correction*, *Not Enough Information* (NEI), *Plan Elaboration*, *Partial Question Answer Pair* (PQAP), *Question Elaboration* (Q-Elab), and *Request Elaboration* (R-Elab).
- Divergent relations include *Correction*, *Counter-evidence*, and *Dis* (in dispute).
- Metatalk relations connect the content of one discourse to the performance of uttering another rather than to its content, including *Consequence**, *Explanation**, *Result**, where the *-ed symbol is used to indicate that this is a metatalk relation and not the normal ones. Refer to Chapter 7.6.5 in Asher and Lascarides (2003) for details.

SDRT provides a more elaborate discussion about relations in dialogues. Upon examining the definitions of different levels of relations, we found that relations such as "cognitive-level", "divergent-level", and "metatalk-level" mostly describe discourse relations between dialogue agents, thus being applicable to dialogue settings. For example, the relations in the dialogue examples (20) and (21) correspond to *Acknowledgement* and *Plan Correction*, respectively.

(20) A: Close the window.
B. Ok.

(21) A: Close the window.
B. I am afraid I can't do that.

SDRT also discusses an important feature in dialogues, which is how people engage in disputes and come to an agreement. The relations that indicate disputes are referred to as "divergent relations". These relations are typically not found in monologues, although self-repair utterances, which are another form of correction, can rarely be observed. In dialogues, the *Correction* relation links discourse units with contradictory contents, as shown in (22) and (23):

(22) A: John distributed the copies.
B: No, Sue distributed the copies.

(23) A: John went to jail. He was caught embezzling funds from the pension plan.
B: No! John was caught embezzling funds, but he went to jail because he was convicted of tax evasion.

To infer rhetorical relations, SDRT uses **axiom schemata**. The general schema is given in (24), where α and β are discourse units; λ is context; connective $>$ means "then normally,". In English, (24) states that β is attached to α in a certain context, and moreover there are "some stuff" (evidence) about α, β and λ, then normally, the discourse relation is R.

(24) $(?(\alpha, \beta, \lambda) \wedge \text{some stuff}) > \mathrm{R}(\alpha, \beta, \lambda)$

Various information is needed to infer the most appropriate relation. Normally, lexical features are good indicators. In a concrete example (25), cue phrases such as "and then" monotonically yield *Narration* relation, with its axiom scheme looks like (26):

(25) π_1. Kim watched TV.
π_2. And then she went out.

(26) $?(\alpha, \beta, \lambda) \wedge \text{and-then}(\alpha, \beta)) \rightarrow \text{Narration}(\alpha, \beta, \lambda)$

Apart from lexical markers, the relation prediction phase also searches for information on punctuation, intonation, and domain knowledge. In dialogue settings, information about speakers' speech acts and (rational) moves could also be useful. The determination of discourse coherence depends on both structural rhetorical information and Gricean reasoning (Benz and Salfner, 2011). According to Asher and Lascarides, Grice's Maxim of Relation (Grice, 1975) is equivalent to discourse coherence. This implies that a new text segment is relevant to a given segment only if there is a rhetorical relation connecting them.

SDRS Update: We have mentioned in the previous part that SDRT differs from other dynamic theories in its sophisticated and more complex discourse updates, in comparison to the "append-based" methods. We have also shown how SDRT makes use of rhetorical relations and right frontier constraint to block odd anaphora attachments (cf pink salmon example (17)). Here, we take one step further in examining how old SDRS absorbs and binds the new information from new SDRS. This process is called **SDRS update**. It contains two tasks: first identify the part of discourse to which the new SDRS will bind; secondly, infer the rhetorical relation.

The formal language used in SDRS update is *glue language*. As the name suggests, this language glues the different logical parts together to form an SDRS form for discourse. Glue language builds up glue logic, which is a logic that supports nonmonotonic inferences. Asher and Lascarides (2003) argue that the relation prediction in discourse analysis should not take the "wait-and-see" strategy that inferring the rhetorical relation only when newly present information monotonically ensures such connection. For instance, *because* is a monotonic clue for relation *Explanation*. Rather, the inference should be made even when the monotonic clues are absent. In (24) we show the syntax of axiom schemata in glue language, namely $?(\alpha, \beta, \lambda)$.

The update tasks contain three steps:

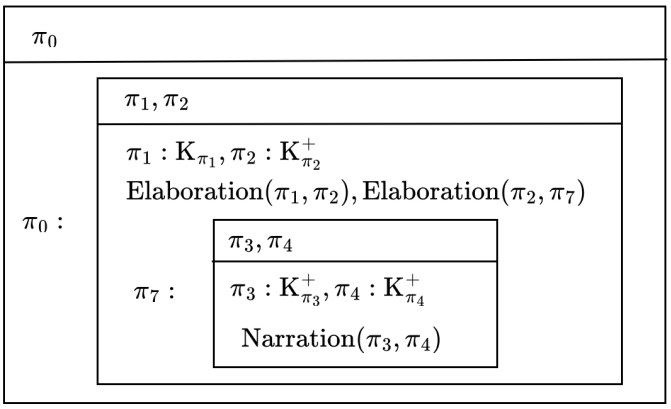
- Build the available subsets of labels from $\langle \mathcal{A}, \mathcal{F} \rangle$, from which discourse β will attach.
- For each previous discourse α, identify a label λ and use glue logic to infer discourse relation(s) between α and β.
- Eliminate other SDRSs obtained in the first step if they fail to meet certain structural constraints (details omitted).

Further, when multiple relations are available, SDRT employs two principles to resolve potential conflicts. The first one is *Specificity Principle*: "when the consequences of default axioms conflict, the axiom with the more specific antecedent win." The intuition behind this principle is that people tend to remember new information when it is specifically linked to a previous context. Another important principle is called *Maximum Discourse Coherence* (MDC) principle. Asher and Lascarides have designed a way to determine which interpretation is more coherent than

another. SDRS update aims for as many relations as possible and as many as *preferred* (simpler and more consistent structures) as possible. By contrast, it does not favor under-specified conditions.

We retake the salmon example (17) to illustrate SDRS update when π_5 is introduced. Before π_5, the SDRS is featured in (27), where K^+_π implies simple SDRS update with no under-specified conditions.

(27)



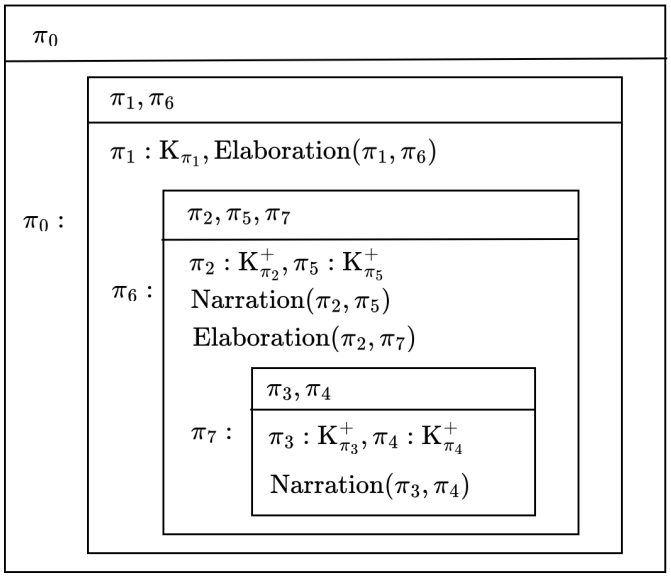
When introducing a new discourse π_5, there are five possible points of attachments: π_0, π_1, π_2, π_7, and π_4. π_3 is excluded since it is blocked by π_4. It is when the *maximum discourse coherence principle* puts into effect. If correctly using this principle, one shall link π_5 to π_2. Due to the *specificity principle*, it is not favorable to attach π_5 to π_0 in the first place. Next, we can rule out π_7 and π_4 as well since the updates will bring under-specified conditions. Another way to test coherence is by comparing different attachment effects. If we attach π_5 to π_4, the sequence of discourse in (28) is much less coherent than that in (29).

(28) He ate salmon. He devoured lots of cheese. He won a dancing competition.

(29) John had a great evening last night. He had a fantastic meal. He won a dancing competition.

Now, we compare the attached points π_1 and π_2. If attached to π_1, we obtain a relation *Elaboration*(π_1, π_5), meaning "the meal" and "winning the dance competition" both contribute to the "great evening". However, since we already know that *Elaboration* holds between π_1 and π_2, which makes two *Elaboration* relations sharing the same first discourse. Under such circumstances, π_1 is viewed as the common topic for a sequence of elaboration (π_2 and π_5), and π_5 is naturally attached to π_2 with *Narration* relation. The updated result is shown in (30). An interesting follow-up is the *constraints on update*: once π_5 is attached to π_2, it also can not attach to π_4. Further, since now LAST is π_5, it blocks the access to π_2, π_6, π_3, and π_4 for future discourse. The pronoun in "* It was a beautiful pink." thus can not be resolved. We have said that the continuation of (17)f to (17)a-d is also strange. Even without the topic change of π_5, SDRS in (27) shows that π_3 is still blocked by π_4, making any further description on π_3 after π_4 not feasible. Previously, we have given theoretical reasoning why (17)f is not a good continuation (right frontier constraint). Here, we provide SDRS updates to prove the oddness of such continuation in practice. By which, we conclude the discussion of the SDRT framework.

(30)



This section provides an overview of two prominent full discourse analysis frameworks: Rhetorical Structure Theory and Segmented Discourse Representation Theory. Both frameworks involve segmenting the text into elementary discourse units (`EDUs`) and identifying relations between them while adhering to certain structural constraints. The relation assignment process is incremental and recursive, with annotators linking `EDUs` to form intermediate discourse units (`CDUs`) that are connected to cover the entire document. Although both frameworks share similarities in the analysis process, they differ in some aspects. For instance, RST produces tree-like structures, whereas SDRT generates graph-like structures. The definition of relations also differs, with RST focusing more on intentions and SDRT on truth conditions. Moreover, SDRT introduces relations at the "divergent-level" and "cognitive-level", which makes it more suitable for dialogue analysis. It should be noted that the SDRT framework is not the only one that can be used for analyzing dialogues. In the GUM corpus (Zeldes, 2017) (that we will present in the next section), for instance, the conversational component is annotated using the RST framework.

There are other discourse analysis frameworks besides SDRT and RST, such as the Lexicalized Discourse Tree Adjoining Grammar (L-DTAG) (Webber, 2004). L-DTAG builds on the Tree-Adjoining Grammar (TAG) (Schabes, 1990; Schilder, 1997; Gardent, 1997) to combine elements in discourse. It was first proposed in Webber and Joshi (1998) and has inspired the creation of the Penn Discourse Treebank (PDTB) annotation project, which is one of the largest datasets of its kind. This leads us to the topic of discourse annotation projects and important discourse corpora.

2.3 Discourse Corpora

Discourse structures for complete documents have been mainly annotated within the Segmented Discourse Representation Theory or the Rhetorical Structure Theory, with the latter leading to the largest corpus and many discourse parsers for monologues, while SDRT is the main theory for dialogue corpora, i.e., STAC (Asher et al., 2016) and Molweni (Li et al., 2020). The issue of data sparsity is not limited to monologues but is even more pronounced in dialogues. Existing discourse-annotated treebanks are scarce and only available in limited domains, such as RST-DT (Carlson et al., 2002a) for news articles (385 documents), SciDTB for scientific abstracts (Yang and Li, 2018) (798 abstracts), STAC for online board game (Asher et al., 2016) (45 games), and

Molweni (Li et al., 2020) for Ubuntu chat log discussion (10,000 short dialogues, in average 9 utterances/dialogue) (Li et al., 2020).

In this section, we describe the corpora annotated within the RST (Section 2.3.1) and the SDRT (Section 2.3.2) frameworks, followed by an introduction of the Penn Discourse Treebank (Section 2.3.3). We then present other discourse-annotated corpora in Section 2.3.4. Furthermore, we provide the results of an investigation of the Molweni corpus which shows some non-negligible annotation flaws (Section 2.3.5).

2.3.1 Corpora in the RST Framework

The RST framework led to the creation of the first annotated discourse corpus, known as the Marcu 1999 RST corpus (Marcu et al., 1999), which aimed to assess the feasibility of conducting RST analysis manually and automatically. Marcu's experimental annotation prompted another significant RST annotation project on the Penn Treebank (Marcinkiewicz, 1994), resulting in the largest and most widely used RST-style discourse corpus, the RST Discourse Treebank (Carlson et al., 2002b).

- The *Marcu 1999 RST Corpus*: The first corpus that was annotated using the RST framework. It is comprised of 90 documents selected from various corpora, including 30 texts from MUC7, 30 from Brown-Learned, and 30 from Wall Street Journal. Although this corpus has not been extensively used for training automatic systems, it demonstrated the feasibility of annotating texts using the RST framework and served as the basis for developing the first annotation guidelines.

- The *RST Discourse Treebank*[8] (RST-DT), developed by Carlson et al. (2002a), comprises 385 news articles from the Wall Street Journal section of the Penn Treebank (Marcinkiewicz, 1994). The Penn Treebank is already manually annotated in syntax. The RST-DT is considered the primary corpus for developing full discourse analysis systems.

- The *GUM Treebank* (Zeldes, 2017) is a continuously growing corpus with a multi-layer annotation that includes POS tagging, sentence segmentation, and RST-style discourse parsing. As of version 8.0, it comprises 193 documents from 12 genres, including interviews, news stories, and travel guides. Both constituency and dependency tree structures are provided for the discourse parses in this corpus.

- The *Instructional Corpus*[9] (Instr-DT) (Subba and Di Eugenio, 2009) consists of 176 documents pertaining to home repair. The corpus features a total of 5,172 annotated rhetorical relations for 5,744 `EDU`s in RST-style constituent trees.

The previously mentioned corpora are all in English. It is important to note that the RST framework has also been utilized for the creation of corpora in other languages. For instance, the Potsdam Commentary Corpus (Stede, 2004) for German, the Spanish RST Discourse Treebank (Da Cunha et al., 2011) for Spanish, the RST Basque Treebank[10] (Iruskieta et al., 2013) for Basque, the Russian RST Treebank (Pisarevskaya et al., 2017) for Russian, and the GCDT (Peng et al., 2022) for Chinese[11].

[8] `https://catalog.ldc.upenn.edu/LDC2002T07`

[9] `https://nlp.lab.uic.edu/resources/`

[10] `https://sites.icmc.usp.br/taspardo/projects.htm`

[11] Refer DISRPT 2023 shared task GitHub for more information: `https://github.com/disrpt/sharedtask2023/tree/main/data`.

2.3.2 Corpora in the SDRT Framework

The initial annotation of the SDRT framework was carried out for the DiSCoR project (Reese et al., 2007), which focused on monologues such as news articles. However, with the release of STAC in 2016, SDRT is now more commonly associated with dialogue, particularly multiparty dialogues. To date, STAC (Asher et al., 2016) and Molweni (Li et al., 2020) corpora are the most widely used datasets for training SDRT-style parsers.

- The *DiSCoR Corpus* (Reese et al., 2007) is the first corpus annotated under the SDRT framework, which was developed to investigate the interaction between rhetorical structures and coreference phenomena. This corpus includes 60 documents from MUC6 (Wall Street Journal) and ACE2 (news articles) corpora, and it has been annotated with 14 discourse relations.

- The *ANNODIS Corpus*[12] (Afantenos et al., 2012a) is a corpus of written French texts from four sources: 39 regional daily news articles, 30 French Wikipedia articles, 25 articles from the proceedings of the *Congrès Mondial de Linguistique Française*, and 32 reports from the *Institut Français de Relations Internationales*. The corpus was created as part of the ANNODIS project (ANNOtation DIScursive). The annotation includes $3k$ elementary discourse units (`EDUs`) and $1.4k$ complex discourse units (`CDUs`) linked by $3k$ rhetorical relations.

- The *STAC Corpus*[13] (Asher et al., 2016) contains 45 online multi-party strategic chat conversations during the board game *The Settlers of Catan*. It is manually annotated and divided into approximately 1000 sub-documents. Although there are different versions of the separation of the training, validation, and test sets, we use the version employed in Shi and Huang (2019) for all our experiments in this thesis (note as "stac_shi2019"). Specifically, this version comprises 1161 short documents (947 for training, 105 for validation, and 109 for testing). It is currently the most commonly used English corpus to train SDRT-style discourse parsers. It is worth noting that the STAC project offers a situated version, which incorporates depictions of non-linguistic events such as game moves. This version is not used in this thesis.

- The *Molweni Corpus*[14] (Li et al., 2020) is a dataset of short multi-party technical chats derived from the larger *Ubuntu Chat Corpus* (Lowe et al., 2015). It consists of $10k$ dialogues, with $9k$ for training, and 500 each for validation and testing. The corpus follows the same annotation schema as STAC, making it compatible with SDRT-style parsers. Despite its large size, Molweni has issues with repetition and annotation accuracy, which are discussed in Section 2.3.5.

In Table 2.4 we show a few key statistics of STAC and Molweni corpus.

Apart from the ANNODIS corpus in French, there is also the Arabic Discourse Treebank (Keskes et al., 2014) annotated under the SDRT framework, which consists of 90 news stories from the Arabic Treebank (ATB).

[12] `http://redac.univ-tlse2.fr/corpus/annodis/`
[13] `https://www.irit.fr/STAC/corpus.html`
[14] `https://github.com/HIT-SCIR/Molweni`

dataset	split			sent/doc		tok/sent		tok/doc		spk/doc		rel
	train	dev	test	max	avg	max	avg	max	avg	max	avg	
Molweni	9,000	500	500	14	8.8	17	11.9	208	105	9	3.5	16
STAC_shi2019	947	105	109	105	11.0	13	4.4	607	50	6	3.0	16

Table 2.4: Some statistics in STAC and Molweni corpora. Numbers of sentences per document (sent/doc), tokens per sentence (tok/sent), tokens per document (tok/doc), speakers per document (spk/doc) are given. Both corpora have 16 relation types.

2.3.3 Penn Discourse Treebank

Unlike corpus in RST- and SDRT-style, *Penn Discourse Treebank* (short PDTB) is not annotated under a theoretical framework for document-level discourse analysis. This project is mainly focused on the identification of local discourse connectives. It has been enriched three times. We show its historical versions and related projects:

- The *PDTB 1.0* (Miltsakaki et al., 2004) was released in 2005. The aim was to produce a large-scale corpus in which discourse connectives and their arguments are annotated. In total this first version contains 30,000 annotations: 10,000 implicit connectives, and 20,000 annotations of the 250 explicit connectives.

- The *PDTB 2.0* (Prasad et al., 2008a) extends the number of annotations of discourse relations and their two abstract object arguments to 35,136, which covers over 1-million words Wall Street Journal corpus. Sense annotation was added for all the explicit, implicit, and AltLex relations. AltLex label refers to the case when the insert of implicit connectives brings redundancy since an alternative non-connective expression is already presented.

- The most recent version *Penn Discourse Treebank 3.0* (Webber et al., 2019) consists of annotations for 53,631 tokens, which is approximately 13,000 more than its predecessor. Additionally, new sense and relation annotations have been included, such as the sense mark for question-response pairs, known as *Hypophora*.

- The *Biomedical Discourse Relation Bank* (*BioDRB*) (Prasad et al., 2011) is a PDTB-style domain-specific corpus. It contains 24 open-access full-text biomedical articles from the GENIA corpus, which counts in a total of 5,859 relation tokens for the four different relation types: Explicit, Implicit, AltLex, and NoRel.

PDTB is widely regarded as one of the largest and most influential treebanks featuring sentence-level discourse information. Similar PDTB-style projects have been established for several other languages, such as the Hindi Discourse Treebank (Prasad et al., 2008b) for Hindi, the Turkish Discourse Treebank (Zeyrek and Webber, 2008; Zeyrek et al., 2009) for Turkish, the French Discourse Treebank (Danlos et al., 2012) for French, the Prague Discourse Treebank 1.0 (Poláková et al., 2013) for Czech, and the Chinese Discourse Treebank 0.5 (Xue et al., 2005) for Chinese. More recently, a multilingual resource known as TED-Multilingual Discourse Bank (Zeyrek et al., 2019) has been released, which features TED-talks annotated at the discourse level in 6 languages: English, Polish, German, Russian, European Portuguese, and Turkish.

2.3.4 Corpora Constructed under Other Frameworks

- The *GraphBank*[15] (Wolf and Gibson, 2005) is a collection of 135 Wall Street Journal newswire texts that are manually annotated with coherence relations, totaling 70,000 words in English. It employs a graph-like structure instead of trees and defines 11 rhetorical relations based on the work of Hobbs (1985). The authors adopted the graph structure because trees are inadequate to represent all discourse structures, including crossing edges.

- The *SciDTB*[16] (Yang and Li, 2018) is a treebank of scientific abstracts annotated in dependency trees. It provides a more flexible and simpler annotation scheme than the RST-style, while still maintaining complete annotation. The treebank comprises 798 abstracts, with a total of 18,978 labeled relations. The size of *SciDTB* is similar to that of *RST-DT*. The relation categories are mainly based on the RST framework and the ISO 24617-8 standard (Bunt and Prasad, 2016), resulting in 17 coarse-grained and 26 fine-grained relation types. This corpus is also in English.

- The *COVID19-DTB*[17] (Nishida and Matsumoto, 2022) adopts the *RST-DT* guidelines for segmenting `EDU`s, and follows the relation types of *SciDTB* and PDTB corpora. The authors simplified the relation types to 14 categories. The corpus consists of 300 English abstracts (6,005 `EDU`s) extracted from the 2020 snapshot of the COVID-19 Open Research Dataset (CORD-19) (Wang et al., 2020).

So far, we have given a brief overview of the languages and treebanks that are available in the RST, SDRT, PDTB, and dependency formalisms. It should be noted that this list is not exhaustive, and for a more comprehensive reference, we recommend consulting the DISRPT shared task website[18].

2.3.5 Investigation of Molweni Corpus

Molweni (Li et al., 2020) is a corpus derived from the Ubuntu Chat Corpus (Lowe et al., 2015), consisting of 10,000 short dialogues with 8 to 15 utterances annotated in the SDRT framework. Due to its size, it is an ideal corpus for supervised learning of discourse structures. Moreover, the corpus contains 30,066 annotated questions and is utilized for Machine Reading Comprehension (MRC) task. Given the complexity of Ubuntu chat logs (e.g., multiple speakers, entangled discussions with various topics), the corpus was examined first. However, we found a significant amount of repetition in sequential documents and inconsistency in discourse annotation for the same utterances.

Clusters: Out of the 500 dialogues in the discourse augmented test set, we discovered 105 "clusters" in total. One particular cluster includes all the documents that have only one or two differing utterances. We hypothesize that this may be due to the previous disentanglement process. For example, documents with ID 10 and 11 are in the same cluster since only the second utterance differs, as illustrated in Figure 2.6. A similar situation is attested in the documents {1, 2, 3}, {7, 8, 9}, {19, 20, 21}, to name a few. The number of similar documents in one cluster varies: with some clusters containing up to 8 highly similar documents.

[15] `https://catalog.ldc.upenn.edu/LDC2005T08`
[16] `https://github.com/PKU-TANGENT/SciDTB`
[17] `https://github.com/norikinishida/biomedical-discourse-treebanks`
[18] `https://github.com/disrpt/sharedtask2023/tree/main`.

cluster	doc id	#doc pair	#same link	#err link	#same rel	#err rel
1	{1, 2, 3}	3	18	2	16	2
2	{7, 8, 9}	3	18	0	18	7
3	{10, 11, 12, 13, 14}	10	80	4	76	25
...						
105	500	676	4,787	284	4,503	606
-	-	-	-	5.9%	-	13.5%

Table 2.5: Investigation of link and relation inconsistency inconsistency in Molweni. A "doc pair" means a pair of two similar documents (e.g., {1, 2}, {1, 3}); "same link": number of links between the same `EDUs`, which should be attached exactly the same way; "same rel": relations between the same `EDUs`, which also should be the same. "err link" and "err rel" are inconsistent links and relation types between the same `EDUs`.

Annotation Inconsistency: Upon closer inspection of the annotation in similar examples, we discovered inconsistencies in both `EDU` attachments and relation types. Specifically, we examined every *document pair* (i.e., two similar documents in the same cluster) in all 105 clusters in the test set. As an example, Figure 2.6 visualizes the inconsistency for documents 10 and 11: we expect the same links and relations among all `EDUs` except for $\texttt{EDU}_2$, but we observed one link inconsistency (in red: $e_8 - e_9$ in document 10, $e_7 - e_9$ in document 11) and two relation inconsistencies (in blue: *Elaboration* for $e_3 - e_6$ in document 10, *Continuation* for $e_3 - e_6$ in document 11), which we refer to as link error and relation error, respectively. In total, we found 6% of link errors (#err link) and 14% of relation errors (#err rel) in the test set, with similar error rates for the validation and train sets. See Table 2.5 for precise scores.

Due to its lengthy and intricate dialogues, the Ubuntu Chat Corpus underwent disentanglement preprocessing, which resulted in a set of shorter, slightly different sub-dialogues. While these may be useful for other dialogue studies such as Machine Reading Comprehension task, our focus on discourse structure requires more various data points with **consistent** discourse annotation. As a result, we decided to exclude this corpus from our experiments in this thesis.

2.4 Discourse in Different Language Settings

Having looked at discourse theories and applied annotation corpus, we now examine the use cases of these theories in different language settings. Typically, we present two dimensional analysis: spoken *vs.* written language, and monologues *vs.* dialogues.

It is generally acknowledged that discourse in speech and writing differs. Although not explicitly stated, the frameworks discussed in Section 2.2 have mostly influenced annotation projects with written language. For example, the RST-DT corpus (Carlson et al., 2002a) is annotated using 385 well-written news articles from the Wall Street Journal (WSJ), which have also been used in the Penn Discourse Treebank (PDTB). There are differing opinions on the complexity of processing discourse in speech and writing. Some linguists believe that oral grammar is simpler than written grammar with sparse vocabulary, resulting in simpler discourse structure. Chafe (1982) is among the advocates of this view. In contrast, opponents argue that discourse structure in speech is much more complex, as stated in Halliday (1994): "Speech is not, in any general sense, 'simpler' than writing; if anything, it is more complex". Halliday further argues that

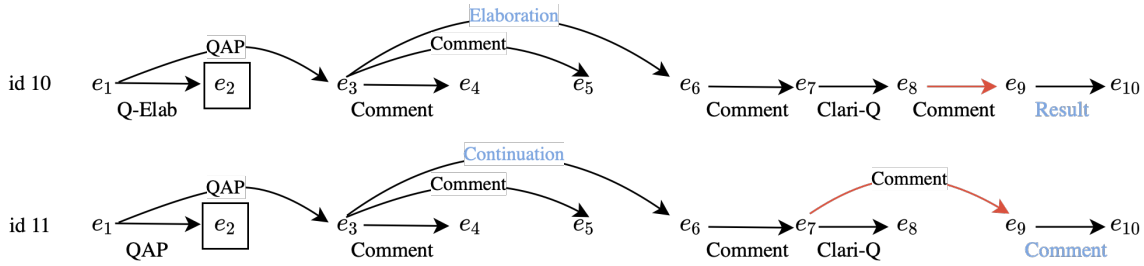


Figure 2.6: Similar documents in the same cluster. Circled `EDUs` are different. In red: inconsistent discourse arcs; in blue: inconsistent rhetorical relation.

test document id 10:
[e_1] matthew99857: so do i need additional hardware to fix it ?
[e_2] vocx: ca n't you disable the raid from the bios ? check your motherboard manual .
[e_3] ikonia: just use the disk as an individual disk
[e_4] sugi: vocxi : oh i am sorry . i misunderstood you . thank i will try it now
[e_5] vocx: you need to word better your answers , seems like nobody in getting you today .
[e_6] sugi: vocx : iso 9660 cd-rom filesystem data udf filesystem data (unknown version , id 'nsr01 ')
[e_7] ikonia: looks like that should work as a loop back file system
[e_8] sugi: -mount -o loop but instead of .iso .mdf ? or the .mds file ?
[e_9] ikonia: try it , linux see 's it as a " image " so it may work
[e_{10}] sugi: vocx : wow it worked , i feel retard for nto

test document id: 11
[e_1] matthew99857: so do i need additional hardware to fix it ?
[e_2] ikonia: no you need to stop using raid
[e_3] ikonia: just use the disk as an individual disk
[e_4] sugi: vocxi : oh i am sorry . i misunderstood you . thank i will try it now
[e_5] vocx: you need to word better your answers , seems like nobody in getting you today .
[e_6] sugi: vocx : iso 9660 cd-rom filesystem data udf filesystem data (unknown version , id 'nsr01 ')
[e_7] ikonia: looks like that should work as a loop back file system
[e_8] sugi: -mount -o loop but instead of .iso .mdf ? or the .mds file ?
[e_9] ikonia: try it , linux see 's it as a " image " so it may work
[e_{10}] sugi: vocx : wow it worked , i feel retard for nto

"speech is grammatically intricate, with meanings related serially". The frequent occurrences of under-specification, implicitness, and ambiguity in spoken language make it more complex compared to written language. Nevertheless, there seems to have an uncontroversial agreement on that spoken and written language differ as for their respective kind of complexity (Crible and Cuenca, 2017).

Another dimension to review the discourse theories is by looking at speech devices: monologue contains one person's speech whereas dialogue is the mixed speeches of two or more speakers. Some theories are more suitable to analyze monologue discourse structure, while others may be extended to dialogue. RST, for instance, with its rooted-tree structure, is mostly used as guideline theory for monologue annotation. RST has also been used in GUM conversation documents annotation (from Santa Barbara Corpus). The annotation schema for PDTB, on the other hand, shows more shallow discourse structure and relations, which makes it more flexible. It has been used for annotation in both monologue and dialogue settings. Tonelli et al. (2010); Riccardi et al. (2016) tested the applicability of PDTB to spontaneous conversations. They applied the schema on an Italian dialogue corpus LUNA and proposed revision suggestions. LUNA is available on the DISRPT website. In an annotation project for SMS message conversation, Xue et al. (2016) made distinction of discourse relations between same-participant and among different-participant, and they adopted PDTB relations for the same-participant part. In total, 44 files have been annotated, with an average 88 messages per file. This corpus is not publicly available online. SDRT, derived from DRT, was initially designed for monologues. In Asher and Lascarides (2003), authors extend the theory to handle dialogue by incorporating questions and requests in discourse structure. We see discourse relations such as *Question Answer Pair*, *Acknowledgement*, *Correction*, etc. The extension makes SDRT adapted in multi-speaker setting. The flexible graph structure (compared to tree structure) also makes it a suitable choice for dialogues.

This section begins by introducing some linguistic peculiarities in Section 2.4.1, namely differences between spoken and written language, and monologues versus dialogues. Subsequently, in Section 2.4.2, we examine research on discourse relations in both domains. Specifically, we explore how monological SDRT relations are expanded to the dialogue setting and the current status of adapting written annotation frameworks to spoken language.

2.4.1 Language Specificities

Discourse in Spoken *vs.* Written Language: Discourse patterns of spoken and written communication are distinct. Compared to written language, spoken language frequently includes ungrammatical and unfinished sentences, disfluencies, fillers, and hesitations (Wang et al., 2017a). The speaker and listener in oral conversations have access to additional channels of information, such as facial expressions, body posture, and eye movement, so the information conveyed solely through words may be incomplete and elliptical at times. On the other hand, spoken communication places emphasis on rapid online processing, leading to shorter sentences and a higher degree of interactivity (Rehbein et al., 2016). Consequently, speakers primarily focus on the current speech turns, leading to a more linear and *a priori* simpler discourse structure.

Here we briefly discuss a typical phenomenon in spoken language: disfluency, which is commonly seen in spontaneous human oral speech, both in monologues and dialogues. At the pragmatic level, disfluencies can communicate valuable information such as hesitation or the introduction of new or unfamiliar information related to the discourse entity being discussed (Yoshida and Lickley, 2010). This phenomenon has attracted considerable attention from researchers, as evidenced by the organization of the Disfluency in Spontaneous Speech (DiSS) workshop, which has been held for over ten editions since 1999. Disfluency in spontaneous speech typically includes

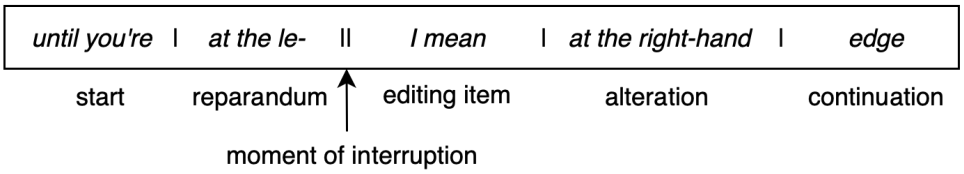


Figure 2.7: General pattern of disfluency, in Ginzburg et al. (2014).

pauses, hesitations, prolongations, truncations, repetitions, self-repairs, and similar phenomena. According to the research by Levelt (1983) and Shriberg (1994), disfluencies in speech tend to follow a regular pattern (see Figure 2.7). Except for the moment of interruption and continuation, all elements in the pattern are optional. This pattern and the relations between its elements can be used to classify disfluencies into different types (Ginzburg et al., 2014). In (31), we provide some examples of disfluencies annotated according to this pattern from the Switchboard corpus (Godfrey et al., 1992). In the examples, the symbol "+" indicates the moment of interruption and separates the reparandum from the alteration, while "{}" brackets represent editing items and filled pauses, and "[]" brackets enclose the disfluency as a whole.

(31) a. { I mean } [[I, + I,] + [there are a lot, + there are so many]] different songs.
b. [We were + I was] lucky too that I only have one brother.

Disfluency examples in Switchboard corpus (Godfrey et al., 1992)

The annotation described above is not easily legible and requires additional efforts for text pre-processing. In addition to disfluencies, spoken discourse often includes arguments that are separated by fragments, as illustrated in example (32) where argument 1 and 2 are separated by "filler words". A filler, filled pause, hesitation marker, or planner is a vocalization or word or sound used by conversation participants to indicate that they are pausing to think but have not yet finished speaking. In this example, the second and third speech turns are filled with filler words. As an extension, we note that there are also a lot of non-verbal signals in oral communication: smiling, frowning, sighing, etc. They may do not have the same effect as words and phrases, but they can directly or indirectly impact the development of discourse. Laughter, for instance, can present propositional content such as repair, implicature, or irony (Ginzburg et al., 2015, 2020).

(32) A : I'm on email every day$_{ARG1}$
A : you know
A : I can
A : I've access to it now$_{ARG2}$

Filler words example from Rehbein et al. (2016).

The abundance of information in spoken language poses processing challenges for discourse analysis. Initially, the complex transcription, as demonstrated in (31), presents difficulties in segmenting the discourse into elementary discourse units (`EDUs`). When provided with a clean

transcription and pre-segmented `EDUs`, discourse relation definition in spoken language could differ from that in written language. For example, when identical or near-identical sequences are repeated, determining the appropriate relation type may not be straightforward (Rehbein et al., 2016). Additionally, certain discourse connectives employed in spoken language may contain different semantic meanings. An example of this is the connective *so*, which often conveys a sense of *conclusion* in spoken language, as we will explore further in Section 2.4.2.

Discourse in Monologues *vs.* Dialogues: Unarguably, dialogue is different and more difficult to analyze than monologue. With the introduction of more than one participant, there emerges the possibility of information exchange, cooperation, agreement, and disagreement (Asher and Lascarides, 2003). A few discourse relations such as questions and informs, directives and commissives must also be incorporated.

Dialogue can have unique structures, such as in multi-party conversations where multiple speakers may give an answer or acknowledgment to the same utterance simultaneously, resulting in a *diamond*-shaped (*losange*-shaped) graph (Asher et al., 2016), as illustrated in Figure 2.8.

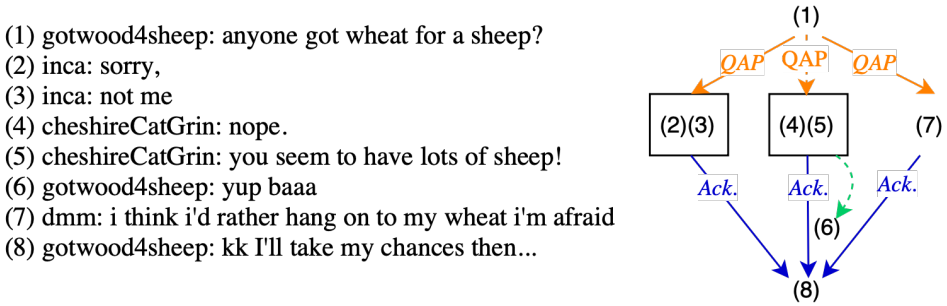
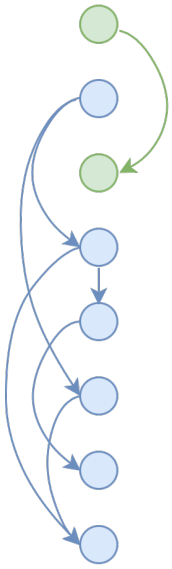


Figure 2.8: *Diamond*-shaped discourse structure from STAC corpus.

Dialogue presents several distinctive properties, including *entangled conversation*, which is a commonly seen phenomenon in online chat forums. In this phenomenon, multiple conversations occur concurrently, and one needs to disentangle the topics to obtain coherent conversations. The "reply-to" indicator is very useful in entangled conversation, showing the link between the current and previous utterances. Figure 2.9 shows an example where Speaker A and B ask two questions independently, forming two sub-conversations. To improve readability, we have colored the speech turns in different sub-conversions with green and blue. Speaker C firstly answers A's question (C_3) and then answers B's question (C_4-C_5). A does not comment back to C's reply but addresses B's question directly. In the end, B replies back to A and C. Thus, the graph structure of this example contains two parts: utterances A_1, C_3 form one independent structure and the other utterances form another structure. Interestingly, we see one message (B_2) that receives multiple responses and one message (B_8) that responds to multiple messages, forming a *losange*-shaped graph that has just been discussed before. The complexity of multi-conversation participation and discourse structure presents a great challenge for discourse analysis in dialogues.

At the pragmatic level, analyzing dialogue requires in-depth analysis and cognitive knowledge, especially when it contains rhetorical interaction. An excerpt from a conversation between a schizophrenia patient and a psychologist from the French corpus SLAM Amblard et al. (2014) illustrates this point. In Example (33), the psychologist (Speaker A) seeks to learn more about the patient's (Speaker B) opinion on management and treatment. The patient mostly responds with *backchannel* utterances, such as "hum" and "yeah", which are typical in spontaneous conversations



A₁ : hehe yes. does Kubuntu have `KPackage'?

B₂ : does anyone know a consoleprog that scales jpegs fast and efficient?.. this digital camera age kills me when I have to sclae photos :s

C₃ : A, yes

C₄ : B, convert

C₅ : part of imagemagick

A₆ : B, ImageMagick

C₇ : B, i used that to convert 100's of photos in one command

B₈ : Oh... I'll have a look.. thx =)

Figure 2.9: Conversation entanglement example, adapted from an online chat in Kummerfeld et al. (2019). Sub-conversations have different colors.

but rarely seen in monologues. These utterances are phatic expressions that encourage more speech and indicate more focus on the conversation. Backchannels are useful discourse markers for discourse analysis, and in some cases, such as with patients with Schizophrenia, they may also indicate adherence and satisfaction with treatment (Howes et al., 2012; Li et al., 2021a). We will explore the usage of backchannel in discourse structure discovery in Chapter 4.

(33) **A₁** : Bon. je sais donc euh je vous rappelle le but. c'est vraiment d'aider au diagnostic et à la prise en charge psychothérapeutique. [*Well, I know euh, so I'll remind you of the goal. It's really to help in the diagnosis and in the psychotherapeutic management.*]

B₂ : Ouais. [*yeah.*]

A₃ : Donc euh / donc voilà. c'est euh / c'est très gentil de / de vous y préter déjà. [*So euh / so yeah. It's euh / it's very kind of you to / to take care of it already.*]

B₄ : Mmh mmh. [*Hum mmh.*]

A₅ : Et euh... vous voudriez parler de quoi. [*And euh... what would you like to talk about.*]

B₆ : Je sais pas. [*I don't know.*]

Backchannel responses example in Li et al. (2021a)

The examples discussed above demonstrate that the discourse structure in dialogue can be more diverse and intricate than that of monologue. Moreover, the interpretation of dialogue often requires pragmatic and para-linguistic factors to be taken into account. In the following section, we discuss the current efforts in discourse analysis across various language settings. Despite the challenges and variations, researches have proposed adaptation methods to overcome these difficulties.

2.4.2 Discourse Relation Adaptation

From Monologue to Dialogue: According to Sacks et al. (1978); Sacks (1992); Mann (1984); Asher and Lascarides (2003), speech turns provide an important clue to discourse structure. By analyzing the speech turns within and between speakers, one can utilize discourse relations (e.g., RST and SDRT relations) to connect these elements for a cohesive semantic implication. In SDRT, each speaker has their interpretation of the dialogue. Even though their beliefs may differ, they follow the same rules of interpretation (Asher and Lascarides, 2003). In other words, they mutually agree on the meanings of rhetorical relations and the default axioms used to infer them. Therefore, dialogue and monologue share some rhetorical relations. To illustrate this point, consider the following example (34), where two speakers (A and B) are discussing another person:

(34) **A_1** : There was this guy. He came to the sessions. He never said anything. Then one day he shows up, and he starts talking, interesting.

B_2 : Why didn't he say anything before?

A_3 : Dunno. Shy maybe.

A_4 : But anyway he's yammerin away and telling these jokes...

Example (5) in Section 7.2.2 in Asher and Lascarides (2003)

This example features an *Elaboration* between speech turns A_1 and A_4 as event "telling these jokes" (E refers to event, therefore: E_{tell_jokes}) follows the event "he starts talking, interesting" ($E_{talking}$), and a *Contrast* across A_3 and A_4 as indicated by the connective "but". It is important to recognize the rhetorical relations in dialogues, since just as in monologues, their truth-conditional entailments can help predict the next event, making the whole dialogue easy to understand (E_{tell_jokes} reinforces the truth of "interesting" in $E_{talking}$).

Apart from the intra-speaker relations *Elaboration* and *Contrast*, we can also find the interactions between different speakers: B_2 gives a *Comment* to A_1 by asking a question, inquiring the reason for the event $E_{talking}$. The pronominal reference "he" indicates that both speakers have the same grounding and "he" refers to the "this guy" in A_1. Then A's response A_3 to B_2 constructs a *Question Answer Pair* (QAP) relation. We can now construct an SDRT representation for this example, as shown in Figure 2.10, where inter-speaker relations are highlighted in orange and those within the same speaker in blue. Precision: this SDRT structure is constructed on the speech-turn level. A standard SDRT analysis should consider segmenting speech turns into `EDUs`.

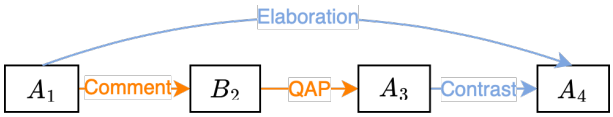


Figure 2.10: SDRT-structure of dialogue example (34).

According to Asher and Lascarides (2003), most rhetorical relations that apply to monologue can also be extended to conversational turns, but they require additional context assumptions. For instance, *Elaboration* and *Narration* must pertain to events that both speakers have observed or agreed upon. Additionally, relations such as *Parallel* and *Contrast* can also apply to dialogues, but they may lead to additional inferences, such as *dispute*, as exemplified in Section 2.2.2 with examples (22) and (23).

From Written to Spoken: Discourse theories discussed in this thesis are initially designed for written language. Recently, there are efforts in adapting these frameworks to spoken language for a more general analysis of discourse. Among these, Tonelli et al. (2010) adapted the PDTB annotation scheme for spontaneous conversation in Italian. Their study addresses two issues: (1) multi-lingual adaptation, they employed PDTB on Italian texts and conducted an analysis of the most common connectives employed, juxtaposing them with their English translation counterparts; (2) relation adaptation, they suggested that certain adaptations to the PDTB scheme were necessary to effectively address particular types of relations in spoken languages, such as implicit connections between non-adjacent arguments (quite often in dialogues). Other adjustments are about the sense hierarchy, typically at the second (or *type*) and third (*subtype*) levels in PDTB. For instance, they added *Goal* as a *type* under CONTINGENCY class, extending the original *Cause* and *Condition*. They also removed *List* in the EXPANSION class since they found that discourse in conversational speech is less structured and in well-written articles. Considering the significant influence of pragmatics in dialogues, they argued that the speaker's intention and implicit connections in a dialogue are fundamental to the discourse structure, thus providing more fine-grained senses for the third-level subtype *pragmatic*.

In a study by Rehbein et al. (2016), two discourse frameworks, PDTB and Cognitive approach to Coherence Relations (CCR) (Sanders et al., 1992, 1993), were compared for their ability to annotate discourse relations in spoken genres such as broadcast interviews and telephone conversations. The researchers found that explicit relations were more prevalent in spoken language compared to written text, and the interpretation of discourse connectives differed. For instance, the connective *so* is typically used for *causal* relation in written language, but often appears as a *conclusion* relation in spoken discourse (example (35)). To address these differences, the authors suggested new categories such as *Alternative Topicalisation* and *Alternative Stress* to express contrast in spoken language.

(35) **A$_1$** : I've already had a meeting hum an update meeting *so* the place hasn't burnt down or anything.

Finally, in a study by Wang et al. (2017a), RST was utilized as a foundation for manual annotations of discourse structure in non-native speakers' monologue speech during an English proficiency assessment (TOEFL). The aim was to examine features extracted from the annotated tree structure to assess discourse coherence and speech proficiency. RST framework was chosen over PDTB to acquire a complete discourse structure. The standard RST-DT annotation method was followed, where `EDUs` were first segmented, and then satellite and nucleus were identified before assigning relations. To handle special cases in speech, the authors created new discourse relations during the annotation process, including *disfluency*, *unfinished-utterance*, and *discourse particle* for filler words such as "you know" or "right". These relations are not strictly rhetorical nor do they convey a specific communicative intention, so they may not fit well within RST's relation inventory. Nonetheless, they address specific linguistic peculiarities in spoken language. As the authors suggest, it would be interesting to investigate how the features perform in an automatic RST parser, possibly trained on written text, and whether the features can be transferred from one text genre to another.

In this chapter, we have explored the theoretical foundations of discourse analysis. We started by discussing the fundamental elements of discourse, including discourse units, connectives, and relations, and then examined the different theories that link these elements together to create a complete discourse structure. We also surveyed several discourse corpora that have been

annotated under various frameworks, highlighting the differences in discourse across different languages and settings.

While the idea of creating a unified framework for discourse analysis is appealing, no effective framework has been established so far. Although proposals have been put forth in Benamara and Taboada (2015); Bunt et al. (2012), they have yet to gain widespread adoption. Even within the written language, there is still disagreement on the categories and number of coherence relations that should be distinguished. Studies presented in the last section have proposed various strategies to adapt from one framework to another or from one language setting to another. Recent efforts have been made to establish connections between different annotation frameworks, corpora, and languages, such as the DISRPT shared task, aimed at creating a unified format for all datasets. This initiative is a significant step towards developing a general and unified discourse annotation scheme. To promote discourse analysis in a broader range of NLP tasks, a unified framework is necessary that can be easily applied to various domains, encompassing both written and spoken language, in both monologue and dialogue settings.

Chapter 3

Discourse Parsing Models & Application for Downstream Tasks

Contents

In the previous chapter, we have learned that discourse examines the relationships between sentences in a document, and we have explored various corpora annotated under different frameworks. This chapter focuses on a particular discourse analysis task called discourse parsing. Discourse parsing aims to produce a comprehensive discourse structure for a given document, which involves connecting individual `EDUs` and assigning labels to their relations. The outcome structure can be advantageous for various NLP tasks, including summarization, sentiment analysis, and topic segmentation, which we will explore towards the end of this chapter.

This chapter is organized as follows: we begin by defining the discourse parsing task and outlining the main steps involved in the RST and SDRT frameworks in Section 3.1. Although RST and SDRT create different forms of structures, they generally follow the same steps. We then discuss various automatic discourse parsers proposed in the literature, which are made possible by the availability of annotated discourse corpora (discussed in Section 2.3). In Section 3.2, we explore different machine learning strategies for discourse parsing, including supervised learning, weakly supervised learning, and unsupervised learning approaches. Due to the scarcity of annotated data, several studies have explored transfer learning and multi-task learning methods. Some of these studies have served as inspiration for our research conducted in Chapter 7. We analyze the effectiveness of these parsing models, discuss their strengths and limitations, and compare their similarities and differences. In Section 3.3, we showcase the applications of discourse information in downstream natural language understanding (NLU) and natural language generation (NLG) tasks. Finally, we provide an in-depth analysis of the utilization of discourse information in downstream tasks, including the discourse features employed, the methods employed for their incorporation, and an evaluation of the performance of discourse-aware models.

3.1 Discourse Parsing Task

Generally speaking, both RST-style and SDRT-style discourse parsing can be divided into three steps:

(1) Discourse Unit Segmentation: Splitting a document into non-overlapping minimal discourse units, also known as `EDUs`.

(2) Link Attachment: Creating attachments among `EDUs`.

(3) Relation Prediction: Predicting a discourse relation for each pair of `EDUs`.

The final result of parsing is a relation-typed tree (RST-style) or graph (SDRT-style), where nodes represent discourse units and edges represent discourse relations, providing a comprehensive discourse structure of a document.

3.1.1 RST-Style Parsing

For RST, except for the three main steps, one more action is required after link attachment, that is to assign *nuclearity* for discourse units. *Nuclearity* tells which part is more important in a linked pair. Let us revisit an example (presented in Section 1.1) and its RST-style parsing structure in Figure 3.1. This text fragment consists of two sentences, which are segmented into four `EDUs`. The first two `EDUs`, denoted by e_1 and e_2, are connected by a mono-nuclear relation called *Consequence*, while e_3 and e_4 are linked by the relation *Circumstance*. In the parsing process, we determine which node is the *nucleus* ("N") and which is the *satellite* ("S") for every pair of nodes. The *nucleus* represents the most salient part of the local relation, while the *satellite*

plays a supplementary role. Note that there are both mono-nuclear types ("N-S" or "S-N") and multi-nuclear type ("N-N") relations are presented in this example. The parsing process can be performed in a *bottom-up* manner, where EDUs are first linked together to create intermediate nodes ($e_{1:2}$ and $e_{3:4}$), and gradually move up to the root. The final result is a binary tree-shaped structure.

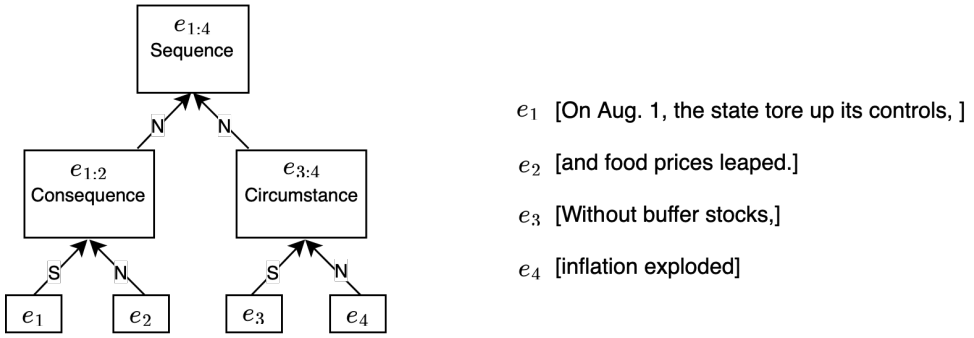


Figure 3.1: RST tree structure (left) for an example extracted from Wall Street Journal (*wsj_1146*) (right).

Two metrics have been employed for evaluating RST-style parsing. The first one is the standard Parserval metric, which originated from syntactic parsing (Black et al., 1991). It examines the label and word span of the parser output and compares it with the gold treebank. In the example illustrated in Figure 3.1, the gold Parserval has three text spans: $e_{1:2}$, $e_{3:4}$, and $e_{1:4}$. However, this metric is quite strict, as it does not distinguish between linguistically more or less significant errors, nor does it take into account cases where the label is accurate but the phrase boundary is slightly incorrect (Rehbein and van Genabith, 2007).

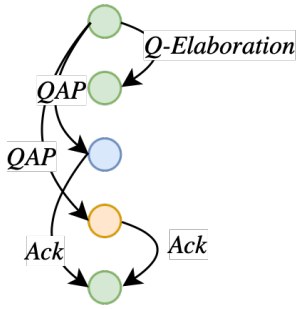
Another commonly used metric is RST-parseval proposed by Marcu (2000), which considers a larger set of nodes to collect all nuclearity and relation labels. All leaves (i.e., EDUs) are included except for the root node. Thus, in the case of Figure 3.1, 7 nodes would be considered for evaluation: three intermediate nodes ($e_{1:2}$, $e_{3:4}$, $e_{1:4}$) and four individual nodes (e_1, e_2, e_3, e_4). However, this metric has an artificial increase in accuracy since every EDU automatically has the correct nuclearity (*nucleus*) and the label (*span*). This convention artificially increases the accuracy for prediction, with four out of seven nodes being correctly predicted by default. As pointed out by Morey et al. (2017), RST-Parseval considers approximately twice as many nodes as the original Parseval would on binarized trees. Since a binarized tree with n EDUs has $n-1$ attachments, and RST-Parseval includes n leaves which results in $2n-1$ nodes. This lack of a unified evaluation metric makes the comparison among parsers difficult, with RST parsers either reporting RST-Parseval or original Parseval scores (or both). Morey et al. (2017) were the first to explicitly use an evaluation procedure for RST parsing that is closer to the original Parseval. They converted all metrics to the original Parseval and found that most gains reported are an artifact of implicit differences in evaluation procedures. They suggested that the original Parseval provides a more accurate picture.

3.1.2 SDRT-Style Parsing

SDRT framework represents DUs in embedded boxes (recall the salmon example (17) in Section 2.2.2), with intermediate boxes representing complex discourse units. However, in the STAC corpus annotation, both EDUs and CDUs are simplified as nodes. Each CDU node is linked to its

constituent EDUs with individual links, resulting in a weakly-connected graph structure[1]. When we mention SDRT-style parsing, we are referring to the representation of graph-like structures.

An example from the STAC corpus is displayed in Figure 3.2. In this example, three speakers "dmm", "inca", and "cheshireCatGrin" are discussing a potential trade of goods during a game. The direction of the links (such as $e_1 \rightarrow e_2$) indicates that e_1 is the *head* and e_2 is the *dependent*. In dialogues, most of the time, there are only *forward* links, i.e., in chronological order, since an utterance cannot be anaphorically or rhetorically dependent on following utterances, as they are previously unknown. This feature is known as the *turn constraint* (Afantenos et al., 2012b) in SDRT-style parsing. Compared to RST-style parsing, SDRT-style parsing demonstrates more flexibility. For example, the SDRT-style parser can establish connections between distant EDUs and allow non-projective links (such as crossing links between e_1 and e_4, and between e_3 and e_5), while RST-style parsing only permits adjacent attachments and restricts links to be projective.



e_1 dmm: I can give a sheep or wood for a wheat.

e_2 dmm: Any takers?

e_3 inca: Sheep would be good.

e_4 cheshireCatGrin: Not here.

e_5 dmm: Okay.

Figure 3.2: SDRT graph structure (left) for a dialogue example (right) from STAC corpus.

Two metrics are commonly used for evaluation: the *Unlabeled Attachment Score* (UAS), which only assesses link attachment without considering relations, and the *Labeled Attachment Score* (LAS), which evaluates whether both attachment and relation type are correctly predicted simultaneously. The latter is also referred to as *Full* performance, a similar assessment is adopted for syntactic dependency parsing. Initially introduced in Afantenos et al. (2015), the common practice is to compute the micro-F_1 score for UAS and LAS performances:

$$\text{Precision} = \text{TP}/\text{predicted links} \tag{3.1}$$

$$\text{Recall} = \text{TP}/\text{gold links} \tag{3.2}$$

$$\text{F}_1 = 2 * \text{Precision} * \text{Recall}/(\text{Precision} + \text{Recall}) \tag{3.3}$$

Previous research on SDRT-style parsing has mainly focused on predicting tree structures, as seen in works like Muller et al. (2012); Afantenos et al. (2015); Shi and Huang (2019). This approach employed a simplification in predicting tree structures instead of graphs and utilized algorithms such as Maximum Spanning Trees (with details in Section 3.2.1). The number of link attachments is thus fixed: for n EDUs, the model always predicts $n - 1$ links. As a result, the evaluation metric UAS actually refers to the recall score.

3.2 Machine Learning Strategies for Discourse Parsing

In this section, we provide an overview of various machine learning approaches for discourse parsers, which we present in the following sequence: **supervised** (Section 3.2.1), **transfer learn-**

[1]For more details on annotation, refer to https://www.irit.fr/STAC/stac_game_graphs/readme.html.

ing (Section 3.2.2), **weakly supervised** (Section 3.2.3), and **unsupervised** (Section 3.2.4) methods.

While it is an intriguing topic in itself, we do not focus on how to categorize different learning approaches in this chapter. Supervised models learn and predict data from the same domain. For transfer learning, we emphasize that knowledge is acquired (partially or extensively) from other tasks or domains. We classify three transfer learning methods: (1) distant supervision, where information is entirely borrowed from auxiliary tasks; (2) domain integration strategy, where the model is trained for the target task but in a different domain; (3) multi-task learning, where auxiliary tasks and the target task are trained together. In other literature, multi-task learning (MTL) is distinct from transfer learning[2]. Here we stress that MTL facilitates learning representations from other tasks, thereby serving as a means to *transfer* knowledge to our main parsing task. For weakly supervised learning, we emphasize the training set's quality, which is frequently noisy and imprecise. Studies employing this learning method often compromise the quality to obtain more annotated data for quantifying purposes. Lastly, if we consider the extreme scenario of transfer learning and learn with few or no labeled instances, we arrive at few-shot, one-shot, or zero-shot learning. Unsupervised learning, also known as zero-shot learning, is the last part of this section's presentation.

3.2.1 Supervised Methods

Our attention in this section is on SDRT-style parsing in dialogues since it is the primary focus of this thesis. We provide a comprehensive summary of supervised parsers to date, along with their performance on STAC and Molweni datasets, in Table 3.1. It is worth noting that some transformer-based parsers employ diverse pre-trained language models as their backbone, making it difficult to evaluate their impact on the final scores. To facilitate better comparison, we report the scores achieved using the base version.

We classify the existing systems based on two key aspects: (1) parsing paradigms (Section 3.2.1.1), whether they are graph-based or transition-based; (2) encoding strategies (Section 3.2.1.2), whether they employ separate encoding or joint encoding. We evaluate a dozen dependency parsers, taking into account their structure, performance, and any unique features they may have.

3.2.1.1 Parsing Paradigms

Existing models can be roughly categorized into graph-based approaches, as in Muller et al. (2012); Afantenos et al. (2015); Perret et al. (2016); Wang et al. (2021a)) and transition-based approaches (also known as sequential or incremental parsing), as in Shi and Huang (2019); Liu and Chen (2021); Yu et al. (2022). A novel way is to combine these two approaches, as done in Fan et al. (2022). We show in Table 3.2 a list of parsers in these three paradigms.

Graph-Based Approaches: The graph-based approach utilizes an edge-factoring algorithm that enables global parameter optimization over the entire tree structure (Sagae, 2009), such as the Maximum Spanning Tree (MST) algorithm (McDonald et al., 2005). For instance, the parser developed by Afantenos et al. (2015) is a representative example, where the authors first employ hand-crafted features to represent the `EDU` pairs f_{ij} and then use maximum entropy (MaxEnt) to estimate the parameters w_{ij}. The values for different parameters $\hat{w}$ are obtained by maximizing the log-likelihood of the training data T:

[2]Refer to Ruder's blog on transfer learning: `https://www.ruder.io/transfer-learning/`

Model	STAC		Molweni	
	Link	Link&Rel	Link	Link&Rel
MST (Afantenos et al., 2015)	68.8	50.4	69.0[§]	48.7[§]
ILP (Perret et al., 2016)	68.9	53.1	67.3[§]	48.3[§]
Deep Sequential (Shi and Huang, 2019)	73.2	55.7	77.3*	54.2*
Struct-Aware GNN (Wang et al., 2021a)	73.5	57.3	81.6	58.5
Hierarchical Transformer-based (Liu and Chen, 2021)	**75.3**[‖]	56.9[‖]	79.7[‖]	55.9[‖]
QA-DP Multi-task (He et al., 2021)	-	-	75.9[†]	56.0[†]
DiscProReco Multi-task (Yang et al., 2021)	74.1*	57.0*	-	-
Distance-Aware Multi-task (DAMT) (Fan et al., 2022)	73.6	57.4	82.5	58.9
SSP+SCIJE (Yu et al., 2022)	73.0	57.4	**83.7**	59.4
Struct-Joint (Chi and Rudnicky, 2022)	74.4	**59.6**	83.5	**59.9**

Table 3.1: Performance of SOTA supervised parsers on STAC and Molweni, micro F_1 scores. "Link" = unlabeled attachment score (UAS); "Link&Rel" = labeled attachment score (LAS).
Upper part parsers use traditional models. MST: maximum spanning tree; ILP: integer linear programming.
Lower part parsers use neural architectures for encoding and/or decoding. QA-DP: question-answering and discourse parsing; DiscProReco: discourse parsing and pronoun recovering; SSP: same-speaker-prediction; SCIJE: speaker-context interaction joint encoding.
§ results come from Chi and Rudnicky (2022). * results are extracted from Fan et al. (2022).
‖ results are taken from "+language backbone" RoBERTa-base setting in Liu and Chen (2021).
† results are from BERT_{base}, refer to He et al. (2021) for results with BERT_{large} and BERT_{wwm}.

Graph-based	Transition-based	Joint
MST, A* (Muller et al., 2012)	Deep Seq (Shi and Huang, 2019)	DAMT (Fan et al., 2022)
MST (Afantenos et al., 2015)	Hierarchical (Liu and Chen, 2021)	
ILP (Perret et al., 2016)	QA-DP Multi-task(He et al., 2021)	
Struct GNN (Wang et al., 2021a)	SSAM (Wang et al., 2021b)	
DiscProReco (Yang et al., 2021)	SSP+SCIJE (Yu et al., 2022)	
Struct-Joint (Chi and Rudnicky, 2022)		

Table 3.2: Graph-based, transition-based, and joint discourse parsers for dialogues.
Left column: MST: maximum spanning tree; A*: decoding strategy, shortest-path searching; ILP: integer linear programming; GNN: graph neural network; DiscProReco: discourse parsing and pronoun recovery multi-task.
Middle column: Deep seq: *Deep Sequential* model; QA-DP: question-answering and discourse parsing multi-task setting; SSAM: structure self attention model; SSP+SCIJE: same-speaker-prediction and speaker-context interaction joint encoding.
Right column: DAMT: distance-aware multi-task.

$$P_{ij} = \frac{1}{Z(c)} \exp \sum_{ij=1}^{m} w_{ij} f_{ij} \tag{3.4}$$

$$\hat{w} = \arg\max_{w} \sum_{ij}^{T} \log P_{ij} \tag{3.5}$$

where ij is a pair of `EDU` and m is the number of features. In the decoding step, they use the Chu-Liu Edmonds (Chu, 1965; Edmonds, 1968) version of the MST algorithm which examines all possible tree structures and chooses the one that has the biggest sum of weight probabilities.

$$T^* = \arg\max_{T \subset G} \sum_{e \in E(T)} w(e) \tag{3.6}$$

$$w(e) = \log(\frac{p(e)}{1 - p(e)}) \tag{3.7}$$

where G is the complete graph of all possible edges and $E(T)$ contains all the edges in candidate tree T. This is very similar in syntactic parsing when the dependencies are established within sentences (Muller et al., 2012; Li et al., 2014c).

Transition-Based Approaches: In contrast, transition-based methods prioritize local optimality by selecting the best action at each step. A typical example is the *Deep Sequential* parser created by Shi and Huang (2019). In their method, after obtaining a structured global representation of each pair of `EDUs`, including both current and previously attached links, denoted as $H_{i,j}$, the link predictor calculates the probability that each `EDU` u_j with is the antecedent of u_i, with $(j < i)$:

$$P(p_i = u_j | H_{i,<i}) = \frac{\exp o_{i,j}}{\sum_{k<i} \exp o_{i,k}} \tag{3.8}$$

$$p_i = \arg\max_{u_j : j<i} P(p_i = u_j | H_{i,<i}) \tag{3.9}$$

where $o_{i,j}$ is the vector representation of pair attachment i, j. Notice the difference between Equations 3.6 and 3.9: the former stresses on the probabilities of all edges, whereas the latter considers historical structures to make the current decision and selects the local maximum at each step.

In terms of time complexity, transition-based approaches are normally quicker – they can finish in linear time, while graph-based algorithms such as Chu-Liu Edmonds have a complexity of $O(n^3)$, with n being the number of `EDUs`. However, one major drawback for transition-based is the error propagation issue, as discussed in Wang et al. (2021a).

Joint Framework: Fan et al. (2022) firstly proposed a joint model (DAMT) that combines the benefits of graph-based and transition-based paradigms. To construct the connection between the transition-based and graph-based semantic representation (H_t and H_g), they used Unidirectional Cross Attention (UCA) layers to create new representations of H_{tc} and H_{gc} in the encoding module:

$$H_{g\to t} = \mathrm{UCA}(W_q H_g, W_k H_t, W_v H_t) \tag{3.10}$$
$$H_{gc} = \mathrm{L}(H_g + H_{g\to t}) \tag{3.11}$$
$$H_{t\to g} = \mathrm{UCA}(W_q H_t, W_k H_g, W_v H_g) \tag{3.12}$$
$$H_{tc} = \mathrm{L}(H_t + H_{t\to g}) \tag{3.13}$$

where W_q, W_k, W_v are the weight matrices for query, key, and value that map vectors to the same feature space; $L(.)$ is a layer normalization function. Structure Self Attention (SSA) is then applied to H_{tc} and H_{gc} to incorporate structural information of conversation. For the decoding part, they use a pointer network and transition-based process to obtain the probability s_{ij} between the current `EDU` i and previous `EDU` j.

$$s_{ij} = H_{tc}^{\top} W h_{dk} + U H_{tc} + V h_{dk} + b \tag{3.14}$$

where W is the weight matrix of bi-linear term; U and V are two weight vectors of the linear term; b is the bias vector; h_{dk} is the k^{th} step output of a Biaffine Attention mechanism (Dozat and Manning, 2016) of the input H_{tc}. Finally, in a multi-task learning setting, the authors aim to minimize the sum of losses of both encoding and decoding, thus integrating the two dependency parsing paradigms. The evaluation demonstrates notable enhancements, particularly for long-distance dependency links. The authors attribute this improvement to the fusion of the transition-based module, which performs better for a link distance greater than one, and the graph-based module, which is competitive with other state-of-the-art parsers for a distance of one.

3.2.1.2 Encoding & Decoding Strategy

We will now examine parsing models by evaluating their encoding and decoding strategies. Initially, approaches used feature engineering techniques to encode `EDU` pairs by incorporating lexical and positional information, and then adopted various decoding strategies such as the Maximum Spanning Tree algorithm (Muller et al., 2012; Li et al., 2014c; Afantenos et al., 2012b) or Integer Linear Programming (Perret et al., 2016). With the introduction of the first neural model *Deep Sequential* (Shi and Huang, 2019), feature engineering has received less attention, and researchers have instead employed Recurrent Neural Networks (such as GRU) (Shi and Huang, 2019; Liu and Chen, 2021; Yu et al., 2022), Graph Neural Networks (Wang et al., 2021b,a; Yang et al., 2021), or Pre-trained Language Models (PLMs) (Liu and Chen, 2021; Yu et al., 2022) to encode contextual information. Additionally, some have used multi-task learning frameworks, such as question-answering or pronoun recovery, to obtain representations, as in Yang et al. (2021); He et al. (2021); Fan et al. (2022).

Most of the aforementioned work treats link attachment and relation prediction as two distinct tasks, with the link predicted before the relation. We categorize these studies under the *Sequential Prediction Group*. In contrast, studies that jointly optimize link attachment and relation prediction are referred to as the *Joint Prediction Group*. Table 3.3 gives a summarization of information about the encoders and decoders in these two groups.

Sequential Prediction: *Group Sequential* contains models that treat link and relation prediction as two separate tasks. The initial work by Afantenos et al. (2015) used traditional feature engineering techniques, mainly incorporating lexical features (such as opinion markers, quantifiers, punctuation presence) and positional features (e.g., the distance between `EDUs`, position in

Model	Encoder	Decoder	Feature	Highlight	Corpus
(a) Group Sequential Link & Rel					
MST (Afantenos et al., 2015)	local	MST	lexi, posit, synt	-	STAC
Deep Seq (Shi and Huang, 2019)	local, global, struct	multiclass classif	-	speaker	STAC
Hierarchical (Liu and Chen, 2021)	global, struct	multiclass classif	-	-	STAC, Molweni
SSP+SCIJE (Yu et al., 2022)	global, struct	multiclass classif	-	speaker	STAC, Molweni
Struct GNN (Wang et al., 2021a)	global	multiclass classif	lexi, posit	-	STAC, Molweni
SSAM (Wang et al., 2021b)	local	/	/	cohesion	STAC
DAMT (Fan et al., 2022)	global, struct	multiclass classif	-	multi-task	STAC, Molweni
(b) Group Joint Link & Rel					
MST, A* (Muller et al., 2012)	local	MST, A*	lexi, synt	-	ANNODIS‖
ILP (Perret et al., 2016)	local	ILP	lexi, posit, synt	-	STAC
Struct-Joint (Chi and Rudnicky, 2022)	global, struct	MST	-	-	STAC, Molweni

Table 3.3: Encoder, decoder, and feature engineering in SOTA dependency parsers.
Model column, SSP+SCIJE: same-speaker-prediction and speaker-context interaction joint encoding; SSAM: structure self attention model; DAMT: distant-aware multi-task model.
Feature column, "lexi": lexical; "synt": syntactic features such as dependency and dialogue act tagging; "posit": positional features such as distance between `EDUs` and the position of the first utterance of the speaker. '-': None; '/': not available.
Highlight column shows if the parser highlights any extra information or jointly learns with other tasks.
Corpus column are the testing corpus in each study. ‖ ANNODIS contains newspaper and Wikipedia articles annotated under SDRT framework in French (Afantenos et al., 2012a).

the dialogue), as well as coarse-grain dialogue act tags (*offer*, *refusal*, etc.) and syntactic dependencies using the Stanford CoreNLP pipeline (Manning et al., 2014). They utilized a maximum entropy model (Berger et al., 1996) as an encoder to estimate the maximum parameters $\hat{w}$ for each pair of `EDUs`, where each `EDU` is represented by a feature vector. For the decoder, they utilized a Maximum Spanning Tree algorithm to obtain the tree with the highest probability for all edges.

The *Deep Sequential* architecture proposed by Shi and Huang (2019) was the first neural architecture based on a hierarchical Gated Recurrent Unit (GRU) that processes segment attachment and relation allocation sequentially. To encode the input, Shi and Huang (2019) used a combination of different representations, including local representations (h_i), non-structured global representations (g_i^{NS}, g_j^{NS}), and structured global representations (g_{j,a_i}^{S}), as shown in Equation 3.15. Non-structured vectors were obtained from the output of an encoder based on Gated Recurrent Units (GRUs) that processed the `EDU` sequence, while structured vectors also incorporated information about previous dependency links and relation types. Additionally, the authors proposed a *Speaker Highlighting Mechanism* that considers speaker information (a_i):

$$H_{i,j} = h_i \oplus g_i^{NS} \oplus g_j^{NS} \oplus g_{j,a_i}^{S} \tag{3.15}$$

The decoding process is performed incrementally, involving multiple choices where the current `EDU` selects its parent with the highest probability, attaches to it, and then determines the most probable relation type. This method resulted in significant improvements in parsing accuracy for STAC (+6%) and Molweni (+7%). The primary advantage of this approach lies in the encoding of global information. Although link and relation predictions are made separately at each step, the previous relation choice is taken into account through the global structured representation, which aids in making subsequent decisions.

The SSP+SCIJE model (Yu et al., 2022) is a recent extension of the *Deep Sequential* model.

In this model, the authors begin by pre-training a language model on the Same-Speaker Prediction (SSP) task. They then incorporated the resulting information ($h^s_{i,j}$) into the `EDU` pair encoding:

$$H_{i,j} = \alpha h^s_{i,j} + (1 - \alpha) h^u_{i,j} \tag{3.16}$$

where α is a hyper-parameter; $h^s_{i,j}$ is the representation from SSP pre-trained model; $h^u_{i,j}$ is the concatenation of representation of two `EDUs`. From Equations 3.15 and 3.16, it is evident that both encoding processes incorporate speaker information (g^S_{j,a_i} and $h^s_{i,j}$). However, the *Deep Sequential* model additionally considers individual `EDU` representation, which is not taken into account by the SSP+SCIJE model.

In a recent study, Wang et al. (2021a) utilized Graph Neural Networks (GNNs) to learn the structure between each pair of `EDUs`. Instead of focusing on `EDU` representation, they explored edge-specific vectors to capture the implicit structure information between `EDU` pairs. To initialize edge vectors, they encoded features such as "if-the-same-speaker", "if-continuous-utterance", and "distance between two `EDUs`". However, this structured GNN model only showed marginal improvement for link prediction on STAC compared to the *Deep Sequential* model (F_1 73.5% vs 73.2%).

Another GNN-based model was proposed by Wang et al. (2021b), where they incorporated additional cohesion information into `EDU` encoding using the WordNet resource and a coreference resolution model to extract lexical and coreference chains. However, it is unclear whether they considered historical decisions and how they carried out the decoding part for link and relation.

Joint Prediction: *Group Joint* contains models that simultaneously optimize link attachment and relation prediction. In an early work by Muller et al. (2012), authors have proposed to jointly calculate the loss for link attachment and relation types:

$$W_{u,v} = -log(P(attach(u,v) = True) \times \max_R P(R|attach(u,v) = True)) \tag{3.17}$$

However, when evaluated on the SDRT-style French corpus ANNODIS (Afantenos et al., 2012a), their results showed poorer performance compared to their sequential model for link attachment and similar performance for full structure prediction. Thus, Muller and colleagues concluded that predicting relations does not improve link attachment.

