Is Partial Linguistic Information Sufficient for Discourse Connective Disambiguation? A Case Study of Concession

Takuma Sato[1,2], **Ai Kubota**[3], and **Koji Mineshima**[4]

[1]Nara Institute of Science And Technology
[2]RIKEN Guardian Robot Project, Kyoto, Japan
[3]The University of Tokyo
[4]Keio University

sato.takuma.sq6@naist.ac.jp aikubota@g.ecc.u-tokyo.ac.jp minesima@abelard.flet.keio.ac.jp

Abstract

Discourse relations are sometimes explicitly conveyed by specific connectives. However, some connectives can signal multiple discourse relations; in such cases, disambiguation is necessary to determine which relation is intended. This task is known as *discourse connective disambiguation* (Pitler and Nenkova, 2009), and particular attention is often given to connectives that can convey both CONCESSION and other relations (e.g., SYNCHRONOUS). In this study, we conducted experiments to analyze which linguistic features play an important role in the disambiguation of polysemous connectives in Japanese. A neural language model (BERT) was fine-tuned using inputs from which specific linguistic features (e.g., word order, specific lexicon, etc.) had been removed. We analyzed which linguistic features affect disambiguation by comparing the model's performance. Our results show that even after performing drastic removal, such as deleting one of the two arguments that constitute the discourse relation, the model's performance remained relatively robust. However, the removal of certain lexical items or words belonging to specific lexical categories significantly degraded disambiguation performance, highlighting their importance in identifying the intended discourse relation.

1 Introduction

Understanding natural language requires correct recognition of discourse relations among sentences (clauses), in addition to correctly understanding the propositional meaning within each sentence (clause). While there are many cases in which discourse relations are not linguistically marked, there are various discourse connectives that explicitly signal discourse relations such as *because*, *although*, and *therefore*. However, even with these connectives, it is not always a simple task to identify the discourse relation, due to the polysemous nature of connectives. For example, *while* in (1) indicates temporal relation, whereas *while* in (2) indicates contrastive relation.

(1) A package arrived while I was away.

(2) John loves to go outside, while Mary prefers to stay home.

In this study, we examine what factors affect the interpretation of polysemous discourse connectives. In particular, we focus on Japanese conjunctions "ながら" (nagara), "つつ" (tsutsu), and "ところで" (tokorode), all of which have both concessive and non-concessive uses.

(3) [$_{\text{Arg1}}$さびしいと思い]ながら [$_{\text{Arg2}}$それを口にしなかった]。 (CONCESSION)
'While [$_{\text{Arg1}}$feeling lonely], [$_{\text{Arg2}}$I did not voice it].'

(4) [$_{\text{Arg1}}$さびしいと思い]ながら [$_{\text{Arg2}}$毎日を過ごした]。 (SYNCHRONOUS)
'While [$_{\text{Arg1}}$feeling lonely], [$_{\text{Arg2}}$I spent every day].'

CONCESSION is a discourse relation that is often expressed with conjunctions such as *but*, *although* and *however*. In prior research, concessions have been considered to have the discourse function of *denial of expectations* (Izutsu, 2008; Kehler, 2002; Winterstein, 2012). Thus, in (3), what is expected is that one would say something if s/he is feeling lonely. Contrary to that expectation, however, the speaker did not do so. On the other hand, there is no such *denial of expectation* in (4).

The purpose of this study is to elucidate what factors are at play in the interpretation of concessions. For this purpose, we conducted experiments to fine-tune transformer-based language models (BERT) using the following types of input: original sentences, sentences with shuffled word order,

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop), pages 977–990
July 28-29, 2025 ©2025 Association for Computational Linguistics

sentences with either Arg1 or Arg2 removed, sentences with words belonging to specific categories removed, and sentences with the semantics of specific vocabulary removed.

Our contributions can be summarized as follows:

- We analyze the transformer-based model's (BERT) behavior using partial linguistic information as input, focusing on the discourse relation recognition task, which has gained little attention in this context.
- Specifically, we focus on the disambiguation of polysemous discourse connectives that can signal CONCESSION, formulating hypotheses based on linguistic research and testing them on an underexplored Japanese dataset.
- Our experiments show that BERT can still perform the task to some extent, even only with partial information.

2 Backgrounds

The difference in the roles of discourse expressions has been discussed as an important topic in semantics and pragmatics. For example, in examples such as (3) and (4), *while* ("ながら", *nagara*) is used as a discourse connective in both cases. However, in (4), the discourse connective merely indicates that Arg1 is an event simultaneous with Arg2, contributing only semantically to the proposition expressed by the entire sentence. In contrast, in (3), as discussed in the previous section, an inferential relation such as *denial of expectations* is encoded, and this connective plays a role in guiding the listener's inference toward the speaker's intended pragmatic interpretation. Building on this kind of distinction made by Blakemore (1987), Wilson and Sperber (1993) referred to the former as *conceptually encoded* and the latter as *procedurally encoded*. Such differences in the roles of discourse expressions continue to be actively discussed to this day (Iten, 2005).

When a single linguistic expression (discourse marker) has two significantly different uses such as these, what linguistic features are useful for disambiguation? This type of question—namely, the method of polysemous discourse disambiguation—has been actively discussed in the fields of theoretical linguistics and computational linguistics. For example, Pitler and Nenkova (2009) demonstrated that syntactic information is to some extent useful for such disambiguation, and Knaebel and Stede (2020) showed that using contextualized embeddings from BERT is effective. However, especially since the advent of neural networks, to the best of our knowledge, there has been no exploratory study that investigates which linguistic features (e.g., lexical semantics, specific POS and word order, etc.) are important by ablating various components. In studies of this kind, connectives that can express CONCESSION are often treated as representative examples (Zufferey and Degand, 2024). Our study, which conducts an analysis focusing on such discourse connectives in Japanese, is within the context of that line of inquiry.

