Probing Discourse Coherence through the Multi-Token Predictions of Language Models

Meinan Liu

Master of Science

Cognitive Science

School of Informatics

University of Edinburgh

2024

Abstract

This projects explores discourse coherence through multi-token predictions in language models. Previous research has mainly focused on single-token connectives, neglecting the challenges of multi-token completion in Masked Language Models [8]. We applied BERT, with an additional Extended Output Prediction Matrix decoder, specifically designed to predict multi-token connectives. Additionally, we developed linguistic resources including a vocabulary of inter-sentential multi-token discourse connectives and their senses from the PDTB-3 Appendix and Connective-Lex. We also created an extended preposed and non-canonical dataset respectively, for further research on discourse relation recognition task. By using the two datasets for model inference, we extended the claim that a preposing structure can help MLMs predict a single token connective in a discourse to multi-token scenario. The preposing structure improves the model's general accuracy and accuracy across genre, and confidence in correct predictions, especially for complex discourse relations including *Arg2-as-detail*, *Arg2-as-instance*, and *Reason*.

Research Ethics Approval

This project was planned in accordance with the Informatics Research Ethics policy. It did not involve any aspects that required approval from the Informatics Research Ethics committee.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Meinan Liu*)

Acknowledgements

Thanks for the generous help and support from my supervisors. I still remember the first meeting with Bonnie before students were assigned their projects. She spoke slowly and clearly, explaining the details of the project in a way that dispelled all of my doubts. I know she is the kind of researcher who truly loves what she does and has embarked on this journey with passion for decades. She is also incredibly responsive. We met every week and exchanged dozens of emails discussing the problems I encountered, which I deeply appreciate. Then I met with Xixian on our weekly meetings. She is just as welcoming and lovely as Bonnie had introduced her to us. She can think of solutions to problems quickly and also has a keen sense of potential issues that may not be immediately obvious. She encouraged us when we felt discouraged by our progress and helped us build a habit of recording what we have down every week. Bonnie and Xixian are both role models for me, and I am so glad that I had the chance to work with them for four months and learn from them.

I also want to thank all the friends I met at the University of Edinburgh. I was a coding rookie at the beginning of the first term and struggled with coursework and lectures. I met them, and we studied together at the Appleton Tower, at the main library, and had lunches and dinners at Chinese restaurants. These memories will last a lifetime, though I wish I had taken more photos. We shared both joy and tears. We stayed up late studying and talked about ghost stories in AT, and we played hard and celebrated the new year together. My boyfriend is one of them. Most of our time together was spent working on coursework, writing our dissertations, and talking about machine learning. Thank you for bringing me so much happiness this year, for watching Chiikawa with me, and for introducing all of your Jellycats to me. We all love the fatty aubergine, or is it just me? You are now really into usagi. :(

Very very lastly, I want to thank my parents. Thank you for your calls, your care, and your financial support throughout my studies and life. There were times when I was busy and didn't want to make frequent calls, but I soon realized that those calls saved me from some of my low moments. They gave me a moment to take a deep breath, relax, and just talk about myself and you. I really enjoyed those calls, even though I never said it out loud.

Table of Contents

Chapter 1

Introduction[1]

A discourse is normally a sequence of clauses, or sentences. Local discourse coherence is considered to be the relation between two adjacent clauses or sentences, which helps the information flow and structure [24, 15]. This coherence can be explicitly signaled by an explicit discourse connective, such as "for example", "but", "and", etc. Ex.(1) illustrates a case where "but" functions as an explicit connective. Conversely, if a discourse relation is not signaled by a discourse connective, then this relation is referred to as implicit. That is, people need to infer the discourse relation based on their understanding of the clauses or sentences being connected and their context. An implicit relation, if it exists, can be made explicit by inserting a discourse connective as shown in Ex.(2). The insertion does not alter the meaning of sentences and clauses, but rather reveals the sentential relation more straightforwardly, allowing any suitable connective to be inserted. For instance, in Ex.(2), "but", "so", or "because" could be used as the inserted implicit connective each conveying a different relation between the clauses. The Penn Discourse Treebank 3.0 (PDTB-3) [16], a corpus of articles from Wall Street Journal with human-annotated discourse relations, and DiscoGeM 1.0 [21], a crowdsourced corpus of implicit discourse relations, are two exemplars of the practice of annotating discourse relations (see Section 2.2).

(1) John left **but**$_{\textbf{explicit connective}}$ Bob stayed.

(2) John left, **[but/so/because]**$_{\textbf{implicit connective}}$ Bob stayed.

Discourse connectives include coordinating conjunctions (such as "and", "but", "or"), subordinating conjunctions (such as "when", "because", "if") and adverbials

[1] Part of the content in the Introduction is from the author's Informatics Project Proposal.

(such as “in addition”, “for example”, “meanwhile”). While subordinating conjunctions typically occur within a sentence (i.e. intra-sententially), connectives and adverbials between sentences (i.e. inter-sententially) are usually coordinating conjunctions. As illustrated in Ex.(3) and Ex.(4), subordinating conjunction “because” connects adjacent clauses within a sentences, and the adverbial “meanwhile” acts across sentences.

(3) Intra-sentential: The federal government suspended sales of U.S. savings bonds **because**$_{\textbf{coordinating conjunction}}$ Congress hasn’t lifted the ceiling on government debt. [wsj_0008, PDTB-3]

(4) Inter-sentential: In another reflection that the growth of the economy is leveling off, the government said that orders for manufactured goods and spending on construction failed to rise in September. **Meanwhile**$_{\textbf{subordinating conjunction}}$, the National Association of Purchasing Management said its latest survey indicated that the manufacturing economy contracted in October for the sixth consecutive month. [wsj_0036, PDTB-3]

Although the discourse connectives are effective in marking the relations between the current sentence and the prior context, they are not the only way; alternatively, marked information structure can also link a previous sentence (i.e. Arg1) with the current one (i.e. Arg2), for example, by a preposed constituent like noun phrase (NP) or preposition phrase (PP) [26] (see more details in Section 2.4). Ex.(5) and Ex.(6) demonstrate cases where at the start of Arg2 — a PP in Ex (5) and an NP in Ex (6) — links with the previous sentence (Arg1).

(5) Preposed PP: He suddenly shivered: he experienced a momentary sensation that he didn’t understand because no one on Earth had ever experienced it before.$_{Arg1}$ **In moments of great stress**$_{\textbf{PP}}$, every life form that exists gives out a tiny sublimal signal.$_{Arg2}$ [Hitchhiker’s Guide to the Galaxy, DiscoGeM 1.0]

(6) Preposed NP: Dudley’s mouth fell open in horror, but Harry’s heart gave a leap.$_{Arg1}$ **Every year on Dudley’s birthday**$_{\textbf{NP}}$, his parents took him and a friend out for the day, to adventure parks, hamburger restaurants, or the movies.$_{Arg2}$ [Harry Potter, DiscoGeM 1.0]

Previous work has shown that preposed structure can help language models predict implicit discourse relations [3]. Specifically, they used a Masked Language Models (MLM) like BERT [2] to insert connectives as a way to predict implicit discourse

relations between sentences. They found that BERT's predictions were more aligned with human annotations when the context included the preposed structure compared to when it did not. However, their work was limited to predicting single-token connective due to the constraint of BERT, which is trained to predict one token at a time.

1.1 Motivation and objectives

Given the limited empirical evidence on whether preposed structures can effectively signal implicit discourse relations—only one study [3],, to our knowledge, has addressed this—and the limitations of previous work focused on single-token prediction, our project aims to explore multi-token connective prediction and test the impact of preposed structures.

Multi-token connectives are expected to often less ambiguous with respect to the discourse relations they can convey. For instance, in Ex.(7), the word "but", with eight possible senses including *Comparison.Concession.Arg2-as-denier*, *Comparison.Contrast* and so on, remains general and open to interpretation. In contrast, the phrase "despite this" in Ex.(8), which carries only one sense in PDTB-3, clearly specifies a "Comparison.Concession.Arg2-as-denier" relation, where the Arg2 denies or contrast with the expectation set by the Arg1, making it less ambiguous compared to the single-token connective "but."

(7) The weather forecast predicted rain. **But** the event continued as planned.

(8) The weather forecast predicted rain. **Despite this**, the event continued as planned.

In our experiments, we will examine whether preposing can help the model predict a multi-token connective.

Furthermore, in previous work [3], while the model generated the top 5 single-token connectives as output, the corresponding probabilities after the Softmax layer were sometimes close to each other, indicating uncertainty in the predictions, and at other times sparse, with the top choice dominating the probability distribution. There were also instances where BERT did not generate any connectives.

Thus, it will be interesting to analyze our multi-token connective predictions, including the mask-fillers, their corresponding senses, and probabilities, and to test some of previous findings [3] in the context of our results.

The dissertation mainly aims to predict multi-token discourse connectives via a fine-tuned MLM, BERT. We qualitatively validate if a multi-token connective prediction is

indeed more specific than a single token connective prediction in Dong et al. [3]'s output. We then compare the probability distribution of our model's predictions on two different test datasets, with and without preposed structures respectively, and examine how the findings from single-token connective predictions apply to multi-token predictions, ultimately drawing a conclusion.

1.2 Results and Contribution

Our experiments show that the model indeed give less ambiguous multi-token connectives as the mask fillers, as we had hoped. Also, a preposed structure in Arg2 within a discourse helped the model to understand the information flow, therefore gives more accurate multi-token connective predictions. While previous researchers validated this claim on single token prediction, we are the first validating the benefits of the preposed structure on multi-token connectives.

We also contributed linguistic resources for further research on discourse relation recognition task. We created a vocabulary of inter-sentential multi-token discourse connectives and their sense-mapping dictionary, and developed two extended datasets—the preposed and non-preposed datasets, each including 1598 samples—derived from both PDTB-3 [16] and DiscoGeM 1.0 [21]. While the PDTB-3 mostly contains news articles, DiscoGeM 1.0 includes political speeches, literature and wikipedia texts.

1.3 Structure of dissertation

The rest of this dissertation contains four chapters. Chapter 2 introduces the background on implicit discourse relation recognition task, two frequently used datasets in this area, and the syntactic preposed structure. Chapter 3 presents the approach to prepare datasets, predict multi-token connectives via a modified BERT model, and the evaluation metrics for further comparison. Chapter 4 provides the results after the implementation of the model. We examines if the multi-token connective is less ambiguous, and compares the model's performance on the preposed and non-preposed test sets. Finally, Chapter 5 discusses the future work that could be done after this project in terms of the datasets, and the model, and draws a conclusion on whether the preposed structure is effective in signaling implicit relations or not.