Another study proposed by Perret et al. (2016) explored the use of Integer Linear Programming for joint decoding, where an objective function is defined based on the scores of attachment and relation:

$$\sum_{i=1}^{n}\sum_{j=1}^{n}(a_{ij}s_a(i,j) + \sum_{k=1}^{m} r_{ijk}s_r(i,j,k)) \tag{3.18}$$

where a_{ij} and r_{ijk} are binary variables for link and relation: equals to 1 if link (ij) or relation k for link (ij) are correct, else 0; $s_a(i,j)$ and $s_r(i,j,k)$ are the scores of attachment and relation obtained from feature engineering. Maximizing this objective function is, in fact, learning the best combination of link and relation.

In addition to the previously mentioned work, A recent study by Chi and Rudnicky (2022) introduced a structured encoding approach where link attachments and relation predictions are jointly optimized on an adjacency matrix. They achieved this by constructing each pair of `EDUs` as a triplet (h, m, r) where h and m represent the indices of the parent and child utterances,

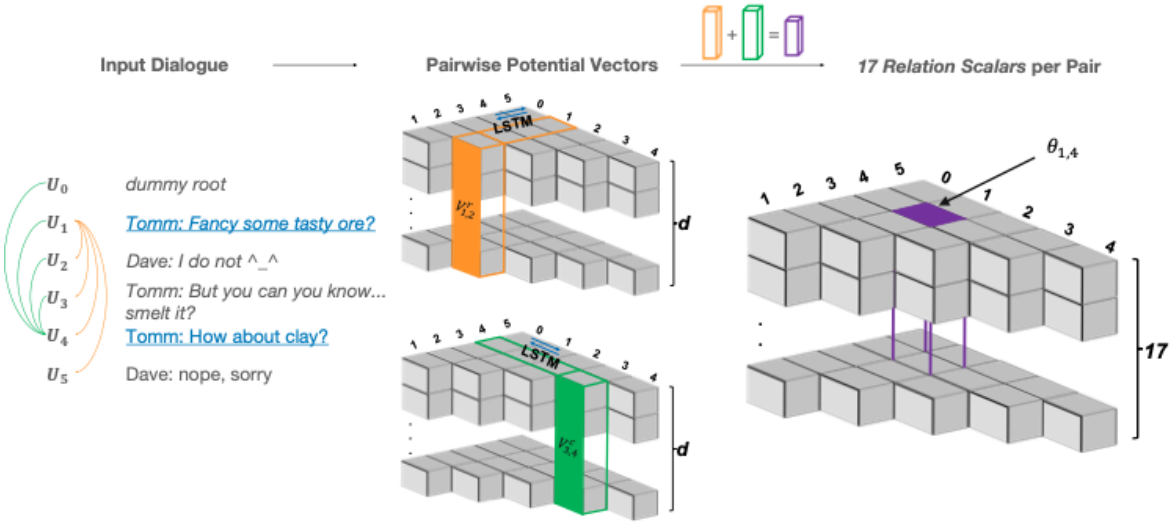


Figure 3.3: Contextual joint encoding process in Chi and Rudnicky (2022). Orange rows mean the 1^{st} utterance can connect to later utterances to choose which one is the child; green columns mean the 4^{th} utterance can have one previous utterance as its parent. V^r and V^c are concatenated together and pass through a linear transformation to obtain the purple vector θ.

respectively, and r represents one of the 17 (16 relations in STAC + *no-relation*) relation types between the two. Figure 3.3 illustrates this process. For a given row h (in orange, where $h = 1$, parent node), the hidden states for all timesteps t that follow h (rows $2 - 5$) are computed and stored in $V^r_{h,t}$. Similarly, the column m^{th} representation (in green, where $m = 4$, child node) is computed by considering all the previous columns:

$$\{V^r_{h,t}\}^n_{t=h+1} = \text{LSTM}(\{V^r_{h,t}\}^n_{t=h+1}) \tag{3.19}$$

$$\{V^c_{t,m}\}^{m-1}_{t=0} = \text{LSTM}(\{V^c_{t,m}\}^{m-1}_{t=0}) \tag{3.20}$$

The transformation of $V^r_{h,t}$ and $V^c_{t,m}$ into individual scores with relation information (as shown in purple) is accomplished by applying a linear transformation layer. This conversion changes the dimensions of V from $\mathbb{R}^{(n+1)\times(n+1)\times 2d}$ to $\mathbb{R}^{(n+1)\times(n+1)\times 17}$, where n is the total number of utterances in a document and d is the token dimension:

$$\theta_{h,m} = \text{Linear}(V^r_{h,m} + V^c_{h,m}) \tag{3.21}$$

With parameterization in Equation 3.21, each `EDU`-pair is aware of neighboring pairs as well as the relation types. For the decoding part, they applied Chu-Liu Edmonds algorithms. The novelty in Chi's work is to directly transform the parent-child vector into a relation-aware vector (Equation 3.21), enabling the joint prediction objective.

To conclude, from Table 3.1, we observe that traditional models (upper part) are significantly outperformed by recent neural models (lower part). Most supervised neural models achieve around 73% and 57% UAS and LAS performances, respectively. Among neural models, the Transformer-based *Deep Sequential* model (Liu and Chen, 2021) achieves the highest UAS score (75.3%), and the joint structured model (Chi and Rudnicky, 2022) obtains the best LAS

Model	Setting	Framework	Output	Synergistic task	Corpus
Distant supervision					
Huber and Carenini (2019)	monologue	RST	struc	sentiment analysis	RST-DT, Instr-DT
Huber and Carenini (2020b)	monologue	RST	struc, nucl	sentiment analysis	RST-DT, Instr-DT
Xiao et al. (2021)	monologue	RST	struc	summarization	RST-DT, Instr-DT, GUM
Jiang et al. (2021a)	monologue	RST	struc, nucl, rel	topic segmentation	RST-DT, MCDTB‖
Domain integration					
Liu and Chen (2021)	dialogue	SDRT	struc, rel	-	STAC, Molweni
Multi-task learning					
Nejat et al. (2017)	monologue	RST	struc, nucl, rel	sentiment analysis	RST-DT
Yang et al. (2021)	dialogue	SDRT	struc, rel	dropped pronoun recovery	STAC, SPDPR†
He et al. (2021)	dialogue	SDRT	struc, rel	machine reading comprehension	Molweni
Fan et al. (2022)*	dialogue	SDRT	struc, rel	-	STAC, Molweni

Table 3.4: Transfer learning strategies in discourse parsing. Output column: for RST: {structure, nuclearity, relation}; for SDRT: {structure, relation}. ‖ MCDTB: Macro Chinese Discourse Treebank (Jiang et al., 2018). † SPDPR: Structure Parsing-enhanced Dropped Pronoun Recovery dataset is a corpus containing 684 multi-party SMS chat files in Chinese (Yang et al., 2021). *: work already presented in Section 3.2.1 but also fit in transfer learning category. -: not applicable.

score (59.6%). The adoption of language backbones significantly elevates scores compared to the original *Deep Sequential* model (Shi and Huang, 2019). However, the differences in adopting different parsing paradigms (graph-based and transition-based) and encoding strategies (*Sequential* or *Joint*) are not obvious.

In contrast to syntactic dependency parsing, supervised models for discourse dependency parsing lag behind in performance[3]. The primary reason for this is the lack of annotated data. The training set is restricted in size and domain, making supervised models trained on STAC (and Molweni) hard to generalize to other domains. Even with domain integration strategies, the study by Liu and Chen (2021) (that we will present in Section 3.2.2.2) shows that inter-domain performance drops by approximately 20% for both UAS and LAS, indicating that supervised models are not yet suitable for wide usage.

3.2.2 Transfer Learning Methods

To tackle data scarcity in discourse parsing, there has been a recent trend towards transfer learning strategies, which has mainly focused on monologues. In this section, we describe three methods to achieve the goal of information transfer: discovery of distant signals from other tasks (Section 3.2.2.1), joint pretraining to help the model adapt to another domain (Section 3.2.2.2), and learning shared representation in a multi-task framework (Section 3.2.2.3). We summarize these studies in Table 3.4, including the domain, framework, model output, auxiliary task used, and testing corpora. Precision: in the previous section, we have already discussed the work of Fan et al. (2022), which uses a multi-task framework to combine graph-based and transition-based parsing paradigms. Since both tasks are discourse parsing and they do not leverage information from another different task, we classify their work in supervised learning. Nonetheless, we still include it under "Multi-task learning" in Table 3.4.

[3]On the leaderboard for syntactic dependency parsing, top parsers on the Penn Treebank achieve > 95% UAS and LAS scores. Please refer to `https://paperswithcode.com/sota/dependency-parsing-on-penn-treebank`.

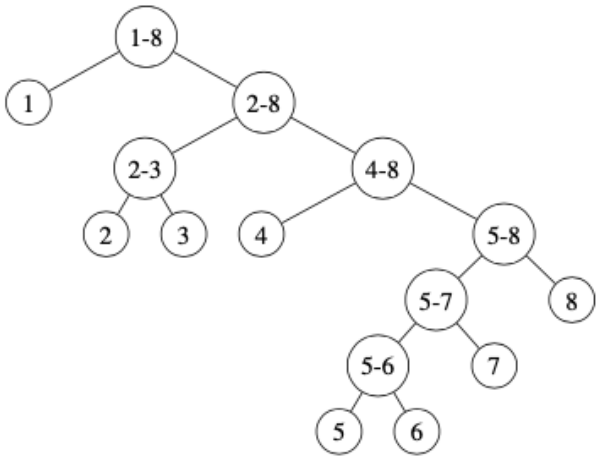


Figure 3.4: Example of the constituent tree for a strongly negative review in Yelp'13 corpus (Tang et al., 2015), from Huber and Carenini (2019). This review contains 8 `EDUs`: [Panera bread wannabes.]$_1$ [Food was okay and coffee]$_2$ [was eh.]$_3$ [Not large portions for the price.]$_4$ [The free chocolate chip cookie was a nice touch]$_5$ [and the orange scone was good.]$_6$ [Broccoli cheddar soup was pretty good.]$_7$ [I would not come back.]$_8$

3.2.2.1 Distant Supervision

The work of Huber and Carenini (2019) drew inspiration from prior studies that used discourse parsing to improve sentiment analysis (Bhatia et al., 2015; Nejat et al., 2017). In contrast, they investigated the potential synergy between sentiment and discourse by exploring the use of multiple-instance learning (MIL) (Angelidis and Lapata, 2018) techniques. Their approach involved smoothing the gold global sentiment label (i.e., the sentiment of the entire document) to local sentiment and attention scores at the `EDU` level, which were then used to construct a discourse tree using a chart-based algorithm like CKY (Jurafsky and Martin, 2014). Figure 3.4 from their paper provides an example, in which a strong negative review in the Yelp'13 (Tang et al., 2015) corpus (food reviews) is parsed. The last `EDU` in the review has the most negative sentiment and is placed at a higher level in the tree, while `EDUs` with positive sentiments, such as 5, 6, and 7, are located at lower levels. Thus, the hierarchical sentiment structure aligns with discourse importance and can be represented in a tree-like form.

The pipeline for generating discourse trees can be divided into two main stages. In the first stage, during training, the input text is segmented into `EDUs` and passed through a hierarchical RNN network. The output of this stage is a vector representation for each `EDU`$_i$, which is used to obtain sentiment scores (S_E) and attention scores (A_E).

$$S_{E_i} = \text{Sigmoid}(\text{FF}(H_{E_i})) \tag{3.22}$$

$$A_{E_i} = \text{Sigmoid}(H_{E_i}) \tag{3.23}$$

where H_{E_i} is hidden-state of `EDU` i and $\text{FF}(\cdot)$ is a feed-forward layer. These two scores are summed up to calculate the final sentiment prediction (O_D) of the MIL model for document D:

$$O_D = \sum_{E_i \in D} S_{E_i} * A_{E_i} \tag{3.24}$$

Next, the MIL model's parameters are optimized by comparing the predicted scores with the gold label. After training the model, step (2) involves using the sentiment scores of each `EDU` to construct a discourse tree using the CKY algorithm. This step is simpler, but it involves computing the scores for all possible tree structures.

The advantages become evident when testing in inter-domain settings: the parsers trained on the Yelp'13 corpus enriched with discourse information outperformed those trained on another human-annotated corpus in a different domain. Specifically, the authors trained a two-stage parser (Wang et al., 2017b) on RST-DT and tested it on Instr-DT, achieving a precision of 73.7; and vice versa, achieving a precision of 74.5. When they trained a two-stage parser on Yelp'13 and tested on Instr-DT and RST-DT, they achieved 74.2 and 77.2, respectively, resulting in a gain of 0.5 and 2.7 points, respectively. This demonstrates the ability to capture more general discourse structure from sentiment information. Their method has the advantage of creating large-scale "silver standard" discourse trees for more general usage, thanks to the abundance and accuracy of sentiment-rich datasets such as Yelp'13 food reviews.

However, there are several drawbacks to this study that should be noted. Firstly, the parser has a limited scope, as it can only generate discourse structure and not nuclearity and relations. While the nuclearity prediction was improved in follow-up work MEGA-DT (Huber and Carenini, 2020b) and Weighted-RST (Huber et al., 2021), the relation prediction remains an unresolved issue. Secondly, the method has limited applicability. As stated in the paper, due to computational power constraints and the non-scalable nature of the CKY algorithm, the authors were only able to process documents with ≤ 20 `EDUs` and were unable to consider inter-sentence relations.

In their follow-up work MEGA-DT (Huber and Carenini, 2020b), authors used averaged attention values a and polarity scores p from the left and right subtrees for internal nodes. For the mono-nucleus class (*N-S* or *S-N*), they assign N to the subtree with a larger a value and S to the node with a lower value. They also created an artificial node *N-N* to tackle multi-nucleus classes. The results, however, are not satisfactory: with parser over-predicted multi-nucleus nodes and low accuracy for the mono-nucleus classes. Finally, authors have proposed a new perspective on nuclearity prediction in Huber et al. (2021) where they argue that binary assessment of this attribute can be replaced by real-valued scores, the so-called "Weighted-RST" framework. They show that the distantly learned weighted discourse trees can better benefit some downstream applications.

After analyzing this line of study, we think that the exploration of better nuclearity inference is valuable since this attribute encodes local importance in a document. Linguistic features, such as discourse markers, can provide assistance in this regard. For instance, connectives like *but* and *however* often imply the emergence of a more significant utterance (*S-N*), while connective *and* indicate equal importance (*N-N*). Regarding relation prediction, it is still a challenging task for both monologues and dialogues, and the authors did not propose a proper solution. In our experiments, we made the initial attempt to predict relations for dialogues, as detailed in Chapter 8.

Another signal for distant supervision in discourse parsing can be obtained from the attention matrices in neural summarizers (Xiao et al., 2021). In this work, authors suggested that attention matrices in summarizers contain structural information that can be used to extract discourse trees. The authors first trained a summarizer and then extracted discourse trees from the summarizers' attention matrices. These trees can be considered a by-product of the summarization task. To extract the trees, they used hierarchical CKY (Jurafsky and Martin, 2014) algorithm for constituent trees, and hierarchical Eisner (Eisner, 1996) and Chu-Liu Edmonds (Chu, 1965; Edmonds, 1968) algorithms for dependency trees. They also experimented with

layer-wise attention, which averages all the attention heads in a layer (the first two layers were tested). However, in our own experiments described in Chapter 7, we examined both layer-wise and head-wise attentions. As a teaser, our results showed that layer-wise aggregation of attention scores is not the best approach for generating discourse trees.

3.2.2.2 Domain Integration

Previously discussed studies have attempted to incorporate discourse information from other tasks, while Liu and Chen (2021) focused on transferring information across different domains[4]. First, the researchers demonstrated that a parser trained solely on STAC performs poorly on Molweni, and vice versa. Consequently, they proposed a solution to the domain integration problem by conducting joint training on both datasets. They modified the supervised *Deep Sequential* model (Shi and Huang, 2019) by incorporating a few changes: (1) they used Pre-trained Language Models as the backbone to encode `EDUs` and, in turn, improved the local `EDU` representation; (2) leveraged Masked Language Modeling (MLM) with joint STAC and Molweni data during model pre-training to enhance domain coverage; and (3) utilized vocabulary refinement techniques to eliminate infrequently occurring vocabulary in both datasets. By using the language model backbone and conducting joint pre-training, the researchers achieved approximately 2% and 10% improvements in cross-domain results (training on STAC and testing on Molweni, and vice versa). It is worth noting that language model implementation contributes the most to these improvements among the three modifications. However, due to the predominantly lexical nature of adaptation strategies, the improved outcomes still fall short of both simple baselines and our semi-supervised outcomes (Chapter 7).

3.2.2.3 Multi-Task Learning

A recent approach to discourse parsing is to incorporate related tasks and leverage shared representations using the Multi-Task Learning (MTL) framework. Here, we discuss two relevant studies in this regard.

Regarding SDRT-style parsing, Yang et al. (2021) suggested utilizing Dropped Pronoun Recovery (DPR) as an additional task. Dropping pronouns like "你*[you]*" and "我*[I]*" is a common phenomenon in oral Chinese conversations. This task is aimed at restoring the dropped pronouns in a conversation. They tested on a Chinese conversational dataset which contains 684 SMS dialogues (Yang et al., 2015). They annotated the dataset under the SDRT framework and obtained $39k$ relations. The idea behind this work is that discourse parsing offers information on linked utterances, which can assist in pronoun recovery. Conversely, the recovered pronouns can complete utterances and be beneficial for discourse parsing. The link attachment part is very much similar to that of Fan et al. (2022) (see Equation 3.14 in Section 3.2.1). They also use a Biaffine Attention Network to obtain the probability of attaching current `EDU` (X_j) to a previous parent (X_i):

$$s_{i,j} = r_i^{(\text{head})} U^{(\text{arc})} r_j^{(\text{dep})} + r_i^{(\text{head})^\top} u^{(\text{arc})} \tag{3.25}$$

$$P_{\text{arc}}(X_j | X_i, C) = \text{softmax}(s_{i,j}^{(\text{arc})}) \tag{3.26}$$

[4]Technically speaking, this work can also be classified as supervised learning. However, the training and test sets were not in the same domain, and the authors leveraged information from another dataset during training. Hence, this work is classified as transfer learning.

The utterance-specific states for the head and dependent are denoted as $r_i^{(\text{head})}$ and $r_i^{(\text{dep})}$, respectively, while $U^{(\text{arc})}$ and $u^{(\text{arc})}$ refer to the weight matrix and bias. The context is represented by C. The prediction of an arc is accomplished using a softmax function, and a relation score distribution $s_{i,j}^{(\text{rel})}$ is calculated for each pair of utterances (X_i, X_j). In the auxiliary task, utterances augmented with discourse structure are utilized for pronoun referent. The overall training objective is to minimize the loss from link and relation prediction, as well as from DPR in a joint manner:

$$loss = \alpha \cdot (\text{loss}_{\text{arc}} + \text{loss}_{\text{rel}}) + \beta \cdot \text{loss}_{\text{dpr}} \tag{3.27}$$

Another study by He et al. (2021) explored the joint training of discourse structure prediction and dialogue comprehension tasks. The evaluation was performed on the Molweni dataset. In contrast to Yang et al. (2021), where the updated utterance representation from discourse parsing is directly passed to the next task, He et al. (2021) initially constructed a representation for the dialogue comprehension task based on QA and then adapted this representation for discourse parsing:

$$S = \text{encode}([\text{CLS}], \text{Q}, [\text{SEP}], \text{D}, [\text{SEP}]) \tag{3.28}$$

$$F_{ij} = (E_{\text{sep}}^i, E_{\text{sep}}^j, E_{\text{sep}}^i - E_{\text{sep}}^j, E_{\text{sep}}^i \cdot E_{\text{sep}}^j) \tag{3.29}$$

In the QA-based dialogue comprehension task, utterances are not encoded in a linear fashion. Instead, they are encoded in the order specified in Equation 3.28, where the question $(Q = w_1 w_2 \dots w_n)$ is followed by [SEP] token and then the dialogue context (i.e., an utterance $D = w_1 w_2 \dots w_m$). This way, each utterance is encoded with information from the question. The resulting utterance features are then passed to the discourse parsing task, as shown in Equation 3.29, where E_{sep}^i is the output feature of the separator for the i^{th} utterance, and $-$ and $\cdot$ represent Euclidean and cosine distances, respectively. The representation for a pair of `EDUs` (ij) is defined by combining the individual `EDU` representations and distance representations.

The training for link and relation prediction in discourse parsing is executed sequentially. While QA in dialogues can benefit DP, particularly for relation types such as QAP, the model performance is not very impressive: 75.9% for link prediction (versus 77.3% using *Deep Sequential* model (Shi and Huang, 2019)) and 56.0% for joint link and relation prediction (versus 54.2% with *Deep Sequential*). The representation of `EDU` pair F_{ij} is quite complex, and it is not clear how question-encoded (Q) utterances would help in finding the appropriate parent node in discourse parsing.

To conclude, in this section, we have discussed several transfer learning approaches that can be utilized to address the issue of data scarcity, ranging from leveraging distant signals in other tasks (Huber and Carenini, 2019; Xiao et al., 2021), to joint training of cross-domain datasets (Liu and Chen, 2021), and finally multi-task learning (He et al., 2021; Yang et al., 2021). Each approach has its own unique use case, depending on the availability of resources and the relevance of the tasks. It is important to note that only closely related tasks can fully benefit from each other. When searching for transferable signals, the source and target settings should also be taken into consideration. For example, while sentiment-augmented datasets are rich in monologues (such as movie and food reviews), they are less commonly seen in dialogues. Although sentiment annotation for individual speech turns may be available, overall judgments for entire conversations are rare. Therefore, the distant learning strategy employed in Huber and Carenini (2019) may not be easily transferable to dialogues. Nonetheless, studying these strategies can provide inspiration and help identify useful tasks that can benefit discourse parsing.

Model	Setting	Framework	Output	Strategy	Corpus
Badene et al. (2019a)	dialogue	SDRT	structure	heuristic rules	STAC
Mihăilă and Ananiadou (2014)	monologue	-	connective	self-training	BioCause[§]
Nishida and Matsumoto (2022)	monologue	RST	struct, nucl, rel	bootstrapping	CORD-19*, COVID19-DTB[‖]
	dialogue	SDRT	struct, rel	bootstrapping	UDC†, Molweni
Chapter 8	dialogue	SDRT	relation	self-training	STAC

Table 3.5: Weakly supervision strategies in discourse parsing. Output column: for RST: {structure, nuclearity, relation}; for SDRT: {structure, relation}.
BioCause[§]: biomedical corpus annotated with causal discourse relation (Mihăilă et al., 2013). CORD-19*: COVID-19 open research dataset (Wang et al., 2020). COVID19-DTB[‖]: COVID-19 corpus proposed in the study (Nishida and Matsumoto, 2022). UDC†: Ubuntu Dialogue Corpus (Lowe et al., 2015). Note that Nishida and Matsumoto (2022) use SciDTB (Yang and Li, 2018) and STAC (Asher and Lascarides, 2003) in monologue and dialogue settings resp. for training.

3.2.3 Weakly Supervised Methods

Instead of relying solely on transfer learning, another approach to address data scarcity is to employ weakly supervised methods. This involves sacrificing some level of quality in exchange for a greater quantity of annotated data, which may be noisier. In this section, we discuss weakly supervised strategies. Table 3.5 provides a summary of related studies.

In dialogue settings, Badene et al. (2019a,b) investigated a weak supervision paradigm where expert-composed heuristics, combined with a generative model, are applied to unseen data. They used a data programming paradigm – introduced by Ratner et al. (2016) with the Snorkel framework (Ratner et al., 2017) to create attachment signals. Precisely, their pipeline includes two steps: (1) Labeling Functions (short in LF) where expert-composed attachment rules are created. For instance, to make an attachment of relation *Result*, one of the rules is to match the starting word of the second *EDU* to a pre-defined result word list (including "so", "accordingly", "as a result", etc.). If a candidate `EDU` contains a result word, then the LF returns the value 1 for "attached" (0: "do not know"; −1: "not attached"). This is a simple example; other rules can be more complex and take into account dialogue act types and different speakers. A total of 17 rules covering 9 relation types have been created[5]. Once the LFs are applied to all the candidates (`EDUs`), step (2) utilized a generative model to calculate probabilities of possible attachments. For this, they built a matrix M_{ij} of size $m \times n$ where m is the number of `EDUs` and n the number of LFs. Each `EDU` receives an attachment score $\phi(\cdot)$:

$$\phi_j(M_i, y_j) := M_{ij} y_j \tag{3.30}$$

$$p_\theta(M, Y) \propto exp(\sum_{i=1}^{m} \sum_{j=1}^{n} \theta_j \phi_j(M_i, j_i)) \tag{3.31}$$

where y_j is the gold label; $\phi(\cdot)$ is score of candidate in matrix M; θ_j are parameters to optimize; p_θ is the probability with parameter θ. The objective, as shown in Equation 3.32, is to minimize the negative log likelihood:

$$\arg\min_\theta - \log \sum_Y p_\theta(M, Y) \tag{3.32}$$

The generative model's performance on STAC testing is comparable to that of the local model in Perret et al. (2016) (F_1 51 vs. 48), but much lower than Perret's Integer Linear Programming

[5]For a comprehensive description of all rules, refer to `https://tizirinagh.github.io/acl2019/`.

approach (68.9). As noted in this study, a major disadvantage of weak supervision is the low precision score, likely due to the imprecise or inaccurate supervision signals provided by the LFs Zhou (2018). This study is valuable for exploring weak supervision in dialogues, but its approach requires domain-specific annotation, carefully designed rules (such as deciding which rule to apply first or whether two rules capture similar information), and a relatively large validation set for rule verification.

Another weakly supervised approach is to increase the size of datasets through self-training. In a study about causal discourse detection in the biomedical domain, Mihăilă and Ananiadou (2014) used an iterative self-training strategy to overcome the problem of a large amount of unannotated data. They trained a classifier μ, or the *teacher*, with a small amount of labeled data and then tested on unannotated data, or the *student*. Only instances with high confidence scores above a pre-set threshold τ were considered gold and added to the labeled data. This process was repeated until all instances were annotated, thereby augmenting the BioCause corpus (Mihăilă et al., 2013). The authors found that, despite the potential noise in the augmented data, more discourse spans were correctly recognized as the dataset size increased (an increase of 4.35 points in F score). In our experiments for discourse relation prediction, we also adopted self-training strategies. However, we found that adding pseudo-labeled examples with high confidence scores did not consistently improve model performance. Instead, we discovered the importance of balancing the label classes for the added examples, as elaborated in Section 8.5.2.

In recent work by Nishida and Matsumoto (2022), self-training was also employed to generate complete discourse structures, including link attachment and relation types. However, unlike Mihăilă and Ananiadou (2014), their goal was to produce annotations for data in a different domain, using unsupervised domain adaptation (UDA). The authors adapted a model trained on a source domain with limited labeled data to a target domain where only unlabeled data was available. The bootstrapping strategy was applied, where one or more *teacher* models generated pseudo-labels for *student* models, and the *students* learned from these pseudo-supervisions, as illustrated in Figure 3.5. In different bootstrapping methods, *teacher* and *student* may refer to the same or different models. The authors compared four methods: (1) Self-Training (ST): *teacher* and *student* are the same model; (2) Co-Training (CT): two models play different roles and switch; (3) Tri-Training (TT): two *teachers* and one *student* are involved, where the latter learns from both and uses an *agreed* ratio to decide whether to include the *teacher's* prediction; (4) Asymmetric Tri-Training (AT): a domain-specific model is only used for inference, and the other two are only for pseudo-label generation.

For the crucial step of pseudo example selection, the authors employed two selection criteria: "rank-above-k" and "rank-diff-k". "Rank-above-k" strategy selects only the top $N \times k$ samples with the highest confidence scores. On the other hand, "rank-diff-k" strategy keeps only those samples whose relative ranking on the *teacher* side is k higher than that on the *student* side. The proposed pipeline is evaluated on both monologues and dialogues, where a domain transfer is performed from scientific papers to biomedicine documents for monologues, and from gaming conversations to technical chat for dialogues. Among the tested models, the co-training approach using shift-reduce model (Nivre, 2004) trained with arc-factored model (McDonald et al., 2005) achieved the best UAS score of 78.8 for monologues. For dialogues, a backward shift-reduce model using co-training achieved the highest UAS and LAS scores of 67.7 and 39.2, respectively. In comparison, the SOTA supervised model presented in Section 3.2.1 obtained UAS and LAS scores of 75.3 and 59.6 for link and relation prediction. Additionally, self-training and tri-training also demonstrated promising results under specific selection criteria, such as "rank-above-0.6" or "rank-diff-100".

This section describes three studies that employ different weak learning strategies: heuristic

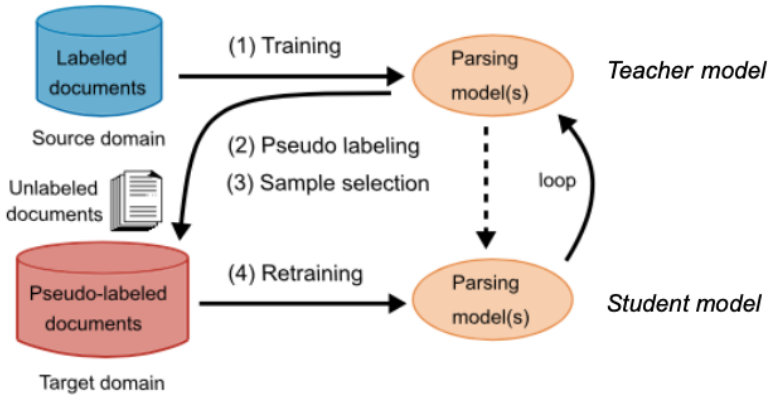


Figure 3.5: Overview of a bootstrapping system for unsupervised domain adaptation in discourse parsing, from Nishida and Matsumoto (2022).

Model	Setting	Framework	Output	Criteria/info	Corpus
Kobayashi et al. (2019)	monologue	RST	structure	(dis)similarity score	RST-DT, PCC 2.0‖
Nishida and Nakayama (2020)	monologue	RST	structure	initialization, Viterbi EM	RST-DT
Chapter 7 Li et al. (2023)	dialogue	SDRT	structure	PLM	STAC

Table 3.6: Unsupervised parsers. for RST: {structure, nuclearity, relation}; for SDRT: {structure, relation}. Criteria/info column: criteria applied to or information relied on for unsupervised parsing. PCC 2.0‖: Potsdam Commentary Corpus (Stede and Neumann, 2014).

rules (Badene et al., 2019a), in-domain self-training (Mihăilă and Ananiadou, 2014), and out-of-domain bootstrapping (Nishida and Matsumoto, 2022). Each of these strategies has its own advantages and is suitable for different scenarios depending on the availability of resources. These approaches are particularly useful for discourse parsing because they demonstrate that even with limited data, it is possible to generate additional (and potentially noisy) training data and improve the performance of our models.

3.2.4 Unsupervised Methods

When there is no annotated data for the main task or similar tasks, we encounter an extreme case of data scarcity. The field of unsupervised discourse parsing has mostly been neglected in the past, likely because of its inferior performance. In this section, we present the results of unsupervised parsers in Table 3.6, mostly applied in the monologue setting.

The use of fully unsupervised methods for RST discourse tree extraction was first explored by Kobayashi et al. (2019). They employed dynamic programming to create discourse trees based on similarity and dissimilarity scores. Furthermore, they investigated three granularities, namely `EDU`-level, sentence-level, and paragraph-level. Figure 3.6 illustrates right branching at different levels. Nishida and Nakayama (2020) conducted a similar study and also used these granularities.

The calculation of similarity in Kobayashi et al. (2019) is simple. They defined similarity between two adjacent spans using pre-defined word embeddings (ELMo (Peters et al., 2018a) and Glove (Pennington et al., 2014)) and an effective sentence vector calculation called *smooth inverse frequency* (SIF), which was originally proposed in Arora et al. (2017). The core calculation for

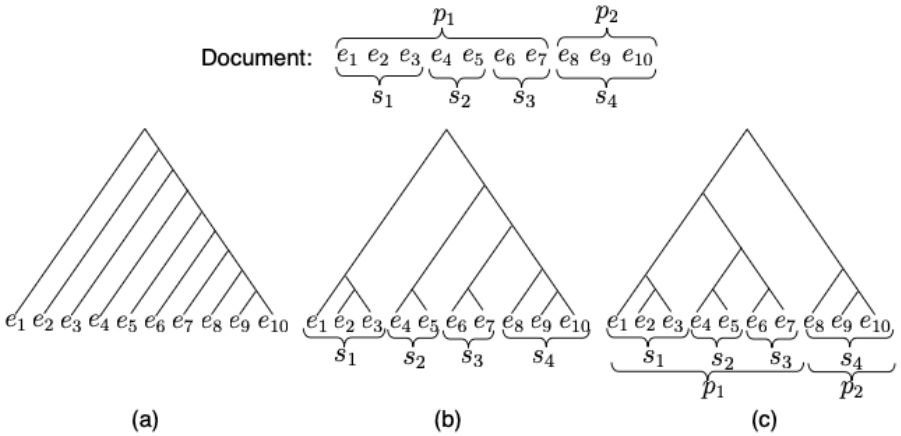


Figure 3.6: Illustration of different levels of granularity during right-branching for an RST-tree, from Kobayashi et al. (2019): (a) doc-2-EDU (*D2E*); (b) doc-2-sent-2-EDU (*D2S2E*); (3) doc-2-parag-2-sent-2-EDU (*D2P2S2E*). This document contains 2 paragraphs, 4 sentences, and 10 EDUs.

the similarity score is shown below :

$$\overrightarrow{u_t} = \sum_{w \in W_t} \frac{a}{p(w) + a} \overrightarrow{w} \tag{3.33}$$

$$\overrightarrow{l_{i:k}} = [\overrightarrow{u_i}; \overrightarrow{u_k}] \tag{3.34}$$

$$sim(\overrightarrow{l_{i:k}}, \overrightarrow{r_{k+1:j}}) = \frac{1}{2}\{\frac{\overrightarrow{l_{i:k}} \cdot \overrightarrow{r_{k+1:j}}}{\|\overrightarrow{l_{i:k}}\| \|\overrightarrow{r_{k+1:j}}\|} + 1\} \tag{3.35}$$

The SIF calculation is applied to each atomic unit $\overrightarrow{u_t}$, which is comprised of words w and their corresponding concatenated word embeddings in ELMo and Glove, denoted as $\overrightarrow{w}$. The span vector $\overrightarrow{l_{i:k}}$ concatenates the leftmost and rightmost atomic unit vectors $\overrightarrow{u_i}$ and $\overrightarrow{u_k}$ to obtain the similarity score of two adjacent spans, $\overrightarrow{l_{i:k}}$ and $\overrightarrow{l_{k+1:j}}$. By using $1 - sim(\cdot)$ as the dissimilarity score, they obtain the split score for the optimal tree. To perform the tree merge or split process, a dynamic CKY programming algorithm is used, which uses a matrix to store scores for all possible sub-spans of a tree at its granularity level and builds complete trees incrementally from the EDU-tree to sentence-tree and paragraph-tree. The algorithm proceeds from coarse to fine levels to construct trees accordingly. The granularity levels for the algorithm were defined as EDU-level, sentence-level, and paragraph-level, as used in other studies (Kobayashi et al., 2019; Nishida and Nakayama, 2020).

The experiments conducted on RST-DT (Carlson et al., 2002a) and Potsdam Commentary Corpus (PCC 2.0) (Stede and Neumann, 2014) yielded highly encouraging results, achieving a maximum micro-F_1 score of 81.1 and 78.4, respectively, on the entire dataset. It should be noted that the then state-of-the-art transition-based supervised parsers achieved 85.6 (Wang et al., 2017b) and 80.2 (Braud et al., 2017) on these two corpora. A comparison of the scores obtained by using different granularities shows that the finest *D2P2S2E* setting yielded a large improvement of about 15% over the coarsest *D2E* setting. This highlights the significance of structural information as parsing a document hierarchically best conforms to its initial skeleton, an observation that has been made in Joty et al. (2013); Feng and Hirst (2014a) and later employed in Nishida and Nakayama (2020); Xiao et al. (2021). However, it is worth noting that although

their approach has shown promising results for RST-style texts, it cannot be directly applied to discourse graph-style for dialogues.

Shortly after, Nishida and Nakayama (2020) proposed an unsupervised RST-style parsing method based on the hypothesis that discourse tree and syntactic tree structures share similar constituent properties, making unsupervised learning algorithms transferable. In contrast to Kobayashi et al. (2019), this study used more intricate features for `EDU` feature extraction and scoring, which involved syntactic cues like the *head word* present in each `EDU`. They employed the Viterbi EM algorithm (Spitkovsky et al., 2010) to train a discourse constituency parser in an unsupervised manner[6]. The approach involved automatically sampling initial discourse trees using prior knowledge (document hierarchy, discourse right-branching tendency, syntax-aware branching tendency, and locality bias) of document structures, followed by alternating E step and M step until reaching the early stopping criteria. The goal of the E step was to perform discourse parsing on the entire dataset and generate pseudo discourse trees:

$$\mathcal{D} = \{(x, \hat{T}) | x \in \mathcal{X}, \hat{T} = \underset{T \in valid(x)}{\operatorname{argmax}}\ s(x, T)\} \tag{3.36}$$

where $\mathcal{D}$ is pseudo treebank with all generated discourse trees; x is one document; $\hat{T}$ is the highest-scoring tree for document x; $valid(x)$ contains all valid trees for x; $s(x, T)$ is a score of the tree T. The scoring function $s(\cdot)$ is the sum of constituent scores over all internal nodes.

In the M step, the model parameters are updated to meet specific constraints. In this scenario, the objective is to optimize the score of the best-parsed tree, ensuring that it outperforms all other potential trees by a large margin (Δ):

$$s(x, \hat{T}) \geq s(x, T') + \Delta(\hat{T}, T') \tag{3.37}$$

where $\hat{T}$ is the best tree; T' is another parse tree in all candidates; $\Delta(\hat{T}, T')$ is the difference between two trees. During the E-M iterations, an early stopping criterion was defined using 30 annotated documents in the validation set. The results on the RST-DT corpus were superior to previous work: the `RST-Parseval` score was 84.3 for structure prediction, compared to 80.0 in Kobayashi et al. (2019), and on par or even better than some supervised models, such as Feng and Hirst (2014a) at 84.4 and Joty et al. (2015) at 82.5.

The success of this approach is largely due to the effective tree initialization step, which increased the model performance by 10 points compared to uniform initialization. However, an analysis of relation classes reveals that initialization rules can also impede the creation of certain relation types, such as *Evaluation* and *Summary*. In general, this work has many similarities with Kobayashi et al. (2019): both studies adopt a bottom-up strategy and use the CKY algorithm for decoding. They follow the hierarchical structure of documents from `EDU`- to sentence- and paragraph-level during tree generation. While both studies show good results on the RST-DT dataset, it is unclear if similar results could be achieved in other domains. While these methods are inspiring, they can only be applied to constituent-style trees, which makes them unsuitable for our intended use in SDRT-style parsing.

This section presents two unsupervised parsing methods for monologues. One method merges (or splits) spans based on similarity (or dissimilarity) scores (Kobayashi et al., 2019), while

[6]Technically, this study is not entirely unsupervised. A few annotated documents are needed to go guide early stopping during training.

the other applies unsupervised syntactic parsing methods, specifically Viterbi EM, to discourse parsing by leveraging transferable properties (Nishida and Nakayama, 2020). However, there is currently no unsupervised study on dialogue parsing. To address this gap, we propose strategies for tackling this issue and report our results in Chapter 7.

3.3 Discourse in Downstream Applications

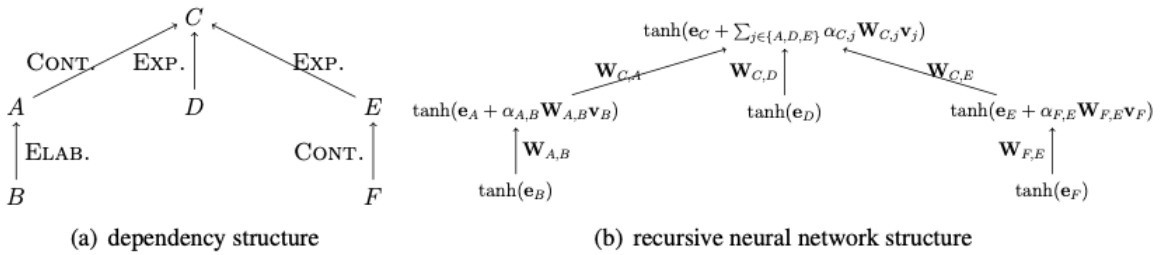
Discourse parsing is a fundamental task in NLP that has been widely applied in Natural Language Understanding (NLU) applications. Examples of these applications include general text categorization (Ji and Smith, 2017), author attribution prediction (Feng and Hirst, 2014a; Ferracane et al., 2017), fake news detection (Karimi and Tang, 2019), political leaning prediction (Devatine et al., 2022), and sentiment analysis (Bhatia et al., 2015; Hogenboom et al., 2015; Huber and Carenini, 2020a). These applications will be discussed in Section 3.3.1. Discourse structure has also been found to be useful in some Natural Language Generation (NLG) tasks, such as summarization (Marcu, 2000; Louis et al., 2010; Yoshida et al., 2014; Li et al., 2016; Liu et al., 2019b) and machine translation (Haenelt, 1992; Mitkov, 1993). Furthermore, with the increasing popularity of online chatting, dialogue machine reading comprehension in the form of question answering has become a hot research topic where discourse also plays a beneficial role (Li et al., 2021b; He et al., 2021). These topics will be covered in Section 3.3.2. Finally, in Section 3.3.3, we will discuss the similarities, usefulness, and limitations of these studies.

3.3.1 Discourse for NLU Tasks

3.3.1.1 Text Categorization

Text classification is a fundamental task in NLU that involves organizing texts into groups, such as sentiment analysis, spam detection, and topic labeling. Earlier methods for this task involved encoding sequences of sentences with sparse embeddings, such as hand-crafted features or lexical clues like n-grams, and passing them through a classifier (Minaee et al., 2021). However, these methods assumed that all parts of a text equally influence categorization. To address this limitation, researchers have sought to weigh different text spans, for example, by using hierarchical structures or attention mechanisms for word- and sentence-level representation (Ko et al., 2004). However, these methods still did not include inter-sentential interaction in sentence encoding.

One approach to incorporating discourse structure in text categorization was proposed by Ji and Smith (2017). They investigated five text categorization tasks, including (1) sentiment analysis on Yelp reviews (Zhang et al., 2015); (2) news article classification on Media Frame Corpus (Card et al., 2015); (3) congressional speaker voting on debate corpus (Thomas et al., 2006); (4) review classification on movie corpus (Pang and Lee, 2004); and (5) legislative bill survival voting on a congressional bill corpus (Yano et al., 2012). They hypothesized that tree-shaped structural information could provide better cues on the importance of different text spans. Using an RST-style discourse parser DPLP (Ji and Eisenstein, 2014), they segmented texts into `EDUs` and constructed an "unlabeled model" (not considering the relations) and a "full model" (with RST relations) where texts composed of `EDUs` are aggregated into trees. Text spans are then passed into a recurrent neural network for classification, as shown in Figure 3.7. Despite being trained on news articles, the DPLP parser has shown to be effective in tasks involving restaurant and movie reviews (on Yelp and Movie corpora, respectively). However, the bill voting prediction task did not benefit from a discourse-aware model, which may be due to its technical legal terms



(a) dependency structure (b) recursive neural network structure

Figure 3.7: Tree structure from RST parser and the process of tree aggregation with RNN in Ji and Smith (2017). RST constituent tree is converted into a dependency tree. A, B, C, D, E, and F are `EDUs`. The original text is: [Although the food was amazing]$_A$ [and I was in love with the spicy pork burrito,]$_B$ [the service was really awful.]$_C$ [We watched our waiter serve himself many drinks.]$_D$ [He kept running into the bathroom]$_E$ [instead of grabbing our bill.]$_F$

and highly specialized conventions that make it the most distant from the news genre. Regarding the "unlabeled model" and "full model", the former yielded better performance than the latter in four tasks, except for Yelp food reviews. The authors did not provide any explanations for this observation, and we speculate that it may be due to inaccurate relations produced by the parser. Degradation studies also showed that parsing performance and text classification results were positively correlated, suggesting that further improvements in discourse parsing could lead to greater gains.

3.3.1.2 Author Attribution:

The task of Author Attribution (AA) is to identify the author of a text, which can be done through binary or multi-class classification.

In Ferracane et al. (2017), the authors explored the AA task by incorporating discourse information, building on the previous work of Feng (2015), which found that above-sentence level discourse information can aid in identifying stylometric cues. To achieve this, they utilized the Entity Grid Model proposed by Feng and Hirst (2014a), where sentences and key entities (noun phrases) form the rows and columns, respectively, and Rhetorical Structure Theory (RST) discourse information is stored in the cells. The RST features represent discourse relations and nuclearity, such as `[definition.N, attribution.S]`, which were obtained using an off-the-shelf RST parser, DPLP (Ji and Eisenstein, 2014). The resulting feature vectors were fed into a Convolutional Neural Network (CNN) for author prediction.

Experimental results on three AA datasets demonstrated the effectiveness of incorporating RST relations into the entity grid, particularly when entities are tracked across the entire document (i.e., in a *global* setting) with a macro-F_1 multi-class classification of 98.8%, compared to 95.3% without discourse. Prior to Ferracane et al. (2017), AA task had been studied by Feng and Hirst (2014b); Feng (2015). However, Feng's work had limitations, such as the local encoding of discourse relations (only adjacent sentences) which brought less remarkable results than those of Ferracane's.

3.3.1.3 Fake News Detection

The task of detecting fake news is becoming increasingly popular in text classification. In Karimi and Tang (2019)'s study, latent discourse-level dependency tree structures were learned and

constructed for fake and real news articles. As there was no available discourse corpus for the fake news domain, the authors built the latent discourse trees in an automated and data-driven manner, inspired by Liu and Lapata (2018). They used a hierarchical bi-LSTM network to obtain sentence representations and constructed a matrix A that stores the parent-child link probabilities. This work did not conduct EDU segmentation. A document-level representation is the average value of each sentence's structure-aware vectors (g_j), which is composed of a parent node (p_j), a child node (c_j), and a fixed bi-LSTM representation (f_j):

$$p_j = r_j \times e_{root} + \sum_{z=1}^{k} A[z, j] \times f_j \tag{3.38}$$

$$c_j = \sum_{z=1}^{k} A[j, z] \times f_j \tag{3.39}$$

$$g_j = \mathcal{G}(W[p_j \| c_j \| f_j] + b) \tag{3.40}$$

where r_j is probability of a sentence j being the root node; e_{root} denotes a special root embedding vector; $\|$ means concatenation operation.

Similar to Ji and Smith (2017), this work assumed that document-level structural-rich representation is beneficial for text classification. Upon examination, the authors found significant property divergence between real and fake news articles, including (1) the number of leaf nodes, (2) the positional difference between the *preorder traversal* of the discourse tree (subtrees are ordered based on when they are added as the child nodes of a parent node) and the original sentential order, and (3) the distance between parent-child nodes. These differences indicated less coherence in fake news texts, according to the conclusions. Later, a study by Ferracane et al. (2019) concluded that Liu and Lapata (2018)'s approach is easily biased by lexical cues so that the extracted latent discourse trees might capture something else than the structure.

3.3.1.4 Political Orientation Prediction

Recently, discourse information has been explored by Devatine et al. (2022, 2023) to predict political orientation, a task that aims to determine the political leaning of an article among three classes: *left*, *center*, and *right*. Following the work of Karimi and Tang (2019), they used the latent discourse structure extraction method introduced in Liu and Lapata (2018) but with some adaptations suggested in Ferracane et al. (2019). These adaptations relate to (1) the pooling operation, (2) the removal of document-level bi-LSTM, and (3) the percolation of descendant trees for the final document representation. Additionally, they used ToNy segmenter (Muller et al., 2019) to perform EDU segmentation as the first step, which is omitted in Karimi and Tang (2019). The authors evaluated their model on the Allsides dataset and compared its performance with that of the model proposed in Baly et al. (2020). Their results demonstrated that the structured attention model outperformed the others by a large margin, achieving an increase of 7 points in accuracy and 6 points in macro-F_1. The increase came from fine-grained discourse (EDU level instead of sentence level) and the consideration of larger context (no token length limitation).

According to their analysis, attention was directed towards distinct lexical fields depending on the political leaning: *health* for *left*, *statistics* for *center*, and *economy* for *right*. Regarding structural analysis, they provided a qualitative assessment indicating that the structures learned were complex and not merely simplistic flat trees. In contrast to Karimi and Tang (2019), this

study did not provide a discussion on variations in structure among the different political classes, which could be due to the absence of such observations. To enhance the current study, it may be useful to conduct further investigation and provide possible explanations for the absence of structural differences.

3.3.1.5 Sentiment Analysis

Sentiment analysis is a popular downstream task where the goal is to determine the overall polarity of a document, usually categorized as binary (positive or negative) or multi-class (such as positive, negative, and neutral).

(Hogenboom et al., 2015) proposed a weighting scheme that incorporates nuclearity and relation information from RST into sentiment analysis. They proposed two ways to calculate the weight: (1) a heavier weight (1 or 1.5) for the nucleus and a lighter weight (0 or 0.5) for the satellite, and (2) a weight that considers both nuclearity and relation. The authors evaluated their approach on a movie review dataset and obtained accuracies between 65% (baseline) and 72% (best model). Their proposed weighting system outperformed a lexicon-based analyzer (Wilson et al., 2005b) by 4 points. However, it was later demonstrated by Bhatia et al. (2015) that a basic classification model based on discourse depth can achieve much better performance on the same dataset. This may be because the weighting scheme doesn't capture inter-sentence information, which could be key to predicting sentiment. Nonetheless, an interesting finding in this work is that finer-grained discourse structures, such as sentence- or `EDU`-level RST trees, are better suited for sentiment analysis than paragraph-level or document-level trees.

Bhatia et al. (2015) conducted a study to improve sentiment prediction by incorporating discourse information at the document level. They revisited the weighting system and proposed a new approach that uses RST parses within a recurrent neural network (RNN). They utilized the DPLP discourse parser (Ji and Eisenstein, 2014) to extract subtrees and gradually compose the constituent parts, similar to the approach taken in Ji and Smith (2017) for text categorization. This study included experiments on two movie review datasets with binary sentiments (Pang and Lee, 2004; Socher et al., 2013a), and they compared their approach with a lexicon-based analyzer and a logistic regression classifier. The rhetorical RNN system significantly outperformed the baselines by $5 - 10$ points. The authors also explored the usefulness of discourse relations and compared the system with and without relations, like the "full model" and "unlabeled model" in Ji and Smith (2017). However, the improvement brought by the relation-enriched RST tree was minor, likely due to the average relation prediction capacity of the DPLP model, which was reported to be only 60% accurate.

Similar to Bhatia et al. (2015)'s RNN model, Tai et al. (2015) developed a tree-LSTM architecture that was later improved with a discourse-LSTM by Kraus and Feuerriegel (2019). The discourse-LSTM merged the information from the tree leaves and propagated it to the higher levels until it reached the root node where a final prediction was made.

In a different approach to using discourse for sentiment prediction, Huber and Carenini (2020a) employed a silver-standard sentiment-leveraged discourse treebank MEGA-DT[7] instead of a human-annotated gold discourse corpus like RST-DT. They hypothesized that the MEGA-DT treebank – obtained using distant supervision but in the same domain as target task – would be more useful than the gold-standard discourse treebank in a inter-domain scenarios. Their approach involved augmenting sentiment annotations with discourse information to improve sentiment predictions. The results showed that their approach was particularly effective for

[7]MEGA-DT is introduced in Huber and Carenini (2020b), as presented in Section 3.2.2.1.

longer documents, but the best-performing model's accuracy was still quite low (approximately 66% for binary classification). While it is interesting to observe how information travels from one task to another, some concerns remain regarding the potential introduction of noise in the long pipeline and the proper evaluation of such noise, as well as the applicability of this approach in other domains.

3.3.2 Discourse for NLG Tasks

3.3.2.1 Machine Translation

Machine Translation (MT) task involves translating a text from one language to another, requiring both text understanding in the source language and text generation in the target language. In this context, discourse information can be useful at both stages of the process.

The use of discourse structure in MT has been discussed for over thirty years. For example, Haenelt (1992) proposed the KONTEXT model, which defined discourse as sequences of transitions between multi-layer information such as sentence structure, referential structure, and thematic structure. Mitkov (1993) introduced the Text Organization Framework Grammar, which maps the source paragraph structures of rhetorical predicates into specific target paragraph structures of rhetorical predicates.

In 2000, Marcu (2000) designed an "analysis-transfer-translate" pipeline for Japanese-English translation, where a Japanese text is first encoded in an RST-style tree and then transferred into an English RST tree, which is used as the base for English sentence generation. Tu et al. (2013) integrated this module into Statistical Machine Translation (SMT) and tested it on Chinese-to-English translation. For RST-tree acquisition, they used hand-crafted features and a Bayesian model to jointly perform EDU segmentation and relation prediction. They annotated around over 1000 complicated sentences in the Chinese Penn Treebank (CTB) (Xue et al., 2005) based on relation types defined in Yue (2008) and trained their parser. The second step was translation rule extraction, where source RST trees were aligned with target language strings, and the final step was decoding the source RST trees into the target language using the extraction rules. While this pipeline has shown to be effective, it does require a significant amount of human-annotated training data for RST-style parser training.

Another study that highlights the potential of integrating discourse structure into MT is presented by Joty et al. (2017). In this study, the authors did not propose a new pipeline for discourse integration but instead designed similarity measures that compare the discourse parse trees of a generated translation and a gold translation. These measurements can provide additional information on the performance of an MT system. Essentially, the more similar the generated RST tree is to the gold RST tree, the better the system is. Furthermore, the authors analyzed the relevance of different elements in RST trees (i.e., attachment, nuclearity, and relation) and demonstrated that all aspects are useful, with nuclearity information being particularly important. This study confirms the usefulness of discourse parsing for MT evaluation.

To gain a better understanding of how the use of discourse devices impacts translation quality, Li et al. (2014b) conducted manual evaluations of translations from Chinese and Arabic to English. They found a strong mismatch in the notion of what constitutes a sentence in Chinese and English, the usage of discourse connectives, and the ambiguity of the connectives. Interestingly, these differences are less present in Arabic-English translations. It appears that discourse usage may affect MT between some language pairs but not others. Other discourse properties such as topic mix, style, coherence patterns (including explicit and implicit rhetorical relations), and the use of anaphora and coreference are essential for producing a more coherent translation.

In this regard, interested readers can refer to an ACL workshop DiscoMT (Webber et al., 2013). This workshop was established in 2011 and seeks to encourage new approaches that incorporate discourse-level features to enhance machine translation. Four successful workshops have been organized to date, with more than 50 accepted papers covering various discourse phenomena such as lexical consistency, lexical cohesion, and implicit relations.

3.3.2.2 Machine Reading Comprehension

Machine Reading Comprehension (MC or MRC) is a task that involves automatically extracting answers from questions based on a given text, in the form of question answering. This technology is highly beneficial in identifying crucial information from various types of text such as Wikipedia pages, stories, essays, and forum discussions like those found on Reddit or Ubuntu.

Prior works that investigated discourse information in QA have mostly relied on hand-annotation of discourse relations, as in Chai and Jin (2004); Verberne et al. (2007b); Jansen et al. (2014). Verberne et al. (2007a,b) focused on answering *why-* questions and found that answers often consisted of propositions spanning multiple sentences and linked by discourse relations such as *cause* and *explanation*. Their proposed method involved extracting text spans with the same proposition as the question topic and then extracting the siblings of those text spans as candidate answers. These answers were then re-ranked using a probability model based on a general language model (Croft and Lafferty, 2003), resulting in a reported success rate of 60%. Jansen et al. (2014) studied non-factoid answer reranking for open-ended questions related to *manner* (*how*-questions) and *reason* (*how*-questions), using both shallow discourse markers (from Hirst and Marcu (1998)'s list) and a discourse parser (Feng and Hirst, 2012) to incorporate discourse information. They found that both shallow and deep discourse representations are useful, and that combining these two strategies led to the best performance. However, both studies relied on supervised discourse parsers. When testing on target domains, the parser may fail to generate trees on different domain data (for instance news→biology cross-domain (Jansen et al., 2014) reported $> 40\%$ failed cases).

In contrast to the presented approaches where a trained parser is used to provide discourse tree structure, Narasimhan and Barzilay (2015) proposed a method for discourse relation induction. They argued that distantly supervised methods can introduce errors due to the mismatch between training and testing data. Instead, they proposed using a probability model to optimize a task-specific objective, thus eliminating the need for explicit annotation. Their approach involved designing a discriminative model that captures relationships between sentences, where a hidden variable $r \in R$ represents the type of relationship (the set R contains *causal*, *temporal*, *explanation*, *other*):

$$P(a, r, z_1, z_2|q) = P(z_1|q) \cdot P(r|q) \cdot P(z_2|z_1, r, q) \cdot P(a|z_1, z_2, r, q) \tag{3.41}$$

where a, q, r, z_1, and z_2 are answer, question, relation, and candidate sentences (z_1, z_2), respectively. They marginalized all the hidden variables and chose the answer that maximizes $P(a|q)$. Their proposed model is particularly advantageous when answering questions that require multiple sentences, as the discourse relation type can help to moderate the relationships between sentences. The component $P(r|q)$ conditions the relation type based on the question, such as when answering a *why*-question that often requires a *causal* relation. An interesting comparison was made with a model that used relations from RST trees produced by a parser (Feng and Hirst, 2012), and the results showed that the RST-based model performed worse. This was because the RST-trained parser over-predicted *elaboration* and failed to provide distinctive inter-sentential relations.

This work is one of the first to investigate unsupervised discourse information injection without relying on discourse parsers. However, the study's consideration of discourse information is limited to the relation level with only four types. As the analysis showed, the model's accuracy was low, with correct predictions for only 50% of *causal* and *other* relations, while *explanation* and *temporal* relations were below 30%. Additionally, model accuracy varied significantly based on question types, with *where-* and *when-* questions having higher accuracy than *why-* and *which-*questions.

In contrast to traditional MRC, multi-party MRC involves a more complex dialogue structure that typically involves two or more people. This makes the task even more challenging. Li et al. (2021b) are the first to propose a discourse-aware graph neural network (GNN) for multi-party MRC. Their approach consists of two primary modules. The first one is the *Discourse Graph module*, which is a GNN responsible for updating the representations of utterances by leveraging the information from their dependency links. The second module is the *MRC module*, which takes the updated utterances representations and combines them with word representations using attention mechanisms, thereby introducing the dialogue discourse graph structure to all the words. The proposed approach was evaluated on the Molweni corpus (Li et al., 2020), which contains both SDRT-style discourse annotation and question-answer pairs. Analysis showed that both discourse structure and relations are helpful in predicting answers – even though meager, increasing the F_1 score by one point. The study demonstrated that the discourse-aware GNN model outperformed the state-of-the-art models such as DialogueGCN (Ghosal et al., 2020) and DialogueRNN (Majumder et al., 2019). Nonetheless, certain concerns must be addressed. Firstly, Molweni, the dataset used for the study, has quality issues such as a high repetition rate and inaccurate annotation (Section 2.3.5), which raises questions about the reliability of the results. Secondly, the generalizability of the study to other domains is unclear since the annotation for discourse parsing and MRC came from the same corpus. It would be beneficial to test the model with a supervised parser trained on another corpus or a parser trained on Molweni but tested on a different MRC corpus.

3.3.2.3 Summarization

The task of summarization involves condensing key information from a lengthy document. Two methods of summarization exist: extractive, which selects the most pertinent sentences from the original text, and abstractive, which creates a summary using new words and sentences. Other types of summarization include extreme summarization – a one-sentence summary of scientific documents –, and lay summarization – a brief summary in layman's terms with less technical jargon that captures the essence of the research paper. Discourse analysis plays a significant role in identifying the most informative sentences in the original text. Discourse trees provide a suitable representation for summaries and can aid in the discovery of informative sentences, with roots and high-level nodes being the most important parts. This idea was first proposed by Marcu et al. (1999) and supported by Carlson et al. (2001); Prasad et al. (2008b). It was further developed by Hirao et al. (2013); Yoshida et al. (2014) who adapted the trees from constituent to dependency form, and by Liu et al. (2019b) who applied it to the entire document representation.

Louis et al. (2010) explored the potential usefulness of discourse information in single document extractive summarization. They investigated two types of discourse information: structure and semantic sense (i.e., relation). To evaluate structural features, they employed a scoring system to determine the relative importance of text spans. The system included a nucleus-satellite penalty (Ono et al., 1994), depth-based and promotion-based scores (Marcu, 1998). Semantic features were evaluated using PDTB (Prasad et al., 2008a) relations. The study found that dis-

course features based on structure were strong indicators of sentence importance, while semantic relations were useful in determining what content should not be included, but did not reliably indicate importance. The best results were obtained by combining both types of features.

Liu et al. (2019b) also tackled the extractive summarization for single document. Their approach involved formulating the task as a multi-root tree induction problem where summary-worthy sentences are the roots and satellite sentences provide additional details as nodes attached to them. The process involved conducting binary classification for each sentence to decide whether they are roots or edges, followed by using a structured attention model to calculate the loss for root and edge prediction iteratively to refine the induced tree. The resulting tree is a latent discourse tree, later work such as Devatine et al. (2022) employed a similar approach for political orientation prediction.

In the domain of abstractive summarization, Gerani et al. (2014) proposed a system that combines multiple product reviews into an aspect-based summary using discourse structure and relation. The authors recognized that while global sentiment summarization is common, very few studies predict fine-grained aspect sentiments. They hypothesized that discourse structure explicitly reveals inter-sentential relations, which could be of help in aspect sentiment detection. The system they proposed consisted of four steps: (1) extraction and pruning of discourse trees, (2) transformation of discourse trees into an Aspect Rhetorical Relation Graph (ARRG), (3) selection of contents, and (4) summary generation. For discourse tree extraction, they used a pre-trained discourse parser (Joty et al., 2013)[8]. They then pruned the parsed trees and retained only aspect words, such as *photo* and *camera*, in the leaves. Using several Aspect-based Discourse Trees (ADTs), they extracted relation tuples and aggregated them into an ARRG. In step (3), they selected only the most important aspects by relying on measurement based on the hierarchical structure of discourse trees and the Weighted Page Rank algorithm (Xing and Ghorbani, 2004). The final step was to generate language based on the extracted sub-graphs (AHT).

When compared to extractive summaries, abstractive summaries were preferred by human raters, and they rated summaries that incorporated discourse-based features higher than those that did not. The feedback provided by raters, such as "very complete" and "related features", indicated that the aspect information was well-aligned with the sentiment thanks to the inclusion of discourse in both step (1) and step (3).

The aforementioned studies highlight the potential of using discourse to improve text summarization in different ways. For extractive summarization, discourse can be leveraged to identify important text spans based on nuclearity and hierarchical information (Louis et al., 2010; Liu et al., 2019b; Hirao et al., 2013). On the other hand, for abstractive summarization, discourse can help discover aspect-based knowledge by exploiting relations such as *elaboration* and the hierarchical structure of the text. This enables the targeting of specific relations between text spans and their relative importance. It would be valuable to conduct a correlation study between parser quality and summarization performance to further enhance this approach.

The field of dialogue summarization is gaining popularity as evidenced by recent studies (Koay et al., 2020; Zhu et al., 2020a; Feng et al., 2021b; Chen and Yang, 2021). In particular, Chen and Yang (2021) proposed a method to explicitly model discourse and action relations (constructed as the "WHO-DOING-WHAT" triplets) into the summarization process. They

[8]The paper did not specify whether the parser was trained with the RST-DT corpus (Carlson et al., 2002a) or Instructional-DT corpus (Subba and Di Eugenio, 2009), and it did not provide any precision.

employed a pre-trained *Deep Sequential* parser (Shi and Huang, 2019) trained on the STAC corpus to generate discourse parse trees for the SAMSum summarization corpus (Gliwa et al., 2019). These trees were then encoded into utterance representations using a Graph Attention Network (Veličković et al., 2018), and the resulting discourse-enhanced graph was injected into the BART model's (Lewis et al., 2020) cross-attention layers for decoding. The results on the SAMSum and ADSC (Misra et al., 2015) test sets demonstrated that incorporating structured information such as discourse and action relations led to improved performance. Interestingly, the performance of structured BART improved with longer conversations, but only up to a certain threshold (when the number of discourse edges was within the range of 8.3 to 9.1). Beyond that threshold (when the number of discourse edges exceeded 9.5), the structured model began to perform poorly. Longer conversations typically involve more long-distance relations that are challenging to predict for discourse parsers, as demonstrated in our analysis in Chapter 7. We speculate that the discourse relation extraction failed to capture these intricate dependencies, which limited its effectiveness for summarization.

Feng et al. (2021b) also proposed utilizing dependency relations to enhance the interaction between different speech turns. To generate discourse trees, they followed the same procedure as described in Chen and Yang (2021). After obtaining the original SDRT-style dependency trees, they performed a Levi graph transformation (Gross and Yellen, 2003), which treated SDRT relations as new vertices. Utterance vertices and relation vertices were connected with additional relation types like *default-in-discourse* and *default-out-discourse*, allowing for explicit modeling of discourse relations and the simultaneous updating of both the utterance and relation vertices. Speaker information was explicitly encoded by combining a one-hot vector representation of the speaker with the utterance vector. They employed Relational Convolutional Graph Networks (Schlichtkrull et al., 2018) and Pointer network (See et al., 2017) for the graph encoder and decoder parts, respectively. Compared to baseline Seq2Seq models, their discourse-enhanced approach outperformed in both the AMI (Carletta et al., 2006) and ICSI (Janin et al., 2003) corpora. Human evaluation on the *relevance* and *informativeness* of the summaries showed a preference for the discourse model. This study conducted some insightful analyses. They found that the higher the quality of the discourse parser, the better the summarization performance. Moreover, the importance of different relation types varied for different testing corpora. For instance, on the AMI corpus, *conditional* and *background* were important relations, while on the academic meeting ICSI corpus, *result* was more beneficial.

3.3.3 Discussion

We show a summary in Table 3.7 that covers five NLU tasks and three NLG tasks presented earlier. In this section, we will discuss various aspects of these discourse-aware models. This includes an examination of the discourse features employed, how they are integrated with other information, and the accuracy and usefulness of these models. Additionally, we provide suggestions for improving the incorporation of discourse in downstream applications.

3.3.3.1 Discourse Feature Consideration

In RST and SDRT frameworks, we consider structure (`EDU` or sentence attachments) and relation as major discourse features. In RST, *nuclearity* is also a crucial component (feature) that determines the relative importance of discourse units. In SDRT, although nuclearity is not explicitly provided, we can infer such information from subordinating (e.g., *Elaboration*) and coordinating (e.g., *Continuation*) relations. However, this information in SDRT has not been well explored.

Reference	Task	Domain/ Setting	Parser	Considered RST/SDRT feature			Other feature	Injection	Architecture
				Struc	Nuclearity	Relation			
NLU application									
(Ji and Smith, 2017)	text categorization	review, news, debate	Ji2014	✓	✓	✗/✓**	-	recursive propagation	RNN
(Ferracane et al., 2017)	author attribution	novel, movie	Ji2014	✗	✓	✓	-	grid-entity composition	CNN
(Karimi and Tang, 2019)	fake news detection	news	-	✓	✗	✗	-	parent-child prob matrix	structured attention†
(Devatine et al., 2022)	political bias	news	-	✓	✗	✗	-	parent-child prob matrix	adapted structured attention‖
(Hogenboom et al., 2015)	sentiment analysis	movie review	HILDA	✗	✓	✓	-	weighting scheme	lexi-based analyzer
(Bhatia et al., 2015)	sentiment analysis	movie review	-	✗	✗	✗	DU position	weighting scheme	lexi-based analyzer, LR classifier
			Ji2014	✓	✓	✓	-	recursive propagation	rhetorical RNN
(Fu et al., 2016)	sentiment analysis	movie review	Surdeanu2015	✓	✓	✓	-	nucleus info+LSTM	RST-LSTM
(Kraus and Feuerriegel, 2019)	sentiment analysis	movie/food review	-	✗	✗	✗	DU position	child-sum, N-ary	Tree-LSTM*
			Ji2014	✓	✓	✓	-	child-sum, N-ary	Discourse-LSTM
(Huber and Carenini, 2020a)	sentiment analysis	food review	Wang2017	✓	✓	✗	-	child-sum+attention	Tree-LSTM*
NLG application									
(Marcu, 2000)	MT	ja→en	-	✓	✓	✓	-	analysis→transfer→translate	-
(Tu et al., 2013)	MT	zh→en	Bayesian joint	✓	✓	✓	-	parse→transfer→translate	-
(Joty et al., 2017)	MT evaluation	multi-language	Joty2015	✓	✓	✓	-	-	-
(Verberne et al., 2007b)	QA	news	Feng2012	✗	✓	✓	-	RST sibling node	LM reranking
(Jansen et al., 2014)	QA	open domain+biology	-	✗	✗	✗	disc markers	marker-based span	discourse marker model (DMM)
			Feng2012	✗	✓	✓	-	relation-based span	discourse parser model (DPM)
(Narasimhan and Barzilay, 2015)	MRC	fiction story	-	✗	✗	✓	lexical, syntactic	-	joint probabilistic model
(Li et al., 2021b)	MRC	chat, multi-party	-	✓	-	✓	-	disc-aware node encoding	graph convolutional network
(Louis et al., 2010)	summarization	news, single doc	-	✓	✓	✗	PDTB-based, other	features for classification	logistic regression classifier
(Liu et al., 2019b)	summarization	news, single doc	-	✗	✗	✗	multi-head attention	-	Transformer
(Hirao et al., 2013)	summarization	news, single doc	HILDA	✓	✓	✓	-	-	tree knapsack model
(Gerani et al., 2014)	summarization	review, multi-doc	Joty2013	✓	✓	✓	-	pruning, aggregation	subgraph extraction
(Chen and Yang, 2021)	summarization	chat, multi-party	Shi2019	✓	✓	✓	-	disc-aware node encoding	graph attention network
(Feng et al., 2021b)	summarization	meeting, multi-party	Shi2019	✓	✓	✓	-	disc-aware node encoding	graph convolutional network

Table 3.7: Summary of discourse information applied on downstream tasks. Parser: discourse parser used to provide discourse structure: Ji2014 DPLP (Ji and Eisenstein, 2014), HILDA: (Hernault et al., 2010), Surdeanu2015 (Surdeanu et al., 2015), Wang2017 (Wang et al., 2017b), Joty2015 CODRA (Joty et al., 2015), Feng2012 (Feng and Hirst, 2012). Joty2013 (Joty et al., 2013), Shi2019 *Deep Sequential* (Shi and Huang, 2019). Considered RST/SDRT features: in RST: {structure, nuclearity, relation}; in SDRT: {structure, relation}. **: one UNLABELLED model without relation and one FULL model with relation in Ji and Smith (2017). Injection: the way discourse information is injected in the system. Architecture: system employed to perform downstream task. †: Liu and Lapata (2018) model. ‖: Liu and Lapata (2018) model adapted with Ferracane et al. (2019). *: Tai et al. (2015) model.

Structure and nuclearity are essential in NLU tasks such as sentiment analysis (Bhatia et al., 2015; Kraus and Feuerriegel, 2019) and text classification (Ji and Smith, 2017). These features are used during the aggregation process to combine sentence-level vectors following the tree structure and taking into account the nuclearity importance. For instance, Bhatia et al. (2015) demonstrated a significant improvement in performance compared to models that do not consider discourse features, even without differentiating between discourse relations. Although incorporating relations leads to further improvements, the gains are fairly modest, likely due to the lower accuracy of relation detection, even for the best systems.

Discourse features are also important in NLG tasks such as question-answering. Jansen et al. (2014) utilized discourse relations and markers to enhance non-factoid question-answering. When incorporating discourse into answer generation, nuclearity proved to be useful because the text span that requires elaboration, evidence, or explanation typically serves as the nucleus of the relation, while the text providing this information acts as the satellite, as demonstrated in Verberne et al. (2007a).

A noteworthy study by Louis et al. (2010) compares the advantages of two discourse features: the structure of the text and the semantic sense of discourse relations. In a single document summarization task, they discovered that structure information is a more robust indicator of importance compared to relations. While relations complement structure information, they alone did not prove to be a strong indicator. The study also compares two different forms of structure features: graph-based (as in Graph Bank) and tree-based (as in RST). The results indicate that both structures are equally valuable. In other works, such as extractive (Hirao et al., 2013) and abstractive summarization (Gerani et al., 2014), and multi-party conversation summarization (Chen and Yang, 2021; Feng et al., 2021b), complete discourse information using pre-trained parsers is used. Both structure and relation play crucial roles in these studies, where structure builds links within utterances and relation provides additional evidence and reasoning.

3.3.3.2 Discourse Information Incorporation

Our presentation of strategies to incorporate discourse information into downstream applications reveals an interesting trend that aligns with the evolution of NLP models: from classical statistical models to more advanced deep neural networks.

(1) Weighting Schema: In NLU tasks such as sentiment analysis (Hogenboom et al., 2015; Bhatia et al., 2015), combination methods for sentences are simple and straightforward: sentiment scores of `EDUs` are weighted based on the tree structure. These weights are normally pre-determined and hand-crafted (Kraus and Feuerriegel, 2019). Hogenboom et al. (2015) considered nuclearity labels i.e., nucleus or satellite) and relation types (e.g., nucleus weights by RST relation). Bhatia et al. (2015) incorporated discourse based on the `EDU` depth in dependency tree: they first converted the constituent tree into a dependency tree following Hirao et al. (2013) and then used a linear function to weight the importance of each unit with d_i the `EDU` depth:

$$\lambda_i = \max(0.5, 1 - d_i/6) \tag{3.42}$$

(2) RNN: A data-driven approach that uses Recursive Neural Networks (RNN) has been developed to combine discourse trees. The RNN recursively propagates the nodes' sentiment scores upwards until the root node is reached, allowing for the representation of the root node to be used for prediction. The hyperbolic tangent function ($tanh(\cdot)$) is commonly used as an aggregating function, as seen in tasks such as sentiment analysis (Bhatia et al., 2015) and text categorization (Ji and Smith, 2017). The representation of internal nodes varies slightly depending on the type of discourse tree used:

$$v_i = \tanh(e_i + \sum_{j \in \text{children(i)}} W_{r_{i,j}} v_j) \tag{3.43}$$

$$v_i = \tanh(W_n^{r_i} v_{n(i)} + W_s^{r_i} v_{s(i)}) \tag{3.44}$$

Equation 3.43 represents the internal node vector v_i in the dependency tree in Ji and Smith (2017). It is composed of its own vector and the sum of all its children's vectors v_j. $W_{r_{i,j}}$ is a relation-specific composition matrix. Equation 3.44 represents the intermediate node composed from relation r_i in constituent tree (Bhatia et al., 2015). The subscripts n and s represent *nucleus* and *satellite* resp.; W_n^r and W_s^s are relation matrices with regards to nucleus and satellite. In multi-nuclear cases, the second component will be simply changed $W_n v_n$.