Investigating which linguistic features are necessary for polysemous discourse disambiguation is important across various domains. For example, in psycholinguistics and theoretical linguistics, identifying the cues that can be used to distinguish such roles is useful for constructing cognitive models of language comprehension and production. In engineering fields such as natural language processing, clarifying the features that enable such distinctions can be beneficial for improving applications like translation and support for foreign language learning.

3 Experimental Setup

3.1 Task Definition

Our task is a multi-class classification task, aiming to determine the correct discourse relation label $L \in l_1, \ldots, l_n$ for a given sequence of input tokens $S = \{w_1, \ldots, w_d\}$. Here, w_i represents the i-th token in the sequence, d denotes the length of the token sequence, l_j $(1 \leq j \leq n)$ refers to the discourse relation label, and n indicates the number of all discourse relation labels in the dataset.

3.2 Dataset

The dataset used in this study is the Japanese discourse relation dataset introduced in Kubota et al. (2024). This dataset contains annotations of discourse relations for sentences connected by the connectives "ながら (*nagara*)," "つつ (*tsutsu*)," and "ところで (*tokorode*)". As Section 1 mentions, these connectives can indicate both concessive and non-concessive discourse relations. Therefore, merely observing discourse markers is insufficient to identify discourse relations in this dataset. The sentences in the dataset were extracted under specific syntactic conditions from the Kainoki Treebank (Kainoki, 2022).

There are five discourse relation labels in total: CONCESSION, SYNCHRONOUS, TIME, LOCATION, and OTHERS. See Kubota et al. (2024) for details on each label. The discourse relations are not necessarily mutually exclusive, and there are cases that can be interpreted as involving multiple discourse relations simultaneously[1]. As examples from Japanese, Muraki (2019) and Kubota et al. (2024) point out that the use of "ながら (*nagara*)" can sometimes appear to simultaneously instantiate both SYNCHRONOUS and CONCESSION relations. Kubota et al. (2024) assigned the label CONCESSION to all sentences in which the meaning of CONCESSION was identified, without allowing co-labeling with SYNCRONOUS. We followed this approach as well. This means that sentences labeled CONCESSION may include instances that could also be interpreted as SYNCHRONOUS, but were not assigned that label. The dataset was split into training, validation, and test sets in an 8:1:1 ratio. Table 1 and 2 shows the statistics.

3.3 Experimental settings

We conducted perturbation experiments to investigate how partial linguistic information, such as word order and specific lexical items, affects model performance in our discourse connective disambiguation task. We fine-tuned the Japanese BERT model[2] using the different manipulation settings below (see also Table 3) to observe the performance under each constraint in the task. The detailed settings for training and related configurations are provided in Appendix (A.1). The following paragraphs show the motivation or hypotheses for each experimental setting.

Original sentence (baseline) Complete sentences are the inputs to the model in this setting. This setting is the same as the standard fine-tuning of BERT. This setup measures BERT's performance on our discourse connective disambiguation task as a baseline without any constraints, serving as the baseline for comparison with the constraints in the following settings.

Word-order ablation In this setting, the input consists of the lemmas of all words in the sentence, shuffled randomly. Shuffling is performed across the entire sentence, beyond the scope of each individual argument. This setup is designed to verify whether the model can accurately disambiguate discourse connectives using only lexical information without the word order of the sentence.

Argument ablation In these settings, we ablated the part before the discourse connective (Arg1) or the part after it (Arg2) from the input text. This setup consists of two sub-settings: Arg1-ablation and Arg2-ablation. Since these settings are equivalent to removing one of two arguments that define discourse *relation*, we expected a significant performance drop from the baseline. Note that in these setups, discourse markers (connectives that signal discourse relations), such as "も (*mo*)" and "ながら (*nagara*)", are also ablated.

Lexical ablation We ablated words classified into specific parts of speech, categories, and functions in these settings. This setting consists of the following five sub-settings: Connective ablation, Function-words ablation, Content-words ablation, *Mo* ablation, and Negation ablation.

Connective ablation is a setting in which we ablate discourse connectives (e.g., "つつ (*tsutsu*)," "ながら (*nagara*)," "ところで (*tokorode*)") from the sentences. This setting transforms our discourse relation recognition (DRR) task from Explicit DRR (EDRR) to Implicit DRR (IDRR). Since IDRR is more challenging than EDRR (Cai et al., 2024), we expected a performance drop from the baseline under this setting.

The Content-words/function-words ablation settings ablate all content words or function words from a sentence, respectively. We defined content-words as noun, verbs, adjectives, and adverbs, and function-words as all words other than content-words[3]. We designed these settings based on previous research that identifies "semantic opposition" between Arg1 and Arg2 as one type of concessive discourse relation, which arises from the presence of antonymous lexical items (Lakoff, 1971; Izutsu, 2008). Since many antonymous lexical items (e.g., tall vs. short) are often content words, the hypothesis underlying this setting is that ablating content words will lead to a more significant performance drop in recognizing concessive relations

[1] We would like to thank the anonymous reviewer who pointed out this issue and provided helpful suggestions.

[2] As the Japanese BERT model, we used tohoku-nlp/bert-base-japanese-v3 (111M parameters), available on Hugging Face (https://huggingface.co/tohoku-nlp/bert-base-japanese-v3).

[3] In this study, we used MeCab (https://taku910.github.io/mecab/) (Kudo et al., 2004) as the morphological analyzer in the BERT tokenizer and UniDic (https://clrd.ninjal.ac.jp/unidic/) (Den et al., 2008) as the dictionary.

Table 1: Data split statistics. We split the entire dataset into train, test, and validation sets in a ratio of 8:1:1. The data we used is label-imbalanced, with relatively few instances of labels other than SYNCHRONOUS.