Chapter 2

Background

Chapter 2 provides the background for understanding the research on multi-token connective prediction. Specifically, Section 2.1 introduces the concept of the implicit discourse relation recognition task, discussing its importance in NLP and the transition from traditional machine learning techniques to neural network approaches. Section 2.2 introduces two widely used corpora in the Implicit Discourse Relation Recognition (IDRR) task. Section 2.3 focuses on Masked Language Models (MLMs), particularly BERT, and their tailored pre-training tasks for sense recognition. Section 2.4 introduces the syntactic preposing strategy, explaining how syntactic structures can enhance the prediction of implicit connectives. Finally, Section 2.5 addresses the challenge of multi-token prediction in MLMs and presents innovative solutions.

2.1 Implicit discourse relation recognition (IDRR)

In both spoken and written communication, it is often the case that no connective is explicitly provided, yet listeners or readers can easily infer the relationship between two segments of text. The task of implicit discourse relation recognition (IDRR) is to detect and identify such "covert" relations when no connective is present in the discourse. This capability is beneficial for various downstream natural language processing (NLP) tasks such as machine translation, question answering, sentiment analysis, etc.

Early research employed traditional machine learning strategies like Naive Bayes for classification, which required hand-crafted features [27]. Linguistic features, including lexical information like one-hot word representations, syntactic information like part-of-speech (POS) [14], and so on, played a crucial role in these models. However, the reliance on selected features and sparse one-hot vectors due to a large vocabulary size

[13] limited the performance of traditional machine learning methods.

In recent years, researchers have increasingly turned to neural networks or deep learning (DL) methods for the IDRR task. In neural networks, inputs are no longer one-hot vectors but word embeddings, which are numerical representations capturing the linguistic information of a token and its context. Neural models have evolved from Convolutional Neural Networks (CNN) [10, 18], Recurrent Neural Networks (RNN) [9], and Long Short Term Memory (LSTM) networks [20] to attention mechanisms [1] and current Large Language Models (LLMs). Using these DL methods, some studies aim for direct sense classification [17, 23], while others propose first predicting and inserting an implicit connective between two texts and then mapping the connective to its sense (i.e., discourse relation) [28, 3], which is a cloze-like task.

2.2 The PDTB and DiscoGeM corpora

This section introduces two corpora used in our experiments: PDTB [16] and DiscoGeM [21]. Other widely used discourse relation corpora include the Georgetown University Multilayer Corpus (GUM) [31], a multilayer corpus for discourse model research released in 2017. While PDTB and DiscoGeM follow the PDTB-style sense annotations, GUM adheres to the Rhetorical Structure Theory (RST) framework [12]. A notable challenge in using GUM for our experiments is that it does not distinguish between inter-sentential and intra-sentential relations, which complicates preprocessing.

2.2.1 PDTB Corpus

The Penn Discourse Treebank (PDTB) corpus [16] is the largest and most widely used resource for discourse relation annotation in the NLP community. The texts in PDTB are sourced from the Wall Street Journal, with discourse relations annotated by professionals. The latest version, PDTB 3.0, was released in 2019 and updated in 2020, containing a total of 53,676 annotated discourse relations. PDTB-3 includes both inter-sentential and intra-sentential, as well as explicit and implicit discourse relations. Before discussing the annotation scheme of PDTB-3, it is important to introduce some key terminologies used in the corpus:

- **Argument**: A text segment containing at least a predicate that expresses an action, event, or state.

- **Connective**: A lexicon or phrase that links together two arguments, signaling the discourse relation between them, such as “if”, “in addition”, etc.
- **Sense**: The type of discourse relation, such as temporal, comparison and so on.
- **Explicit/Implicit connective**: If there is a connective in an argument, then it is an explicit connective. Otherwise, it is implicit because the connective doesn’t exist but the discourse relation is there.

Most argument-pairs (Arg1 and Arg2) are annotated with an explicit or inserted implicit connective, along with a sense and other metadata. Note that annotators can insert two connectives and their corresponding senses, if they feel that both senses are conveyed implicitly. To maintain consistency in identifying different types of relations, PDTB-3 employs a hierarchical sense classification with three levels (see in Appendix A). Level 1 includes four main classes: Temporal, Contingency, Comparison, and Expansion. These are further subdivided into types (level 2) and subtypes (level 3). As shown in Figure 2.1(a), if the connective “Instead” is explicit within the arguments, the corresponding sense “Expansion.Substitution.Arg2-as-subst” is labeled between the adjacent discourse units. For implicit relations, as illustrated in Figure 2.1(b), the connective “By contrast” is inserted, and its sense “Comparison.Contrast” is annotated. This insertion should be both semantically and syntactically appropriate and natural within the context. It is worth noting that an argument-pair may sometimes exhibit more than one discourse relation, leading to two connectives and senses being annotated. However, such instances are rare (less than 10 samples in our test dataset) and are not the focus of our experiments.

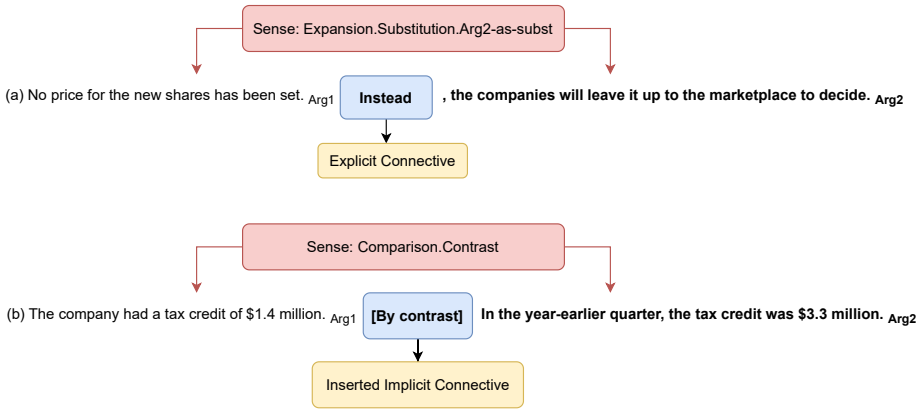

Figure 2.1: Examples of corpus annotation for connectives and their senses. Explicit connectives are present in the raw text with their senses annotated directly, while implicit connectives are added during annotation, with their senses annotated separately.

2.2.2 DiscoGeM corpus

The DiscoGeM corpus [21, 30] is a crowdsourced corpus of genre-mixed inter-sentential implicit discourse relations, annotated in the PDTB-style. The DiscoGeM 1.0 corpus [21], which is exclusively in English, includes 6,505 implicit discourse relations. Some of the texts are not original English texts but are translated into English from other languages. As shown in Table 2.1, the DiscoGeM 1.0 contains texts from three distinct genres: political speeches (Europarl), literature, and encyclopedic (Wikipedia). The updated DiscoGeM 2.0 is a parallel corpus that supports multiple languages [30], but for our experiments, we concentrated exclusively on original English texts, which led us to select 1741 samples from DiscoGeM 1.0 as one of our test data sources.

The annotation process in DiscoGeM is similar to that in PDTB-3, involving the insertion of implicit connectives, with sense annotations that match the PDTB-3 style. The developers tested four aggregation methods of combining the choices from their ten crowd workers to best represent the discourse relations in the data. They recommended two methods as shown in Table 2.2: (1) the CrowdTruth soft label, which kept crowd workers' annotations with probabilities of different sense types via Dumitrache et al.'s CrowdTruth 2.0 method [4], and (2) the majority-single label, where the sense with the most votes was chosen as the gold label. If there was a tie, a single sense was picked randomly and recorded. We followed their suggestion and used the majority-single label as the gold label in our experiments since it is preferred when only one label is needed.

genre	**Arg1**	**Arg2**
wikipedia	Analytical chemistry studies and uses instruments and methods used to separate, identify, and quantify matter.	In practice, separation, identification or quantification may constitute the entire analysis or be combined with another method.
europarl	You will be aware from the press and television that there have been a number of bomb explosions and killings in Sri Lanka.	One of the people assassinated very recently in Sri Lanka was Mr Kumar Ponnambalam, who had visited the European Parliament just a few months ago.
novel	After the horses came Muriel, the white goat, and Benjamin, the donkey.	Benjamin was the oldest animal on the farm, and the worst tempered.

Table 2.1: DiscoGeM 1.0 data with 3 genres.

majoritylabel_sampled	**crowdtruth_softlabel**
arg2-as-detail	arg2-as-detail:0.46157664794362013 conjunction:0.45508605763343674 precedence:0.04635423525438645 arg1-as-detail:0.03698305916855685 arg1-as-cond:0.0 arg1-as-denier:0.0 arg1-as-goal:0.0 ... synchronous:0.0

Table 2.2: Example Table with majoritylabel_sampled, crowdtruth_softlabel, arg1, and arg2.

2.3 Masked language models for IDRR

As discussed in Section 2.1, some researchers treat sense recognition as a cloze task. Consequently, recent studies have favored pre-trained Masked Language Models (MLMs), particularly BERT [2] and its variants, such as RoBERTa [11], spanBERT [7], due to their strong performance. BERT, in its off-the-shelf form, learns contextual word embeddings during pre-training, which can be fine-tuned for various downstream tasks.

Moreover, BERT's pretraining task, next sentence prediction, is particularly well-suited for discourse relation recognition, as it enhances BERT's sensitivity to the relationships between sentences.

2.4 Syntactic preposing

In some IDRR tasks, researchers predict a connective and then map it to its sense. Among them, Dong et al. [3] discovered that the syntactic structure of preposing in the second text (Arg2) improved the accuracy of MLMs, specifically BERT, in predicting implicit connectives. Preposing can take various forms, including the preposing of noun phrases (NP), prepositional phrases (PP), verb phrases (VP), adjective phrases (AP), and adverbial phrases (AdvP). Their study focused exclusively on preposed PP and NP examples, as illustrated in Ex.(5) of PP and Ex.(6) of NP in Chapter 1, respectively.

In their experiments, two datasets were created by Dong herself from PDTB-3 : a preposed set, where the NP/PP is sentence-initial in Arg2 as shown in Ex.(6) and Ex.(5), and a canonical set, where the NP/PP is re-positioned to the end of the first main clause in Arg2 to create a typical sentence structure as shown in Ex(10) when "from an administrative point of view" is right-moved to the end. The Arg2 can extend over several clauses or even several sentences. A [MASK] token was inserted between the two arguments, and the model was tasked with predicting a single-token connective. The results indicated that BERT performed better on the preposed set than the canonical set, making this study the first to empirically validate that preposing can help signal discourse relations. Considering that many connectives consist of a single token and given BERT's limitation to predict only a single token as the mask filler, their research primarily focused on single-token discourse connectives, as opposed to multi-token connectives such as "for example" or "on the other hand."