(3) LSTM: Since RST-style discourse structures are trees, some researchers have employed LSTM for discourse information injection, such as RST-LSTM (Fu et al., 2016) and Discourse-LSTM (Kraus and Feuerriegel, 2019), both were employed for sentiment analysis task. The RST-LSTM model utilized nuclearity information from RST parse trees to explicitly model the importance of different text segments[9]. The Discourse-LSTM, proposed by Kraus and Feuerriegel (2019), extended RST-LSTM by incorporating the relation type between two nodes and replacing the weight matrices with tensor-based weights[10].

(4) GNN: For summarization tasks, dependency structure is commonly used, especially in the multi-party dialogue scenarios (Chen and Yang, 2021; Feng et al., 2021b). Discourse structure is represented as a dependency graph with utterances as nodes and edges as rhetorical relations. For each utterance, its representation is updated by its neighbour nodes and the relation in a graph convolutional network (Equation 3.45 in Feng et al. (2021b)) or a graph attention network (Equation 3.46 and 3.47 in Chen and Yang (2021)):

$$h_i^{(l+1)} = \sigma(\sum_{r \in \mathbb{R}} \sum_{v_j \in \mathbb{N}_i^r} \frac{1}{|\mathbb{N}_i^r|} W_r^{(l)} h_j^{(l)}) \tag{3.45}$$

$$a_{ij} = \frac{exp(\sigma(a^T [W_{v_i} || W_{v_j} || W_r e_{i,j}]))}{\sum_{k \in \mathbb{N}_i} exp(\sigma(a^T [W_{v_i}] || W_{v_k} || W_r e_{i,k}))} \tag{3.46}$$

$$h_i = \sigma(\sum_{j \in \mathbb{N}_i} a_{ij} W_{v_j}) \tag{3.47}$$

where σ is the activation function; $\mathbb{N}_i$ is the set containing node i's neighbours; $[\cdot||\cdot]$ is concatenation symbol; W and W_r are learnable node- and relation- specific parameters; $^{(l)}$ represents the h-th layer in convolutional network. Note that in Feng et al. (2021b), authors applied Levi transformation on relations, transforming them into nodes, which explains the sum over all relations r in Equation 3.45. Despite slight differences, the encoding processes are much alike.

We have presented various methods for discourse information incorporation. Since the granularity of tasks is different – some tasks more focus on document-level prediction, while others focus on local interaction (rhetorical relations between sentence pairs), the integration of discourse information varies. In the work by Verberne et al. (2007a) and Jansen et al. (2014) for instance, authors focus on extracting answers with the help of discourse markers (shallow discourse level) and parsed discourse relation (deeper level). When the text spans are extracted, they used a ranking system to select the final answer spans, which is very different from the aggregation methods in NLU.

[9] Refer to Equations $20 - 26$ for details in Fu et al. (2016).

[10] Refer to Equations $17 - 18$ and $19 - 35$ for details in Kraus and Feuerriegel (2019).

3.3.3.3 Pipeline Design

Typically, the process for incorporating discourse information into downstream applications involves using an off-the-shelf discourse parser, like DPLP (Ji and Eisenstein, 2014), HILDA (Hernault et al., 2010), or Two-stage (Wang et al., 2017b). The parser is trained on a gold annotation corpus, such as RST-DT or STAC, and then used on the target test set to generate discourse trees. However, this approach has two drawbacks: First, the pre-trained parser may introduce errors due to the cross-domain mismatch between training and testing data. Second, the choice of discourse framework with regards to the downstream task is not clear given the wide spectrum of discourse frameworks available (RST (Mann and Thompson, 1988), PDTB (Prasad et al., 2008a), Graph Bank (Wolf and Gibson, 2005), SDRT (Asher and Lascarides, 2003)). Additionally, the different types of relation variants within the same framework further add complexity. For instance, while the GUM corpus is annotated with the same RST framework as RST-DT, the relation classes differ between these two datasets. This makes it challenging to directly transfer a parser trained on GUM to be tested on RST-DT or vice versa.

In contrast, some studies, such as Narasimhan and Barzilay (2015), opt not to rely on externally trained parsers but to induce relations between sentences while optimizing a task-specific objective. They proposed a joint probabilistic model to identify single or multiple relevant sentences given a question and established a rhetorical relation between them.

To address the issues of cross-domain supervised parsing, Liu and Lapata (2018) proposed a method for automatically inducing structural dependencies of text. They enlarged the sentence-level attention mechanism to document level, capturing the interaction among sentences and creating a latent discourse structure for a document. This approach has been adopted by several subsequent studies for single document summarization (Liu et al., 2019b; Karimi and Tang, 2019) and bias detection (such as fake news and political standing prediction (Devatine et al., 2022)). However, the method has been criticized by Ferracane et al. (2019), who found that the generated tree structures were often shallow and trivial and not well-aligned with human annotation.

At the end of this chapter, we offer some thoughts on how to better utilize discourse for downstream applications and drive advancements in discourse-aware NLP. There are several aspects for potential improvements, for instance:

(1) Discourse Parsing Performance: Admittedly, discourse parsing by itself is a hard task. The state-of-the-art RST and SDRT parsers are now achieving respectively low 50 (Parseval metric in Nguyen et al. (2021)) and low 60 (micro F_1 score, presented in Table 3.1) on full parsing. In the study conducted by Ji and Smith (2017), authors explored the relationship between parsing performance and the gains observed in text classification. Through training on different sizes of subsets of annotated data, they discovered a positive correlation, suggesting that enhancing discourse parsing, either by using larger annotated datasets or improving the models, could yield greater improvements in downstream applications.

(2) Domain Adaptation Methods: To bridge the gap between training and target domains, adaptation methods such as direct discourse annotation for genres of interest, as suggested in Ji and Smith (2017), could be an efficient approach.

(3) Discourse Information Incorporation: Further investigation is needed to determine which aspects of discourse information are necessary for a given task (Section 3.3.3.1), as well as how to best integrate it (Section 3.3.3.2). The discourse community would benefit from more studies such as Ji and Smith (2017) and Louis et al. (2010).

(4) Hybrid Data Annotation: The lack of large and unified annotated datasets for discourse is a major factor contributing to the gap between syntactic and discourse parsing. While the Universal Dependencies serve as a substantial resource for syntactic parsing, the training dataset for discourse parsing is still limited in size and domain. As a result, training high-quality discourse parsers is a challenging task. One potential is to leverage the powerful GPT-like large language models (such as ChatGPT[11] and other similar models like InstructGPT (Ouyang et al., 2022) and GPT-4 (OpenAI, 2023)) to assist in the annotation process, with human intervention being reserved for more challenging cases.

[11] `https://openai.com/blog/chatgpt`.

Part II

Discourse Structure Discovery

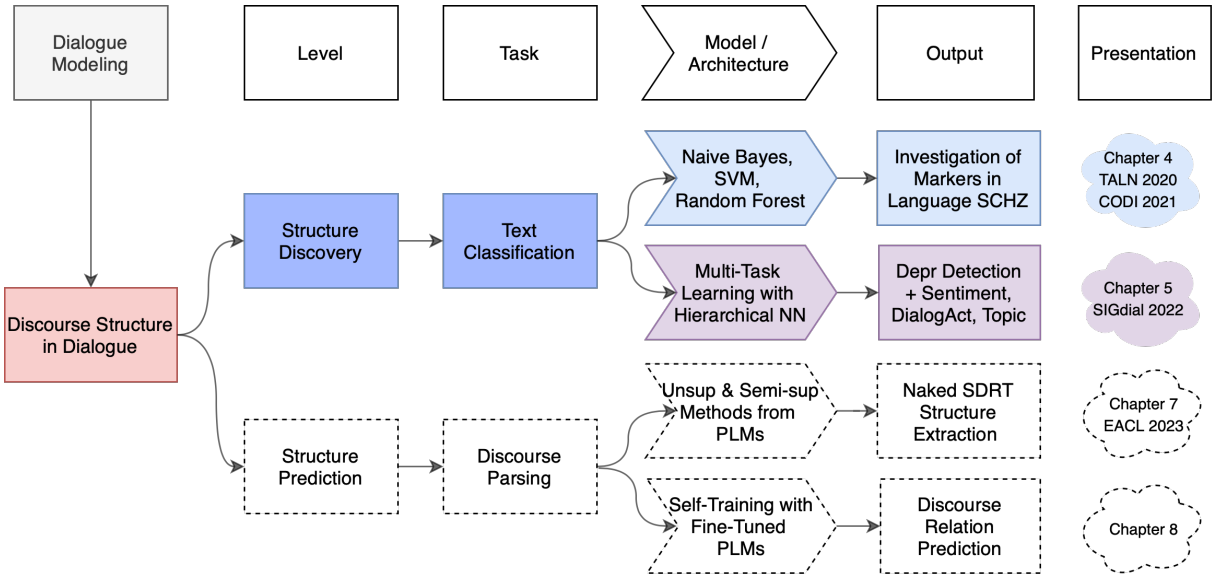
In Part I, we have provided an introduction to the fundamental theories and state-of-the-art models for discourse parsing. With this groundwork laid, we now turn our attention to addressing the first research question of this thesis, namely:

RQ1 How can we effectively use discourse and structural information as linguistic features in text classification tasks for dialogue, such as mental disorder illness detection?

Our focus lies in discovering discourse structure in dialogues, a task that presents both similarities and differences to that of monologues. To achieve this, we initially concentrate on a specific domain: the detection of mental disorders. Through an analysis of linguistic features present in the language production of patients, we hope to unveil structural information in dialogues.

Chapter 4 describes our efforts to identify language specificities of Schizophrenia, using a small dataset of 41 clinical interviews in French. We extract linguistic features at different levels and discover a significant lexical bias when using lexical features, highlighting the importance of corpus study in similar genres. Although not yielding the best performance, we promote the utilization of discourse and dialogue level information, as they uncover intriguing language particularities. In Chapter 5, we address the challenge of interaction modeling in dialogues by switching to another mental disorder – depression detection – and using a larger corpus with neutral data collection process: DAIC-WOZ. We employ a multi-task neural network system and consider different levels of information: from speech turns to the entire dialogue. To leverage information from semantic- and structural-related tasks, we utilize emotion classification in conversation, dialogue act and topic classification in the DailyDialog dataset. Our aim is to explore the potential of hierarchical neural networks in capturing different levels of information and to learn dialogue structure through various tasks.

Along the way, we encounter various challenges, such as data scarcity, innate bias, and interaction modeling. We devise several strategies to address each of these challenges and receive promising results.



Chapter 4

Investigating Language Markers of Schizophrenia in Dialogues

Contents

This chapter aims to address the query of how to utilize linguistic markers such as discourse in classification tasks within dialogue settings. Specifically, we focus on the identification of the language used by individuals with Schizophrenia during spontaneous conversations. Schizophrenia is defined as a severe mental illness (APA, 2015) that comes with varied symptoms, ranging from delirium to hallucinations. Among these symptoms, there are language disorders, especially the so-called *positive thought disorder* (i.e., disorganized language output such as *derailment* and

tangentiality) [1] and *negative thought disorder*[2]. Schizophrenia affects about 1% of the world's adult population, with cognitive troubles for around $70 - 80\%$ of the patients (Potvin et al., 2017). Since the symptoms often affect language skills, several studies proposed using NLP techniques on patients' productions (details in Section 4.1). These studies can help to identify what is affected in language, thus understand better the disease and its symptoms and how language works in general. Another goal of such studies could be to design systems that would help psychiatrists with diagnosis, by giving them additional indices through simple discussions, possibly alleviating the need for the patients to go through several cognitive tests, but this is a long-reach goal. In this study, we explore linguistic markers, especially discourse related markers through feature exploration within a classification system. We do so on spontaneous dialogues in French where all the previous work was in English and most used social media data or monologues. Replicating state-of-the-art results allows us to confirm some previous findings of specific features of the language of Schizophrenia.

Our study focuses on two aspects: carefully exploring data representations and investigating preliminary modeling of dialogues, both with scarce data. Using spontaneous conversations makes for a realistic scenario – the patient is merely talking with her clinician. However, representing dialogues is a challenging task: preliminary experiments indicate that it is easy to distinguish between the speech turns of clinicians when they talk with patients or controls (detailed scores in Table 4.2), possibly due to their proficiency in adapting the topic of conversation according to the participant. To mitigate any bias from clinician's speech, we thus restrict ourselves to patients' speech turns, and test varied context windows to tackle data sparsity. Additionally, we compare several representations and confirm that lexicon is a good indicator, making for high-performing models with at best 93.7% (accuracy). Nevertheless, our analysis demonstrates that it probably corresponds to a bias in our data caused by the constraints imposed during the collection process. Most of the datasets are likely biased the same way. This analysis leads us to delexicalized models while focusing on dimensions presumed to be affected in Schizophrenia: morpho-syntactic, syntactic, dialogue, and discourse information are therefore considered. Our best delexicalized model gets 77.9% accuracy and shows the importance of morpho-syntactic information and high-level features in dialogue.

When dealing with medical data, ethical questions arise (Le Glaz et al., 2021). The diagnosis of Schizophrenia is complex and relies on many indices. In Martinez-Martin et al. (2018), authors raise concerns about the ethical implications of using machine learning prognostic estimates to treat psychosis. They question whether the model's validity could be affected by local context variables such as differences in psychiatric practice and social support. It is evident that AI systems cannot replace the expertise of a human in diagnosing medical conditions. They cannot provide diagnoses, but only human can. We concur with Martinez-Martin et al. (2018) that linguistic cues, while essential, must be understood in the context of a patient's social environment. In this study, we only focus on linguistic factors. In future research, we suggest incorporating both linguistic and socioeconomic criteria when designing machine learning algorithms, thus creating more objective AI tools for psychiatric research and practice.

This chapter is adapted from two publications: one paper at the 27th French National Conference of NLP (TALN 2020) (Amblard et al., 2020) and one paper at the 2nd Workshop on Computational Approaches to Discourse (CODI 2021) (Li et al., 2021a). It is structured as follows: we begin with an overview of previous studies that utilized NLP and machine learning

[1] *Derailment*: spontaneous speech that tends to slip off track. *Tangentiality*: reply to a question in an oblique or irrelevant manner.

[2] *Negative thought disorder*: poverty of speech (known as *alogia*) and poverty of content.

techniques for mental health in Section 4.1, with a specific focus on Schizophrenia detection. Through examining the key aspects of these related works, we observe that our study differs from its predecessors in that the latter primarily utilize social media data and/or do not address the issue of lexical bias. In Section 4.2, we introduce our methods, providing details on dialogue modeling and data representation. Our experiments are conducted on a French corpus created under the project SLAM, which we present in Section 4.3. Results of our experiments are presented in Section 4.4[3], followed by a series of analyses in Section 4.5. Finally, we conclude our study in Section 4.6.

4.1 Related Work

4.1.1 Mental Disorder & Linguistic Clues

A long line of work in psychiatry, starting in the 1960s, proposed descriptions of language output of patients with Schizophrenia, as reviewed in Kuperberg (2010b). Psychiatrists rely on language and speech behavior as one of the main clues in psychiatric diagnosis (Ratana et al., 2019). They found that these patients' speech tends to be less predictable (Salzinger et al., 1964, 1970; Salzinger, 1979), with a poorer vocabulary (Salzinger and Hammer, 1963; Manschreck et al., 1991). It has also been found that their language productions tend to be more grammatically deviant (Hoffman and Sledge, 1988) and less syntactically complex than that of controls (Fraser et al., 1986; Morice and Ingram, 1982). At discourse level, they associate words within a larger context than controls (Maher et al., 2005) with often more diffuse associations (Chaika, 1974; Elvevåg et al., 2007). They also present referential impairments – categorized as vagueness, missing information, or confusing reference (Rochester, 2013; Docherty et al., 1996) –, and specific discontinuities at the discourse level (Musiol and Trognon, 2000; Rebuschi et al., 2014).

On the other hand, many researchers have used NLP methods to help to identify varied mental disorders, such as depression (De Choudhury et al., 2011, 2013b,a; Schwartz et al., 2013; Nguyen et al., 2014; Sekulić and Strube, 2019; Howes et al., 2014; Guntuku et al., 2019), post-traumatic stress disorder (PTSD) (Pedersen, 2015; He et al., 2017; Kleim et al., 2018), suicide risk (Coppersmith et al., 2016; Benton et al., 2017; Coppersmith et al., 2018), Alzheimer's disease (AD) (Orimaye et al., 2014; Fraser et al., 2016; Gosztolya et al., 2019), and autism (Goodkind et al., 2018; Sakishita et al., 2019). We also investigate depression detection in the following chapter.

4.1.2 Detection of Schizophrenia in NLP

The automatic detection of Schizophrenia is an active field of research, with studies focusing mainly on two types of characteristics: biomedical signals such as electroencephalography (EEG) and magnetic resonance imaging (MRI) (Greenstein et al., 2012; Sabeti et al., 2011). Although it is evident that the language produced by individuals with mental illness patients differ from that of others, research based on linguistic data is relatively scarce. It is only in recent years that we have observed a trend in the use of NLP techniques for the automatic detection of various disorders, such as depression (Pestian et al., 2017) or in combination with other disorders such as PTSD (Pedersen, 2015) and pre-symptoms of Alzheimer's disease (Jarrold et al., 2010). Early line of work has mainly focused on lexical information (Hong et al., 2012; Mitchell et al., 2015; Birnbaum et al., 2017a; Xu et al., 2019), with different types of data including that generated with the help of practitioners such as interviews and questionnaires, as well as data freely generated

[3]Our code is available on: `https://github.com/chuyuanli/non-lexical-markers-scz-conv`.

Work	Corpus	#Instances	Lang	Feature	Result
Strous et al. (2009)	Essay writing	72	Hebrew	lexical	Acc 83.3%
Mitchell et al. (2015)	Tweets	348 × 2800*	English	lexical	Acc 82.3%
Kayi et al. (2017)	Essay writing	373	English	synt, semantic	F_1 70.3%
	Tweets	348 × 2800*	English	synt, pragm	F_1 81.7%
Allende-Cid et al. (2019)	Oral narratives	189	English	morpho-synt	F_1 82.8%
Amblard et al. (2020)	Clinic interview	41	French	lexical	Acc 93.7%

Table 4.1: Related work in identification of Schizophrenia. *: 348 users with average 2800 tweets per user. Kayi et al. (2017) utilize the same data set of Mitchell et al. (2015).

by patients on social media. In interviews, practitioners have some control over the content, for example, they can lead the discussion towards recent treatment. On the other hand, social media data is completely free text created by self-stated diagnosed patients and is often selected with specific hashtags such as "#stress" or "#depression" (Le Glaz et al., 2021).

Regarding the NLP techniques used in Schizophrenia research, we find a limited number of relevant studies, which are presented in Table 4.1, with key information such as data type, the size of corpora, language ("lang"), feature employed, and classification results.

Before discussing these studies, we need to acknowledge that comparing the corpora used in related works is challenging due to several reasons. Firstly, they differ in size. For instance, the datasets in Strous et al. (2009) and Kayi et al. (2017) contain a varying number of essays composed by patients and controls, each with different length requirements (in Strous, $300-500$ words; in Kayi "two-paragraphs"). The Twitter dataset in Mitchell et al. (2015) includes a certain number of users and their average tweets per user, where each tweet is limited to 140 characters at the time of collection (2008-2015). As for the oral narratives in Allende-Cid et al. (2019), there is no information about the length of the documents. Secondly, all datasets are not publicly available due to confidentiality issues, including the one we use (Amblard et al., 2020). When attempting to create our Twitter dataset, we follow the acquisition method in Mitchell et al. (2015) and extract public tweets from around 600 users, each with an average of 1,890 tweets. We perform an initial human annotation phase but have to discontinue the project due to GDPR regulations in 2020[4], which prohibit using newly acquired data from social networks. To overcome the limitations caused by the size and type of data, we decide to explore other cognitive impairments detection using larger and readily available datasets, such as depression and DAIC-WOZ dataset (DeVault et al., 2014). We discuss this further in the next chapter.

In the first study dedicated to Schizophrenia detection problem, Strous et al. (2009) used written documents from individuals with Schizophrenia to build classification systems based on lexical information and achieved an accuracy of 83.3%. They observed specific traits in individuals with Schizophrenia such as a more restricted use of prepositions and an over-representation of the first person. Then, several studies were conducted using Twitter messages written by individuals self-identifying as having Schizophrenia. Mitchell et al. (2015) collected data for 174 patients (at most 3200 tweets/user, in average 2800/user) and tested different sets of lexical traits, such as semantic categories from a lexicon or Brown clusters: they presented classification systems (SVM) with a best accuracy of 82.3%. This study was extended in Birnbaum et al. (2017a) using 1.9 million tweets collected for 146 patients. They also achieved high scores, with 90.0% accuracy, using lexical traits, particularly categories from the Linguistic Inquiry and Word

[4]`https://commission.europa.eu/law/law-topic/data-protection/eu-data-protection-rules_en`

Count (LIWC) lexicon (Pennebaker et al., 2001). They observed, as previously mentioned, an increased use of first person pronouns and terms belonging to the health lexical field.

Unlike ours, these studies rely on LIWC categories – psycho-metrically validated lexicon mapping words to psychological concepts, Latent Dirichlet Allocation (LDA) (Blei et al., 2003) – inferring topics in each document, and Brown clustering (Brown et al., 1992) – grouping contextually similar words into the same cluster. Despite the good performance, most of these resources are only available in English.

More recent approaches considered syntactic, semantic, and pragmatic (level of committed belief and sentiment) information: Kayi et al. (2017) compared syntactic (POS and syntactic parses), semantic (semantic role labelling), lexical (LDA, clusters), and sentiment annotations on tweets and also narrative texts written by patients (373 LabWriting essays). They are the first to investigate whether patients exhibit more negative sentiment than controls. For this purpose, they employed Stanford Sentiment Analysis Tool (Socher et al., 2013b) and Columbia sentiment classifier trained for social media (Rosenthal and McKeown, 2013) on essay writing and tweets, respectively. The first tool yields a 5-way classification (*very negative*, *negative*, *neutral*, *positive*, and *very positive*) as well as intensity scores. The second tool gives a 3-way (*negative*, *neutral*, and *positive*) sentiment without intensity information. The combination of syntactic, semantic and lexical information give best scores (70.3% in F_1 for essays), and POS tags alone could achieve 69.8% in F_1. On tweets, using sentiment information seems to help, allowing to obtain 81.6% in F_1 when combined with lexical and semantic features. However, adding syntactic features leads to a small drop (78.6% F_1). Allende-Cid et al. (2019) used narrative texts written by patients to explore representations not directly involving the tokens, *i.e.*, POS tags (160 categories) – including gender and number –, and meta-POS (12 categories), the latter being more general categories (e.g., Noun, Verb and Determiner). They compared morpho-syntactic features to lexical ones, using bag-of-words (BOW), with four different classifiers: KNN, Random Forest, SVM and Adaptive boosting. They reported 87.5% in F_1 for BOW features, and observed a drop in performance for non lexicalized models, with, nonetheless, performance higher than chance: 75.1% for meta-POS and 82.8% for POS.

Aforementioned studies are based on narrative texts (essays and tweets). They show good performance with morpho-syntactic features, especially with Part-Of-Speech tags. We here demonstrate that some findings can generalize to **spontaneous conversations**.

In the dialogue setting, Howes et al. (2012, 2013) investigated linguistic features in the transcripts of conversations between patients of Schizophrenia and clinicians. Their studies focused on patient satisfaction and adherence to treatment prediction. Their dataset consists of 131 outpatient consultations, with an average length of 2,706 words per document. They used high-level features of the dialogue structure, – such as backchannels or overlap –, lexical features using pure words, and topics using LDA. For data representation, they worked on the concatenation of speech turns of the patient, the same way as we do in the task of Schizophrenia detection (Amblard et al., 2020; Li et al., 2021a). The results of including lexical features showed good performance, but their generalizability may be limited. On the other hand, using high-level features alone did not yield significant performance. As a result, the researchers concluded that although high-level dialogue factors were helpful in predicting symptoms, they were not good indicators for adherence or satisfaction measures.

Our position with respect to related work involves conducting two consecutive studies to detect Schizophrenia within dialogue settings. In the first study (Amblard et al., 2020), we follow similar studies in monologues by initially examining lexical features. We then expand on this by addressing data sparsity, analyzing the impact of context through varying dialogue window sizes, and dealing with lexical bias by comparing features at different levels, including

Features	SVM	LR	NB	RF	Perc	Best
bow	**99.02**	98.05	96.83	98.54	96.10	SVM
ngram	95.85	95.85	**97.32**	96.59	92.44	NB
POS	80.73	79.76	55.37	**85.12**	58.78	RF
POS+2-3-*treelet*	89.02	86.83	65.61	**90.98**	69.51	RF
Connectives	**92.44**	90.73	65.85	88.29	78.78	SVM
Connectives D	**90.73**	90.98	69.51	90.24	79.76	SVM

Table 4.2: Classification results of the speech of psychologist (talking to patients vs. talking to controls) using different classification algorithms. SVM: supported vector machine. LR: logistic regression. NC: naive bayes. RF: random forest. Perc: perceptron. Best: best algorithm.

sequences of POS tags, finer tree representations, and dialogue-level information, in a subsequent research (Li et al., 2021a). Due to the scarcity and dissimilarity of the data in comparison to monologues, we explore a model for detecting Schizophrenia symptoms using higher-level less lexicalized features (Section 4.2), inspired by the work of Howes et al. (2012).

4.2 Method: Better Data Representation & Feature Engineering

4.2.1 Task Simplification

Our dataset consists of 41 dialogues between a psychologist and a patient (either with Schizophrenia or as a control). These dialogues are lengthy, with an average of $6,013$ words per conversation and 268 speech turns per patient/control, each averaging $2,811$ words (details in Section 4.3.1). Ideally, we would modeling the interactions in dialogues and considering the speech turns from both parties. However, taking into account the features from psychologists' speech turns is a tricky task, since they may bring bias and influence the results. To examine such impact, we conduct experiments using psychologist's speech turns when they talk to patients or controls. The results are shown in Table 4.2. Clearly, high accuracy can be achieved using either lexical features (bag-of-words, n-grams) or syntactic features (part-of-speech tags, *treelets*). The simple merging of subjects' and psychologists' speech turns is evidently not an ideal way for data representation. As a result, we exclude the psychologist's speech turns in all dialogues to reduce their influence on classification and only focus on patients' speech turns in this study. In the following chapter, we take one step further and consider the interactions between two parties by using a neural network architecture, but for another classification task.

4.2.2 Varying Dialogue Size

To examine the effect of context length on our model, we create shorter sub-dialogues by dividing the original dialogues into smaller parts and using these as individual instances for classification. In the **Full** setting, we concatenate all of a patient/control's speech turns into a single large document, which may be difficult for the system to process due to the small number of classification instances (i.e., 41). The **Indiv.** setting classifies each speech turn individually, resulting in more instances (10,319) but losing the context of neighboring speech turns. Some of these speech turns are also quite short, with an average of 11 words. The **W-**n setting (where $n \in 128, 256, 512, 1024$) is a middle ground, using a window of at least n words (always extend-

	#Doc	#Speech turn/doc			#Token/doc		
Setting	total	min	max	avg	min	max	avg
Indiv.	10,319	1	1	1	1	274	11
W-128	893	1	34	11	128	317	145
W-256	443	1	72	20	256	424	271
W-512	209	2	129	42	512	609	530
W-1024	132	8	202	83	703	1,088	873
Full	41	76	555	268	703	6,778	2,811

Table 4.3: Document, speech turn, and token length per document in different dialogue settings in our corpus. All the numbers exclude the production of clinicians. "Indiv" regards every speech turn as an instance. "W-n" setting takes at most n tokens in an instance. "Full" takes one complete document as an instance.

ing to the end of the current speech turn) to identify distinctive features in smaller blocks of conversation. Values of n are selected in such a way that the length of the context gradually increases, almost doubling that of the previous setting. This configuration results in 893, 443, 209, and 132 instances, with an average of 11, 20, 42, and 83 speech turns, respectively. Key statistics of each setting is presented in Table 4.3.

4.2.3 Comparing Representations

We consider three different types of feature representation: (1) lexical, including bag-of-words (BOW) and n-grams; (2) syntactic (i.e., non-lexical) features such as part-of-speech (POS) tags and syntactic relation chains *n-treelet*; (3) semantic and pragmatic features where we consider three markers commonly seen in dialogues: *Open Class Repair* (OCR), *Backchannel* (BC), and discourse connectives.

(1) Lexicons: Existing work investigating language particularities for Schizophrenia demonstrated the importance of lexical features (Mitchell et al., 2015; Kayi et al., 2017). For French, as for many languages, we do not have access to a resource such as LIWC. We thus propose to simply include Bag-Of-Words (BOW) and n-grams ($n \in \{2, 3\}$) to our models as a proxy for topic identification. BOW representation is the simplest and serves as a reference system, while n-grams can contain words that span multiple turns of a single speaker, and thus encode part of the dialogue context.

(2) Syntactic Markers: To build more general models, we test the two following non-lexical features: part-of-speech (POS) tags and *treelets*. Allende-Cid et al. (2019) demonstrated that POS tags are effective features. We also test for larger patterns with sequences, POS n-gram with $n \in \{1, 2, 3\}$. Kayi et al. (2017) only used the dependencies as syntactic features. We extend to *treelet* features.

A *treelet* refers to subtrees obtained from a syntactic parse. This feature has been utilized in multiple NLP tasks, such as machine translation (Quirk et al., 2005) and language modeling (Pauls and Klein, 2012). In the first study, the authors suggested combining conventional statistical models with phrasal *treelets* that have linguistic generality. In (Pauls and Klein,

Figure 4.1: Examples of two forms of 3-*treelet*, adapted from Johannsen et al. (2015). The left *treelet* has one head and two dependents. The right *treelet* has a chain of dependencies.

2012), the authors utilized *treelets* to create a generative syntactic language model. The model outperformed traditional n-gram models in grammaticality tasks and achieved better fluency.

In our case, precisely, we use UDPipe (Straka and Straková, 2017) to obtain morpho-syntactic tagging and dependency parsing results. Since our data is dialogue and not monologue, pre-trained models perform poorly. Therefore, we re-train a UDPipe model[5] using a French spoken language corpus Spoken-French 2.5[6]. Preprocessing includes removing punctuation and minimal segmentation (e.g. adding a space for apostrophes). To encode the syntactic features, we use the method proposed in Johannsen et al. (2015), which consists of extracting all sub-trees (*treelet*) of at most 3 tokens. A *treelet* of 1 token is simply the associated POS tag, such as `VERB` and `NOUN`. A 2-*treelet* corresponds to 2 tokens with a syntactic relation between a head and a dependent, e.g., "`VERB`→`nsubj`→`NOUN`" where the head is `VERB` and the dependent is `NOUN`. Finally, a 3-*treelet* corresponds to 3 tokens with 2 syntactic relations: it could be one head dominates two dependents or a chain of dependencies, e.g., "`NOUN`←`nsubj`←`VERB`→`dobj`→`NOUN`" or "`PRON`←`poss`←`NOUN`←`nsubj`←`VERB`", as illustrated in Figure 4.1.

(3) Discourse & **Dialogue Markers:** Finally, we also test higher-level features that involve discourse and dialogue information. Howes et al. (2012) showed the importance of features specific to spontaneous dialogues that do involve lexicon but in a more generic way: *Open Class Repair* (OCR) initiators such as *pardon?*, *huh?*; *Backchannel* (BC) responses such as *yeah*, *hum mm*.

To reflect text organization, we also include discourse features by extracting the forms (without disambiguation) corresponding to connectives (*but*, *because*, *since*) as identified in LexConn (Roze et al., 2012), as well as the disambiguated connectives. Connectives differ in terms of their specificity and can be ambiguous at two levels (Laali and Kosseim, 2017): (1) they can be used in *discourse-usage* or *non-discourse-usage*. The word *and*, for instance, can signal a very unspecific addition – which is nonetheless distinct from the mere juxtaposition of clauses without connective (Dipper and Stede, 2006). (2) They may be used to signal more than one discourse relation. For instance, connectives *however* and *but* can signal *Contrast* or *Concession* relations. In our initial extraction, we find a few unspecific connectives with relatively large proportion: *et [and]* (12%), *donc [so]* (7%). It thus motivates us to disambiguate these connectives. We use an off-the-shelf discourse parser [7] from Laali and Kosseim (2017) which achieves $> 93\%$ accuracy of disambiguation on French Discourse Treebank (FDTB1) (Steinlin et al., 2015). We distinguish raw connectives with disambiguated connectives with respectively "Conn" and "ConnD" in the following sections.

In order to improve reproducibility, we give the full list of tokens used for OCR (Table 4.4) and *Backchannel* (Table 4.5), as well as their corresponding translation in English aside. The

[5] `https://ufal.mff.cuni.cz/udpipe/2/models#universal_dependencies_26_models`.
[6] Renamed as Rhapsodie treebank: `https://tinyurl.com/UniversalDependencies-French-S`
[7] `https://github.com/mjlaali/french-dc-disambiguation`

French version is obtained by translating the list given by the authors Howes et al. (2012) who predicted adherence to treatment for Schizophrenia from dialogue transcripts, and by adding a few additional terms specific to French with the help of a psychologist in the team.

French	English	French	English
pardon vous disiez	pardon you said	ah vous parlez pardon	ah sorry you were saying
pardon	pardon	excusez-moi	excuse-me
excuse moi	excuse-me	bon je suis désolée	i am sorry
désolé(e)	sorry	(ah) ouais ?	(oh) yes?
ah bon ?	oh really?	c'est vrai ?	is it true?
c'est euh ?	it's euh?	hum ?	huh?
de quoi	of what	c'est quoi ?	what is it?
c'est-à-dire	which means	euh ?	euh?
dites moi plus	tell me more	mais encore	but still

Table 4.4: Open Class Repair initiators list (original French, with English translation).

4.2.4 Feature Selection

Our learning problem is confronted with high-dimensional features (*treelet* $> 16k$ vocabulary, n-gram $> 118k$ vocabulary) and rare training instances (41 documents), which generally leads to overfitting and lack of generalization of the models. We include feature selection during training with a method implemented in Scikit-Learn[8] (Pedregosa et al., 2011). By calculating the weights (or coefficients) assigned by a model to each feature and keeping only those whose weight is above a threshold, this method allows us to select important features. We test without selection (threshold "None"), then with a threshold corresponding to the mean and median of the obtained weights, as well as 10 values regularly distributed between $1e-5$ (the default value in the used implementation) and the weight of the 50^{th} most important feature. This maximum value chosen *a priori* ensures that at least 50 features are retained in the model. Feature selection allows us to drastically reduce the size of the vocabulary, especially for lexical (bow and n-gram) and syntactic features (*treelet* and its n-gram combination). In Table 4.6, we present the original and selected feature sizes in the W-1024 setting. The selection is carried out using the SVM classifier. We can observe that the original number of 2-grams and 3-grams was more than $118k$, but it reduces to around 100 after the selection. Similarly, the selected features for POS+2-3-*treelet* are also reduced to less than one percent of its original size.

4.3 Experimental Setting

4.3.1 Dataset

SLAM Project: Our corpus is developed as part of the SLAM project. The interviews are conducted in a hospital setting with patients diagnosed by psychiatric doctors and psychologists from the host institution. The interview is accompanied by neuropsychological tests to measure the patients' abilities in various areas (working memory capacity, verbal fluency, attention, motor speed, executive functions, etc.). In addition, the patients' verbal interactions with a psychologist

[8] `https://scikit-learn.org/: feature_selection.SelectFromModel`.

French	English	French	English
oui	yes	ouais	yeah
ouais voilà	yeah that's it	oui c'est ça	yes that's it
oui bah oui	yea euh yes	oui... forcément	yes... for sure
bah ouais	euh yeah	hum (hum)	hum (hum)
muh mmh	muh mmh	mmh/mmhh	mmh/mmhh
d'accord	okay	ok	ok
voilà	that's it	c'est ça	that's it
c'est vrai	that's true	c'est sûr	(yes) (for) sure
ça c'est clair	that's clear/clearly/definitely	eh bien sûr	euh of course
carrément	completely	bien sûr	of course
super	super	ok... bon	ok... then
d'accord ça marche	okay it works	certes	certainly
mais hein	but hein	je comprends	i understand
vraiment	really	bien	well
bon	good	très bien	very good
quand même	still	tout à fait	exactly
certainement	certainly/sure	exactement	exactly
tant mieux	all the better	oh	oh
ah	ah	ben	well...
alors ben	well...	ah d'accord	ah okay
ah ça euh	ah (this) euh	eh bah c'est bien	euh well that's good

Table 4.5: Backchannel response list (original French, with English translation).

are recorded during a semi-directed interview. The patients' participation is voluntary and the elements collected during the experiment are not used by the medical team for the patient's follow-up. There is therefore real freedom in the interview. The themes addressed remain simple: patient's daily life, medical history, history before hospitalization, etc.. These interviews are recorded with a double eye-tracker system, but we did not utilize the eye movements as features in our experiments. The interviews are conducted by a psychologist who is not personally involved in the dialogue. Therefore, this is not a symmetrical everyday interaction situation, the patient's speech is closer to a monologue. This explains our choice to extract the speaker's linguistic production and isolate it as a coherent whole. Amblard et al. (2014) explains why the distribution of SLAM data is difficult due to the content of the interviews giving many geographical and biographical elements of the patient and their surroundings so that anonymization does not make sufficiently opaque.

Corpus Description: The corpus consists of 41 documents, 18 people with Schizophrenia and 23 controls for the control group. The transcripts are standardized and follow a transcription guide (Rebuschi et al., 2014). Only one psychologist interviews these 41 subjects. Each of these groups contains 15 male subjects, the rest (3 and 8) being female. This distribution therefore presents a bias. It is accepted that there are significant differences according to gender (clinical and paraclinical aspects) (Douki Dedieu et al., 2012). At the moment, the majority of studies focus mainly on male subjects and we think that these differences will have to be taken into account in the diagnostic approach. The groups are balanced with age, intelligence quotient

Feature	#Original	Threshold	#Selected	%Selected
bow	6,504	median	3,254	50.0
2-3-gram	118,473	8	98	0.1
1-2-3-POS	2,031	5	198	9.7
2-*treelet*	879	7	92	10.5
3-*treelet*	14,103	7	155	1.1
POS+2-3-*treelet*	14,996	7	124	0.8

Table 4.6: Original and selected numbers of features using SVM classifier, in W-1024 setting.

	In #doc	#Speech turn/doc	#Word/doc	#Word/speech turn	%Gram words
Patient	18	200	2676	13.4	56%
Psy-scz	18	200	1815	9.1	50%
Control	23	342	3305	10.8	51%
Psy-control	23	307	4779	15.6	54%

Table 4.7: SLAM corpus statistics of different participants. "Psy-scz": psychologist's statistics when talking to patients; "Psy-control": psychologist's statistics when talking to controls. "Gram words": grammatical words $\notin \{noun, verb, adv, adj\}$.

(IQ) score, years of studies, and three cognitive tests' results (WAIS-III, TMT, CVLT)[9].

We show in Table 4.7 key statistics in regards to speech production of different participants: patients, controls, psychologist when she talks to patients or controls. Not surprisingly, people with Schizophrenia have, on average, the same number of speech turns per document as the psychologist (200). However, they speak more (2676 words per document) and their sentences are longer (13.4 words per sentence) compared to the psychologist (1815 words per document, 9.1 words per sentence). The controls express themselves significantly more (342 speech turns and 3305 words per document) with shorter sentences (10.5 words per sentence). People with Schizophrenia also have a higher rate of use of grammatical words (also known as function words which not belonging to the categories: noun, verb, adverb or adjective) than the psychologist or the controls: SCZ 56% vs. controls 51% vs. psychologist 50%, as observed in Hoffman and Sledge (1988). The grammatical deviance could also explain the good results when using POS tags as features in previous studies (Kayi et al., 2017; Allende-Cid et al., 2019).

For illustration, we show two translated excerpts with commonly seen themes in example (36) and (37), where the acronyms "SCZ", "PSY", and "CON" refer to Schizophrenia patient, psychologist, and control, respectively. The high-lighted words are typical theme terms that appear in different groups, which we will discuss more in Section 4.5.1.

(36) Psychologist - Schizophrenia

PSY: So now you are going to a workshop hum, what is it?

SCZ: Yes, so I went to a therapeutic workshop... what do they call it...

PSY: Therapeutic education... right

[9]WAIS-III: Wechsler Adult Intelligence Scale (WAIS) is an IQ test designed to measure intelligence and cognitive ability in adults and older adolescents. Trail Making Test (TMT) is a widely used test to assess executive abilities in patients. California Verbal Learning Test (CVLT) measures episodic verbal learning and memory.

(37) Psychologist - Control

PSY: What do you want to do after?

CON: Uh I would like to do the master of psychopathy of the cognition and the interactions.

PSY: Mmh mmh.

4.3.2 Implementation Details

We compare several classification algorithms: Support Vector Machines (SVM), Logistic Regression (LR), Random Forest (RF), Perceptron (Perc), and Naive Bayes (NB), without and with feature selection based on importance weight (Section 4.2.4), all implemented in Scikit-Learn library (Pedregosa et al., 2011). We have tested the following hyper-parameters:

- Naive Bayes: smoothing $\alpha \in V = \{0.001, 0.005, 0.01, 0.1, 0.5, 1, 5, 10, 100\}$;
- Logistic Regression: L_2 and regularization $C \in V$;
- SVM with linear kernel: L_2 and regularization $C \in V \cup \{1000\}$;
- Perceptron: L_2 and $\alpha \in V$;
- Random Forest: max_depth $\in \{2, None\}$;

For the thresholds employed for feature selection, recall that we use 10 values equally distributed from $1e-5$ to the weight of the 50^{th} most important feature (thus allowing to keep at least 50 features), as well as the mean and median values of the weights.

Since our dataset is minimal (41 documents), we use nested cross-validation to assess the performance of our system: we tune hyper-parameters on $K-1$ folds and then evaluate on the left-out fold; we repeat the whole process M times ($K = M = 5$). We report average accuracy over the M out folds. Best hyper-parameters values and algorithms are given in Appendix A.2.

4.4 Results

In Table 4.8, we present all the baseline results in the first row. In all settings, the majority class is the control group. The full setting includes the initial groups with 23 dialogues for the control group and 18 for the patient group. As the control group produces more utterances on average, the Indiv. setting highlights a more pronounced imbalance, with a baseline of 65%. We showcase lexical, syntactic, and dialogue and discourse features in different blocks. The last block consists of selected combinations of features, particularly those that performed well individually, such as n-POS and BC.

Single feature-wise, lexical features perform the best, with BOW achieving outstanding scores of 93.7% (92.2% in F_1) and 72.4% (71.6% in F_1) for the Full and Indiv. settings respectively. The best algorithm is Naive Bayes. Using SVM, we obtain accuracy of 90.98% and 70.2%. These results are superior to those presented in Allende-Cid et al. (2019) (87.50% in F_1), which also use a BOW representation and SVM, but with a larger corpus, and also higher than those presented in Birnbaum et al. (2017b), which achieved 90% accuracy with Random Forest classifier on a larger Twitter dataset, using n-grams ($n = 1, 2, 3$, corresponding to BOW+n-gram in our case) and semantic categories from LIWC lexicon. This suggests potential lexical bias in our dataset. We will provide more discussion on this point in Section 4.5.1.

As for discourse markers, BC and connectives are good indicators, with 74.5% and 76.7% in their best settings. BC is more influenced by the context length compared to connectives, likely due to the limited vocabulary in short contexts. In the extreme case of the Indiv. setting, where single speech turns are considered as instances, which excludes inter-sentence connectives, the accuracy is very low, even lower than the baseline. The distinction between regular connectives and disambiguated connectives (ConnD) is large, with the latter obtained through an off-the-shelf discourse parser proposed in Laali and Kosseim (2017). We will delve deeper into the disambiguation aspect in the following Section 4.5.3. Syntactic features, such as POS tags and *treelets*, are effective markers, particularly for longer chains ($n = 2, 3$). These are lexical-free features that allow us to discern the language usage between two groups, as we shall see in Section 4.5.2. A consistent trend is observed, where the best performance is achieved with longer window sizes between 512 and 1024, using SVM, which is known to perform better with longer context and sparser data. We provide further details on the performance of classifiers in Section 4.5.6.

We evaluate various combinations of features within and across different feature groups, and the most effective ones are presented in the last block in Table 4.8. When combining POS and treelet, the performance increase is minimal. However, when adding BC, we observe a notable improvement: 3-POS+BC being the highest performing system with an accuracy of 77.86, followed closely by 2-POS+BC at 76.6. Similar to the syntactic features, SVM is the best algorithm. The second-best combination is POS and disambiguated connectives, with 3-*treelet*+ConnD yielding 76.6 and 75.7 in the Full and W-1024 settings respectively. It is worth noting that replacing 3-*treelet* with 2-*treelet* also yields impressive results: 75.2 and 74.2 (details in Appendix A.1). Longer syntactic chains tend to capture more precise language specificities and therefore provide stronger clues for classification. We invite readers to refer to Appendix A.1 for the performances with all the features and algorithms in each context setting.

4.5 Analysis

4.5.1 Lexical Features & Bias

Building on our work (Amblard et al., 2020), we compare different representations for Full and Indiv. settings - the most similar to long narrative texts or short Twitter messages. As in previous work, we find that lexical information is very effective (first sub-part in Table 4.8) with at best 93.66% in accuracy. However, analysis from precise studies suggested a potential issue: Mitchell et al. (2015) reported that *health-related* lexicon is more represented in Twitter dataset, and Howes et al. (2012) showed that the most predictive unigrams are about *conditions*, *treatment*, and *medication*. Similarly, we conduct an examination of the most commonly used words for people with Schizophrenia and controls (as in examples (37) and (38)). We observe the following themes:

- For people with Schizophrenia: typically words related to pain such as “disease”, “hospitalization”, and “hallucinations”. This corresponds to the *Catastrophe* label among the top semantic features observed by Kayi et al. (2017) who present linguistic traits that are predictive of people with Schizophrenia in writing. From this empirical analysis, we can see the conversational context in which patients were indirectly led to mention the onset of their illness.

- For control subjects: words related to education such as “master”, “thesis”, and “degree” and to psychology such as “psychiatrist” and “psychologist” stand out significantly. It happens

Features	Full	Indiv.	W-128	W-256	W-512	W-1024
Majority baseline	56.1	65.4	60.9	60.1	60.2	59.8
Lexical features						
bow	**93.66**	**72.43**	-	-	-	-
ngram	85.61	69.59	-	-	-	-
Syntactic features						
POS	53.66	55.80	60.63	60.48	60.09	57.18
2-POS	67.36	56.33	64.85	68.53	71.74	71.11
3-POS	71.65	56.53	65.39	70.66	72.55	71.71
2-*treelet*	69.19	56.73	65.02	70.11	74.19	74.63
3-*treelet*	66.78	55.34	63.95	66.39	69.03	70.31
Dialogue & discourse features						
OCR	60.62	50.17	52.43	55.19	59.28	67.26
BC	74.48	54.79	62.01	66.89	67.86	63.82
Connectives	72.44	55.28	64.05	69.68	73.57	76.73
Connectives D	67.11	53.79	58.61	65.13	67.15	70.67
Feature combination						
1-2-3-POS	69.01	58.36	66.19	72.03	72.67	72.52
POS + 2-3-*treelet*	66.59	57.77	65.52	69.11	72.39	71.43
3-POS + BC	74.93	57.46	**69.92**	**73.75**	**77.86**	**75.20**
3-POS + ConnD	74.52	57.11	65.05	72.54	72.68	74.23
3-*treelet* + ConnD	76.61	56.04	63.70	69.28	69.67	75.74

Table 4.8: Majority baseline and best averaged accuracy for Full, Individual, W-{128, 256, 512, 1024} settings. OCR: *Open Class Repair*; BC: *backchannel*; Connnectives D or ConnD: desambiguaged connectives. In bold: the best score for each column; underlined text: the best score for each row.

Vocabulary (fr)	Translation (en)	ρ	p-value
Douleur	Pain		
maladie	disease	0.540	$< 1e-3$
hospitalisé	hospitalized	0.509	$< 1e-3$
hallucinations	hallucinations	0.420	0.006
Éducation	Education		
master	master	-0.505	$< 1e-3$
concours	competition	-0.496	$< 1e-3$
fac	college	-0.490	0.001
Psycho	Psycho		
psychologie	psychology	-0.536	$< 1e-3$
psychologue	psychologist	-0.453	0.002
Déictique	Deictic		
j' / je	i	0.635	$< 1e-5$
mon	my	0.613	$< 1e-5$
t' / tu	you	-0.467	0.002
nous	we	-0.342	0.028

Table 4.9: ρ- and p-value of Spearman test for BOW lexical features in SLAM.

that the control subjects are mostly first or second year students enrolled in a humanities program.

We run Spearman correlation test to rank lexical features and find similar results with the p-value< 0.05 and the coefficient $|\rho| > 0.3$. The terms linked to the condition are in top ranks for Schizophrenia (*maladie [disease]*, *traitement [treatment]*, *médecin [doctor]*), while terms related to studies (*licence [bachelor]*, *thèse [PhD]*) and social life (*vacances [holidays]*, *monde [world / people]*) are correlated with controls, as shown in Table 4.9. This finding is due to the nature of our data: patients talk about their disease with a clinician, and controls talk more about their everyday life. These features perform well because they reflect a lexical bias in data collection. However, the models can not learn a lot about language specificities about this disease and they will not be usable in the wild.

Furthermore, we find that the subjects with Schizophrenia use more references to the first person, as seen with deictic words (*j' [I]*, *mon [my]+masculine object*, *ma [my]+feminin object*, *mes [my]+plural*) as well as forms of auxiliaries (*suis [am]* and *ai [have]*) while controls use more references to the second person (*tu [you]*, *es [are]*, and *as [have]*). We also evaluate the impact of these features in the models: by ignoring *je [I]* and *tu [you]* (and the elided forms *j' [I]* and *t' [you]*), we observe a slight drop in accuracy with NB (-0.49%) but a significant drop with SVM (-6.59%). These observations align with the conclusions of previous studies: Strous et al. (2009) argue that a greater use of first person deictic words and fewer references to third person subjects, accompanied by lexical repetitions, are characteristics of subjects self-centered. Other studies have also claimed that the use of the singular first person is associated with negative affective states such as depression (Rude et al., 2004; Chung and Pennebaker, 2007). Of course, this type of result should be appreciated in relation to the contextual and conversational conditions under which the data is collected.

4.5.2 POS Tags & Syntax

Sequences of POS tags (2-POS and 3-POS) and of *treelet* (2-*treelet* and 3-*treelet*) are fully non-lexicalized features. They capture some internal structure of the interaction. We obtain our best scores with the longest sequences (3-POS, 72.55% accuracy, 74.34% F_1). These scores are higher than the ones reported by Kayi et al. (2017) on tweets (69.20% F_1) or essays (69.76% F_1) with simple POS tags and a lot more documents, and are very close to Allende-Cid et al. (2019) with meta-POS (75.1% in F_1). This confirms the predictive power of POS for the task.

We find that patients with Schizophrenia use more verbs than controls: 2-POS such as VERB-ADP and 3-POS such as PRON-AUX-VERB, where ADP stands for *adposition* and it covers *preposition* and *postposition*. As in Kayi et al. (2017), we also observe a higher proportion of adverbs. Statistics of 2-*treelet* tend to indicate that individuals with Schizophrenia use more verbal groups and less nominal groups. Thus, the 2-*treelet* "VERB→aux→AUX" and "VERB→nsubj→PRON" are the most discriminatory features of individuals with Schizophrenia. We show examples for these patterns:

(38) Pattern "VERB→aux→AUX" in SCZ group

- *(j')ai fait [(i) have done]*
- *(c')est (pas) gagné [(it) is not won]*)

(39) Pattern "VERB→nsubj→PRON" in SCZ group

- *ça va [it's fine]*
- *(je) sais pas [(i) don't know]*

Further, we observe that the usage of adverbs of time (*parfois [sometimes], plus maintenant [not anymore], quasiment jamais [almost never]*), of place (*ici déjà [here already]*) and of frequency and manner (*beaucoup plus [much more], beaucoup mieux [much better]*) is higher than that of controls - this is possibly linked to the exchange about their (current) heath condition.

On the other hand, controls employ a higher portion of linking adverbs (*enfin [finally], donc [so], quand même [anyway]*). They tend to use more complicated syntactic structures, such as those with SCONJ (subordinating conjunction) and CCONJ (coordinating conjunction), confirmed by our analysis of discourse connectives. Syntactic features confirm these observations (Table 4.10), the most predictive being verbal structures, followed by adverbial modifiers such as *advmod* and *advcl*. *Advmod* is a (non-clausal) adverb or adverbial phrase; *advcl* is an adverbial clause modifier. They serve to modify a verb or other predicate. This goes along with Kayi et al. (2017), in which the top parse tag is *advmod*, and confirms clinician's descriptions (Morice and Ingram, 1982; Fraser et al., 1986) on the use of less complex syntactic structures for patients with Schizophrenia.

4.5.3 Dialogue & Discourse

We now move to dialogue and discourse feature analysis. Figure 4.2 presents results on selected subsets of non- or less- lexicalized features for the six splits (Indiv, W-$\{128, 256, 512, 1024\}$, Full) of our data. Horizontal lines correspond to the majority vote baseline in each setting.

Following Howes et al. (2012), we test OCR and *Backchannel* (BC). Concerning dialogue features, OCR gives poor results mostly behind the baseline, while BC is above with 74.48% (Full). Moreover, combining with BC to another feature set almost consistently allows improvements

treelet	SCZ	ρ	Control	ρ
1-*token*	verb	0.21	noun	−0.17
2-*token*	verb→aux→aux	0.41	pron→nsubj→pron	−0.64
	verb→nsubj→pron	0.37	cconj→nsubj→pron	−0.46
	aux→advcl→verb	0.34	propn→conj→pron	−0.46
3-*token*	pron→nsubj→verb←iobj←pron	0.51	pron→obj→verb←mark←sconj	−0.66
	aux→aux→verb←obl←pron	0.49	adp→mark→verb←det←det	−0.39
	adj→advcl→verb←nsubj←pron	0.47	verb→expl→noun→case→adp	−0.36

Table 4.10: Typical syntactic features in Schizophrenia and control groups (p-value < 0.05 for *2-tokens* and *3-tokens*).

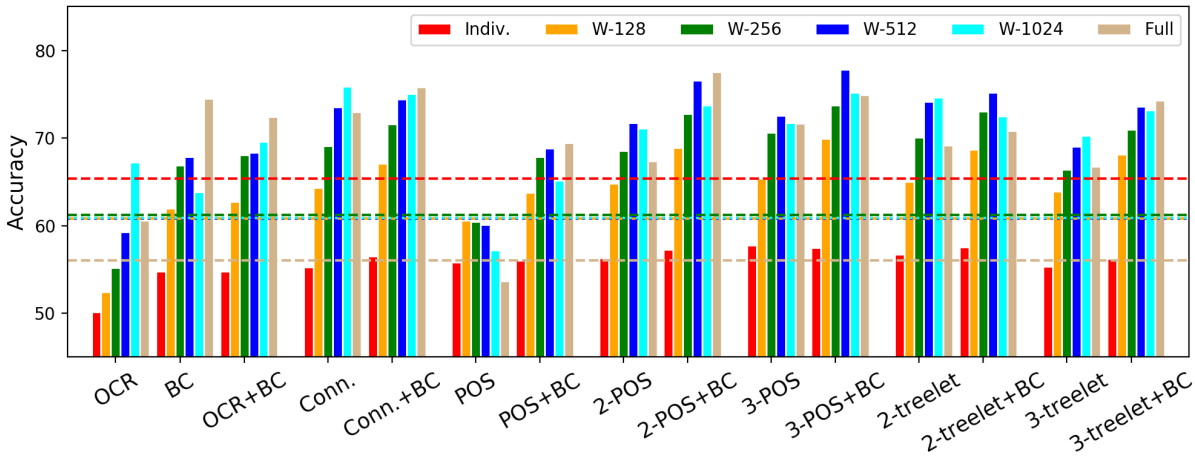


Figure 4.2: Accuracy of BC and the combination with other features in different context windows. OCR = *Open Class Repair*, BC = *Backchannel response*, Conn = connectives. Indiv., Full, and W-n are context window size.

(not the case with OCR). These features are good indicators, contrary to what was reported in Howes et al. (2012). Note that we directly use the tokens as features rather than the proportion of BC per word, which allows more refined analysis. Looking at the models weights and Spearman correlation values, we find that the most informative features for controls are phatic expressions. At the same time, patients with Schizophrenia are correlated with more ambiguous expressions which are also used in non-phatic contexts (i.e., less BC responses): this supports that the patients are less prone to maintain the conversation in Howes's paper.

(40) Informative BC employed in SCZ group

- *je comprends [I understand]*
- *bien sûr [of course]*
- *exactement [exactly]*

(41) Informative BC employed in control group

- *ah*, *ok*, *hum-hum*

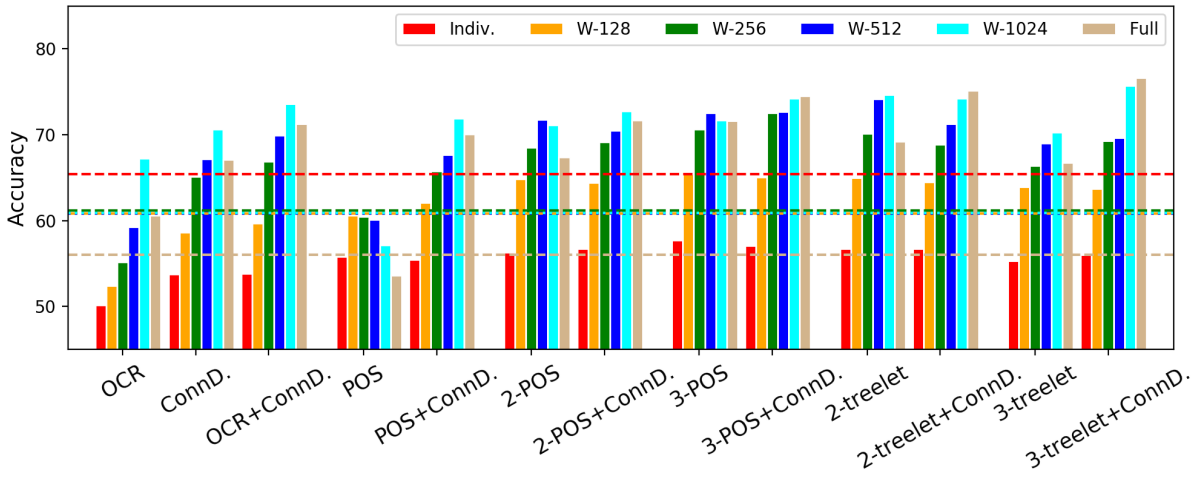


Figure 4.3: Accuracy of the combinations of connectives with syntactic features. *ConnD* = connectives disambiguated. Indiv., Full, and W-n are context window size.

- *vraiment [really]*
- *c'est ça [that's right / yeah, right]*

Connectives also give promising results, at best 76.7% (in W-1024 setting). However, the simple look-up from LexConn list (Roze et al., 2012) is a very coarse way of connective extraction. Disambiguation with the discourse parser in Laali and Kosseim (2017) results in much fewer connectives: connective types drop from 142 to 103 and total count drops from 14.5k to 6k. For unspecific connectives such as *et [and]*, the parser excludes over half of the cases. After the disambiguation process, we run experiments for the combination of connectives with other features, with results in Figure 4.3. It is clear that connectives are beneficial when combined with syntactic features, especially with 3-*treelet* (76.6%).

In terms of the context length, W-512 and W-1024 are the two best settings, with W-512 giving the best scores in the combination with OCR, POS, 2-POS, and W-1024 best with 3-POS, 2-*treelet* and 3-*treelet*. We show the correlation of connectives with Schizophrenia (positive ρ values) and control classes (negative ρ values) in Figure 4.4. Trend shows that controls use longer connectives (*jusqu'à ce que [until that]*, *au point de [to the point that]*) compared to patients (*donc [so]*, *puis [then]*). Connectives linked to the present moment (*maintenant (que) [now (that)]*, *depuis que [ever since]*) are also highly correlated to Schizophrenia group, which might refer to changes after treatment.

4.5.4 Context Window Size

Our experiments were also designed to test the impact of the context when dealing with dialogues. Figure 4.2 and 4.3 demonstrate that, in general, the larger the window, the better the scores. Individual speech turns are too small and contain no context. However, using the whole conversation most often leads to a drop in performance compared to large window sizes (W-512 and W-1024) due to the data sparsity, as we can observe for connectives, n-POS and n-*treelet*. OCR and backchannels do not follow this trend, meaning that they are probably less sparse. The best window sizes are W-512 and W-1024, respectively. They perform better than other window sizes for almost all syntactic and discourse features.

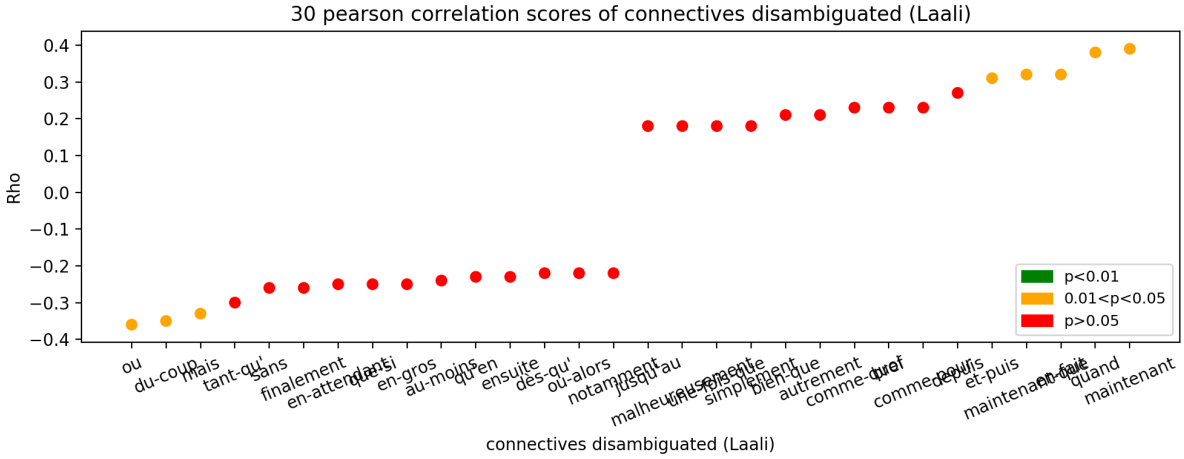


Figure 4.4: Correlation of disambiguated connectives in two groups. Positive ρ values stands for Schizophrenia class and negative for control class. The color of each point indicate its p-value: green has $p < 0.01$, yellow has $0.01 < p < 0.05$, red has $p > 0.05$. For readability, we only present the most representative connectives for both parties ($|\rho| > 0.2$). Other connectives within the gap are omitted in this plot.
Connectives and the translation (from left to right): *ou [or], du-coup [therefore], mais [but], tant-qu' [so-that], sans [without], finalement [finally], en-attendant [while-waiting], en-gros [roughly], au-moins [at least], qu'en [in], ensuite [then], dès-qu' [as soon as], ou-alors [or], notamment [in particular], jusqu'au [until], malheureusement [unfortunately], une-fois que [once that], simplement [simply], bien-que [although], autrement [otherwise], comme-quoi [as what], bref [in short], comme-pour [as for], depuis [since], et-puis [and then], maintenant-que [now that], en-fait [in fact], quand [when], maintenant [now].*

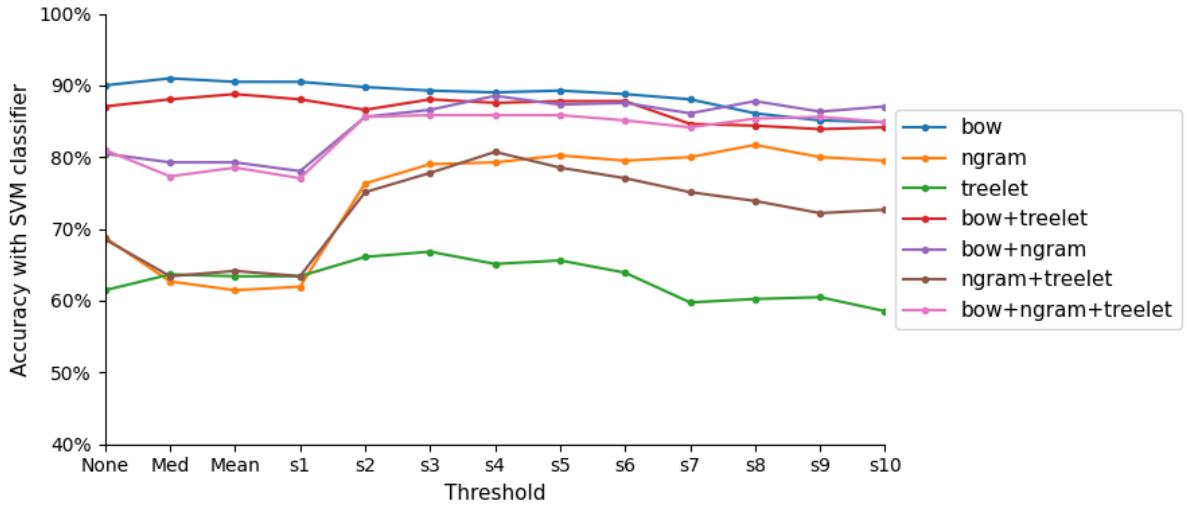


Figure 4.5: Accuracy scores in terms of different feature selection thresholds on lexical features bow, n-gram, syntactic feature *treelet*, and their combinations.

These experiments demonstrate that using the block of conversation is relevant – the models find enough information to make accurate classification –, while allowing to increase the number of classification instances artificially.

4.5.5 Influence of Feature Selection

Given the limited number of test instances and the large vocabulary, especially for lexical and syntactic features, we investigate the impact of different threshold levels on the model's performance. As seen in Figure 4.5, we evaluate the performance of BOW, n-grams ($n = 2, 3$), n-*treelet* ($n = 1, 2, 3$), and the combination thereof. However, we observe that combining these features does not lead to any improvement or may even decrease performance. The lexical features appear to be redundant, as combining BOW and *treelet* has no impact on performance when utilizing a threshold selection (red line). Without selection, the combination exceeds n-grams alone but not BOW alone. Furthermore, the n-grams representation is not suitable in this context, as it performs worse than BOW, possibly due to overlapping speech turns.

As demonstrated in Figure 4.5, the feature selection step is essential in optimizing performance, particularly for n-grams (orange line) and its combination with BOW and *treelet* (BOW+n-grams+treelet, BOW+n-grams, n-grams+treelet as seen in pink, purple and brown lines respectively). The utilization of *treelet* also improves through feature selection, with optimal scores achieved at lower threshold levels.

4.5.6 Best Algorithm

Among the 5 classifiers, NB generally performs well when dealing with word counts (in Full and Indiv.), while SVM and LR are generally better in other cases.

In Table 4.11, we show the best algorithm for non-lexical single (apart from BOW and `ngram`) and combined features in all context settings. More precisely, SVM performs better when the context window is relatively large, and the data sparsity is more pronounced (Full, W-1024). At the same time, logistic regression (LR) is better at dealing with small to medium-sized contexts

Setting	Best single feature			Best comb. features		
	Feature	Accuracy	Classifier	Feature	Accuracy	Classifier
Full	BC	74.48	SVM	2-POS+BC	77.54	SVM
Indiv.	3-POS	56.53	LR	1-2-3-POS	58.36	LR
W-128	3-POS	65.39	SVM	3-POS+BC	69.92	LR
W-256	3-POS	70.66	LR	3-POS+BC	73.75	LR
W-512	3-POS	72.55	SVM	3-POS+BC	77.86	LR
W-1024	3-POS	71.71	SVM	3-*treelet*+*ConnD*	75.74	SVM

Table 4.11: Best algorithm for the single (except for BOW) and combined features in different window settings.

(Indiv., W-128, W-256 settings). In general, random forest (RF) and Perceptron are performing poorly compared to SVM and LR. In all window settings, they show a much lower accuracy with syntactic features – while they are the most important features in the best systems (2-POS+BC in Full and 3-POS+BC in W-512). However, RF shows better performances with discourse markers such as *backchannel* feature alone, with detailed information in Appendix A.2.

4.6 Conclusion

In this chapter, we explore language specificities associated with mental impairment in French. To do so, we use conversations in SLAM project which involves patients with Schizophrenia in order to learn about language features associated with the disease. We test various representations, including lexical, syntactic, discourse, and dialogue level information. To deal with data scarcity issue, we compare different context length settings to represent dialogues: from non-context (Indiv.), a certain length of context (W-n), and whole dialogue (Full); and further employ feature selection techniques to reduce vocabulary size. Our best system uses only lexical information and achieves an accuracy of 93.66% with the NB classifier. Our experiments replicate performances as high as previous studies in English. However, the analysis of our data and model highlights possible lexical biases in our corpus, especially because the control group's vocabulary is centered on academic studies, and patients are used to describing their medical surroundings. This could make our models less resilient. We suspect that other corpora employed in dialogue settings for this task might be similarly biased. Therefore, we suggest that exploring alternative types of information is crucial for creating a robust model that provides insights into language specifics instead of being limited by the data collection process.

Our results show that non-lexical features such as syntactic tags (POS and *treelet*) and discourse level markers (*backchannel*, connectives) are good indicators. The combination of BC with syntactic features yields the best results. Further analysis of these features shows interesting findings: patients in the Schizophrenia group tend to use more ambiguous expressions (less BC responses) than the control group. They also use more verbs and more superficial syntactic structures. On the other hand, the controls express more apparent acknowledgment responses during the conversation, and they employ a higher portion of linking adverbs that loosely indicate more coherent speeches.

In this study, we do not to employ a neural architecture due to two reasons: first of all, our dataset is limited, and a neural model may be overly complex relative to the amount of training data available, leading to poor performance on unseen data and may only learn biased lexicons

instead of the real language specificity. Therefore, a lexical-free model is preferred to enhance generalizability. Secondly, we aim to investigate the relationship between linguistic features and mental impairment by analyzing the feature importance at various linguistic levels. This allows us to provide clear explanations of the underlying correlations. In contrast, neural models are more challenging to interpret.

One of the limitations of this study is the lack of interaction between the patients and psychologists. Unsurprisingly, psychologists seem to actively adapt their way of speaking when facing different interlocutors. Thus, directly adding the linguistic features of psychologists may add more bias to the system. The interaction of speeches should be designed in a more sophisticated way. We will provide interaction modeling in the next chapter.

For NLP practitioners, we hope that this study will remind us of the importance of looking for bias in data and exploring higher-level information (i.e., less language dependent), such as discourse information, to produce robust systems and draw more general conclusions. For research on Schizophrenia using NLP methods, we manage to replicate results on another language and modality, thus confirming that these are features specific to the disease. We plan to continue the study by exploring other datasets and more sophisticated features, with a new reflection on the bias for studies on other conditions.

Chapter 5

Multi-Task Learning for Depression Detection in Dialogues

Contents

In the previous chapter, we have discussed using linguistic features at various levels for Schizophrenia language detection in dialogue settings. One of the limitations we recognize is the lack of speech interaction engineering. In this chapter, we aim to overcome this limitation by proposing a hierarchical neural architecture within a multi-task learning framework. Our focus is on identifying depressive subjects during guided conversations. As we encounter difficulties in data size and in confidentiality issues in the previous study, we switch to a larger and publicly available dataset DAIC-WOZ for depression detection. This allows us to test more sophisticated models and compare them with related works.

Depression is a serious mental disorder that affects around 5% of adults worldwide[1]. It comes with multiple causes and symptoms, leading to major disability, but is often hard to diagnose, with about half the cases not detected by primary care physicians (Cepoiu et al., 2008).

[1] `https://www.who.int/news-room/fact-sheets/detail/depression`

Automated detection of depression, sometimes associated to other mental health disorders, has been the topic of several studies recently, with a particular focus on social media data and online forums (Coppersmith et al., 2015; Benton et al., 2017; Guntuku et al., 2017; Yates et al., 2017; Song et al., 2018; Akhtar et al., 2019; Ríssola et al., 2021). The ultimate goal of such system is to complement expert assessments, but such empirical studies are also valuable to better understand how communication is affected by health disorders. In this chapter, we propose to investigate depression detection within dialogues, a scenario less studied but more similar to the interviews with clinicians, which allows to examine the interactions in conversation.

Like Schizophrenia and other cognitive impairment detection tasks, depression detection also faces the challenge of data scarcity. As we see, using social media data is a way to tackle this issue, including considering data generated by self-diagnosed users – a method that leads to potentially noisy data and comes with ethical (Chancellor et al., 2019) and privacy (GDPR regulations) issues. Since our focus is on analyzing dialogue structures, we choose to study a dataset of 189 clinical interviews called the DAIC-WOZ (Gratch et al., 2014). This dataset is specifically collected by experts to aid in the diagnosis of distress conditions. It includes identification of whether participants are depressive or not, as well as a severity score for those who are. This dataset is in English, and it has been extensively studied with various modalities, such as audio (Al Hanai et al., 2018; Williamson et al., 2016), visual (Haque et al., 2018) and textual (Haque et al., 2018; Dinkel et al., 2019; Mallol-Ragolta et al., 2019).

In our work, we draw inspiration from previous studies (Qureshi et al., 2019, 2020) and propose to adopt the Multi-Task Learning (MTL) framework in order to enable our model to leverage information from multiple sources. We believe that MTL can be highly beneficial for our model, especially given the limited size of our dataset. We incorporate three auxiliary tasks, including *Emotion Classification*, which is naturally tied to mental health states, *Dialogue Act Identification*, which serves as an indicator of local coherence in a dialogue, and *Topic Classification*, which provides an indication of the global information in a dialogue. In the previous chapter, we have explored several discourse features such as *backchannel responses* (BC), *open class repair* (OCR), and discourse connectives. These features are considered *shallow* because they remain on the surface level and reflect less structural information. In this chapter, we use dialogue acts and topics as *shallow* information to refer to the discourse structure in dialogues. By considering these shallow structures, we hope to gain insights into the overall organization of the dialogue and how it relates to the mental state of the participants. It is important to note that, considering the limited performance of existing discourse parsers, we choose not to consider parsed structures into MTL. Nonetheless, this presents an intriguing aspect for future exploration.

