Predicting Antonyms in Context using BERT

Ayana Niwa [†], **Keisuke Nishiguchi**[‡], **Naoaki Okazaki**[†]

[†] Tokyo Institute of Technology
[‡] CyberAgent, Inc.
ayana.niwa at nlp.c.titech.ac.jp
nishiguchi_keisuke at cyberagent.co.jp
okazaki at c.titech.ac.jp

Abstract

We address the task of antonym prediction in a context, which is a fill-in-the-blanks problem. This task setting is unique and practical because it requires contrastiveness to the other word and naturalness as a text in filling a blank. We propose methods for fine-tuning pre-trained masked language models (BERT) for context-aware antonym prediction. The experimental results show that these methods have positive impacts on the prediction of antonyms within a context. Moreover, human evaluation reveals that more than 85% of the predictions using the proposed method are acceptable as antonyms.

1 Introduction

Antonymy is a relationship between two words that express contrasting or opposite meanings (e.g., "agree–disagree"). Capturing antonymy is directly helpful for downstream applications such as sentiment transfer (Li et al., 2018) and claim generation (Hidey and McKeown, 2019). Further, semantically contrasting expressions with antonyms are utilized in advertising slogans (Katrandjiev et al., 2016), political speeches (Heritage and Greatbatch, 1986), and Chinese poetry (Yan et al., 2016).

As antonymy is one of the relations of lexical semantics, such as synonymy and hyponymy, antonymy can be modeled using a similar approach to lexical knowledge acquisition. Most of the published studies on this topic have focused on the prediction of the relation between a given word pair (Barkan et al., 2020; Shwartz and Dagan, 2016), or a target (tail) for a given word (head) and its relation (Camacho-Collados et al., 2018; Rimell et al., 2017). However, predicting an antonym is challenging because multiple types of words are plausible as antonyms for a word. This is because a word can have semantic contrastiveness to the other as long as the word contains at least one feature contrasting to the other (Leech, 1976). For example, *dual*, *double*, and *multiple* can be antonyms for *single* because they all have the contrasting features of AMOUNT or NUMBER. Additionally, the appropriateness of an antonym varies depending on the context. For example, *double*, *dual*, and *multiple* are used for a bed, nationality, and the number of meanings of a word, respectively. Hence, antonym prediction must be considered within the context.

In this study, we consider the new task of antonym prediction, that is, **the fill-in-the-blanks problem for antonyms in context**. For example, in the sentence, "A ____ bed is better than *single* for me," we expect to fill the blank with the words "*double*" or "*king-sized*." The fill-in-the-blanks setting requires the prediction of context-aware antonyms by capturing the contrasting features between the word pair. The task also requires a consideration of the naturalness of a text when filling the blank, which is necessary for applications of generating text with antonyms.

In recent years, pre-training and fine-tuning approaches have achieved high performance in various NLP tasks (Devlin et al., 2019; Yang et al., 2019). Therefore, we use Bidirectional Encoder Representations from Transformers (BERT) as a pre-trained model to predict antonyms in a context. However, it is not easy to collect training data for fine-tuning the model, that is, text containing antonym pairs with a contrastive context.

Therefore, we focus on the rhetorical device that effectively employs antonymy, that is, *antithesis*, which juxtaposes words or phrases in a similar structure with contrasting meanings. An antithesis is suitable for data creation because it ensures that a text has one or more antonym pairs in a contrastive context. For example, the sentence, "My mother who [is *sensitive* to the *pension*] [is *insensitive* to the *insurance*]," has an antithesis structure with two antonym pairs. We propose four methods to fine-tune BERT for antonyms: (1) domain adaptation using an antithesis corpus, (2) contrastive masking to focus on antonym prediction, (3) antithesis po-

Proceedings of the 14th International Conference on Natural Language Generation (INLG), pages 48–54,
Aberdeen, Scotland, UK, 20-24 September 2021.

sitional encodings to capture antithesis structures, and (4) pseudo-supervision data collected by automatic annotation using an antonym dictionary.