	SYNCHRONOUS	CONCESSION	TIME	LOCATION	OTHERS	**total**
Train	1002	218	8	42	65	1336
Valid	120	32	4	3	8	167
Test	111	41	2	4	10	168
Total	**1233**	**291**	**14**	**49**	**83**	**1670**

Table 2: Data statistics for each connective. All three are polysemous connectives that can convey CONCESSION; however, the discourse relations they signal other than CONCESSION differ for each.

Connective	**Discourse Relation**	**Counts**
nagara	CONCESSION	213
	SYNCHRONOUS	1,047
	OTHERS	65
tsutsu	CONCESSION	51
	SYNCHRONOUS	186
tokorode	CONCESSION	27
	TIME	14
	LOCATION	49
	OTHERS	18

than ablating function words.

The *Mo* ablation setting removes the particle "も (*mo*)" when it is attached to "ながら (*nagara*)" or "つつ (*tsutsu*)". In the Japanese language, when the "も (*mo*)" particle follows "ながら (*nagara*)" or "つつ (*tsutsu*)," the discourse relation can always be classified as Concession (Kubota et al., 2024). Based on this, "も (*mo*)" in this context is considered an important local lexical cue for recognizing CONCESSION. We conducted the experiment in this setting under the hypothesis that ablating this "も (*mo*)" would decrease performance.

The negation ablation setting removes various negation expressions in Japanese from sentences. The target expressions for removal include "ない (nai)," "なし (*nashi*)," "非 (*hi*)," "不 (*hu*)," "無 (*mu*)," "未 (*mi*)," "反 (*han*)," and "異 (*i*)." Corpus linguistics research has confirmed that negation appears with statistically significant frequency in concessive sentences (Torabi Asr and Demberg, 2015; Crible, 2021). From this observation, we hypothesized that ablating negation as a local lexical cue will decrease performance scores. This setting is intended to test this hypothesis.

Semantic ablation In these settings, we replaced words classified into specific POS with nonsensical imaginary words. This setting consists of three sub-settings: Content-words semantic ablation, Function-words semantic ablation, and All-words semantic ablation. Table 5 in the appendix shows the correspondence between each word's POS and its substitute imaginary words. We implemented these settings to ablate the target words' lexical semantics while holding the sentences' syntactic structure to a certain extent. This experiment was conducted under the expectation that subword segmentation in BERT's tokenizer captures the morphological characteristics of each part of speech (POS) in Japanese (e.g., adjectives typically end with "い"), and that even for non-existent words, certain POS and syntactic information would be preserved to some extent depending on the surrounding context.

Content/function-words semantic ablation are settings where all content/function words in a sentence are replaced with nonsense words. The paragraph on Lexical ablation provides the definitions of content and function words. All-words semantic ablation is a setting where we replace all words in a sentence with nonsense words.

4 Results and Analyses

4.1 Results

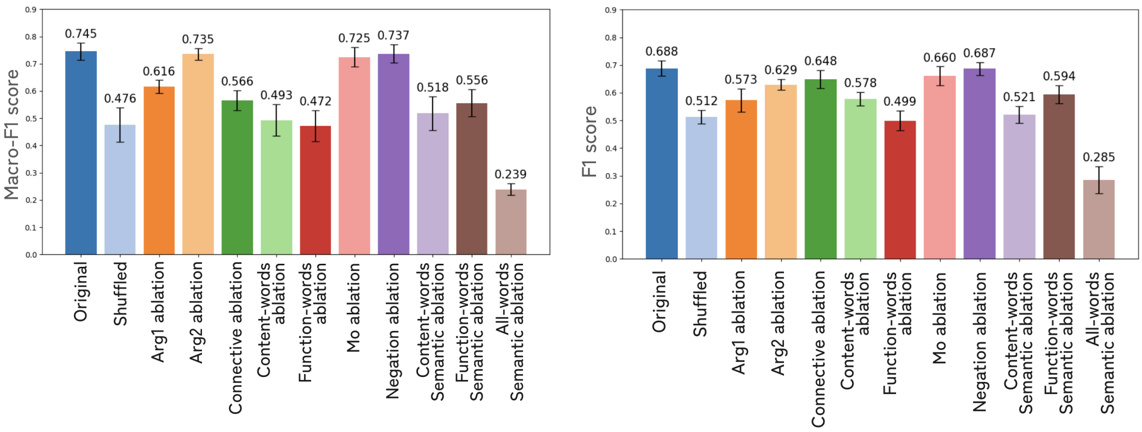
The results of fine-tuning BERT under each experimental setting are shown in Figure 1. Inference on the test set was performed 10 times for each setting using the fine-tuned BERT model, and we report the mean F1 Score along with the 95% confidence interval. Also, one of this study's research questions was whether the model can disambiguate discourse connectives using only partial linguistic information [4]. To answer this, figure 1b presents the F1 score for CONCESSION label of the fine-tuned BERT model after fine-tuning.

Note that the number of manipulated words significantly varies across experimental settings

[4] Additionally, we show macro F1 scores per connectives in Table 7 in Appendix.

Table 3: Examples of manipulations in experimental settings. In each experimental setting, words with ~~strikethrough~~ were deleted, while words highlighted in magenta were replaced with nonsense words.