Our neural network architecture is a traditional one, employing the hard-parameter sharing technique (Ruder, 2017) and is less complex than the shared-private architecture proposed in Qureshi et al. (2020). Despite its simplicity, our model demonstrates remarkable efficacy (Section 5.5). Recognizing the importance of dialogue organization, we propose a hierarchical architecture that is tailored specifically for dialogue processing. This architecture includes tasks performed at both the speech turn level and the dialogue level. By using this approach, we aim to capture the structure and flow of a conversation more accurately. We believe that this will improve the model's ability to detect patterns and relationships between different aspects of the conversation.

This chapter is adapted from one publication at the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGdial 2022) (Li et al., 2022). The structure of this chapter is outlined as follows: Firstly, in Section 5.1, we provide an overview of previous studies in the field, including the use of the MTL framework in health-related prediction tasks and its

effectiveness in dealing with data scarcity issues. We examine several works that have employed MTL in precise depression detection using the DAIC-WOZ dataset (Gratch et al., 2014) and identify potential areas for improvement. Next, we introduce a few classic multi-task learning structures and our proposal in Section 5.2, the latter includes a baseline model and an MTL model[2]. In Section 5.3, we introduce two types of dialogue corpora, which we utilize to achieve our objective of integrating mental illness detection in dialogues with auxiliary tasks. The first type of corpus consists of conversations with patients with cognitive impairment, among which we select DAIC-WOZ for our main task of depression detection. The second corpus pertains to the subtask of *emotion recognition* in dialogues. Experimental setup is presented in Section 5.4. Following this, we present the results and analysis of our experiments in Section 5.5. Finally, we conclude our study in Section 5.6.

5.1 Related work

5.1.1 Multi-Task Learning on Health-Related Prediction Task

Within multi-task learning, a model has to learn shared representations to generalize the target task better. It improves performance over single-task learning (STL) by leveraging commonalities or correlations between tasks. Recent years have witnessed a series of successful applications in various NLP tasks, such as Part-Of-Speech (POS) tagging, syntactic chunking[3], Named Entity Recognition (NER), Semantic Role Labeling (SRL), etc., as in Collobert and Weston (2008); Søgaard and Goldberg (2016); Ruder (2017); Ruder et al. (2019), which demonstrate the effectiveness of MTL in learning information from different but related sources. It also tackles the data scarcity issue and reduces the risk of overfitting (Mishra et al., 2017; Benton et al., 2017; Bingel and Søgaard, 2017).

Joshi et al. (2019) demonstrated the benefit of MTL for specific pairs of close health prediction tasks on tweets. In this research, the authors explored the advantages of employing MTL in three specific health informatics pairs, namely (1) symptoms that overlap for the same classification, such as classifying influenza and several other symptoms like cold, fever, and diarrhea, (2) medical concepts that overlap, such as vaccination behavior and drug usage, and (3) related classification problems, like detecting vaccination intention and vaccination relevance. The authors claim that since the symptoms overlap, these tasks are related, making them suitable for the MTL framework. Their model consists of an embedding layer, a shared representation layer, such as bi-LSTM, convolutional, or bi-LSTM+convolutional, a dropout layer, and 2 dense layers for task outputs, and it is fully shared. The corpus they used comprises 5 Twitter datasets that track medical information, including flu, drug usage, vaccines, and others, consisting of approximately $40k$ tweets. In comparison to single-task learning, the results showed that the shared bi-LSTM layer and bi-LSTM+convolutional shared layer aided the three tasks. However, this improvement was not observed when the convolutional layer was used as a shared representation. The enhancement was around $2-4\%$ for all pairs wherever applicable. The authors observed that the benefits of MLT depend on the type of shared layers and how related the tasks were.

In another study, Benton et al. (2017) utilized MTL on social media data to improve the prediction of various mental health signals, including *neuroatypicality* (atypical mental health), *suicide attempts*, *anxiety*, *depression*, *eating disorders*, *panic attacks*, *schizophrenia*, *bipolar disorder*, and *post-traumatic stress disorder* (PTSD). The authors employed a fully-shared layer for

[2]Our code is available at `https://github.com/chuyuanli/MTL4Depr`.

[3]Chunking is also known as shallow syntactic parsing, one word receives one syntactic tag such as a begin-chunk (e.g. B-NP) or inside-chunk tag (e.g. I-NP).

all tasks and an additional per-task hidden layer. They trained the first hidden layer jointly for 5,000 iterations and then trained the second hidden layer for another 1,000 iterations. Gender prediction was also included as an auxiliary task. Their corpus consisted of multiple Twitter datasets, with 9,611 users and in average 3,521 tweets each, totaling over 33.8 million tweets. AUC was used as the main metric. The results showed improved predictions for all mental health conditions except *schizophrenia* – the only case where STL model outperformed MTL. Notably, *anxiety*, PTSD, and *bipolar disorder* showed pronounced gains in detection. While adding gender as an auxiliary task led to more predictive models, the difference was not statistically significant for most tasks. Interestingly, in small datasets, modeling the common mental health conditions with the most data (in their case, *depression* and *anxiety*) helped in detecting rare conditions such as *bipolar disorder* and PTSD. Like the findings in Joshi et al. (2019), the authors also verified the significance of selecting a suitable set of related tasks. However, unlike in Joshi's work, they did not evaluate the effect of different shared layer types. Their shared layer is a multilayer perceptron.

5.1.2 Multi-Task Learning on Depression Detection

With a focus on depression detection, the shared task AVEC 2016 (Valstar et al., 2016) and AVEC 2017 (Ringeval et al., 2017) have brought out a series of multi-modal studies using vocal and visual features on the DAIC-WOZ dataset (Gratch et al., 2014). Some of which also explored text-level features: Williamson et al. (2016) used regression model with *semantic content* features such as *question answer pairs* and reported a SOTA score on the validation set (F score at 0.76). Al Hanai et al. (2018) and Haque et al. (2018) learned sentence embeddings with an LSTM network. However, their results on textual features are lower than SOTA by a large margin. In their study, Dinkel et al. (2019) evaluated and compared different techniques for text embedding. These included word-level methods such as Word2Vec (Mikolov et al., 2013) and FastText (Bojanowski et al., 2017), as well as sentence-level methods such as BERT (Devlin et al., 2019a) and ELMo (Peters et al., 2018a). To represent a text, they used the average of word-level embeddings for Word2Vec and FastText, the penultimate layer embedding for BERT, as well as the average of all three layer embeddings for ELMo. Their model consisted of a 3-layer bidirectional gated recurrent unit (GRU) followed by a linear transformation layer. They also experimented with different pooling strategies and determined that the most effective approach was to use mean pooling with ELMo embeddings.

We compile a table of the state-of-the-art studies that use textual modality information, as shown in Table 5.1. Each study is categorized according to its "text embedding", "model structure" (model architecture and whether they are trained in a multi-task framework), and "highlight" properties, such as the use of attention mechanisms. Most studies focus on binary classification, predicting whether the subject is depressed or not. While Qureshi et al. (2019) perform a multi-class classification task on the severity of depression levels (middle section of the table). Since the gold labels for test set are not published until recently, most of the work evaluates their models on the development set (top and middle sections of the table). Only Mallol-Ragolta et al. (2019) and Xezonaki et al. (2020) present results on the test set (bottom section of the table). These are the two primary works that we compare our systems with. Note that the test set in Mallol-Ragolta et al. (2019) contains only 9 depressed documents, while the official set contains 14. It is not clear yet how their partition differs from the official one. It is worth noting that in Williamson et al. (2016), the semantic models achieved significantly better performance compared to other models. Upon closer examination, we discover that their *semantic content* features not only utilize lexical features such as *question answer pairs*, but also

non-verbal cues such as [laughter], [sigh], and [sniffle], which are not included in other studies. Additionally, their *semantic context* features use rule-based queries to calculate indicator points. For example: if the patient responds with the keyword "suicid", then the context feature +1 point. These queries are tailored to the specific corpus and rely solely on extracting keywords, making it unlikely for this model to be effectively generalized to other contexts.

As depicted in Table 5.1, different works utilize various modeling strategies. While some work focuses on data representation comparison, such as Dinkel et al. (2019), where authors test various embeddings, others emphasize architecture modeling, as in Xezonaki et al. (2020); Mallol-Ragolta et al. (2019). When it comes to text classification, a crucial aspect to consider is the *word-to-document* transformation, which involves representing a text in a vector space for classification purposes. The table demonstrates that two primary transformations have been proposed. The first approach involves using embeddings and an average strategy to aggregate word-level embeddings to sentence- or document-level, as in Al Hanai et al. (2018); Haque et al. (2018); Dinkel et al. (2019). Some works, however, directly use powerful sentence-level embeddings such as BERT and ELMo to exploit contextual information encoded in these models. The second approach employs the convolutional network such as GRU and LSTM, as in Mallol-Ragolta et al. (2019); Xezonaki et al. (2020). In these models, word- and sentence-level information is integrated to finally aggregate onto the document level. Our proposed method, presented in this chapter, aligns with the second strategy, where we focus on **modeling the dialogue structure** and place less emphasis on effective text embedding strategies.

In the middle part of Table 5.1, we present work by Qureshi et al. (2019, 2020) who employ similar strategy as ours, i.e., multi-task learning with emotion prediction as auxiliary task. A resume of their work in presented in Table 5.2. We explain the different MTL schemes (fully-shared, shared-private, adversarial shared-private) in Section 5.2.1. Precisely, the authors add emotion intensity and depression severity (DLR, a regression problem) prediction to the main depression classification (DLC) task. However, they find that the emotion-unaware model achieves the best results for the DLC task, with an accuracy of 66.7% on the development set. It should be noted that they use a monologue corpus CMU-MOSEI for the emotion task, which may introduce domain bias to harm the performance. For the depression regression (DLR) and emotion intensity regression tasks, the best results are obtained with the emotion-aware model. They are also the first to conduct a thorough class-wise analysis of depression severity and show that a multi-task model can be beneficial for some classes (such as *moderate*) while failing for others (such as *mild*). In conclusion, due to mixed results, the authors state that no definitive conclusions can be drawn regarding whether emotion-aware MTL helps with depression classification/regression. On the contrary, we hypothesize that emotional information would benefit depression detection.

At the bottom part of Table 5.1, we resume two works that we can directly compare: Mallol-Ragolta et al. (2019) use a hierarchical contextual attention network with static word embeddings within a single-task setting and then combined representations at the word and sentence levels. They report at best 63% in F_1. Recently, Xezonaki et al. (2020) present even better results, 70% in F_1, by augmenting the attention network with a conditioning mechanism based on external lexicons, including LIWC (Pennebaker et al., 2001), RC Emotion Lexicon (Emolex) (Mohammad and Turney, 2013), Twitter sentiment lexicon (Kiritchenko et al., 2014), and Opinion lexicon (Wilson et al., 2005a). They also incorporate the summary associated with each interview. We instead rely on MTL in our work, where incorporating external sources is more direct.

Finally, another work that is loosely related to ours is Cerisara et al. (2018). In this study, the authors examine MTL using *sentiment*[4] and *dialogue act* prediction on a Mastodon corpus,

[4]Sentiment and emotion are closely related but with different functions and/or granularity, as discussed in

Model	Embedding	Structure	Highlight	Performance			
				F_1	Prec.	Rec.	Acc.
Dev set, binary classification							
Williamson et al. (2016)	GloVe	SVM	+semantic content	0.76	-	-	-
	GloVe	SVM	+semantic context	**0.81**	-	-	-
Al Hanai et al. (2018)	Doc2Vec[¶]	logistic regression	-	0.59	0.71	0.50	-
	Doc2Vec[¶]	LSTM	response sequences	0.67	0.57	**0.80**	-
Haque et al. (2018)	W2V[‡]	single vector+linear	-	-	-	0.65	-
	D2V[§]	single vector+linear	-	-	-	0.68	-
Dinkel et al. (2019)	BERT	MT GRU	+severity regression	0.55	0.66	0.56	0.66
	FastText	MT GRU	+severity regression	0.60	0.59	0.62	0.68
	W2V[‡]	MT GRU	+severity regression	0.61	0.61	0.64	0.70
	ELMO	MT GRU	+severity regression	0.64	**0.73**	0.66	**0.72**
Mallol-Ragolta et al. (2019)	GloVe	NHN[**]	-	0.40	-	0.50	-
	GloVe	HLGAN[*]	+local-global attention	0.60	-	0.60	-
	GloVe	HCAN[‖]	+attention	0.51	-	0.54	-
Xezonaki et al. (2020)	GloVe	HAN[††]	+attention	0.46	-	0.48	-
	GloVe	HAN[††]+L	+attention, lexicon	0.62	-	0.63	-
Dev set, multi-class classification							
Qureshi et al. (2019)	USE[†]	LSTM	-	0.45	-	-	0.60
Qureshi et al. (2020)	USE[†]	MT LSTM	+emo prediction	0.51	-	-	0.61
Test set, binary classification							
Mallol-Ragolta et al. (2019)	GloVe	NHN[**]		0.45	-	0.50	-
	GloVe	HLGAN[*]	+local-global attention	0.35	-	0.33	-
	GloVe	HCAN[‖]	+attention	0.63	-	0.66	-
Xezonaki et al. (2020)	GloVe	HAN[††]	+attention	0.62	-	0.63	-
	GloVe	HAN[††]+L	+attention, lexicon	**0.70**	-	**0.70**	-

Table 5.1: Comparison of different models' performance on DAIC-WOZ development and test sets. Best scores in development and test sets are in **bold**.
Embedding column: Doc2Vec[¶]: self-trained embedding using Python Gensim library. W2V[‡]: pre-trained Word2Vec (Mikolov et al., 2013). D2V[§]: pre-trained Paragraph Vector (Le and Mikolov, 2014). USE[†]: pre-trained Universal Sentence Encoder (Cer et al., 2019).
Structure column: HLGAN[*]: Hierarchical Local-Global Attention Network. NHN[**]: Naive Hierarchical Network. HCAN[‖]: Hierarchical Contextual Attention Network. MT: multi-task learning framework. HAN[††]: Hierarchical Attention-based Network. "+L": add lexical features from six resources, refer to Xezonaki et al. (2020) paper for details.
Highlight column: "+semantic content": *question answer pairs* and non-verbal cues such as laughter and sigh markers. "+semantic context": coarse contextual indicators such as previous diagnoses and ongoing therapy, use rule-based queries to accumulate points, not directly utilize speech. "+attention": + attention mechanism. "+severity regression": auxiliary task with depression severity prediction. "+emo": + auxiliary task with emotion prediction. "-": not reported.

Model	Architecture	DLC		DLR		EIR
		Acc %	F_1	RMSE	MAE	MSE
Qureshi et al. (2019)	ST (DLC)	60.6	0.54*	-	-	-
	ST (DLR)	-	-	4.90	3.99	-
	MT (DLC+DLR), fully-shared	**66.7**	**0.62***	4.96	3.89	-
	MT (DLC+DLR), shared-private	60.6	0.42	4.70	3.81	-
Qureshi et al. (2020)	MT (all), fully-shared	57.57	0.46	4.83	4.03	**6.96**
	MT (all), shared-private	63.64	0.58	**4.56**	**3.79**	7.02
	MT (all), adversarial shared-private	60.61	0.60	4.61	3.69	7.11

Table 5.2: Results from work Qureshi et al. (2019, 2020) on multi-task learning on depression classification (DLC), depression regression (DLR), and emotion intensity regression (EIR). DLC and DLR use DAIC-WOZ dataset, EIR use CMU-MOSEI dataset (Zadeh et al., 2018). ST: single task, MT: multi-task. MT (all): multi-task learning with all three tasks (DLC+DLR+EIR). RMSE: root mean square error; MAE: mean average error; MSE: global metric that averages the squared errors. In **bold**: the best score for each column. '-': not applicable. *: results extracted from Qureshi et al. (2020), scores reported in Qureshi et al. (2019) for ST and fully-shared MT are 0.45 and 0.53, respectively.

a social networking platform with microblogging features similar to Twitter, where both annotations are available. They discover a positive correlation between these two tasks. Although this work does not address depression detection, it provides evidence for the relevance of using dialogue act and sentiment prediction, which are tasks that we believe are pertinent to our primary depression task in dialogue settings.

Upon a thorough review of related studies, it is evident that none of them have explored the potential connection between depression and dialogue structure. Therefore, we believe that our work is the first to address the detection of depression in dialogue transcriptions using the MTL approach and incorporating tasks related to the structure of the conversation.

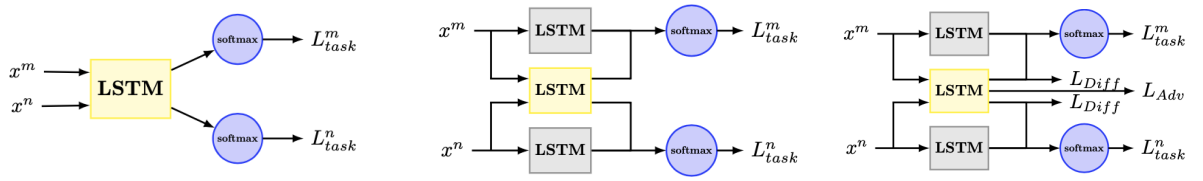
5.2 Model Architecture

5.2.1 Multi-Task Learning Schemes

The objective of MTL is to learn the common and task-invariant features by training shared layers. There are several MTL architecture designs available, with two common sharing schemes being the **fully-shared** and **shared-private** schemes (Caruana, 1993, 1997). Ruder (2017) refers to these schemes as *hard parameter sharing* and *soft parameter sharing*, respectively. In the fully-shared scheme (short in FS or FS-MTL), the hidden layers are shared across all tasks, while only task-specific output layers are maintained, as shown in Figure 5.1a. x^m and x^n are input representations of task m and n. The learned shared representation for the task m is formulated as follows:

$$s_t^m = \text{LSTM}(x_t^m, s_{t-1}^m, \theta_s) \tag{5.1}$$

Munezero et al. (2014). Cerisara et al. (2018) used three labels for sentiment: *positive*, *negative*, *neutral*. In this study, we use seven emotional classes: *anger*, *disgust*, *fear*, *happiness*, *sadness*, *surprise*, *neutral*.



(a) Fully-Shared Model (b) Shared-Private Model (c) Adversarial Shared-Private

Figure 5.1: Three Multi-Task Learning architectures, from Liu et al. (2017). m and n are different tasks.

where s_{t-1} is the previous hidden state and θ represents all the parameters in the network LSTM. Conversely, in the shared-private scheme (short in SP or SP-MTL), each task has its own set of parameters (gray LSTM blocks in Figure 5.1b). The tasks share a network that encourages the parameters to be similar (yellow LSTM block). Formally, for any task m, it has a shared representation s^m (equation 5.1) as well as a task-specific representation h^m (equation 5.2). The final features are concatenated from both representations.

$$\begin{aligned} s^m_t &= \text{LSTM}(x^m_t, s^m_{t-1}, \theta_s) \\ h^m_t &= \text{LSTM}(x^m_t, h^m_{t-1}, \theta_h) \end{aligned} \tag{5.2}$$

The FS model may ignore the fact that some features are task-dependent. As for SP scheme, there is no guarantee that sharable features are trained in the shared space and task-specific features in the private spaces. To address the limitations of previous models, Liu et al. (2017) suggested a new sharing scheme that incorporates adversarial training known as **adversarial shared-private** (ASP), as shown in Figure 5.1c. Final feature for a task m is still the combination of its shared and private representations, while the training process is enhanced with two new losses. One loss is an additional task adversarial loss L_{Adv} which prevents task-specific features from intruding into the shared space. This is a min-max optimization, and we leave the precise formulation to the readers. To make the features in shared and private spaces more differentiable, an orthogonality constraint is added L_{diff}. Finally, the loss function of an adversarial shared-private model is the sum of all three losses, with hyper-parameters λ and γ.

$$L = L_{Task} + \lambda L_{Adv} + \gamma L_{Diff} \tag{5.3}$$

A variant of the ASP model is *Cross-Stitch Network*, proposed by Misra et al. (2016). They started with a shared-private network and introduced cross-stitch units between the private networks. These units enable the model to determine the way one network utilizes the knowledge of another by learning a linear combination. Qureshi et al. (2020) utilized the ASP network. However, they failed to exhibit any improvement in performance compared to non-adversarial models, as shown in Table 5.2, which could be attributed to the insufficient training examples for a complex model.

Over time, hierarchical MTL networks have been developed. For example, a *Fully-Adaptive Feature Sharing Network* is proposed by Lu et al. (2017), in which a network grows like a tree, with different sub-network parameters dedicated to different tasks, and similar tasks are grouped under the same branch. However, their greedy algorithm for tree growing sometimes results in one task per branch, leading to a model that fails to learn shared parameters. Another hierarchical

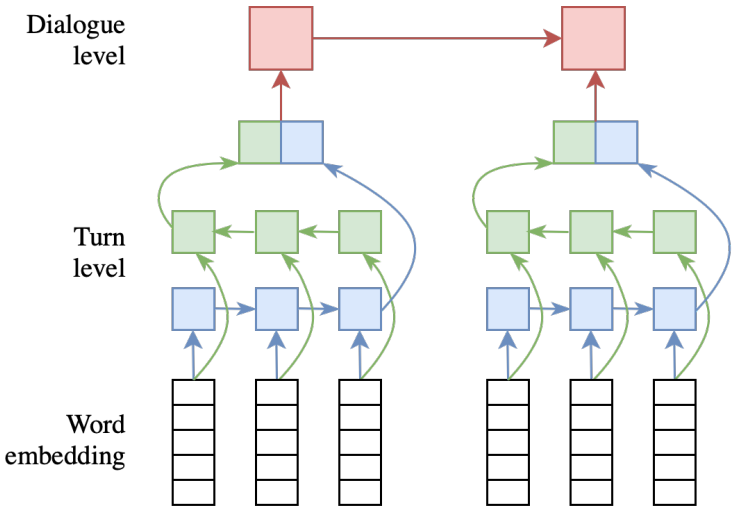


Figure 5.2: Baseline two-level recurrent network. *Turn*-level utilizes bi-LSTM network and *document*-level utilizes a RNN network. Adapted from Cerisara et al. (2018).

model proposed by Ruder et al. (2017) is the *Sluice Network*, which is a generalization of the fully-shared model, cross-stitch network, and hierarchical models. Their model determines where in the sub-network and at what layer the representation of a task is best learned.

After the introduction of different MTL schemes, we see that fully-shared and shared-private models are the basic models of all subsequent variations. The fully-shared scheme has the innate disadvantage of ignoring task-specific information, but its shared architecture also reduces the risk of overfitting, creating a model that is easier to generalize (Ruder, 2017; Baxter, 1997). Other models, with more complex architectures and more training parameters, may not be suitable for small corpora like ours. Therefore, we start with the basic fully-shared scheme in this work.

5.2.2 Our Models

One condition generally assumed for the success within MTL is that the primary and auxiliary tasks should be related (Ruder, 2017). The emotion-related task is thus a natural choice since it is linked to mental states. We hypothesize that depressive disorder can also affect how people interact with others during conversations. We thus take a first step toward linking dialogue structure and depression by examining shallow signals: dialogue acts and topics. In addition, since the information comes at different levels, we propose hierarchical modeling, from speech turns to documents.

Baseline Model: Our basic model is a two-level recurrent network, similar to the one in Cerisara et al. (2018), as shown on the left in Figure 5.2. The input words are mapped to vectors using word embeddings from scratch. The first level (*turn*-level) takes the embeddings into a bi-LSTM network to obtain one vector for each turn. The second level (*dialogue*-level) takes a sequence of turns into a RNN network, and the output is finally passed into a linear layer for depression prediction.

MTL Model: As outlined in Section 5.2.1, we advocate for the simple structure, namely the *fully-shared* structure in our experiments. Our MTL architecture comprises shared hidden layers and task-specific output layers (Figure 5.3) and aligns with the *hard parameter sharing* approach

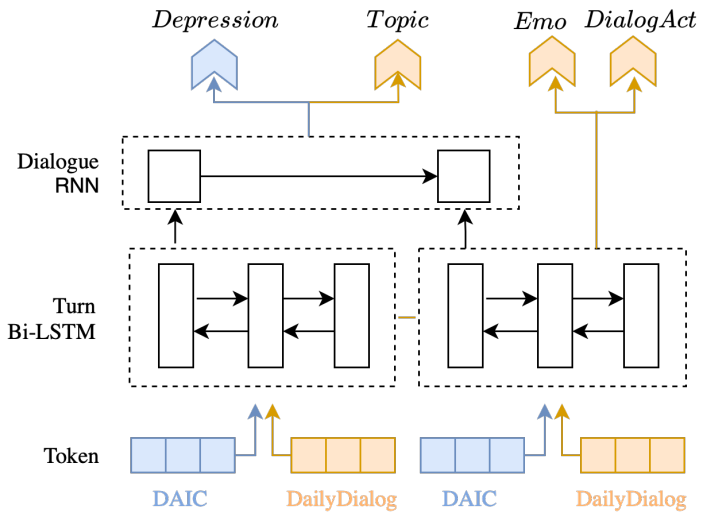


Figure 5.3: Multi-task fully shared hierarchical structure. Flow in light blue stands for depression prediction task (DAIC dataset); flow in orange represents auxiliary tasks: emotion, dialogue act, and topic prediction (Dailydialog dataset). The information flow inside RNN and LSTM networks are simplified for better readability.

(Caruana, 1993, 1997; Ruder, 2017). Since some auxiliary tasks are annotated at the speech-turn level (i.e., emotion, dialogue act) while others are at the document level (i.e., depression, topic), our architecture is hierarchical and organizes task-specific output layers (MLP) at two levels.

The training process operates as follows: when learning emotion prediction and depression prediction, as these two tasks are not at the same level, we train different levels of networks. When processing emotion-annotated utterances, we tune the *turn*-level LSTM network. When processing depression-annotated utterances, both the *turn*-level LSTM and the *document*-level GRU networks are trained. Overall, sentence level information (emotion, dialogue act) can be learned in the *turn*-level LSTM network and transferred upwards to help depression and topic prediction. Conversely, higher-level information can be backpropagated to update the network at the lower level. The loss is simply the sum of the losses for each task (Equation 5.4). Regarding the MTL setting, we set equal weight for each task as the standard choice.

$$L = L_{Depr} + L_{Emo} + L_{DA} + L_{Topic} \tag{5.4}$$

5.3 Datasets

In the previous chapter, our study is limited by the size and bias of the corpus. In this study, we are taking a more cautious approach to corpus selection. We aim to use a relatively large dialogue dataset in the field of cognitive impairment, allowing us to evaluate the effectiveness of our proposed hierarchical structure for dialogue modeling. As French corpora are limited in number, we turn to English corpora and expand our scope to include various mental illnesses. In Section 5.3.1, we introduce the DAIC-WOZ corpus (DeVault et al., 2014) and another candidate corpus, the Carolinas Conversation Collection (CCC), which focuses on Alzheimer's disease. Although we do not use CCC corpus in this study, it may be of interest to test in future work.

Ellie Who's someone that's been a positive influence in your life?
Part Uh my father.
Ellie Can you tell me about it?
Part Yeah, he is uh
Part He's a very he's a man of few words
Part And uh he's very calm
Part Slow to anger
Part And um very warm very loving man

Figure 5.4: Left: the setting of Wizard-of-Oz interview where a participant talks to Ellie, the virtual interview who is actually controlled by two humans. Right: An excerpt from a WoZ interview (Gratch et al., 2014). "Part": participant.

Additionally, we conduct research on auxiliary tasks that could benefit dialogue modeling, including emotion recognition in conversation (ERC), machine reading comprehension (MRC), and dialogue act classification. We discover that emotion recognition is a suitable task due to its task relevance and the growing availability of publicly available corpora (Busso et al., 2008; McKeown et al., 2011; Li et al., 2017). In Section 5.3.2, we present the chosen corpus, the DailyDialog corpus (Li et al., 2017), and extend to other candidate ERC corpora for interested readers in Section 5.3.3.

5.3.1 Mental Illness Dialogue Corpora

DAIC-WOZ: It is a subset of the DAIC corpus Gratch et al. (2014) which contains 189 sessions (one session is one dialogue, in average 250 speech turns) of two-party interviews. All the conversations are publicly available[5]. The experiment involves a human participant engaging in a conversation with a computer-generated interviewer named Ellie (as depicted in Figure 5.4), whose non-verbal behavior is controlled by one wizard and verbal responses by another. The interview follows a semi-structured format, with the initial questions being designed to establish rapport and gradually moving towards specific questions about symptoms and events related to depression. Ellie's responses are predetermined and pre-recorded. The use of Ellie as the interviewer ensures a relatively neutral lexical bias in the conversation, making it feasible to incorporate her utterances into dialogue modeling. This feature brings the major difference from the approach employed in the previous chapter. On the right side of Figure 5.4, we show an interview excerpt presented in Gratch et al. (2014).

Table 5.3 gives the partition of train (107), development (35), and test (47) sets. We show the PHQ-9 scores and binary partition. Originally, patients are associated with a score related to the Patient Health Questionnaire (PHQ-9): a patient is considered depressive if PHQ-9 ≥ 10 (Kroenke and Spitzer, 2002). For binary classification, labels 0 and 1 represent none-minimal depression and depression presented, respectively. For the multi-class classification, we denote $[0-4]$ (label 0): none-minimal, $[5-9]$ (label 1): mild, $[10-14]$ (label 2): moderate, $[15-19]$ (label 3): moderately severe, and $[20-27]$ (label 4): severe. Note that in Qureshi et al. (2020), the

[5] `https://dcapswoz.ict.usc.edu`

	Total	PHQ-9 binary		PHQ-9 multi-class				
		$0\|[0-9]$	$1\|[10-27]$	$0\|[0-4]$	$1\|[5-9]$	$2\|[10-14]$	$3\|[15-19]$	$4\|[20-27]$
Train	107	77[6]	30	47	29	20	7	4
Dev.	35	23	12	17	6	5	6	1
Test	47	33	14	22	11	5	7	2
Total	189	133	56	86	46	30	20	7

Table 5.3: DAIC-WOZ dataset binary and multi-class partitions.

authors conducted multi-class classification on train and development sets without precising the partition. They used in total 138 documents for experiments. It is unknown which 5 documents were missing. For our experiments, we follow the original splits and utilize all 189 documents.

Carolinas Conversation Collection (CCC): Additionally, we discover the Carolinas Conversation Collection (Pope and Davis, 2011), which contains conversations with patients suffering from Alzheimer's Disease (AD) as well as elderly individuals with different medical conditions (controls). The corpus consists of 125 conversations with AD patients and an additional 125 dialogues with the control group. Several studies have already been conducted on this corpus (Luz et al., 2018; Nasreen et al., 2019). For example, Nasreen et al. (2019) used hand-annotated Dialogue Acts (DA) information to distinguish patients with AD. They focused on the types of questions asked by both groups, how they were answered, and whether any significant patterns appeared to differentiate the groups by investigating features combined with dialogue acts (such as *clarification question* and *signal non-understanding*), confusion rates (*question ratio* and *confusion ratio*), and other conversational information (such as the average number of words per minute and the number of speech turn switches per minute). Their best model was an SVM with n-gram dialogue acts as features. This study is noteworthy because it provides evidence for the potential multi-task learning of AD detection and dialogue act prediction. Unfortunately, we are unable to obtain the CCC corpus due to the Institutional Review Board (IRB) approval process.

5.3.2 Multi-Layer Annotation Corpus: DailyDialog

The task of Emotion Recognition in Conversations (ERC) lies at the intersection of dialogue modeling and emotion prediction. It has become increasingly popular in recent years, mainly thanks to the increasing number of publicly available corpora (Busso et al., 2008; McKeown et al., 2011; Li et al., 2017; Chen et al., 2018; Poria et al., 2018; Zhang et al., 2018b; Chatterjee et al., 2019). Several criteria are taken into consideration to choose an ERC corpus to be jointly trained with DAIC-WOZ, including corpus size, conversational topic, and corpus modality. DailyDialog (Li et al., 2017) is favored since it is a relatively large dataset and contains multi-layer annotations, including emotion, dialogue act, and topic, which enables us to investigate information from other auxiliary tasks in dialogues. We first present our choice of corpus for emotion prediction task, and then extend to other candidate corpora in the next section.

DailyDialog (Li et al., 2017) is a human-written corpus that contains $13,118$ two-party dialogues (with an average of 7.9 speech turns per dialogue). The corpus is publicly available[7].

[6]Participant#409 had PHQ-9 score at 10 but is given a binary score of 0. With respect to the original label we kept this instance in the class 0.

[7]`http://yanran.li/Dailydialog`.

Emotion	Train		Dev		Test	
	#	%	#	%	#	%
0-no emotion	72,143	82.8	7,108	88.1	6,321	81.7
1-anger	827	0.9	77	1.0	118	1.5
2-disgust	303	0.3	3	0.04	47	0.6
3-fear	146	0.2	11	0.1	17	0.2
4-happiness	11,182	12.8	684	8.5	1019	13.2
5-sadness	969	1.1	79	1.0	102	1.3
6-surprise	1,600	1.8	107	1.3	116	1.5
Utt. Total	87,170	100.0	8,069	100.0	7,740	100.0

Table 5.4: Emotion distribution in train, development and test sets in DailyDialog.

Dialog Act	Train		Dev		Test	
	#	%	#	%	#	%
1-inform	39,873	45.7	3,125	38.7	3,534	45.7
2-question	24,974	28.6	2,244	27.8	2,210	28.6
3-directive	12,242	16.3	1,775	22.0	1,278	16.5
4-commissive	8,081	9.23	925	11.5	718	9.3
Utt. Total	87,170	100.0	8,069	100.0	7,740	100.0

Table 5.5: Dialog act distribution in train, development and test sets in DailyDialog.

Three expert-annotated information are provided: 7 emotions ((Ekman, 1999) BigSix Theory: *happiness*, *surprise*, *sadness*, *anger*, *disgust*, *fear*, and *neutral*), and 4 coarse-grained dialogue acts (DA: *questions*, *inform*, *directives*, and *commissives*) at speech-turn level, and 10 topics at document level. We follow the original separation of the train (11,118), validation (1,000), and test (1,000) sets. Detailed statistics on emotions, dialogue acts, and topics are given in Table 5.4, 5.5, and 5.6, respectively. To enhance the modeling of dialogues, it would be beneficial for future research to examine more fine-grained annotations than the 4-way dialogue act annotations in DailyDialog. For instance, considering the SWBD-DAMSL tagset in Switchboard Corpus (Jurafsky, 1997) would provide a more fine-grained analysis.

5.3.3 Other Emotion-Enriched Conversational Corpora

Apart from DailyDialog, we discovered at least 6 candidate Emotion Recognition in Conversations (ERC) corpora, whose key information is shown in Table 5.7.

IEMOCAP (Busso et al., 2008): Interactive Emotional Dyadic Motion Capture Database, short in IEMOCAP, is one of the most important datasets for studying emotions in conversation. It consists of videos among pairs of 10 speakers spanning 10 hours of various dialogue scenarios (both scripted and spontaneous). During the conversation, markers were placed on the face, head, and hands to record information about facial expressions and hand movements. The emotion classes are slightly different from those of DailyDialog: {*anger*, *happiness*, *sadness*,

Topic	Train		Dev		Test	
	#	%	#	%	#	%
1-ordinary life	2,975	26.8	418	41.8	252	25.2
2-school life	453	4.1	0	0	34	3.4
3-culture & education	50	0	0	0.0	5	0.5
4-attitude & emotion	616	5.5	1	0.0	50	0.5
5-relationship	3,879	34.9	129	12.9	384	38.4
6-tourism	860	7.7	124	12.4	79	7.9
7-health	205	1.8	41	4.1	21	2.1
8-work	1,574	14.2	215	21.5	135	1.4
9-politics	105	0.9	13	1.3	13	1.3
10-finance	399	3.6	59	5.9	27	2.7
Total	11,118	100.0	1,000	100.0	1,000	100.0

Table 5.6: Topic distribution in train, development and test sets in DailyDialog.

Corpus	Modality	Topic	Participant	#Doc	#Utt
IEMOCAP (Busso et al., 2008)	multimodal	theater+daily talks	dyadic	151	7,433
SEMAINE (McKeown et al., 2011)	multimodal	emotional scenarios	dyadic	95	5,798
EmotionLines (Chen et al., 2018)	spoken+script	Friends TV series	multi-party	1,000	14,503
MELD (Poria et al., 2018)	spoken+script	Friends TV series	multi-party	1,433	13,708
Persona-Chat (Zhang et al., 2018b)	spoken	spontaneous talks	dyadic	10,907	162,064
DailyDialog (Li et al., 2017)	written	daily talks	dyadic	13,118	103,607
EmoContext (Chatterjee et al., 2019)	written	Twitter Q-A pairs	dyadic	38,421	115,263

Table 5.7: Key information of 7 ERC corpora. All corpora are in English.

neutral, *excitement*, *frustration*}. The dataset is available at `https://sail.usc.edu/iemocap/iemocap_release.htm`.

SEMAINE (McKeown et al., 2011): This is an audiovisual database used to create Sensitive Artificial Listener (SAL) agents that can engage in emotionally colored conversations with a person. The database was recorded in a Wizard-of-Oz setting where a human user talks to an artificially intelligent agent controlled by a human operator. The conversations revolve around topics that are emotionally significant to the participants and encourage them to express their emotions strongly. The database contains 959 conversations that are approximately 5 minutes long, recorded with 150 participants. The data was annotated with four real-valued affective attributes: *valence* ($[-1, 1]$), *arousal* ($[-1, 1]$), *expectancy* ($[-1, 1]$), and *power* ($[0, \infty]$). The dataset is available at `https://ibug.doc.ic.ac.uk/resources/semaine-database2/`.

EmoContext (Chatterjee et al., 2019): This is a collection of tweets (Twitter-Qs) and their corresponding responses (Twitter-As), spanning four years from 2012 to 2015. It focuses on three emotion categories: *happy*, *sad*, and *angry*. Unfortunately, the dataset is not currently accessible online.

EmotionLines (Chen et al., 2018): This dataset is derived from two sources: Friends TV scripts and private conversations on Facebook Messenger, known as EmotionPush Chat Logs (Wang et al., 2016). Each source contains 1000 dialogues. Every utterance in the dataset is labeled with one of Ekman's six basic emotions plus the neutral emotion. The annotations were obtained using Amazon Mechanical Turkers, and a total of $29,245$ utterances have been annotated. The dataset is available at `http://doraemon.iis.sinica.edu.tw/emotionlines/download.html`.

MELD (Poria et al., 2018): MELD is a multi-party dataset that incorporates multiple modalities such as audio, visual, and textual. The conversations in MELD are also extracted from the Friends series, similar to EmotionLines, but this dataset has undergone thorough revision and has removed outliers present in EmotionLines. Additionally, a dyadic version of the dataset is available where dialogues are divided into several two-party sub-dialogues. Compared to IEMOCAP and SEMAINE, MELD contains a greater number of labeled utterances. The dataset is available at `https://affective-meld.github.io/`.

Persona-Chat (Zhang et al., 2018b): This is a spoken dialogue dataset created through crowd-sourcing, where each participant assumes an assigned persona (there are 1155 possible personas, each with at least 5 profile sentences). The goal of the conversation is simply to chat and get to know each other naturally. The dataset comprises a total of $162,064$ utterances across $10,907$ dialogues. This dataset has been used to train next sentence prediction models based on the dialogue history. The dataset and trained dialogue models can be found on the ParlAI platform (Miller et al., 2017) at `https://github.com/facebookresearch/ParlAI`.

In summary, IEMOCAP and SEMAINE are two datasets that are commonly used in multimodal emotion recognition, but they have special topics and designed emotional scenarios related to theatre. On the other hand, EmoContext is a Twitter-based dataset with limited context length and is not easily accessible. When it comes to annotated emotions, MELD, and DailyDialog follow Ekman's BigSix Theory {*happiness*, *surprise*, *sadness*, *anger*, *disgust*, *fear*}, including

Label	DailyDialog	MELD	EmotionLines	IEMOCAP	EmoContext
Neutral	85572	6436	6530	1708	-
Happiness/Joy	12885	2308	1710	648	4669
Surprise	1823	1636	1658	-	-
Sadness	1150	1002	498	1084	5838
Anger	1022	1607	772	1103	5954
Disgust	353	361	338	-	-
Fear	74	358	255	-	-
Frustrated	-	-	-	1849	-
Excited	-	-	-	1041	-
Other	-	-	-	-	21960

Figure 5.5: Statistics of emotion class distribution in five ERC datasets, from Poria et al. (2019).

	Dailydialog				DAIC-WOZ				Daily+DAIC	
	#doc	#utt	#utt/doc	#tok/doc	#doc	#utt	#utt/doc	#tok/doc	#doc	#utt
Train	11,118	87,215	7.8	107	107	25,519	239	1,931	11,225	112,734
Dev.	1,000	7,806	7.8	109	35	9,326	267	2,061	1,035	17,132
Test	1,000	7,958	7.9	107	47	12,569	267	2,175	1,047	20,527
Total	13,118	102,979	7.9	107	189	47,417	251	2,016	13,307	150,393

Table 5.8: Number of documents and utterances in DAIC-WOZ and DailyDialog corpora.

neutral, while EmoContext only provides three categories {*happiness*, *anger*, *sadness*}. IEMOCAP covers three emotions with two additional classes: *frustration* and *excitement*. However, the distribution of emotion classes is unbalanced across different datasets, as depicted in Table 5.5 (taken from Poria et al. (2019)).

5.3.4 Our Combined Dataset

We choose DAIC-WOZ and DailyDialog as our primary and auxiliary datasets, respectively. Table 5.8 provides statistics for both datasets, including the number of documents, speech turn lengths, and token counts. DailyDialog has an average of 7.9 utterances per document, resulting in a total of $102k$ utterances, while DAIC-WOZ has an average of 250 utterances per document and a total of $47k$ utterances. There is an imbalance between the document length of two datasets: DAIC documents are almost 30 times longer than those in DailyDialog. DailyDialog has slightly longer sentences with an average of 13 tokens per sentence, while DAIC-WOZ has an average of 8 tokens per sentence.

In addition, we consider a *resize strategy* that cuts long documents in DAIC-WOZ into shorter sub-documents (8 speech turns per document to match the length in DailyDialog), thus artificially increases the number of instances while maintaining the document length. We only resize the training set while keeping the development and test sets unchanged. This strategy is tested for jointly training depression detection and emotion classification tasks.

5.4 Experimental setup

Baselines: We compare our MTL results with: (1) Majority class where the model predicts all subjects positive (i.e. depressive); (2) Baseline single-task model described in Section 5.2.2); (3) State-of-the-art results on the test set reported by Mallol-Ragolta et al. (2019) and Xezonaki et al. (2020). Note that for main results, we do not compare to Williamson et al. (2016); Haque et al. (2018); Al Hanai et al. (2018); Dinkel et al. (2019); Qureshi et al. (2020) who only report on the development set.

Evaluation Metrics: For depression classification, we follow Dinkel et al. (2019) and report accuracy, macro-F_1, precision, and recall scores. For emotion analysis, we report macro-F_1 score, following Cerisara et al. (2018).

Implementation Details: We implement our model with AllenNLP library (Gardner et al., 2018). We use the original separation of train, validation, and test sets for both corpora. The model is trained for a maximum of 100 epochs with early stopping. For STL as well as for MTL scenarios, we optimize on macro-F_1 metric for depression classification. We use cross-entropy loss. The batch size is 4 for Dailydialog and 1 for DAIC (within the limit of GPU Video Random Access Memory). We use the tokenizer from SpaCy Library (Honnibal et al., 2020) and construct the word embeddings by default with a dimension of 128. The *turn*-level has one hidden layer and 128 output neurons. We tune *document*-level RNN layers in $\{1, 2, 3\}$ and hidden size in $\{128, 256, 512\}$. Model parameters are optimized using Adam (Kingma and Ba, 2014) with the learning rate at $1e-3$. The dropout rate is set to 0.1 for both *turn*-level and *document*-level encoders. In summary:

- Learning rate: $\{1e-3, 1e-4, 2e-3\}$
- Dropout rate: $\{0.1, 0.2\}$
- Word embedding dimension: 128
- *Turn*-level layer: 1
- *Turn*-level hidden size: 128
- *Document*-level layers: $\{1, 2, 3\}$
- *Document*-level hidden size: $\{128, 256, 512\}$

5.5 Results and Analysis

5.5.1 Main Results

We show the results using MTL hierarchical structure for depression detection in Table 5.9, which are compared to majority vote baseline and SOTA models (at the top). Our baseline model is a single-task naive hierarchical model which obtains similar results (F_1 44) as the baseline model (NHN) in Mallol-Ragolta et al. (2019) (F_1 45).

Using the multi-task architecture, we get improvements when adding each task separately. We see more than a +11.5% increase in F_1 when adding emotion ('+Emotion') or topic ('+Topic') classification task and, at best, +16.9% with dialogue acts ('+DialogAct'). This demonstrates

	F_1	Precision	Recall	Accuracy
Baseline majority vote	41.3	35.1	50.0	70.2
State-of-the-art models				
NHN (Mallol-Ragolta et al., 2019)	45	-	50	-
HCAN (Mallol-Ragolta et al., 2019)	63	-	66	-
HAN+L (Xezonaki et al., 2020)	70	-	70	-
Ours				
STL Depression	43.9	44.5	47.5	63.8
MTL +Emotion	55.5	56.2	61.6	70.2
MTL +Topic	55.6	55.9	56.8	59.6
MTL +DialogAct	60.8	60.6	61.4	66.0
MTL +Emotion+Topic	64.4	64.4	64.4	70.2
MTL +DialogAct+Topic	63.7	**78.1**	62.8	**76.6**
MTL +Emotion+DialogAct+Topic	**70.6***	70.1	**71.5***	74.5

Table 5.9: Depression detection results on DAIC-WOZ. NHN: naive hierarchical network; HCAN: hierarchical contextual attention network; HAN+L: hierarchical attention network with external lexicons.
STL: single-task using DAIC-WOZ only; MTL: multi-task using DAIC-WOZ and adding classification for Emotion (+Emotion), Topic (+Topic), dialogue Act (+DialogAct) from Dailydialog.
*Significantly better than SOTA performance with p-value < 0.05.

the relevance of each task to the primary problem of depression detection, especially the interest of dialogue acts. When adding topics, we observe a small drop in accuracy compared to STL while the F_1 is better, meaning that the prediction for minority class (non-depressive) improves. Interestingly, in terms of accuracy, the tasks at different levels (depression '+Emotion' and depression '+DialogAct') seem to help more. We deduce that they help build a better local representation (speech turns) before the global representation. The '+DialogAct+Topic' model achieves the highest accuracy of 76.6. However, there is a significant gap between the recall and precision scores, indicating that the model tends to predict more negative classes (non-depressive subjects) while struggling with positive ones. This could lead to a failure to identify depression in real-life situations.

When jointly learning all four tasks – combining depression detection with three auxiliary tasks ('+Emo+Diag+Top') –, all metrics improve. We obtain our best system with an improvement of +26.7% in F_1 compared to STL baseline, outperforming the state-of-the-art with a +7.6% increase compared to the best system in Mallol-Ragolta et al. (2019) and about +0.5% compared to Xezonaki et al. (2020). Depressed people tend to express specific emotions; it is thus natural to think that emotion is beneficial for the main task. These results indicate that both emotion and dialogue structure help as they provide complementary information, paving the way for new research directions with more fine-grained modeling of dialogue structure for tasks in conversational scenarios.

5.5.2 Performance on Auxiliary Tasks

To better understand our model, we look at the performance of emotion, dialogue act, and topic classification tasks. Directly comparing the results of our MTL approach ('+Emo+Diag+Top')

Model	DailyDialog emo				DAIC-WOZ depr			
	F_1	Prec.	Rec.	Acc.	F_1	Prec.	Rec.	Acc.
ST Emo	38.3	45.2	35.9	80.8	-	-	-	-
MT Emo + Depr	40.0	42.5	37.3	80.7	55.5	56.2	61.6	70.2
MT Emo + Depr(resized)	41.3	46.4	38.9	80.3	68.3	72.8	66.9	76.6

Table 5.10: Classification results on emotion prediction on DailyDialog, with single-task (ST) and multi-task (MT) settings. Emo: emotion prediction; Depr: depression prediction; Depr(resized): resized train set in DAIC-WOZ to match the length of speech turn in DailyDialog.

with a STL architecture for each task, however, seems unfair. The optimized objective and structural complexity are different: the former is optimized on the depression detection task on two levels, while the latter is tuned on the target auxiliary task with either speech turn (emotion and dialogue act) or full dialogue (topic). Unsurprisingly, the results show that the MTL system underperforms the basic STL structure for dialogue acts and topics, with at best 67.8 in F1 (MTL) vs. 68.8 (STL) for dialogue acts, and 52.0 (MTL) vs. 52.4 (STL) for topic classification.

On the other hand, our MTL system achieves an F_1 score of 40.0 for emotion, compared to 38.3 for the STL baseline, demonstrating the benefits of joint learning of both tasks (Table 5.10). Resized training strategy shows further improvements for both emotion prediction (41 vs 40) and depression detection (68 vs 56). The performance breakdown for each emotion is depicted in Figure 5.6. It should be noted that the distribution of emotion classes in DailyDialog is highly imbalanced, with one dominant class, *neutral*, occupying more than 80% of the dataset, while five rare classes together account for less than 5%. The F score for the *neutral* class remains consistently high in both ST and MT settings, with values above 88%. The *happiness* class, the second largest with approximately 12% of the training set, shows modest improvement from the depression task (+0.7). Among the four negative emotion classes, three show clear improvement in the MT setting: *anger*, *disgust*, and *sadness*, with F score gains of 5%, 6%, and 1%, respectively. The improvement is even greater $(7\%, 16\%, 6\%)$ when training with the resized DAIC set. The *fear* class is the smallest in the corpus (with a proportion of 0.2%) and the most challenging one to predict. The result demonstrates a 2% increase with the resized DAIC set. Surprisingly, *surprise* appears to be the only class that does not benefit from the additional task, with a decrease of 1.5% in performance. Resizing the DAIC set does not show any benefits for this class either. However, overall, the MTL model proves to be beneficial for the emotion prediction task, with $1-3$ points improvement on F_1 scores compared to the ST baseline, whether using the original or resized train sets.

We believe that the augmentation of emotion-related utterances in a shared network task is the primary reason for the improved performance. To test our hypothesis, we manually analyze the dialogue acts of the utterances from Ellie, and categorize them into high-level classes: *Backchannel*, *Comment*, *Opening*, *Other*, and *Question*. We use a different set of dialogue acts than those in DailyDialog to better align with Ellie's speech intentions. The annotation is carried out by a single annotator. We discover that approximately 13% of the utterances are emotion-related, including queries such as "things that make you mad", "things you feel guilty about", and "last time you felt really happy". Additionally, mentions of topics related to happiness or regret appear in almost all the interviews. Furthermore, as the original DAIC-WOZ conversations are long, several emotion-related utterances are included in one document. By reshaping the training set, we not only increase the size of learning instances but also reduce the complexity of

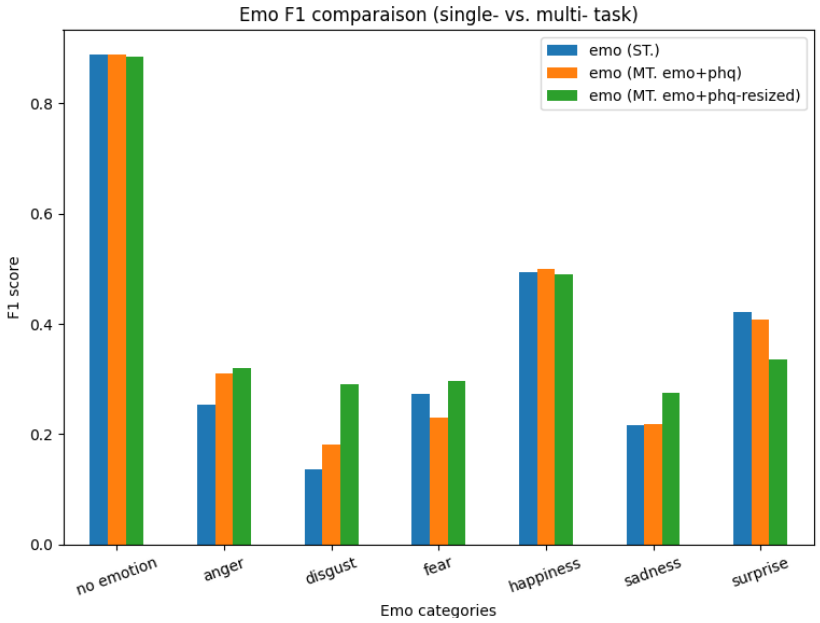


Figure 5.6: Class-wise emotion performance in single-task (ST) and multi-task (MT) settings. "phq": PHQ-8 score, used to indicate depression.

High-level DA	#	%	Sub-category	#	%
Question	7,907	53%	Emo	1,054	13%
			Non-emo	6,853	87%
Backchannel	3,231	22%	-	-	-
Comment	3,074	20%	-	-	-
Opening	611	4%	-	-	-
Other	171	1%	-	-	-

Table 5.11: High-level dialogue act (DA) distribution of Ellie's speech in DAIC-WOZ. # and % represent the number and percentage of Ellie's utterances, respectively.

learning multiple emotions in a single document, thus leading to better results. The distribution of dialogue act annotations is presented in Table 5.11. The annotation is available for free use for future research.

5.6 Conclusion

In this chapter, we continue the discussion on the topic of discourse structure in dialogues within the context of mental illness. Along the way, we face various challenges, including data scarcity, interaction modeling, and dialogue structure modeling. To overcome the data scarcity issue, we conduct a thorough investigation of publicly available corpora for our primary cognitive impairment detection task and auxiliary tasks. The selection process leads us to opt for the commonly used English corpus, DAIC-WOZ, for depression detection, and we also identify a potential corpus containing conversations with Alzheimer's Disease patients. To address the

drawback of lacking interaction in the previous chapter, we develop a hierarchical neural network architecture designed to model speech turns from two parties. Lastly, we believe that modeling dialogue structure requires the consideration of different levels of information, from speech turns to the entire dialogue. As such, we propose leveraging information from dialogue act and topic modeling from another dialogue dataset, DailyDialog.

We demonstrate a correlation between depression and emotion, and show the importance of dialogue structures through the use of shallow markers like dialogue acts and topics. To improve our approach, we may consider incorporating other features, such as speaker identity (Qin et al., 2020) and common-sense knowledge (Ghosal et al., 2020). Our next goal is to investigate more advanced dialogue structure modeling, potentially using discourse parsing. However, discourse parsing by itself is a challenging task, with limited domain applicability and data scarcity issues. We aim to address these challenges in the upcoming chapters (Chapter 7 and Chapter 8) by presenting novel strategies to overcome insufficient training data and creating a general discourse parsing model for future use. We also plan to extend our work beyond binary depression classification to include severity classification using a cascading structure: first, detect depression and then classify the severity. To ensure the stability of our model, we intend to refine our work and report on cross-validation splits of the data, which is especially important when dealing with sparse data that may not be representative. A further step will be to investigate the generalization of our model to other mental health disorders, such as Schizophrenia and Alzheimer's Disease.

Part III

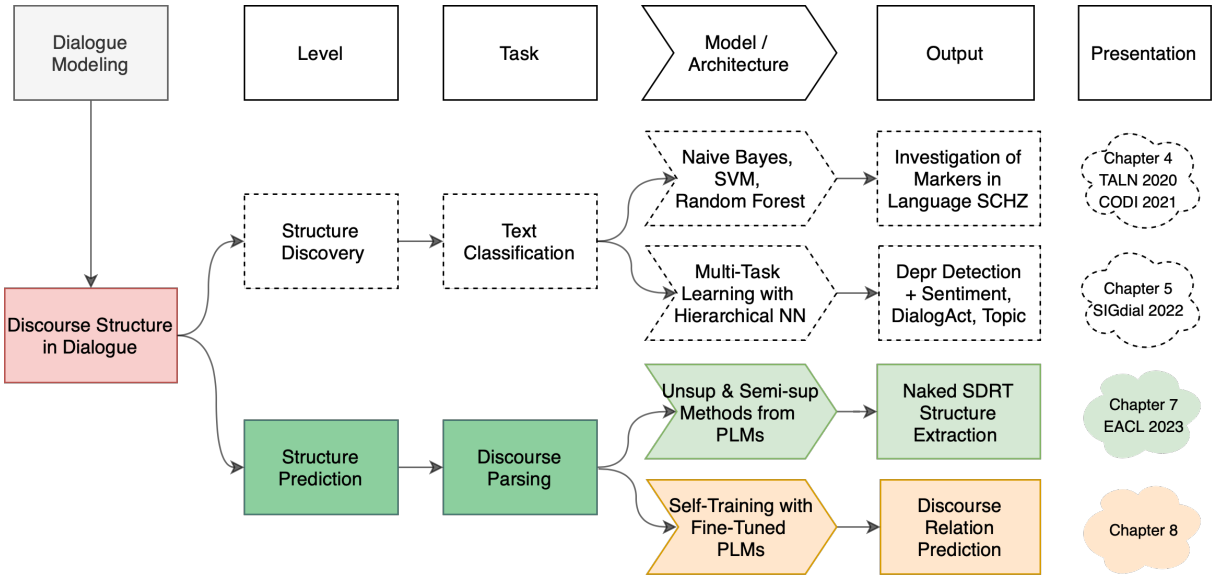
Discourse Structure Prediction

After discussing two studies on discourse structure discovery in part II, we now move towards the second research question of this thesis, namely:

RQ2 How can we generate discourse structures with machine learning techniques using minimal supervision to achieve the greatest applicability in real-life scenarios?

Our focus in this part is on discourse parsing in dialogues, specifically using the SDRT framework to infer both `EDU` attachment and relation prediction. We are aware of the issue of data sparsity in discourse parsing, which we have addressed in Chapter 3 by discussing various distant and transfer learning strategies. However, these strategies are primarily designed for monologues, and it is unclear how well they generalize to other domains and how dependent they are on the availability of annotated data. In contrast, language modeling can be regarded as an additional task that captures general linguistic knowledge without requiring annotation. Pre-trained language models (PLMs) such as BERT Devlin et al. (2019a), BART Lewis et al. (2020), and the GPT family Radford et al. (2019); Brown et al. (2020) have become increasingly popular and have been applied to various NLP tasks. The popularity of these large models has spawned a subfield of research called "BERTology", which seeks to understand what implicit representations are learned by these models. Previous studies reveal that LMs capture certain aspects of language dependency, such as subject-verb agreement Goldberg (2019); Jawahar et al. (2019) and syntactic dependency Tenney et al. (2019); Hewitt and Manning (2019). Nevertheless, the discourse aspect of PLMs has not been thoroughly explored.

For a better understanding of LMs, we first establish the basics and related BERTology work in Chapter 6. We then present experimental results on the use of PLMs for discourse parsing in dialogues. Our approach involves a *structure-then-relation* pipeline for tackling this problem, as detailed in Chapter 7 and Chapter 8, respectively. It is not unusual in real-life scenarios to have a few dozen short documents annotated by experts. Thus, in this part, we utilize 50 annotated documents for supervision. In extreme cases, we also present experiments that require no supervision and instead rely solely on the text itself and the attention mechanism in PLMs.



Chapter 6

Pre-Trained Language Models & Discourse

Contents

In this chapter, our focus is on Pre-Trained Language Models (short in PLMs). These models have demonstrated remarkable performance on a wide range of NLP tasks, including sentiment analysis (Gu et al., 2021), machine reading comprehension (Yang et al., 2019a), relation extraction (Alt et al., 2019), and semantic role labeling (Shi and Lin, 2019), among others. Although it is evident that PLMs excel in these tasks, the underlying reasons for their success remain less understood. Unlike classical machine learning models such as logistic regression or decision trees, PLM's architecture is much more complex. The significant size of the parameters and gradient calculation make a hard task to unveil the reasoning process inside the model. Researchers are fascinated by the mechanism behind these models. It is evident that the utilization of large-scale PLMs can be advantageous for machine comprehension and information extraction (Liu et al., 2022). However, the extent to which they encode discourse-level information is a relatively unexplored area. The purpose of this chapter is to investigate the capability of PLMs to capture and encode discourse information.

To achieve this goal, we start by giving a brief history of word embeddings – the predecessors of contextualized representations in PLMs in Section 6.1. We then introduce the basic concepts of PLMs, including their architectures, training languages (mono-lingual or multi-lingual), and learning objectives in Section 6.2. Then, we delve into the field of "BERTology" – a field of study that investigates the inner workings of Transformer-based models – and discuss the knowledge representation in these models in Section 6.3. In Section 6.4, we focus on studies that explore discourse information encoded in PLMs, typically via probing tasks, and how to extract these information.

6.1 From Word Embeddings to Pre-trained Language Models

Word embeddings are fixed-length vectors that are dense and distributed representations for words, based on the distributional hypothesis (Almeida and Xexéo, 2019). The concept of word embeddings can be traced back to the 1950s, with the introduction of distributional semantics (Harris, 1954; Firth, 1957), which is based on the idea that the meaning of a word can be inferred from the context in which it appears.

In NLP, word embeddings have emerged as a useful tool for transforming words into numerical vector spaces. This approach is particularly advantageous as computers are better equipped to directly handle numbers, and the resulting vectors can be subjected to useful mathematical operations such as addition, concatenation, and distance measures. These vectors are also well-suited for various tasks such as measuring the semantic similarity between words, phrases, and documents (Turney and Pantel, 2010). Salton et al. (1975)'s **Vector Space Model** (VSM) is considered to be one of the most influential models in information retrieval (IR) history. In VSM, each document is represented by a vector where each dimension corresponds to a specific feature or term. The values in the vector indicate the presence or importance of the corresponding feature in the document. Each document is shown as a point in a vector space. The proximity of points in this space reflects the semantic similarity, with close points being semantically similar and distant points being semantically different. The success of VSM in IR soon extends to other tasks in NLP. For example, in Rapp (2003), vector-based representations of word meaning achieve a 92.5% accuracy on multiple-choice synonym questions from the Test of English as a Foreign Language (TOEFL); Turney (2006) use a vector-based representation of semantic relations and score 56% on multiple-choice analogy questions from the SAT college entrance test, which is comparable to the human score of 57%.

In the early 2000s, researchers begin to develop computational methods for creating word embeddings automatically. The first widely used method is **Latent Semantic Analysis** (LSA) (Dumais et al., 2004), which applies singular value decomposition to a co-occurrence matrix of words to obtain a reduced-dimensionality representation.

In 2013, Mikolov et al. (2013) introduce **Word2Vec**, which is a single-layer neural network based on the inner product between two word vectors. The core idea is that a word can be represented by a set of words that appear nearby. Word2Vec has two models, namely the *continuous bag-of-words* (CBoW) and *Skip-gram* models. CBoW learns the context words and predict one target word, while Skip-gram uses the target word to predict its surrounding words. Word2vec is said to use *prediction-based* approach since it is based on teaching the word vectors to predict the contexts in which the words reside (Baroni et al., 2014; Almeida and Xexéo, 2019). The embeddings generated from Word2Vec are *static*, meaning that they are fixed vectors. Another method to create word embeddings is *count-based*, which creates word vectors upon word occurrences statistics, a well-known model is **Global Vectors** (GloVe) (Pennington et al., 2014). GloVe uses a global context window to calculate word-word co-occurrences, in comparison to Word2Vec which uses window-based methods to scan the context across the entire corpus. The resulting embeddings of GloVe show interesting linear substructures of the word in vector space, such as "Paris to France" is close to "Rome to Italy". In the following years, NLP community has witnessed emergence of more word embeddings models. **FastText** (Bojanowski et al., 2017), for instance, innovatively incorporate character-level n-grams rather than word-level tokens and to address the out-of-vocabulary problem appeared in the previous models.

Static embeddings face challenges in representing polysemy, as they provide only one representation for a word regardless of its linguistic context. To illustrate this, let us consider the sentence "I left my pen on the left side of the table". The word "left" appears twice in the sentence

with different meanings, but static embeddings cannot capture this distinction. In recent years, significant advancements have been made with the introduction of *contextual embeddings*, such as **Embeddings from Language Models** (ELMo) developed by Peters et al. (2018a). ELMo is considered a milestone in the area of word embeddings, following the success of Word2Vec. ELMo vectors are derived from a bidirectional LSTM trained with a coupled language model objective, using a large text corpus. These representations are deep, as they are based on all internal layers of the bidirectional LSTM. Another notable development is **Bidirectional Encoder Representations from Transformers** (BERT) introduced by Vaswani et al. (2017). BERT utilizes the Transformer architecture, which we will explain shortly, as an alternative to the recurrent neural network used in ELMo. By applying bidirectional training of the Transformer model to language modeling, BERT can effectively learn the contextual information of a word by considering its surrounding context. These contextual embedding models have revolutionized the field of NLP by capturing the nuances of word meanings within different contexts. The development of contextual embeddings has also stimulated research in neural network-based language modeling.

In addition, there are some interesting research in comparing and incorporating static and contextual embeddings. Bommasani et al. (2020) propose to interpret contextual embeddings with static embeddings since the latter have more mature interpretability methods, i.e., convert BERT embeddings back to static vectors in Word2Vec and GloVe. We refer this study to readers who are interested.

6.2 Basics of Pre-Trained Language Models

Language modeling is a concept that has been present since early stages. In the work of Bengio et al. (2000), a probabilistic language model is defined as a model that calculates the probability of the next token based on all the previous tokens in a sequence. Over time, with the advancements in neural networks (NN), NN-based language models have gained significant popularity due to their remarkable performance in various NLP tasks. In this section, we mainly discuss NN-based language models.

A pre-trained language model, in simple term, is a type of machine learning model trained on a large corpus of text data in an unsupervised manner. During pre-training, the model learns general prediction tasks, such as *masked language modeling* – predict missing words in a text sequence, and *next sentence prediction* – generate coherent text based on a given prompt. The goal of pre-training is to teach the model linguistic knowledge and to generate meaningful representations. There are various terms used to refer to pre-trained language models, such as Large Language Models (LLMs), Neural Language Models (NLMs), Language Models (LMs), Foundation Models (FMs), and Pre-trained Language Models (PLMs). In this thesis, we use the term PLMs. These models have great generalization ability and can be fine-tuned for specific tasks and new domains. Some examples of popular pre-trained language models include BERT (Devlin et al., 2019a), RoBERTa (Liu et al., 2019a), and GPT-3 (Brown et al., 2020).

Since their presence, PLMs are at the base of many state-of-the-art approaches in NLP field. The common procedure is first "pre-training" a model and then "fine-tuning" it to adapt to different tasks and domains. We set the foundation by exploring some fundamental concepts in PLMs.

Architecture & **Schema:** We start by introducing a revolutionary architecture, known as **Transformer**. This architecture is introduced with the **self-attention mechanism** by

Vaswani et al. (2017), and soon becomes the core component in upcoming PLMs. Self-attention is a mechanism that enables the model to weigh the importance of different parts of the input sequence. It is called "self" because the model attends to the input sequence itself, rather than attending to a separate sequence or context. We can visualize this mechanism in a matrix $\mathcal{A}$ of size $n \times n$, with n being the number of tokens of an input sequence. Each token can interact with each other and decide who they should pay more attention to and put a value in corresponding case in matrix $\mathcal{A}_{ij}$[1]. In this way, self-attention helps the model to focus on the most relevant words or tokens in the input sequence.

Transformer is a stacked self-attention layers. A standard Transformer structure contains an **encoder** and a **decoder** layer. The encoder layer takes in a sequence of input tokens and generates a sequence of hidden states, where each hidden state represents the input token at that position along with its context in the sequence. While the decoder layer takes in the encoded sequence from the encoder layer and generates a sequence of output tokens *autoregressively*, meaning that it generates one token at a time by attending to the previously generated tokens. It also uses multi-head self-attention to attend to the encoded sequence, along with encoder-decoder attention to capture the alignment between the input and output sequences.

Transformer-based PLMs can be classified into different **schemes**: *encoder-only* such as BERT (Devlin et al., 2019a) and RoBERTa (Liu et al., 2019a); *decoder-only* such as the GPT family (Radford et al., 2018, 2019; Brown et al., 2020); and finally, *encoder-decoder* structure (*sequence-to-sequence*) such as BART (Lewis et al., 2020). We visualize these schemes in Figure 6.1.

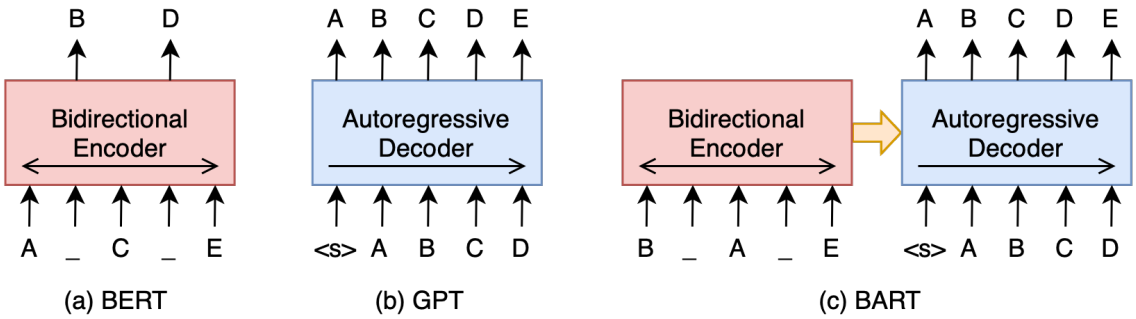


Figure 6.1: A schematic comparison between BERT, GPT, and BART. Adapted from Lewis et al. (2020). (a) BERT utilizes only bidirectional encoder networks; (b) GPT is composed of autoregressive (uni-directional) decoder networks; (c) BART contains both bidirectional encoder and autoregressive decoder.

Encoder-only PLMs such as BERT can be used for various Natural Language Understanding (NLU) tasks. Without auto-regressive decoder layers, the missing tokens are predicted individually, which make this schema of language models not easily used for generation (Lewis et al., 2020). On the other hand, decoder-only PLMs such as GPT are often used for Natural Language Generation (NLG) tasks that require coherent and fluent text production. The recent incredibly powerful ChatGPT model is based on the GPT-3.5 architecture. Lastly, BART is an encoder-decoder model that excels in text generation and summarization tasks, and its encoder layers make it suitable for comprehension tasks as well. Notably, the attention matrices in the encoder layers of BART have demonstrated higher proficiency in capturing discourse informa-

[1]We leave the detail calculation process with key, query, and value matrices for interested readers. A nice explanation can be found in this blog: `https://jalammar.github.io/illustrated-transformer/`.

tion compared to those in encoder-only PLMs. We present related research on this topic in Section 6.4.

Although Transformer architecture dominates PLMs, it is worth noting that not all PLMs use this architecture. Some of them is composed of recurrent neural networks (RNNs) which also allow them to learn contextual representations. ELMo (Peters et al., 2018a), which we have discussed earlier, is based on bidirectional LSTMs.

Training Objectives: Two commonly used training objectives for language model pre-training are *Masked Language Modeling* (MLM) and *Causal Language Modeling* (CLM). MLM, introduced in Vaswani et al. (2017), involves predicting a masked token within a sequence, allowing the model to attend to tokens bidirectionally. This objective is utilized in models such as BERT, RoBERTa, and BART. In contrast, CLM predicts next word in a sequence and can only consider the words appearing on the left side, making it unidirectional. Examples of pre-trained models with this objective include the GPTs. Another pre-training objective is *Translation Language Modeling* (TLM), which gives rise to Cross-lingual Language Models (XLMs) (Conneau and Lample, 2019). TLM extends MLM to parallel sentences in two different languages and masks words in both sentences. By considering both languages, a model trained using TLM can predict a word in one language by attending to its context and the translation, facilitating better alignment of different representations.

Figure 6.2 shows different training objectives. In (a) MLM, random tokens are replaced by masks ("_") and the model learns to predict the missing tokens during pre-training. The masked token can weigh the representation of every other input word to learn its representation (α is the attention weight). TLM is very similar to that of MLM, except that they extend MLM to pairs of parallel sentences. For example, to predict a masked English word "curtains", the model can attend to both the English sentence and its French translation, as shown in (b). In (c) CLM, tokens are generated one step at a time: given "<s>" the model predicts "the"; given "<s> the", it predicts "curtains", etc., until the final token "blue". A CLM uses a special end-of-sentence token to indicate the end of the sequence, such as *<eos>* or *</s>*.

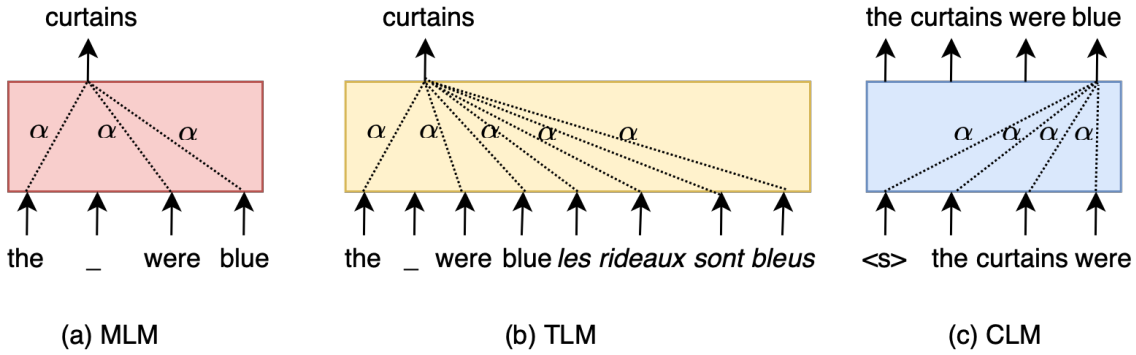


Figure 6.2: The comparison of MLM, TLM, and CLM pre-training objectives, adapted from Conneau and Lample (2019). α are attention weights. In (b) TLM, French tokens are italicized.

In addition, pre-training objectives can be combined. BART (Lewis et al., 2020), for instance, is a combination of MLM and denoising autoencoding. The input to encoder is corrupted text with missing tokens and shuffled text spans (that is why it is called denoising), and the model needs to generate the original text out from the decoder.

Mono-lingual & **Multi-lingual PLMs:** Although most PLMs are trained with English texts, much efforts have been put into multilingual models. We see examples such as multilingual BERT (mBERT) (Devlin et al., 2019a): pre-trained on 104 highest-resource languages in Wikipedia, and XLM-RoBERTa (XLM-R) (Conneau et al., 2020): masked language model trained on 100 languages over two terabytes of filtered CommonCrawl data.

