BERT-based Cohesion Analysis of Japanese Texts

Nobuhiro Ueda
Kyoto University
ueda@nlp.ist.i.kyoto-u.ac.jp

Daisuke Kawahara
Waseda University
dkw@waseda.jp

Sadao Kurohashi
Kyoto University, CREST
kuro@nlp.ist.i.kyoto-u.ac.jp

Abstract

The meaning of natural language text is supported by cohesion among various kinds of entities, including coreference relations, predicate-argument structures, and bridging anaphora relations. However, predicate-argument structures for nominal predicates and bridging anaphora relations have not been studied well, and their analyses have been still very difficult. Recent advances in neural networks, in particular, self training-based language models including BERT (Devlin et al., 2019), have significantly improved many natural language processing tasks, making it possible to dive into the study on analysis of cohesion in the whole text. In this study, we tackle an integrated analysis of cohesion in Japanese texts. Our results significantly outperformed existing studies in each task, especially about 10 to 20 point improvement both for zero anaphora and coreference resolution. Furthermore, we also showed that coreference resolution is different in nature from the other tasks and should be treated specially.

1 Introduction

The meaning of natural language text is supported by cohesion among various kinds of entities. Such cohesion includes coreference relations, predicate-argument structures, and bridging anaphora relations. For example, there are various relations between entities in the Japanese text (two sentences) of Figure 1. Clarifying these relations is indispensable for computers to understand natural language texts.

Among these relations, coreference relations and predicate-argument structures for verbs have been actively studied. However, predicate-argument structures for nominal predicates and bridging anaphora relations have not been studied well, and their analyses have been still very difficult. Recent advances in neural networks, in particular self training-based language models including BERT (Devlin et al., 2019), have significantly improved many natural language processing (NLP) tasks. By using these techniques, it is now possible to dive into the study on analysis of cohesion in the whole text, including nominal predicate-argument structures and bridging anaphora relations. In this study, we tackle an integrated analysis of cohesion in Japanese texts. To the best of our knowledge, no study has focused on such an integrated analysis of text cohesion.

First, we explain the characteristics of each relation in Japanese texts. A coreference relation is a relation between nouns that refer to the same real-world entity, and the task of revealing this relation is called coreference resolution (CR). In Figure 1, “author” and “he” refer to the same entity. There is no Japanese specific phenomenon in CR.

A predicate-argument structure consists of a predicate and its arguments that fill each case of the predicate, such as *who* does/did *what* to *whom*. The task of clarifying these relations is called predicate-argument structure analysis (PAS analysis). In Japanese, an argument is usually labeled by a case maker

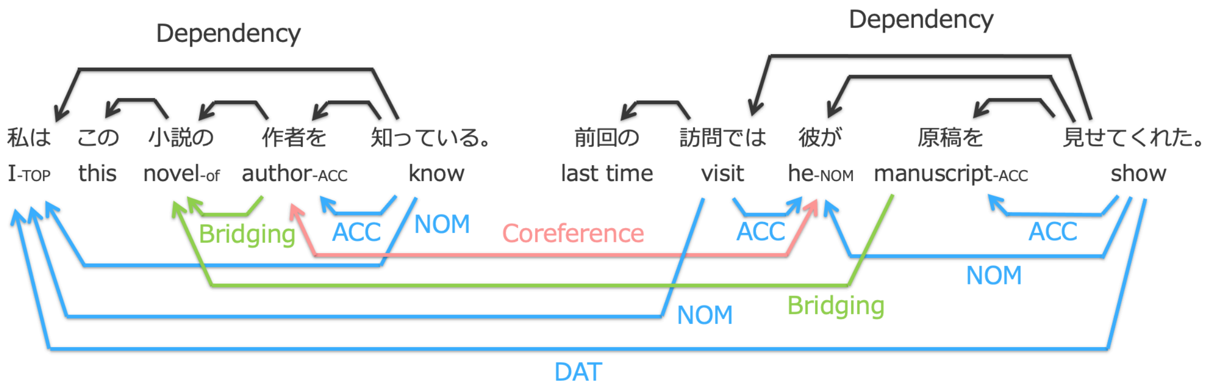

Figure 1: An example of Japanese semantic relations. Its English translation is “I know the author of this novel. He showed me his manuscript on my last visit.” The upper edges represent dependency relations. The lower edges represent verbal and nominal predicate-argument structures, bridging anaphora relations, and coreference relations. “NOM”, “ACC”, and “DAT” represent the nominative, accusative, and dative arguments, respectively.

such as “が” (*ga*) “を” (*wo*), and “に” (*ni*), which roughly correspond to nominative (NOM), accusative (ACC), and dative (DAT), respectively. When an argument is labeled by a case marker and has a dependency relation with a predicate, like the situation of “author-ACC” and “know” in Figure 1, their relation is *overt*. On the other hand, even if an argument and a predicate has a dependency relation, the case marker is sometimes hidden when, for example, a topic maker is attached to the argument. In that situation, we need to clarify their relation: NOM, ACC, or DAT, like the NOM relation between “I-TOP” and “know” in Figure 1. We call it case analysis.