Category	Type	Example
Original	**Original**	[Arg1 さびしいと思い] ながら[Arg2 も、それを口にしなかった] (While [Arg1 I felt lonely], [Arg2 I did not say it].)
Word-order ablation	—	た ない ながら に。 それ する を も 口 さびしい 思う と、 (not did while . it , say I lonely felt I)
Argument ablation	**Arg1-ablation**	[Arg1 ~~さびしいと思い~~] ~~ながら~~ [Arg2 も、それを口にしなかった。] (~~While~~ [Arg1 ~~I felt lonely~~], [Arg2, I did not say it.])
	Arg2-ablation	[Arg1 さびしいと思い] ~~ながら~~ [Arg2 ~~も、それを口にしなかった~~] (~~While~~ [Arg1 I felt lonely], [Arg2 ~~, I did not say it.~~])
Lexical ablation	**Connective ablation**	[Arg1 さびしいと思い] ~~ながら~~ [Arg2 も、それを口にしなかった。] (~~While~~ [Arg1 I felt lonely,] [Arg2 I did not say it.])
	Content-words ablation	[Arg1 ~~さびしい~~と~~思い~~]ながら[Arg2 も、~~それ~~を~~口~~に~~し~~なかった。] (While [Arg1 ~~I felt lonely~~] [Arg2, ~~I~~ did not ~~say~~ it.])
	Function-words ablation	[Arg1 さびしい~~と~~思い] ~~ながら~~ [Arg2 ~~も、~~それ~~を~~口~~に~~し~~なかった。~~] (~~While~~ [Arg1 I felt lonely][Arg2, I ~~did not~~ say ~~it.~~])
	***Mo* ablation**	[Arg1 さびしいと思い]ながら[Arg2 ~~も、~~ それを口にしなかった] (While [Arg1 I felt lonely], [Arg2 I did not say it].)
	Negation ablation	[Arg1 さびしいと思い]ながら[Arg2 、それを口にし~~なかっ~~た] (While [Arg1 I felt lonely], [Arg2 I did ~~not~~ say it].)
Semantic ablation	**Content-words semantic ablation**	[Arg1 もさらいとたゆねる]ながら[Arg2 も、彼女をミョガパスにたゆねるなかった。]
	Function-words semantic ablation	[Arg1 さびしいがが思い]でありく[Arg2 が。彼女が口がししだだ。]
	All-words semantic ablation	[Arg1 もさらいがたゆねるが]でありく[Arg2 が。彼女がミョガパスがたゆねるだだ。]



(a) Macro-F1 score for all labels.

(b) F1 score for CONCESSION label.

Figure 1: F1-scores on the test set after fine-tuning BERT on each input format. Each bar represents the mean score on the test set across 10 fine-tuning iterations, and the error bars indicate the 95% confidence interval.

(see Table 6 in Appendix for the exact count). To account for this variation in analysis, we computed the performance (F1 score for CONCESSION) drop per manipulated word. The results are presented in Figure 2 as a bar graph, with the y-axis set to a logarithmic scale. For each experimental setting $e \in E$ (where E is the set of all experimental settings), let s_e denote the CONCESSION-only F1 score for that setting and c_e denote the number of manipulated words in that setting. We then calculated the performance drop per manipulated word as $\frac{s_{original} - s_e}{c_e}$ where $s_{original}$ is the score of the original (baseline) setting.

4.2 Interpreting results for each setting

Original sentence (baseline) Firstly, an examination of the scores achieved by the baseline model reveals that the BERT model can disambiguate discourse connectives when the inputs are complete sentences. This model exhibits significantly higher scores than the chance rates for both all discourse relation labels (0.2354) and the CONCESSION label alone (0.3077). Kubota et al. (2024) reported that the kappa-values for the annotation were 0.72, 0.46, and 0.75 for "ながら (*nagara*)," "つつ (*tsutsu*)," and "ところで (*tokorode*)", respectively. This indicates that the task is inherently complicated, often with no definitive answer. Given this difficulty, the

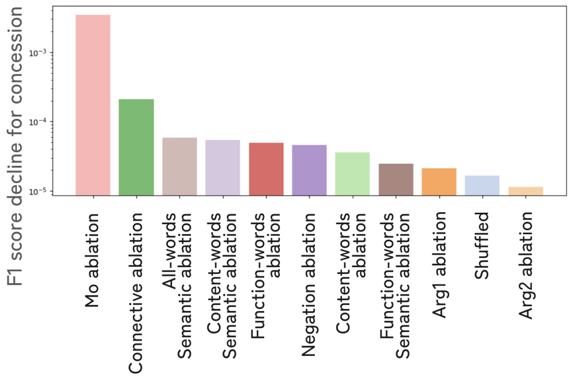


Figure 2: The performance degradation per manipulated word in each experimental setting. It means the decrease in F1 score for the CONCESSION label from the baseline, divided by the number of words manipulated in each setting. The Y-axis is on a logarithmic scale.

BERT model can be said to be able to solve it when given original sentences as inputs.

Word-order ablation In this setting, a relatively large performance drop was observed compared to the baseline; however, the decline was not catastrophic enough to reach the chance rate. This suggests that even when syntactic and word order information is removed and the disambiguation task is performed solely based on the lexical information, a certain level of performance can still be achieved. Additionally, when comparing the scores across all labels with those specific to CONCESSION, the latter exhibited a smaller decline in performance. The performance degradation per manipulated word for the CONCESSION label is also relatively small. This suggests that even when the syntactic structure is disrupted, the model can still make somewhat correct judgments by using lexical semantics as a cue.

Argument ablation In this setting, we observed a performance drop from the baseline, but the extent of the decline was relatively small. Additionally, the ablation of Arg1 had a more negative impact on performance than the ablation of Arg2. The performance degradations per manipulated word were also relatively small for both Arg1 and Arg2. This result suggests that even when one of the two arguments constituting discourse relations is removed, BERT can still perform the discourse connective disambiguation task to a certain extent. Given that discourse *relations* are defined between two textual arguments (Arg1 and Arg2), it may be counter-intuitive that the model can perform well in our disambiguation task even when one of the two elements that define the relation is excluded. However, there may be linguistic clues left in either Arg1 or Arg2. For example, it has been reported that the discourse relation tends to be CONCESSION if the predicate of Arg1 has a stative predicate or a verb of thought or perception such as “思う” (*to think*) (Muraki, 2019; Japanese Descriptive Grammar Research Group, 2008). Of course, this is only a trend and not a decisive factor in determining discourse relations. Nevertheless, it should be noted that such linguistic clues are very likely to influence interpretation.