Other studies focus on creating monolingual BERT in different languages, among which we find BERT in French: FlauBERT (Le et al., 2020) and CamemBERT (Martin et al., 2020), in German: GottBERT (Scheible et al., 2020) and German BERT[2], in Dutch: BERTje (de Vries et al., 2019), in Spanish: BETO (Cañete et al., 2020), in Russian: Ru-BERT (Kuratov and Arkhipov, 2019), in Finnish: Fin-BERT (Virtanen et al., 2019), as well as in Portugese (Polignano et al., 2019) and in Japanese (Kikuta, 2019).

The effectiveness of multilingual models in comparison to monolingual models on tasks such as neural machine translation (NMT) has recently garnered considerable attention. For instance, a recent study by Xu et al. (2021) revealed a decline in performance and relatively minor improvements when employing multilingual models for English $\rightleftharpoons$ German machine translation. This is possibly caused by the *curse of multilinguality* (Conneau et al., 2020) where low-resource language performance can be improved by adding higher-resource languages during pre-training; on the contrary, high-resource performance suffers and degrades. A possible compromise approach is to create bilingual PLMs with special focus on two involving languages.

6.3 BERTology: A Probe into BERT

With the increasing dominance of PLMs in the field of NLP, there has been a significant focus on studying the inner workings of large-language models such as BERT. This research area, commonly referred to as "BERTology", has garnered considerable attention. The primary objective of BERTology is to gain insights into the types of information captured by these models and explore ways to effectively utilize them. Probing tasks, also referred to as diagnostic classifiers, usually involves designing a separate task that focuses on extracting specific linguistic information from the model's internal representations.

Clark et al. (2019) examine the behavior of attention heads in general and probe each attention head for linguistic phenomena. They observe that attention heads tend to focus on delimiters such as the "[SEP]" token or punctuation, and that heads within the same layer exhibit similar behavior. They conduct single-head experiments and discover that certain heads specialize in particular aspects of syntax, such as identifying noun modifiers and possessive pronouns in dependency parsing, and exhibit similar behavior in coreference resolution tasks.

Tenney et al. (2019) investigate how BERT captures linguistic information and examines whether it follows the traditional NLP pipeline order of POS tagging, semantic parsing and SRL, and coreference resolution - in increasing difficulty. The authors propose two metrics to assess this: the *scalar mixing weights* measure which layers in combination are most relevant to the task, while *cumulative scoring* calculates the additional gain when adding another layer in the probing test, indicating at which layer the target can be accurately predicted. Notably, their findings suggest that BERT exhibits a consistent trend across both metrics for linguistic patterns, consistent with those observed in Peters et al. (2018b). Additionally, the study demonstrates that syntactic information tends to be concentrated in a few layers, indicating that it is more localized, while semantic information is generally distributed throughout the entire network.

[2] `https://deepset.ai/german-bert`

Rogers et al. (2020) conduct a comprehensive investigation of 150 studies of the BERT model. Their findings show precise knowledge that BERT learns or fails, mostly in syntactic and semantic domains. For instance, they find that BERT representations for syntactic tasks are hierarchical rather than linear and that it does not understand negation. As for semantic information, they observe that BERT has some knowledge of semantic roles while struggling with number representations; and that they can not reason based on the world knowledge. They also present various proposals on how to optimize the training process and model architecture, and suggest a few future research paths on BERTology.

Although the initial focus of BERTology was on probing tasks specifically for BERT, it has expanded to include other PLMs such as ELMo, GPT, and XLM. For example, Hewitt and Manning (2019) propose a *structural probe* for extracting syntax trees from BERT and ELMo. Zhu et al. (2020b) compare rhetorical capabilities (linguistic features linked to RST such as `EDU` length, discourse tree properties) of BERT-based models, GPT, and XLNet (Yang et al., 2019b). Koto et al. (2021) also investigate discourse capacities by comparing BERT-like models with GPT-2, BART, and T5 (Raffel et al., 2020).

Admittedly, probing tasks can provide analysis in different linguistic aspects, such as syntactic structures (Hewitt and Manning, 2019; Kim et al., 2019; Mareček and Rosa, 2019), agreement (Goldberg, 2019; Jawahar et al., 2019), ontologies (Michael et al., 2020), and semantic roles (Ettinger, 2020; Tenney et al., 2019). They also have shortcomings. As noted by Tenney et al. (2019), the absence of a linguistic pattern in a probing classifier does not guarantee its absence, and the presence of a pattern does not indicate how it is used. Rogers et al. (2020) also warn that different probing methods can lead to contradictory results, so relying on a single test is insufficient. To address these issues, Elazar et al. (2021) propose an alternative method called *amnestic probing*, which involves removing a property in a given task and measuring its influence, offering new directions for future probing research. As suggested in Rogers et al. (2020), the key message is that we have more questions than answers about the workings of BERT. While our current understanding is limited, the immense potential of PLMs should not be overlooked and further detailed studies are needed to unravel their intricacies. Moreover, it is important to be careful of the language used to describe these models. Pre-Reinforcement Learning from Human Feedback (RLHF) models such as BERT and BART are only exposed to *forms* during training. Hence, these models are not supposed to "understand the meaning" as our human do (Bender and Koller, 2020). Recent AI advancement introduces RLHF techniques into LLMs such as InstructGPT Ouyang et al. (2022) and ChatGPT (OpenAI, 2023), aligning the model's training objectives to that of complex human values and preferences. We can expect a better understanding of the world knowledge from these systems.

6.4 Discourse Information Exploration with PLMs

The growing importance of incorporating discourse information in various downstream tasks, such as summarization, argument mining, and machine translation (discussed in Section 3.3), has led researchers to explore the extent to which PLMs capture discourse information. In this section, we present studies that utilize probing tasks to assess the presence of discourse within PLMs, including tasks such as `EDU` segmentation, discourse connective detection, and relation identification. Furthermore, we delve into a recent study that demonstrates the direct extraction of discourse structure from PLMs, which serves as a source of inspiration for our work in Chapter 7. A summary of the relevant studies is provided in Table 6.1.

Model	Setting	Framework	Probing/Parsing	Language	PLMs
Probing					
Zhu et al. (2020b)	monologue	RST	(1) rhetorical relation occu.	en	BERT, RoBERTa
			(2) tree depth	en	GPT-2, XLM, XLNet
			(3) EDU length	en	
Pandia et al. (2021)	monologue	PDTB	(1) connective prediction	en	BERT, RoBERTa
			(2) causal *vs* concessive	en	ALBERT
			(3) temporal implicature	en	
Koto et al. (2021)	monologue	RST	(1) 4-way NSP	en, zh, de, es	BERT, RoBERTa
			(2) sentence ordering	en, zh, de, es	ALBERT, ELECTRA
			(3) discourse connective	en, zh, de	GPT-2, BART, T5
			(4) RST nuclearity	en, zh, de, es	
			(5) RST relation	en, zh, de, es	
			(6) RST EDU segmentation	en, zh, de, es	
			(7) cloze story test	en	
Wu et al. (2020)	monologue		dependency parsing	en	BERT
Self-supervised					
Huber and Carenini (2020c)	monologue	RST	structure	en	auto-encoder
Huber and Carenini (2022)	monologue	RST	structure	en	BERT, BART, +ft
Li et al. (2023)	dialogue	SDRT	(1) structure	en	BART, +ft
			(2) EDU seg + structure	en	

Table 6.1: Summary of discourse probing (upper part) and self-supervised discourse parsing (lower part) tasks in BERTology. "NSP": next sentence prediction. In "Language" column: en=English, zh=Chinese, de=German, es=Spanish. In "PLMs" column: +ft: fine-tuned PLMs.

6.4.1 Discourse Probing Tasks

Since the emergence of BERTology research, much attention has been put on exploring syntactic (such as grammaticality, dependency structure) and semantic (such as semantic role labeling, coreference resolution) information. Only until recently, efforts have been put in semantic and pragmatic levels.

In one of the earliest studies exploring the rhetorical capabilities of PLMs, Zhu et al. (2020b) examine the inter-sentential rhetorical knowledge. They evaluate several PLMs, including BERT-based models (BERT, BERT-m, RoBERTa), GPT, and XLM, using 24 features grouped into three categories: tree properties (depth and Yngve depth), EDU length, and the frequency of discourse relations (such as *attribution* and *background*). The probing task is formulated as an optimization problem in which an oracle RST-parser (Feng and Hirst, 2014a) is used to provide parsed trees and a probing matrix is used to extract the aforementioned features. This study reveals that BERT-based language models outperform GPT and XLM models in terms of stability across tasks and layers, as well as distribution of features across layers. The researchers suggest that BERT-based models perform better due to their ability to incorporate rhetorical information from both directions. However, this study only demonstrates shallow discourse capabilities in PLMs, and it remains unclear whether PLMs can encode structural information such as tree structure.

Pandia et al. (2021) aim to infer inter-sentential pragmatic knowledge through the prediction of discourse connectives. They formulate their experiments as *cloze tests*, i.e., no fine-tuning or any supervised training of PLMs, only to see how well these models have already encoded pragmatic knowledge. They select $\approx 17k$ instances from PDTB-2 (Prasad et al., 2008a) with explicit one-word connectives in order to satisfy the masked single-word prediction setting. Since connectives depend very much on the left and right side contexts, authors explore three masked

language models (MLM): BERT, RoBERTa, and ALBERT (Lan et al., 2019). For the cloze tasks, they set three scenarios ranging from the most naturally occurring setting to more controlled one. In the first scenario, the authors replace the connective between two sentences with a mask token (<mask>) and ask the model to predict the probabilities of all candidate connectives (66 in total), while keeping the two sentences intact. All models achieved an accuracy above 50%, with RoBERTa outperforming others at 66%. Accuracy increases with the model size across different models. However, the scores breakdown revealed a different story: the *Conjunction* connectives such as "and" are excessively predicted, while other categories such as *Causal: result* and *Concession* are significantly under-predicted. It is difficult to ascertain whether PLMs truly comprehend the implications of these connectives or merely give trivial predictions of major connectives. In the second and third settings, Linguistic pairs are constructed with nearly identical syntax and word content but a subtle difference in context. These tests can be difficult even for humans, as they test the pragmatic abilities of models in reducing the impact of shallow syntactic and lexical cues - which models are more likely to have learned to prioritize. Not surprisingly, the results demonstrate significant failure for all models in the second and third scenarios, with accuracy hovering around 0 or 25%. These scores suggest that this form of pragmatic competence is still lacking in PLMs.

Koto et al. (2021) investigate the ability of PLMs to process discourse information through seven probing tasks, including next sentence prediction (NSP), sentence ordering, connective prediction, `EDU` segmentation, nuclearity prediction, relation prediction, and a cloze story test which requires selecting the best ending for a four-sentence story. They experiment with seven PLMs (BERT, RoBERTa, ALBERT, ELECTRA (Clark et al., 2020), GPT-2, BART, and T5) and evaluate on four languages: English, Chinese, Spanish, and German. This study provides a comprehensive examination of the pragmatic capabilities of PLMs. The results suggest that BERT and BART are better than other models at capturing discourse information, especially in their encoder networks. GPT-2, a pure language model, struggles in this regard. Among the tasks examined, sentence ordering and RST relation prediction pose greater challenges for all models. These results serve as a foundation for future research, as highlighted in Huber and Carenini (2022); Li et al. (2023), where researchers are encouraged to utilize BERT and BART as primary PLMs for discourse structure extraction.

The aforementioned studies use probing tasks to explore PLMs discourse capabilities. However, this approach is undermined by the uncertainty of the amount of knowledge that is learnt by the probe itself: do the Language Models genuinely encode linguistic information, or is it the probe that learns the task itself? In order to reduce the impact of the probe, Wu et al. (2020) propose a **perturbed masking** method to analyze PLMs. They measure the impact a word x_j has on predicting another word x_i and build an impact matrix which is then used to induce syntactic and discourse structures. For the discourse task, they generate an `EDU`-level impact matrix $\mathcal{F}$ and use Eisner and CLE algorithms to extract dependency structures. Their experiments on the SciDTB dataset (Yang and Li, 2018) demonstrate that the Eisner algorithm and Euclidean distance perform the best (achieving a UAS of 34.2), although this is nearly 7 points below the left-chain baseline. As a point of reference, a supervised graph-based parser (Li et al., 2014c) achieves a UAS of 57.6 on the same dataset. This study stands out from other works in two ways. Firstly, it employs parameter-free probing methods. The impact matrix does not add any new parameters, which allows for a more straightforward examination of the encoded linguistic information. Secondly, the authors evaluate the efficacy of new probes on document-level structure rather than relying on shallow discourse signals like `EDU` length or connectives. They aim to reconstruct the internal structures from the impact matrix.

6.4.2 Discourse Inference via Self-Supervised Learning

In continuation with the research on discourse probing tasks, Huber and Carenini (2022) introduce a new approach to encode entire long documents into PLMs and extract their RST-style discourse structures. When using Transformer-based models, a major limitation is the input length (for instance, BERT-base is limited to 512 sub-tokens, while BART can handle up to 1024 sub-tokens). Despite the development of larger language models like Transformer-XL (Dai et al., 2019), or sparse pattern models such as Longformer (Beltagy et al., 2020) and Big Bird (Zaheer et al., 2020), many systems still encounter difficulties with document-length inputs. To offer a solution that is model-agnostic and applicable to any transformer-based architecture, the authors suggest an approach called the "sliding window" method. This method involves dividing the long document of m sub-word tokens into multiple sequences with a maximum length of t_{max}, and then sliding down one token at a time until the end of the document is reached. By doing this, $(m - t_{max} + 1)$ of partial sequences are generated and are put into LMs to obtain partial attention matrices M_P. Document-level matrix M_D is obtained from the addition of all M_P matrices. Finally, by dividing M_D to its frequency-tracking matrix M_F, a frequency normalized self-attention M_A is achieved. This process is illustrated in Figure 6.3.

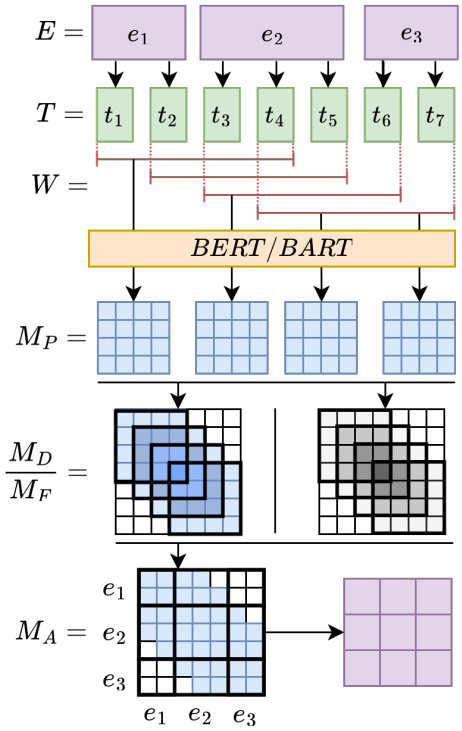


Figure 6.3: An illustration of the sliding window approach proposed in Huber and Carenini (2022). e_n are `EDUs`; t_n are sub-word tokens; red lines are input text spans with maximum length limit (t_{max}); M_P is partial square self-attention matrix (size $t_{max} \times t_{max}$); M_D is document-level matrix obtained from M_P matrices' addition; M_F is frequency matrix that tracks the number of overlaps in M_D; M_A is frequency normalized attention matrix.

The authors utilize original BERT-base and BART-large models for the discourse inference, and extend to seven fine-tuned versions that include tasks such as sentiment analysis, natural language inference, summarization, and question-answering. Since this is not a probing task,

Supervision	Model	RST-DT		GUM		STAC		
		Architecture	Span	Architecture	Span	Architecture	Link	L&R
Monologue RST-style parsing								
Sup	Wang et al. (2017b)	Two-stage	72.0†	Two-stage	58.6†	-	-	
Inter-d sup	Huber and Carenini (2022)	Two-stage$_{\text{GUM}}$	65.4	Two-stage$_{\text{RST-DT}}$	54.0	-	-	
Self-sup	Huber and Carenini (2022)	BERT	35.7	BERT	33.0	-	-	
		BART+CNN-DM	39.1	BART+CNN-DM	32.7	-	-	
Dialogue SDRT-style parsing								
Sup	Shi and Huang (2019)	-	-	-	-	Hierarchical	71.4*	55.7
	Chi and Rudnicky (2022)	-	-	-	-	Structured	74.4	59.6
Inter-d sup	Liu and Chen (2021)	-	-	-	-	Hierarchical	48.3	26.6
	Chi and Rudnicky (2022)	-	-	-	-	Structured	50.6	31.6
Semi-sup	Li et al. (2023)	-	-	-	-	BART	57.6	-
		-	-	-	-	BART+SO-STAC	59.5	38.6‖

Table 6.2: Performance of SOTA supervised models (Sup), supervised models with inter-domain integration (Inter-d sup), and self-/semi- supervised models (self-/semi- sup) discourse parsing. Evaluation metric for monologues (RST-DT and GUM) is original parseval (span), for dialogue is micro-F_1 for both link attachment (Link) and link & relation prediction (L&R).
†: result taken from Huber and Carenini (2022). *: taken from Li et al. (2023). ‖: result from Chapter 8. PLM+x: language model fine-tuned on task x. CNN-DM: news summarization task. SO-STAC: sentence ordering task trained on STAC. - means not applicable.

no additional layers are added. The authors input the `EDUs` into the language model, extract all attention matrices, and convert them into potential discourse trees using the CKY and Eisner algorithms. They then analyze each self-attention matrix individually and compare their alignment with discourse information.

The authors conduct experiments on GUM (Zeldes, 2017) and RST-DT (Carlson et al., 2002a) datasets, and compare the results with chain baselines (left-branching and right-branching) as well as a distant-supervised models in Xiao et al. (2021). The performance of the self-supervised approach, using BERT-base and BART-large, is much lower compared to the supervised model presented in Wang et al. (2017b) (with a gap of 20 points). Nevertheless, discourse structures inferred by the PLMs outperform the chain baselines by a large margin (greater than 10), and exhibit significant improvement when compared to those inferred from neural summarizers (Xiao et al., 2021).

The analysis reveals that the higher layers of the models capture mostly constituent structures, whereas dependency structures are more evenly distributed throughout the layers. The behavior of original and fine-tuned LMs is similar, indicating that both pre-trained and fine-tuned LMs can effectively capture discourse information. Interestingly, the study finds that over 16% of the correctly predicted dependency structures are not captured by supervised models, indicating that PLMs capture some complementary information. Overall, the captured discourse information is found to be both local and general, and consistent with the information obtained from supervised models.

In Chapter 3, we have discussed supervised discourse parsing models. Now, with the introduction of self-supervised discourse parsing, it is interesting to compare the performances of different learning strategies. To provide a clear comparison of model performances and the applicability of different learning strategies, we present a brief comparison of supervised, inter-domain supervised, and self-supervised learning in discourse parsing in Table 6.2. The top part of the table is

reserved to monologues, while the bottom part is for dialogues. As a teaser, we also include the results of our semi-supervised methods for naked structure extraction and full discourse parsing for dialogues in the last two lines of the table, which we will discuss in detail in Chapter 7 and Chapter 8. As expected, self-supervised models achieve lower results than supervised models in both monologue and dialogue settings, highlighting the difficulty of self-supervised learning. Notably, inter-domain supervised models also under-perform supervised models by a considerable margin ($4 - 7\%$ for monologues, $> 20\%$ for dialogues), especially in the dialogue setting, which indicates the limited generalization capacity of supervised models. Interestingly, in the dialogue setting, our semi-supervised model outperforms the inter-domain supervised model by a significant margin (for link attachment: 57.6 vs 50.6 and 48.3; for link+rel: 38.6 vs 31.6 and 26.6), suggesting that our proposed strategies could perform better than inter-domain integration in scenarios where no / few annotated data is available.

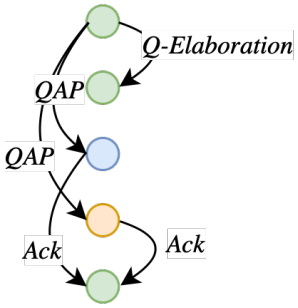
To conclude, in this chapter, we provide an overview of pre-trained language models and preview our upcoming work related to these models. We begin by tracing the evolution from word embeddings to language models and discussing the study of BERTology while focusing on discourse information exploration. These studies show that discourse information is encoded in PLMs, but the challenge lies in how to extract it and enhance its presence. While some studies explore the possibility of extracting discourse structure from PLMs (Wu et al., 2020; Huber and Carenini, 2022), none of them test on dialogues. In the next chapter, we continue our discussion on PLMs and discourse, specifically on the discourse structure extraction in dialogues.

Chapter 7

Naked Discourse Structure Extraction from PLMs

Contents

Our main focus in this chapter is the automatic extraction of the naked structures in dialogues, using SDRT-annotated corpus STAC (Asher et al., 2016). In STAC corpus, discourse structures are represented as dependency graphs with arcs linking `EDUs` and semantico-pragmatic relations (e.g. *Acknowledgment*, *Contrast* or *Question-Answer Pair*), as shown in the following example:



e_1 dmm: I can give a sheep or wood for a wheat.

e_2 dmm: Any takers?

e_3 inca: Sheep would be good.

e_4 cheshireCatGrin: Not here.

e_5 dmm: Okay.

As we know, data scarcity has always been an issue for discourse parsing, especially in dialogues. In Chapter 6, we investigate a few BERTology studies related to discourse. Notably, large language models such as BERT (Devlin et al., 2019a) and BART (Lewis et al., 2020) may encode discourse structure information to a certain extent, as evidenced by Koto et al. (2021); Pandia et al. (2021). Our hypothesis is that the attention matrices in these models can capture the dependency relations between `EDUs`, and that fine-tuning tasks related to discourse can further enhance this information. This is supported by recent research conducted on monologues (Huber and Carenini, 2022). However, there are still several open questions that need to be addressed, such as which PLMs to use, what discourse-related fine-tuning tasks to employ, how to extract dependency structure from attention matrices, and how to identify the most "discourse-rich" attention heads. We intend to provide answers to all of these questions in this chapter.

This chapter is adapted from one publication at the 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2023) (Li et al., 2023). It is organized as follows: in Section 7.1, we provide an overview of related studies in discourse parsing, with a focus on semi-supervised and unsupervised methods. By considering these studies, we can infer why they are not readily applicable for our objectives. Our method of structure extraction is then presented in Section 7.2. For the critical task of attention head selection, we propose both semi-supervised and unsupervised strategies. Results obtained on the STAC corpus and detailed analysis are presented in Section 7.4 and Section 7.5. In addition, we conduct experiments on the GUM corpus (Zeldes, 2017) specifically on its conversation segment, and the results are reported in Section 7.6. While most prior research on discourse parsing begins with manually segmented `EDUs`, this approach is not practical. We take a step further and use predicted `EDUs` instead, and discuss the deployed results in Section 7.7. Finally, we present efforts in extending tree structure to graph structure in Section 7.8, and we conclude the chapter in Section 7.9.

7.1 Overview of Discourse Parsing Methods

As presented in Chapter 3, early approaches to discourse parsing on STAC use supervised methods with varied decoding strategies (Section 3.2.1), such as Maximum Spanning Tree algorithm (Muller et al., 2012; Li et al., 2014c; Afantenos et al., 2012b) or Integer Linear Programming (Perret et al., 2016). Shi and Huang (2019) first proposed a neural architecture based on hierarchical Gated Recurrent Unit (GRU) which processes segment attachment and relation allocation

sequentially. They reported 73.2% F_1 on STAC for naked structures. Recently, Wang et al. (2021a) adopted Graph Neural Networks (GNNs) and reported marginal improvements for link prediction (73.8% F_1). Chi and Rudnicky (2022) also adopted GNN structure but with a joint framework for structure and relation prediction, their model achieved a F score at 74.4%.

Lately, a new trend towards semi-supervised and unsupervised discourse parsing has emerged, primarily because of the problem of data scarcity. However, this trend has been mostly restricted to monologues (Section 3.2.2, 3.2.3). In RST framework: Huber and Carenini (2019, 2020b) leveraged sentiment information and showed promising results in cross-domain settings with the silver-standard labeled corpus. Xiao et al. (2021) extracted discourse trees from neural summarizers and confirmed the existence of discourse information in self-attention matrices. Although these studies are intriguing, their effectiveness is yet to be proven in dialogue settings. For example, in Huber and Carenini (2019), sentiment information was utilized. They smoothed document-level sentiment to sentence-level sentiment and attention scores through the Multiple-Instance Learning (Angelidis and Lapata, 2018) strategy, and subsequently employed local attention scores to construct discourse trees. While sentiment-annotated monologues are prevalent, such as in food and movie reviews, it is challenging to assign global sentiment labels for dialogues, since different speakers may have different emotions, making it almost impossible to establish a "unified tone" for a dialogue. In the case of summarization tasks, as only vital information is extracted, it is unclear how the remaining parts of the documents interact with each other. In a dependency-tree-style discourse structure, our aim is not to build a hierarchical tree but to create *flat* connections among all the `EDUs`. The information leveraged solely from summarization is also unsuitable for our purposes.

Another line of work proposed to enlarge training data with a combination of several parsing models, as done in Jiang et al. (2016); Kobayashi et al. (2021); Nishida and Matsumoto (2022). In a fully unsupervised setting, Kobayashi et al. (2019) used similarity and dissimilarity scores for discourse tree creation, a method that can not be directly used for discourse graphs though. As for dialogues, transfer learning approaches are rare. Badene et al. (2019a,b) investigated a weak supervision paradigm where expert-composed heuristics, combined with a generative model, are applied to unseen data. Their method, however, requires domain-dependent annotation and a relatively large validation set for rule verification. Still, it suffers from low recall due to uneven coverage of various linguistic phenomena. Another study by Liu and Chen (2021) focused on cross-domain transfer using STAC (chats in a game) and Molweni (chats in Ubuntu forum) for training and testing interchangeably. They applied simple adaptation strategies (mainly lexical information) on a SOTA discourse parser (Shi and Huang, 2019) and show improvement compared to bare transfer (train on Molweni and test on STAC F_1 increase from 42.5% to 50.5%). Yet, their model failed to surpass simple baselines.

Very recently, Nishida and Matsumoto (2022) proposed unsupervised methods for domain adaptation in discourse parsing (Section 3.2.4). They investigated bootstrapping methods to adapt pre-trained BERT-based parsers to out-of-domain data with some success. Although effective, their method mandates pre-training discourse parsers on a comparatively large hand-annotated domain (in their instance, they used nearly 900 STAC documents for training), which is not applicable to our goals. Additionally, their method necessitates well-tuned confidence measures and exact sample selection criteria.

In Chapter 6, we present the latest BERTology research on discourse study. Our approach is largely inspired by Huber and Carenini (2022)'s work, where authors introduced a novel way to encode long documents and explored the effect of different fine-tuning tasks on PLMs, confirming that pre-trained and fine-tuned PLMs both can capture discourse information. However, this study differs from our research in two aspects. Firstly, it primarily focuses on discourse parsing in

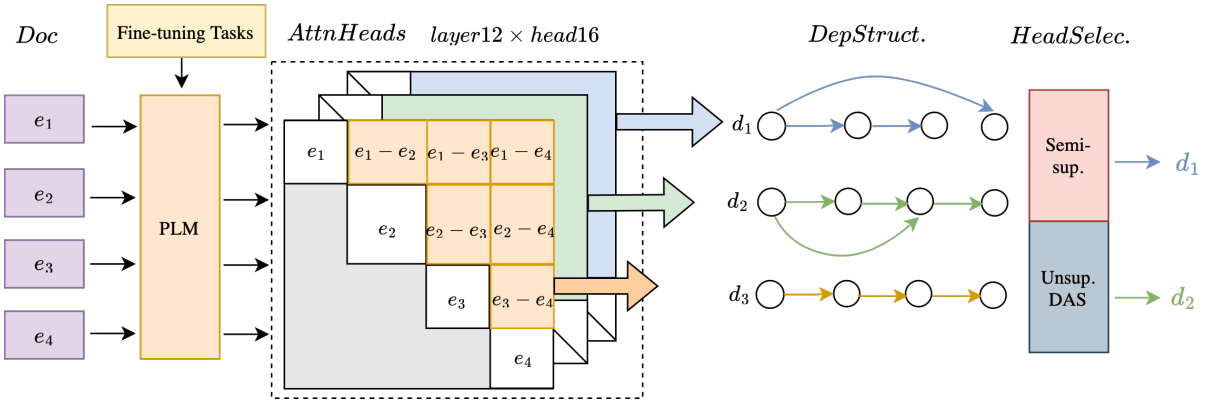


Figure 7.1: Pipeline for discourse structure extraction.

monologues, rather than dialogues. Secondly, it does not address the question of how to identify the attention heads that are rich in discourse information. Our objective, on the other hand, is not just to demonstrate the existence of discourse information, but also to present effective methods for its extraction.

After reviewing the previous work, we identify a gap in the field of discourse structure extraction in dialogues, particularly with regards to semi-supervised and unsupervised methods. Our proposed solution aims to address this gap.

7.2 Method: From Attention Matrix To Discourse Tree

7.2.1 Problem Formulation and Simplifications

Figure 7.1 shows the overview of the pipeline: given a dialogue with n *Elementary Discourse Units* (`EDUs`), which are the minimal spans of text (mostly clauses, at most a sentence) to be linked by discourse relations: $D = \{e_1, e_2, e_3, ..., e_n\}$, the goal is to extract a Directed Acyclic Graph (DAG) connecting the n `EDUs` that best represents its SDRT discourse structure from attention matrices in PLMs. We conduct an extensive investigation of every attention head in the PLM to obtain a vast number of potential structures, as represented by d_1, d_2, and d_3 in the figure. Subsequently, we employ semi-supervised and unsupervised methods to identify the most effective attention head for extracting the discourse structure. In this study, we make a few simplifications, partially adopted from previous work.

(1) We do not deal with SDRT *Complex Discourse Units* (`CDUs`) attachments following Muller et al. (2012); Afantenos et al. (2015). In Muller et al. (2012), the algorithm of `CDU` transformation is as follows: the head of a `CDU` is the highest in its subgraph and leftmost `DU` in the discourse if there is more than one. The algorithm of finding the head is recursive until an `EDU` is reached. Initially, we address the extraction of *flat* dependency structure, with the subsequent task of deducing hierarchical structure left for further exploration.

(2) Similar to Shi and Huang (2019), our solution only generate projective discourse trees. Projective trees contain no crossing edges. In STAC, we observe ≈ 6% of non-projective edges. Our approximation of projective trees is thus reasonable for the initial step. Approximately 5% of the nodes have multiple incoming edges, while the remaining nodes follow the single-parent principle observed in tree structures. We outline methods to enhance our

tree algorithm by introducing additional edges to generate graph structures in Section 7.8. Although the enhancement may not result in significant improvements (at most ≈ 1 point), it is a valuable effort to generate real SDRT-graph structures.

(3) We break down the discourse parsing task into two steps. This chapter specifically deals with the structure extraction step, while the assignment of relation types will be addressed in Chapter 8.

7.2.2 Which Kinds of PLMs to Use?

7.2.2.1 Pre-Trained Models

We explore both vanilla and fine-tuned PLMs, as they are both shown to contain discourse information for monologues (Huber and Carenini, 2022). We choose BART (Lewis et al., 2020) as our pre-trained language model for two reasons.

Firstly, BART exhibits exceptional abilities in comprehending complex structures due to its pre-training objectives. BART is a large language model based on the standard sequence-to-sequence Transformer architecture (Vaswani et al., 2017). Similar to BERT (Devlin et al., 2019a), it employs bidirectional encoders, and like GPT (Radford et al., 2018), it has autoregressive decoders. What sets BART apart is its training method that involves corrupting documents with various noised transformations, including token masking (similar to BERT), sentence permutation, document rotation, and text infilling (akin to SpanBERT (Joshi et al., 2019)). Comparison of different pre-training objectives shows that the noising techniques in BART surpass those used in other PLMs such as BERT, GPT, and XLNet. BART has exhibited excellent performance on various downstream tasks, particularly in dialogue task ConvAI2 (Dinan et al., 2020), outperforming previous work on conversational response generation by a significant margin. Given its superior performance on dialogue tasks, we believe that it is well-suited to our purposes. BART model contains three kinds of attention matrices: encoder, decoder and cross attention. We use the encoder attention in this work to derive discourse trees, since it has been shown to capture most discourse information (Koto et al., 2021; Huber and Carenini, 2022) and outperformed the other alternatives in preliminary experiments on a validation set. We present the method to generate trees from attention heads in the following Section 7.2.3.

Secondly, we conduct preliminary experiments to compare the performance of BART against other alternatives, including DialoGPT (Zhang et al., 2020) and DialogLM (Zhong et al., 2022), which are pre-trained with conversational data. Our results show that BART outperformed these models. In Section 7.4.3, we present additional results on the performance of further language models and provide our reasoning for why they do not perform as well as BART.

Although we mainly test on BART model, please note that our approach is generally model agnostic and can be applied to any transformer-based architecture.

7.2.2.2 Fine-Tuning Tasks

We fine-tune BART on three discourse-related tasks. The first task document summarization is inspired from the promising results in monologues (Huber and Carenini, 2022). It is also one of the synergistic tasks for discourse parsing, as mentioned in Section 7.1. The second task is question answering (QA). In Chapter 3, we present a multi-task learning framework (He et al., 2021), where the authors jointly learn QA-based machine reading comprehension and discourse parsing. In order to perform the QA task correctly, models need to handle different discourse relationships in plain text. This intuitively strengthens the process of structure understanding in

dialogue data. The final task is our proposal sentence ordering, which considers the specificities of dialogues and doesn't require any extra human annotation.

Summarization: We use BART fine-tuned on the popular CNN-DailyMail (CNN-DM) news corpus (Nallapati et al., 2016), which gives the biggest increase of discourse performance compared to pre-trained model in Huber and Carenini (2022). Since BART has no pre-training dialogue data, we wish to improve the model's performance on dialogues and fine-tune it on an abstractive dialogue summarization dataset SAMSum (Gliwa et al., 2019).

Question-Answering (QA): Using question answering as a machine reading comprehension task is an effective way to assess a model's ability to comprehend the relationships between speech turns in dialogues (He et al., 2021). One popular type of question answering is span-based QA (Rajpurkar et al., 2016, 2018), which requires the model to extract a continuous text span from the original dialogue. To enhance BART's capability in capturing relational structures, we fine-tune it on the most recent version of the Stanford Question Answering Dataset (SQuAD 2.0) (Rajpurkar et al., 2018).

Sentence Ordering: We fine-tune BART on the sentence ordering task, reordering a set of shuffled sentences to their original order. This task is challenging, especially for long documents. According to a state-of-the-art model for sentence ordering (Chowdhury et al., 2021), the authors found that as the length of a document increases from 5 to 20 sentences, the model's performance drops from above 80 to less than 40 (measured in accuracy), and this trend holds for various types of documents, including scientific papers and narratives. In the case of STAC, the average number of speech turns is 13, making it a relatively difficult scenario for this task. Additionally, we observe in this study that the model's performance is affected by the "effect of shuffling" - a metric defined by the minimum number of swaps needed to reconstruct the ordered sequence. When a shuffled document is significantly different from the original one, the model perceives the task as more difficult and performs poorly. Conversely, a lower degree of shuffling results in a more coherent and meaningful input, leading to an easier task. Therefore, we devise various shuffling methods to control the effect of shuffling and ensure a more gradual and effective learning process. Specifically, as shown in Figure 7.2, we explore:

(a) *partial-shuf*: randomly picking 3 utterances (2 for short dialogues with less than 4 utterances) in a dialogue and shuffling them. This permutation is supposed to be the easiest one since we keep the most of the context unchanged.

(b) *minimal-pair-shuf*: shuffling minimal pairs, comprising of a pair of speech turns from 2 different speakers with at least 2 utterances. A speech turn represents the beginning of a new speaker. We shuffle these pairs with respect to the original order inside the pair. This shuffling is more difficult than *partial-shuf* with a larger shuffling effect. "Local" contexts are supposed to be coherent, and the model needs to find the inconsistency in larger contexts.

(c) *block-shuf*: shuffling a block containing multiple speech turns. We divide one dialogue into $[2, 5]$ blocks based on the number of utterances and shuffle between blocks. Block size is designed to be as twice or 3 times bigger than "min-pair", we thus set criteria aiming to have ≈ 6 `EDUs` per block: $|utt.| < 12 : b = 2$, $|utt.| \in [12, 22] : b = 3$, $|utt.| \in [22, 33] : b = 4$, $|utt.| \geq 33 : n = 5$. This shuffling method also emphasizes on maintaining consistent local contexts, similar to *minimal-pair-shuf*. However, we increase the range of local context and aim to make the sentence reordering task easier.

(d) *speaker-turn-shuf*: grouping all speech productions of one speaker together. The sorting task consists of ordering speech turns from different speakers' production. This shuffling

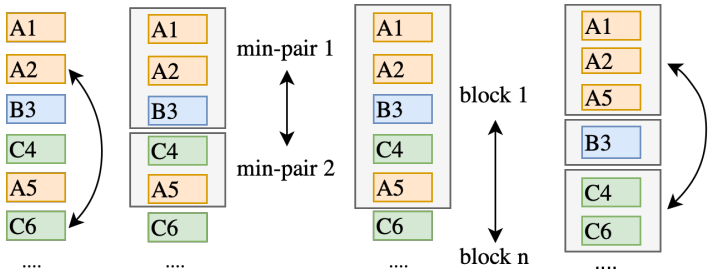


Figure 7.2: Sentence Ordering shuffling strategies (from left to right: partial, minimal-pair, block, speaker-turn) on a sequence of utterances 1 to 6, with A, B, C as the speakers.

strategy aims to capture the interaction between dialogue participants by keeping the consistency of speeches from the same speaker. It requires the model to learn how to maintain coherence in the interaction among speakers.

We evenly combine all permutations mentioned above to create our **mixed-shuf** data set and conduct the SO task as the third auxiliary task to fine-tune BART. We can see that every shuffling strategy poses a unique emphasis of coherence and presents its own level of difficulty. We are of the opinion that incorporating diverse permutations can enhance the model's ability to understand the structure in dialogues. The initial subpar test results obtained using random shuffling lend further support to our proposal of mixed stuffing.

7.2.3 How To Derive Trees From Attention Heads?

Given an attention matrix $A^t \in \mathbb{R}^{k \times k}$ where k is the number of tokens in the input dialogue, we derive the matrix $A^{edu} \in \mathbb{R}^{n \times n}$, with n the number of `EDUs`, by computing $A^{edu}(i, j)$ as the average of the submatrix of A^t corresponding to all the tokens of `EDUs` e_i and e_j, respectively. As a result, A^{edu} captures how much `EDU` e_i depends on `EDU` e_j and can be used to generate a tree connecting all `EDUs` by maximizing their dependency strength. Concretely, we find a Maximum Spanning Tree in the fully-connected dependency graph A^{edu} using the Eisner algorithm (Eisner, 1996). Conveniently, since an utterance cannot be anaphorically and rhetorically dependent on following utterances in a dialogue, as they are previously unknown (Afantenos et al., 2012b), we can further simplify the inference by applying the following hard constraint to remove all backward links from the attention matrix A^{edu}: $a_{ij} = 0, \text{if } i > j$.

We present an example in Figure 7.3 to illustrate our approach. The document consists of three `EDUs`, with the first containing two tokens, the second containing three, and the third containing two. We input these tokens into BART to obtain a token-level attention matrix A^t with dimensions of 7×7. To obtain A^{edu} of size 3×3, we average the attention scores within each `EDU` to form sub-matrices, denoted by bold-line borders. By imposing the *forward-link* constraint, we obtain a half-matrix highlighted in blue. Finally, we apply the Eisner algorithm to this half-matrix.

7.2.4 How To Find the Best Heads?

Pioneering work led by Raganato and Tiedemann (2018) showed that specific attention heads mark different syntactic and semantic dependency relations. Authors confirmed that higher layers tend to encode more semantic information. Recently, Xiao et al. (2021) and Huber and Carenini (2022) showed that discourse information is not evenly distributed between heads and

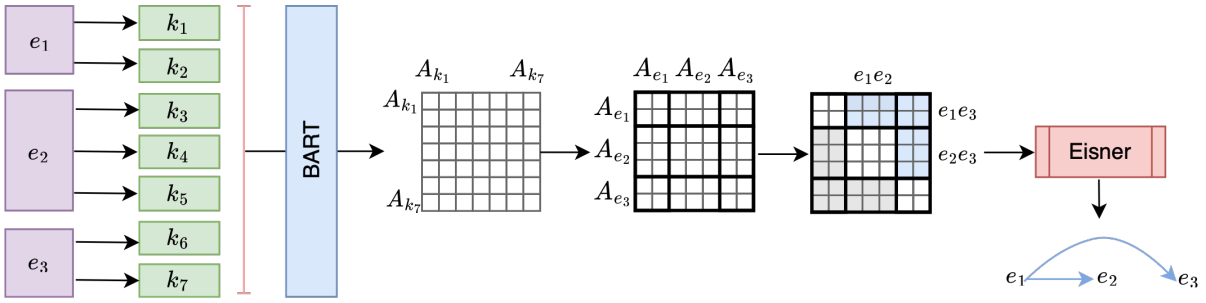


Figure 7.3: An illustration of dependency tree extraction from attention matrix. e_n are `EDUs`; k_n are sub-word tokens in corresponding `EDUs`; A_k is a token-level attention matrix; A_e is an `EDU`-level attention matrix; e_ie_j is the attention score between `EDUs` e_i and e_j. Only the upper-right part of the attention matrix (in blue) is utilized for MST calculation.

layers. Precisely, the authors in Xiao et al. (2021) utilized the average attention scores across all heads in a layer, and found that there were differences in performance across layers within the same model. They tested a 2-layer 1-head model and a 6-layer 8-head model. When comparing these two models, they observed that the performance gap between layers decreased with more layers, possibly due to the information being distributed across different layers. It is, however, unclear which layer contains more discourse information, as the performance is not consistent among different testing corpora. In contrast, instead of aggregating attention scores across layers, Huber and Carenini (2022) inspected each attention head in all layers. They discovered the "locality" property across different fine-tuned LMs and that higher layers predominantly capture constituency tree structures, whereas dependency structures are more uniformly distributed across layers. In our study, we observe that dependency discourse structures are consistently located in deeper layers (Section 7.5.1), which is consistent with the findings in (Raganato and Tiedemann, 2018).

The previous studies mentioned are insightful, but they do not provide any approach to identify the head or heads containing the most discourse information. To address this issue, we propose two effective methods for selecting the most discourse-rich heads with minimal supervision, including a fully unsupervised and a semi-supervised method. Our goal is to conduct a comprehensive investigation of the encoder representation by analyzing both head-wise and layer-wise attention heads.

In our unsupervised approach, we follow the approach in Huber and Carenini (2022) by examining each attention head individually and distinguish between the local and global best head (refer to Section 7.2.4.1). As for our semi-supervised approach, we use a few annotated examples to select the heads with relatively more discourse information. We conduct head-wise examination and further layer-wise aggregation, similar to Xiao et al. (2021) (Section 7.2.4.2).

7.2.4.1 Unsupervised Best Head Selection

Dependency Attention Support Measure (DAS): Loosely inspired by the confidence measure in Nishida and Matsumoto (2022), where the authors define the confidence of a *teacher* model based on predictive probabilities of the decisions made, we propose a DAS metric measuring the degree of support for the maximum spanning (dependency) tree (MST) from the attention matrix. Formally, given an attention matrix A^g (i.e., A^{edu} for the dialogue g) with n `EDUs`, the MST T^g is built by selecting $n-1$ attention links l_{ij} from A^g based on the tree

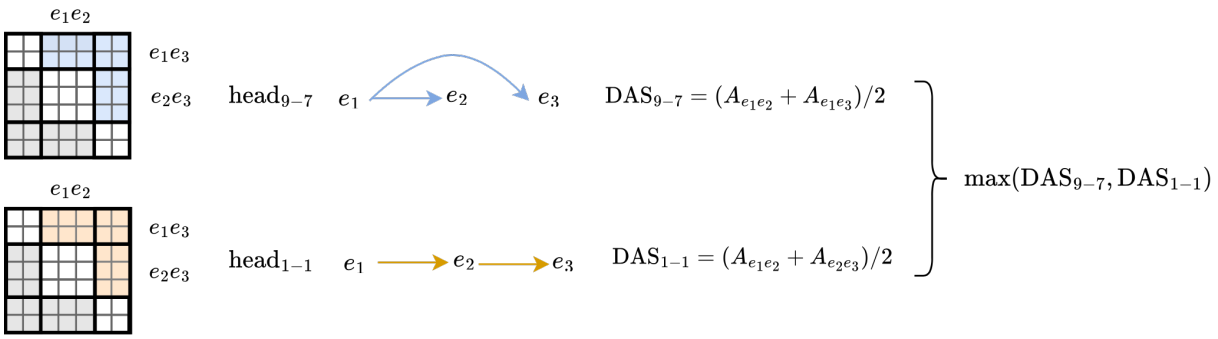


Figure 7.4: An illustration of DAS calculation. Head_{9-7} and head_{1-1} are two example heads.

generation algorithm.

Please note that DAS can be easily adapted for a general graph by removing the restriction to $n-1$ arcs. DAS measures the strength of all those connections by computing the average score of all the selected links:

$$DAS(T^g) = \frac{1}{n-1} \sum_{i=1}^{n} \sum_{j=1}^{n} Sel(A^g, i, j) \tag{7.1}$$

with $Sel(A^g, i, j) = A^g_{ij}$, if $l_{ij} \in T^g$, 0 otherwise.

For illustration, we present the calculation of DAS score of two imaginary heads layer9-head7 and layer1-head1 in Figure 7.4. With attention scores in the highlighted region, the Eisner algorithm gives two different tree structures, from which we can calculate two DAS scores.

Selection Strategy: With DAS, we can now compute the degree of support from each attention head h on every single example g for the generated tree $DAS(T^g_h)$. We therefore propose two strategies to select attention heads based on the DAS measure, leveraging either global or local support. The **global** support strategy selects the head with the highest averaged DAS score over all the data examples:

$$H_{global} = \arg\max_h \sum_{g=1}^{M} DAS(T^g_h) \tag{7.2}$$

where M is the number of examples. In this way, we select the head with a generally good performance on the target dataset.

The second strategy is more adaptive to each document, focusing only on the **local** support. It does not select one specific head for the whole dataset, but instead selects the head/tree with the highest support for every single example g, i.e.,

$$H^g_{local} = \arg\max_h DAS(T^g_h) \tag{7.3}$$

7.2.4.2 Semi-Supervised Best Head / Layer Selection

Here we consider both head-wise and layer-wise selection using a few annotated examples. In conformity with real-world situations where labeled data is scarce, we sample three small subsets with $\{10, 30, 50\}$ data points (i.e., dialogues) from the validation set. For **layer-wise** attention

Dataset	#Doc	#Utt/doc	#Tok/doc	#Spk/doc	Domain
DailyDialog	13,118	13	119	2	Daily
STAC	1,161	11	50	3	Game
GUM-conv	9	209	1,161	3	Daily

Table 7.1: Key statistics of datasets. Utt = sentences in DailyDialog or `EDUs` in STAC and GUM; Tok = tokens; Spk = speakers.

matrices, we average 16 attention heads for every layer which gives 12 candidate layers. For **head-wise** attention matrices, we take each attention matrix individually which results in 12 layers $\times$16 heads (= 192) candidate matrices for each dialogue. Then, the head with the highest micro-F_1 score (the best layer ("L") and the best head ("H")) on the validation set is selected to derive trees in the test set.

7.3 Experimental Setup

7.3.1 Datasets

We use the multi-party dialogue STAC corpus[1] (Asher et al., 2016), annotated following the SDRT framework, to evaluate our approach on the discourse dependency structure prediction task. Including 300 strategic conversations of players trading goods during the board game *The Settlers of Catan*, this corpus contains some high-frequency game-related words such as *sheep, clay* and *wood*.

To evaluate a variety of fine-tuned PLMs (see Section 7.2.2), we use publicly available HuggingFace models for the summarization and question-answering tasks. For the newly proposed sentence ordering (SO) task, we train the BART model on two dialogue datasets: (1) the STAC corpus itself (raw text), in line with the final structural evaluation, however, limiting the input data to plain texts. (2) DailyDialog (Li et al., 2017), a human-written corpus covering various topics for English learners (10 categories), from ordinary life to finance. We select this corpus due to its large size, diversity of topics and high quality. We summarize the key dataset statistics for STAC and DailyDialog in Table 7.1. STAC has a separation of 82%, 9%, 9% for train, validation, and test sets resp.; DailyDialog 85%, 8%, 8%. Focusing on the STAC corpus in our main evaluation, we report additional results on the *conversational subset* of GUM (Zeldes, 2017) in Section 7.6. We purposely exclude the Molweni corpus (Li et al., 2020) in this work, due to major quality issues found in preliminary dataset exploration in Section 2.3.5.

7.3.2 Baselines and Supervised Dialogue Discourse Parsers

We compare against the simple yet strong unsupervised yet powerful LAST baseline (Schegloff, 2007), attaching every `EDU` to the previous one. Furthermore, to assess the gap between our

[1]Precision on STAC corpus: the STAC project includes two corpora, the *linguistic only* corpus which only contains textual conversation information, and the *situated* corpus which includes conversational texts as well as descriptions of server messages and UI logs (Badene, 2021). In our experiments, we utilize the *linguistic only* corpus. Since STAC has been updated several times, not all the studies have employed the same version. Shi and Huang (2019) released the version for their experiments: `https://github.com/shizhouxing/DialogueDiscourseParsing`. This version has been used in subsequent studies, including Wang et al. (2021b); Liu and Chen (2021); Fan et al. (2022); Yu et al. (2022); Chi and Rudnicky (2022). For all of our experiments utilizing the STAC corpus in this thesis, we use the *shi2019* version.

Model
BART-large `https://huggingface.co/facebook/bart-large`
BART-large-cnn `https://huggingface.co/facebook/bart-large-cnn`
BART-large-samsum `https://huggingface.co/linydub/bart-large-samsum`
BART-large-finetuned-squad2 `https://huggingface.co/phiyodr/bart-large-finetuned-squad2`
RoBERTa-large `https://huggingface.co/roberta-large`
DialoGPT-small `https://huggingface.co/microsoft/DialoGPT-small`
DialogLED-large-5120 `https://huggingface.co/MingZhong/DialogLED-large-5120`

Table 7.2: Huggingface models and URLs.

approach and supervised dialogue discourse parsers, we compare with the *Deep Sequential* model by Shi and Huang (2019) and the Structure Self-Aware (SSA) model by Wang et al. (2021a).

7.3.3 Evaluation Metrics

We report the micro-F_1 for discourse parsing and the Unlabeled Attachment Score (UAS) for the generated naked dependency structures.

7.3.4 Implementation Details

We base our work on the transformer HuggingFace library (Wolf et al., 2020) and follow the *text-to-marker* framework proposed in Chowdhury et al. (2021) for the SO fine-tuning procedure. We use the original separation of train, validation, and test sets; set the learning rate to $5e-6$; use a batch size of 2 for DailyDialog and 4 for STAC, and train for 7 epochs. All other hyper-parameters are set following Chowdhury et al. (2021). We do not do any hyperparameter tuning. We omit 5 documents in DailyDialog during training since the document lengths exceed the token limit. We replace speaker names with markers (e.g. Sam $\rightarrow$ "spk1"), following the preprocessing pipeline for dialogue utterances in PLMs. Table 7.2 shows the models and the sources we obtained from Huggingface library.

7.4 Results

7.4.1 Unsupervised Head Selection

Results using our novel unsupervised DAS method on STAC are shown in Table 7.3 for both the global (H_g) and local (H_l) head selection strategies. These are compared to: (1) the unsupervised LAST baseline (at the top), which only predicts local attachments between adjacent `EDUs`. LAST is considered a strong baseline in discourse parsing (Muller et al., 2012), but has the obvious disadvantage of completely missing long-distance dependencies which may be critical in downstream tasks. (2) The supervised *Deep Sequential* parser by Shi and Huang (2019) and Structure Self-Aware model by Wang et al. (2021a) (center of the table), trained on STAC,

Model			
Unsupervised Baseline			
LAST			56.8
Supervised Models			
Deep-Sequential (Shi and Huang, 2019)			71.4
SSA-GNN (Wang et al., 2021a)			73.8
Unsupervised PLMs	H_g	H_l	H_{ora}
BART	56.6	56.4	57.6
+ CNN	56.8	56.7	57.1
+ SAMSum	56.7	56.6	57.6
+ SQuAd2	55.9	56.4	57.7
+ SO-DD	56.8	57.1	58.2
+ SO-STAC	56.7	**57.2**	59.5

Table 7.3: Micro-F_1 on STAC for supervised SOTA models and PLMs. H_g: global best head. H_l: local best heads. H_{ora}: oracle head. Best (non-oracle) score in the 3^{rd} block in bold.

reaching 71.4% and 73.8% in F_1, respectively. We re-train the *Deep Sequential* model using the released code. The obtained scores are slightly lower as in the paper, a similar observation is reported in Wang et al. (2021a).

In the last sub-table of Table 7.3, we show unsupervised scores from pre-trained and fine-tuned LMs on three auxiliary tasks: summarization, question-answering, and sentence ordering (SO) with the mixed shuffling strategy. We present the global head (H_g) and local heads (H_l) performances selected by the DAS score (see Section 7.2.4.1). The best possible scores using an oracle head selector (H_{ora}) are presented for reference.

Comparing the values in the bottom sub-table, we find that the pre-trained BART model under-performs LAST, with global and local heads achieving similar performance. Noticeably, models fine-tuned on the summarization task ("+CNN", "+SAMSum") and question-answering ("+SQuAD2") only add marginal improvements compared to BART. In the last two lines of the sub-table, we explore our novel sentence ordering fine-tuned BART models. We find that the BART+SO approach, trained on DailyDialog (DD) and STAC itself, surpasses LAST when using local heads. As commonly the case, the intra-domain training performs best, which is further strengthened in this case due to the special vocabulary in STAC. Importantly, our PLM-based unsupervised parser can capture some long-distance dependencies compared to LAST (Section 7.5.2). Additional analysis regarding the chosen heads is in Section 7.5.1.

7.4.2 Semi-Supervised Head Selection

While the unsupervised strategy only delivered minimal improvements over the strong LAST baseline, Table 7.4 shows that if a few annotated examples are provided, it is possible to achieve substantial gains. In particular, we report results on the vanilla BART model, as well as BART model fine-tuned on DailyDialog ("+SO-DD") and STAC itself ("+SO-STAC"). We execute 10 runs for each semi-supervised setting ($[10, 30, 50]$) with head-wise ("H") and layer-wise ("L") attention matrices, and report average scores and the standard deviation.

With oracle attention heads (Gold H in the table), all three models achieve superior performance compared to LAST. Further, using a small scale validation set (50 examples) to select the

Train on → Test with ↓	BART F_1	+ SO-DD F_1	+ SO-STAC F_1
LAST BSL	56.8	56.8	56.8
H_{ora}	57.6	58.2	59.5
Unsup H_g	<u>56.6</u>	56.8	56.7
Unsup H_l	56.4	<u>57.1</u>	<u>57.2</u>
Semi-sup 10 L	$55.8_{0.8}$	$55.7_{1.0}$	$55.6_{0.9}$
Semi-sup 30 L	$55.8_{0.6}$	$56.5_{0.4}$	$56.3_{0.4}$
Semi-sup 50 L	$56.2_{0.2}$	$56.4_{0.7}$	$56.4_{0.1}$
Semi-sup 10 H	$57.0_{1.2}$	$57.2_{1.2}$	$57.1_{2.6}$
Semi-sup 30 H	$57.3_{0.5}$	$57.3_{1.3}$	$59.2_{0.9}$
Semi-sup 50 H	$\mathbf{57.4_{0.4}}$	$\mathbf{57.7_{0.5}}$	$\mathbf{59.3_{0.7}}$

Table 7.4: STAC micro-F_1 scores from BART and fine-tuned models with unsupervised and semi-supervised approaches. $\{10, 30, 50\}$ are number of annotated datapoints. H = *head-wise*, L = *layer-wise*. The best semi-supervised score is in bold. Subscription is the standard deviation.

best attention head remarkably improves the F_1 score from 56.8% (LAST) to 59.3% (+SO-STAC) with head-wise attention matrix. F_1 improvements across increasingly large validation-set sizes are consistent, accompanied by smaller standard deviations, as would be expected.

Our results reveal that the performance of the **head-wise** semi-supervised method is consistently better than that of the **layer-wise** method. While the best layer-wise performance is 56.4, slightly underperforming the LAST baseline, the best head-wise performance improves to 59.3. Since different attention heads capture varying amounts of discourse information, averaging them may cancel out the informative cues. This observation suggests that layer-wise aggregation is not an optimal method for extracting discourse information. In contrast, the head-wise results are very encouraging. With only 30 annotated examples, we already achieve performances close to the oracle results, and further improvements can be made with more examples.

7.4.3 Experiments with Other PLMs

To consider pre-trained models with different architectures, we present the results of experiments using RoBERTa (Liu et al., 2019a), a bidirectional encoder model, and DialoGPT (Zhang et al., 2020), an autoregressive decoder model. To account for the influence of training data, we also incorporate DialogLED - DialogLM (Zhong et al., 2022) with Longformer (Beltagy et al., 2020) architecture.

Table 7.5 demonstrates that the decoder-only model has the lowest oracle head performance (56.2), whereas models with encoder networks perform similarly: BART with a score of 57.6, RoBERTa with 57.4, and DialogLED with 57.2. These results are consistent with the findings in Koto et al. (2021), where the authors concluded that RoBERTa and BART are the most effective models in capturing discourse information in their encoder layers.

Despite DialogLED being pre-trained on a large amount of dialogue data, its performance being similar to BART is surprising. Fine-tuning with dialogue data leads to a significant improvement in BART's performance from 57.6 to 59.5, whereas DialogLED's performance only slightly improves from 57.2 to 58.4. This suggests that the effect of our sentence ordering task on DialogLED is less pronounced, likely due to the model's pre-training on dialogue-related permu-

Model		Unsup		Semi-sup		
	H_{ora}	H_g	H_l	Semi10	Semi30	Semi50
BART	57.6	56.6	56.4	$57.0_{1.2}$	$57.3_{0.5}$	$57.4_{0.4}$
+ SO-DD	58.2	56.8	57.1	$57.2_{1.2}$	$57.3_{1.3}$	$57.7_{0.5}$
+ SO-STAC	59.5	56.7	57.2	$57.1_{2.6}$	$59.2_{0.9}$	$\underline{\mathbf{59.3}}_{0.7}$
RoBERTa	57.4	56.8	56.8	$55.6_{1.3}$	$56.8_{0.2}$	$\underline{56.9}_{0.3}$
DialoGPT	56.2	42.7	36.2	$52.9_{4.3}$	$55.1_{1.7}$	$\underline{56.2}_{0.0}$
DialogLED	57.2	56.8	56.7	$54.6_{2.6}$	$54.7_{2.1}$	$\underline{56.6}_{1.9}$
+ SO-DD	57.7	56.4	56.6	$55.0_{2.8}$	$56.1_{2.4}$	$\underline{57.3}_{0.9}$
+ SO-STAC	58.4	56.8	57.1	$57.7_{0.1}$	$\underline{58.2}_{0.5}$	$57.7_{0.1}$

Table 7.5: Micro-F_1 on STAC with other PLMs. H_{ora}: oracle head. H_g: global best head. H_l: local best heads. Best score (except H_{ora}) in each row is underlined.

tation tasks. Additionally, we observe that the high-performing attention heads are located in the deeper layers of DialogLED, similar as in BART, whereas in RoBERTa, they are more uniformly distributed across the layers, and some even appear in the shallow layers. This observation is consistent with the findings in Huber and Carenini (2022).

7.5 Analysis

7.5.1 Effectiveness of DAS

We now take a closer look at the performance degradation of our unsupervised approach based on DAS in comparison to the upper bound defined by the performance of the oracle-picked head. Figure 7.5 shows the DAS score matrices (left) for three models with the oracle heads and DAS-selected heads highlighted in green and yellow, respectively. It becomes clear that the oracle heads do not align with the DAS-selected heads. Making a comparison between models, we find that discourse information is consistently located in deeper layers, with the oracle heads (light green) consistently situated in the same head for all three models, which in line with observations for monologues in Huber and Carenini (2022). However, while not aligning with the oracle, the top-performing DAS heads (in yellow) are among the top 10% best heads in all three models, as shown in the box plot on the right. Hence, we confirm that the DAS method is a reasonable approximation to find discourse intense self-attention heads among the 12×16 attention matrices.

7.5.2 Document and Arc Lengths

The inherent drawback of the simple, yet effective LAST baseline is its inability to predict indirect arcs. To test if our approach can reasonably predict distant arcs of different lengths in the dependency trees, we analyze our results in regard to the arc lengths. Additionally, since longer documents tend to contain more distant arcs, we also examine the performance across different document lengths compared to LAST.

Arc Distance: To examine the discourse parsing performance for data sub-sets with specific arc lengths, we present the UAS score plotted against different arc lengths on the left side in Figure 7.6. Our analysis thereby shows that direct arcs achieve high UAS score ($> 80\%$),

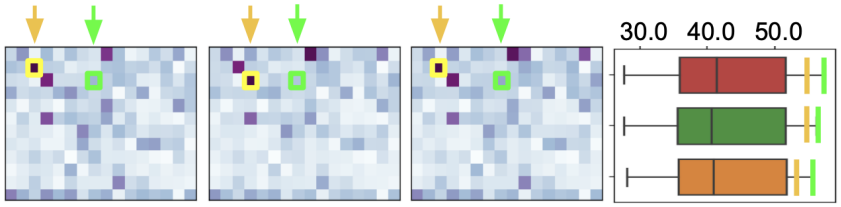


Figure 7.5: Heatmaps: DAS score matrices (layers: top to bottom=12 to 1, heads: left to right=1 to 16) for BART, BART+SO-DD, BART+SO-STAC. Darker purple=higher DAS score. Boxplot: Head-aggregated UAS scores for model BART (orange), BART+SO-DD (green), and BART+SO-STAC (red). Light green=head with highest UAS. Yellow=head with the highest DAS score.

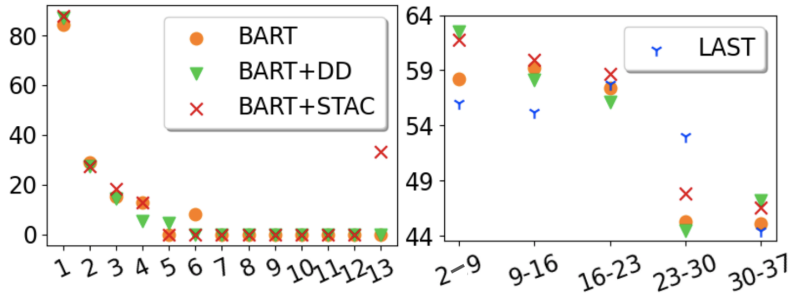


Figure 7.6: Left: UAS and arcs' distance. x axis: arc distance. Right: averaged UAS for different lengths of documents. x axis: #EDUs in a document. y axis: UAS.

independent of the model used. We further observe that the performance drops considerably for arcs of distance two and onwards, with almost all models failing to predict arcs longer than 6. BART+SO-STAC model correctly captures an arc of distance 13. Please note that the presence for long-distance arcs (≥ 6) is limited, accounting for less than 5% of all arcs.

We further analyze the precision and recall scores when separating dependency links into *direct* (adjacent forward arcs) and *indirect* (all other non-adjacent arcs), following Xiao et al. (2021). Precision and recall scores of direct and indirect arcs in the test set are shown in Figure 7.7. For direct arcs, all models perform reasonably good. The precision is higher ($\approx$ +6%) and recall is lower than the baseline (100%), indicating that our models predict less direct arcs but more precisely. For indirect arcs, the best model is BART+SO-STAC (20% recall, 44% prec.), closely followed by original BART model (recall at 20%, precision at 41%).

Document Length: Longer documents tend to be more difficult to process because of the growing number of possible discourse parse trees. Hence, we analyze the UAS performance of documents in regards to their length, here defined as the number of `EDUs`. Results are presented on the right side in Figure 7.6, comparing the UAS scores for the three selected models and LAST for different document lengths. We split the document length range into 5 even buckets between the shortest (2 `EDUs`) and longest (37 `EDUs`) document, resulting in 60, 25, 16, 4 and 4 examples per bucket. We also calculate the LAST baseline for each group, presented in the blue trident.

For documents with less than 23 `EDUs`, all fine-tuned models perform better than LAST, with BART fine-tuned on STAC reaching the best result. We note that PLMs exhibit an increased

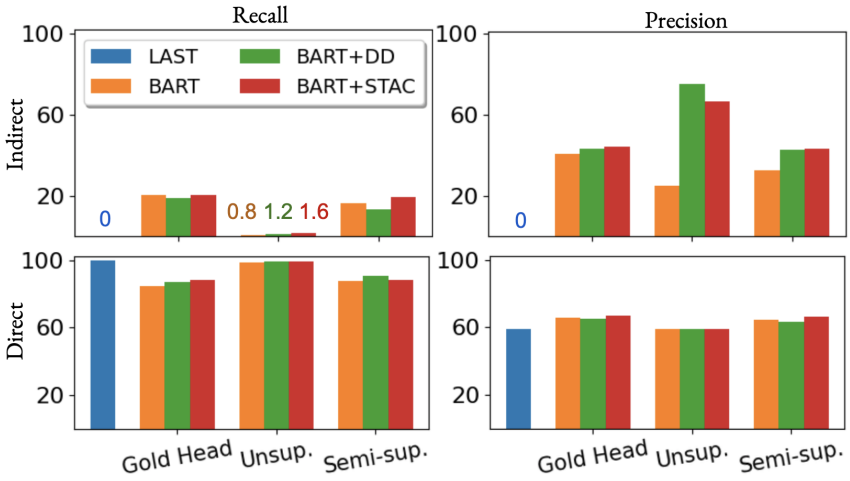


Figure 7.7: Comparison of recall (left) and precision (right) of indirect (top) and direct (bottom) links in LAST baseline and SO fine-tuned models on STAC.

		#EDUs		#Arcs	
	#Doc	Single-in	Multi-in	Proj.	N-proj.
(1) Non-Tree	48	706	79	575	170
(2) Tree	61	444	0	348	35
- **Proj. tree**	**48**	314	0	266	0

Table 7.6: STAC test set ground-truth tree and non-tree statistics. "Single-in" and "multi-in" means `EDU` with single or multiple incoming arcs. "Proj" and "N-proj" means projective and non-projective arcs.

capability to predict distant arcs in longer documents. However, in the range of $[23, 30]$, the PLMs are inclined to predict a greater number of false positive distant arcs, leading to under-performance compared to the LAST baseline. As a result, we see that longer documents (≥ 23) are indeed more difficult to predict than short documents, with even the performance of our best model (BART+STAC) strongly decreasing.

7.5.3 Projective Trees Examination

Given the fact that our method only extracts projective tree structures, we now conduct an additional analysis, exclusively examining the subset of STAC containing projective trees, on which our method could *in theory* achieve perfect accuracy. Table 7.6 gives key statistics for this subset ("proj. tree"). For the 48 extracted tree examples, the document length decreases from an average of 11 to 7 `EDUs`, however, still contains $\approx 40\%$ indirect arcs, keeping the parsing difficulty comparable.

Parsing Results: Discourse parsing results are presented in Table 7.7. We show the performances of oracle heads (Gold H), unsupervised global and local heads ("Unsup H_g", "Unsup H_l"), and semi-supervised layer-wise and head-wise heads ("Semi-sup n L", "Semi-sup n H").

As shown, all three unsupervised models outperform LAST (62%). The best model is still

Train on → Test with ↓	BART F_1	+ SO-DD F_1	+ SO-STAC F_1
LAST BSL	62.0	62.0	62.0
H_{ora}	64.8	67.4	68.6
Unsup H_g	62.5	62.5	62.1
Unsup H_l	62.1	62.9	63.3
Semi-sup 10 L	$59.4_{2.8}$	$60.6_{2.9}$	$58.3_{1.8}$
Semi-sup 30 L	$62.1_{0.2}$	$61.8_{1.2}$	$59.8_{0.9}$
Semi-sup 50 L	$62.1_{0.0}$	$62.3_{0.3}$	$59.9_{0.6}$
Semi-sup 10 H	$54.6_{5.8}$	$59.2_{4.7}$	$61.6_{5.6}$
Semi-sup 30 H	$60.3_{4.7}$	$60.3_{4.4}$	$65.6_{4.3}$
Semi-sup 50 H	$\mathbf{64.8_{0.0}}$	$\mathbf{66.3_{2.3}}$	$\mathbf{68.1_{1.4}}$

Table 7.7: Micro-F_1 scores on STAC projective tree subset with BART and SO fine-tuned BART models. "L"=layer-wise, "H"=head-wise. The best score in the semi-supervised approach is bold.

BART fine-tuned on STAC, followed by the inter-domain fine-tuned +SO-DD and BART models. Using the semi-supervised approach and head-wise attention, we see further improvement with the F_1 score reaching 68% (+6% than LAST). Conversely, we found that aggregating attentions layer-wise was not better than LAST, which is consistent with our results on the entire test set (Table 7.4).

Direct & **Indirect Arcs Performance:** Similarly, as in Section 7.5.2, we take a look at the performance of indirect and direct arcs prediction in the tree subset. Degradation for direct and indirect edges' precision and recall scores are presented in Figure 7.8 (BART model), Figure 7.9 (BART+SO-DD), and Figure 7.10 (BART+SO-STAC). Further, we compare the performance on the whole test set and projective tree subset. Darker colored bars are the results for the whole test set and lighter colored bars tree subset. We find that the recall of indirect edges improves the most in all three models.

Predicted *vs.* Gold Tree Properties: Following Ferracane et al. (2019), we analyze key properties of the 48 gold trees compared to our extracted structures using the semi-supervised method. To test the stability of the derived trees, we use three different seeds to generate the shuffled datasets to fine-tune BART. Table 7.8 presents the averaged scores and the standard deviation of the trees. In essence, while the extracted trees are generally "thinner" and "taller" than gold trees and contain slightly less branches, they are well aligned with gold discourse structures and don't contain "vacuous" trees, where all nodes are linked to one of the first two `EDUs`.

7.5.4 Qualitative Analysis

We provide qualitative analysis of inferred structures. Among all the predicted tree structures, we randomly selected 2 well predicted trees (F score > 60%) with BART-SO-STAC model, as shown in Figure 7.11 and Figure 7.12. Every prediction is compared with the gold-standard tree ("Ground turth"). In these figures, red arrows are *false positive* attachments and blue ones are

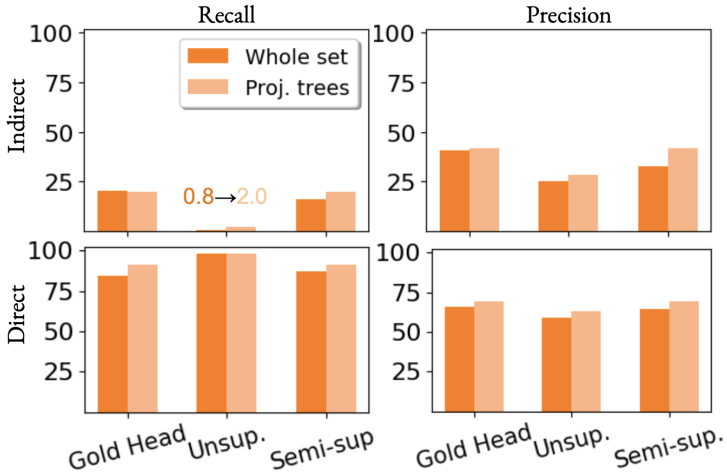


Figure 7.8: Recall and precision metrics in whole test set (darker color) *vs.* projective tree subset (brighter color), with BART model.

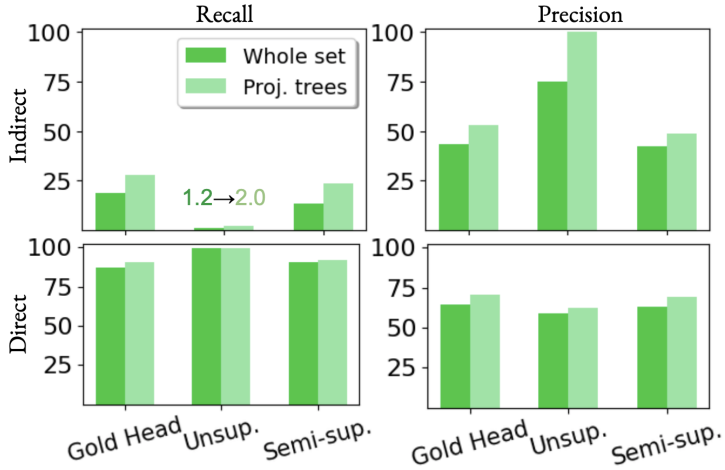


Figure 7.9: Recall and precision metrics in whole test set (darker color) *vs.* projective tree subset (brighter color), with BART+SO-DD model.

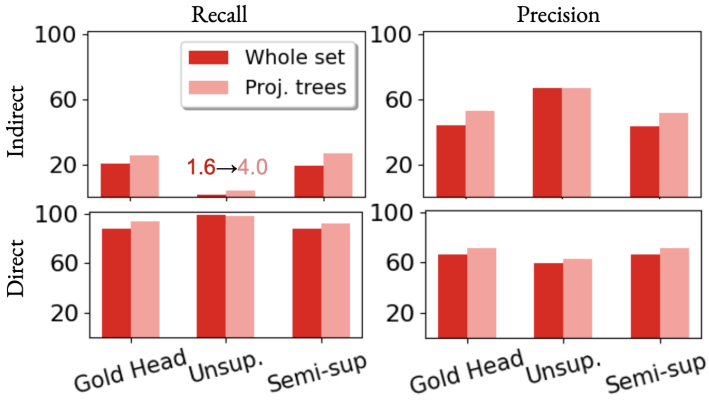


Figure 7.10: Recall and precision metrics in whole test set (darker color) *vs.* projective tree subset (brighter color), with model BART+SO-STAC.

	Avg.branch	Avg.height	%leaf	Norm. arc
GT	1.67	3.96	0.46	0.43
BART	1.20	5.31	0.31	0.34
+SO-DD	$1.32_{0.01}$	$5.31_{0.15}$	$0.32_{0.02}$	$0.37_{0.00}$
+SO-STAC	$1.27_{0.08}$	$5.28_{0.05}$	$0.32_{0.01}$	$0.35_{0.02}$

Table 7.8: Statistics for ground truth projective trees and extracted trees from oracle attention heads in BART and fine-tuned BART models.

false negative attachments. Speech turns are provided for reference.