Japanese NLP has been long suffering from zero anaphora resolution because arguments are very often omitted in Japanese texts. In Figure 1, for example, the DAT argument of “show” is omitted in the second sentence, and “I” in the first sentence should be referred to.

This study does not handle overt situations, since it can be analyzed with 95% or higher accuracy based on the recent accurate parsing technology. When we compare our method with related work, we discuss case analysis and zero anaphora resolution separately. When we consider an integrated cohesion model, we show their combined accuracy.

PAS is usually considered for verbal predicates (PAS analysis for verbal predicates is called VPA, hereafter). An eventive noun is syntactically noun, but it takes arguments and they usually appear in the context. Such eventive nouns are called *nominal predicates*. This study also considers PAS analysis for *nominal predicates*, and it is called nominal predicate-argument structure analysis (NPA). There is no overt situation for NPA. Even if an argument and a nominal predicate has a dependency relation like “visit of me”, we have to do case analysis to detect NOM. In most cases, like “visit” in Figure 1, we need to do zero anaphora resolution between entities with no dependency.

A bridging anaphora relation is an anaphoric relation between an anaphor and its antecedent, where the antecedent makes up a kind of semantic insufficiency of the anaphor. For example, “roof” is the roof of a building, “price” is the price of a product, and “author” is the author of a work. In Figure 1, the antecedent of “author” is “novel”, and that of “manuscript” is again “novel”. Bridging anaphora resolution (BAR) does not classify the relation between an anaphor and its antecedent, but just judge if they have a relation or not. Because of that, if they have a dependency like “novel-of author” in Figure 1, they are considered as overt and are not analyzed or evaluated by this study, like an overt situation of VPA. On the other hand, the relation between “manuscript” and “novel” are analyzed like zero anaphora resolution.

Considering relations among these tasks, VPA and NPA are very similar, and BAR is also similar to them since all three tasks consider important relations between different entities. On the other hand, coreference is a relation between identical entities and essentially different from VPA/NPA/BAR. Taking the difference into account, this paper proposes integrated cohesion analysis models and discusses their performance experimentally.

The contribution of this study is three-fold:

- We propose a framework for analyzing multiple semantic relations included in cohesion all together.
- We improve the performance of these tasks greatly without input features by using BERT.
- We focus on the differences of coreference resolution from other semantic tasks and demonstrated that it is effective to treat CR specially.

2 Related Work

Although many studies have focused on Japanese VPA (Omori and Komachi, 2019; Shibata and Kurohashi, 2018; Kurita et al., 2018; Matsubayashi and Inui, 2017; Ouchi et al., 2017; Matsubayashi and Inui, 2018), few studies worked on solving other related tasks simultaneously. Shibata and Kurohashi (2018) introduced a mechanism called an *entity buffer* to capture the saliency of entities in a document, and performed multi-task learning of VPA and CR. They showed that although CR and VPA are heterogeneous tasks, CR is improved by performing VPA. This is because VPA captures the saliency of entities and improves the representation of the entity buffer since entities referenced more by VPA are considered to be more salient. Omori and Komachi (2019) treated not only VPA but also NPA simultaneously and showed that there was a gain on both sides. However, none of the studies performed both CR and NPA. In this study, we perform VPA, NPA, BAR, and CR simultaneously.

In English, Semantic Role Labeling (SRL) is a task similar to Japanese VPA. Li et al. (2020) achieved F1 score of 88.03 in SRL using RoBERTa (Liu et al., 2019). Unlike in English, argument omission frequently occurs in Japanese, and thus Japanese VPA is more difficult. In Shibata and Kurohashi (2018), the performance of zero anaphora resolution is 58.1% (F1 score), while case analysis, a task of finding arguments that have a relation of *case*, can be solved with F1 score of 89.5%.

There are not many studies on bridging anaphora resolution and coreference resolution in Japanese. For BAR, Sasano et al. (2004) constructed a dictionary of expressions like “A の B” (“B of A” or “A’s B”) from a large raw corpus, and used this dictionary for the analysis. Since BAR is a very difficult task, their BAR performance is an F1 score of 42.7.