(9) Preposed argument: From an administrative point of view, the formalisation is a good thing.

(10) Canonical argument: The formalisation is a good thing from an administrative point of view.

2.5 Multi-token completion

Multi-token completion is a significant challenge in the use of MLMs for sentence completion [8]. Typically, these models are constrained to single-token mask fillers or must predict a sequence of [MASK] tokens simultaneously. While the latter is technically feasible, it requires pre-determining the length of the span or incorporating additional supervision during training.

Exploring solutions beyond MLMs, recent pre-trained seq2seq models like T5 [19] are capable of performing the IDRR task. However, these models are computationally intensive, demanding substantial resources and time for training and inference. To address this in the context of question-answering, Kalinsky et al. [8] propose a straightforward yet effective solution for multi-token completion: the Extended-Matrix (EMAT) decoder, which outputs promising results and achieves state-of-the-art accuracy for named entity recognition (NER), thereby enabling multi-token names to serve as answers to questions. Among the MLMs they evaluated, including BERT[2], RoBERTa[11], SpanBERT[7], T5[19], BERT with the EMAT strategy outperformed all other models, therefore was chosen and adapted to our experiments.

This approach maintains the original MLM encoder to generate contextual embeddings for new multi-token phrases, treating them as a single mask filler. For instance, "New York", "Prime Minister", "the United Kingdom" will all be seen as single mask fillers. The output prediction matrix is then extended to incorporate the embeddings of these new phrases, making the model's size dependent on the expanded vocabulary. For example, if the model learns the embedding of "Prime Minister" during its training, then this embedding is added to the prediction matrix. By fine-tuning the model to learn the embeddings of these new phrases, it was expected to predict a multi-token named entities when seeing a [MASK] token in question-answering contexts. As shown in Ex.(11), when the masked text is fed as input, the model is expected to output "Prime Minister". However, our experiments focused on predicting multi-token connectives instead of named entities, requiring us to modify their method by fine-tuning the MLM on connectives to suit our specific task.

(11) The **[MASK]** of the United Kingdom at the moment is Keir Starmer.

Output: **Prime Minister**

Chapter 3

Approach and Implementation

Chapter 3 provides an overview of our approach to the multi-token connective prediction task. Section 3.1 details the process of collecting the connective vocabulary and their associated senses, and creating the training and test datasets. Section 3.2 illustrates how the Extended-Matrix (EMAT) solution [8] for named entity recognition is adapted for our specific task, using a concrete example to demonstrate the structure of the model. Section 3.3 presents the evaluation metrics—accuracy, precision, surprisal, and entropy—used to compare the performance of the preposed and canonical sets.

3.1 Data collection and preprocessing

3.1.1 Connective vocabulary and sense mapping

A vocabulary of 70 multi-token connectives was collected from the explicit and implicit connectives listed in appendices A and C of the PDTB-3 Annotation Manual [16], and Connective-Lex [25]. The Connective-Lex, released in 2017, complements newly-created lexicons of discourse connectives. The Connective-Lex for English contains connectives which are not present in PDTB-3. Connectives that typically function as subordinating conjunctions between clauses, rather than coordinating conjunctions between adjacent sentences, were manually excluded. Concurrently, a mapping between each connective to its potential senses listed in the two resources was established (see Appendix B), which would be used to calculate accuracy during the model's inference stage, as the accuracy of our experiments is based on predicting the sense inserted by the human annotators instead of predicting the one and only correct connective label. That is, if the model predicts "by contrast", it will be considered a correct prediction when

the annotated sense is "Comparison.Contrast," even if the human-inserted connective is "by comparison." This is because both "by comparison" and "by contrast" can signal "Comparison.Contrast".

3.1.2 Training and development datasets

For training purposes, ~838K argument-pairs where the second sentence (Arg2)[1] begins with a multi-token connective were extracted from the Wikipedia English dataset (20220301.en) available on Huggingface [5]. We masked out the multi-token connective in each Arg2. Together with Arg1, they formed into a masked text, formatted as "Arg1 [SEP][MASK], Arg2". We added a comma to separate the connective and the argument for simplicity. Each sentence pair is concatenated using [SEP], a special token that separates two sentences in BERT. The purpose of the training is just to enable the model to fill the [MASK] token with multi-token connectives, therefore we do not need annotations of discourse relations during the training. Table 3.1 shows an example in the training dataset and how it is organized in the CSV file with specified columns. The dataset was randomly split into training (80%) and development (20%) sets, with ~671K and ~167K samples respectively.

span	span_lower	range	text	freq	masked_text
As a result	as a result	[153,165]	At that time, people who had confirmed COVID-19 cases in Alberta, had recently returned from trips to "Iran, Egypt, Spain, Washington state and Mexico." As a result, the province requested that "all travellers returning from Italy" self-isolate for two weeks.	83171	At that time, people who had confirmed COVID-19 cases in Alberta, had recently returned from trips to "Iran, Egypt, Spain, Washington state and Mexico." [SEP][MASK], the province requested that "all travellers returning from Italy" self-isolate for two weeks.

Table 3.1: Example of the training dataset with columns including span: multi-token connectives, span_lower, range: where this connective is located, text: original concatenated text of Arg1 and Arg2, freq: how many times a connective appears in the dataset, and masked_text.

[1] Note that the arguments collected here are sentences, not the *Argument* satisfying the strict definition (see Section 2.2.1).

3.1.3 Test dataset

For inference, two additional datasets were used: PDTB-3 [16] and DiscoGeM 1.0 [21], both of which contain human annotations of implicit relations and inserted connectives. We constructed a preposed and canonical test set from the two corpora, each including 1598 samples. A section of the preposed and canonical test CSV files can be referred in Appendix C.

For the DiscoGeM 1.0 dataset, preposed structures were identified using the spaCy, NLTK, and constituent treelib libraries [6] in Python. We treat each sentence in the corpus as an individual argument. For instance, as shown in Ex.(12), the sentence is parsed to generate a constituency tree, which outlines the syntactic structure of the sentence by organizing it into hierarchical components such as S (sentence), PP, and etc. The constituency tree reveals that the phrase "All through that summer" is a PP located at the beginning of the sentence. Dependency parsing is subsequently applied to determine if the preposed NP or PP serves as the grammatical subject of the argument, identified by labels such as "nsubj," "nsubjpass," or "expl". If the phrase does not function as the subject (as in this example), it is classified as a preposed phrase. This method effectively isolates non-subject phrases that have been fronted in the sentence, often for emphasis or to provide context.

(12) **Example Arg**: All through that summer the work of the farm went like clockwork. [Animal Farm, DiscoGeM 1.0]

Constituency Tree:

```
(S
  (PP (ADVP (DT All)) (IN through) (NP (DT that) (NN summer)))
  (NP (NP (DT the) (NN work)) (PP (IN of) (NP (DT the) (NN farm))))
  (VP (VBD went) (PP (IN like) (NP (NN clockwork))))
  (. .))
```

Preposed phrase:

```
(PP All through that summer)
```

Similar to the training data, each masked text in the preposed set was formatted as "Arg1 [SEP][MASK], Arg2" before being input into the model. As illustrated in the

preposed Ex.(13), the PP "by the light of the match" is sentence-initial in Arg2, while in the canonical Ex.(14), the canonical masked text is constructed by right-moving the preposed phrase (either NP or PP) to the end of Arg2.

(13) Preposed masked text: He heard a slight groan. [SEP][MASK], **by the light of the match**$_{\text{preposed PP}}$ he saw a heavy shape moving slightly on the floor.$_{\text{Arg2}}$ [Animal Farm, DiscoGeM 1.0]

(14) Canonical masked text: He heard a slight groan. [SEP][MASK], he saw a heavy shape moving slightly on the floor **by the light of the match**$_{\text{canonical PP}}$.

The *Argument* that satisfies the strict definition that it is a text segment including at least a predicate is not marked in DiscoGeM 1.0. In fact, the arguments they collected are all complete sentences which may include multiple clauses. Ex.(15) shows an argument with a relative clause in DiscoGeM. The simple method of moving the preposed phrase "for most of its history" to the end of the clause not *Argument* can sometimes result in a canonical sentence that lacks natural flow, as in Ex.(16).

(15) **For most of its history AI research has been divided into sub-fields**$_{\text{argument}}$, which often fail to communicate with each other$_{\text{clause}}$. [Wikipedia, DiscoGeM 1.0]

(16) Unsatisfactory Canonical Arg2: AI research has been divided into sub-fields, which often fail to communicate with each other, **for most of its history**$_{\text{canonical phrase}}$ [Wikipedia, DiscoGeM 1.0]

In contrast, positioning "for most of its history" immediately after the *Argument* "AI research has been divided into sub-fields" and excluding the clause at the same time, as shown in Ex.(17), produces a more coherent sentence than moving it to the end of the sentence when involving a clause introduced by a complementizer such as "which" or "that". Although we noticed such a straightforward method may create less coherent canonical Arg2, these cases only account for less than 2% of the canonical set including samples from both corpus.

(17) Satisfactory Canonical Arg2: AI research has been divided into sub-fields **for most of its history**$_{\text{canonical phrase}}$. [Wikipedia, DiscoGeM 1.0]

In addition to the masked text and the annotated sense, metadata for each sample was recorded, including corpus, data source, genre, the inserted connective, and the

preposed phrase.[2] In the end, the preposed set and canonical set from DiscoGeM 1.0 each contains 157 inter-sentential discourse relation samples.

In terms of the preposed and canonical sets from PDTB-3, we directly used Dong's datasets [3], each comprising 1,441 inter-sentential implicit discourse relations. The PDTB-3 marks arguments within sentences, which constructs a satisfactory canonical Arg2 like Ex.(17) without redundant clauses.

Therefore, our contributions include the creation of a vocabulary of inter-sentential multi-token discourse connectives and their sense-mapping dictionary, the development of two extended mix-genre datasets—the preposed and canonical datasets, each including 1598 samples—derived from both PDTB-3 and DiscoGeM 1.0, while PDTB-3 contains news articles and DiscoGeM includes political speeches, literature and wikipedia texts.