In the two examples our model achieves over 88% accuracy in predicting projective arcs, including those spanning across 4 `EDUs`, on all three STAC examples. This is noteworthy as it indicates that our method can predict non-linear and non-trivial attachments. These results offer promising evidence that our approach is capable of accurately extracting discourse structures.

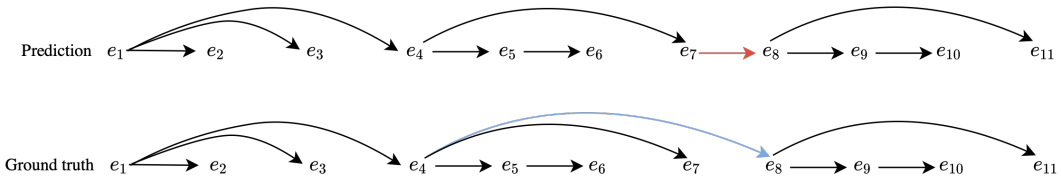


Figure 7.11: Well predicted example: *pilot02-4*. UAS: 90%. In red: FP arcs; in blue: FN arcs.
[e_1] Cat: anyone would give me clay?
[e_2] Thomas: none here
[e_3] william: no
[e_4] Cat: I have one wood to exchange
[e_5] Cat: any takers?
[e_6] william: no
[e_7] Cat: for sheep, wheat or clary
[e_8] Thomas: can I buy a sheep for two ore?
[e_9] william: have none
[e_{10}] Thomas: kk
[e_{11}] Cat: no sheep

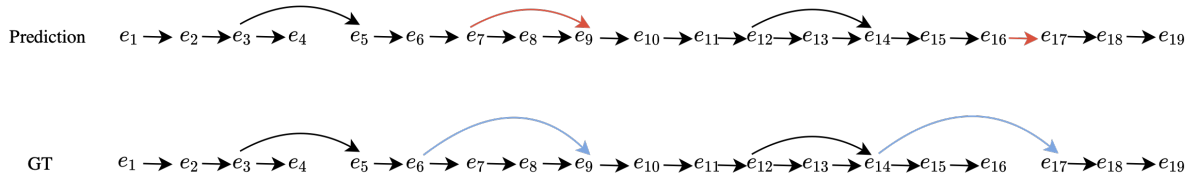


Figure 7.12: Well predicted example: *pilot02-18*. UAS: 89%. In red: FP arcs; in blue: FN arcs.
[e_1] william: hi markus.
[e_2] william: how many people are we waiting for?
[e_3] Thomas: think it's 1 more
[e_4] william: ok
[e_5] Markus: yes, one more
[e_6] Markus: seems there's a hickup logging into the game ...
[e_7] Thomas: that's ok, I not on a schedule
[e_8] Thomas: *I'm
[e_9] Markus: I guess you two had no problems joining the game?
[e_{10}] william: nope
[e_{11}] Markus: Ah great!
[e_{12}] Markus: So, one of you can now start the game.
[e_{13}] Markus: Have fun!
[e_{14}] william: the arrow is pointing at me
[e_{15}] william: but i cant press roll
[e_{16}] william: oh sorry
[e_{17}] Thomas: u can place a settlement
[e_{18}] Thomas: first
[e_{19}] Thomas: u roll later

On the other hand, we have identified some patterns from poorly predicted structures. For instance, in Figure 7.13, our model fails to predict the *losange-shape*, which is a common error in STAC.

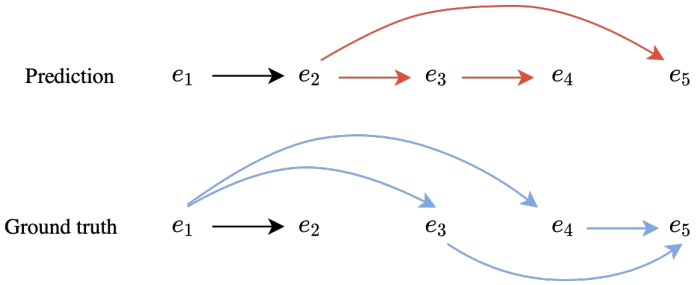


Figure 7.13: Badly predicted example: *s2-leagueM-game4*. UAS: 20%.
[e_1] dmm: i can give a sheep or wood for a wheat.
[e_2] dmm: any takers?
[e_3] inca: sheep would be good.
[e_4] CheshireCatGrin: Not here.
[e_5] dmm: okay.

In Figure 7.14, our model produces chain-style structure instead of distant attachments:

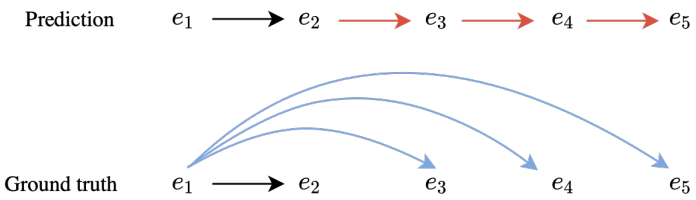


Figure 7.14: Badly predicted example: *s1-league3-game3*. UAS: 25%.
[e_1] nareik15: anyone have ore.
[e_2] nareik15: I have some wood to trade.
[e_3] yiin: no sorry.
[e_4] inca: nope, sorry.
[e_5] Gaeilgeoir: no, sorry.

In the case where the model predicts distant arcs, we find examples with low precision, particularly for long documents, as the one shown in Figure 7.15. This also requires further improvement.

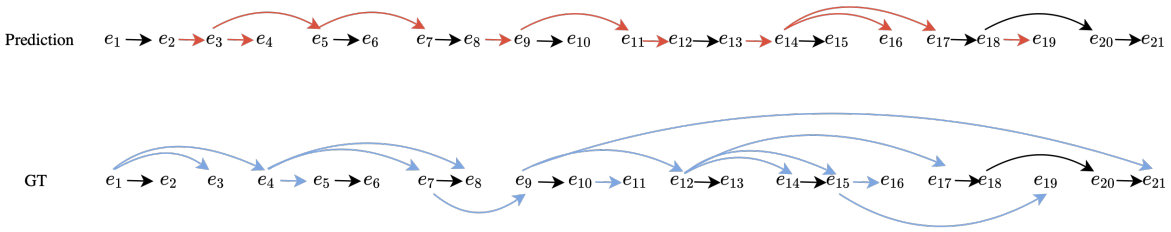


Figure 7.15: Badly predicted example: *s1-league4-game2*. UAS: 30%.
[e_1] Shawnus: need wheat
[e_2] Shawnus: want..clay?
[e_3] ztime: you odo?
[e_4] ztime: yer..
[e_5] ztime: I need clay..
[e_6] ztime: can give wheat
[e_7] Shawnus: k
[e_8] Shawnus: this might be where i lose my road card a?
[e_9] ztime: er..
[e_{10}] ztime: I think the trade is wrong?
[e_{11}] ztime: did you want wheat?
[e_{12}] Shawnus: yes
[e_{13}] Shawnus: for clay
[e_{14}] ztime: it said you wanted clay...
[e_{15}] somdechn: We all want wheat man
[e_{16}] somdechn: and clay...
[e_{17}] ztime: ok
[e_{18}] ztime: thanks..
[e_{19}] Shawnus: haha
[e_{20}] Shawnus: thanks
[e_{21}] somdechn: That happens in the real game as well.

Finally, we showcase two random examples in Figure 7.16 and Figure 7.17.

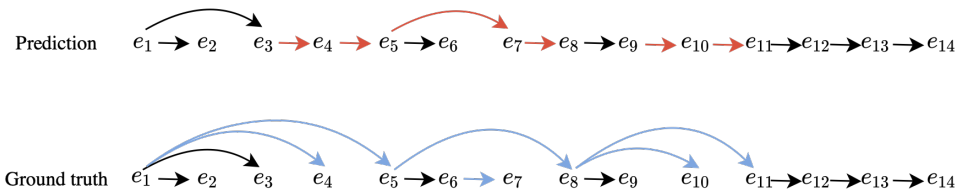


Figure 7.16: Random example: *s2-league4-game2*. UAS: 53.9%.
[e_1] ztime: 7!!!!
[e_2] somdechn: Yeah right...
[e_3] ztime: what... is this a fix?
[e_4] Shawnus: hahaha
[e_5] ztime: ok anyone want wheat?
[e_6] Shawnus: nope
[e_7] Shawnus: just someone to roll 9's..
[e_8] somdechn: Yes
[e_9] somdechn: I can give you wood.
[e_{10}] ztime: was that yes to a trade somdech?
[e_{11}] ztime: OK.. cool.. for 1 wheat?
[e_{12}] somdechn: and an ore.. :)
[e_{13}] ztime: err.. don't have ore..
[e_{14}] ztime: thanks..

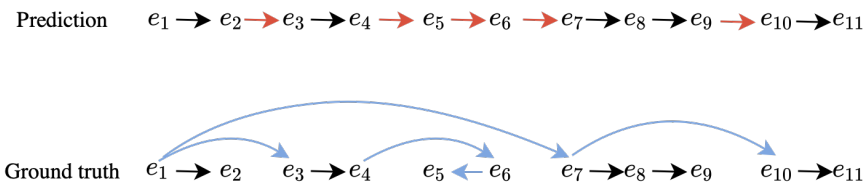


Figure 7.17: Random example: *s1-league3-game3*. UAS: 50%.
[e_1] nareik15: anyone have wood to trade. I have sheep
[e_2] yiin: no
[e_3] Gaeilgeoir: Sorry,
[e_4] Gaeilgeoir: I need wood too
[e_5] Gaeilgeoir: I have wheat
[e_6] Gaeilgeoir: if you want
[e_7] inca: do you have wheat kieran?
[e_8] inca: if so
[e_9] inca: i can trade wood
[e_{10}] nareik15: sorry,
[e_{11}] nareik15: plenty of sheep though :)

7.6 Additional Results on GUM-conv Subset

Our main evaluation and analysis are applied on STAC corpus. Additionally, we extend our experiment and test on another discourse-augmented corpus: GUM[2] (Zeldes, 2017). Note that GUM is initially annotated under RST-framework. To have a direct comparison, we convert constituent trees from GUM into dependency trees using the algorithm proposed in Li et al. (2014a).

GUM-conv Subset: GUM corpus is a growing corpus with rich syntactic and semantic annotation. In its version 8.0, it contains 12 different communicative settings (interviews, textbooks, *etc.*) and a total number of 193 documents. We experiment only with "Conversation" subset (originated from Santa Barbara Corpus (Bois et al., 2000)) which contains 9 recordings of naturally occurring conversations: SBC027 Atoms Hanging Out, SBC001 Actual Blacksmithing, SBC048 Mickey Mouse Watch, SBC025 The Egg which Luther Hatched, SBC031 Tastes Very Special, SBC042 Stay out of It, SBC002 Lambada, SBC011 This Retirement Bit, SBC024 Risk. We use these documents for inter-domain evaluation. Note that GUM conversation subset is very different from that in STAC or DailyDialog: texts are generally much longer (209 utterances per document versus 11 in STAC, 13 in DailyDiaog); contains informal responses (laughs, disfluences); 3 documents are quasi-monologues with one speaker dominants the talk.

Unsupervised Results: We show in Table 7.9 unsupervised results using DAS measurement with global and local heads, as well as with oracle head. We show unsupervised baseline LAST for comparison. Note that in the RST framework, we do not have *Turn constraint* as in the SDRT, so that the transformation can result in links in both directions. LAST baseline only count forward links.

The initial BART model performs slightly worse than the LAST baseline. However, after fine-tuning on CNN-DM and SAMSum, the performance improves. This is consistent with the findings in Huber and Carenini (2022), which show that models fine-tuned on the CNN-DM dataset consistently outperform the BART baseline. On the other hand, for the question-answering fine-tuning task, the results are worse than those of the original BART on GUM. In the Sentence Ordering (SO) task, fine-tuning on DailyDialog yields better results than on STAC. This may be due to the fact that STAC is much shorter than GUM and has a very different vocabulary, creating a significant representation gap.

In comparison to the average performance on STAC, the results on GUM-conv are approximately 20% lower, and we believe that there are at least two reasons for this. Firstly, the documents in GUM-conv are much longer, with an average `EDU` length of 209 compared to 11 in STAC. Previous analysis has shown that when a document contains more than 23 `EDUs`, the F_1 score drops below 50% (see Figure 7.6). By directly comparing our scores to those reported in Huber and Carenini (2022), we found that their best score on the whole GUM test set was 41.8%, which is approximately 10% higher than ours. However, the overall average document length in their study is much smaller at only 107 utterances per document.

Secondly, the nature of the documents in GUM-conv is different from those in STAC. The former contains oral recordings of real-life conversations, with shorter, more informal utterances that are often filled with specific language markers such as laughs, hesitations, and `backchannel` responses. These markers may require additional fine-tuning on oral documents to improve model

[2] `https://gucorpling.org/gum/`.

Model			
Unsupervised Baseline			
LAST			32.1
Unsupervised PLMs	H_g	H_l	H_{ora}
BART	30.4	30.8	↓ 31.8
+ CNN	32.1	**32.2**	↑ 33.0
+ SAMSum	30.5	30.5	↑ 32.2
+ SQuAd2	30.3	30.4	↓ 31.3
+ SO-DD	30.0	30.0	↑ 32.6
+ SO-STAC	31.0	31.0	↑ 31.3

Table 7.9: Micro-F_1 scores on GUM-conv subset with unsupervised PLMs. H_g: global best head. H_l: local best head. H_{ora}: oracle head. Arrows indicate higher or lower scores compared to LAST. Best score in bold.

performance. Moreover, at least 3 documents in GUM-conv are monologue-like, which makes models fine-tuned on dialogue settings less suitable.

We present experiments on oral dialogues that are extremely long (more than 200 utterances) in this part. While the results are not very satisfying, BART+CNN-DM achieved a new state-of-the-art score of 33% for unsupervised discourse parsing on GUM-conv. This is noteworthy because when compared to LAST, the recall and precision for indirect arcs increased from 0 to 7% and 22%, respectively.

7.7 Deployed Discourse Tree Extraction

Following previous work of discourse parsing, all our experiments have started with gold-standard `EDU` annotations. However, as mentioned in Zeldes et al. (2019), this would not be possible in a realistic setting. To assess the performance of a deployed system, we conduct additional experiments in which we first perform `EDU` segmentation and then feed the predicted `EDUs` to our methods.

`EDU` Segmentation Model: We employ the DisCoDisCo model (Gessler et al., 2021) – the top-performing system in DISRPT 2021 for `EDU` segmentation shared task – pre-trained on a random sample of 50 dialogues from train set. We repeat this process three times to accommodate instability. Our average F-score is 94.8, as shown in Table 7.10. In Gessler et al. (2021), authors used 900 training instances and experimented over 5 runs. They obtained an F score of 94.9. Precision: DISRPT shared task employs the original STAC version (45 long documents), as in `https://www.irit.fr/STAC/corpus.html`. In our experiments, we use the *shi2019* version, where 45 long documents have been divided into 1160 sub-documents, as in `https://github.com/shizhouxing/DialogueDiscourseParsing`. The splits of training and testing in DISRPT and *shi2019* version are different. We are cautious in choosing the random training examples.

In the pre-training phase, we utilize all 12 hand-crafted features, including for instance POS tags, dependency relations (UD deprel), and sentence lengths, and opt for treebanked data (available from DISRPT Github) for enhanced performance: 94.9 for treebanked vs. 91.9 for plain text data in DisCoDisCo paper.

	Gold #	Predicted #	Precision %	Recall %	F_1 %
Run1	1155	1115	96.8	93.4	95.1
Run2	1155	1189	92.4	95.2	93.8
Run3	1155	1081	98.9	92.6	95.6
Avg	1155	1081	96.0	93.4	94.8

Table 7.10: EDU segmentation results on STAC test set using DisCoDisCo model (Gessler et al., 2021) re-trained on 50 random dialogues from the validation set. Scores are averaged over three runs.

Evaluation Method: We observe two kinds of mistakes during EDU segmentation: (1) incorrectly separates one EDU into two or more EDUs – false positive or FP, and (2) fails to recognize two separate EDUs as such – false negative or FN.

We illustrate these two mistakes with a toy example in Figure 7.18: a document contains 5 speech turns and 6 gold-standard discourse units. In the prediction, the segmentation system fails to recognize $\texttt{EDU}_3$ and $\texttt{EDU}_4$ in speech turn 3, and incorrectly separates speech turn 4 into two EDUs. In the end, the system gives 6 EDUs, but the internal segmentation is not accurate.

speech turn	gold segmetation	prediction
s_1	$\texttt{EDU}_1$	$\texttt{EDU}_1$
s_2	$\texttt{EDU}_2$	$\texttt{EDU}_2$
s_3	$\texttt{EDU}_3, \texttt{EDU}_4$	$\texttt{EDU}_3$
s_4	$\texttt{EDU}_5$	$\texttt{EDU}_4, \texttt{EDU}_5$
s_5	$\texttt{EDU}_6$	$\texttt{EDU}_6$

Figure 7.18: A document with 5 speech turns and 6 EDUs. For simplification, we use labels s and EDU instead of texts.

To evaluate link attachment performance with the predicted EDUs, we borrow the analysis pipeline in Joty et al. (2015) and adapt it for SDRT-style parsing. We illustrate the measurements in Table 7.11. Precisely, in a false positive case where the system separates one EDU into two (x and y) or more elements, we regard the first element x as the *head*, so that all the incoming and outgoing edges from EDU ($x - y$) should now go to and come out from x to be count as correct attachment. Also, other elements should be linearly attached to each other: y linked to x, z linked to y, etc. In a false negative scenario where the system fails in separating a speech turn to two EDUs (x and y), if a discourse parser predicts an incoming link pointing to the union ($x - y$) while the gold attachment indeed has an incoming link pointing to the *head* x, then we consider it a correct attachment. The same logic applies to outgoing links.

Deployed Structure Extraction Results: Results of structure extraction are shown in Table 7.12, with comparison of using predicted and gold EDUs. The best head (i.e., H_{ora}) perfor-

	Incoming	Outgoing
Human	$i \rightarrow (\mathbf{x} - y)$	$(\mathbf{x} - y) \rightarrow j$
System	$i \rightarrow x$, $x \rightarrow y$	$x \rightarrow j$, $x \rightarrow y$

Table 7.11: Evaluation in the case of false positive EDUs. The head of an EDU is bold.

mance decreases by ≈ 7 points, from 59.5 to 52.6, as well as unsupervised and semi-supervised results. Despite the drop, our unsupervised and semi-supervised models still outperform the LAST baseline. A recent full parser for RST-style discourse parsing is proposed by Nguyen et al. (2021). They report a higher F score of 96.3 for EDU segmentation on RST-DT, compared to ours 94.8 on STAC. However, they also observe a drop of approximately 6 points in structure prediction when using predicted EDU with pretrained models, from 74.3 to 68.4.

			Unsupervised		Semi-supervised		
	LAST	H_{ora}	H_g	H_l	semi-10	semi-30	semi-50
Gold EDUs	56.8	59.5	56.7	57.2	$57.4_{0.4}$	$57.7_{0.5}$	$\mathbf{59.3_{0.7}}$
Pred EDUs Avg	48.9	52.6	50.8	51.1	$50.6_{2.0}$	$52.1_{0.7}$	$\mathbf{52.2_{0.4}}$
Details with predicted EDUs							
Run1	48.8	52.9	50.2	50.9	$51.0_{2.0}$	$52.4_{0.6}$	$52.5_{0.4}$
Run2	49.6	50.5	50.3	50.3	$48.3_{2.0}$	$49.8_{0.6}$	$49.9_{0.5}$
Run3	48.4	54.5	51.9	52.1	$52.4_{2.3}$	$54.0_{0.8}$	$54.2_{0.3}$

Table 7.12: Top part: gold EDUs and predicted EDUs parsing results with BART+SO-STAC model. Scores for predicted EDUs are averaged over three runs. Bottom part: relation prediction result of each run.

7.8 Extension to Graph Structure

Our method only extracts tree structures. Although Maximum Spanning Tree algorithm such as Eisner covers approximately 94% of edges, we aim to produce SDRT graph-like structure. It is important to note that MST algorithms generate exactly $n - 1$ edges for a document with n EDUs. This is a strong constraint that directly forbids a commonly presented discourse structure in STAC – "losange" shaped structure resulting from multiple speakers giving an answer to or acknowledging the same utterance (Asher et al., 2016). To overcome the tree algorithm's constraints, we explore various extension methods using the attention scores of *unselected* edges.

In an effort to construct graphs directly, we opt not to impose any restrictions on the number of attaching edges, with the sole criterion being that every node should have at least one incoming edge, thereby ensuring a connected graph. While in the STAC case, the number of edges in a document is almost always close to $n - 1$, no such information is available for data in other domains. We sort the attention scores in descending order and proceed to make the attachment one by one until a connected graph is formed. However, this method turns out to be far less effective than the Eisner tree algorithm. The algorithm is excessively greedy, resulting in linking too many edges and yielding a near-perfect recall rate but an abysmally low precision rate. Moreover, the attention scores are not sorted in a manner that facilitates graph building, and as

a result, too many false positives are included. Consequently, we dismiss this approach as being impractical.

The second approach we tested is based on the tree structure. We add high-scoring unattached edges back to the established structure. This method involves sorting the attention scores of all unattached edges in descending order and creating additional links for the top k edges, where k is a hyperparameter in the range of $[1, n]$. This is because for a document with n `EDUs`, we need to have at least $n-1$ edges for the graph to be connected and at maximum we can have $n \times (n-1)$ edges. In contrast to the first approach, the second method has the advantage of having the base structure, which makes it easier to tune the value of k. In this method, we choose to use small values of k, similar to the *graph density constraint* in Perret et al. (2016). However, our experiments show that for all values of k tested, increase the value of k leads to higher edge recall but rapidly decreasing precision, resulting in lower F-scores when compared to Eisner algorithm. Therefore, we decide not to adopt this approach as well.

We propose a third method that also relies on the established tree structure, but instead of relying solely on attention scores, we incorporate other dialogue information to train a binary classifier with feature engineering. The goal is to predict whether an additional edge should be included in the post-processing step. After examining a small annotated validation set of 50 documents, we find that longer documents typically contain more negotiation phases and therefore tend to have more losange shapes. Additionally, we notice that the relation types *question answer pair* and *acknowledgement* are frequently missing. Based on these observations, we propose four empirically motivated features to train the classifier:

(1) Attention value (A): Attention score of an *unselected* edge.

(2) Distance (D): Distance between two `EDUs` (normalized by the total count of `EDUs`).

(3) Relation type (R): Probabilities of predicted relation types using the DisCoDisCo model (Gessler et al., 2021), pre-trained with 50 dialogues in the validation set.

(4) Document length (L): Total number of `EDUs`.

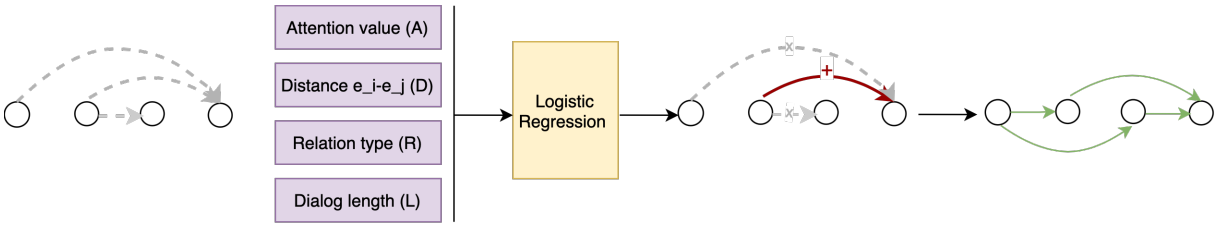


Figure 7.19: Extension to graph structure by adding extra edges.

The pipeline for tree structure post-processing is shown in Figure 7.19. After training a logistic regression classifier with a small number of annotated examples from the validation set, we apply this classifier on the *unselected* edges in the test set. The results are displayed in Table 7.13. Our results reveal that the combination of attention value, `EDUs` distance, and relation type ("+A+D+R") produce a noticeable increase of 1.1 points in F_1 (60.4% versus 59.3%), while incorporating all four features ("+A+D+R+L") shows a rise of 0.9 (60.2%) in F_1. Nevertheless, adding more edges to the set elevate the recall rate but brings about a decline in the precision rate, which only yield a marginal improvement.

	Eisner	+ x Features			
		+A	+A+D	+A+D+R	+A+D+R+L
F_1	59.5	59.5	59.6	60.4	60.2
#TP	-	0	2	24	38
#FP	-	0	0	23	64

Table 7.13: Tree growing strategy results in micro-F_1, true positive (TP) and false positive arcs (FP). A=attention value, D=normalized distance between two `EDUs`, R=relation type, L=dialog length in terms of `EDU` count.

In general, the extension of tree structures to graphs poses great challenges. Firstly, determining the appropriate number of additional edges to be added remains uncertain. Secondly, we notice that the attention scores for *unselected* edges are very close, making it difficult to identify the correct ones. This observation suggests that relying solely on attention scores may not be enough at this stage. Additional signals or cues are likely required for a more precise selection process.

7.9 Conclusion

This chapter delves into the extraction of naked discourse structures, which is a challenging task due to the high degree of data scarcity that characterizes discourse parsing. In addition, existing distant learning techniques that are effective for monologues are not readily transferable to dialogues. Consequently, we investigate methods for building naked discourse structures using attention matrices in pre-trained language models (Section 7.2).

Previous studies on BERTology have mainly focused on discovering discourse information through different probing tasks, as discussed in the previous chapter. However, our goal is to extract discourse structures from PLMs at scale. To achieve this, we design a simple yet effective sentence ordering task that does not require human annotation and can be applied to any domain. By comparing it with other discourse-related fine-tuning tasks, we demonstrate that sentence ordering is the most effective. Selecting the best attention head is a key issue in using PLMs for document-level discourse information extraction. We are the first to address this issue in dialogues by proposing both unsupervised and semi-supervised approaches. The unsupervised approach is based on a novel metric called "Dependency Attention Support" (DAS), which measures the degree of support for the dependency trees generated by each head. We choose high-DAS heads. Meanwhile, the semi-supervised approach selects heads based on their performance on a small annotated validation dataset.

Experiments on the STAC dataset reveal that our unsupervised and semi-supervised methods outperform a strong baseline LAST (F_1 56.8%): unsupervised method gives at best 57.2% and semi-supervised at best 59.3%, delivering substantial gains on the complete STAC dataset (Section 7.4).

Interestingly, discourse is consistently captured in deeper PLMs layers, and more accurate for shorter links. Analysis on projective trees shows that our method is especially effective to extract treelike discourse structures, with promising results competitive to some supervised methods (Section 7.5). In order to create a more realistic scenario, we propose a two-step approach where we first perform `EDU` segmentation and subsequently employ the predicted `EDUs`

for constructing the discourse structure, as discussed in Section 7.7. However, due to the potential for error propagation in this process, the performance of the deployed system is observed to be approximately 7 points lower than that of the gold standard.

Although we achieve promising results on tree structures, we still intend to explore the possibility of extracting graph-like structures from attention matrices (Section 7.8). Our current approach builds upon the established tree structure and extends it with additional arcs of high attention scores, but the results are not satisfactory. Thus, we plan to investigate alternative graph construction algorithms in the future. Furthermore, we aim to expand our shuffling strategies for sentence ordering and explore other auxiliary tasks. Moving forward, our goal is to incorporate the prediction of rhetorical relation types into the naked structure, which we will address in the upcoming chapter.

Chapter 8

Discourse Relation Prediction using Self-Training

Contents

After discussing our work on discourse structure extraction in Chapter 7, we now turn our attention to discourse relation prediction. This task can be accomplished either in sequence after the structure construction phase or concurrently with link attachment. However, in our case, we

opt for a two-step parsing approach: first, we build `EDU` attachment, and then we assign relation types to each pair of `EDUs`. We again employ semi-supervised strategies, but using different supervisions.

We are mostly inspired by the strategy of *bootstrapping* (or *pseudo-labeling*) proposed in Nishida and Matsumoto (2022). Bootstrapping can also be referred to as semi-supervised and/or weakly-supervised learning. For one thing, the target data may contain unseen labels from the train set, and for another, the learning signals could be noisy. In bootstrapping, an initial model is trained with limited gold-standard data and used to give *pseudo labels* on a large non-annotated dataset. The model is then retrained on the gold standard and auto-labeled data together to give inference on the remaining part of non-annotated data. This is an iterative process so that the model can be trained with several rounds of auto-labeled data. The rationale behind the process is that at each stage, the current learning model can give *a priori* highly-confident predictions on (at least partially) unseen data so that the next learning model could benefit from the data augmentation to increase its performance. By convention, we call the model that gives pseudo prediction the *teacher* model, and the one that learns the task *student* model.

Depending on the number of *teacher* and *student* models and how they "teach" and "learn" from each other, bootstrapping systems can be further divided into various training paradigms: (1) *self-training* (Yarowsky, 1995): one single model is used which is both the *teacher* and the *student*; (2) *co-training* (Blum and Mitchell, 1998): involves two models that teach each other. The two models have different inductive biases and can learn knowledge from one another. They start to give predictions independently on the same dataset. During inference, however, two models are retrained with different newly added data filtered with certain selection criteria. The aim is to retrain each model with the knowledge that it has not yet learned; (3) *tri-training* (Zhou and Li, 2005): consists of three models which are initially trained on the same set. Different from *co-training* where the *student* learns from one *teacher*, tri-training uses two *teachers* at the same time. The pseudo-labeled data need to meet an agreement (i.e., selection criteria) among different *teachers* in order to provide to the *student*. This paradigm is supposed to provide more reliable pseudo-labels by accommodating different predictions. Other paradigms include asymmetric tri-training (Saito et al., 2017), mean teacher (Tarvainen and Valpola, 2017), etc. We leave the training process to interested readers.

Among different training paradigms, self-training is the most commonly used strategy in classic semi-supervised learning scenarios (Rosenberg et al., 2005). The training process is the simplest with only one model that plays the role of both *teacher* and *student*. In this section, we explore bootstrapping strategy by first investigating the self-training paradigm. We choose to use fine-tuned pre-trained language models (PLMs) as a continuation of the discourse structure extraction study in the previous chapter. Another reason is that the state-of-art discourse relation prediction model is also based on pre-trained BERT (Gessler et al., 2021). In the third chapter, we have shown that transfer learning (Section 3.2.2) and weakly-supervised (Section 3.2.3) methods have been applied to structure extraction in discourse parsing, a few showing promising results. But such strategies have not been fully explored in relation prediction. We are, to the best of our knowledge, the first to propose combining self-training with PLMs in discourse relation prediction.

This chapter is organized as follows: in Section 8.1 we present similar classification tasks using a self-training strategy, what are the choice of *teacher* models, the sample selection criteria, and the design of the learning loop. We also walk through a few studies using supervised methods for relation prediction. Due to the data scarcity issue and heavily unbalanced label distribution (details in Section 8.3), relation prediction remains a difficult and under-explored task. We present our methodology in Section 8.2 where we make a few simplifications on the task and

propose a novel sample selection criteria on pseudo-labeled data, with the implementation detail in Section 8.3. We present preliminary results in Section 8.4 and provide analysis in Section 8.5.

8.1 Related Work

Discourse relation prediction is no longer a novel task. In the past decades, various relation classification studies have been proposed under different theoretical frameworks: Rhetorical Structure Theory (Marcu et al., 1999), Segmented Discourse Representation Theory (Asher and Lascarides, 2003), and the Penn Discourse Treebank's framework (Prasad et al., 2008a). Different frameworks bring out various annotation forms and relation inventories, creating even finer strands of relation classification tasks. The well-known ones include explicit connective classification (Nie et al., 2019) and implicit relation identification (Rutherford et al., 2017; Kim et al., 2020; Xiang and Wang, 2023), mostly using the PDTB dataset.

For relation prediction in the SDRT framework – precisely with the STAC corpus (Asher et al., 2016), there have not been many studies. We discover two systems DisCoDisCo (Gessler et al., 2021) and DiscRel (Varachkina and Pannach, 2021) that have presented results on STAC. Both systems were proposed under the DISRPT shared task[1]. The first one, DisCoDisCo, utilized a Transformer-based pre-trained language model as backbones and is further enforced with manually extracted categorical features (such as speaker information, the distance between `EDUs`). Using fully supervised training, DisCoDisCo archived 59% accuracy with the base version and 65% with feature engineering. The second system DiscRel used sentence embeddings from SBERT (Reimers and Gurevych, 2020) to compute Euclidean distance between discourse units, and then applied a Random Forest classifier to predict relation labels. This approach showed better results for Chinese and Spanish discourse datasets but was 11 points behind DisCoDisCo on STAC.

As one of the crucial tasks for discourse parsing, relation prediction only gives its best performance at low $60s$, leaving room for further improvement. One possible reason is the data scarcity issue, as discussed in the previous chapter. The most commonly used SDRT-style corpus STAC contains only 45 gaming documents and $\approx 10k$ `EDUs`[2], compared to 385 documents and $21.8k$ discourse units in RST-DT (Carlson et al., 2002a). Further, the number of relation classes is important – RST-DT has 18 coarse-grained relations and STAC 16; the class distribution is also significantly unbalanced. All these factors make relation classification a hard task. Recent studies show that *infrequent* classes suffer from underfitting in supervised learning (Jiang et al., 2016; Kobayashi et al., 2021), probably the main reason for unsatisfying classification results.

To increase training examples for relation prediction, various semi-supervised and weakly-supervised methods have been proposed. Braud and Denis (2014) proposed to combine the natural (human-annotated) and artificial (extraction using heuristic rules) examples in order to improve the implicit relation identification. They tested on a small French corpus `ANNODIS` (Afantenos et al., 2012a) which contains merely 3000 annotated pairs and showed 4.4 points of improvement on a 4-way classification. Shi et al. (2019) leveraged multi-lingual resources from parallel corpora to augment the numbers of implicit relation pairs. Using back-translation, they acquired more reliable implicit discourse relation instances. Results show promising, but we can not follow their strategy due to the lack of such parallel corpora.

[1] `https://sites.google.com/georgetown.edu/disrpt2021`.

[2] The official number of documents in Asher et al. (2016) is 45. In later versions, these long documents have been divided into 1000 smaller sub-documents with an average turn length at 13, as in Shi and Huang (2019); Li et al. (2023), etc.

Apart from artificial data creation methods, models can also *teach* themselves with limited supervision. Self-training, as proposed in Rosenberg et al. (2005); Lee et al. (2013), is an effective technique for refining models when the gold annotation is limited. It involves incorporating unlabeled data into the training process by assigning them pseudo-labels, which helps to enhance the model's ability to generalize. An extension of self-training is co-training (Blum and Mitchell, 1998), which contains different models and agreement tuning with prediction decisions. Studies on co-trained models have proved to be effective in information retrieval (Blum and Mitchell, 1998) and sentence simplification task (Li and Nenkova, 2015). In relation prediction, Jiang et al. (2016) aimed to improve the *infrequent* relation prediction. They co-trained two discourse models CODRA (Joty et al., 2015) and Shifit-Reduce parser (Ji and Eisenstein, 2014) and applied a *filtering* step to select only "high quality" pseudo labels. Results on RST-DT showed considerable improvements for low-frequency relations but require careful tuning for filtering thresholds. Very recently, Nishida and Matsumoto (2022) applied several bootstrapping methods including self-training, co-training, and tri-training for unsupervised discourse domain adaptation. They implemented SOTA discourse parsers and conducted comprehensive comparisons among different bootstrapping strategies. They discovered that the current bottleneck for self-training is the low coverage of accurately predicted pseudo labels, and that self-training enhanced by *active learning* (Settles, 2009) could be a future solution to this problem.

More recently, studies on PLMs such as BERT (Devlin et al., 2019a), BART (Lewis et al., 2020), and GPT (Radford et al., 2019; Brown et al., 2020) show strong performance on various NLP tasks, such as document and relation classification (Shi and Demberg, 2019; Meng et al., 2020; Arslan et al., 2021). These models were pre-trained with hundreds and millions of texts and are capable of producing contextualized word-level or document-level embeddings. These vectors can be used as both the general knowledge source for text understanding and feature representation for classification tasks (Meng et al., 2020). In the context of semi- and weakly-supervised learning, PLMs have been used as reliable classifiers to produce pseudo labels. For instance, in text classification, Meng et al. (2020) first used PLM to collect high-quality genre-specific words (e.g., economy, sport, business) in the unlabeled corpus, and then retrained itself on this distinctive information. Self-trained LM showed stronger generalization ability than other weakly-supervised models. Yu et al. (2021) proposed a contrastive learning framework for fine-tuning PLMs with weak supervision (semantic rules). They tackled the noise contamination issue in self-training and presented significant improvements in sequence-, token-, and sentence-level classification tasks.

Inspired by the self-training paradigm and the outstanding generalization capacity of pre-trained language models, we base our research on the crossroad of self-training and PLMs. To the best of our knowledge, we are the first to propose this combination in discourse relation prediction.

8.2 Methods

8.2.1 Problem Formulation and Simplifications

In the context of SDRT-style discourse representation, a document is represented as a Directed Acyclic Graph (DAG), where every vertex is an elementary discourse unit (`EDU`) – the minimal spans of text – except for the root node, has a single head (one incoming link). Every vertex can have multiple dependents (outgoing links). Every edge in the DAG is typed with a relation.

Formally, given a document $\mathcal{D}$ represented as n non-overlapping sequential `EDUs` and the established attachments: $\{(h, d) \mid 0 \leq h \leq n, 0 \leq d \leq n\}$, where h represents the *head* and d the

dependent, our goal is to predict a relation r to every linked pair (h, d): $Y = \{r \mid (h, d), 0 \leq h \leq n, 0 \leq d \leq n, r \in \mathcal{R}\}$, where $\mathcal{R}$ is the inventory of relations in SDRT.

In our experiments, we make two simplifications:

(1) We assume that all the attachments are already given so that we only focus on the relation prediction task. We regard this problem as a multi-class classification problem. In our case, the number of classes is 16. Note that the distribution of classes is unbalanced (details in Section 8.3).

(2) We treat every `EDU` pair individually, without considering its neighboring context in the document, the same setting as in DISRPT 2021 and 2023 shared tasks[3]. Admittedly, it is more natural to interpret the rhetorical relation of a pair inside a document, especially for long-distance pairs (non-adjacent `EDUs`). However, individual relation pairs invoke local coherence, we consider it as the first step towards global relation coherence building. In the long term, we plan to consider a larger context.

It is also worth mentioning that discourse structure construction and relation prediction are not necessarily two-stage tasks. In Chapter 3 we have presented methods that jointly learn from both tasks and predict a full discourse structure gradually, such as the work by Chi and Rudnicky (2022). Our two-stage approach, on the other hand, gives a clearer picture of the performance of each task. We believe that it is beneficial for full discourse parsing.

8.2.2 Self-Training Loop

We illustrate the training loop within Figure 8.1: self-training starts with a single model $\mathcal{M}$ trained on a small dataset of gold-standard annotation (X, Y_g) (shown as green database in the Figure). In our case, the BERT-base model is fine-tuned with 700 relation pairs. The fine-tuned BERT ($\mathcal{M}$) is used to provide pseudo relation labels on large unannotated data in the same domain (X^t, Y_p^t). Under pre-defined selection criteria, a subset from (X^t, Y_p^t) is sampled (orange database) and merged with the original 700 pairs to retrain BERT ($\mathcal{M}^t$). At each training round (red dashed arrows), we use the previous model to provide prediction on remaining unannotated data and fine-tune a new BERT model with gold and pseudo-labeled data. BERT is both the *teacher* and the *student* for itself.

8.2.3 Classification Module

Our relation classification module has a simple architecture (module "classifier" in Figure 8.1): a base version BERT model is used, and we fine-tune it with gold-standard relation set or the combination of gold and pseudo-labeled so that it outputs 16 relation scores. A softmax layer is employed to give normalized probabilities.

We select BERT, not only because it is the base of the state-of-the-art relation classifier DisCoDisCo (Gessler et al., 2021), but also because of the Next Sentence Prediction (NSP) pre-training task. Previous studies show that the NSP task is helpful for inference tasks; recent work on discourse (Gessler et al., 2021; Shi and Demberg, 2019) further confirms its advantage for relation classification. For this reason, our encoding of relation pairs is to fit the NSP pattern in BERT: a [CLS] token starts the pair, followed by the first `EDU`, a [SEP] marker, and finally the second `EDU`. We keep speaker information at the beginning of each `EDU`, but replace the speaker

[3]DISRPT 2021: `https://sites.google.com/georgetown.edu/disrpt2021/home`. DISRPT 2023: `https://sites.google.com/view/disrpt2023/home`

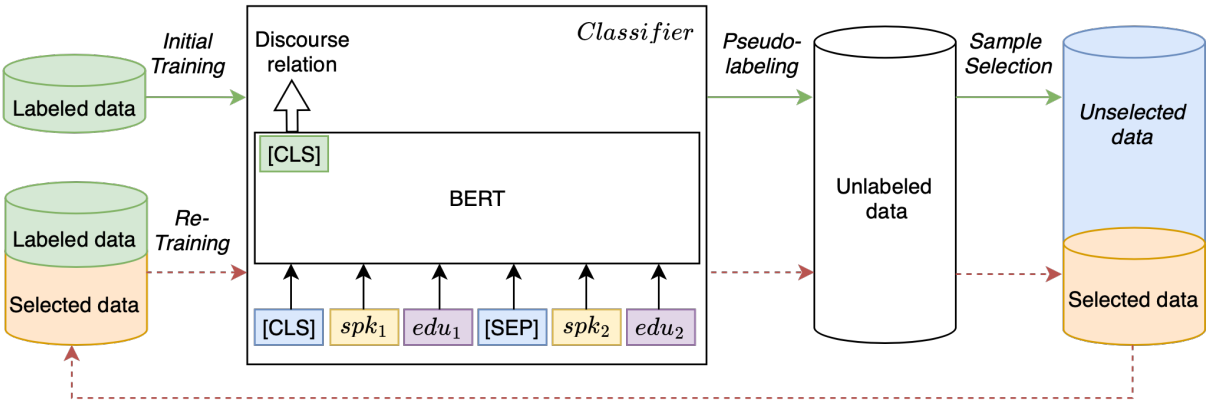


Figure 8.1: An overview of our relation prediction pipeline with self-training. Green solid arrows show initial training and pseudo-labeling. Red dashed arrows indicate iterative training with the combination of gold and selected examples.

names with markers ("spk$n$"). As an example, we show a *question-answer* relation pair from STAC:

[CLS] spk3: anyone need wheat? [SEP] spk2: no thanks

This is a similar representation as in Shi and Demberg (2019) where authors encode pair of arguments in PDTB (Prasad et al., 2008a) for implicit relation prediction. Gessler et al. (2021) did not encode speaker markers in their BERT-base DisCoDisCo model, we will show in Section 8.4 that our encoding yields better results. In STAC, there is no strict separation of same-participant relations and different-participant relations. But in practice, each relation has its preferred use case: for instance, *acknowledgment* and *question answer pair* are common relations among different participants; while *explanation* happens with the same speaker. By showing speaker information, we expect that the model learns such nuances in its native feature space.

8.2.4 Sample Selection Strategy

When selecting a subset from pseudo-labeled examples (X^t, Y_p^t), two questions arise: what are the relatively well-predicted examples, and how to assemble them? The first question corresponds to *confidence measurement* of a prediction model, and the second, *selection strategy* given the confidence.

Confidence Measurement: When using a pre-trained language model as a relation classifier, the raw output is a list of logit values. By using softmax normalization, we obtain a probability distribution of the given n classes. The predicted class thus goes to the one with the highest probability. Conventionally, we can loosely translate the probability of the winning class as the *confidence* of such prediction. Recall that in the previous chapter, we utilized PLM's attention matrices as an indication of dependency support among `EDUs`. While in self-training, we directly interpret the output of PLM as a *confidence* measure.

A model can be *confident* about a correct prediction, in which case the model is confident and reliable. On the contrary, the model can also be *confident* about a wrong prediction, in which case the model is confident but not reliable. The study on the correlation between a model's

predicted probabilities and the probabilities of correctness is known as the **calibration** property (Jiang et al., 2021b). Much work has been dedicated to the probabilistic calibration of deep neural networks, as in Jiang et al. (2012); Jagannatha and Yu (2020); Desai and Durrett (2020); Jiang et al. (2021b). Desai and Durrett (2020) found that pre-trained models are generally more accurate and calibrated. They evaluated the posterior calibration of BERT and RoBERTa on the tasks Natural Language Inference, paraphrase detection, and commonsense reasoning. In both in-domain and out-of-domain settings, pre-trained models appear to be more reliable than the baselines.

Interestingly, in the case of multiple *teachers* bootstrapping (such as co-training and tri-training), the comparison of confidence scores among different models can tell the story of which model is better at handling which predictions (Nishida and Matsumoto, 2022). In our experiment, we utilize only one model BERT and we regard the normalized output as *confidence* measure.

Selection Criteria: With the confidence measurement, a key challenge is how to choose high-confident pseudo-labeled inferences to enhance the initial model. The answer may be more complex than just taking the top confident examples. A simple reason is that highly confident examples tend to be the easiest to predict. If only sample these examples, we manually bring in bias and break the balance of class distribution. The new models will be better and better at certain (easy) classes while worse at other less present (and more difficult) examples. On the other hand, the confidence score is the only source we know about the unknown dataset. There is clearly a trade-off between narrow-but-reliable and large-but-noisy data selection.

Inspired by the work of Steedman et al. (2003) and Du et al. (2021), we define two selection criteria for sample selection. The first one mainly focuses on the reliability of the chosen data, i.e., data with high confidence; the second one is a combination of reliability and variety, choosing highly confident examples while taking care of class distribution in the new sample:

(a) **Top-k**: This is a reliability-oriented criterion. We rank the confidence score of predicted examples in descending order and take the top k pseudo-labeled examples into the next round of training, with $k \in [0, N]$, N is the total number of unannotated data. In our experiments, we started k at 800 and tested k with an interval of 1000. This selection process is quite similar to that of Nishida and Matsumoto (2022) where authors proposed to use a percentage threshold of top-ranking examples (called "rank-above-k", with $k \in [0.0, 1.0]$). Our method, on the other hand, hard-code k values. With $N = 10k$, our proposal in fact corresponds to the 0.1 interval in Nishida's calculation.

(b) **Top-class-k**: This is a variety-oriented criterion. From the pseudo-labeled dataset, we rank confidence scores and choose a subset that satisfies the same class distribution as in the gold-standard set. For each class, we select the examples with a higher confidence score. This is a compromise between high-*confidence* and class variety. Note that there is no guarantee that pseudo labels are correct, especially for the ones with lower *confidence* scores. Nevertheless, we regard this as an approximation to the initial set distribution. The sample class distribution may still be (slightly) different from the initial train set, given that the model could fail (completely) to predict some difficult classes. This issue could be eased when providing more unannotated data.

There are different ways of sample selection. For instance, another common way to select k is to test k at a specific *confidence* ranges so that one can be certain of taking reliable samples regardless of the size of the selection, as in Braud and Denis (2014). In their case for implicit relation prediction, they tested $k \in [0.3, 0.85]$ with an increment of 0.1 until 0.5 for the lower

	Labeled train		Validation		Testing	
Relation	#	%	#	%	#	%
Question_answer_pair	175	25.0	152	22.89	305	27.04
Comment	108	15.43	110	16.57	165	14.63
Acknowledgment	86	12.29	87	13.1	148	13.12
Continuation	65	9.29	69	10.39	113	10.02
Elaboration	64	9.14	52	7.83	101	8.95
Question_elaboration	36	5.14	30	4.52	72	6.38
Result	26	3.71	29	4.37	29	2.57
Contrast	32	4.57	29	4.37	44	3.9
Explanation	34	4.86	31	4.67	31	2.75
Clarification_question	23	3.29	20	3.01	33	2.93
Parallel	10	1.43	14	2.11	15	1.33
Correction	12	1.71	11	1.66	21	1.86
Alternation	5	0.71	8	1.2	19	1.68
Narration	8	1.14	7	1.05	13	1.15
Conditional	12	1.71	10	1.51	18	1.6
Background	4	0.57	5	0.75	1	0.09
Total	700	100.0	664	100.0	1,128	100.0

Table 8.1: Rhetorical relations and frequencies in labeled train (seed-27), labeled validation, and test sets in STAC.

bound and of 0.05 until 0.95 for the upper bound. In our case, we wish to have direct control of the pseudo-labeled sample size. For this reason, we test the Top-k and Top-class-k methods.

8.3 Experimental Setup

8.3.1 Relation Distribution in STAC

We utilize the multi-party dialogue corpus STAC, as a continuation of discourse structure extraction. This corpus contains 1,161 short dialogues, with in average 11 speech turns per document. The initial separation of train, validation, and test is set at 82%, 9%, and 9%, respectively. In the self-training scenario, we take a small subset from the train set (700 relation pairs, $\approx$ 50 documents) as the source labeled dataset and all the remaining examples as unlabeled target dataset (9,400 relation pairs, $\approx$ 890 documents). We use a small subset (664 relation pairs, 50 documents) from the development dataset for validation. We keep the 1128 relation pairs (109 documents) in the test set for testing. Our starting point for the initial BERT fine-tuning thus contains 100 annotated documents (half for train and half for validation).

Table 8.1 shows complete statistics of class distribution for each set. For clarification, we only show the labeled part of the train. The information on the unlabeled part is kept unknown during the experiment. To accommodate instability, we randomly choose 700 labeled examples five times. In this table, we present one train group. As mentioned in problem formulation (Section 8.2), we work on speech-pair level relation prediction and not document level, thus the number shown in the table refers to relation pairs.

8.3.2 Baselines and Evaluation

One unsupervised baseline is the majority class vote. As shown in Table 8.1, the class distribution in STAC is heavily unbalanced, with three majority classes *question answer pair*, *comment*, and *acknowledgment* occupying around half of all relations, where *question answer pair* alone represents 27% in the test set.

Our baseline model use pre-trained BERT as a classifier ("BERT-clf"). The base architecture is simply BERT with a linear projection and a softmax layer on top of the pooling layer. We do not tune any parameters in BERT in this model. We compare against DisCoDisCo system (Gessler et al., 2021): where a BERT-based model is enhanced with hand-crafted features. This system includes a feature vector situated between the [CLS] token and the token in the first `EDU`. The features are numerical such as the distance between two `EDUs` and categorical such as *same-speaker*. In comparison, our system has a different input representation form and we did not apply any feature engineering in the self-training process. To have a direct comparison with DisCoDisCo, we employ accuracy as an evaluation metric for relation prediction. In the analysis section, we show accuracy, recall, and F scores for each relation class.

8.3.3 Implementation Details

In our experiments, we use the uncased base BERT model (Devlin et al., 2019b) provided by Huggingface library (Wolf et al., 2020). The base BERT model is first trained on the labeled source dataset with the following parameters: the batch size of 2, learning rate at $2e-5$, AdamW optimizers with a weight decay rate at 0.01. We fine-tune BERT for a total of 10 epochs and picked the one with the best performance on the validation set. For self-training, we keep the same parameters but give more training epochs: the maximum is set at 20 with early stopping at 5, based on the performance on validation set.

To accommodate instability, we run 5 times fine-tuning with random sample data: for the initial BERT fine-tuning, we choose five groups of labeled examples to retrain BERT; we then keep these examples unchanged and add pseudo labels at the self-training stage. All the training groups are seeded for reproducibility[4]. For evaluation, we report average accuracy scores with the standard deviation.

8.4 Results

We assume that self-training is an effective semi-supervised strategy for discourse relation prediction. In this section, we present experimental results to verify our hypothesis. To begin with, we show a few systems of comparison by showing how different data representations influence the final result. We then choose one optimal setting for self-training. In the self-training part, we compare results with two sample selection criteria and show further improvement with iterative loops.

8.4.1 Preliminary Results with Supervised Learning

8.4.1.1 Systems of Comparison

We have two BERT-base models: BERT classifier ("BERT-clf") and fine-tuned BERT ("BERT-ft"). At the first stage, we train our models using the same data separation as in DisCoDisCo

[4]Code is available on `https://github.com/chuyuanli/DisRel-w-selftraining`.

Unsupervised baseline		Accuracy
Majority class		27.04
DisCoDisCo model		Accuracy
w/o feats		59.67
w/ feats		65.03
Supervised model	Input	Accuracy
(1) BERT-clf	(a) w/o spk	55.80
	(b) w spk	61.20
(2) BERT-ft	(a) w/o spk	59.36
	(b) w spk	64.88

Table 8.2: Systems of comparison. *Supervised models* use the same train, validation, and test sets as in DisCoDisCo.

model. Since DisCoDisCo does not encode speaker information in the speech turns, we further compare two data representations: (a) DisCoDisCo encoding ("w/o spk"); (b) our encoding ("w spk"). We also show the majority class percentage as the baseline score. The results are shown in Table 8.2.

(a) [CLS] anyone need wheat? [SEP] no thanks [SEP]

(b) [CLS] spk3: anyone need wheat? [SEP] spk2: no thanks

The basic version in DisCoDisCo gives 59.7 accuracy with $\approx 9k$ training relations. When adding extra features, the performance increases by 5 points. For our models, we observe a similar gap between the with- and without-speaker input settings both for BERT classifier and fine-tuned BERT, which suggests that the improvement bring by feature engineering in DisCoDisCo largely comes from the speaker information. While our BERT-ft model does not explicitly encode such information, a simple concatenation of speaker markers and the speech turns seems to do the job. Clearly, BERT classifier is not as good as fine-tuned BERT, with an accuracy of ≈ 4 points lower. The performance gap is even more pronounced with fewer training data: with 400 and 700 training pairs, BERT-ft achieves an accuracy of 51 and 57, respectively, while BERT-clf only gives 38 and 40, respectively. Based on such observation, we decide to use BERT-ft and "w speaker" encoding as our principle model (setting 2b) for self-training.

8.4.1.2 Impact of Training Size

Before diving into self-training results, we show the evolution of prediction accuracy within supervised learning setting. Starting from 700 gold relation pairs (≈ 50 documents), we augment training size by adding 1000 relation pairs gradually. We run 5 groups of randomly chosen train data with BERT-ft, and show the average accuracy and standard deviation in Table 8.3. To our expectation, model performance consistently increases with more gold-standard training data: from 56% to 68%, accompanied by a smaller standard deviation. In a realistic scenario, we assume having ≈ 50 annotated documents, and from this point, we test self-learning with pseudo-labeled data.

Note that the training and validation examples employed here are different from those in Table 8.2, which explains the difference in accuracy scores. In the previous section 8.4.1.1, in

Train size	Accuracy
700	$56.61_{0.99}$
1,500	$60.46_{2.99}$
2,500	$63.44_{1.02}$
5,000	$65.76_{0.88}$
7,500	$66.67_{1.10}$
10,000	$68.14_{0.88}$

Table 8.3: BERT-ft model supervised performance with different sizes of training data. Accuracy is averaged with 5 groups of randomly selected train data; subscription is the standard deviation.

order to have a fair comparison with the DisCoDisCo model, we follow the subset separation in the DISRPT shared task. However, in our own experiments (structure prediction and relation classification), we apply the data separation in Shi and Huang (2019). This is also the most commonly used setting for full discourse parsing to the best of our knowledge. For clarification, all the results (except those in Table 8.2) in this chapter are based on the same data separation as in Shi and Huang (2019).

8.4.2 Results with Self-Training

8.4.2.1 Influence of Selection Criteria

We compare two sample selection strategies: top-k and top-class-k. Both selections rank the pseudo labels based on their confidence scores and select resp. without and with consideration of the label distribution in the gold-standard set. We test k values gradually, adding 200 to 7,800 pseudo-labeled data with intervals of 200, 400, and 1,000 (from $k = 800$ and onwards). Our total unannotated data size is 9,300. The 1000 interval loosely correspond to the 0.1 in "rank-above-k" criteria in Nishida and Matsumoto (2022), where $k \in [0.1, 1.0]$.

BERT-ft supervised		
700		$56.61_{0.99}$
BERT-ft self-train	Top-k	Top-class-k
+ 200	$54.73_{1.29}$	$\underline{55.76_{3.57}}$
+ 400	$54.01_{2.26}$	$\underline{57.07_{1.27}}$
+ 800	$54.11_{3.05}$	$\underline{57.66_{1.17}}$
+ 1,800	$53.58_{3.62}$	$\underline{57.34_{1.66}}$
+ 2,800	$55.71_{1.91}$	$\underline{57.62_{0.38}}$
+ 3,800	$56.60_{2.14}$	$\underline{57.62_{1.69}}$
+ 4,800	$56.84_{0.58}$	$\underline{57.75_{1.22}}$
+ 5,800	$\underline{\mathbf{58.23_{0.86}}}$	$\mathbf{58.01_{0.77}}$
+ 6,800	$57.82_{1.09}$	$\underline{57.89_{0.98}}$
+ 7,800	$\underline{57.80_{0.71}}$	$56.97_{2.37}$

Table 8.4: BERT-ft self-training with Top-k and Top-class-k sample selection criteria. The best score per row is underlined. The best score per column is bold. Subscription is the standard deviation.

The results are shown in Table 8.4: the accuracy score of 56.6 in the first row is our starting point – supervised learning results with 700 gold-standard relations (50 documents). The second part presents self-training results. Both selection criteria bring obvious improvement in performance compared to supervised learning: at best 58.2 and 58.1 for top-k and top-class-k respectively. Compared to top-k selection, top-class-k consistently gives gains regardless of the pseudo-labels' size. From +400 point and onwards, self-training over-performs supervised learning. The improvement is stable in a large k range $[800, 7800]$. As for top-k selection, when k is small ($k < 2800$), the number and variety of selected pseudo-labeled data are small, resulting in lower accuracy. When k is relaxed, the coverage of different classes of data increases, and the performance hit the highest point at 58.2. After this point, the accuracy slightly decreases, probably due to the noise of pseudo-labeled data.

In general, we observe that both selection strategies improve model performance using pseudo-labeled data, which is a positive signal. However, the tuning process of k value requires extra effort. Top-k selection, for instance, only shows its advantage when k is relatively large, while smaller k harms the model fine-tuning. In comparison, top-class-k selection shows more stable performance, probably because it follows the class distribution and proportionally increases training examples. The training process shows that top-class-k selection is less prone to overfitting.

8.4.2.2 Evolution with Iterative Training

We have shown the effectiveness of self-training in Table 8.4. We now explore the influence of iterative self-training. With a great amount of unlabeled data, the self-training process can be repeated many times: at each loop, k pseudo-labeled examples are selected and combined with previous train examples; we then fine-tune a new BERT model with this larger train set and make a prediction on the test set. We pre-define a stopping criterion at 3 loops following Nishida and Matsumoto (2022).

BERT-ft supervised						
700						$56.61_{0.99}$
BERT-ft self-train	Loop 1		Loop 2		Loop 3	
	Actual k	Acc	Actual k	Acc	Actual k	Acc
+ 800	756	$57.66_{1.17}$	784	$55.94_{1.16}$	783	$\mathbf{58.07_{1.23}}$
+ 1,800	1,686	$57.34_{1.66}$	1,729	$\mathbf{58.35_{1.24}}$	1,764	$57.43_{2.38}$
+ 2,800	2,595	$57.62_{0.38}$	2,718	$57.48_{1.56}$	2,696	$\mathbf{58.05_{2.28}}$

Table 8.5: BERT-ft 3-loop iterative self-training results with Top-class-k sample selection. The actual selected k is shown in each loop. The best score per row is bold. Subscription is the standard deviation.

We use Top-class-k selection in iterative self-training since it shows superior results than the Top-k selection in the first round. Since this selection strategy emphasizes similar label distribution, small classes accumulate more training examples as we increase training loops, such as *correction*, *conditional*, and *alternation*. It is expected that iterative training can help increase the recall of these relations. We test three groups of k values: $k \in [800, 1800, 2800]$.

Table 8.5 shows the results. For clarification, at each loop's prediction, some classes do not receive enough prediction, resulting in a smaller k than the theoretical number. We show the

actual k values in the table for reference. The best performance is in bold for each setting (+ 800, + 1800, + 2800). We observe that all three settings receive extra gains compared to the first loop, validating the benefits of a larger amount of training data even though they might be noisy. From the actual number of k, we also notice that more *distribution-aligned* examples have been predicted in the second and third loops than the first one (for instance, in +1800 setting, loop 3 merges with 1768 relations compared to 1686 in the loop 1), indicating the model tend to predict more *infrequent* classes. This is encouraging, which suggests that with more loops, the coverage (i.e. *recall*) of infrequent classes is increasing.

For settings +800 and $+2,800$, the best performance comes with the last loop. The setting $+1,800$ hits the peak at the second loop. There is no strong evidence showing that the deeper the loop, the better the performance. We assume that there exists a trade-off between the coverage and the precision of the predictions. With more iterations, the model sees more training data and improves its generalization ability. But what comes alone is the risk of noise contamination. The best score comes when coverage and precision reach an optimal point. We investigate more on this point in the analysis part.

At this stage, we confirm the effectiveness of self-training and further prove the benefits of iterative training. Nevertheless, our best score 58.4 is still much lower than the top score in the same training scale in supervised learning ($2,500$ examples, 63.4 accuracy). In the next section, we decompose the results into class-wise and try to find the bottleneck for further improvement.

8.5 Analysis

8.5.1 Is Confident Model Reliable and/or Biased?

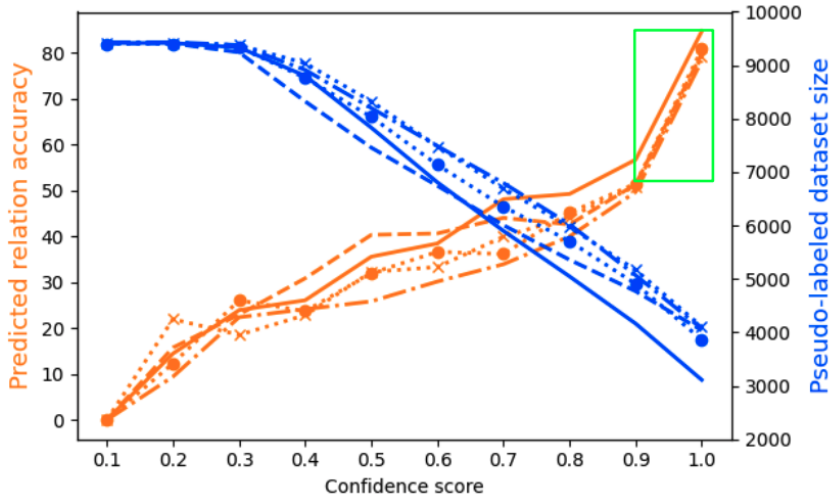


Figure 8.2: Source-only model (i.e. BERT-ft with 700 gold-standard examples) prediction accuracy and confidence on unannotated train set. Each line style represents a different seed of initial BERT fine-tuning. We use 5 seeds.

We used two selection criteria that prioritize high confidence scores to select pseudo-labeled examples. We examine whether the hypothesis that high confidence in data correlates with accurate predictions holds true. Figure 8.2 displays the results for each seeded training group, where the x-axis represents confidence scores, the left y-axis shows prediction accuracy (orange

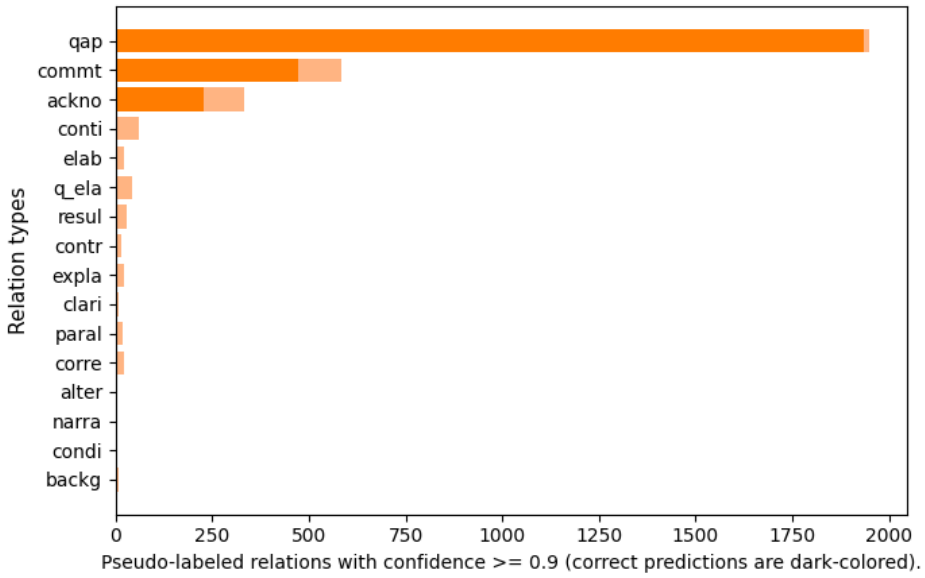


Figure 8.3: Pseudo-labeled class distribution under high confidence prediction (confidence score $>= 0.9$). *Question answer pair* label is overly predicted than the other classes.

lines), and the right y-axis indicates the number of selected examples (blue lines). The accuracy lines exhibit a clear upward trend as confidence scores increase, with the most confident examples achieving over 80% accuracy. The consistency in the trend across all training groups suggests that confidence scores positively correlate with model performance. Thus, our findings confirm the reliability of our basic model.

The examples with high-confidence $[0.9, 1.0]$ are well predicted, however, when adding these examples for self-training, model performance is not improved. In order to understand the inner reason, we zoom in on this high-confidence part (light-green circled area) and show the label distribution in Figure 8.3. Classes are ranked according to their proportion in the train set, light orange shows the predicted numbers and dark orange shows the correct prediction. It turns out that *question answer* relation gets the most credit for the model's high accuracy: more than 60% pseudo-predicted labels are *qap* with 99% precision. 18% and 10% of the predictions go to the second and third frequent relations *comment* and *acknowledgment*, with 80% and 70% of precision. Sadly, all other labels receive very few predictions and most of which are merely false positives. The model is clearly biased.

The confidence score tells two sides of a story. It helps us select well-predicted examples but in a biased way. This finding suggests that the bottleneck of the self-training system is the low coverage of pseudo-predictions. To create a less biased training set, we can loosen the confidence threshold and let in more noisy and diverse data. As shown in Table 8.4, the best score (58.2%) in self-training is not achieved with the highest confidence point, but 0.6 (i.e. when $k = 5800$), which confirms the best k point as in Nishida and Matsumoto (2022).

8.5.2 Is There a Trade-off between Reliability and Variety?

Our second sample selection criterion gives more consideration to class variety (or *coverage*) by selecting high-ranking pseudo labels from each class. The question is: are there reliable examples for each class? To answer this question, we decompose the accuracy line in Figure 8.2

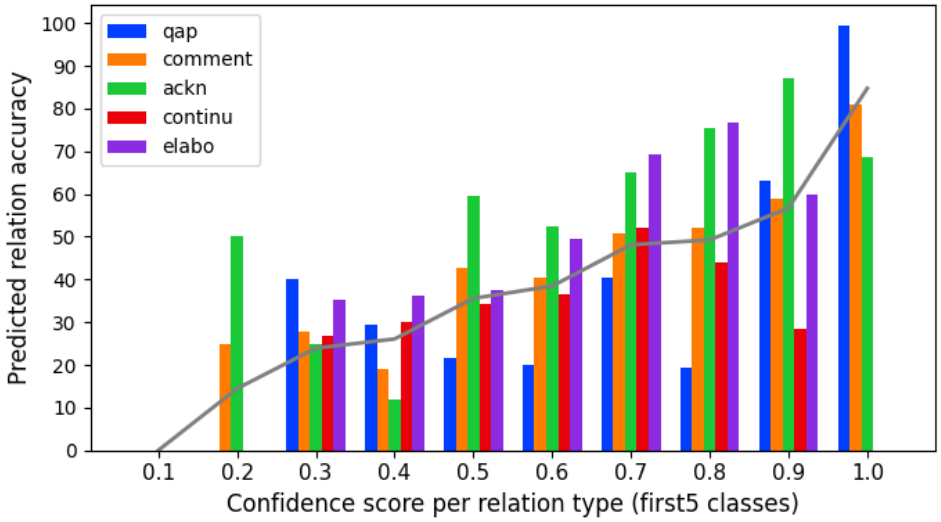


Figure 8.4: Five major classes accuracy and confidence score distribution. The gray line is the combination of all 16 relation classes.

to 16 relations, and make every 5 or 6 relations into one group, shown in Figure 8.4, Figure 8.5, and Figure 8.6, respectively. For simplicity, we call these three groups "First5", "Mid5", and "Last6".

"First5" group in Figure 8.4 contains *frequent* relations: *qap*, *comment*, *acknowledgment*, *continuation*, and *elaboration*. The gray line indicates the global trend of all relations. Frequent relations show roughly a positive correlation between confidence and accuracy, but not all of them strictly achieve the best precision with the highest confidence score. *Elaboration*, for instance, performs better with $k \in [0.7, 0.8]$; *continuation* is best with $k \in [0.6, 0.7]$.

Relations in "middle5" group (Figure 8.5) is composed of *question elaboration*, *result*, *contrast*, *explanation*, and *clarification question*. These relations have a frequency less than 10% and higher than 2% in the labeled train set. Bars are more scarce in the plot, with missing bars indicating the complete failure of predicting such a relation. *Result* for instance, has only been correctly predicted once at a low confidence range $k \in [0.2, 0.3]$. The density of the bars moves more centered compared to that in the "first5" frequent relations, suggesting that the model is less *confident* to give predictions on these relations.

The final group "last6" contains six *infrequent* relations (Figure 8.6). They are the least present and the most difficult to be sampled. From the figure, we see that *parallel*, *correction*, *narration*, and *background* are completely missing, while *alternative* and *conditional* are predicted only with very low confidence ($[0.2, 0.3]$).

To answer the question at the beginning of this section, selected examples are not necessarily "reliable" under the "top-class-k" selection criterion. Less frequent relations can be chosen even if they have low confidence scores. By adding these examples into self-training, we hope to give positive reinforcement in re-training process. The models should not only be fed with good-and-biased or noisy-and-diverse examples. How to find a good balance between reliability and variety is another bottleneck in self-training.

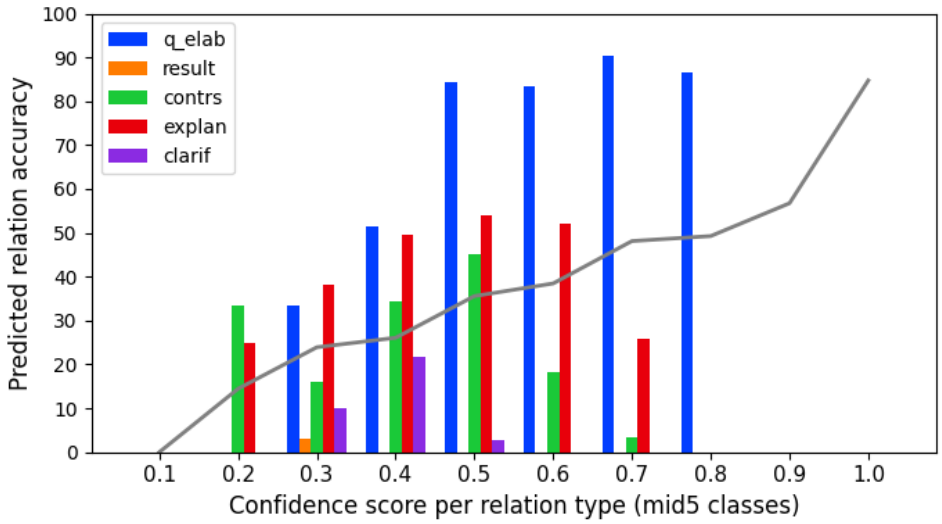


Figure 8.5: Five middle classes accuracy and confidence score distribution. The gray line is the combination of all 16 relation classes.

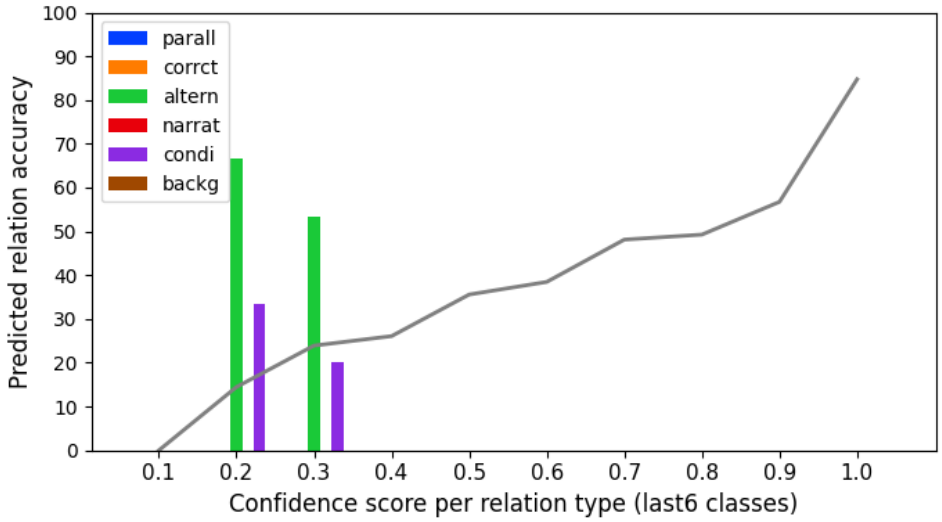


Figure 8.6: Six small classes accuracy and confidence score distribution. The gray line is the combination of all 16 relation classes.