In English, BAR and CR are performed in a span-based method (Hou, 2020; Wu et al., 2020; Yu and Poesio, 2020), while in Japanese, they are performed in a dependency-based method. Although it cannot be directly compared with our study for this reason, Hou (2020) and Wu et al. (2020) have improved the accuracy of BAR and CR by employing a QA framework. Yu and Poesio (2020) showed the effectiveness of multi-task learning of CR and BAR. Although both are heterogeneous tasks each other, both have one thing in common in English: both of them need to extract the spans corresponding to mentions first. In contrast to CR, the corpus size of the English BAR is very small. Therefore, NN was not able to perform span extraction accurately. The improvement in the accuracy of BAR with multi-task learning of CR is partly due to the fact that the CR provides a good span representation for BAR.

3 Proposed Method

In this study, we perform VPA, NPA, BAR, and CR all together using BERT. This model is called Cohesion Analysis Model (CAModel). In this section, we first describe our Base Model for performing only VPA, and then describe the CAModel for multi-task learning of all four tasks. Finally, we describe the Coreference-aware Cohesion Analysis Model (CorefCAModel), which deals with CR specially.

3.1 Base Model

Our Base Model using BERT is shown in Figure 2. This figure shows the analysis of nominative (NOM) of the predicate t_i in VPA. The predictions are performed by an argument selection method, following Shibata and Kurohashi (2018) and Kurita et al. (2018). When the predicate t_i is the target, the model calculates the probability that a word is the nominative argument of t_i for all other words in the

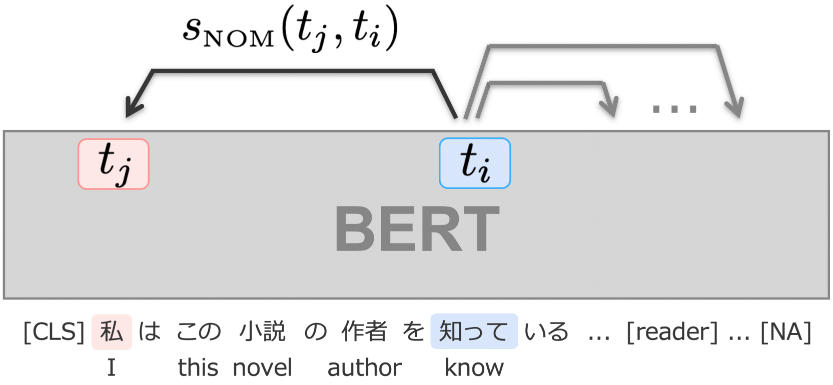
$s_{\mathrm{NOM}}(t_j, t_i)$
...
t_j
t_i
BERT
[CLS] 私 は この 小説 の 作者 を 知って いる ... [reader] ... [NA]
I this novel author know

Figure 2: Our Base Model (in the case of analyzing the nominative (NOM) of verbal predicate t_i).

documents. The one with the highest probability among them is adopted as the nominative argument of t_i. This calculation is done for all the other cases, such as accusative (ACC) and dative (DAT), and for all predicates in the document.

3.1.1 Input Representation

The segmentation of a document consists of three steps: sentence, word, and subword segmentation. Sentence and word segmentation is annotated in the corpora. For subword segmentation, we use BPE (Sennrich et al., 2016) following the segmentation method used at the pre-training stage. Following Devlin et al. (2019), we insert `[CLS]` and `[SEP]` tokens at the beginning and end of a document, respectively. In addition, we insert five special tokens at the end of input sequence: `[author]`, `[reader]`, `[unspecified person]`, `[NULL]`, and `[NA]`. `[author]`, `[reader]`, and `[unspecified person]` are used in exophora resolution. In anaphora resolution, an anaphor sometimes refers to an entity that does not appear in the document. This phenomenon is referred to as exophora. In this study, *author*, *reader*, and *unspecified:person* are taken into consideration as the targets of exophora, and we use `[author]`, `[reader]`, and `[unspecified person]` as proxy-targets for exophora. `[NULL]` and `[NA]` mean that a predicate takes no argument and that a mention has no other coreferring mentions, respectively.

3.1.2 Output Layer

We put Multi Layer Perceptron (MLP) on top of BERT as an output layer. This MLP calculates the probability by using the output of BERT. Specifically, the probability that a subword t_j, which corresponds to an argument candidate, is the c-case argument of a subword t_i, which corresponds to a predicate, is calculated as follows:

$$P(t_j|t_i, c) = \frac{\exp(s_c(t_j, t_i))}{\sum_k \exp(s_c(t_k, t_i))} \tag{1}$$

$$s_c(t_j, t_i) = \boldsymbol{v}^{\mathrm{T}} \tanh(W_c \boldsymbol{t}_j + U_c \boldsymbol{t}_i), \tag{2}$$

where W_c and U_c are weight matrices for each relation, and $\boldsymbol{v}$ is a weight vector shared across relations. $\boldsymbol{t}_i$ and $\boldsymbol{t}_j$ denote hidden vectors of the BERT's final layer corresponding to the subword t_i and t_j.