Lexical ablation First, in the Connective ablation setting, moderate performance declines from the baseline were observed. This result indicates that transforming an Explicit Discourse Relation Recognition (EDRR) task into an Implicit Discourse Relation Recognition (IDRR) task increases its difficulty even for polysemous connectives. Focusing on the CONCESSION label, the drop was relatively small. This is a natural outcome, considering that all the connectives targeted in our experiment can serve as markers for CONCESSION. The performance degradation per manipulated word was the second largest, suggesting that the type of connective functions as a local lexical cue for the model’s recognition of CONCESSION.

Next, in the Content/function-words ablation setting, ablating function words caused a greater performance drop than ablating content words. We consider this to be an interesting result as it contradicts our initial experimental hypothesis. A similar trend was observed in the performance degradation per manipulated word, indicating that the omission of function words has a more significant negative impact on the model’s judgment than the omission of content words.

Next, a performance drop was observed in the *Mo* ablation setting, although its extent was relatively small. However, it is important to note that this setting manipulates only a tiny number of words. Consequently, the performance drop per manipulated word was the largest among all experimental settings. Therefore, our experimental hypothesis—that “も (*mo*) ” (when attached to discourse markers) serves as an important local lexical cue for recognizing CONCESSION—is primarily supported by the results.

In the negation ablation setting, the performance drop was minimal, and the performance drop per manipulated word was also not substantial. This result contradicts our hypothesis, based on previous research, that negation functions as an important local lexical cue for identifying CONCESSION.

Semantic ablation First, in the content/function-words semantic ablation experiment, a certain degree of performance degradation was observed for both content and function words compared to the baseline. When comparing this with the Content/function-words ablation experiment, the performance degradation for content words was smaller in the semantic ablation settings when considering scores for all labels. However, when focusing only on the CONCESSION label, the degradation was smaller in the lexical ablation settings. For function words, the semantic ablation settings exhibited a smaller degradation across both scoring metrics. We observed a similar trend when analyzing the degree of performance degradation per manipulated word. Since we designed these experiments to eliminate lexical semantics while preserving the syntactic structure of sentences as much as possible, we expected the performance degradation to be smaller than experiments within the lexical ablation settings. The results for both function and content words in the all-label score align with this expectation, suggesting that BERT utilizes syntactic structure to some extent for discourse relation recognition, even in the absence of lexical semantics. However, the fact that an unexpected result emerged in the CONCESSION-only score for content words is particularly intriguing.

Next, in the All-words semantic ablation setting, the model achieved scores that were either close to or even lower than the chance rate for both all-label scores and the CONCESSION-only scores. This result suggests that the model is unlikely to effectively utilize the minimal remaining syntactic (part-of-speech) information in the sentences. However, since this operation does not necessarily guarantee a complete extraction of syntactic information, a more refined experimental design would be required to draw a definitive conclusion.

4.3 Error Analysis

We conduct an error analysis on several characteristic cases to gain a concrete understanding of the model's judgment. Table 4 shows the correctness of the model's outputs under each experimental setting for the three cases below.

The first case is an example where the model appears to classify CONCESSION by using "も (*mo*)" as a local lexical cue.

Table 4: The correctness of the model's outputs for each experimental setting under each selected instance. ✓ indicates that the model's classification was correct, while × indicates that the classification was incorrect.

	(5)	(6)	(7)
Original	✓	✓	✓
Shuffled	✓	✓	✓
Arg1 ablation	✓	✓	✓
Arg2 ablation	✓	✓	✓
Connective ablation	✓	×	✓
Content-words ablation	✓	✓	×
Function-words ablation	×	✓	✓
Mo ablation	×	✓	✓
Negation ablation	✓	×	✓
Content-words semantic ablation	✓	✓	×
Function-words semantic ablation	✓	✓	✓
All-words semantic ablation	×	✓	×

(5) [$_{\text{Arg1}}$気がつくと、がれきに囲まれ]ながら[$_{\text{Arg2}}$も息ができる状態でした。] (CONCESSION)
I found myself able to breathe while being surrounded by rubble.

In this example, even when "も (*mo*) " is removed, the model should still be able to correctly recognize CONCESSION if it understands the semantic content of the sentence.[5] However, the model fails to make the correct classification when "も (*mo*)" is excluded from the input.

The second case is an example where the model fails to correctly classify CONCESSION under the negation ablation setting.

(6) [$_{\text{Arg1}}$この問題をいまさら議論した]ところで[$_{\text{Arg2}}$無意味でしょう。] (CONCESSION)
Even if we discuss this issue at this point, it would not be meaningful.

In this setting, the character "無 (*mu*)" in "無意味 (*muimi*: meaningless)" in Arg2 was excluded. When this character is removed, the denial of expectation—where the expectation could be like "engaging in a discussion is usually meaningful"—no longer holds. We are inferring that the model failed in classification due to this factor.

In the third example, from a lexical semantics perspective, the polarity shift between the positive connotation of "学がある (being knowledgeable)" and the negative connotation of "翻弄される

[5]It is somewhat acceptable to interpret this case as a denial of an expectation, such as "If one were surrounded by rubble, they would normally be unable to breathe." Moreover, interpreting it as SYNCHRONOUS would not be natural.

(been tossed around)" serves as a key clue for identifying CONCESSION.

(7) [Arg1 学があり]ながら[Arg2 運命の手に翻弄されてきた男、という印象を全体から感じる。] (CONCESSION)
The overall impression is of a man who, despite being knowledgeable, has been tossed around by the hands of fate.