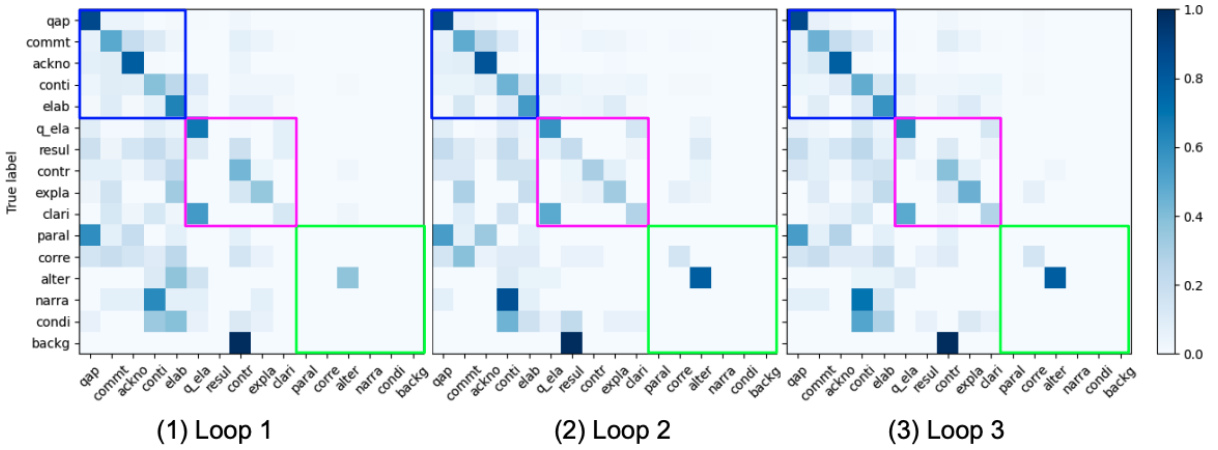


Figure 8.7: Evolution of three-loop self-training, model reinforced with 1800 pseudo-labeled examples at each loop; confusion matrices of 1128 test examples. Blue, pink, and green circled zone are resp 5 high-frequent, 5 middle, and 6 low-frequent relations. Initial train set seed 78. Rows are true labels and columns are predicted labels. Results are normalized.

8.5.3 Is Iterative Training a Good Reinforcement?

Results in table 8.5 show increasing accuracy with iterative training and among the different sizes of pseudo-selected labels. Moderate may be the numbers, all the improvements are effective (with *t-test* $p < 0.3$[5]).

The idea of multi-loop self-training is to improve the model's performance by adding more training examples for the *infrequent* classes, thus gradually easing the issue of underfitting. We reveal the results by showing the evolution of confusion matrices in Figure 8.7 during three loops. All the models are tested on 1128 STAC test set. We separate the matrix into three zones for a clearer presentation: the blue circled zone is the "first5" high-frequent relations; the pink circled zone are "middle5" group; the green circled zone has the "last6" low-frequent relations. Each model is trained with 1800 more pseudo-labeled examples than the previous one.

A clear observation is that the "last6" (light green circled) zone has some recall improvement with iterative self-training, typically for *correction* and *alternation*. In the "middle5" (pick circled) zone, *question elaboration*, *explanation*, and *clarification question* relations also have higher recall and better precision. In the "first5" (blue circled) zone, iterative training does not bring many changes, probably because the model is familiar with these relations and performs well.

Another good signal is that the color beneath the high-frequent classes is becoming lighter with more loops, indicating that the model is not over-fitted with high-frequent labels as we continue self-training. On the other hand, the model keeps miss-predicting *narration* with *continuation*, and *background* by *contrast* (or *result* in loop 2).

We have demonstrated that using "top-class-k" selection method, self-training helps to improve *infrequent* class recall. However interesting, the improvement is not apparent. Will more pseudo-labeled data further boost the augmentation?

Results in Table 8.4 show that when k equals to 5800, self-training result is the best (58% vs 56.6 with *source-only* model). To investigate the performance for infrequent classes, we visualize

[5] p value is not as small in the convention value since we can only compare groups of 5 values. With more seeded groups, we expect to get more significant results.

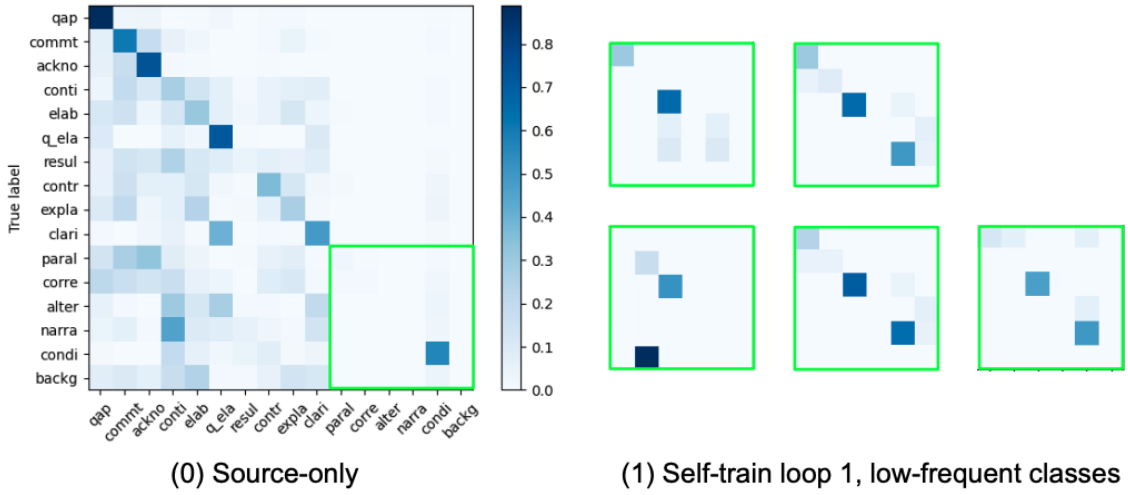


Figure 8.8: Comparison of confusion matrices between *source-only* (left) and first loop self-trained models (right, 5 runs of +5800 pseudo-labeled examples). We highlight the right-bottom part for low-frequent classes. Rows are true labels and columns are predicted labels. Results are normalized.

the confusion matrices of all models trained with $5,800$ pseudo labels at the first self-training loop (blocks on the right side in Figure 8.8). In comparison, we show the confusion matrix without self-training on the left side. Focusing on the green blocks, we confirm that the improvement for "last6" infrequent classes are more pronounced than that in Figure 8.7, when $k = 1800$.

Due to the limited size of unlabeled relations ($\approx$ 9300 in total), increasing the k value can no longer guarantee the proportional increase of infrequent classes, but only adding more high-frequent pseudo-labeled relations such as *question answer pair* and *acknowledgment*, thus harming the retraining process. A simple and straightforward method is to add more in-domain documents in the unlabeled data pool, a tentative point to test as further work.

8.5.4 Human-in-the-Loop at Rescue?

Until now, we have analyzed the impact of sample selection criteria, the impact of iterative training, and the size of the unlabeled data. We show that when adding more distribution-similar pseudo labels, the improvement is more pronounced. Another important question is: how accurate are these pseudo-labeled data? If we dispose of some human resources, can human annotation help create more reliable predictions?

In the mind-opening work from Nishida and Matsumoto (2022), authors show significant improvement ($> 6\%$) in performance when adding actively-labeled data: they sampled 100 documents with the *worst* confidence scores and provide human annotation. Inspired by their work, we test two hypotheses tailored to our scenario:

(1) Human feedback should not only be given to the least confidence predictions, rather they should be given to a subset that follows a certain label distribution.

(2) Small amount of human feedback is effective in showing significant improvement. By "small", we suggest a few hundred of relations (roughly corresponding to $10-50$ documents in STAC dataset).

BERT-ft supervised	
source-only 700	$56.61_{0.99}$
BERT-ft self-train	
$+1,800$ top-class-k	$57.34_{1.66}$
Self-train with human feedback	
$+1,800$ top	$57.96_{1.30}$
$+1,800$ bottom	$62.75_{3.26}$
$+1,800$ top-bottom	$62.23_{1.65}$
$+1,800$ random	$63.44_{2.22}$

Table 8.6: Part 1 and part 2: supervised model and self-training model performance. Part 3: comparison of different ways ("top", "bottom", "top-bottom", "random") to provide human feedback. All the scenarios are tested on $1,128$ STAC test set.

For the first hypothesis, we investigate 4 ways to inject human annotation. Supposing we have 1800 gold relations, we investigate when giving the most confidence ("top"), the least confident ("bottom"), the equal combination of the most and least confident ("top-bottom"), or the perfect distribution-satisfied ("random") gold examples, which scenario bring the greatest improvement. Results are given in the third section of Table 8.6. For comparison, the first two rows give *source-only* and self-trained models' performance: 56.6 and 57.3.

When giving corrections to the confident examples, we see very little improvement compared to pure self-training (57.9 vs 57.3), which suggests that "top-ranked" pseudo-labels are already of high precision. When it comes to the least confident examples, we see a big increase (+5 points compared to self-training), aligning with Nishida and Matsumoto (2022). A similar enhancement is also observed in the compromise point with half-confident and half-unconfident examples. When it comes to the easiest selection way – random, the highest performance is achieved (63.4), suggesting that in a pool full of unlabeled data of the same domain, the best strategy for human feedback is by randomly providing annotations. Anecdotal as it seems, this discovery tells that feedback for good prediction and bad prediction is both useful for model improvement.

For the second hypothesis, we include human feedback on the pseudo-labeled examples. Precisely, when the source model makes inferences, we first select a subset using "top-class-k" selection strategy, we then manually check and correct the predictions if necessary, to finally incorporate these examples with the original gold-annotated data for retraining. We test different ranges of human annotations, from 200 to 7800 relations pairs. Note that the true benefits come with a small amount of annotation.

Figure 8.9 shows the comparison of self-training (blue line) and self-training with human enforcement (or human-feedback, "HF", orange line) with different sizes of data. We also present supervised learning results for reference (light green line). At small data ranges (i.e. $k \in [200, 400, 800]$), human feedback does not seem to give much influence on self-training. This is mainly because the pseudo-labeled data at this stage are highly accurate. The corrections made are not sufficient enough to tune the models' prediction on small and difficult examples. Starting from 1800, we see an evident increase compared to self-training. The gap between self-training and human-feedback self-training grows wider with more data correction. When $k = 5800$, self-training attends its highest point and only goes down afterward while HF continues to improve

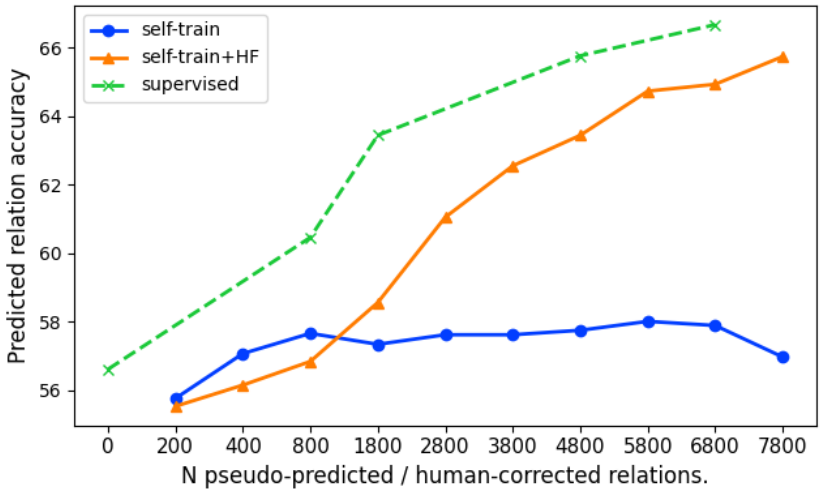


Figure 8.9: Comparison of supervised training (green line), self-training with 700 supervised and k pseudo-labeled examples (blue line), self-training with k human-feedback (HF) examples (orange line). Evaluated on 1,128 examples on STAC test set.

the performance. We have analyzed previously that at this point, the model is contaminated by too much noise and uneven label distribution. But with human annotation, we can minimize the noise to the greatest extent, so that the model continues to learn accurate information.

Notice that the gap between supervised learning and self-training with HF is decreasing. However, self-training does not surpass full supervision even with heavy human intervention. This is probably because of the pre-defined sample selection process in self-training. By using the confidence measure, we are prioritizing the examples that are more similar to those in the initial train set (that is why they gain high confidence). Models trained with these examples – even though correctly annotated –, perform less well with unseen data.

To this point, our second hypothesis on "small amount of human feedback is enough" does not hold. Self-training can bring limited improvement, and human effects only start to show strong support with a considerable amount of annotation ($>$ 130 documents with the size of STAC dialogues). On the other hand, Nishida and Matsumoto (2022) states that with only 100 actively-labeled documents, they gain at least 6 points compared to pure self-training. We reason that this is mainly due to the different test scenarios: their goal is domain adaptation and the model will benefit more when providing gold annotation in the target domain.

8.6 Towards Full Discourse Parsing

System Composition & Results: Taking one step further, we introduce our full discourse parsing system that performs complete parsing, from `EDU` segmentation to structure attachment, and finally, relation prediction. The system comprises three modules, as shown in Figure 8.10. Remarkably, we train the system using only 50 documents, with an average of 13 `EDUs` per document, making it the first semi-supervised discourse parsing system for dialogues.

Let's go through the three modules and present step-by-step performance: The first module, DisCoDisCo (Gessler et al., 2021), achieves a F score of 94.8% for `EDU` segmentation. Next, the predicted `EDUs` are put into a fine-tuned BART model for structure extraction. This model is fine-tuned using Sentence Ordering, as described in Section 7.2.2. Using only 50 annotated examples,

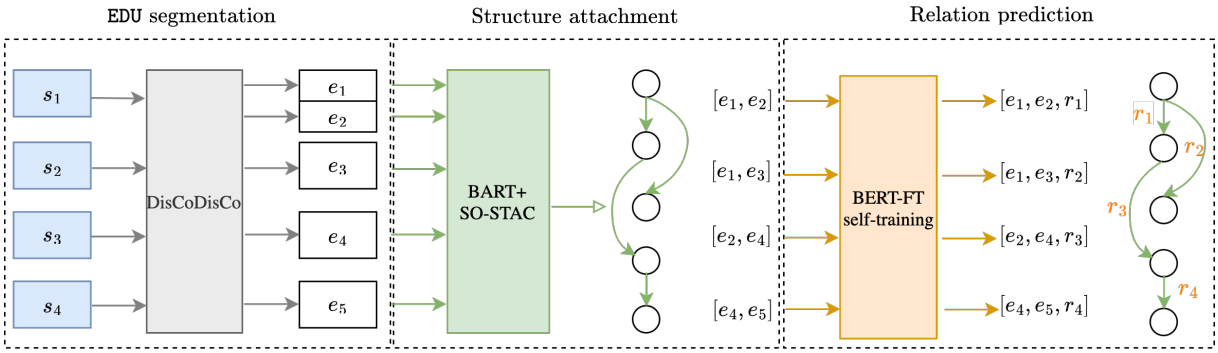


Figure 8.10: Pipeline of our proposed full discourse parsing system.

we determine the best performing attention head and extract tree structures, resulting in a performance of 52.2%. Lastly, we employ a fine-tuned BERT model to predict rhetorical relations based on the extracted structures. This model is iteratively re-trained using a combination of gold (50 documents) and pseudo-labeled data, as outlined in Section 8.2. The final parsing result, considering all three steps, is 32.8%, as displayed in the last line of Table 8.7.

Model	N. train	EDU seg	Link	Relation
Liu and Chen (2021) (sup)	1,091	-	75.3	56.9
Chi and Rudnicky (2022) (sup)	1,091	-	74.4	59.6
Ours w. gold EDUs and link	50	-	-	$58.4_{1.3}$
Ours w. gold EDUs	50	-	$59.3_{0.7}$	$38.6_{0.7}$
Ours w. pred EDUs	50	94.8	$52.2_{0.4}$	$32.8_{0.9}$

Table 8.7: SDRT-style full parsing results. N. train: number of training examples in STAC. Top: state-of-the-art supervised parsers. Bottom: our semi-supervised parser setp-by-step parsing results. Scores are micro-F_1. "-" means using the gold data.

At the top of Table 8.7, we present state-of-the-art models that utilize gold EDUs as a starting point for link and relation prediction. To the best of our knowledge, there are no supervised models that report results based on predicted EDUs. When comparing our results with these supervised models, we observe a performance gap of approximately 15 points for link attachment (using gold EDUs) and approximately 18 points for link+relation prediction. This difference in performance can largely be explained by the huge difference of training size. It's worth noting that supervised models, such as the one proposed by Chi and Rudnicky (2022), perform joint link and relation tasks, enabling the model to leverage relation information to aid in link prediction. In contrast, our approach does not provide relation information during the link prediction step. While these supervised models perform better within specific domains, studies have shown a significant drop in their inter-domain capabilities (Liu and Chen, 2021; Nishida and Matsumoto, 2022). In contrast, our parsing pipeline is built upon models trained with distant and weak supervisions, making it more adaptable to other domains compared to supervised models. Regarding RST-style parsing, we find a full parsing system proposed by Nguyen et al. (2021), which is trained and tested on the RST-DT corpus, thus not directly comparable to our results. Nevertheless, we observe a similar gap of 20 points from link attachment to relation prediction, confirming the inherent challenge in discourse relation prediction.

Step-by-Step Error Analysis: The *Structure-then-relation* framework is susceptible to error propagation. Our system achieve a performance of 59.3 for structure attachment, which is a relatively low starting point for the following relation prediction task. We are intrigued by the rhetorical relations in these "missing structures" and how can we use this information this information to improve future joint frameworks.

To achieve this goal, we divide the gold-standard relation pairs into four categories, depicted in Figure 8.11. The gray and green blocks represent the "missing relations" (i.e., false negatives) that arise from the initial two stages of the process: `EDU` segmentation and structure attachment. The orange and red boxes, on the other hand, represent the potential relation pairs for the relation prediction phase. The most substantial issue arises from the structure attachment phase, which accounts for nearly 39% of errors. When combined with errors from `EDU` segmentation, almost 44% of relation pairs remain unattached. Regarding the attached pairs, relation prediction accuracy is $61.2_{1.6}$ with predicted `EDUs` and $62.0_{1.2}$ with gold `EDUs`, resulting in the final full parsing score of 32.8.

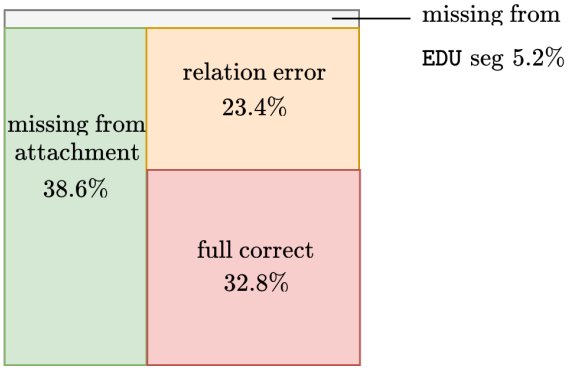


Figure 8.11: Full parsing system error accumulation in different tasks: relation pairs missing from `EDU` segmentation (gray) and structure attachment (green), relation prediction error (orange). The red part is correct in full parsing.

Are unattached pairs difficult to predict due to the rarity of the relationship? Which relationship is the most challenging to predict even with the correct attachment? To answer these questions, we examine the relation composition in each task block in Figure 8.11 and show the result in Figure 8.12. In Figure 8.12, each relation bar is composed of the number of correct predictions (red), wrong predictions (orange), missed predictions due to unattachment (green), and missed predictions due to segmentation (gray). The exact scores are reported in Table 8.8. The top three relations with the most missing pairs are *Question answer pair*, *Comment*, and *Acknowledgment*, accounting for 127, 63, and 71 missing pairs, respectively. Proportionally, almost all *infrequent* classes suffer from low link attachment. In *frequent* classes, *Question elaboration* and *Continuation* suffer the most from missing attachment (at 57% and 48.7%), followed by *Acknowledgment* and *Question answer pair*, as 48% and 41.6%, respectively.

In terms of the precision of predicted relations (red vs orange), *Question answer pair*, *Acknowledgment*, and *Elaboration* are among the best-performing relations, with percentage at 46.9%, 40.5%, and 45.5%, respectively. This means that once given, they are highly likely to be correctly predicted. This high precision in relation prediction provides an opportunity to recover missing attachments of the same type through joint learning. This is a promising direction for future work.

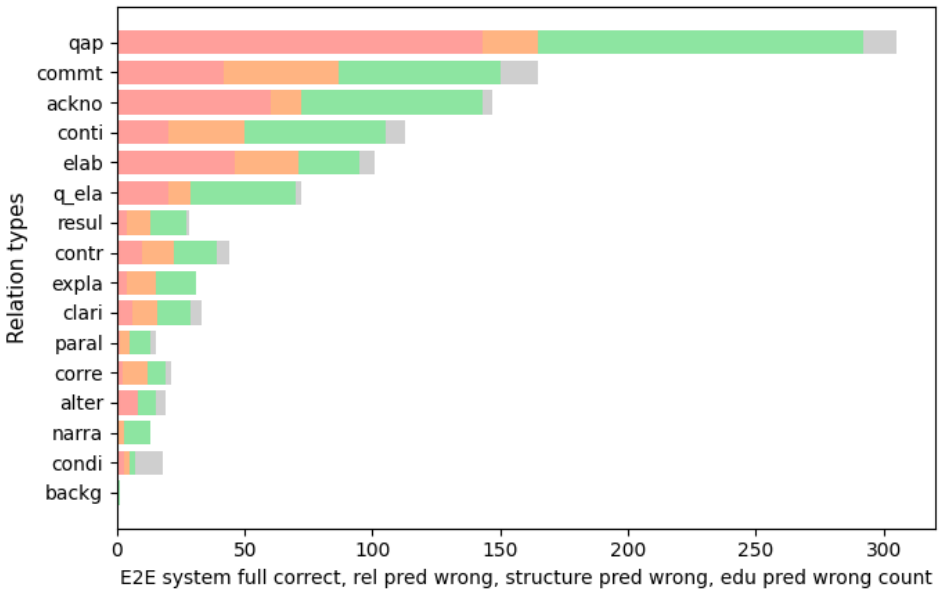


Figure 8.12: Step-by-step parsing results decomposed in relation types. Red: correctly predicted relation in full parsing; Orange: wrongly predicted relation in full parsing; green: false negative errors in structure attachment; gray: false negative errors in EDU segmentation.

	Relationship	Correct prediction	Wrong prediction	Missing attachment	Missing segmentation	Total (100%)
First5	Question answer pair	143 (46.9)	22 (7.2)	127 (41.6)	13 (4.3)	305
	Comment	42 (25.5)	45 (27.3)	63 (38.2)	15 (9.1)	165
	Acknowledgement	60 (40.5)	13 (8.8)	71 (48.0)	4 (2.7)	148
	Continuation	20 (17.7)	30 (26.5)	55 (48.7)	8 (7.1)	113
	Elaboration	46 (45.5)	25 (24.8)	24 (23.8)	6 (5.9)	101
Mid5	Question elaboration	20 (27.8)	9 (12.5)	41 (57.0)	2 (2.8)	72
	Result	5 (17.2)	9 (31.0)	14 (48.3)	1 (3.5)	29
	Contrast	10 (22.7)	12 (27.3)	17 (38.6)	5 (11.4)	44
	Explanation	4 (12.9)	11 (35.5)	16 (51.6)	0 (0)	31
	Clarification question	6 (18.2)	10 (30.3)	13 (39.4)	4 (12.1)	33
Last6	Parallel	1 (6.7)	4 (26.7)	8 (53.3)	2 (13.3)	15
	Correction	2 (9.5)	10 (47.6)	7 (33.3)	2 (9.5)	21
	Alternation	8 (42.1)	0 (0)	7 (36.8)	4 (21.1)	19
	Narration	0 (0)	3 (23.1)	10 (76.9)	0 (0)	13
	Conditional	3 (16.7)	2 (11.1)	2 (11.1)	11 (61.1)	18
	Background	0 (0)	0 (0)	1 (100)	0 (0)	1

Table 8.8: Full parsing system relation decomposition in each module. We show absolute numbers and percentages (%). "First5" are 5 frequent relation classes; "Mid5" are 5 middle classes; "Last6" are 6 infrequent classes.

8.7 Conclusion

In this chapter, we investigate discourse relation prediction task. Following the set-up in 2021 and 2023 DISRPT shared tasks, we treat this problem as multi-class classification problem. We use gold-standard `EDU` attachments so that every pair has a gold relation. In the SDRT-style parsing approach, the number of relation labels is 16. However, the distribution of these classes is highly uneven, with the top three most common relations, namely, *question answer pair*, *comment*, and *acknowledgment*, accounting for more than half of all the relations.

Our relation classifier is a BERT-based model fine-tuned with 700 gold-standard relation pairs. We choose BERT because the Next Sentence Prediction pre-training task has shown beneficial for discourse relation classification. The pipeline is to produce pseudo labels on unannotated data using fine-tuned BERT, and employ sub-sampling to select reliable examples for the next rounds of retraining. We propose to take the high-confident examples in each relation as a way to converse the class diversity. We also investigate iterative training and find that infrequent relation classes benefit particularly from iterations. The overall model performance is better when giving more pseudo-labeled data at each loop. Inspired by active learning in Settles (2009), we investigate the combination of self-training with human feedback. Typically, we propose two hypotheses to verify the effectiveness of *human-in-the-loop* training process. We find that human efforts put on relatively low-confidence examples can help to boost the performance, but only after a certain amount of annotations.

Moving forward, we present a full discourse parsing pipeline in dialogues (Section 8.6), which is the first of its kind. We combine the structure prediction module presented in Chapter 7 and the self-training relation prediction module in Section 8.4. This combination yields a F score of 32.8, which indicates great room for improvement in future research.

For future work, we aim to tackle the relation prediction problem in a larger context by considering the global discourse structure. The current approach is effective for adjacent speech turns, but for long-distance attachment, we need more contextual information. The second step is to investigate joint strategies for both link attachment and relation prediction. As discussed in Chapter 7, tree extraction algorithms such as Eisner are constrained in generating multi-incoming edges (recall the "losange" shape). These multi-outgoing and multi-incoming edges often correspond to specific relation types such as "question-answer pair" and "acknowledgment". One potential approach to address this issue is to enhance the structures by incorporating related information. For example, if a relation with high confidence is provided between two `EDUs` but no attachment is previously made, we could perform post-hoc refinement to add back the attachment.

Chapter 9

Conclusion

Contents

This thesis addresses a crucial and relatively unexplored area in NLP, namely discourse analysis in dialogues, motivated by the pressing need for reliable and versatile discourse parsers and the scarcity of available resources. Our primary objective is to propose effective machine learning techniques, such as improved data representation and feature engineering, and relevant distant and weak supervision signals, to overcome the scarcity of data in discourse analysis. In pursuit of this aim, we formulate two research questions:

RQ1 How can we use discourse information as deployed linguistic features in text classification tasks such as mental disorder illness detection?

RQ2 How can we generate discourse structures with machine learning techniques using less supervision for the greatest applicability in real-life scenarios?

which we have subsequently answered in part II and III of this thesis. Let us revisit the diagram presented in Chapter 1, where we address each research question with two projects. In the following, we provide a brief overview of these projects in Section 9.1; we discuss their limitations and suggest possible future improvements in Section 9.2. Dealing with actual data and large pre-trained language models can raise ethical concerns, which we address in Section 9.3.

9.1 Presented Results

9.1.1 Discourse Structure Discovery

In Part II, we provide a response to the first research question (RQ1) that concerns the incorporation of structural information in text classification tasks. The task of identifying cognitive impairment presents a realistic challenge where the issues of lexical biases and data scarcity are prevalent. These challenges do not have any established solutions, and we believe that our

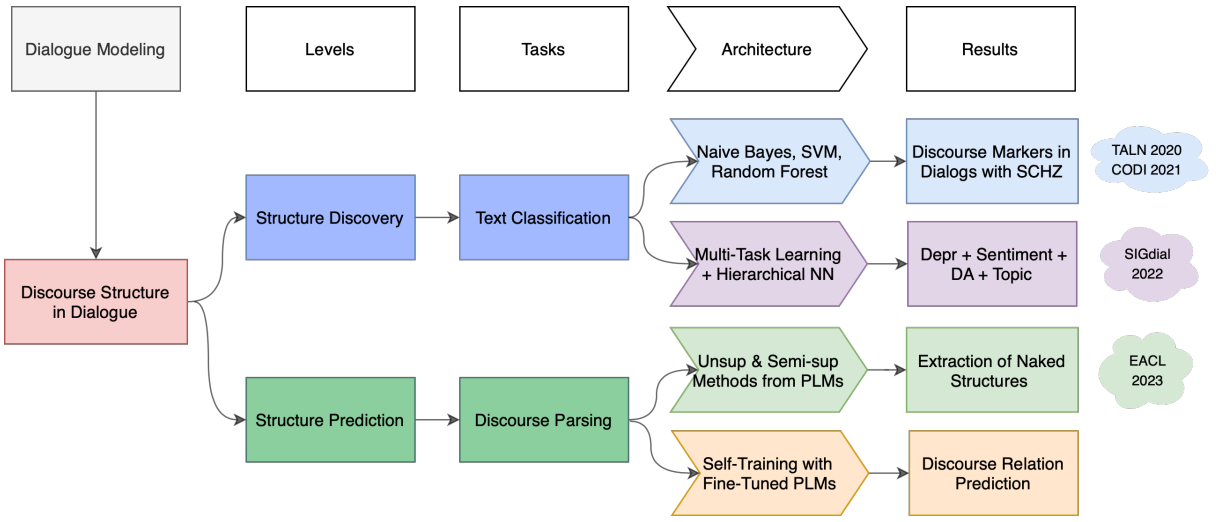


Figure 9.1: Thesis projects overview.

efforts towards answering RQ1 highlight the significance of exploring higher-level, less language-dependent features to create robust systems and derive more universal conclusions from conversational data.

Schizophrenia Language Identification: The aim of this project is to investigate linguistic markers related to schizophrenia through feature exploration using a classification system. The study focuses on spontaneous dialogues in French (Rebuschi et al., 2014) and proposes two methods to address the issue of data sparsity. The first method involves exploring different levels of feature engineering, including lexical (*bag-of-words*), syntactic (POS tagging), and discourse (*Backchannel* response, *Open Class Repairs*, connectives) features. The second method involves modeling dialogues by restricting analysis to patients' speech turns and testing various context windows to improve data representation. The study compares several classification algorithms and finds that Naive Bayes performs well with lexical counts, while SVM and LR are better suited for scarce data and high-dimensional features. The analysis reveals that patients tend to voluntarily discuss their illness and treatment, resulting in disease-related topics, which heavily biases the lexicon. Delexicalized models, which emphasize morpho-syntactic information and high-level discourse features, are more generalizable. The study also uncovers interesting findings related to the characteristics of schizophrenia patients, such as their use of more verbal and adverbial phrases and less phatic expressions, which is consistent with previous studies.

Depression Detection: The first project is limited in its ability to model interactions. To address this issue, we undertake a second project that investigates the hierarchical structure of discourse in dialogues and its potential for depression detection. To mitigate the issue of sparse data, we draw inspiration from the Multi-Task Learning (MTL) framework and learn features jointly from multiple related tasks. We consider three auxiliary tasks: *emotion classification*, *dialogue act*, and *topic classification*, to explore how shallow information about dialogue structure can enhance performance. We adopt a classic *hard-parameter sharing* architecture, which is simpler than the shared-private architecture used in Qureshi et al. (2020) but has proven effective. To incorporate dialogue organization, we propose a dialogue-specific hierarchical architecture, where two tasks (emotion and dialogue act classification) are performed at the speech turn level,

while two others (depression detection and topic classification) are performed at the document level. We observe significant improvements when adding each task separately. Jointly learning all four tasks results in an improvement in all metrics (F score +27 points). Our ablation studies show that emotion and depression detection mutually benefit each other. The positive results for shallow markers, such as dialogue acts and topics, also indicate their relevance to the dialogue structure.

9.1.2 Discourse Structure Prediction

In part III, we aim to answer the second research question (RQ2) on generating full discourse structures in practical settings. Our work highlights the potential of the PLMs for both structure extraction and relation prediction tasks. PLMs exhibit excellent generalization abilities, and we demonstrate that with tailored fine-tuning tasks such as Sentence Ordering, we can improve the encoding of structural information in dialogues. Although the full parsing results presented in Chapter 8 are only 32.8, which is far from ideal, they represent a precious first step towards developing general full discourse parsers. Our proposed pipeline is, to the best of our knowledge, the first attempt to address this challenging task in the context of dialogue settings.

Structure Extraction from PLMs: The third project focuses on the extraction of discourse structure, particularly within the Segmented Discourse Representation Theory (SDRT) (Asher and Lascarides, 2003), which is commonly used in dialogue settings. The project pioneers the use of semi-supervised and unsupervised methods to address data scarcity issues in dialogues and extract discourse information from pre-trained language models (PLMs). We examine the robustness and locality of discourse structures in PLMs by analyzing the captured information across self-attention heads and diverse fine-tuning tasks. Choosing the best attention head is a critical issue when using PLMs to extract document-level discourse information. Experimental results on the STAC corpus (Asher et al., 2016) show that unsupervised and semi-supervised methods outperform a strong baseline (F_1 56.8%), delivering substantial gains on the complete dataset (F_1 59.3%) and further improvements on the tree-structured subset (F_1 68.1%). Qualitative analysis of inferred structures reveals that our model successfully predicts more than 82% of projective arcs, some of which span across four EDUs. This is encouraging, suggesting that our approach is capable of extracting reasonable discourse structures with minimal supervision.

Relation Prediction with Self-Training: The final project involves the second phase of discourse parsing, which is relation prediction. This work builds on the structure extraction project and focuses on leveraging pre-trained language models (PLMs) through self-training. We examine various techniques for selecting pseudo-labeled data, and find that selecting samples based solely on confidence scores is not sufficient. While self-training can enhance model performance, the improvement is modest (around 1 point). The key challenge of self-training lies in generating precise and diverse pseudo labels. To overcome this limitation, we investigate the potential of a "human-in-the-loop" strategy by providing gold annotation for uncertain examples with low-confidence scores. Our findings suggest that human efforts can be beneficial, but require a considerable amount of annotation. However, in practical settings, it can be difficult to obtain such extensive supervision.

9.2 Limitations & Perspectives

In Part II, we use conversations involving patients with mental disorders in order to learn language features associated with the disease. Our experiments replicate performances as high as previous studies in English (Mitchell et al., 2015; Kayi et al., 2017; Allende-Cid et al., 2019) for Schizophrenia identification, and surpass the previous SOTA models in depression detection (Mallol-Ragolta et al., 2019; Xezonaki et al., 2020).

Although the results are promising, there is still room for improvement. One of the challenges in studying language in Schizophrenia is the lack of interaction. In our preliminary studies, instead of using the speech of patients and controls for classification, we use psychologists' speech turn. However, psychologists tend to adapt their way of speaking when interacting with different participants, making it a potential source of bias. To avoid introducing further biases, we do not include psychologists' speech in model training, and our classification models only capture local contexts, which is not the best approach for modeling dialogue data. To address this limitation, neural networks could be used instead of classic probabilistic models. Special markers could be used to indicate the beginning and end of speech turns of different participants, thus considering a complete multi-speaker interaction.

To mitigate the bias while keeping the interaction, one possibility is to use **adversarial learning** within a neural model. In adversarial learning, an adversarial model is trained to maximize a loss function that is opposite to the original model's loss function. By introducing this adversarial component, the original model is forced to learn more generalizable features that are less susceptible to bias (Zhang et al., 2018a). We can draw inspiration from the work that tackles gender bias, as in Bordia and Bowman (2019); Liu et al. (2020). One potential solution could involve developing a component model that can distinguish whether the psychologist is communicating with a patient or not. By doing so, we can preserve the interaction while reducing the influence of any bias introduced by the psychologist's speech.

However exciting, there are some practical difficulties in implementing this idea. Firstly, the size of the corpus is a concern since our target dataset is extremely small, consisting of only 41 documents with an average length of approximately 260 speech turns (Rebuschi et al., 2014). Due to the lack of data, it is unlikely that the model can efficiently learn and converge, which could result in either overfitting or underfitting. Additionally, the opaque decision-making process of deep neural models (Iyer et al., 2018) presents another challenge, which could require more effort to interpret the results. While recent techniques have proposed various methods for interpretability (Linardatos et al., 2020), there is yet no consistent and credible approach.

Our second project focuses on depression detection and aims to address the interaction issue by using a hierarchical bi-LSTM model. The model first encodes each sentence and then captures interactions among sentences. This approach is made possible by the neutral data collection process in the DAIC-WOZ dataset (DeVault et al., 2014), where participants speak to an animated virtual interviewer using standardized questions. This dataset contains a larger number of documents (189) compared to our previous project on language in Schizophrenia, allowing for better training of the model. By using a hierarchical structure, we are able to demonstrate the correlation between depression and emotion and show the relevance of features such as dialogue acts and topics. However, our approach to modeling dialogue structure is over-simplified, relying solely on dialogue act prediction. In the auxiliary dataset DailyDialog (Li et al., 2017), dialogue acts are annotated into four broad categories, namely *inform*, *questions*, *directives*, and *commissives*. Although we expect the model to learn the bi-turn dialogue flow, such as *questions-inform* and *directives-commissives*, to partially reflect the structure of a document, these flows are in-

sufficient to reflect the true discourse structure due to the coarse granularity of the dialogue act annotations.

One possibility to incorporate structure information is to explore more detailed methods of **modeling dialogue structures**, potentially relying on discourse parsing. However, this approach poses a direct challenge due to the lack of general and robust discourse parsers. Parsers that are pre-trained on STAC corpus, such as *Deep sequential* (Shi and Huang, 2019) and *Structural-joint* (Chi and Rudnicky, 2022), have limitations in vocabulary and require careful domain adaptation strategies to be applied to other domains (Liu and Chen, 2021). Another challenge is designing a sub-task that can learn discourse structure. Discourse parsing is a complex task that involves `EDU` attachment and relation prediction. Incorporating such complex procedures into multi-task learning directly may be difficult. Surrogate tasks such as `EDU` attachment prediction could be considered, where the model predicts whether a pair of `EDUs` should be linked together.

We are interested in investigating the task of **classifying depression severity** as an extension to binary classification. A potential approach to achieve this is through a cascading structure, where the model first detects depression and subsequently performs severity classification. Cascading methods have not been extensively utilized in the mental disorder field, and we can take cues from the application of these methods in sentiment analysis on Twitter, as seen in Calvo and Juárez Gambino (2018). To ensure the robustness of our proposed method, we plan to refine our work and evaluate it using cross-validation splits of the data. This is particularly important as our dataset is scarce and may suffer from issues of representativeness. A further step will be to investigate the generalization of our model to other mental health disorders, hopefully with better structure modeling.

In Part III, while we show promising initial results on the ability to capture valid discourse structures from semi-supervised and self-training methods, the performance of our proposed methods is still limited, especially compared to fully supervised systems on the intra-domain parsing models: for link attachment 59% *versus* 74%; for full parsing 33 *versus* 59 (Chi and Rudnicky, 2022), calling for further improvements.

There are several unanswered questions that require further investigation in our future work. The first one pertains to enhancing the extracted discourse structure so that it aligns better with the SDRT-style graphs. In Chapter 7, we mainly focus on generating projective tree structures, which is consistent with previous studies (Muller et al., 2012; Afantenos et al., 2015). This approach covers the majority of the links ($\approx 94\%$) and can serve as a foundation for accurately inferring the remaining non-projective links in future work. In Section 7.8, we experiment with extending tree-like structures to graph-like structures by utilizing the "growing tree" strategy to add edges to the established tree structure. The improvement, however, is modest, increasing the F score by only one point. An alternative approach would be to re-implement the *Integer Linear Programming* methods presented in Perret et al. (2016) but with pre-trained language models as backbones.

The second open question concerns the two-step approach employed for discourse parsing, which is addressed separately in Chapter 7 and Chapter 8. Our current approach involves link attachment and subsequent relation prediction, but it is susceptible to error propagation. On the other hand, joint models for discourse parsing are not uncommon in the literature. For instance, in RST-style parsing, the CODRA framework (Joty et al., 2015) integrates the structure and label of a discourse tree constituent jointly in probabilistic discriminative parsing models (Conditional Random Fields). Feng and Hirst (2012) confirm that considering sequential dependencies improve the performance of the discourse parser. In Chapter 7, we discuss the limitations of tree extraction

algorithms like Eisner, which are unable to produce multi-incoming edges and, therefore, cannot capture structures like the *losange* shape. While it is challenging to predict such structures based on their structural properties alone, multi-outgoing and multi-incoming edges are often associated with specific relations like *Question-answer pair* and *Acknowledgment*. In this case, a simple proposition of the joint model is to augment the discourse structures with additional information. For instance, if a high-confidence relation is identified between two `EDUs` that have not been attached, we could perform post-hoc refinement and add the missing attachment back into the structure.

Thirdly, after showing all the synergistic downstream applications in Section 3.3, there is significant potential to apply our discourse parsers to new domains and apply them to other tasks. In automatic generation tasks, structural document-level representations of semantic relations have shown benefits in aiding abstractive dialogue summarization, as in Chen and Yang (2021). Given that discourse annotated corpora in English are limited to a few domains, mainly gaming (Asher et al., 2016) and online technical forum chats (Li et al., 2020), our semi-supervised approaches are currently the most effective resources to produce discourse structures for raw documents.

Finally, since we work with large language models and investigate every single attention head in structure extraction experiments, computational efficiency is a concern. We conduct experiments on a machine with 4 GPUs. For structure extraction, the calculation for one discourse tree on one head costs approximately 0.75 seconds (in STAC the averaged dialogue length is 11 `EDUs`), which quickly sums up to 4.5 hours with only 100 data points for all the candidate trees in one language model (192 in BART). When dealing with much longer documents, for example, AMI (Carletta et al., 2006) and conversational section in GUM (in average $>$ 200 utterances/dialogue) (Zeldes, 2017), our estimation shows that one dialogue takes up to $\approx$ 2 minutes, which means 6.5 hours for 192 candidate trees. Even though we use parallel computation, the exhaustive "head" computation results in a tremendous increase in time and running storage. If we would like to conduct similar experiments with long documents, the exhaustive research process should be optimized. One possibility is to investigate only those "discourse-rich" heads, mainly in the deeper layers, for future work.

9.3 Ethical Considerations

Given the sensitive nature of some of our experiments, involving cognitive impairment detection tasks and the utilization of large pre-trained language models, we find it essential to discuss the ethical implications of our work.

In the experiments regarding cognitive impairments detection (Chapter 4, 5), we claim that the goal of automatic systems is NOT to replace human healthcare providers. All these systems may be used only in support of human decisions. The principle of leaving the decision to the machine would imply major risks for decision-making in the health field, a mistake that in high-stakes healthcare settings could prove detrimental or even dangerous. Another issue is the representativeness of the data. Currently, it is very complex to access patients in order to have more examples. The institutional complexity leads researchers to systematically use the same data set, creating a bias between the representation of the pathology, in particular for mental ones whose expression can take very varied forms. This also implies defining a variation in relation to a normative use of language that comes with a strong risk in this type of approach.

As for the discourse parsing tasks (Chapter 7, 8), since we are investigating the nature of the discourse structures captured in large PLMs, our work can be seen as making these models

more transparent. This will hopefully contribute to avoiding unintended negative effects when the growing number of NLP applications relying on PLMs are deployed in practical settings.

On the resource level, we carefully select the dialogue corpora to control for potential biases, hate speech, and inappropriate language by using human-annotated corpora and professionally curated resources. We only work with interview transcription, with no audio or visual information. Further, we consider the privacy of dialogue partners in the selected datasets by replacing names with generic user tokens.

In terms of the environmental cost, the experiments described in Li et al. (2023) make use of Nvidia RTX 2080 Ti GPUs for tree extraction and Nvidia A100 GPUs for BART fine-tuning. We use up to 4 GPUs for the parallel computation. The experiments on structure extraction take up to 1.2 hours for one language model, and we test a dozen models. In the relation prediction experiments with self-training, we use 2 GPUs for the parallel computation on Nvidia A100 for a cumulative 80 hours. We note that while our work is based on exhaustive research on all the attention heads and parameter tuning in PLMs to obtain valuable insights, future work will able to focus more on discourse-rich heads, which can help to avoid the quadratic growth of computation time for longer documents.

Appendix A

Investigating Language Markers of Schizophrenia in Dialogues

A.1 Performance with Different Features and Window Settings

Features	SVM	LR	NB	RF	Perc	Best
bow	90.98	87.07	**93.66**	84.88	79.76	NB
ngram	**85.61**	83.66	65.61	71.95	75.12	SVM
OCR	60.5	**60.62**	45.83	59.02	53.14	LR
BC	**74.48**	54.44	68.19	70.41	61.91	SVM
Connectives	68.78	66.83	62.68	**72.44**	62.44	RF
Connectives D	64.63	63.22	53.17	**67.11**	60.98	RF
POS	50.49	49.51	**53.66**	50.00	50.24	NB
2-POS	**67.36**	58.64	49.88	59.71	56.01	SVM
3-POS	**71.65**	68.04	55.46	63.47	60.93	SVM
1-2-3-POS	**69.01**	55.19	50.28	62.74	54.05	SVM
2-*treelet*	67.34	66.04	50.63	**69.19**	56.05	RF
3-*treelet*	**66.78**	65.51	53.94	60.52	62.27	SVM
POS + 2-3-*treelet*	**66.59**	60.98	58.05	65.85	52.20	SVM
POS + OCR	59.9	58.18	46.96	49.71	54.37	SVM
2-POS + OCR	**65.19**	59.59	51.36	56.99	53.48	SVM
3-POS + OCR	**67.62**	59.78	56.11	62.19	60.74	SVM
POS + BC	65.11	61.12	53.95	**69.46**	63.96	RF
2-POS + BC	**77.54**	64.77	56.32	64.49	63.76	SVM
3-POS + BC	**74.93**	67.17	58.79	63.82	68.8	SVM
2-*treelet* + BC	79.03	68.93	54.29	**70.86**	67.15	RF
3-*treelet* + BC	**74.28**	69.13	57.78	61.14	67.09	SVM
OCR + BC	69.67	64.44	46.94	**72.44**	59.47	RF
Conn + BC	71.53	62.22	68.85	**75.85**	68.41	RF
OCR + ConnD	70.23	**71.28**	50.54	60.6	65.86	LR
POS + ConnD	**70.04**	69.19	55.24	55.72	65.87	SVM
2-POS + ConnD	**71.73**	70.86	56.91	58.77	68.09	SVM
3-POS + ConnD	**74.52**	70.36	57.68	62.05	69.31	SVM
2-*treelet* + ConnD	**75.17**	72.66	58.18	67.61	70.86	SVM
3-*treelet* + ConnD	**76.61**	72.69	56.58	57.93	69.15	SVM

Table A.1: Full setting results with individual and combination features using 5 classifieurs. LR = Logistic Regression; NB = Naive Bayes; RF = Random Forest; Perc = Perceptron. "Connectives D" or "ConnD": desambiguated connectives.

Features	SVM	LR	NB	RF	Perc	Best
bow	70.20	70.29	**72.43**	64.65	52.53	NB
ngram	**69.59**	68.99	66.99	66.45	51.52	SVM
OCR	50.04	49.93	50.0	49.97	**50.17**	Perc
BC	53.73	50.32	54.49	**54.80**	50.69	RF
Connective	55.2	**55.28**	52.38	54.99	50.73	LR
Connective D	53.75	**53.79**	50.78	53.27	50.61	LR
POS	55.32	**55.8**	50.0	53.4	50.69	LR
2-POS	56.24	**56.33**	50.22	54.51	51.21	LR
3-POS	56.53	**56.53**	52.55	54.82	50.8	LR
1-2-3-POS	58.24	**58.36**	53.34	55.33	51.27	LR
2-*treelet*	56.58	**56.73**	51.03	54.35	51.33	LR
3-*treelet*	55.25	**55.34**	53.12	54.68	50.97	LR
POS + 2-3-*treelet*	**57.77**	57.35	53.96	51.46	wait	SVM
POS + OCR	**55.4**	55.39	50.0	53.61	50.74	SVM
2-POS + OCR	56.29	**56.33**	50.31	54.38	51.32	LR
3-POS + OCR	**56.58**	56.57	53.14	54.74	50.46	SVM
POS + BC	55.92	**56.0**	51.26	53.81	51.16	LR
2-POS + BC	57.21	**57.38**	51.99	55.58	51.37	LR
3-POS + BC	57.3	**57.46**	54.35	54.33	54.93	LR
2-*treelet* + BC	57.34	**57.55**	53.24	54.98	51.59	LR
3-*treelet* + BC	**56.23**	55.74	53.92	54.89	51.52	SVM
OCR + BC	54.29	54.52	50.29	**54.78**	50.45	RF
Conn + BC	55.99	53.19	**56.51**	54.9	51.06	NB
OCR + ConnD	**53.82**	53.81	51.19	53.32	50.5	SVM
POS + ConnD	55.38	**55.46**	51.31	53.98	50.95	LR
2-POS + ConnD	56.49	**56.75**	51.77	54.48	51.34	LR
3-POS + ConnD	57.05	**57.11**	53.35	54.95	50.98	LR
2-*treelet* + ConnD	56.62	**56.75**	52.76	54.58	51.62	LR
3-*treelet* + ConnD	55.94	**56.04**	53.45	54.72	50.87	LR

Table A.2: Indiv. setting results with individual and combination features using 5 classifieurs. LR = Logistic Regression; NB = Naive Bayes; RF = Random Forest; Perc = Perceptron. "Connectives D" or "ConnD": desambiguated connectives.

Features	SVM	LR	NB	RF	Perc	Best
OCR	52.41	**52.43**	52.25	52.36	51.2	LR
BC	61.77	62.0	55.73	**62.01**	53.9	RF
Connective	**64.05**	63.97	54.45	62.81	56.34	SVM
Connective D	58.45	**58.61**	53.97	57.62	53.5	LR
POS	57.84	58.03	50.0	**60.63**	51.7	RF
2-POS	63.7	**64.85**	53.26	59.29	56.55	LR
3-POS	**65.39**	65.35	61.73	59.96	58.11	SVM
1-2-3-POS	65.62	**66.19**	59.05	60.88	55.88	LR
2-*treelet*	64.71	**65.02**	54.99	60.7	56.35	LR
3-*treelet*	63.21	**63.95**	57.52	60.21	55.52	LR
POS + 2-3-*treelet*	65.17	**65.52**	58.86	61.7	56.0	LR
POS + OCR	58.79	58.67	52.11	**60.75**	52.34	RF
2-POS + OCR	63.84	**65.09**	54.04	59.69	57.65	LR
3-POS + OCR	**65.7**	65.68	62.0	60.14	58.0	SVM
POS + BC	**63.76**	63.67	56.52	63.6	54.29	SVM
2-POS + BC	68.64	**68.88**	59.56	62.32	59.02	LR
3-POS + BC	69.46	**69.92**	65.24	61.64	59.23	LR
2-*treelet* + BC	68.59	**68.73**	59.75	63.26	58.41	LR
3-*treelet* + BC	67.81	**68.16**	60.98	62.02	57.37	LR
OCR + BC	62.27	**62.72**	56.6	62.07	54.16	LR
Conn + BC	**67.1**	66.8	60.13	64.43	56.92	SVM
OCR + ConnD	59.29	**59.68**	54.7	58.4	53.15	LR
POS + ConnD	58.98	59.14	54.49	**62.13**	52.99	RF
2-POS + ConnD	**64.43**	64.3	58.11	59.89	57.11	SVM
3-POS + ConnD	**65.05**	64.93	61.66	60.28	57.57	SVM
2-*treelet* + ConnD	64.2	**64.47**	57.79	61.39	56.91	LR
3-*treelet* + ConnD	63.66	**63.7**	57.94	61.05	56.08	LR

Table A.3: W-128 setting results with individual and combination features using 5 classifieurs. LR = Logistic Regression; NB = Naive Bayes; RF = Random Forest; Perc = Perceptron. "Connectives D" or "ConnD": desambiguated connectives.

Features	SVM	LR	NB	RF	Perc	Best
OCR	54.58	54.84	**55.19**	55.14	50.93	NB
BC	66.19	66.54	55.09	**66.89**	55.99	RF
Connectives	**69.68**	68.89	53.66	66.54	57.89	SVM
Connectives D	**65.13**	63.73	57.11	59.06	56.95	SVM
POS	58.0	58.26	50.0	**60.48**	51.75	RF
2-POS	66.07	**68.53**	51.76	61.75	58.73	LR
3-POS	70.48	**70.66**	62.66	62.46	62.65	LR
1-2-3-POS	71.72	**72.03**	58.24	62.74	57.21	LR
2-*treelet*	**70.11**	70.01	53.57	64.46	60.02	SVM
3-*treelet*	**66.39**	66.28	59.27	61.26	58.44	SVM
POS + 2-3-*treelet*	68.47	**69.11**	60.45	63.23	57.61	LR
POS + OCR	58.77	59.3	54.57	**61.63**	53.12	RF
2-POS + OCR	65.83	**68.42**	55.09	62.26	57.85	LR
3-POS + OCR	70.57	**70.84**	63.84	62.85	62.27	LR
POS + BC	67.3	67.49	55.7	**67.89**	56.05	RF
2-POS + BC	72.43	**72.76**	60.05	66.29	60.34	LR
3-POS + BC	73.29	**73.75**	66.88	64.54	62.41	LR
2-*treelet* + BC	**73.05**	72.83	60.86	66.34	60.32	SVM
3-*treelet* + BC	**70.96**	70.82	62.69	65.13	61.1	SVM
OCR + BC	67.68	**68.1**	59.43	67.7	56.81	LR
Conn + BC	**71.63**	70.9	59.97	68.57	59.59	SVM
OCR + ConnD	**66.91**	65.49	58.74	60.86	57.67	SVM
POS + ConnD	**65.78**	64.82	57.62	62.05	57.12	SVM
2-POS + ConnD	**69.12**	68.68	61.25	61.86	60.87	SVM
3-POS + ConnD	**72.54**	71.98	65.9	63.16	63.29	SVM
2-*treelet* + ConnD	**68.87**	68.05	60.24	64.37	61.03	SVM
3-*treelet* + ConnD	**69.28**	68.25	62.07	61.41	61.07	SVM

Table A.4: W-256 setting results with individual and combination features using 5 classifieurs. LR = Logistic Regression; NB = Naive Bayes; RF = Random Forest; Perc = Perceptron. "Connectives D" or "ConnD": desambiguated connectives.

Features	SVM	LR	NB	RF	Perc	Best
OCR	59.27	58.83	**59.28**	58.52	53.47	NB
BC	64.7	59.78	64.6	**67.86**	58.25	RF
Connectives	71.87	70.77	52.81	**73.57**	61.01	RF
Connectives D	**67.15**	66.92	58.35	64.91	60.59	SVM
POS	**60.09**	59.61	50.0	59.15	52.43	SVM
2-POS	71.48	**71.74**	50.43	60.12	59.3	LR
3-POS	**72.55**	71.99	64.08	62.41	64.31	SVM
1-2-3-POS	**72.67**	71.84	55.26	61.41	55.83	SVM
2-*treelet*	72.88	**74.19**	52.62	63.99	59.08	LR
3-*treelet*	69.01	**69.03**	61.65	62.15	60.73	LR
POS + 2-3-*treelet*	71.8	**72.39**	60.11	63.48	57.37	LR
POS + OCR	61.71	60.13	58.83	**61.98**	53.45	RF
2-POS + OCR	70.21	**71.49**	59.33	61.06	58.99	LR
3-POS + OCR	**73.01**	72.67	66.45	62.4	64.24	SVM
POS + BC	67.23	67.0	59.28	**68.85**	58.07	RF
2-POS + BC	**76.6**	75.74	61.34	64.11	63.45	SVM
3-POS + BC	**77.86**	77.58	68.75	62.8	65.6	SVM
2-*treelet* + BC	**75.2**	74.65	61.42	64.79	62.4	SVM
3-*treelet* + BC	**73.6**	73.46	64.83	62.66	64.61	SVM
OCR + BC	68.16	66.99	64.73	**68.32**	60.5	RF
Conn + BC	**74.42**	62.71	74.28	74.01	62.37	SVM
OCR + ConnD	69.63	**69.93**	61.61	67.62	64.65	LR
POS + ConnD	67.16	**67.69**	58.82	65.3	60.51	LR
2-POS + ConnD	**70.53**	70.29	61.6	63.83	63.53	SVM
3-POS + ConnD	72.55	**72.68**	66.21	64.46	65.73	LR
2-*treelet* + ConnD	70.01	**71.25**	62.03	68.65	64.6	LR
3-*treelet* + ConnD	**69.67**	69.44	62.22	65.4	64.19	SVM

Table A.5: W-512 setting results with individual and combination features using 5 classifieurs. LR = Logistic Regression; NB = Naive Bayes; RF = Random Forest; Perc = Perceptron. "Connectives D" or "ConnD": desambiguated connectives.

Features	SVM	LR	NB	RF	Perc	Best
OCR	64.58	64.58	62.87	**67.26**	55.6	RF
BC	63.0	**63.82**	59.89	61.87	58.28	LR
Connectives	75.8	**76.73**	51.11	76.42	61.0	LR
Connectives D	70.12	**70.67**	58.18	63.38	67.48	LR
POS	56.95	54.55	50.0	**57.18**	51.45	RF
2-POS	70.55	**71.11**	50.0	60.87	56.94	LR
3-POS	**71.71**	71.1	63.76	60.55	65.8	SVM
1-2-3-POS	**72.52**	67.85	51.27	58.67	53.36	SVM
2-*treelet*	**74.63**	73.58	50.94	65.08	59.08	SVM
3-*treelet*	**70.31**	70.04	66.8	66.14	64.33	SVM
POS + 2-3-*treelet*	**71.43**	68.71	61.68	65.69	58.39	SVM
POS + BC	63.65	**65.19**	60.17	61.78	59.51	LR
2-POS + BC	72.32	**73.76**	61.67	61.24	60.94	LR
3-POS + BC	75.13	**75.2**	70.35	61.2	69.34	LR
2-*treelet* + BC	72.25	**72.5**	61.54	64.33	64.01	LR
3-*treelet* + BC	71.85	**73.2**	71.24	63.33	65.93	LR
OCR + BC	67.27	**69.61**	69.42	64.74	59.66	LR
Conn + BC	73.55	75.06	63.78	**75.09**	66.32	RF
OCR + ConnD	72.36	**73.58**	62.31	66.25	66.57	LR
POS + ConnD	**71.9**	70.75	58.88	64.25	66.04	SVM
2-POS + ConnD	72.06	**72.72**	62.37	62.91	68.81	LR
3-POS + ConnD	73.48	**74.23**	70.13	62.49	71.79	LR
2-*treelet* + ConnD	73.9	**74.21**	60.68	66.97	70.58	LR
3-*treelet* + ConnD	**75.74**	75.57	65.12	65.64	70.67	SVM

Table A.6: W-1024 setting results with individual and combination features using 5 classifieurs. LR = Logistic Regression; NB = Naive Bayes; RF = Random Forest; Perc = Perceptron. "Connectives D" or "ConnD": desambiguated connectives.

A.2 Hyper-Parameters

Full setting				
Feature	Accuracy	Algo	Hyper-parameter	Threshold
bow	93.66	NB	$\alpha = 0.001$	9
ngram	85.61	SVM	$C = 5$	4
OCR	60.62	LR	$C = 0.001$	$1e-5$
BC	74.48	SVM	$C = 0.1$	9
Connectives	72.44	RF	$max\_depth = 2$	2
POS	53.66	NB	$\alpha = 0.001$	5
2-POS	67.36	SVM	$C = 100$	3
3-POS	71.65	SVM	$C = 100$	2
2-*treelet*	69.19	RF	$max\_depth = 2$	$1e-5$
3-*treelet*	66.78	SVM	$C = 100$	8
1-2-3-POS	69.01	SVM	$C = 1000$	1
POS+2-3-*treelet*	66.59	SVM	$C = 1000$	4
3-POS + BC	74.93	SVM	$C = 100$	8

Table A.7: Best scores (averaged accuracy), best algorithms (Algo), corresponding hyper-parameters, and thresholds for full documents (Full) setting.

Indiv. setting				
Feature	Accuracy	Algo	Hyper-parameter	Threshold
bow	72.43	NB	$\alpha = 0.1$	$1e-5$
ngram	69.59	SVM	$C = 5$	2
OCR	50.17	PERC	$\alpha = 0.001$	mean
BC	54.79	RF	$max\_depth = 2$	$1e-5$
Connectives	55.28	LR	$C = 100$	5
POS	55.80	LR	$C = 1$	$1e-5$
2-POS	56.33	LR	$C = 1$	$1e-5$
3-POS	56.53	SVM	$C = 0.5$	$1e-5$
2-*treelet*	56.73	LR	$C = 5$	$1e-5$
3-*treelet*	55.34	LR	-	$1e-5$
1-2-3-POS	58.36	LR	$C = 100$	$1e-5$
POS+2-3-*treelet*	57.77	SVM	$C = 0.5$	$1e-5$
3-POS + BC	57.46	LR	$C = 5$	$1e-5$

Table A.8: Best scores (averaged accuracy), best algorithms (Algo), corresponding hyper-parameters, and thresholds for individual documents (Indiv) setting.

W-512 *setting*				
Feature	Accuracy	Algo	Hyper-parameter	Threshold
OCR	59.28	NB	$\alpha = 0.001$	2
BC	67.86	RF	$max\_depth = None$	1
Connectives	73.57	RF	$max\_depth = 2$	median
POS	60.09	SVM	$C = 100$	$1e-5$
2-POS	71.74	LR	$C = 100$	mean
3-POS	72.55	SVM	$C = 100$	1
2-*treelet*	74.19	LR	$C = 100$	6
3-*treelet*	69.03	LR	$C = 100$	6
1-2-3-POS	72.67	SVM	$C = 100$	7
POS+2-3-*treelet*	72.39	LR	$C = 100$	4
3-POS + BC	77.86	SVM	$C = 100$	$1e-5$

Table A.9: Best scores (averaged accuracy), best algorithms (Algo), corresponding hyper-parameters, and thresholds for window size 512 tokens (W-512) setting.

W-1024 *setting*				
Feature	Accuracy	Algo	Hyper-parameter	Threshold
OCR	67.26	RF	$max\_depth = 2$	5
BC	63.82	LR	$C = 100$	2
Connectives	76.73	LR	$C = 100$	mean
POS	57.18	RF	$max\_depth = None$	1
2-POS	71.11	LR	$C = 100$	7
3-POS	71.71	SVM	$C = 5$	$1e - 5$
2-*treelet*	74.63	SVM	$C = 100$	7
3-*treelet*	70.31	SVM	$C = 100$	7
1-2-3-POS	72.52	SVM	$C = 1000$	5
POS+2-3-*treelet*	71.43	SVM	$C = 100$	7
3-POS + BC	75.20	LR	$C = 100$	1

Table A.10: Best scores (averaged accuracy), best algorithms (Algo), corresponding hyper-parameters, and thresholds for window size 1024 tokens (W-1024) setting.

Bibliography

Afantenos, S., Asher, N., Benamara, F., Bras, M., Fabre, C., Ho-Dac, L.-M., Le Draoulec, A., Muller, P., Pery-Woodley, M.-P., Prévot, L., et al. (2012a). An empirical resource for discovering cognitive principles of discourse organisation: the annodis corpus. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2727–2734.

Afantenos, S., Asher, N., Benamara, F., Cadilhac, A., Dégremont, C., Denis, P., Guhe, M., Keizer, S., Lascarides, A., Lemon, O., et al. (2012b). Modelling strategic conversation: model, annotation design and corpus. In *Proceedings of the 16th Workshop on the Semantics and Pragmatics of Dialogue (Seinedial), Paris*.

Afantenos, S., Kow, E., Asher, N., and Perret, J. (2015). Discourse parsing for multi-party chat dialogues. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 928–937, Lisbon, Portugal. Association for Computational Linguistics.

Akhtar, S., Ghosal, D., Ekbal, A., Bhattacharyya, P., and Kurohashi, S. (2019). All-in-one: Emotion, sentiment and intensity prediction using a multi-task ensemble framework. *IEEE transactions on affective computing*.

Al Hanai, T., Ghassemi, M. M., and Glass, J. R. (2018). Detecting depression with audio/text sequence modeling of interviews. In *Interspeech*, pages 1716–1720.

Al-Saif, A. and Markert, K. (2011). Modelling discourse relations for arabic. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 736–747.

Alemany, L. A., Masalles, I. C., and Cirera, L. P. (2002). Lexicón computacional de marcadores de discurso. *Procesamiento del lenguaje natural*, 29.

Allende-Cid, H., Zamora, J., Alfaron-Faccio, P., and Alonso, M. (2019). A machine learning approach for the automatic classification of schizophrenic discourse. *IEEE Access*, pages 45544–45554.

Almeida, F. and Xexéo, G. (2019). Word embeddings: A survey. *arXiv preprint arXiv:1901.09069*.

Alt, C., Hübner, M., and Hennig, L. (2019). Fine-tuning pre-trained transformer language models to distantly supervised relation extraction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1388–1398.

Amblard, M., Braud, C., Li, C., Demily, C., Franck, N., and Musiol, M. (2020). Investigation par méthodes d'apprentissage des spécificités langagières propres aux personnes avec schizophrénie

(investigating learning methods applied to language specificity of persons with schizophrenia). In *Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Volume 2: Traitement Automatique des Langues Naturelles*, pages 12–26.

Amblard, M., Fort, K., Musiol, M., and Rebuschi, M. (2014). L'impossibilité de l'anonymat dans le cadre de l'analyse du discours. In *Journée ATALA éthique et TAL*, Paris, France.

Amblard, M. and Pogodalla, S. (2014). Modeling the dynamic effects of discourse: Principles and frameworks. *Interdisciplinary Works in Logic, Epistemology, Psychology and Linguistics: Dialogue, Rationality, and Formalism*, pages 247–282.

Angelidis, S. and Lapata, M. (2018). Multiple instance learning networks for fine-grained sentiment analysis. *Transactions of the Association for Computational Linguistics*, 6:17–31.

APA, A. P. A. (2015). *DSM-5-Manuel diagnostique et statistique des troubles mentaux*. Elsevier Masson.

Arora, S., Liang, Y., and Ma, T. (2017). A simple but tough-to-beat baseline for sentence embeddings. In *International conference on learning representations*.

Arslan, Y., Allix, K., Veiber, L., Lothritz, C., Bissyandé, T. F., Klein, J., and Goujon, A. (2021). A comparison of pre-trained language models for multi-class text classification in the financial domain. In *Companion Proceedings of the Web Conference 2021*, pages 260–268.

Asher, N. (1993). *Reference to abstract objects in discourse*, volume 50. Springer Science & Business Media.

Asher, N., Hunter, J., Morey, M., Farah, B., and Afantenos, S. (2016). Discourse structure and dialogue acts in multiparty dialogue: the STAC corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2721–2727, Portorož, Slovenia. European Language Resources Association (ELRA).

Asher, N. and Lascarides, A. (2003). *Logics of conversation*. Cambridge University Press.

Badene, S. (2021). *Weak supervision for learning discourse structure in multi-party dialogues*. PhD thesis, Université Paul Sabatier-Toulouse III.

Badene, S., Thompson, K., Lorré, J.-P., and Asher, N. (2019a). Data programming for learning discourse structure. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 640–645, Florence, Italy. Association for Computational Linguistics.

Badene, S., Thompson, K., Lorré, J.-P., and Asher, N. (2019b). Weak supervision for learning discourse structure. In *EMNLP*.

Baly, R., Da San Martino, G., Glass, J., and Nakov, P. (2020). We can detect your bias: Predicting the political ideology of news articles. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4982–4991.

Baroni, M., Dinu, G., and Kruszewski, G. (2014). Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 238–247.

Baxter, J. (1997). A bayesian/information theoretic model of learning to learn via multiple task sampling. *Machine learning*, 28:7–39.

Beltagy, I., Peters, M. E., and Cohan, A. (2020). Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Benamara, F. and Taboada, M. (2015). Mapping different rhetorical relation annotations: A proposal. In *Fourth Joint Conference on Lexical and Computational Semantics (* SEM 2015)*, pages 147–152.

Bender, E. M. and Koller, A. (2020). Climbing towards nlu: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 5185–5198.

Bengio, Y., Ducharme, R., and Vincent, P. (2000). A neural probabilistic language model. *Advances in neural information processing systems*, 13.

Benton, A., Mitchell, M., and Hovy, D. (2017). Multitask learning for mental health conditions with limited social media data. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 152–162, Valencia, Spain. Association for Computational Linguistics.

Benz, A. and Salfner, F. (2011). Discourse relations and relevance implicatures: A case study. In *Logic, Language, and Computation: 8th International Tbilisi Symposium on Logic, Language, and Computation, TbiLLC 2009, Bakuriani, Georgia, September 21-25, 2009. Revised Selected Papers 8*, pages 182–196. Springer.

Berger, A., Della Pietra, S. A., and Della Pietra, V. J. (1996). A maximum entropy approach to natural language processing. *Computational linguistics*, 22(1):39–71.

Bhatia, P., Ji, Y., and Eisenstein, J. (2015). Better document-level sentiment analysis from RST discourse parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2212–2218, Lisbon, Portugal. Association for Computational Linguistics.

Bingel, J. and Søgaard, A. (2017). Identifying beneficial task relations for multi-task learning in deep neural networks. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 164–169, Valencia, Spain. Association for Computational Linguistics.

Birnbaum, M. L., Ernala, S. K., Rizvi, A. F., De Choudhury, M., and Kane, J. M. (2017a). A collaborative approach to identifying social media markers of schizophrenia by employing machine learning and clinical appraisals. *J Med Internet Res*, 19(8):e289.

Birnbaum, M. L., Ernala, S. K., Rizvi, A. F., De Choudhury, M., and Kane, J. M. (2017b). A collaborative approach to identifying social media markers of schizophrenia by employing machine learning and clinical appraisals. *Journal of medical Internet research*, 19(8):e289.

Black, E., Abney, S., Flickinger, D., Gdaniec, C., Grishman, R., Harrison, P., Hindle, D., Ingria, R., Jelinek, F., Klavans, J. L., et al. (1991). A procedure for quantitatively comparing the syntactic coverage of english grammars. In *Speech and Natural Language: Proceedings of a Workshop Held at Pacific Grove, California, February 19-22, 1991*.

Blakemore, D. (1987). Semantic constraints on relevance.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Blum, A. and Mitchell, T. (1998). Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory*, pages 92–100.

Bois, J. W. D., Wallace, L., Meyer, C., Thompson, S. A., Englebretson, R., and Martey, N. (2000). Santa barbara corpus of spoken american english, parts 1-4. *Philadelphia: Linguistic Data Consortium*.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5:135–146.

Bommasani, R., Davis, K., and Cardie, C. (2020). Interpreting pretrained contextualized representations via reductions to static embeddings. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4758–4781.

Bordia, S. and Bowman, S. (2019). Identifying and reducing gender bias in word-level language models. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 7–15.

Braud, C. (2015). *Identification automatique des relations discursives implicites à partir de corpus annotés et de données brutes*. PhD thesis, Universite Paris Diderot-Paris VII.

Braud, C., Coavoux, M., and Søgaard, A. (2017). Cross-lingual rst discourse parsing. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 292–304.

Braud, C. and Denis, P. (2014). Combining natural and artificial examples to improve implicit discourse relation identification. In *coling*.

Brown, P. F., Della Pietra, V. J., Desouza, P. V., Lai, J. C., and Mercer, R. L. (1992). Class-based n-gram models of natural language. *Computational linguistics*, 18(4):467–480.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Bunt, H., Alexandersson, J., Choe, J.-W., Fang, A. C., Hasida, K., Petukhova, V., Popescu-Belis, A., and Traum, D. R. (2012). Iso 24617-2: A semantically-based standard for dialogue annotation. In *LREC*, pages 430–437.

Bunt, H. and Prasad, R. (2016). Iso dr-core (iso 24617-8): Core concepts for the annotation of discourse relations. In *Proceedings 12th joint ACL-ISO workshop on interoperable semantic annotation (ISA-12)*, pages 45–54.

Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J. N., Lee, S., and Narayanan, S. S. (2008). Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335–359.

Calvo, H. and Juárez Gambino, O. (2018). Cascading classifiers for twitter sentiment analysis with emotion lexicons. In *Computational Linguistics and Intelligent Text Processing: 17th International Conference, CICLing 2016, Konya, Turkey, April 3–9, 2016, Revised Selected Papers, Part II 17*, pages 270–280. Springer.

Cao, S., da Cunha, I., and Iruskieta, M. (2018). The rst spanish-chinese treebank. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 156–166.

Cao, S., Xue, N., da Cunha, I., Iruskieta, M., and Wang, C. (2017). Discourse segmentation for building a rst chinese treebank. In *Proceedings of the 6th Workshop on Recent Advances in RST and Related Formalisms*, pages 73–81.

Card, D., Boydstun, A., Gross, J. H., Resnik, P., and Smith, N. A. (2015). The media frames corpus: Annotations of frames across issues. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 438–444.

Carletta, J., Ashby, S., Bourban, S., Flynn, M., Guillemot, M., Hain, T., Kadlec, J., Karaiskos, V., Kraaij, W., Kronenthal, M., et al. (2006). The ami meeting corpus: A pre-announcement. In *International workshop on machine learning for multimodal interaction*, pages 28–39. Springer.

Carlson, L. and Marcu, D. (2001). Discourse tagging reference manual. *ISI Technical Report ISI-TR-545*, 54(2001):56.

Carlson, L., Marcu, D., and Okurovsky, M. E. (2001). Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory. In *Proceedings of the Second SIGdial Workshop on Discourse and Dialogue*.

Carlson, L., Okurowski, M. E., and Marcu, D. (2002a). *RST discourse treebank*. Linguistic Data Consortium, University of Pennsylvania.

Carlson, L., Okurowski, M. E., and Marcu, D. (2002b). *RST discourse treebank*. Linguistic Data Consortium, University of Pennsylvania.

Caruana, R. (1993). Multitask learning: A knowledge-based source of inductive bias. In *Machine learning: Proceedings of the tenth international conference*, pages 41–48.

Caruana, R. (1997). Multitask learning. *Machine learning*, 28(1):41–75.

Cañete, J., Chaperon, G., Fuentes, R., Ho, J.-H., Kang, H., and Pérez, J. (2020). Spanish pre-trained bert model and evaluation data. In *PML4DC at ICLR 2020*.

Cepoiu, M., McCusker, J., Cole, M. G., Sewitch, M., Belzile, E., and Ciampi, A. (2008). Recognition of depression by non-psychiatric physicians—a systematic literature review and meta-analysis. *Journal of general internal medicine*, 23(1):25–36.

Cer, D., Yang, Y., Kong, S.-y., Hua, N., Limtiaco, N., John, R. S., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., et al. (2019). Universal sentence encoder.

Cerisara, C., Jafaritazehjani, S., Oluokun, A., and Le, H. T. (2018). Multi-task dialog act and sentiment recognition on mastodon. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 745–754, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Chafe, W. (1982). Integration and involvement in speaking, writing, and oral literature. *Spoken and written language: Exploring orality and literacy*, pages 35–54.

Chai, J. and Jin, R. (2004). Discourse structure for context question answering. In *Proceedings of the Workshop on Pragmatics of Question Answering at HLT-NAACL 2004*, pages 23–30.