Note that the prediction by this model is based on subword units, while in Japanese a basic phrase, which consists of one content word and zero or more function words, is the basic unit of the four tasks. Thus, we adopt the head subword of the content word in a basic phrase as the representative of the basic phrase.

3.2 Cohesion Analysis Model

In addition to VPA, we also perform NPA, BAR, and CR simultaneously. This model is called Cohesion Analysis Model (CAModel). All of these analyses are performed by the argument selection method

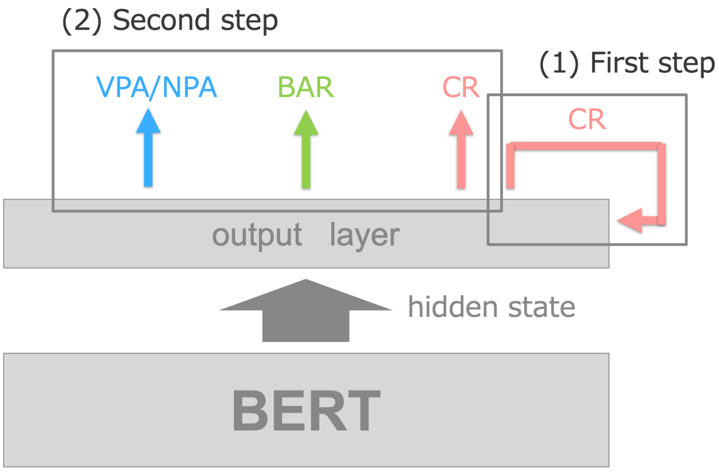

Figure 3: CorefCAModel. This model solves CR first (1) and then solves all tasks using CR features (2).

in the same way as the Base Model. For CR, we select mentions that share the entity for a target noun as we select a nominative argument for a target predicate in VPA. For BAR, we select a noun which has a bridging anaphora relation with a target noun. While CR and BAR are performed on a different network from VPA, NPA is performed on the same network as VPA because the set of relations to be analyzed is the same as that of VPA.

3.3 Coreference-aware Cohesion Analysis Model

CR is different in nature from other tasks. For this reason, in simple multi-task learning, CR is not expected to contribute to the performance. We will demonstrate this by experiments in Section 4. To effectively take advantage of CR information, we propose the Coreference-aware Cohesion Analysis Model (CorefCAModel), which treats CR specially. An overview of this model is shown in Figure 3. The prediction of the CorefCAModel consists of two steps. In the first step it performs only CR, and in the second step it analyzes all tasks using the CR results obtained in the previous step as features. Specifically, the probability from the CorefCAModel $P'(t_j|t_i, c)$ is calculated as follows:

$$P'(t_j|t_i,c) = \frac{\exp(s'_c(t_j,t_i))}{\sum_k \exp(s'_c(a_k,t_i))} \tag{3}$$

$$s'_c(t_j,t_i) = \boldsymbol{v}^{\mathrm{T}} \tanh(\boldsymbol{h}_{i,j,c}) \tag{4}$$

$$\boldsymbol{h}_{i,j,c} = W_c \boldsymbol{t}_j + U_c \boldsymbol{t}_i + \boldsymbol{h}_j^{coref} \tag{5}$$

$$\boldsymbol{h}_j^{coref} = \sum_k P_{coref}(t_k|t_j) V \boldsymbol{t}_k, \tag{6}$$

where V is a weight matrix shared across relations and $\boldsymbol{t}_i$, $\boldsymbol{t}_j$, and $\boldsymbol{t}_k$ denote hidden vectors of the BERT's final layer corresponding to the subword t_i, t_j, and t_k. $P_{coref}(t_k|t_j)$ is the output of the first step and represents the probability that t_j has a coreference relation with t_k. This is calculated in the same way as Base Model as follows:

$$P_{coref}(t_k|t_j) = \frac{\exp(s'_{coref}(t_k,t_j))}{\sum_l \exp(s'_{coref}(t_l,t_j))} \tag{7}$$

$$s'_{coref}(t_k,t_j) = \boldsymbol{v}^{\mathrm{T}} \tanh(W_{coref} \boldsymbol{t}_k + U_{coref} \boldsymbol{t}_j). \tag{8}$$

[1] This score does not include inter-sentential zero anaphora resolution since Kurita et al. (2018) does not consider it.

	train	dev	test
Web	12,258	1,585	2,195
News	8,348	1,145	1,783

Table 1: The number of sentences in the Web and News corpora.