3.2 Mask-filling

Figure 3.1 illustrates the architecture with the MLM encoder [2] and the Extended-Matrix (EMAT) decoder [8], which were built to predict an implicit discourse connective filling the [MASK] in each argument-pair. We first obtained the contextual embedding of each multi-token connective in our predefined connective list (see Section 3.1.1) via the MLM encoder and then these embeddings were fed into the extended-matrix decoder. We trained the EMAT decoder for three epochs on the whole training dataset containing argument-pair examples extracted from Wikipedia, and mapped all word vectors including these new phrases' embeddings to the output prediction matrix. Note that all new vectors will be only added to the prediction matrix instead of to the base model's vocabulary, which avoid retraining BERT. During inference, the formatted input, as illustrated in Figure 3.1 with a [MASK] token representing the inserted implicit connective, was tokenized and processed to compute the contextual embedding of the [MASK] token. The model subsequently generated predictions along with their probabilities. If the prediction corresponds to a multi-token connective in our predefined vocabulary, it is mapped to its respective senses and compared to the gold sense. For instance, in Figure 3.1, the sense "Comparison.Contrast" of the connective "In fact" is highlighted in red, indicating a match with the gold sense.

[2]To meet the default format of the Kalinsky et al.'s test data [8], two additional columns, span: default and span lower: default, were added in our test CSV files.

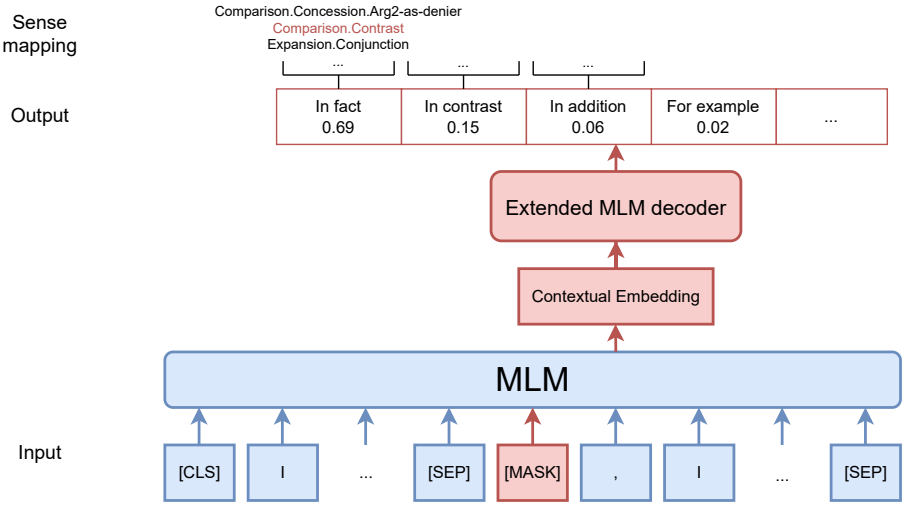

Figure 3.1: Model Architecture

3.3 Evaluation metrics

3.3.1 Accuracy and precision

Since our predictions are connectives rather than senses, we calculate the accuracy in a specific way. Multi-token connective predictions of the model would be mapped to all relation senses according to our sense-mapping dictionary and if there is a corresponding sense that matches the human-annotated sense, then the prediction will be counted as correct. The average accuracy of the model's top N predictions over the dataset, a@N, is computed within the following equation:

$$\text{a@N} = \frac{1}{k}\sum_{i=1}^{k} \max_{x \in \text{pred}_i^N} \left(\mathbb{1}_{\{sense(x)=gold_i\}}\right), \tag{3.1}$$

where the model's top N predictions for sample i are represented as pred_i^N. A single lexical entry x is considered correct if it can convey the gold sense $gold_i$ as per the sense-mapping dictionary $sense(x)$. If a prediction is correct, which means the subscript $sense(x) = gold_i$ is satisfied, thereby $\mathbb{1}_{\{sense(x)=gold_i\}} = 1$, otherwise $\mathbb{1}_{\{sense(x)=gold_i\}} = 0$. Since we only consider if there is a correct answer across the top N predictions rather than how many of them are correct, therefore a max function is applied. If any of the top N predictions is correct, the entire prediction for sample i is deemed correct. We compute the dataset's accuracy by averaging over all k samples.

Precision measures how many of the predictions are correct relative to how many are made. For instance, if both of the top 2 predictions are correct, p@2 is 100%, and if only one of them is correct, then p@2=50%. The average precision of the model's top N predictions over the dataset, p@N, is computed as the following equation:

$$\text{p@N} = \frac{1}{k}\sum_{i=1}^{k}\frac{\sum_{x\in \text{pred}_i^N} \mathbb{1}_{\{sense(x)=gold_i\}}}{N}, \tag{3.2}$$

where items are similarly denoted as in Eq.(3.1), for a@N. For each sample i we count how many predictions are correct out of the top N predictions $pred_i^N$ and divide by N. This calculation is also averaged across all k samples.

3.3.2 Surprisal and entropy

The model's prediction certainty can be quantified using two statistical measures: surprisal, in the context of information theory [22], and entropy. Surprisal measures the unexpectedness or unpredictability of the human-annotated sense for the model. Specifically, it quantifies how "surprised" the model is by a specific prediction. If the model is less surprised by the human annotated gold sense, this means that the model assigns a high probability to a connective which can convey that sense, indicating that the model's prediction is close to the gold label. Therefore, a small surprisal can suggest that the model is confident and that the prediction is expected and matches the gold label. It is worthwhile to explore how preposing influences this certainty. The model generates a probability distribution over the entire vocabulary for each sample. Surprisal for a dataset is defined as the summed negative log likelihood (NLL) over all samples:

$$NLL = -\sum_{i=1}^{k}\sum_{x\in V}\log p_i(x)\cdot \mathbb{1}_{\{sense(x)=gold_i\}}, \tag{3.3}$$

where x denotes a lexical entry within the vocabulary V, and the summation extends over k samples in the dataset.[3] A lexical entry x is considered correct if it is a connective and can convey the annotated implicit sense.

Ex.(18) illustrates a low surprisal case. The model's top 2 predictions are "For example" and "For instance", which can both convey the gold sense *Arg2-as-instance* according to our sense-mapping dictionary (see Appendix B). When the model's correct predictions hold the larger portion of the total probability in the distribution, indicating a greater certainty on correctness, the output's surprisal will be low as this example.

[3] Here, we do not limit the predictions to top N, but use the whole vocabulary instead to compute surprisal and the followed entropy.

(18) **Low surprisal example**: Jim Beam print ads, however, strike different chords in different countries.$_{\text{Arg1}}$ [SEP][MASK], in Australia, land of the outback, a snapshot of Jim Beam lies on a strip of hand-tooled leather.$_{\text{Arg2}}$ [wsj_1274, PDTB-3]

Gold sense: *Expansion.Instantiation.Arg2-as-instance*

Model output over the vocabulary (mask-filler with its probability in descending order):

```
[({'For example'}, 0.8127207), ({'For instance'}, 0.18605553),
('In particular', 0.0003257261), ('In fact', 0.00030983146),
('In contrast', 0.00013416453), ...]
```

Surprisal: 0.00047635453

In contrast, Ex.(19) shows a high surprisal case. We observe that the top 5 can not signal the gold sense and they hold a larger portion of the total probability, thus the correct connectives that can convey the sense only account for small probabilities, resulting in a high surprisal. This suggests that the model is confused and uncertain about which prediction is correct.

(19) **High surprisal example**: But they didn't lose touch with the U.S. issuers.$_{\text{Arg1}}$ [SEP][MASK], since 1985, Japanese investors have bought nearly 80% of $10 billion in Fannie Mae corporate debt issued to foreigners.$_{\text{Arg2}}$ [wsj_0274, PDTB-3]

Gold sense: *Expansion.Substitution.Arg2-as-subst*

Model output over the vocabulary (mask-filler with its probability):

```
[('For example', 0.44564933), ('In fact', 0.18052544),
('In addition', 0.10665635), ('For instance', 0.09934871),
('As a result', 0.047118817),...]
```

Surprisal: 12.167008

Entropy, on the other hand, measures the model's general certainty across all its predictions, regardless of correctness. It assesses the spread of the probability distribution over all possible predictions (i.e. the model's vocabulary). A higher entropy

value suggests a more dispersed or even probability distribution, indicating greater uncertainty in the model's predictions. Conversely, a lower entropy value signifies a more concentrated probability distribution when top predictions are generated with significantly higher probabilities, implying a higher degree of certainty about the predictions, nevertheless it does not specify whether these predictions are correct. The entropy for a dataset is calculated as follows:

$$H = -\sum_{i=1}^{k} \sum_{x \in V} p_i(x) \log p_i(x), \tag{3.4}$$

with all parameters similarly defined as in Eq.(3.3). The overall entropy for the dataset is obtained by summing across all k samples.

The following is an example of low entropy. The first prediction' probability is 0.94, almost accounting for the total probability share, and the remaining probabilities would be accordingly small. This suggest that the model is very confident about its prediction, so it assigns the first prediction a probability of 0.94, no matter the result is correct or wrong.

(20) **Low entropy example**: He's currently in the midst of a 17-city U.S. tour with Yehudi Menuhin and the Warsaw Sinfonia, with stops including Charleston, S.C. (Oct. 25), Sarasota, Fla. (Oct. 28), Tampa, Fla. (Oct. 29) and Miami (Oct. 31).$_{\text{Arg1}}$ [SEP][MASK], later this season he gives a recital at Washington's Kennedy Center, and appears as soloist with several major orchestras.$_{\text{Arg2}}$ [wsj_1388, PDTB-3]

Model output over the vocabulary (mask-filler with its probability):

```
[('In addition', 0.9442712), ('As well', 0.025426337),
('In fact', 0.0075221206), ('At the same time', 0.0028811994),
('After that', 0.0020392046),...],
```

Entropy: 0.33977264

Ex.(21) is a high entropy example. The top 1 prediction only has a relatively low probability at 0.13, and the other illustrated probabilities are all very small ranging from 0.09 to 0.06. The remaining probability mass which is not present is definite to have a probability no greater than 0.06. This suggest that the model is perplexed, and it does

not sure which prediction is right or wrong, so it makes an even guess given the possible answers it can give.

(21) **High entropy example**: She used the market's wild swings to buy shares cheaply on the sell-off.$_{\text{Arg1}}$ [SEP][MASK], on the comeback, Ms. Del Signore unloaded shares she has been aiming to get rid of.$_{\text{Arg2}}$ [wsj_1208, PDTB-3]

Model output over the vocabulary (mask-filler with its probability):

```
[('In fact', 0.12930626), ('At the same time', 0.08424396),
('As a result', 0.08413355), ('In addition', 0.059824232),
('In the end', 0.05716916)],
```

Entropy: 3.2894974

Chapter 4

Results and Evaluation

In this chapter, we compared the model's prediction result between the preposed set and the canonical set. Section 4.1 analyzes the predicted mask-fillers. Section 4.2 evaluates the performance between the two sets from different perspectives: their accuracy, precision, surprisal, and entropy, the sense types that the predicted connectives are matched with. We also identified which sense types benefit the most from the preposed structure, and examined if preposing is helpful across genre.