Chaika, E. (1974). A linguist looks at "schizophrenic" language. *Brain and language*, 1(3):257–276.

Chancellor, S., Birnbaum, M. L., Caine, E. D., Silenzio, V. M., and De Choudhury, M. (2019). A taxonomy of ethical tensions in inferring mental health states from social media. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 79–88.

Chatterjee, A., Gupta, U., Chinnakotla, M. K., Srikanth, R., Galley, M., and Agrawal, P. (2019). Understanding emotions in text using deep learning and big data. *Computers in Human Behavior*, 93:309–317.

Chen, J. and Yang, D. (2021). Structure-aware abstractive conversation summarization via discourse and action graphs. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1380–1391, Online. Association for Computational Linguistics.

Chen, S.-Y., Hsu, C.-C., Kuo, C.-C., Ku, L.-W., et al. (2018). Emotionlines: An emotion corpus of multi-party conversations. *arXiv preprint arXiv:1802.08379*.

Chi, T.-C. and Rudnicky, A. (2022). Structured dialogue discourse parsing. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 325–335.

Chowdhury, S. B. R., Brahman, F., and Chaturvedi, S. (2021). Is everything in order? a simple way to order sentences. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10769–10779.

Chu, Y.-J. (1965). On the shortest arborescence of a directed graph. *Scientia Sinica*, 14:1396–1400.

Chung, C. and Pennebaker, J. W. (2007). The psychological functions of function words. In Fiedler, K., editor, *Social Communication*, volume 1, chapter 12, pages 343–359. Psychology Press.

Clark, K., Khandelwal, U., Levy, O., and Manning, C. D. (2019). What does bert look at? an analysis of bert's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286.

Clark, K., Luong, M.-T., Le, Q. V., and Manning, C. D. (2020). Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

Collobert, R. and Weston, J. (2008). A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pages 160–167.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, É., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Conneau, A. and Lample, G. (2019). Cross-lingual language model pretraining. *Advances in neural information processing systems*, 32.

Coppersmith, G., Dredze, M., Harman, C., Hollingshead, K., and Mitchell, M. (2015). Clpsych 2015 shared task: Depression and ptsd on twitter. In *Proceedings of the 2nd workshop on computational linguistics and clinical psychology: from linguistic signal to clinical reality*, pages 31–39.

Coppersmith, G., Leary, R., Crutchley, P., and Fine, A. (2018). Natural language processing of social media as screening for suicide risk. *Biomedical informatics insights*, 10:1178222618792860.

Coppersmith, G., Ngo, K., Leary, R., and Wood, A. (2016). Exploratory analysis of social media prior to a suicide attempt. In *Proceedings of the third workshop on computational linguistics and clinical psychology*, pages 106–117.

Crible, L. and Cuenca, M.-J. (2017). Discourse markers in speech: characteristics and challenges for corpus annotation. *Dialogue and Discourse*, 8(2):149–166.

Croft, B. and Lafferty, J. (2003). *Language modeling for information retrieval*, volume 13. Springer Science & Business Media.

Da Cunha, I., Torres-Moreno, J.-M., and Sierra, G. (2011). On the development of the rst spanish treebank. In *Proceedings of the 5th Linguistic Annotation Workshop*, pages 1–10.

Dai, Z., Yang, Z., Yang, Y., Carbonell, J. G., Le, Q., and Salakhutdinov, R. (2019). Transformer-xl: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2978–2988.

Danlos, L., Antolinos-Basso, D., Braud, C., and Roze, C. (2012). Vers le fdtb: French discourse tree bank. In *TALN 2012: 19ème conférence sur le Traitement Automatique des Langues Naturelles*, volume 2, pages 471–478. ATALA/AFCP.

Davidson, D. (2001). *Essays on Actions and Events: Philosophical Essays Volume 1*. Clarendon Press.

De Choudhury, M., Counts, S., and Czerwinski, M. (2011). Identifying relevant social media content: leveraging information diversity and user cognition. In *Proceedings of the 22nd ACM conference on Hypertext and hypermedia*, pages 161–170.

De Choudhury, M., Counts, S., and Horvitz, E. (2013a). Social media as a measurement tool of depression in populations. In *Proceedings of the 5th Annual ACM Web Science Conference*, pages 47–56.

De Choudhury, M., Gamon, M., Counts, S., and Horvitz, E. (2013b). Predicting depression via social media. *Icwsm*, 13:1–10.

De Groote, P. (2006). Towards a montagovian account of dynamics. In *Semantics and linguistic theory*, volume 16, pages 1–16.

de Vries, W., van Cranenburgh, A., Bisazza, A., Caselli, T., van Noord, G., and Nissim, M. (2019). Bertje: A dutch bert model. *arXiv preprint arXiv:1912.09582*.

Demberg, V., Scholman, M. C., and Asr, F. T. (2019). How compatible are our discourse annotation frameworks? insights from mapping rst-dt and pdtb annotations. *Dialogue & Discourse*, 10(1):87–135.

Desai, S. and Durrett, G. (2020). Calibration of pre-trained transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 295–302.

Devatine, N., Muller, P., and Braud, C. (2022). Predicting political orientation in news with latent discourse structure to improve bias understanding. In *Proceedings of the 3rd Workshop on Computational Approaches to Discourse*, pages 77–85.

Devatine, N., Muller, P., and Braud, C. (2023). An integrated approach for political bias prediction and explanation based on discursive structure. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11196–11211.

DeVault, D., Artstein, R., Benn, G., Dey, T., Fast, E., Gainer, A., Georgila, K., Gratch, J., Hartholt, A., Lhommet, M., et al. (2014). Simsensei kiosk: A virtual human interviewer for healthcare decision support. In *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*, pages 1061–1068.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019a). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019b). Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT (1)*.

Dinan, E., Logacheva, V., Malykh, V., Miller, A., Shuster, K., Urbanek, J., Kiela, D., Szlam, A., Serban, I., Lowe, R., et al. (2020). The second conversational intelligence challenge (convai2). In *The NeurIPS'18 Competition: From Machine Learning to Intelligent Conversations*, pages 187–208. Springer.

Dinkel, H., Wu, M., and Yu, K. (2019). Text-based depression detection on sparse data. *arXiv preprint arXiv:1904.05154*.

Dipper, S. and Stede, M. (2006). Disambiguating potential connectives. In *Proceedings of KONVENS*, volume 6, pages 167–173. Citeseer.

Docherty, N. M., DeRosa, M., and Andreasen, N. C. (1996). Communication disturbances in schizophrenia and mania. *Archives of General Psychiatry*, 53(4):358–364.

Douki Dedieu, S., Ouali, U., and Nacef, F. (2012). Schizophrénie et genre. In Daléry, J., d'Amato, T., and Saoud, M., editors, *Pathologies schizophréniques*, Psychatrie, pages 199–205. Lavoisier.

Dozat, T. and Manning, C. D. (2016). Deep biaffine attention for neural dependency parsing. *arXiv preprint arXiv:1611.01734*.

Du, J., Grave, É., Gunel, B., Chaudhary, V., Celebi, O., Auli, M., Stoyanov, V., and Conneau, A. (2021). Self-training improves pre-training for natural language understanding. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5408–5418.

Dumais, S. T. et al. (2004). Latent semantic analysis. *Annu. Rev. Inf. Sci. Technol.*, 38(1):188–230.

Edmonds, J. (1968). Optimum branchings. *Mathematics and the Decision Sciences, Part*, 1(335-345):25.

Eisner, J. (1996). Three new probabilistic models for dependency parsing: An exploration. In *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*.

Ekman, P. (1999). Basic emotions. *Handbook of cognition and emotion*, 98(45-60):16.

Elazar, Y., Ravfogel, S., Jacovi, A., and Goldberg, Y. (2021). Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175.

Elvevåg, B., Foltz, P. W., Weinberger, D. R., and Goldberg, T. E. (2007). Quantifying incoherence in speech: an automated methodology and novel application to schizophrenia. *Schizophrenia research*, 93(1-3):304–316.

Ettinger, A. (2020). What bert is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

Ezzabady, M. K., Muller, P., and Braud, C. (2021). Multi-lingual discourse segmentation and connective identification. In *2nd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2021)*, pages 22–32. Association for Computational Linguistics.

Fan, Y., Li, P., Kong, F., and Zhu, Q. (2022). A distance-aware multi-task framework for conversational discourse parsing. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 912–921.

Feltracco, A., Jezek, E., Magnini, B., and Stede, M. (2016). Lico: A lexicon of italian connectives. In *CLiC-it/EVALITA*.

Feng, V. W. and Hirst, G. (2012). Text-level discourse parsing with rich linguistic features. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 60–68.

Feng, V. W. and Hirst, G. (2014a). A linear-time bottom-up discourse parser with constraints and post-editing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 511–521, Baltimore, Maryland. Association for Computational Linguistics.

Feng, V. W. and Hirst, G. (2014b). Patterns of local discourse coherence as a feature for authorship attribution. *Literary and Linguistic Computing*, 29(2):191–198.

Feng, W. V. (2015). *RST-style discourse parsing and its applications in discourse analysis*. University of Toronto (Canada).

Feng, X., Feng, X., and Qin, B. (2021a). A survey on dialogue summarization: Recent advances and new frontiers. *arXiv preprint arXiv:2107.03175*.

Feng, X., Feng, X., Qin, B., and Geng, X. (2021b). Dialogue discourse-aware graph model and data augmentation for meeting summarization. In Zhou, Z.-H., editor, *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 3808–3814. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Ferracane, E., Durrett, G., Li, J. J., and Erk, K. (2019). Evaluating discourse in structured text representations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 646–653, Florence, Italy. Association for Computational Linguistics.

Ferracane, E., Wang, S., and Mooney, R. (2017). Leveraging discourse information effectively for authorship attribution. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 584–593.

Firth, J. (1957). A synopsis of linguistic theory, 1930-1955. *Studies in linguistic analysis*, pages 10–32.

Fraser, B. (1990). An approach to discourse markers. *Journal of pragmatics*, 14(3):383–398.

Fraser, K. C., Meltzer, J. A., and Rudzicz, F. (2016). Linguistic features identify alzheimer's disease in narrative speech. *Journal of Alzheimer's Disease*, 49(2):407–422.

Fraser, W. I., King, K. M., Thomas, P., and Kendell, R. E. (1986). The diagnosis of schizophrenia by language analysis. *The British Journal of Psychiatry*, 148(3):275–278.

Frege, G. (1988). *Die Grundlagen der Arithmetik: eine logisch mathematische Untersuchung über den Begriff der Zahl*, volume 366. Felix Meiner Verlag.

Fu, X., Liu, W., Xu, Y., Yu, C., and Wang, T. (2016). Long short-term memory network over rhetorical structure theory for sentence-level sentiment analysis. In *Asian conference on machine learning*, pages 17–32. PMLR.

Gardent, C. (1997). Discourse tree adjoining grammars. *Claus report*, (89).

Gardner, M., Grus, J., Neumann, M., Tafjord, O., Dasigi, P., Liu, N. F., Peters, M., Schmitz, M., and Zettlemoyer, L. (2018). AllenNLP: A deep semantic natural language processing platform. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pages 1–6, Melbourne, Australia. Association for Computational Linguistics.

Gerani, S., Mehdad, Y., Carenini, G., Ng, R. T., and Nejat, B. (2014). Abstractive summarization of product reviews using discourse structure. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1602–1613, Doha, Qatar. Association for Computational Linguistics.

Gessler, L., Behzad, S., Liu, Y. J., Peng, S., Zhu, Y., and Zeldes, A. (2021). Discodisco at the disrpt2021 shared task: A system for discourse segmentation, classification, and connective detection. In *Proceedings of the 2nd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2021)*, pages 51–62.

Ghosal, D., Majumder, N., Poria, S., Chhaya, N., and Gelbukh, A. (2020). Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation. In *EMNLP-IJCNLP 2019-2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*.

Ginzburg, J., Breitholtz, E., Cooper, R., Hough, J., and Tian, Y. (2015). Understanding laughter. In *Proceedings of the 20th amsterdam colloquium. University of Amsterdam. http://semanticsarchive. net/Archive/mVkOTk2N/AC2015-proceedings. pdf*.

Ginzburg, J., Fernández, R. M., and David, S. (2014). Disfluencies as intra-utterance dialogue moves. *Semantics and Pragmatics*, 7(9):64.

Ginzburg, J., Mazzocconi, C., and Tian, Y. (2020). Laughter as language. *Glossa: a journal of general linguistics (2021-...)*, 5(1).

Gliwa, B., Mochol, I., Biesek, M., and Wawer, A. (2019). SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 70–79, Hong Kong, China. Association for Computational Linguistics.

Godfrey, J. J., Holliman, E. C., and McDaniel, J. (1992). Switchboard: Telephone speech corpus for research and development. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, volume 1, pages 517–520. IEEE Computer Society.

Goldberg, Y. (2019). Assessing bert's syntactic abilities. *arXiv preprint arXiv:1901.05287*.

Goodkind, A., Lee, M., Martin, G. E., Losh, M., and Bicknell, K. (2018). Detecting language impairments in autism: A computational analysis of semi-structured conversations with vector semantics. *Proceedings of the Society for Computation in Linguistics*, 1(1):12–22.

Gosztolya, G., Vincze, V., Tóth, L., Pákáski, M., Kálmán, J., and Hoffmann, I. (2019). Identifying mild cognitive impairment and mild alzheimer's disease based on spontaneous speech using asr and linguistic features. *Computer Speech & Language*, 53:181–197.

Gratch, J., Artstein, R., Lucas, G., Stratou, G., Scherer, S., Nazarian, A., Wood, R., Boberg, J., DeVault, D., Marsella, S., Traum, D., Rizzo, S., and Morency, L.-P. (2014). The distress analysis interview corpus of human and computer interviews. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3123–3128, Reykjavik, Iceland. European Language Resources Association (ELRA).

Greenstein, D., Weisinger, B., Malley, J. D., Clasen, L., and Gogtay, N. (2012). Using multivariate machine learning methods and structural MRI to classify childhood onset schizophrenia and healthy controls. *Frontiers in psychiatry*, 3.

Grice, H. P. (1975). Logic and conversation. In *Speech acts*, pages 41–58. Brill.

Grimes, J. E. and Grimes, R. E. (1975). *The thread of discourse*, volume 207. Walter de Gruyter.

Groenendijk, J. and Stokhof, M. (1990). Dynamic montague grammar.

Gross, J. L. and Yellen, J. (2003). *Handbook of graph theory*. CRC press.

Grosz, B. J. and Sidner, C. L. (1986). Attention, intentions, and the structure of discourse. *Computational linguistics*, 12(3):175–204.

Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., and Poon, H. (2021). Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.

Guntuku, S. C., Preotiuc-Pietro, D., Eichstaedt, J. C., and Ungar, L. H. (2019). What twitter profile and posted images reveal about depression and anxiety. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 13, pages 236–246.

Guntuku, S. C., Yaden, D. B., Kern, M. L., Ungar, L. H., and Eichstaedt, J. C. (2017). Detecting depression and mental illness on social media: an integrative review. *Current Opinion in Behavioral Sciences*, 18:43–49.

Haenelt, K. (1992). Towards a quality improvement in machine translation: Modelling discourse structure and including discourse development in the determination of translation equivalents. In *Proceedings of the Fourth Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages*.

Halliday, M. A. K. (1994). Spoken and written modes of meaning. *Media texts: Authors and readers*, 7:51–73.

Haque, A., Guo, M., Miner, A. S., and Fei-Fei, L. (2018). Measuring depression symptom severity from spoken language and 3d facial expressions. *arXiv preprint arXiv:1811.08592*.

Harris, Z. S. (1954). Distributional structure. *Word*, 10(2-3):146–162.

He, Q., Veldkamp, B. P., Glas, C. A., and de Vries, T. (2017). Automated assessment of patients' self-narratives for posttraumatic stress disorder screening using natural language processing and text mining. *Assessment*, 24(2):157–172.

He, Y., Zhang, Z., and Zhao, H. (2021). Multi-tasking dialogue comprehension with discourse parsing. In *Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation*, pages 551–561, Shanghai, China. Association for Computational Lingustics.

Heim, I. (1983). File change semantics and the familiarity theory of definiteness. *Semantics Critical Concepts in Linguistics*, pages 108–135.

Heim, I. R. (1982). *The semantics of definite and indefinite noun phrases*. University of Massachusetts Amherst.

Hernault, H., Prendinger, H., du Verle, D. A., and Ishizuka, M. (2010). Hilda: A discourse parser using support vector machine classification. *Dialogue & Discourse*, 1(3):1–33.

Hewitt, J. and Manning, C. D. (2019). A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.

Hirao, T., Yoshida, Y., Nishino, M., Yasuda, N., and Nagata, M. (2013). Single-document summarization as a tree knapsack problem. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1515–1520.

Hirst, G. and Marcu, D. (1998). The rhetorical parsing, summarization, and generation of natural language texts.

Hobbs, J. R. (1979). Coherence and coreference. *Cognitive science*, 3(1):67–90.

Hobbs, J. R. (1985). On the coherence and structure of discourse.

Hoffman, R. E. and Sledge, W. (1988). An analysis of grammatical deviance occurring in spontaneous schizophrenic speech. *Journal of neurolinguistics*, 3(1):89–101.

Hogenboom, A., Frasincar, F., De Jong, F., and Kaymak, U. (2015). Using rhetorical structure in sentiment analysis. *Communications of the ACM*, 58(7):69–77.

Hong, K., Kohler, C. G., March, M. E., Parker, A. A., and Nenkova, A. (2012). Lexical differences in autobiographical narratives from schizophrenic patients and healthy controls. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 37–47.

Honnibal, M., Montani, I., Van Landeghem, S., and Boyd, A. (2020). spacy: Industrial-strength natural language processing in python.

Howes, C., Purver, M., and McCabe, R. (2013). Using conversation topics for predicting therapy outcomes in schizophrenia. *Biomedical informatics insights*, 6:BII–S11661.

Howes, C., Purver, M., and McCabe, R. (2014). Linguistic indicators of severity and progress in online text-based therapy for depression. *ACL 2014*, page 7.

Howes, C., Purver, M., McCabe, R., Healey, P., and Lavelle, M. (2012). Predicting adherence to treatment for schizophrenia from dialogue transcripts. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 79–83.

Huber, P. and Carenini, G. (2019). Predicting discourse structure using distant supervision from sentiment. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2306–2316, Hong Kong, China. Association for Computational Linguistics.

Huber, P. and Carenini, G. (2020a). From sentiment annotations to sentiment prediction through discourse augmentation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 185–197, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Huber, P. and Carenini, G. (2020b). MEGA RST discourse treebanks with structure and nuclearity from scalable distant sentiment supervision. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7442–7457, Online. Association for Computational Linguistics.

Huber, P. and Carenini, G. (2020c). Unsupervised learning of discourse structures using a tree autoencoder. *arXiv preprint arXiv:2012.09446*.

Huber, P. and Carenini, G. (2022). Towards understanding large-scale discourse structures in pre-trained and fine-tuned language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2376–2394.

Huber, P., Xiao, W., and Carenini, G. (2021). W-RST: Towards a weighted RST-style discourse framework. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3908–3918, Online. Association for Computational Linguistics.

Iruskieta, M., Aranzabe, M. J., de Ilarraza, A. D., Gonzalez, I., Lersundi, M., and de Lacalle, O. L. (2013). The rst basque treebank: an online search interface to check rhetorical relations. In *4th workshop RST and discourse studies*, pages 40–49.

Iyer, R., Li, Y., Li, H., Lewis, M., Sundar, R., and Sycara, K. (2018). Transparency and explanation in deep reinforcement learning neural networks. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 144–150.

Jagannatha, A. and Yu, H. (2020). Calibrating structured output predictors for natural language processing. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2078–2092, Online. Association for Computational Linguistics.

Janin, A., Baron, D., Edwards, J., Ellis, D., Gelbart, D., Morgan, N., Peskin, B., Pfau, T., Shriberg, E., Stolcke, A., et al. (2003). The icsi meeting corpus. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).*, volume 1, pages I–I. IEEE.

Jansen, P., Surdeanu, M., and Clark, P. (2014). Discourse complements lexical semantics for non-factoid answer reranking. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 977–986.

Jarrold, W. L., Peintner, B., Yeh, E., Krasnow, R., Javitz, H. S., and Swan, G. E. (2010). Language analytics for assessing brain health: Cognitive impairment, depression and pre-symptomatic alzheimer's disease. In *International Conference on Brain Informatics*, pages 299–307. Springer.

Jawahar, G., Sagot, B., and Seddah, D. (2019). What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy. Association for Computational Linguistics.

Ji, Y. and Eisenstein, J. (2014). Representation learning for text-level discourse parsing. In *Proceedings of the 52nd annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 13–24.

Ji, Y. and Smith, N. A. (2017). Neural discourse structure for text categorization. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 996–1005.

Jiang, F., Fan, Y., Chu, X., Li, P., Zhu, Q., and Kong, F. (2021a). Hierarchical macro discourse parsing based on topic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13152–13160.

Jiang, F., Xu, S., Chu, X., Li, P., Zhu, Q., and Zhou, G. (2018). Mcdtb: a macro-level chinese discourse treebank. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3493–3504.

Jiang, K., Carenini, G., and Ng, R. (2016). Training data enrichment for infrequent discourse relations. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2603–2614, Osaka, Japan. The COLING 2016 Organizing Committee.

Jiang, X., Osl, M., Kim, J., and Ohno-Machado, L. (2012). Calibrating predictive model estimates to support personalized medicine. *Journal of the American Medical Informatics Association*, 19(2):263–274.

Jiang, Z., Araki, J., Ding, H., and Neubig, G. (2021b). How can we know when language models know? on the calibration of language models for question answering. *Transactions of the Association for Computational Linguistics*, 9:962–977.

Jiang, Z., Xu, F. F., Araki, J., and Neubig, G. (2020). How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438.

Johannsen, A., Hovy, D., and Søgaard, A. (2015). Cross-lingual syntactic variation over age and gender. In *Proceedings of the nineteenth conference on computational natural language learning*, pages 103–112.

Joshi, A., Karimi, S., Sparks, R., Paris, C., and MacIntyre, C. R. (2019). Does multi-task learning always help?: An evaluation on health informatics. In *Proceedings of the The 17th Annual Workshop of the Australasian Language Technology Association*, pages 151–158, Sydney, Australia. Australasian Language Technology Association.

Joty, S., Carenini, G., Ng, R., and Mehdad, Y. (2013). Combining intra-and multi-sentential rhetorical parsing for document-level discourse analysis. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 486–496.

Joty, S., Carenini, G., and Ng, R. T. (2015). Codra: A novel discriminative framework for rhetorical analysis. *Computational Linguistics*, 41(3):385–435.

Joty, S., Guzmán, F., Màrquez, L., and Nakov, P. (2017). Discourse structure in machine translation evaluation. *Computational Linguistics*, 43(4):683–722.

Jurafsky, D. (1997). Switchboard swbd-damsl shallow-discourse-function annotation coders manual. *Institute of Cognitive Science Technical Report*.

Jurafsky, D. and Martin, J. H. (2014). Speech and language processing. vol. 3. *US: Prentice Hall*.

Kamp, H. (1981). Evénements, représentations discursives et référence temporelle. *Langages*, (64):39–64.

Kamp, H. and Reyle, U. (2013). *From discourse to logic: Introduction to model theoretic semantics of natural language, formal logic and discourse representation theory*, volume 42. Springer Science & Business Media.

Karimi, H. and Tang, J. (2019). Learning hierarchical discourse-level structure for fake news detection. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3432–3442.

Kayi, E. S., Diab, M., Pauselli, L., Compton, M., and Coppersmith, G. (2017). Predictive linguistic features of schizophrenia. In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (* SEM 2017)*, pages 241–250.

Keskes, I., Zitoune, F. B., and Belguith, L. H. (2014). Learning explicit and implicit arabic discourse relations. *Journal of King Saud University-Computer and Information Sciences*, 26(4):398–416.

Kikuta, Y. (2019). Bert pretrained model trained on japanese wikipedia articles. `https://github.com/yoheikikuta/bert-japanese`.

Kim, N., Feng, S., Gunasekara, C., and Lastras, L. (2020). Implicit discourse relation classification: We need to talk about evaluation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5404–5414.

Kim, T., Choi, J., Edmiston, D., and Lee, S.-g. (2019). Are pre-trained language models aware of phrases? simple but strong baselines for grammar induction. In *International Conference on Learning Representations*.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kiritchenko, S., Zhu, X., and Mohammad, S. M. (2014). Sentiment analysis of short informal texts. *Journal of Artificial Intelligence Research*, 50:723–762.

Kleim, B., Horn, A. B., Kraehenmann, R., Mehl, M. R., and Ehlers, A. (2018). Early linguistic markers of trauma-specific processing predict post-trauma adjustment. *Frontiers in Psychiatry*, 9:645.

Knott, A. and Dale, R. (1994). Using linguistic phenomena to motivate a set of coherence relations. *Discourse Processes*, 18(1):35–62.

Ko, Y., Park, J., and Seo, J. (2004). Improving text categorization using the importance of sentences. *Information processing & management*, 40(1):65–79.

Koay, J. J., Roustai, A., Dai, X., Burns, D., Kerrigan, A., and Liu, F. (2020). How domain terminology affects meeting summarization performance. In *Proceedings of the 28th International Conference on Computational Linguistics (COLING)*.

Kobayashi, N., Hirao, T., Kamigaito, H., Okumura, M., and Nagata, M. (2021). Improving neural rst parsing model with silver agreement subtrees. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1600–1612.

Kobayashi, N., Hirao, T., Nakamura, K., Kamigaito, H., Okumura, M., and Nagata, M. (2019). Split or merge: Which is better for unsupervised RST parsing? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5797–5802, Hong Kong, China. Association for Computational Linguistics.

Koto, F., Lau, J. H., and Baldwin, T. (2021). Discourse probing of pretrained language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3849–3864.

Kraus, M. and Feuerriegel, S. (2019). Sentiment analysis based on rhetorical structure theory: Learning deep neural networks from discourse trees. *Expert Systems with Applications*, 118:65–79.

Kroenke, K. and Spitzer, R. L. (2002). The phq-9: a new depression diagnostic and severity measure.

Kummerfeld, J. K., Gouravajhala, S. R., Peper, J. J., Athreya, V., Gunasekara, C., Ganhotra, J., Patel, S. S., Polymenakos, L. C., and Lasecki, W. (2019). A large-scale corpus for conversation disentanglement. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3846–3856.

Kuperberg, G. R. (2010a). Language in schizophrenia part 1: an introduction. *Language and linguistics compass*, 4(8):576–589.

Kuperberg, G. R. (2010b). Language in schizophrenia part 1: an introduction. *Language and linguistics compass*, 4(8):576–589.

Kuratov, Y. and Arkhipov, M. (2019). Adaptation of deep bidirectional multilingual transformers for russian language. *arXiv preprint arXiv:1905.07213*.

Laali, M. and Kosseim, L. (2017). Automatic disambiguation of french discourse connectives. *arXiv preprint arXiv:1704.05162*.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., and Soricut, R. (2019). Albert: A lite bert for self-supervised learning of language representations. In *International Conference on Learning Representations*.

Lascarides, A. and Asher, N. (1993). Temporal interpretation, discourse relations and commonsense entailment. *Linguistics and philosophy*, 16(5):437–493.

Le, H., Vial, L., Frej, J., Segonne, V., Coavoux, M., Lecouteux, B., Allauzen, A., Crabbé, B., Besacier, L., and Schwab, D. (2020). Flaubert: Unsupervised language model pre-training for french. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2479–2490.

Le, Q. and Mikolov, T. (2014). Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196. PMLR.

Le Glaz, A., Haralambous, Y., Kim-Dufor, D.-H., Lenca, P., Billot, R., Ryan, T. C., Marsh, J., Devylder, J., Walter, M., Berrouiguet, S., et al. (2021). Machine learning and natural language processing in mental health: systematic review. *Journal of Medical Internet Research*, 23(5):e15708.

Lee, D.-H. et al. (2013). Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, page 896.

Levelt, W. J. (1983). Monitoring and self-repair in speech. *Cognition*, 14(1):41–104.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. (2020). BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting*

of the Association for Computational Linguistics, pages 7871–7880, Online. Association for Computational Linguistics.

Li, C., Amblard, M., Braud, C., Demily, C., Franck, N., and Musiol, M. (2021a). Investigating non lexical markers of the language of schizophrenia in spontaneous conversations. In *Proceedings of the 2nd Workshop on Computational Approaches to Discourse*, pages 20–28, Punta Cana, Dominican Republic and Online. Association for Computational Linguistics.

Li, C., Braud, C., and Amblard, M. (2022). Multi-task learning for depression detection in dialogs. In *SIGDIAL 2022-The 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*.

Li, C., Huber, P., Xiao, W., Amblard, M., Braud, C., and Carenini, G. (2023). Discourse structure extraction from pre-trained and fine-tuned language models in dialogues. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 2517–2534.

Li, J., Li, R., and Hovy, E. (2014a). Recursive deep models for discourse parsing. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2061–2069.

Li, J., Liu, M., Kan, M.-Y., Zheng, Z., Wang, Z., Lei, W., Liu, T., and Qin, B. (2020). Molweni: A challenge multiparty dialogues-based machine reading comprehension dataset with discourse structure. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2642–2652, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Li, J., Liu, M., Zheng, Z., Zhang, H., Qin, B., Kan, M.-Y., and Liu, T. (2021b). Dadgraph: A discourse-aware dialogue graph neural network for multiparty dialogue machine reading comprehension. *arXiv preprint arXiv:2104.12377*.

Li, J. and Nenkova, A. (2015). Fast and accurate prediction of sentence specificity. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29.

Li, J. J., Carpuat, M., and Nenkova, A. (2014b). Assessing the discourse factors that influence the quality of machine translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 283–288.

Li, J. J., Thadani, K., and Stent, A. (2016). The role of discourse units in near-extractive summarization. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 137–147.

Li, M., Zhang, L., Ji, H., and Radke, R. J. (2019). Keep meeting summaries on topic: Abstractive multi-modal meeting summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2190–2196, Florence, Italy. Association for Computational Linguistics.

Li, S., Wang, L., Cao, Z., and Li, W. (2014c). Text-level discourse dependency parsing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 25–35, Baltimore, Maryland. Association for Computational Linguistics.

Li, Y., Feng, W., Sun, J., Kong, F., and Zhou, G. (2014d). Building chinese discourse corpus with connective-driven dependency tree structure. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2105–2114.

Li, Y., Su, H., Shen, X., Li, W., Cao, Z., and Niu, S. (2017). Dailydialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995.

Lin, Z., Ng, H. T., and Kan, M.-Y. (2014). A pdtb-styled end-to-end discourse parser. *Natural Language Engineering*, 20(2):151–184.

Linardatos, P., Papastefanopoulos, V., and Kotsiantis, S. (2020). Explainable ai: A review of machine learning interpretability methods. *Entropy*, 23(1):18.

Liu, H., Wang, W., Wang, Y., Liu, H., Liu, Z., and Tang, J. (2020). Mitigating gender bias for neural dialogue generation with adversarial learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 893–903.

Liu, P., Qiu, X., and Huang, X.-J. (2017). Adversarial multi-task learning for text classification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1–10.

Liu, Y. and Lapata, M. (2018). Learning structured text representations. *Transactions of the Association for Computational Linguistics*, 6:63–75.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019a). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Liu, Y., Titov, I., and Lapata, M. (2019b). Single document summarization as tree induction. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1745–1755.

Liu, Z. and Chen, N. (2021). Improving multi-party dialogue discourse parsing via domain integration. In *Proceedings of the 2nd Workshop on Computational Approaches to Discourse*, pages 122–127, Punta Cana, Dominican Republic and Online. Association for Computational Linguistics.

Liu, Z., Krishnaswamy, P., and Chen, N. F. (2022). Domain-specific language pre-training for dialogue comprehension on clinical inquiry-answering conversations. In *Multimodal AI in healthcare: A paradigm shift in health intelligence*, pages 29–40. Springer.

Louis, A., Joshi, A., and Nenkova, A. (2010). Discourse indicators for content selection in summarization. In *Proceedings of the SIGDIAL 2010 Conference*, pages 147–156, Tokyo, Japan. Association for Computational Linguistics.

Lowe, R., Pow, N., Serban, I., and Pineau, J. (2015). The Ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 285–294, Prague, Czech Republic. Association for Computational Linguistics.

Lu, Y., Kumar, A., Zhai, S., Cheng, Y., Javidi, T., and Feris, R. (2017). Fully-adaptive feature sharing in multi-task networks with applications in person attribute classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5334–5343.

Luz, S., Garcia, S. D. L. F., and Albert, P. (2018). A method for analysis of patient speech in dialogue for dementia detection. In *Resources and ProcessIng of linguistic, para-linguistic and extra-linguistic Data from people with various forms of cognitive impairment*, pages 35–42. European Language Resources Association (ELRA).

Ma, X., Zhang, Z., and Zhao, H. (2021). Enhanced speaker-aware multi-party multi-turn dialogue comprehension. *arXiv e-prints*, pages arXiv–2109.

Maher, B. A., Manschreck, T. C., Linnet, J., and Candela, S. (2005). Quantitative assessment of the frequency of normal associations in the utterances of schizophrenia patients and healthy controls. *Schizophrenia Research*, 78(2-3):219–224.

Majumder, N., Poria, S., Hazarika, D., Mihalcea, R., Gelbukh, A., and Cambria, E. (2019). Dialoguernn: An attentive rnn for emotion detection in conversations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 6818–6825.

Mallol-Ragolta, A., Zhao, Z., Stappen, L., Cummins, N., and Schuller, B. W. (2019). A hierarchical attention network-based approach for depression detection from transcribed clinical interviews. *Proc. Interspeech 2019*, pages 221–225.

Mann, W. C. (1984). Discourse structures for text generation. Technical report, UNIVERSITY OF SOUTHERN CALIFORNIA MARINA DEL REY INFORMATION SCIENCES INST.

Mann, W. C. and Thompson, S. A. (1987). *Rhetorical structure theory: Description and construction of text structures*. Springer.

Mann, W. C. and Thompson, S. A. (1988). Rhetorical structure theory: Toward a functional theory of text organization. *Text-interdisciplinary Journal for the Study of Discourse*, 8(3):243–281.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J. R., Bethard, S., and McClosky, D. (2014). The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pages 55–60.

Manschreck, T. C., Maher, B. A., Rosenthal, J. E., and Berner, J. (1991). Reduced primacy and related features in schizophrenia. *Schizophrenia Research*, 5(1):35–41.

Marcinkiewicz, M. A. (1994). Building a large annotated corpus of english: The penn treebank. *Using Large Corpora*, 273.

Marcu, D. (1997). The rhetorical parsing of unrestricted natural language texts. In *35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*, pages 96–103.

Marcu, D. (1998). To build text summaries of high quality, nuclearity is not sufficient. In *Working Notes of the AAAI-98 Spring Symposium on Intelligent Text Summarization*, pages 1–8.

Marcu, D. (2000). *The theory and practice of discourse parsing and summarization*. MIT press.

Marcu, D., Amorrortu, E., and Romera, M. (1999). Experiments in constructing a corpus of discourse trees. In *Towards Standards and Tools for Discourse Tagging*.

Mareček, D. and Rosa, R. (2019). From balustrades to pierre vinken: Looking for syntax in transformer self-attentions. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 263–275, Florence, Italy. Association for Computational Linguistics.

Martin, L., Muller, B., Suárez, P. J. O., Dupont, Y., Romary, L., de la Clergerie, É. V., Seddah, D., and Sagot, B. (2020). Camembert: a tasty french language model. In *ACL 2020-58th Annual Meeting of the Association for Computational Linguistics*.

Martinez-Martin, N., Dunn, L. B., and Roberts, L. W. (2018). Is it ethical to use prognostic estimates from machine learning to treat psychosis? *AMA journal of ethics*, 20(9):E804.

McDonald, R., Pereira, F., Ribarov, K., and Hajic, J. (2005). Non-projective dependency parsing using spanning tree algorithms. In *Proceedings of human language technology conference and conference on empirical methods in natural language processing*, pages 523–530.

McKeown, G., Valstar, M., Cowie, R., Pantic, M., and Schroder, M. (2011). The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE transactions on affective computing*, 3(1):5–17.

McKeown, K. R. (1985). Discourse strategies for generating natural-language text. *Artificial intelligence*, 27(1):1–41.

Meng, Y., Zhang, Y., Huang, J., Xiong, C., Ji, H., Zhang, C., and Han, J. (2020). Text classification using label names only: A language model self-training approach. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9006–9017.

Michael, J., Botha, J. A., and Tenney, I. (2020). Asking without telling: Exploring latent ontologies in contextual representations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6792–6812.

Mihăilă, C. and Ananiadou, S. (2014). Semi-supervised learning of causal relations in biomedical scientific discourse. *Biomedical engineering online*, 13(2):1–24.

Mihăilă, C., Ohta, T., Pyysalo, S., and Ananiadou, S. (2013). Biocause: Annotating and analysing causality in the biomedical domain. *BMC bioinformatics*, 14(1):1–18.

Mikolov, T., Yih, W.-t., and Zweig, G. (2013). Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies*, pages 746–751.

Miller, A., Feng, W., Batra, D., Bordes, A., Fisch, A., Lu, J., Parikh, D., and Weston, J. (2017). Parlai: A dialog research software platform. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 79–84.

Miltsakaki, E., Prasad, R., Joshi, A. K., and Webber, B. L. (2004). The penn discourse treebank. In *LREC*.

Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., and Gao, J. (2021). Deep learning–based text classification: a comprehensive review. *ACM Computing Surveys (CSUR)*, 54(3):1–40.

Mírovskỳ, J., Synková, P., Rysová, M., and Poláková, L. (2017). Czedlex-a lexicon of czech discourse connectives. *The Prague Bulletin of Mathematical Linguistics*, 109(1):61.

Mishra, A., Dey, K., and Bhattacharyya, P. (2017). Learning cognitive features from gaze data for sentiment and sarcasm classification using convolutional neural network. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 377–387, Vancouver, Canada. Association for Computational Linguistics.

Misra, A., Anand, P., Fox Tree, J. E., and Walker, M. (2015). Using summarization to discover argument facets in online idealogical dialog. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 430–440, Denver, Colorado. Association for Computational Linguistics.

Misra, I., Shrivastava, A., Gupta, A., and Hebert, M. (2016). Cross-stitch networks for multi-task learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3994–4003.

Mitchell, M., Hollingshead, K., and Coppersmith, G. (2015). Quantifying the language of schizophrenia in social media. In *Proceedings of the 2nd workshop on Computational linguistics and clinical psychology: From linguistic signal to clinical reality*, pages 11–20.

Mitkov, R. (1993). How could rhetorical relations be used in machine translation? In *Intentionality and structure in discourse relations*.

Mohammad, S. M. and Turney, P. D. (2013). Crowdsourcing a word–emotion association lexicon. *Computational intelligence*, 29(3):436–465.

Montague, R. (1970). English as a formal language.

Montague, R. (1973). The proper treatment of quantification in ordinary english. In *Approaches to natural language: Proceedings of the 1970 Stanford workshop on grammar and semantics*, pages 221–242. Springer.

Morey, M., Muller, P., and Asher, N. (2017). How much progress have we made on RST discourse parsing? a replication study of recent results on the RST-DT. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1319–1324, Copenhagen, Denmark. Association for Computational Linguistics.

Morice, R. D. and Ingram, J. C. (1982). Language analysis in schizophrenia: Diagnostic implications. *Australian and New Zealand Journal of Psychiatry*, 16(2):11–21.

Muller, P., Afantenos, S., Denis, P., and Asher, N. (2012). Constrained decoding for text-level discourse parsing. In *Proceedings of COLING 2012*, pages 1883–1900, Mumbai, India. The COLING 2012 Organizing Committee.

Muller, P., Braud, C., and Morey, M. (2019). ToNy: Contextual embeddings for accurate multilingual discourse segmentation of full documents. In *Proceedings of the Workshop on Discourse Relation Parsing and Treebanking 2019*, pages 115–124, Minneapolis, MN. Association for Computational Linguistics.

Munezero, M., Montero, C. S., Sutinen, E., and Pajunen, J. (2014). Are they different? affect, feeling, emotion, sentiment, and opinion detection in text. *IEEE transactions on affective computing*, 5(2):101–111.

Musiol, M. and Trognon, A. (2000). *Eléments de psychopathologie cognitive: le discours schizophrène*. A. Colin.

Muskens, R. (1996). Combining montague semantics and discourse representation. *Linguistics and philosophy*, pages 143–186.

Nallapati, R., Zhou, B., dos Santos, C., Gulçehre, Ç., and Xiang, B. (2016). Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, Berlin, Germany. Association for Computational Linguistics.

Narasimhan, K. and Barzilay, R. (2015). Machine comprehension with discourse relations. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1253–1262.

Nasreen, S., Purver, M., and Hough, J. (2019). A corpus study on questions, responses and misunderstanding signals in conversations with alzheimer's patients. In *Proceedings of the 23rd Workshop on the Semantics and Pragmatics of Dialogue-Full Papers. SEMDIAL, London, United Kingdom (Sep 2019), http://semdial. org/anthology/Z19-Nasreen semdial*, volume 13.

Nejat, B., Carenini, G., and Ng, R. (2017). Exploring joint neural model for sentence level discourse parsing and sentiment analysis. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 289–298.

Nguyen, T., Phung, D., Dao, B., Venkatesh, S., and Berk, M. (2014). Affective and content analysis of online depression communities. *IEEE Transactions on Affective Computing*, 5(3):217–226.

Nguyen, T.-T., Nguyen, X.-P., Joty, S., and Li, X. (2021). Rst parsing from scratch. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1613–1625.

Nie, A., Bennett, E., and Goodman, N. (2019). Dissent: Learning sentence representations from explicit discourse relations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4497–4510.

Nishida, N. and Matsumoto, Y. (2022). Out-of-domain discourse dependency parsing via bootstrapping: An empirical analysis on its effectiveness and limitation. *Transactions of the Association for Computational Linguistics*, 10:127–144.

Nishida, N. and Nakayama, H. (2020). Unsupervised discourse constituency parsing using viterbi em. *Transactions of the Association for Computational Linguistics*, 8:215–230.

Nivre, J. (2004). Incrementality in deterministic dependency parsing. In *Proceedings of the workshop on incremental parsing: Bringing engineering and cognition together*, pages 50–57.

Nivre, J., De Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajic, J., Manning, C. D., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., et al. (2016). Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666.

Ono, K., Sumita, K., Research, S. M., Center, D., Komukai-Toshiba-cho, T. C., et al. (1994). Abstract generation based on rhetorical structure extraction. *arXiv preprint cmp-lg/9411023*.

OpenAI (2023). Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Orimaye, S. O., Wong, J. S.-M., and Golden, K. J. (2014). Learning predictive linguistic features for alzheimer's disease and related dementias using verbal utterances. In *Proceedings of the workshop on computational linguistics and clinical psychology: From linguistic signal to clinical reality*, pages 78–87.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Pandia, L., Cong, Y., and Ettinger, A. (2021). Pragmatic competence of pre-trained language models through the lens of discourse connectives. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 367–379.

Pang, B. and Lee, L. (2004). A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, pages 271–es.

Passonneau, R. J. and Litman, D. (1997). Discourse segmentation by human and automated means. *Computational Linguistics*, 23(1):103–139.

Pauls, A. and Klein, D. (2012). Large-scale syntactic language modeling with treelets. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 959–968.

Pedersen, T. (2015). Screening twitter users for depression and ptsd with lexical decision lists. In *Proceedings of the 2nd workshop on computational linguistics and clinical psychology: from linguistic signal to clinical reality*, pages 46–53.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

Pelletier, F. J. (1994). The principle of semantic compositionality. *Topoi*, 13(1):11–24.

Peng, S., Liu, Y. J., and Zeldes, A. (2022). GCDT: A Chinese RST Treebank for Multigenre and Multilingual Discourse Parsing. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing*, pages 382–391, Online only. Association for Computational Linguistics.

Pennebaker, J., Francis, M., and Booth, R. (2001). *Linguistic inquiry and word count (LIWC)*.

Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Perret, J., Afantenos, S., Asher, N., and Morey, M. (2016). Integer linear programming for discourse parsing. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 99–109, San Diego, California. Association for Computational Linguistics.

Péry-Woodley, M.-P., Afantenos, S., Ho-Dac, L.-M., and Asher, N. (2011). La ressource annodis, un corpus enrichi d'annotations discursives. *revue TAL*, 52(3):71–101.

Pestian, J. P., Sorter, M., Connolly, B., Bretonnel Cohen, K., McCullumsmith, C., Gee, J. T., Morency, L.-P., Scherer, S., Rohlfs, L., and Group, S. R. (2017). A machine learning approach to identifying the thought markers of suicidal subjects: a prospective multicenter trial. *Suicide and Life-Threatening Behavior*, 47(1):112–121.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018a). Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Peters, M. E., Neumann, M., Zettlemoyer, L., and Yih, W.-t. (2018b). Dissecting contextual word embeddings: Architecture and representation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1499–1509.

Pisarevskaya, D., Ananyeva, M., Kobozeva, M., Nasedkin, A., Nikiforova, S., Pavlova, I., and Shelepov, A. (2017). Towards building a discourse-annotated corpus of russian. In *Computational Linguistics and Intellectual Technologies: 23rd International Conference on Computational Linguistics and Intellectual Technologies "Dialogue"*, pages 194–204.

Pitler, E. and Nenkova, A. (2009). Using syntax to disambiguate explicit discourse connectives in text. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 13–16.

Poláková, L., Mírovskỳ, J., Nedoluzhko, A., Jínová, P., Zikánová, Š., and Hajicová, E. (2013). Introducing the prague discourse treebank 1.0. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 91–99.

Polanyi, L. (1985). A theory of discourse structure and discourse coherence in papers from the general session at the twenty-first regional meeting. In *CLS. Papers from the General Session at the... Regional Meeting*, volume 21, pages 306–322.

Polanyi, L. (1988). A formal model of the structure of discourse. *Journal of pragmatics*, 12(5-6):601–638.

Polanyi, L., Culy, C., Van Den Berg, M., Thione, G. L., and Ahn, D. (2004). A rule based approach to discourse parsing. In *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue at HLT-NAACL 2004*, pages 108–117.

Polanyi, L. and Scha, R. (1984). A syntactic approach to discourse semantics. In *10th International Conference on Computational Linguistics and 22nd Annual Meeting of the Association for Computational Linguistics*, pages 413–419.

Polignano, M., Basile, P., de Gemmis, M., Semeraro, G., and Basile, V. (2019). AlBERTo: Italian BERT Language Understanding Model for NLP Challenging Tasks Based on Tweets. In *Proceedings of the Sixth Italian Conference on Computational Linguistics (CLiC-it 2019)*, volume 2481. CEUR.

Pope, C. and Davis, B. H. (2011). Finding a balance: The carolinas conversation collection.

Poria, S., Hazarika, D., Majumder, N., Naik, G., Cambria, E., and Mihalcea, R. (2018). Meld: A multimodal multi-party dataset for emotion recognition in conversations. *arXiv preprint arXiv:1810.02508*.

Poria, S., Hazarika, D., Majumder, N., Naik, G., Cambria, E., and Mihalcea, R. (2019). Meld: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536.

Potvin, S., Aubin, G., and Stip, E. (2017). L'insight neurocognitif dans la schizophrénie. *L'Encéphale*, 43(1):15–20.

Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A., and Webber, B. (2008a). The penn discourse treebank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*.

Prasad, R., Husain, S., Sharma, D. M., and Joshi, A. (2008b). Towards an annotated corpus of discourse relations in hindi. In *Proceedings of the 6th Workshop on Asian Language Resources*.

Prasad, R., McRoy, S., Frid, N., Joshi, A., and Yu, H. (2011). The biomedical discourse relation bank. *BMC bioinformatics*, 12(1):1–18.

Qin, K., Wang, L., and Kim, J. (2017). Joint modeling of content and discourse relations in dialogues. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 974–984, Vancouver, Canada. Association for Computational Linguistics.

Qin, L., Li, Z., Che, W., Ni, M., and Liu, T. (2020). Co-gat: A co-interactive graph attention network for joint dialog act recognition and sentiment classification. *arXiv preprint arXiv:2012.13260*.

Quirk, C., Menezes, A., and Cherry, C. (2005). Dependency treelet translation: Syntactically informed phrasal smt. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 271–279.

Qureshi, S. A., Dias, G., Hasanuzzaman, M., and Saha, S. (2020). Improving depression level estimation by concurrently learning emotion intensity. *IEEE Computational Intelligence Magazine*, 15(3):47–59.

Qureshi, S. A., Saha, S., Hasanuzzaman, M., and Dias, G. (2019). Multitask representation learning for multimodal estimation of depression level. *IEEE Intelligent Systems*, 34(5):45–52.

Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al. (2018). Improving language understanding by generative pre-training.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P. J., et al. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67.

Raganato, A. and Tiedemann, J. (2018). An analysis of encoder representations in transformer-based machine translation. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. The Association for Computational Linguistics.

Rajpurkar, P., Jia, R., and Liang, P. (2018). Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. (2016). Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392.

Rapp, R. (2003). Word sense discovery based on sense descriptor dissimilarity. In *Proceedings of Machine Translation Summit IX: Papers*.

Ratana, R., Sharifzadeh, H., Krishnan, J., and Pang, P. (2019). A comprehensive review of computational methods for automatic prediction of schizophrenia with insight into indigenous populations. *Frontiers in psychiatry*, 10:659.

Ratner, A., Bach, S. H., Ehrenberg, H., Fries, J., Wu, S., and Ré, C. (2017). Snorkel: Rapid training data creation with weak supervision. In *Proceedings of the VLDB Endowment. International Conference on Very Large Data Bases*, volume 11, page 269. NIH Public Access.

Ratner, A. J., De Sa, C. M., Wu, S., Selsam, D., and Ré, C. (2016). Data programming: Creating large training sets, quickly. *Advances in neural information processing systems*, 29.

Rebuschi, M., Amblard, M., and Musiol, M. (2014). Using SDRT to analyze pathological conversations. Logicality, rationality and pragmatic deviances. In Rebuschi, M., Batt, M., Heinzmann, G., Lihoreau, F., Musiol, M., and Trognon, A., editors, *Interdisciplinary Works in Logic, Epistemology, Psychology and Linguistics: Dialogue, Rationality, and Formalism*, volume 3 of *Logic, Argumentation & Reasoning*, pages 343 – 368. Springer.

Reese, B., Hunter, J., Asher, N., Denis, P., and Baldridge, J. (2007). Reference manual for the analysis and annotation of rhetorical structure (version 1.0). *URL: http://timeml.org/jamesp/annotation_manual. pdf (last access: 7.9. 2015)*.

Rehbein, I., Scholman, M., and Demberg, V. (2016). Annotating discourse relations in spoken language: A comparison of the pdtb and ccr frameworks. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1039–1046.

Rehbein, I. and van Genabith, J. (2007). Evaluating evaluation measures. In *Proceedings of the 16th Nordic Conference of Computational Linguistics (NODALIDA 2007)*, pages 372–379, Tartu, Estonia. University of Tartu, Estonia.

Reimers, N. and Gurevych, I. (2020). Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525.

Riccardi, G., Stepanov, E. A., and Chowdhury, S. A. (2016). Discourse connective detection in spoken conversations. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6095–6099.

Ringeval, F., Schuller, B., Valstar, M., Gratch, J., Cowie, R., Scherer, S., Mozgai, S., Cummins, N., Schmitt, M., and Pantic, M. (2017). Avec 2017: Real-life depression, and affect recognition workshop and challenge. In *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*, pages 3–9.

Ríssola, E. A., Losada, D. E., and Crestani, F. (2021). A survey of computational methods for online mental state assessment on social media. *ACM Transactions on Computing for Healthcare*, 2(2):1–31.

Rochester, S. (2013). *Crazy talk: A study of the discourse of schizophrenic speakers*. Springer Science & Business Media.

Rogers, A., Kovaleva, O., and Rumshisky, A. (2020). A primer in bertology: What we know about how bert works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Rosenberg, C., Hebert, M., and Schneiderman, H. (2005). Semi-supervised self-training of object detection models.

Rosenthal, S. and McKeown, K. (2013). Columbia nlp: Sentiment detection of subjective phrases in social media. In *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 478–482.

Rouchota, V. (1996). Discourse connectives: what do they link. *UCL Working Papers in Linguistics*, 8(199–214).

Roze, C., Danlos, L., and Muller, P. (2012). Lexconn: a french lexicon of discourse connectives. *Discours. Revue de linguistique, psycholinguistique et informatique. A journal of linguistics, psycholinguistics and computational linguistics*, (10).

Rude, S., Gortner, E.-M., and Pennebaker, J. (2004). Language use of depressed and depression-vulnerable college students. *Cognition & Emotion*, 18(8):1121–1133.

Ruder, S. (2017). An overview of multi-task learning in deep neural networks. *arXiv e-prints*, pages arXiv–1706.

Ruder, S., Bingel, J., Augenstein, I., and Søgaard, A. (2017). Sluice networks: Learning what to share between loosely related tasks. *arXiv preprint arXiv:1705.08142*, 2.

Ruder, S., Bingel, J., Augenstein, I., and Søgaard, A. (2019). Latent multi-task architecture learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 4822–4829.

Rutherford, A., Demberg, V., and Xue, N. (2017). A systematic study of neural discourse models for implicit discourse relation. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 281–291.

Sabeti, M., Katebi, S., Boostani, R., and Price, G. (2011). A new approach for eeg signal classification of schizophrenic and control participants. *Expert Systems with Applications*, 38(3):2063–2071.

Sacks, H. (1992). Lectures on conversation: Volume i. *Malden, Massachusetts: Blackwell*.

Sacks, H., Schegloff, E. A., and Jefferson, G. (1978). A simplest systematics for the organization of turn taking for conversation. In *Studies in the organization of conversational interaction*, pages 7–55. Elsevier.

Sagae, K. (2009). Analysis of discourse structure with syntactic dependencies and data-driven shift-reduce parsing. Technical report, UNIVERSITY OF SOUTHERN CALIFORNIA LOS ANGELES.

Saito, K., Ushiku, Y., and Harada, T. (2017). Asymmetric tri-training for unsupervised domain adaptation. In *International Conference on Machine Learning*, pages 2988–2997. PMLR.

Sakishita, M., Ogawa, C., Tsuchiya, K. J., Iwabuchi, T., Kishimoto, T., and Kano, Y. (2019). Autism spectrum disorder's severity prediction model using utterance features for automatic diagnosis support. In *International Workshop on Health Intelligence*, pages 83–95. Springer.

Salton, G., Wong, A., and Yang, C.-S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620.

Salzinger, K. (1979). Ecolinguistics: A radical behavior theory approach to language behavior.

Salzinger, K. and Hammer, M. (1963). Some formal characteristics of schizophrenic speech as a measure of social deviance. *Annals of the New York Academy of Sciences*.

Salzinger, K., Portnoy, S., and Feldman, R. S. (1964). Verbal behavior of schizophrenic and normal subjects. *Annals of the New York Academy of sciences*, 105(15):845–860.

Salzinger, K., Portnoy, S., Pisoni, D. B., and Feldman, R. S. (1970). The immediacy hypothesis and response-produced stimuli in schizophrenic speech. *Journal of Abnormal Psychology*, 76(2):258.

Sanders, T. J., Spooren, W. P., and Noordman, L. G. (1992). Toward a taxonomy of coherence relations. *Discourse processes*, 15(1):1–35.

Sanders, T. J., Spooren, W. P., and Noordman, L. G. (1993). Coherence relations in a cognitive theory of discourse representation. *Cognitive Linguistics*, 4(2).

Schabes, Y. (1990). *Mathematical and computational aspects of lexicalized grammars*. University of Pennsylvania.

Schegloff, E. A. (2007). *Sequence organization in interaction: A primer in conversation analysis I*, volume 1. Cambridge university press.

Scheible, R., Thomczyk, F., Tippmann, P., Jaravine, V., and Boeker, M. (2020). Gottbert: a pure german language model. *arXiv preprint arXiv:2012.02110*.

Schilder, F. (1997). Tree discourse grammar, or how to get attached to a discourse. In *Proceedings of the Second International Workshop on Computational Semantics (IWCS-II)*, pages 261–273. Citeseer.

Schlichtkrull, M., Kipf, T. N., Bloem, P., Berg, R. v. d., Titov, I., and Welling, M. (2018). Modeling relational data with graph convolutional networks. In *European semantic web conference*, pages 593–607. Springer.

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., Shah, A., Kosinski, M., Stillwell, D., Seligman, M. E., et al. (2013). Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one*, 8(9):e73791.

See, A., Liu, P. J., and Manning, C. D. (2017). Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083.

Sekulić, I. and Strube, M. (2019). Adapting deep learning methods for mental health prediction on social media. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 322–327.

Settles, B. (2009). Active learning literature survey.

Shelmanov, A., Pisarevskaya, D., Chistova, E., Toldova, S., Kobozeva, M., and Smirnov, I. (2019). Towards the data-driven system for rhetorical parsing of russian texts. In *Proceedings of the Workshop on Discourse Relation Parsing and Treebanking 2019*, pages 82–87.

Shi, P. and Lin, J. (2019). Simple bert models for relation extraction and semantic role labeling. *arXiv preprint arXiv:1904.05255*.

Shi, W. and Demberg, V. (2019). Next sentence prediction helps implicit discourse relation classification within and across domains. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 5790–5796.

Shi, W., Yung, F., and Demberg, V. (2019). Acquiring annotated data with cross-lingual explicitation for implicit discourse relation classification. *NAACL HLT 2019*, page 12.

Shi, Z. and Huang, M. (2019). A deep sequential model for discourse parsing on multi-party dialogues. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7007–7014.

Shriberg, E. E. (1994). *Preliminaries to a theory of speech disfluencies*. PhD thesis, Citeseer.

Sileo, D., Van De Cruys, T., Pradel, C., and Muller, P. (2019). Mining discourse markers for unsupervised sentence representation learning. In *Proceedings of NAACL-HLT*, pages 3477–3486.

Sileo, D., van de Cruys, T., Pradel, C., and Muller, P. (2020). Discsense: Automated semantic analysis of discourse markers. In *12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 991–999.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A. Y., and Potts, C. (2013a). Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A. Y., and Potts, C. (2013b). Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642.

Søgaard, A. and Goldberg, Y. (2016). Deep multi-task learning with low level tasks supervised at lower layers. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 231–235, Berlin, Germany. Association for Computational Linguistics.

Song, H., You, J., Chung, J.-W., and Park, J. C. (2018). Feature attention network: Interpretable depression detection from social media. In *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation*, Hong Kong. Association for Computational Linguistics.

Soria, C. and Ferrari, G. (1998). Lexical marking of discourse relations-some experimental findings. In *Discourse Relations and Discourse Markers*.

Spitkovsky, V. I., Alshawi, H., Jurafsky, D., and Manning, C. D. (2010). Viterbi training improves unsupervised dependency parsing. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, pages 9–17.

Stede, M. (2004). The potsdam commentary corpus. In *Proceedings of the Workshop on Discourse Annotation*, pages 96–102.

Stede, M. (2011). Discourse processing. *Synthesis Lectures on Human Language Technologies*, 4(3):1–165.

Stede, M. and Neumann, A. (2014). Potsdam commentary corpus 2.0: Annotation for discourse research. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 925–929.

Stede, M., Scheffler, T., and Mendes, A. (2019). Connective-lex: A web-based multilingual lexical resource for connectives. *Discours. Revue de linguistique, psycholinguistique et informatique. A journal of linguistics, psycholinguistics and computational linguistics*, (24).

Stede, M. and Umbach, C. (1998). Dimlex: A lexicon of discourse markers for text generation and understanding. In *COLING 1998 Volume 2: The 17th International Conference on Computational Linguistics*.

Steedman, M., Hwa, R., Clark, S., Osborne, M., Sarkar, A., Hockenmaier, J., Ruhlen, P., Baker, S., and Crim, J. (2003). Example selection for bootstrapping statistical parsers. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 236–243.

Steinlin, J., Colinet, M., and Danlos, L. (2015). Fdtb1: Repérage des connecteurs de discours en corpus. In *Actes de la 22e conférence sur le Traitement Automatique des Langues Naturelles. Articles courts*, pages 34–40.

Straka, M. and Straková, J. (2017). Tokenizing, pos tagging, lemmatizing and parsing ud 2.0 with udpipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99, Vancouver, Canada. Association for Computational Linguistics.

Strous, R. D., Koppel, M., Fine, J., Nachliel, S., Shaked, G., and Zivotofsky, A. Z. (2009). Automated characterization and identification of schizophrenia in writing. *The Journal of nervous and mental disease*, 197(8):585–588.

Subba, R. and Di Eugenio, B. (2009). An effective discourse parser that uses rich linguistic information. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 566–574.

Surdeanu, M., Hicks, T., and Valenzuela-Escárcega, M. A. (2015). Two practical rhetorical structure theory parsers. In *Proceedings of the 2015 conference of the North American chapter of the association for computational linguistics: Demonstrations*, pages 1–5.

Taboada, M. and Mann, W. C. (2006). Applications of rhetorical structure theory. *Discourse studies*, 8(4):567–588.

Tai, K. S., Socher, R., and Manning, C. D. (2015). Improved semantic representations from tree-structured long short-term memory networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1556–1566.

Tang, D., Qin, B., and Liu, T. (2015). Document modeling with gated recurrent neural network for sentiment classification. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 1422–1432.

Tarvainen, A. and Valpola, H. (2017). Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30.

Tenney, I., Das, D., and Pavlick, E. (2019). Bert rediscovers the classical nlp pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601.

Thomas, M., Pang, B., and Lee, L. (2006). Get out the vote: Determining support or opposition from congressional floor-debate transcripts. In *COLING• ACL 2006*, page 327.

Tofiloski, M., Brooke, J., and Taboada, M. (2009). A syntactic and lexical-based discourse segmenter. In *Proceedings of the ACL-IJCNLP 2009 conference short papers*, pages 77–80.

Tonelli, S., Riccardi, G., Prasad, R., and Joshi, A. (2010). Annotation of discourse relations for conversational spoken dialogs. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*.

Tu, M., Zhou, Y., and Zong, C. (2013). A novel translation framework based on rhetorical structure theory. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 370–374.

Turney, P. D. (2006). Similarity of semantic relations. *Computational Linguistics*, 32(3):379–416.

Turney, P. D. and Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, 37:141–188.

Valstar, M., Gratch, J., Schuller, B., Ringeval, F., Lalanne, D., Torres Torres, M., Scherer, S., Stratou, G., Cowie, R., and Pantic, M. (2016). Avec 2016: Depression, mood, and emotion recognition workshop and challenge. In *Proceedings of the 6th international workshop on audio/visual emotion challenge*, pages 3–10.

Varachkina, H. and Pannach, F. (2021). A unified approach to discourse relation classification in nine languages. In *Proceedings of the 2nd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2021)*, pages 46–50.

Vargas, F., Jonas, D., Rabinovich, Z., Benevenuto, F., and Pardo, T. (2022). Rhetorical structure approach for online deception detection: A survey. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5906–5915.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Veličković, P., Casanova, A., Lio, P., Cucurull, G., Romero, A., and Bengio, Y. (2018). Graph attention networks.

Verberne, S., Boves, L., Oostdijk, N., and Coppen, P. (2007a). Discourse-based answering of why-questions.

Verberne, S., Boves, L., Oostdijk, N., and Coppen, P.-A. (2007b). Evaluating discourse-based answer extraction for why-question answering. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 735–736.

Virtanen, A., Kanerva, J., Ilo, R., Luoma, J., Luotolahti, J., Salakoski, T., Ginter, F., and Pyysalo, S. (2019). Multilingual is not enough: Bert for finnish. *arXiv preprint arXiv:1912.07076*.

Wang, A., Song, L., Jiang, H., Lai, S., Yao, J., Zhang, M., and Su, J. (2021a). A structure self-aware model for discourse parsing on multi-party dialogues. In *Proceedings of the Thirtieth International Conference on International Joint Conferences on Artificial Intelligence*.

Wang, J., Zhang, L., and Kong, F. (2021b). Multi-level cohesion information modeling for better written and dialogue discourse parsing. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 40–52. Springer.

Wang, L. L., Lo, K., Chandrasekhar, Y., Reas, R., Yang, J., Burdick, D., Eide, D., Funk, K., Katsis, Y., Kinney, R. M., et al. (2020). Cord-19: The covid-19 open research dataset. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*.

Wang, S.-M., Lee, C.-H. S., Lo, Y.-C., Huang, T.-H., and Ku, L.-W. (2016). Sensing emotions in text messages: An application and deployment study of emotionpush. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations*, pages 141–145.

Wang, X., Bruno, J., Molloy, H., Evanini, K., and Zechner, K. (2017a). Discourse annotation of non-native spontaneous spoken responses using the rhetorical structure theory framework. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 263–268.

Wang, Y., Li, S., and Wang, H. (2017b). A two-stage parsing method for text-level discourse analysis. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 184–188, Vancouver, Canada. Association for Computational Linguistics.

Webber, B. (2004). D-ltag: extending lexicalized tag to discourse. *Cognitive Science*, 28(5):751–779.

Webber, B., Popescu-Belis, A., Markert, K., and Tiedemann, J. (2013). Proceedings of the workshop on discourse in machine translation. In *Proceedings of the Workshop on Discourse in Machine Translation*.

Webber, B., Prasad, R., Lee, A., and Joshi, A. (2019). The penn discourse treebank 3.0 annotation manual. *Philadelphia, University of Pennsylvania*, 35:108.

Webber, B., Stone, M., Joshi, A., and Knott, A. (2003). Anaphora and discourse structure. *Computational linguistics*, 29(4):545–587.

Webber, B. L. and Joshi, A. K. (1998). Anchoring a lexicalized tree-adjoining grammar for discourse. In *Discourse Relations and Discourse Markers Workshop*, pages 86–92. Association for Computational Linguistics.

Williamson, J. R., Godoy, E., Cha, M., Schwarzentruber, A., Khorrami, P., Gwon, Y., Kung, H.-T., Dagli, C., and Quatieri, T. F. (2016). Detecting depression using vocal, facial and semantic communication cues. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, pages 11–18.

Wilson, D. and Sperber, D. (2012). Linguistic form and relevance. *Wilson & Sperber (Eds.), Meaning and Relevance*, pages 149–168.

Wilson, T., Hoffmann, P., Somasundaran, S., Kessler, J., Wiebe, J., Choi, Y., Cardie, C., Riloff, E., and Patwardhan, S. (2005a). Opinionfinder: A system for subjectivity analysis. In *Proceedings of HLT/EMNLP 2005 Interactive Demonstrations*, pages 34–35.

Wilson, T., Wiebe, J., and Hoffmann, P. (2005b). Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of human language technology conference and conference on empirical methods in natural language processing*, pages 347–354.

Wolf, F. and Gibson, E. (2005). Representing discourse coherence: A corpus-based study. *Computational linguistics*, 31(2):249–287.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Wu, Z., Chen, Y., Kao, B., and Liu, Q. (2020). Perturbed masking: Parameter-free probing for analyzing and interpreting bert. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4166–4176.

Xezonaki, D., Paraskevopoulos, G., Potamianos, A., and Narayanan, S. (2020). Affective conditioning on hierarchical attention networks applied to depression detection from transcribed clinical interviews. In *INTERSPEECH*, pages 4556–4560.

Xiang, W. and Wang, B. (2023). A survey of implicit discourse relation recognition. *ACM Computing Surveys*, 55(12):1–34.

Xiao, W., Huber, P., and Carenini, G. (2021). Predicting discourse trees from transformer-based neural summarizers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4139–4152, Online. Association for Computational Linguistics.

Xing, W. and Ghorbani, A. (2004). Weighted pagerank algorithm. In *Proceedings. Second Annual Conference on Communication Networks and Services Research, 2004.*, pages 305–314. IEEE.

Xu, H., Van Durme, B., and Murray, K. (2021). Bert, mbert, or bibert? a study on contextualized embeddings for neural machine translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6663–6675.

Xu, J., Gan, Z., Cheng, Y., and Liu, J. (2020). Discourse-aware neural extractive text summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5021–5031.

Xu, S., Yang, Z., Chakraborty, D., Tahir, Y., Maszczyk, T., Chua, Y. H. V., Dauwels, J., Thalmann, D., Thalmann, N. M., Tan, B.-L., et al. (2019). Automated lexical analysis of interviews with individuals with schizophrenia. In *Proceedings of the 9th International Workshop on Spoken Dialogue System Technology*, pages 185–197. Springer.

Xue, N., Ng, H. T., Pradhan, S., Prasad, R., Bryant, C., and Rutherford, A. (2015). The conll-2015 shared task on shallow discourse parsing. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning-Shared Task*, pages 1–16.

Xue, N., Su, Q., and Jeong, S. (2016). Annotating the discourse and dialogue structure of sms message conversations. In *Proceedings of the 10th Linguistic Annotation Workshop held in conjunction with ACL 2016 (LAW-X 2016)*, pages 180–187.

Xue, N., Xia, F., Chiou, F.-D., and Palmer, M. (2005). The penn chinese treebank: Phrase structure annotation of a large corpus. *Natural language engineering*, 11(2):207–238.

Yang, A. and Li, S. (2018). Scidtb: Discourse dependency treebank for scientific abstracts. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 444–449.

Yang, A., Wang, Q., Liu, J., Liu, K., Lyu, Y., Wu, H., She, Q., and Li, S. (2019a). Enhancing pre-trained language representations with rich knowledge for machine reading comprehension. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2346–2357.

Yang, J., Xu, K., Xu, J., Li, S., Gao, S., Guo, J., Xue, N., and Wen, J.-R. (2021). A joint model for dropped pronoun recovery and conversational discourse parsing in chinese conversational speech. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1752–1763.

Yang, Y., Liu, Y., and Xue, N. (2015). Recovering dropped pronouns from chinese text messages. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 309–313.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., and Le, Q. V. (2019b). Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

Yano, T., Smith, N. A., and Wilkerson, J. (2012). Textual predictors of bill survival in congressional committees. In *proceedings of the 2012 conference of the north American chapter of the Association for Computational Linguistics: human language technologies*, pages 793–802.

Yarowsky, D. (1995). Unsupervised word sense disambiguation rivaling supervised methods. In *33rd annual meeting of the association for computational linguistics*, pages 189–196.

Yates, A., Cohan, A., and Goharian, N. (2017). Depression and self-harm risk assessment in online forums. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2968–2978, Copenhagen, Denmark. Association for Computational Linguistics.

Yoshida, E. and Lickley, R. J. (2010). Disfluency patterns in dialogue processing. In *DiSS-LPSS Joint Workshop 2010*.

Yoshida, Y., Suzuki, J., Hirao, T., and Nagata, M. (2014). Dependency-based discourse parser for single-document summarization. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1834–1839.

Yu, N., Fu, G., and Zhang, M. (2022). Speaker-aware discourse parsing on multi-party dialogues. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 5372–5382.

Yu, Y., Zuo, S., Jiang, H., Ren, W., Zhao, T., and Zhang, C. (2021). Fine-tuning pre-trained language model with weak supervision: A contrastive-regularized self-training approach. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1063–1077.

Yue, M. (2008). Rhetorical structure annotation of chinese news commentaries. *Journal of Chinese Information Processing*, 22(4):19–23.

Zadeh, A. B., Liang, P. P., Poria, S., Cambria, E., and Morency, L.-P. (2018). Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2236–2246.

Zaheer, M., Guruganesh, G., Dubey, K. A., Ainslie, J., Alberti, C., Ontanon, S., Pham, P., Ravula, A., Wang, Q., Yang, L., et al. (2020). Big bird: Transformers for longer sequences. *Advances in Neural Information Processing Systems*, 33:17283–17297.

Zeldes, A. (2017). The gum corpus: Creating multilayer resources in the classroom. *Language Resources and Evaluation*, 51(3):581–612.

Zeldes, A., Das, D., Maziero, E. G., Antonio, J., and Iruskieta, M. (2019). The disrpt 2019 shared task on elementary discourse unit segmentation and connective detection. In *Proceedings of the Workshop on Discourse Relation Parsing and Treebanking 2019*, pages 97–104.

Zeldes, A., Liu, Y. J., Iruskieta, M., Muller, P., Braud, C., and Badene, S. (2021). The disrpt 2021 shared task on elementary discourse unit segmentation, connective detection, and relation classification. In *Proceedings of the 2nd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2021)*, pages 1–12.

Zeyrek, D., Mendes, A., Grishina, Y., Kurfali, M., Gibbon, S., and Ogrodniczuk, M. (2019). Ted multilingual discourse bank (ted-mdb): a parallel corpus annotated in the pdtb style. *Language Resources and Evaluation*, pages 1–38.

Zeyrek, D., Turan, U. D., Bozsahin, C., Cakıcı, R., Sevdik-Calli, A. B., Demirsahin, I., Aktaş, B., Yalçinkaya, I., and Balaban, H. Ö. (2009). Annotating subordinators in the turkish discourse bank. In *Proceedings of the Third Linguistic Annotation Workshop (LAW III)*, pages 44–47.

Zeyrek, D. and Webber, B. (2008). A discourse resource for turkish: Annotating discourse connectives in the metu corpus. In *Proceedings of the 6th workshop on Asian language resources*.

Zhang, B. H., Lemoine, B., and Mitchell, M. (2018a). Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340.

Zhang, S., Dinan, E., Urbanek, J., Szlam, A., Kiela, D., and Weston, J. (2018b). Personalizing dialogue agents: I have a dog, do you have pets too? *arXiv preprint arXiv:1801.07243*.

Zhang, X., Zhao, J., and LeCun, Y. (2015). Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28.

Zhang, Y., Sun, S., Galley, M., Chen, Y.-C., Brockett, C., Gao, X., Gao, J., Liu, J., and Dolan, B. (2020). DIALOGPT : Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278, Online. Association for Computational Linguistics.

Zhong, M., Liu, Y., Xu, Y., Zhu, C., and Zeng, M. (2022). Dialoglm: Pre-trained model for long dialogue understanding and summarization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11765–11773.

Zhou, Z.-H. (2018). A brief introduction to weakly supervised learning. *National science review*, 5(1):44–53.

Zhou, Z.-H. and Li, M. (2005). Tri-training: Exploiting unlabeled data using three classifiers. *IEEE Transactions on knowledge and Data Engineering*, 17(11):1529–1541.

Zhu, C., Xu, R., Zeng, M., and Huang, X. (2020a). A hierarchical network for abstractive meeting summarization with cross-domain pretraining. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 194–203.

Zhu, Z., Pan, C., Abdalla, M., and Rudzicz, F. (2020b). Examining the rhetorical capacities of neural language models. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 16–32, Online. Association for Computational Linguistics.