	Web			News		
	case analysis	zero anaphora resolution	coreference resolution	case analysis	zero anaphora resolution	coreference resolution
S&K (2018)	89.2	58.1	68.5	**89.5**	35.6	54.1
Kurita et al. (2018)	**92.0**	58.4[1]	-	-	-	-
CAModel (ours)	88.3 (±0.5)	**70.3** (±0.7)	**77.0** (±0.6)	85.9 (±0.5)	**56.7** (±0.5)	**57.6** (±0.8)

Table 2: Our CAModel performance (F-measure) compared with Shibata and Kurohashi (2018) and Kurita et al. (2018).

4 Experiments

4.1 Experimental Settings

In our experiments, we used CAModel and CorefCAModel. CAModel was trained on various combinations of tasks. We fine-tuned both the models for 4 epochs using cross entropy loss following Devlin et al. (2019). Since CorefCAModel cannot perform CR with sufficiently high accuracy in the early stage of training, we mixed gold coreference data with the first stage prediction and gradually reduced the gold ratio inspired by Scheduled Sampling (Bengio et al., 2015).

We used two kinds of datasets for our experiment. One is the Kyoto University Web Document Leads Corpus (Web corpus) (Hangyo et al., 2012), and the other is the Kyoto University Text Corpus (News corpus) (Kawahara et al., 2002). Verbal predicate-argument relations, nominal predicate-argument relations, coreference relations, and bridging anaphora relations are manually annotated in both the corpora. Table 1 lists the number of sentences in each corpus. In our experiment, training was performed on a mixture of both corpora[2]and the evaluation was done on each corpus.

We used the NICT BERT Japanese pre-trained model (with BPE).[3] This model was trained after morphological and subword segmentation using the full text of Japanese Wikipedia for approximately 1 million steps. At the fine-tuning stage, we set the maximum sequence length to 128. The maximum sequence length of the Web corpus was shorter than 128. In the News corpus, there are many documents with sequence lengths exceeding 128, and one document is divided into multiple parts for training. To do this, we divided a document so that it had as many preceding contexts as possible.

For VPA, we extracted all predicates in a document, and analyzed them in terms of four cases of NOM, ACC, DAT and NOM2.[4] The Japanese dependency parser KNP (Kurohashi and Nagao, 1994) was used for predicate extraction. For NPA, we analyzed nouns that KNP judged to have arguments. For both of VPA and NPA, we used arguments that have *case* or *zero* relation for training and evaluation.[5] BAR and CR were performed on nouns. Following Shibata and Kurohashi (2018), we consider *author*, *reader*, and *unspecified person* as targets of exophora, and the evaluation of VPA, NPA, BAR, and CR was relaxed using a gold coreference chain.[6]

[2] In our preliminary experiments, we have verified that mixing the corpora leads to better performance than using them alone.

[3] https://alaginrc.nict.go.jp/nict-bert/index.html

[4] In Japanese, a predicate sometimes has two nominative arguments, so we use nominative2 (NOM2) case to distinguish them.

[5] In our preliminary experiments, solving *overt* arguments together slightly worsened the performance of zero anaphora resolution.

[6] Our code can be found in this repository: https://github.com/nobu-g/cohesion-analysis

	VPA (Web)	VPA (News)
CAModel (VPA)	76.91 (±0.36)	68.70 (±0.70)
CAModel (VPA + CR)	77.04 (±0.25)	68.77 (±0.24)
CAModel (VPA + NPA)	77.38 (±0.17)	69.55 (±0.29)
CAModel (VPA + BAR)	76.42 (±0.47)	69.34 (±0.39)
CAModel (VPA + NPA + BAR)	77.41 (±0.25)	**69.76** (±0.38)
CAModel (VPA + NPA + BAR + CR)	77.15 (±0.62)	69.37 (±0.26)
CorefCAModel (CR → VPA + NPA + BAR + CR)	**77.56** (±0.52)	69.27 (±0.25)

Table 3: Performance (F-measure) of verbal predicate-argument structure analysis (**VPA**) evaluated on the test set of the Web and News corpora.

	NPA (Web)	NPA (News)
CAModel (NPA)	58.82 (±1.98)	54.52 (±0.86)
CAModel (NPA + CR)	59.58 (±0.56)	53.15 (±0.94)
CAModel (NPA + VPA)	**61.99** (±1.11)	56.27 (±0.44)
CAModel (NPA + BAR)	58.64 (±2.01)	54.15 (±1.18)
CAModel (NPA + VPA + BAR)	61.16 (±0.73)	**56.33** (±0.30)
CAModel (NPA + VPA + BAR + CR)	61.82 (±0.80)	55.77 (±0.52)
CorefCAModel (CR → NPA + VPA + BAR + CR)	61.45 (±0.50)	55.43 (±0.95)

Table 4: Performance of nominal predicate-argument structure analysis (**NPA**) evaluated on the test set of the Web and News corpora.