We assume that the intervention on content words likely resulted in the loss of this information, leading to the model's misclassification.

5 Discussion and Future Direction

5.1 What does BERT need to recognize CONCESSION?

Previous studies have pointed out that antonymous lexical items and negation are important in the identification of CONCESSION concerning *denial of expectation* (Lakoff, 1971; Izutsu, 2008; Crible, 2021). While this partially aligns with our findings, our experiments on Lexical ablation and Semantic ablation suggest that complete disambiguation is not necessarily impossible without these elements. Furthermore, from the perspective of *denial of expectation*, it may seem possible to hypothesize that the removal of Arg1/Arg2 would have a fatal impact. However, our results do not support such a conclusion, and it is possible that statistical machine learning models like BERT can distinguish CONCESSION to some extent using only surface-level information.

Additionally, previous studies have reported that word order and lexical semantics are often redundant (Papadimitriou et al., 2022; Sinha et al., 2021a; Clouatre et al., 2022), but our results do not lead to such a conclusion. In our experiments, the loss of either one resulted in a certain degree of performance degradation. However, a previous study also reported that linguistic information's importance varies depending on the task (Zhao et al., 2024). Therefore, determining to what extent we generalize our experimental results to tasks beyond the recognition of CONCESSION requires further research.

5.2 Do BERT and humans make similar inferences?

We suspect that humans wouldn't be able to achieve the identical scores as BERT when relying on only partial information. For instance, even considering just the examples in Table 3, it seems unlikely that humans could correctly recognize CONCESSION (without mere guesswork) in sentences like those found in Arg1-ablation, Content-words ablation, or Content-words semantic ablation. This suggests that transformer-based language models like BERT may be handling our discourse connective disambiguation task in a way that differs from human processing. However, this remains a hypothesis, and drawing a definitive conclusion would require conducting experiments in which humans attempt the same task as our study.

5.3 Shift of Discourse Relation during Ablation

In some cases, performing ablation can cause the ground-truth discourse relation to change[6]. For example, considering the removal of "も (*mo*)" in (5), it may no longer be the case that only CONCESSION is the correct discourse relation—judging it as SYNCHRONOUS may not necessarily be incorrect either. Liu et al. (2024) point out that in discourse relation recognition, such a shift in discourse relation can occur when connectives are removed, and this is a possible reason why models trained on Explicit Discourse Relation Recognition tasks fail in Implicit tasks.

Such cases are likely included in our data and experiments to some extent, but we predict that their number is small. By conducting future analyses using explainability methods other than ablation (e.g., Integrated Gradients (Sundararajan et al., 2017), LIME (Ribeiro et al., 2016), SHAP (Lundberg and Lee, 2017), etc.), it may be possible to compensate for this weakness in our experimental methodology.

6 Related Works

6.1 Discourse Relation Recognition

Discourse relation recognition (DRR) is an NLP task that aims to determine the semantic relation between two textual arguments (Xiang and Wang, 2022; Kishimoto et al., 2020). The Penn Discourse Treebank (PDTB) is widely used as a dataset annotated with discourse relations (Prasad et al., 2008).

In PDTB, Prasad et al. (2008) categorized discourse relations as explicit or implicit. When a connective conveys a relation, it is Explicit

[6] We would like to thank the anonymous reviewer who pointed this out.

Discourse Relation Recognition (EDRR); otherwise, it is Implicit Discourse Relation Recognition (IDRR) (Wang, Chenxu and Jian, Ping and Wang, Hai, 2023). Among these two, IDRR (Implicit Discourse Relation Recognition) has attracted attention because it is expected to be widely applicable to downstream tasks in NLP, such as text generation and summarization (Wang, Chenxu and Jian, Ping and Wang, Hai, 2023), yet remains challenging even with transformer-based pre-trained models (Cai et al., 2024).

6.2 Partial Linguistic Information for NLU

Various studies have analyzed the importance (or lack thereof) of different types of information in NLU tasks by observing model performance under different manipulations and ablations applied to the original input. One particularly notable type of partial information is word order. Papadimitriou et al. (2022); Sinha et al. (2021a); Clouatre et al. (2022) argue that word order is often redundant with lexical information, and knowing the set of words in a sentence is often sufficient for NLU tasks. Their findings show that fine-tuning models on shuffled word order does not significantly degrade performance.

Research on partial information in model judgments has been active in the Natural Language Inference (NLI) task, which judges whether a *premise* entails, contradicts, or is neutral to a *hypothesis*. Many NLI datasets contain annotation artifacts, allowing models to perform well without truly learning sentence relationships (Poliak et al., 2018; Gururangan et al., 2018; Tsuchiya, 2018). Studies also show Transformer models achieve high accuracy on permuted NLI examples, which means they are insensitive to word order (Sinha et al., 2021b; Gupta et al., 2021). Conversely, Ettinger (2020) noted BERT's performance degrades for some, but not all, word order perturbations.

In NLI, high accuracy with shuffled or partial input often indicates model or dataset biases, highlighting limitations in generalization. In contrast, in DRR and disambiguation, local lexical clues can serve as genuine linguistic signals. Compared to NLI, fewer studies have explored partial or shuffled input in DRR. Some works (Sileo et al., 2019; Kim et al., 2020) show that simple lexical cues can often detect discourse relations, even implicit ones, without syntactic or semantic analysis. In particular, Sileo et al. (2019) explores how discourse markers can enhance sentence representation learning in an unsupervised manner. They extract sentence pairs with discourse markers from large corpora, using them as positive examples to create datasets for capturing semantic relationships without labeled data. Both studies demonstrated that simple lexical features, such as individual words or phrases, can often suffice to detect discourse relations, extracting significant information about discourse structure without syntactic or semantic analysis.