4.1 Predicted mask-fillers

The output of each prediction consist of top N lexical entries and their corresponding probabilities. Before analyzing these predictions, we assume that all lexical entries that could serve as connectives are indeed connectives, even if they might also hold other syntactic roles. For instance, "on the other hand" could also mean the hand of a person or imply that someone is wearing or carrying something on that hand.

Top 5 mask fillers The model's output is organized in descending order from the most to the least likely predictions. Table 4.1 shows the average probabilities of each of the top 5 predictions for both the preposed and canonical sets. Notably, the first prediction accounts for nearly half of the total probability in both sets. The cumulative average probability for the top 5 predictions in each set approximates 80%, with the residual probability mass distributing over less likely predictions.

Most importantly, all of the top 5 predictions are multi-token connectives, which gives an evidence for the effectiveness of the fine-tuning process. In contrast, Dong's results on single-token experiments [3] showed that only about 60% of the top 5

N^{th} Prediction	Average Probability	
	Preposed Set	Canonical Set
1	0.459621	0.414769
2	0.163837	0.161387
3	0.089585	0.093956
4	0.058879	0.064359
5	0.041498	0.046896
Top 5 Cumulative Average Prob	**0.813419**	**0.781367**

Table 4.1: Average probabilities in preposed set vs. canonical set.

predictions in the preposed set and 55% in the canonical sets can function as connectives, and in 4% of the preposed samples and 13% of the canonical samples, BERT failed to include any connectives in the top 5 predictions.

Ambiguity One of the reason why we extend Dong et al.[3]'s work to multi-token scenario is that we hoped that the model's predictions can be more specific than the single-token connective experiments' results, because multi-token connectives are less ambiguous. Therefore, the preposed set's results from Dong's single-token experiments and our experiments are compared qualitatively to validate our belief. Ex.(22) and Ex.(23) are illustrated for analysis between the single token and multi-token output. For simplicity, we only present the top 5 predictions, and the rest of predictions are represented by "...".

In Ex.(22), only "and" in the top 5 predictions from Dong's results is a connective and can convey the gold sense *Expansion.Conjunction*, while the ambiguous single-token "and" actually holds 11 possible senses. Conversely, our model's top 5 predictions are all connectives. Among them, "In addition" with two possible senses, "In fact" with 11 possible senses, "At the same time" with 2 possible senses can all signal the gold sense *Expansion.Conjunction*. Comparatively, our multi-token predictions are less ambiguous since their possible senses are limited and specific.

(22) **Masked_text**: The peninsula comes off the vast southeastern alluvial plain with fields of rice and cotton and sorghum as far as the eye can see. Near the coast there are dense coverts of live oak interspersed with marshes and prairies. Deer, wild hog, armadillos and alligators are the glamour quadrupeds and the birds are innumerable, especially the herons and the spoonbills. [SEP][MASK], above

the blossoms of lantana and scarlet pea the inky-brown and golden palamedes butterfly floats on its lazy wingbeat. [wsj_1323, PDTB-3]

Gold sense: *Expansion.Conjunction*

Dong et al's output:

```
[('high', 0.4602), (just', 0.1779),
('and', 0.0686), ('far', 0.025),
('up', 0.0216),...]
```

Our output:

```
[('In addition', 0.33142963), ('On the other hand', 0.14200377),
('On the other', 0.11193432), ('In fact', 0.10518605),
('At the same time', 0.09251382),...]
```

Another example is illustrated in Ex.(23), when the gold sense is *Expansion.Level-of-detail.Arg2-as-detail*. We still observe that among Dong's predictions, "and" is correct but too general with 11 sense types. On the contrary, "As a result", "In the end", "In addition", and "For example" with three, eight, two, three sense types respectively, are less ambiguous.

(23) **Masked_text**: And pressure by big investors forced Donaldson Lufkin & Jenrette Securities Corp. to sweeten Chicago & North Western's $475 million junk bond offering. [SEP][MASK], after hours of negotiating that stretched late into Thursday night, underwriters priced the 12-year issue of resettable senior subordinated debentures at par to yield 14.75%, higher than the 14.5% that had been expected. [wsj_1464, PDTB-3]

Gold sense: *Expansion.Level-of-detail.Arg2-as-detail*

Dong et al's output:

```
[('but', 0.2355), ('and', 0.2304),
('so', 0.07), ('finally', 0.0698),
('"', 0.0526),...]
```

Our output:

```
[('As a result', 0.22718893), ('In the end', 0.18287785),
('In addition', 0.12972623), ('For example', 0.09070829),
('At the same time', 0.04366458),...]
```

4.2 Preposed set vs. canonical set

4.2.1 Accuracy and Precision

Table 4.2 compares the model's predictions for the preposed and the canonical set in terms of a@N as computed in Eq.(3.1), and p@N as computed in Eq.(3.4), where N is 1, 2, 3, 4, 5. The preposed set consistently achieves higher accuracy across all N values, with accuracy increasing from 58.95% at N = 1 to 91.05% at N = 5. The results shows that the model's predictions align more closely with human annotations for the preposed set than for the canonical set, suggesting that a preposed structure can provide hints for recognizing discourse relations.

For precision, the preposed set also consistently outperforms the canonical set, but the difference in performance becomes marginally less pronounced as N increases. At N=1 and N=2, precision in the preposed set is markedly higher than in the canonical set. By N = 5, the preposed set still leads (48.82% vs. 45.21%), but the gap marginally narrows, indicating a decrease in the relative advantage of the preposed structure for predicting more precisely when more predictions are considered for a sample.

	Preposed Set		Canonical Set	
N	a@N	p@N	a@N	p@N
1	**58.95%**	**58.95%**	54.38%	54.38%
2	**75.84%**	**56.79%**	71.09%	51.56%
3	**84.79%**	**55.09%**	79.97%	49.97%
4	**88.99%**	**51.60%**	85.48%	47.50%
5	**91.05%**	**48.82%**	88.24%	45.21%

Table 4.2: a@N and p@N in preposed set vs. canonical set.

4.2.2 Surprisal and entropy

Table 4.3 provides a comparison between the preposed and canonical sets concerning average surprisal as computed in Eq.(3.3), and entropy as computed in Eq.(3.2) for all discourse relation samples. The preposed set consistently demonstrates an advantage: both the average surprisal and the entropy are substantially lower in the preposed set compared to its canonical counterpart. This indicates that the model is not only more certain about its general predictions, irrespective of its correctness, but also shows greater certainty when its predictions align with human annotations, particularly when the text has a preposed structure. This difference validates that a preposed structure can improve the model's performance in terms of the general certainty across the dataset and the certainty on correct predictions.

Metric	Preposed Set	Canonical Set
Average Surprisal	**1.118**	1.228
Average Entropy	**1.888**	2.054

Table 4.3: Average surprisal and entropy in preposed set vs. canonical set.

Ex.(24) of the preposed set and Ex.(25) of the canonical set qualitatively compares model's surprisal result on the same sample in the two sets, respectively. Among the predictions present in Ex.(24), the 2^{nd}, 4^{th}, 5^{th} prediction can convey the gold sense *Comparison.Contrast*, and some of the remaining predictions may signal the gold sense as well. A cumulative of these correct predictions results in a low surprisal of 0.95.

(24) **Preposed masked_text**: In late afternoon New York trading yesterday, the dollar stood at 1.8415 West German marks, up from 1.8340 marks late Monday, and at 142.85 yen, up from 141.90 yen late Monday. [SEP][MASK], **a month ago**$_{\textbf{preposed phrase}}$, a similar survey predicted the dollar would be trading at 1.8690 marks and 139.75 yen by the end of October. [wsj_0301, PDTB-3]

Gold sense: *Comparison.Contrast*

Model output over the vocabulary (mask-filler with its probability):

```
[('In addition', 0.35656312), ('By comparison', 0.12282629),
('Since then', 0.09480629), ('In fact', 0.0916691),
('In contrast', 0.0638354),...]
```

Surprisal: 0.95212275

Comparatively, in Ex.(25) of the canonical set, the 3^{rd}, 4^{th}, 5^{th}, and other correct predictions with a lower probabilities produce a higher surprisal of 1.47.

(25) **Canonical masked_text**: In late afternoon New York trading yesterday, the dollar stood at 1.8415 West German marks, up from 1.8340 marks late Monday, and at 142.85 yen, up from 141.90 yen late Monday. [SEP][MASK], a similar survey predicted the dollar would be trading at 1.8690 marks and 139.75 yen by the end of October **a month ago**$_{\textbf{preposed phrase}}$. [wsj_0301, PDTB-3]

Gold sense: *Comparison.Contrast*

Model output over the vocabulary (mask-filler with its probability):

```
[('In addition', 0.400831), ('At the same time', 0.2552863),
('In comparison', 0.059462074), ('By comparison', 0.04943322),
('In fact', 0.04680012),...]
```

Surprisal: 1.4653729

Ex.(26) and Ex.(27) qualitatively compares model's entropy result on the same sample for two sets, one with a preposed structure, and one without. The model testing on the preposed masked_text produces a probability distribution where the top 1 prediction enjoys the largest share at 0.61, while the first prediction's probability on the canonical masked_test is 0.48. This results in sparse distribution in the former and a more evenly distribution in the latter. After a summation of $plogp$ (see Eq.(3.4) in Section 4.2.2), the distribution of the preposed text achieves a lower entropy compared with the canonical text.

(26) **Preposed masked_text**: Morgenzon has long been a special domain of Afrikanerdom. [SEP][MASK], **according to Mr. Verwoerd**$_{\textbf{preposed phrase}}$ the early Afrikaner pioneers were the first people to settle in the eastern Transvaal, even before the blacks. [wsj_1760, PDTB-3]

Model output over the vocabulary (mask-filler with its probability):

```
[('In fact', 0.6146381), ('For example', 0.25532994),
('For instance', 0.09028673), ('In particular', 0.012352365),
```

```
('In addition', 0.0058183167),...]
```

Entropy: 1.0934856

(27) **Canonical masked_text**: Morgenzon has long been a special domain of Afrikanerdom. [SEP][MASK], the early Afrikaner pioneers were the first people to settle in the eastern Transvaal, even before the blacks **according to Mr. Verwoerd**$_{\textbf{preposed phrase}}$. [wsj_1760, PDTB-3]

Model output over the vocabulary (mask-filler with its probability):

```
[('In fact', 0.48055366), ('For example', 0.32374576),
('For instance', 0.14666279), ('In particular', 0.018653061),
('After all', 0.0057877456),...]
```

Entropy: 1.2639269

4.2.3 Prediction certainty across sense types

Figure 4.1 presents a scatter plot of the prediction probabilities when both the preposed and canonical sets give a correct top 1 prediction. Each point represents a sample, with the x-axis indicating the probability of correct prediction by the preposed set and the y-axis for the canonical set. The red line, which represents y=x, is used as a reference to evaluate the consistency between the two prediction sets.