	BAR (Web)	BAR (News)
CAModel (BAR)	61.53 (±0.32)	**55.77** (±0.61)
CAModel (BAR + CR)	60.19 (±0.80)	54.00 (±0.95)
CAModel (BAR + VPA)	**62.03** (±1.19)	55.76 (±0.73)
CAModel (BAR + NPA)	60.98 (±1.40)	54.91 (±0.97)
CAModel (BAR + VPA + NPA)	61.10 (±0.83)	55.00 (±0.96)
CAModel (BAR + VPA + NPA + CR)	60.39 (±1.12)	53.93 (±1.26)
CorefCAModel (CR → BAR + VPA + NPA + CR)	60.14 (±0.92)	54.32 (±1.00)

Table 5: Performance of bridging anaphora resoluton (**BAR**) evaluated on the test set of the Web and News corpora.

	CR (Web)	CR (News)
CAModel (CR)	77.81 (±0.62)	**59.94** (±0.75)
CAModel (CR + VPA)	77.36 (±0.49)	58.89 (±0.48)
CAModel (CR + NPA)	77.30 (±0.33)	58.40 (±0.58)
CAModel (CR + BAR)	**77.88** (±0.39)	58.13 (±0.77)
CAModel (CR + VPA + NPA)	77.46 (±0.84)	58.74 (±0.50)
CAModel (CR + VPA + NPA + BAR)	77.02 (±0.60)	57.55 (±0.76)
CorefCAModel (CR → VPA + NPA + BAR + CR)	76.56 (±0.58)	57.58 (±1.28)

Table 6: Performance (F-measure) of coreference resolution (**CR**) evaluated on the test set of the Web and News corpora.

4.2 Experimental Results

Table 2 shows comparisons with existing studies with an F-measure. The 95% confidence intervals of the results for five runs with different random seeds are shown in the parentheses. Our model greatly improved the performance of zero anaphora resolution, which is considered to be particularly difficult, and coreference resolution. The performance of case analysis is worse than the existing studies. This is probably because our study, unlike the existing studies, does not use features such as the dependency structure of the input sentence and selectional preferences. Hereafter, case analysis and zero anaphora resolution are collectively referred to as VPA.

Tables 3, 4, 5, and 6 show the results of VPA, NPA, BAR, and CR, respectively. VPA, NPA, BAR, and CR in parentheses denote the tasks performed at the training stage.

First, we focus on the effect of VPA in multi-task learning. Regarding NPA (Table 4), multi-task learning with VPA improved the performance. This also holds for BAR in Table 5.

Next, we focus on the effect of NPA in multi-task learning. Table 3 shows that of the four tasks, NPA contributed the most to the performance of VPA. On the other hand, in Table 5 we can see that solving NPA slightly worsened the performance of BAR.[7]

Next, we focus on the effect of BAR in multi-task learning. In Table 3, comparing CAModel (VPA + NPA) and CAModel (VPA + NPA + BAR), solving BAR, in addition to VPA and NPA, slightly improved the performance of VPA. BAR had little effect on NPA (Table 4).

Finally, we focus on the effect of CR in multi-task learning. Regarding VPA (Table 3), while NPA and BAR contributed to the performance of VPA, solving CR together reduces the score. Although the score of CAModel (VPA + CR) is slightly higher than that of CAModel (VPA), the score of CAModel (VPA + NPA + BAR + CR) is much lower than that of CAModel (VPA + NPA + BAR). CR also worsened the performance of BAR (Table 5). Table 6 shows that VPA, NPA, and BAR do not contribute to the performance of CR, and solving CR by itself is the best in CAModel.

CorefCAModel outperformed all the other models in VPA on the Web corpus. In other tasks, CorefCAModel performs slightly worse than other models.

4.3 Discussion

The results show that VPA, NPA, and BAR are generally beneficial mutually in multi-task learning, but CR is not. CR reduces the F1 score of VPA and BAR, and has no benefit of multi-task learning with other tasks. While VPA, NPA, and BAR are tasks to analyze relations between entities (or predicates), CR is a task to find the same entity. Our experimental results confirm the fact that CR is different in nature from other tasks.