Our study aims to contribute further to this line of work by focusing on a specific linguistic phenomenon and a non-English language and investigating how well partial linguistic information can help disambiguate discourse connectives.

7 Conclusion

In this study, we demonstrated that BERT can perform discourse connective disambiguation with a certain level of accuracy using only partial linguistic information in complex discourse relations. Specifically, we focus on Japanese polysemous connectives that are sometimes but not always interpreted as CONCESSION. We fine-tuned BERT using inputs in which word order, arguments, specific words, or their lexical semantics were ablated from the original sentences and observed the model's performance. By calculating the performance drop per manipulated word for each experiment, we analyzed which linguistic elements significantly impact the model's performance in this task. The results showed that the model mainly exhibited a certain level of performance in complex discourse connective disambiguation even without observing complete sentences, relying only on partial information. We hope this study contributes to advancing empirical approaches from NLP and computational linguistics toward understanding language and the nature of linguistic phenomena.

Limitations

Since this study is linguistically motivated and aims to provide a detailed analysis and insights into specific linguistic phenomena, the size of the dataset used in the experiments is limited. As described in Sec. 2, the experiments and analyses in this study focused on discourse connectives capable of conveying CONCESSION; however, by conducting similar evaluations over a broader range of discourse relations, new findings can be expected. Additionally, we used BERT as a representative transformer-based model, but conducting experiments with decoder-only models such as GPT would also be beneficial for further extended investigations. In our experimental methodology, if encoder-only models and decoder-only models exhibit different behaviors, exploring those differences would also be beneficial from a model-analysis perspective. To ascertain whether the implications of this study can be generalized, it would be beneficial to conduct broader experimentation.

Not only expanding the experiments, but also employing different analytical methods would be effective. This time, we examined the importance of various linguistic features by applying perturbations to the model inputs; however, employing representative analytical techniques in machine learning, such as LIME (Ribeiro et al., 2016) and SHAP (Lundberg and Lee, 2017), also represents a promising direction for enhancing the robustness of our analysis.

Besides, this study is conducted with a corpus in the Japanese language. As mentioned above, it is a promising direction for future research to verify whether the findings of this study are applicable to other languages.

Ethical Statement

Our research does not involve manual experiments and is unlikely to lead to harmful applications. However, we must exercise utmost caution, as our findings may be overly generalized to less widely spoken languages, which could foster indifference toward those languages and cultures, further disadvantaging them.

Acknowledgement

We thank the anonymous reviewers for their helpful comments and suggestions. This work is supported by JST, CREST Grant Number JPMJCR2114.

In this study, AI assistants, including ChatGPT, Copilot, and DeepL, were used in accordance with the ACL Policy on AI Writing Assistance. We primarily used them to assist with coding and writing, but all code and text outputs were manually reviewed. The authors take full responsibility for all of them.

Training Details The fine-tuning performed in this study took approximately two days, and it was executed using a single GPU with around 16GB of memory.

References

Diane Blakemore. 1987. *Semantic Constraints on Relevance*. Blackwell, Oxford.

Mingyang Cai, Zhen Yang, and Ping Jian. 2024. Improving implicit discourse relation recognition with semantics confrontation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8828–8839, Torino, Italia. ELRA and ICCL.

Louis Clouatre, Prasanna Parthasarathi, Amal Zouaq, and Sarath Chandar. 2022. Local structure matters most: Perturbation study in NLU. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3712–3731, Dublin, Ireland. Association for Computational Linguistics.

Ludivine Crible. 2021. Negation cancels discourse-level processing differences: Evidence from reading times in concession and result relations. *Journal of Psycholinguistic Research*, 50(6):1283–1308.

Yasuharu Den, Junpei Nakamura, Toshinobu Ogiso, and Hideki Ogura. 2008. A proper approach to japanese morphological analysis: Dictionary, model, and evaluation. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*. European Language Resources Association (ELRA).

Allyson Ettinger. 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.

Ashim Gupta, Giorgi Kvernadze, and Vivek Srikumar. 2021. BERT & family eat word salad: Experiments with text understanding. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(14):12946–12954.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Corinne Iten. 2005. *Linguistic Meaning, Truth Conditions and Relevance: The Case of Concessives*. Palgrave Macmillan.

Mitsuko Narita Izutsu. 2008. Contrast, concessive, and corrective: Toward a comprehensive study of opposition relations. *Journal of Pragmatics*, 40(4):646–675.

Japanese Descriptive Grammar Research Group. 2008. *Gendai Nihongo Bunpo 6 (Modern Japanese Grammar 6)*. Kuroshio Publishing.

Ed Kainoki. 2022. The Kainoki treebank – a parsed corpus of contemporary Japanese. https://kainoki.github.io. Accessed: 2024-04-01.

Andrew Kehler. 2002. Coherence, reference, and the theory of grammar. *CSLI Publications*.

Najoung Kim, Song Feng, Chulaka Gunasekara, and Luis Lastras. 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, Online. Association for Computational Linguistics.

Yudai Kishimoto, Yugo Murawaki, and Sadao Kurohashi. 2020. Adapting BERT to implicit discourse relation classification with a focus on discourse connectives. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1152–1158, Marseille, France. European Language Resources Association.

René Knaebel and Manfred Stede. 2020. Contextualized embeddings for connective disambiguation in shallow discourse parsing. In *Proceedings of the First Workshop on Computational Approaches to Discourse*, pages 65–75, Online. Association for Computational Linguistics.

Ai Kubota, Takuma Sato, Takayuki Amamoto, Ryota Akiyoshi, and Koji Mineshima. 2024. Annotation of Japanese discourse relations focusing on concessive inferences. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1215–1224, Torino, Italia. ELRA and ICCL.

Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. 2004. Applying conditional random fields to japanese morphological analysis. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 230–237.

Robin Lakoff. 1971. If's, and's and but's about conjunction. In Charles J. Fillmore and D. Terence Langendoen, editors, *Studies in linguistic semantics*, pages 3–114. Irvington.

Wei Liu, Stephen Wan, and Michael Strube. 2024. What causes the failure of explicit to implicit discourse relation recognition? In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2738–2753, Mexico City, Mexico. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization.

Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 4768–4777, Red Hook, NY, USA. Curran Associates Inc.

Shinjiro Muraki. 2019. The various usages of "nagara" (in Japanese). In *Lexicology and Grammar*, volume 156 of *Hitsuji Kenkyū Sōsho (Linguistics Series)*, chapter 5, Part 1. Hitsuji Shobō.

Isabel Papadimitriou, Richard Futrell, and Kyle Mahowald. 2022. When classifying grammatical role, BERT doesn't care about word order... except when it matters. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 636–643, Dublin, Ireland. Association for Computational Linguistics.

Emily Pitler and Ani Nenkova. 2009. Using Syntax to Disambiguate Explicit Discourse Connectives in Text. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 13–16, Suntec, Singapore. Association for Computational Linguistics.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191, New Orleans, Louisiana. Association for Computational Linguistics.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn Discourse TreeBank 2.0.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1135–1144, New York, NY, USA. Association for Computing Machinery.

Damien Sileo, Tim Van De Cruys, Camille Pradel, and Philippe Muller. 2019. Mining discourse markers for unsupervised sentence representation learning. 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 3477–3486, Minneapolis, Minnesota. Association for Computational Linguistics.

Koustuv Sinha, Robin Jia, Dieuwke Hupkes, Joelle Pineau, Adina Williams, and Douwe Kiela. 2021a. Masked language modeling and the distributional hypothesis: Order word matters pre-training for little. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2888–2913, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Koustuv Sinha, Prasanna Parthasarathi, Joelle Pineau, and Adina Williams. 2021b. UnNatural Language Inference. 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 7329–7346, Online. Association for Computational Linguistics.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 3319–3328. JMLR.org.

Fatemeh Torabi Asr and Vera Demberg. 2015. Uniform surprisal at the level of discourse relations: Negation markers and discourse connective omission. In *Proceedings of the 11th International Conference on Computational Semantics*, pages 118–128, London, UK. Association for Computational Linguistics.

Masatoshi Tsuchiya. 2018. Performance impact caused by hidden bias of training data for recognizing textual entailment. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Wang, Chenxu and Jian, Ping and Wang, Hai. 2023. Numerical semantic modeling for implicit discourse relation recognition. *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Deirdre Wilson and Dan Sperber. 1993. Linguistic form and relevance. *Lingua*, 90:1–25.

Grégoire Winterstein. 2012. What but-sentences argue for: An argumentative analysis of but. *Lingua*, 122(15):1864–1885.

Wei Xiang and Bang Wang. 2022. A survey of implicit discourse relation recognition. *ACM Comput. Surv.*

Qinghua Zhao, Jiaang Li, Lei Li, Zenghui Zhou, and Junfeng Liu. 2024. Word order's impacts: Insights from reordering and generation analysis.

Sandrine Zufferey and Liesbeth Degand. 2024. *Connectives and Discourse Relations*. Key Topics in Semantics and Pragmatics. Cambridge University Press.

A Appendix

A.1 Configurations of Training

In fine-tuning, we used AdamW (Loshchilov and Hutter, 2019) as the optimizer and the scheduler created by `get_linear_schedule_with_warmup` from the Hugging Face `Transformers` library[7], which are the default settings of the `Trainer` class. For training, we used an early stopping setting where training was terminated if no increase in the F1-score on the validation set was observed for three consecutive epochs. The maximum number of epochs was set to 30.

A.2 Detailed Experimental Settings, Statistics, and Results

Table 5: The substitute imaginary words for each POS in lexical replacement. For pronouns, prenoun-adjectival, and other POS that belong to highly limited grammatical categories, actual existing words are used.

Part of Speech	Substitute Word
Noun	ミョガパス
Pronoun	彼女
Adjectival-noun	さもらか
Prenoun-adjectival	この
Adverb	もさらく
Conjunction	でありく
Interjection	わあ
Verb	たゆねる
Adjective	もさらい
Auxiliary-verb	だ
Particle	が
Prefix	ふら
Suffix	ぼね
Auxiliary-symbol	-

Table 6: The number of manipulated words in each experimental setting.

Experimental setting	Count
Shuffled	6,931
Arg1 ablation	3,408
Arg2 ablation	3,548
Connective ablation	179
Content-words ablation	3,070
Function-words ablation	3,861
Mo ablation	8
Negation ablation	35
Content-words semantic ablation	3,070
Function-words semantic ablation	3,861
All-words semantic ablation	6,931

[7] https://huggingface.co/docs/transformers/v4.42.0/en/main_classes/optimizer_schedules#transformers.get_linear_schedule_with_warmup

Table 7: The macro-F1 scores for each connective

	つつ (***tsutsu***)	ところで (***tokorode***)	ながら (***nagara***)
Original (baseline)	0.736	0.604	0.789
Shuffled	0.629	0.249	0.499
Arg1-ablation	0.736	0.660	0.620
Arg2-ablation	0.705	0.706	0.523
Connective ablation	0.478	0.518	0.459
Content-words ablation	0.661	0.243	0.559
Function-words ablation	0.452	0.535	0.355
Mo ablation	0.705	0.814	0.695
Negation ablation	0.736	0.482	0.777
Content-words semantic ablation	0.736	0.417	0.741
Function-words semantic ablation	0.625	0.408	0.530
All-words semantic ablation	0.705	0.067	0.372