It is observed from the distribution that the majority of the points are scattered along the red line, suggesting that both sets are likely to give a correct prediction for a sample with a similar probability. However, more points appear below the red line, indicating that in many instances, when the model makes a correct prediction in both sets, it assigns a higher probability to its prediction in the preposed set than in the canonical set. This implies that the model is more confident in samples with a preposed structure.

To further assess the model's performance on the two sets across various sense types, we evaluated the number of correct top 1 predictions for each set. Chi-square tests were performed between the sets for certain sense types. Table 4.4 lists counts for top 8 sense types each represented by over 100 samples in the test dataset, sorted by descending order of frequency, and the corresponding number of correct predictions for each set.

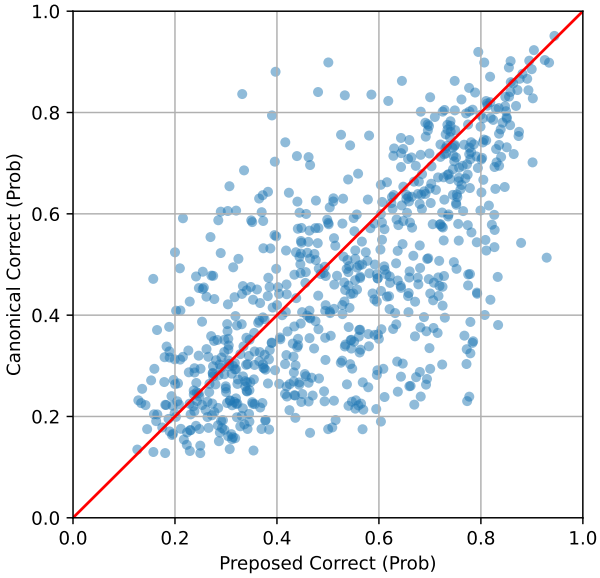

Figure 4.1: Scatter plot of the probabilities of samples when both the preposed and the canonical set predict correctly. Only Top 1 predictions are considered.

The results revealed significant differences between the two sets for three specific sense types: *Expansion.Level-of-detail.Arg2-as-detail*, *Expansion.Instantiation.Arg2-as-instance*, and *Contingency.Cause.Reason*, all at a significance level of 0.05. These findings on multi-token connective predictions are consistent with results from Dong et al. [3], who reported significant prediction differences in four sense types: *Expansion.Conjunction* ($p_{conjunction} = 0.05$ in our findings), *Expansion.Level-of-detail.Arg2-as-detail*, *Expansion.Instantiation.Arg2-as-instance*, and *Contingency.Cause.Reason* between the two sets. Notably, the preposed set of our experiments also consistently predicted more samples correctly for the three significant sense types.

In light of the surprisal discussed in Section 4.2.2, which suggests that the model predicted with greater certainty on correct predictions in the preposed set generally, we explored whether this trend also applied specifically to the aforementioned sense types in Dong et al.[3] and our work. Figure 4.2 employs a Kernel Density Estimate (KDE) plot of top 1 prediction probabilities across four sense types, where the y-axis represents the density of predictions' probability rather than the count of a certain probability. KDE plot is a smoothed version of a histogram. Here, it smooths the distribution of top 1 predictions' probabilities, providing a continuous probability density curve, which shows the likelihood of a prediction's probability falling at different values along the x-

Sense Type	N	Preposed	Canonical	χ^2	p
Expansion.Conjunction	341	176	202	3.71	.05
Expansion.Level-of-detail.Arg2-as-detail	241	226	202	11.03	*
Expansion.Instantiation.Arg2-as-instance	191	178	159	8.16	*
Contingency.Cause.Reason	191	134	99	12.72	*
Contingency.Cause.Result	184	79	88	0.70	.40
Comparison.Contrast	139	74	59	2.83	.09
Temporal.Asynchronous.Precedence	131	16	13	0.16	.69
Comparison.Concession.Arg2-as-denier	101	42	31	2.15	.14

Table 4.4: Correct top 1 predictions for senses (with more than 100 samples) in preposed set vs. canonical set: counts, and χ^2 test results. N is the frequency of each sense type in the dataset

axis. The area under the entire curve sums to one. Take Figure 4.2(a) as an example, the red curve is labeled as Canonical Incorrect (X), which represents the canonical samples whose top 1 prediction is incorrect, and we collect these samples' top 1 probabilities to draw a KDE plot. Mode is the point where the smoothed density is highest, which represents the most frequent probability in our case. The mode of the red curve is around 0.25, which means that this label's probability mass is packed around 0.25, a low probability, and therefore the model is less certain about this label. What's more, the steepness and flatness can also provide some information about predictions' probability distribution. We observe that the red curve is steep with a low variance, indicating the model's uncertainty is applied to many samples in this label.

Arg2-as-detail Analyzing the Figure 4.2(a), we see different distributions of prediction probabilities for the four labels in the figure legend. The Preposed Correct (V), in blue, which means that the preposed set gives a correct prediction, has the largest mode around 0.7 among four labels, indicating the preposed set tends to make correct predictions with high confidence. In contrast, the Preposed Incorrect (X) curve, colored orange, has a mode around 0.3, suggesting that although the model still makes some incorrect predictions, it is generally made with low confidence. As we mentioned before, the red curve for Canonical Incorrect (X) is steep with its mode at 0.25, suggesting many predictions' of this label are made with low probabilities.

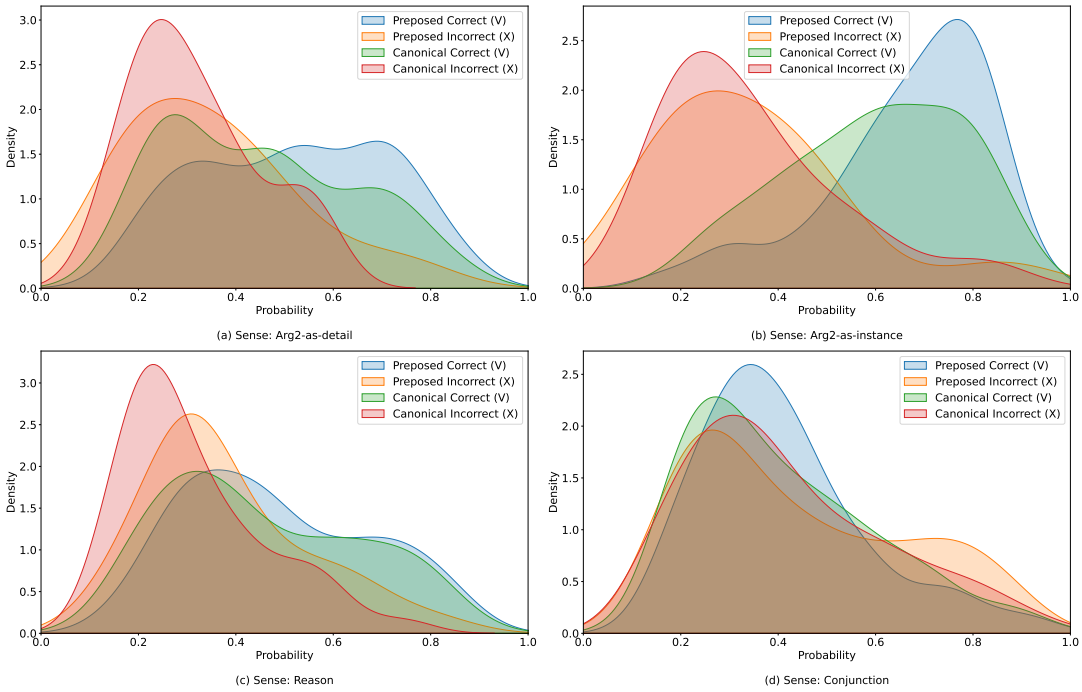

Figure 4.2: Kernel Density Estimate (KDE) plots of top 1 prediction probabilities for sense types: *Arg2-as-detail*, *Arg2-as-instance*, *Reason, Conjunction*.

Arg2-as-instance In Figure 4.2(b), the trend we noticed previously is more pronounced. The Preposed Correct (V), represented by the blue left-skewed curve, peaks sharply with a large mode around 0.8. This pattern suggests that when the preposed set makes correct predictions, it does so with a high degree of confidence. This finding validates that the preposing structure help significantly on recognizing the sense type of *Arg2-as-instance*. Furthermore, this observation collaborates with Ward and Birner's analysis [26], which discusses how a preposed constituent following a preceding argument (Arg1) typically represents old or previously mentioned information. This structural choice not only emphasizes known information but also improves clarity in communication within a discourse. In the specific context of *Arg2-as-instance*, the preposed constituent often relates to a hierarchical relationship, such as a set in Arg1 and its elements in Arg2, assisting a clearer understanding of the discourse relation.

Reason While the four curves in Figure 4.2(c) are closely aligned, it is still notable that the mode of the Preposed Correct (V) curve is marginally larger than those of the other three labels. Although the peaks of incorrect sets are higher than the other two, the density is not equivalent of counts, which does not mean that incorrect predictions are more than correct ones. We can only observe the curves of incorrect sets are more

steep with a low variance, while the other two are more flat with a high variance, which is a good sign, indicating that when the model is perplexed, it assigns a low probability to its top 1 prediction.

Among the three sense types in Figure 4.2(a), (b), and (c), a consistent pattern is that the canonical labels always has a smaller probability mode compared to its preposed counterpart, and the model is more certain on the preposed set's predictions with the largest mode when its prediction is correct. This observation is also compatible with our earlier comparative analysis of surprisal across all discourse relations in the two dataset (see Section 4.2.2), where the preposed set has a smaller average surprisal.

Conjunction The last but not least, our findings diverge from those of Dong et al. [3] in *Conjunction*. As detailed in Table 4.4, the preposed set demonstrates fewer correct predictions compared to the canonical set, marking a deviation across all sense types examined. Upon further investigation in Figure 4.2(d), we observe that despite fewer correct predictions, the Preposed Correct (V) label still has the largest mode relative to the other three labels.

4.2.4 Analysis of genre

Our previous analysis mainly consider top 1 predictions, but since our genre-mix samples are limited, therefore treating each of the top 5 predictions individually can quintuple the samples for the count and give us a broader view of prediction accuracy across different genres. Examples for each genre in DiscoGeM 1.0 has been given in Table 2.1 in Section 2.2.2.