Therefore, CR should be treated differently from VPA, NPA, and BAR. In this study, we proposed a CorefCAModel which performs only CR first and uses its results as features for all the tasks. We confirmed that CorefCAModel performs better in some tasks than CAModel, which simply performs multi-task learning. However, in most tasks, CorefCAModel does not perform better than CAModel. CR information could potentially help the other tasks, and it is our future work to consider a better combined model of the four tasks.

Figure 4 shows an analysis example of a document in the Web corpus. In this example, most of the relations of VPA, BAR, and CR were correctly analyzed. While the correct NOM argument of “study” and “worried” is “girl”, it was analyzed as “I”. This is because the topic in the first sentence is “I” with a topic marker but is superseded by “girl” in the second sentence. This kind of discourse structure could not be captured by the current model. Figure 5 shows an analysis example of a document in the News corpus. The model correctly analyzed nominal predicate-argument structures such as “China-NOM aid” and “base-ACC construction”, which can be predicted from relatively local cohesion. On the other hand, “China-of navy” and “Myanmar-DAT construction” were not analyzed correctly. Capturing these relations requires a deeper understanding of the document. For example, “China-of navy” could be correctly analyzed if the information about other relations such as “base-ACC construction” and “navy-of base” were used. In the future, we plan to take advantage of the information about other relations for the analysis target.

[7] Because NPA has not been studied much, there may be problems with the quality of the annotation.

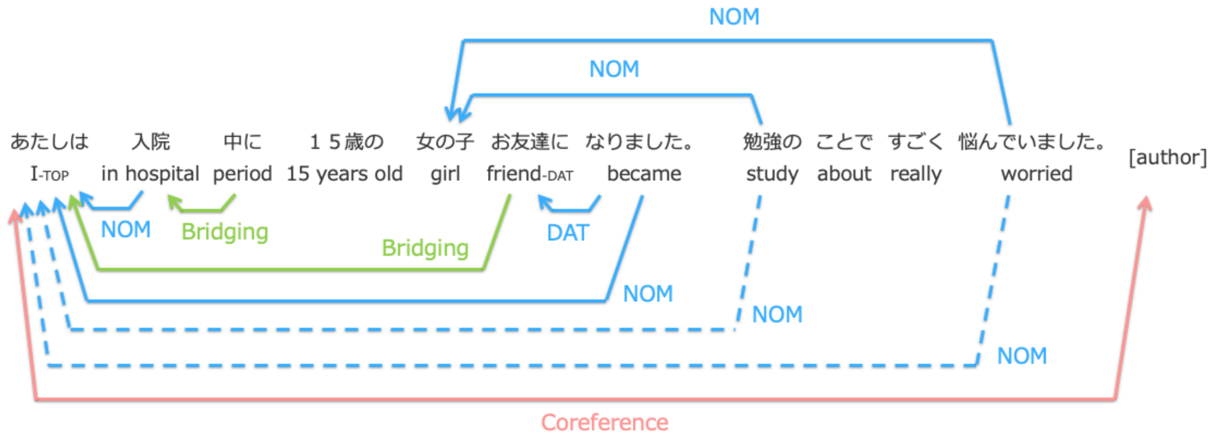

Figure 4: An example of cohesion analysis from the Web corpus. The English translation is "When I was in the hospital, I became friends with a 15-year-old girl. She was very worried about her studies." The lower edges represent system outputs. The broken lines represent incorrect system outputs. The upper edges represent gold data that the system could not predict correctly.

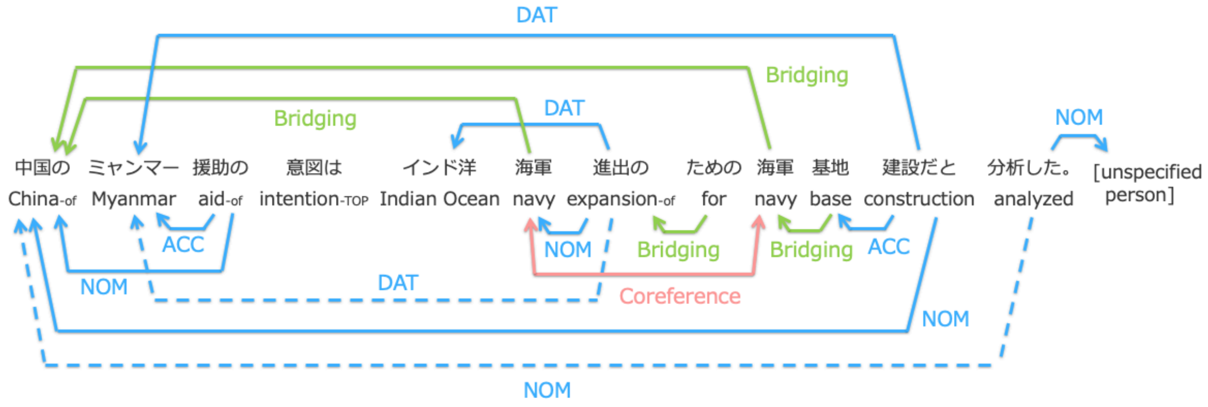

Figure 5: An example of cohesion analysis from the News corpus. The English translation is "(They) analyzed that China's intention to aid Myanmar is to build a naval base for the Indian Ocean naval expansion." The lower edges represent system outputs. The broken lines represent incorrect system outputs. The upper edges represent gold data that the system could not predict correctly.

5 Conclusion

We proposed multi-task learning of cohesion analysis including verbal predicate-argument structure analysis, nominal predicate-argument structure analysis, coreference resolution, and bridging anaphora resolution using BERT, and investigated the effect of each task on multi-task learning. Our model significantly outperformed existing studies in each task, especially about 10 to 20 point improvement for zero anaphora resolution. Furthermore, we also showed that coreference resolution is different in nature from the other tasks and should be treated specially. In the future, based on this fact, we would like to create a model that can better utilize coreference information for other tasks.

References

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. 2015. Scheduled sampling for sequence prediction with recurrent neural networks. *CoRR*, abs/1506.03099.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL*, pages 4171–4186.

Masatsugu Hangyo, Daisuke Kawahara, and Sadao Kurohashi. 2012. Building a diverse document leads corpus annotated with semantic relations. In *Proceedings of PACLIC*, pages 535–544.

Yufang Hou. 2020. Bridging anaphora resolution as question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1428–1438, Online, July. Association for Computational Linguistics.

Daisuke Kawahara, Sadao Kurohashi, and Kôiti Hasida. 2002. Construction of a Japanese relevance-tagged corpus. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*. European Language Resources Association (ELRA).

Shuhei Kurita, Daisuke Kawahara, and Sadao Kurohashi. 2018. Neural adversarial training for semi-supervised Japanese predicate-argument structure analysis. In *Proceedings of ACL*, pages 474–484.

Sadao Kurohashi and Makoto Nagao. 1994. A syntactic analysis method of long japanese sentences based on the detection of conjunct ive structures. *Computational Linguistics*, 20(4).

Tao Li, Parth Anand Jawale, Martha Palmer, and Vivek Srikumar. 2020. Structured tuning for semantic role labeling. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8402–8412, Online, July. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Yuichiroh Matsubayashi and Kentaro Inui. 2017. Revisiting the design issues of local models for Japanese predicate-argument structure analysis. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 128–133, Taipei, Taiwan, November. Asian Federation of Natural Language Processing.

Yuichiroh Matsubayashi and Kentaro Inui. 2018. Distance-free modeling of multi-predicate interactions in end-to-end Japanese predicate-argument structure analysis. In *COLING2018*, pages 94–106.

Hikaru Omori and Mamoru Komachi. 2019. Multi-task learning for Japanese predicate argument structure analysis. In *Proceedings of NAACL*, pages 3404–3414.

Hiroki Ouchi, Hiroyuki Shindo, and Yuji Matsumoto. 2017. Neural modeling of multi-predicate interactions for Japanese predicate argument structure analysis. In *Proceedings of ACL*, pages 1591–1600.

Ryohei Sasano, Daisuke Kawahara, and Sadao Kurohashi. 2004. Automatic construction of nominal case frames and its application to indirect anaphora resolution. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 1201–1207, Geneva, Switzerland, aug 23–aug 27. COLING.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *ACL2016*, pages 1715–1725.

Tomohide Shibata and Sadao Kurohashi. 2018. Entity-centric joint modeling of Japanese coreference resolution and predicate argument structure analysis. In *Proceedings of ACL*, pages 579–589.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Wei Wu, Fei Wang, Arianna Yuan, Fei Wu, and Jiwei Li. 2020. CorefQA: Coreference resolution as query-based span prediction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6953–6963, Online, July. Association for Computational Linguistics.

Juntao Yu and Massimo Poesio. 2020. Multi-task learning based neural bridging reference resolution. *ArXiv*, abs/2003.03666.

A Hyper parameters

The hyperparameters are listed in Table 7.

[8] This scheduler is implemented in Transformers (Wolf et al., 2019) and we used it.

Parameter name	Parameter value
Optimizer	AdamW
Learning rate	5×10^{-5}
Optimizer eps	1×10^{-8}
Weight decay	0.01
Dropout rate (BERT layer)	0.1
Dropout rate (output layer)	0.0
LR scheduler	linear_schedule_with_warmup[8]
Scheduler warmup proportion	0.1
Batch size	8

Table 7: Hyperparameters.