Table 4.5 compares correct top 5 predictions in the preposed and canonical sets across four different genres: News articles (WSJ), Wikipedia, Literature, and Political speeches (Europarl). The model achieved high accuracy on news articles, approximately half of the predictions being correct, which may due to the structured, formal style of facts writing. Wikipedia follows a similar trend, benefiting from its encyclopedic and descriptive nature. The genre of political speeches, due to the smallest sample size (a narrower range because only six discourse relations are involved in this genre), makes them easier to predict, thereby showing the highest accuracy. In contrast, Literature presents a lower accuracy, reflecting the genre's complexity with nuanced and diverse language styles, including figurative expressions and intricate narrative forms that challenge the models ability to recognize the relation between sentences.

When comparing the two datasets across these genres, the canonical set, while competitive, generally falls short of the preposed set's performance, indicating that a preposed structure improves model's prediction accuracy across genres and this improvement is more evident in structured genres.

Genre	**Frequency (N)**	**Preposed Set**		**Canonical Set**	
		N	**%**	**N**	**%**
News articles	7205	3563	49.5	3305	45.9
Wikipedia	460	205	44.6	184	40.0
Literature	250	91	36.4	82	32.8
Political speeches	75	42	56.0	41	54.7

Table 4.5: Correct top 5 predictions for four genres in preposed set vs. canonical set: N: count, %: proportion.

Chapter 5

Future Work and Conclusions

We concluded our project as follows:

1. We adapted the extended output prediction matrix decoder solution raised by Kalinsky et al. [8] to train a multi-token mask-filler on the implicit discourse relation recognition task. The top 5 outputs from our model are all less ambiguous multi-token connective with an accuracy at ~90%, proving that this strategy is effective not only in their original named entity recognition task but also on other multi-token prediction scenarios.

2. We extended the claim that a preposing structure can help MLMs predict a single token connective in a discourse [3] to multi-token scenario. The preposing structure improves the model's general accuracy, accuracy across genres, and certainty on correct predictions, specifically on the three sense types: *Expansion.Level-of-detail.Arg2-as-detail*, *Expansion.Instantiation.Arg2-as-instance*, and *Contingency.Cause.Reason*.

In our future work, we can consider three directions: more discourse relations, mores senses and connectives, larger high-quality datasets.

Our study focused on inter-sentential relations, but intra-sentential relations is also worthy of more research. In choice of the gold label, a single sense was preferred, but there are research showing that sometimes a discourse may hold more than one relations simultaneously. The PDTB-3 also provides two connectives and senses if they exist in the discourse, therefore can be used to apply on more cases rather than limiting on single sense scenario.

In terms of the test data, we mentioned another corpus, namely GUM, annotated in RST-style. If converting the RST-style into PDTB style, we can have having more

discourse relations sample in the inference. We used the PDTB-3 and DiscoGeM 1.0, while the latter did not mark the *Argument*, therefore we could use the NLP toolkit to construct satisfactory canonical arguments.

Moreover, our study applied the extended decoder matrix strategy on predicting connectives, therefore our training data only includes argument-pairs with a multi-token connective and the model dominantly gave a multi-token prediction, showcasing a successful fune-tuning. However, if we expects to generate a natural prediction distribution, the training data should be single-token inclusive. Can MLMs capture the nuances between a single-token connective and a multi-token connective when their sense is similar or even same. Specifically, does their predictions favor a common and light single-token connective such as "and" or "but," rather than a heavy and sophisticated multi-token connective like "at the same time" or "in contrast," or vice versa because multi-token connectives are more unambiguous? Limited studies [29] discussed the deviation on connective selection between language models and humans, or ever validated that a well-trained model can differentiate similar (single- and multi-token) connectives and use them appropriately as humans do.

Bibliography

[1] Hongxiao Bai and Hai Zhao. Deep enhanced representation for implicit discourse relation recognition. *arXiv preprint arXiv:1807.05154*, 2018.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[3] Yunfang Dong, Xixian Liao, and Bonnie Webber. Syntactic Preposing and Discourse Relations. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2790–2802, 2024.

[4] Anca Dumitrache, Oana Inel, Lora Aroyo, Benjamin Timmermans, and Chris Welty. Crowdtruth 2.0: Quality metrics for crowdsourcing with disagreement. *arXiv preprint arXiv:1808.06080*, 2018.

[5] Wikimedia Foundation. Wikimedia downloads. https://dumps.wikimedia.org, 2024.

[6] Oren Halvani. Constituent Treelib - A Lightweight Python Library for Constructing, Processing, and Visualizing Constituent Trees. https://github.com/Halvani/constituent-treelib, April 2024.

[7] Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. Spanbert: Improving pre-training by representing and predicting spans, 2020.

[8] Oren Kalinsky, Guy Kushilevitz, Alexander Libov, and Yoav Goldberg. Simple and Effective Multi-Token Completion from Masked Language Models. In Andreas Vlachos and Isabelle Augenstein, editors, *Findings of the Association for*

Computational Linguistics: EACL 2023, pages 2356–2369, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics.

[9] Yudai Kishimoto, Yugo Murawaki, and Sadao Kurohashi. A knowledge-augmented neural network model for implicit discourse relation classification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 584–595, 2018.

[10] Yang Liu, Sujian Li, Xiaodong Zhang, and Zhifang Sui. Implicit discourse relation classification via multi-task neural networks. *Proceedings of the AAAI conference on artificial intelligence*, 30(1), 2016.

[11] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019.

[12] William C Mann and Sandra A Thompson. *Rhetorical structure theory: A theory of text organization*. University of Southern California, Information Sciences Institute Los Angeles, 1987.

[13] T Mikolov. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[14] Joonsuk Park and Claire Cardie. Improving implicit discourse relation recognition through feature set optimization. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 108–112, 2012.

[15] Rashmi Prasad, Bonnie Webber, and Aravind Joshi. Reflections on the Penn Discourse TreeBank, Comparable Corpora, and Complementary Annotation. *Computational Linguistics*, 40(4):921–950, December 2014.

[16] Rashmi Prasad, Bonnie Webber, Alan Lee, and Aravind Joshi. Penn discourse treebank version 3.0. *LDC2019T05*, 2019.

[17] Lianhui Qin, Zhisong Zhang, and Hai Zhao. A stacking gated neural architecture for implicit discourse relation classification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2263–2270, 2016.

[18] Lianhui Qin, Zhisong Zhang, Hai Zhao, Zhiting Hu, and Eric P Xing. Adversarial connective-exploiting networks for implicit discourse relation classification. *arXiv preprint arXiv:1704.00217*, 2017.

[19] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.

[20] Attapol Rutherford, Vera Demberg, and Nianwen Xue. 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, 2017.

[21] Merel C. J. Scholman, Tianai Dong, Frances Yung, and Vera Demberg. Discogem: A crowdsourced corpus of genre-mixed implicit discourse relations. In *Proceedings of the Thirteenth International Conference on Language Resources and Evaluation (LREC'22)*, Marseille, France, June 2022. European Language Resources Association (ELRA).

[22] Claude Elwood Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.

[23] Wei Shi and Vera Demberg. Learning to explicitate connectives with seq2seq network for implicit discourse relation classification. *arXiv preprint arXiv:1811.01697*, 2018.

[24] Manfred Stede. *Discourse Processing*. Morgan & Claypool Publishers, 2012.

[25] Manfred Stede, Tatjana Scheffler, and Amália Mendes. 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), 2019.

[26] Gregory Ward and Betty Birner. Information structure and non-canonical syntax. *The handbook of pragmatics*, pages 152–174, 2006.

[27] Wei Xiang and Bang Wang. A survey of implicit discourse relation recognition. *ACM Computing Surveys*, 55(12):1–34, 2023.

[28] Yu Xu, Man Lan, Yue Lu, Zheng Yu Niu, and Chew Lim Tan. Connective prediction using machine learning for implicit discourse relation classification. In *The 2012 international joint conference on neural networks (ijcnn)*, pages 1–8. IEEE, 2012.

[29] Frances Yung, Merel Scholman, and Vera Demberg. A practical perspective on connective generation. In *Proceedings of the 2nd Workshop on Computational Approaches to Discourse*, pages 72–83, 2021.

[30] Frances Yung, Merel C. J. Scholman, Sarka Zikanova, and Vera Demberg. Discogem 2.0: A parallel corpus of english, german, french and czech implicit discourse relations. In *Proceedings of the Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING'24)*, Turin, Italy, May 2024. European Language Resources Association (ELRA) and International Committee on Computational Linguistics (ICCL).

[31] Amir Zeldes. The GUM corpus: Creating multilayer resources in the classroom. *Language Resources and Evaluation*, 51(3):581–612, 2017.

Appendix A

PDTB-3 Sense Hierarchy

The sense hierarchy is from the PDTB-3 Annotation Manual [16].

Level-1	**Level-2**	**Level-3**
TEMPORAL	SYNCHRONOUS	–
	ASYNCHRONOUS	PRECEDENCE
		SUCCESSION
CONTINGENCY	CAUSE	REASON
		RESULT
		NEGRESULT
	CAUSE+BELIEF	REASON+BELIEF
		RESULT+BELIEF
	CAUSE+SPEECHACT	REASON+SPEECHACT
		RESULT+SPEECHACT
	CONDITION	ARG1-AS-COND
		ARG2-AS-COND
	CONDITION+SPEECHACT	–
	NEGATIVE-CONDITION	ARG1-AS-NEGCOND
		ARG2-AS-NEGCOND
	NEGATIVE-CONDITION+SPEECHACT	–
	PURPOSE	ARG1-AS-GOAL
		ARG2-AS-GOAL
COMPARISON	CONCESSION	ARG1-AS-DENIER

Continued on next page

Level-1	**Level-2**	**Level-3**
		ARG2-AS-DENIER
	CONCESSION+SPEECHACT	ARG2-AS-DENIER+SPEECHACT
	CONTRAST	–
	SIMILARITY	–
EXPANSION	CONJUNCTION	–
	DISJUNCTION	–
	EQUIVALENCE	–
	EXCEPTION	ARG1-AS-EXCPT
		ARG2-AS-EXCPT
	INSTANTIATION	ARG1-AS-INSTANCE
		ARG2-AS-INSTANCE
	LEVEL-OF-DETAIL	ARG1-AS-DETAIL
		ARG2-AS-DETAIL
	MANNER	ARG1-AS-MANNER
		ARG2-AS-MANNER
	SUBSTITUTION	ARG1-AS-SUBST
		ARG2-AS-SUBST

Appendix B

Multi-token Connectives and Their Senses

Connectives	Senses
after all	Contingency.Cause+Belief.Reason+Belief Expansion.Conjunction Expansion.Level-of-detail.Arg2-as-detail
after that	Temporal.Asynchronous.Succession
along with	Expansion.Conjunction
and then	Expansion.Disjunction
as a consequence	Contingency.Cause.Result
as a result	Contingency.Cause.Result Contingency.Cause+Belief.Result+Belief Expansion.Level-of-detail.Arg2-as-detail
as an alternative	Expansion.Disjunction
as it turns out	Contingency.Cause.Result Expansion.Conjunction
as part of that	Expansion.Instantiation.Arg2-as-instance
as such	Contingency.Cause+Belief.Result+Belief Contingency.Cause.Result
as well	Comparison.Similarity Expansion.Conjunction
at that point	Temporal.Synchronous
at that time	Temporal.Synchronous

at the same time	Temporal.Synchronous Expansion.Conjunction
at the time	Temporal.Synchronous
because of that	Contingency.Cause.Result
before that	Temporal.Asynchronous.Succession
but then again	Comparison.Concession.Arg2-as-denier
but then	Comparison.Concession.Arg2-as-denier
by comparison	Comparison.Contrast Comparison.Concession.Arg2-as-denier Expansion.Conjunction
by contrast	Comparison.Contrast Comparison.Concession.Arg2-as-denier
by doing so	Expansion.Manner.Arg1-as-manner
by the way	Comparison:Contrast Expansion.Conjunction
by then	Temporal.Asynchronous.Succession|Contingency.Cause.Reason Temporal.Asynchronous.Succession
despite this	Comparison.Concession.Arg2-as-denier
during that time	Temporal.Synchronous
even before then	Temporal.Asynchronous.Succession |Comparison.Concession.Arg2-as-denier
even before	Temporal.Asynchronous.Precedence|Comparison.Concession.Arg1-as-denier
even then	Temporal.Asynchronous.Precedence|Comparison.Concession.Arg2-as-denier
for example	Expansion.Instantiation.Arg2-as-instance Contingency.Cause.Reason Expansion.Level-of-detail.Arg2-as-detail
for instance	Expansion.Instantiation.Arg2-as-instance Expansion.Conjunction Expansion.Level-of-detail.Arg2-as-detail
for one thing	Expansion.Instantiation Contingency.Cause.Reason Expansion.Conjunction

	Expansion.Instantiation.Arg2-as-instance Expansion.Level-of-detail.Arg2-as-detail
for one	Expansion.Instantiation Expansion.Instantiation.Arg2-as-instance
for that purpose	Contingency.Purpose.Arg1-as-goal
for that reason	Contingency.Cause.Result
in addition	Expansion.Conjunction Expansion.Level-of-detail.Arg2-as-detail
in any case	Comparison.Concession.Arg2-as-denier
in any event	Expansion.Conjunction Expansion.Level-of-detail.Arg1-as-detail
in comparison	Comparison.Contrast
in contrast	Comparison.Contrast
in essence	Expansion.Conjunction
in fact	Comparison.Concession.Arg2-as-denier Comparison.Contrast Expansion.Conjunction Expansion.Instantiation.Arg2-as-instance Expansion.Level-of-detail.Arg1-as-detail Expansion.Level-of-detail.Arg2-as-detail Contingency.Cause+Belief.Reason+Belief Contingency.Cause+Belief.Result+Belief Contingency.Cause.Reason Contingency.Cause.Result Expansion.Equivalence
in general	Expansion.Level-of-detail.Arg1-as-detail
in more detail	Expansion.Level-of-detail.Arg2-as-detail
in other words	Expansion.Equivalence Comparison.Similarity Contingency.Cause.Reason Contingency.Cause.Result Expansion.Conjunction Expansion.Level-of-detail.Arg1-as-detail Expansion.Level-of-detail.Arg2-as-detail

in particular	Expansion.Instantiation.Arg2-as-instance Expansion.Level-of-detail.Arg2-as-detail Expansion.Conjunction
in response	Contingency.Cause.Result Expansion.Conjunction
in short	Expansion.Level-of-detail.Arg1-as-detail Contingency.Cause+SpeechAct.Result+SpeechAct Contingency.Cause.Reason Contingency.Cause.Result Expansion.Conjunction Expansion.Equivalence Expansion.Level-of-detail.Arg2-as-detail
in sum	Expansion.Level-of-detail.Arg1-as-detail Expansion.Conjunction Expansion.Equivalence Expansion.Level-of-detail.Arg2-as-detail
in the end	Comparison.Concession.Arg2-as-denier Comparison.Contrast Contingency.Cause.Result Expansion.Conjunction Expansion.Level-of-detail.Arg1-as-detail Expansion.Level-of-detail.Arg2-as-detail Temporal.Asynchronous.Precedence Expansion.Equivalence
in the meantime	Temporal.Asynchronous.Succession Temporal.Synchronous—Comparison.Contrast Temporal.Synchronous Temporal.Synchronous
in the meanwhile	Temporal.Synchronous
in this case	Expansion.Instantiation.Arg2-as-instance
in this way	Contingency.Cause.Result
in turn	Temporal.Asynchronous.Precedence Contingency.Cause.Result Expansion.Conjunction

	Expansion.Level-of-detail Temporal.Asynchronous
later on	Temporal.Asynchronous.Precedence
more accurately	Expansion.Substitution.Arg2-as-subst
more specifically	Expansion.Level-of-detail.Arg2-as-detail
more to the point	Expansion.Level-of-detail.Arg2-as-detail
no matter	Comparison.Concession.Arg1-as-denier
on the contrary	Comparison.Contrast Expansion.Level-of-detail.Arg2-as-detail
on the other hand	Comparison.Concession.Arg2-as-denier Comparison.Contrast
on the other	Comparison.Concession.Arg2-as-denier Comparison.Contrast
on the whole	Expansion.Conjunction Expansion.Level-of-detail.Arg1-as-detail Expansion.Level-of-detail.Arg2-as-detail
prior to this	Temporal.Asynchronous.Succession
quite the contrary	Expansion.Substitution
since then	Temporal.Asynchronous.Precedence
that is	Expansion.Equivalence Expansion.Level-of-detail.Arg2-as-detail Contingency.Cause.Reason Contingency.Cause.Result Expansion.Conjunction Expansion.Level-of-detail.Arg1-as-detail
to this end	Contingency.Cause.Result
what's more	Expansion.Conjunction

Appendix C

Test datasets

corpus	datasource	genre	connective	range	text	masked_text	sense	preposed_phrase	span	span_lower
PDTB3	wsj_0414	wsj	Thus	[64,68]	The supply of experienced civil engineers, though, is tighter. In recent months, California's Transportation Department has been recruiting in Pennsylvania, Arizona and Texas for engineers experienced in road and bridge design.	The supply of experienced civil engineers, though, is tighter. [SEP][MASK], in recent months, California's Transportation Department has been recruiting in Pennsylvania, Arizona and Texas for engineers experienced in road and bridge design.	Contingency.Cause.Result	In recent months	default	default
PDTB3	wsj_1629	wsj	By comparison	[97,110]	net income for the quarter was $5.9 million, or 71 cents a share, on revenue of $145.4 million. For the year-earlier period, the company reported a loss of $520,000 or six cents a share.	Net income for the quarter was $5.9 million, or 71 cents a share, on revenue of $145.4 million. [SEP][MASK], for the year-earlier period, the company reported a loss of $520,000 or six cents a share.	Comparison.Contrast	For the year-earlier period	default	default
DiscoGeM1.0	0013_Christianity	wikipedia	In addition	[145,156]	Christianity played a prominent role in the development of Western civilization, particularly in Europe from late antiquity and the Middle Ages. Following the Age of Discovery (15th–17th century), Christianity was spread into the Americas, Oceania, sub- Saharan Africa, and the rest of the world via missionary work.	Christianity played a prominent role in the development of Western civilization, particularly in Europe from late antiquity and the Middle Ages. [SEP][MASK], following the Age of Discovery (15th–17th century), Christianity was spread into the Americas, Oceania, sub- Saharan Africa, and the rest of the world via missionary work.	Expansion.Conjunction	Following the Age of Discovery (15th –17th century)	default	default
DiscoGeM1.0	Harry_Potter_and_the_Philospher_Stone_EN_paragraph_09	novel	Afterwards	[122,132]	She let Harry watch television and gave him a bit of chocolate cake that tasted as though she'd had it for several years. That evening, Dudley paraded around the living room for the family in his brand-new uniform.	She let Harry watch television and gave him a bit of chocolate cake that tasted as though she'd had it for several years. [SEP][MASK], that evening, Dudley paraded around the living room for the family in his brand-new uniform.	Temporal.Asynchronous.Precedence	That evening	default	default

Figure C.1: A small section of the preposed test data for illustration.

corpus	datasource	genre	connective	range	text	masked_text	sense	preposed_phrase	span	span_lower
PDTB3	wsj_1506	wsj	But	[194,197]	The guideline wasn't a law, but a joint interpretation of how the U.S. might operate during foreign coups in light of the longstanding presidential order banning a U.S. role in assassinations. In fact, yesterday the administration and Congress were still differing on what had been	The guideline wasn't a law, but a joint interpretation of how the U.S. might operate during foreign coups in light of the longstanding presidential order banning a U.S. role in assassinations. [SEP][MASK], yesterday the administration and Congress were still differing on what had been agreed to in fact.	Comparison.Concession.Arg2-as-denier	In fact	default	default
PDTB3	wsj_0776	wsj	While	[46,51]	About eight firms will get the lion's share. At the others, there are going to be a lot of disappointments, after all those promises and all that big money that's been paid to people	About eight firms will get the lion's share. [SEP][MASK], there are going to be a lot of disappointments, after all those promises and all that big money that's been paid to people at the others.	Expansion.Conjunction	At the others	default	default
DiscoGeM1.0	0027_Arctic Ocean	wikipedia	Consequently	[71,83]	In September 2012, the Arctic ice extent reached a new record minimum. Compared to the average extent (1979- 2000), the sea ice had diminished by 49%.	In September 2012, the Arctic ice extent reached a new record minimum. [SEP][MASK], the sea ice had diminished by 49% compared to the average extent (1979- 2000).	Comparison.Contrast	Compared to the average extent (1979- 2000)	default	default
DiscoGeM1.0	europarl-original-en-ep-00-03-17.txt	europarl	Considering the fact that	[38,63]	Mr President, I welcome this measure. From an administrative point of view the formalisation is a good thing.	Mr President, I welcome this measure. [SEP][MASK], the formalisation is a good thing from an administrative point of view.	Contingency.Cause.Reason	From an administrative point of view	default	default

Figure C.2: A small section of the canonical test data for illustration.