The experimental results with the Japanese slogan corpus demonstrate that the proposed fine-tuning methods contribute to the adaptation of BERT to the context-aware antonym prediction task. An automatic evaluation based on a single correct answer is improper because there are multiple acceptable answers. However, the manual evaluation revealed that more than 85% of the words predicted by the proposed method are appropriate as antonyms and that fine-tuned BERT is highly capable of capturing antonymy in a context.

2 Method

2.1 Model

Given a sequence of n tokens, $x_1, \ldots, x_n$, with a `[MASK]` (blank) token at position m ($1 \leq m \leq n$), the conditional probability of token y_m for filling the blank can be modeled using a BERT (Devlin et al., 2019) (illustrated in Figure 1),

$$P(y_m | x_1, \ldots, x_m, \ldots, x_n). \quad (1)$$

Based on the bidirectional contexts of the input, BERT considers the surrounding context of `[MASK]`. By fine-tuning BERT on a text corpus with an antithesis structure (described in Section 2.2), we can expect that the model will eventually consider antonymy in an input text by domain adaptation because an antithesis contains more than one antonym pair and the contrastive context.

To utilize a small corpus for adapting BERT for antonyms, we explore two approaches. First, we create supervision data for fine-tuning by replacing a token with `[MASK]` such that the `[MASK]` token is likely to have a counterpart in the text. For example, given a text, "[Starts with the reckoning], [ends with the relish]," we obtain two training instances, that is, "`[MASK]` with the `[MASK]`, ends with the relish" and "Starts with the reckoning, `[MASK]` with the `[MASK]`." These `[MASK]` tokens are chosen because they do not appear in the counterpart phrase, whereas "with" and "the" do. We refer to this strategy as *contrastive masking*. This strategy selectively creates supervision data for filling antonyms more efficiently than the default strategy for BERT (deciding `[MASK]` positions randomly).

Second, we extend the positional encodings in BERT to indicate an antithesis structure in an input. Consider a text that includes two spans $[i, j)$ and $[k, l)$ ($1 \leq i < j \leq k < l \leq n$) forming an antithesis structure and the span $[i, j)$ includes `[MASK]` tokens. To indicate that span $[i, j)$ corresponds to $[k, l)$, we compute an index specialized for the antithesis structure,

$$a_t = \begin{cases} k + \left\lfloor \frac{(l-k)(t-i)}{j-i} \right\rfloor & (a \in [i, j)) \\ t & (\text{otherwise}) \end{cases}. \quad (2)$$

The index a_t represents the position of the token x_t if $t \notin [i, j)$ but that of the counterpart span $[k, l)$ if $t \in [i, j)$. We used the mean of absolute positional encoding (used in the original BERT) and *antithesis positional encodings* (indexed by a_t) as the positional encodings for BERT.

2.2 Supervision data

For the domain of the training data, we selected advertising slogans in which antitheses were likely to be used frequently. We used a corpus of advertising slogans that consisted of 111,295 Japanese slogans collected from existing books (Taniyama, 2007; Nakahata, 2008; Aota et al., 2007; Umeda, 2016; Sendenkaigi Award Committee, 2003–2018) to construct the supervision data. With the slogans, we constructed an antithesis corpus manually by crowd-sourcing two subtasks: (1) filtering out slogans that do not contain antithesis structures with strict criteria, and (2) annotating antithesis spans. This process yielded 7,457 slogans with annotated spans of antitheses[1]. Additional information is provided in the Appendix.

2.3 Pseudo-supervision data

The number of instances in the supervision data may be small for fine-tuning BERT. Thus, we also explore an approach for automatically annotating a text in a manner similar to distant supervision. Specifically, we find slogans that include pairs of antonyms included in the antonym dictionary (Sanseido Editorial Office, 2017). This process resulted in 1,894 slogans that were not included in the dataset explained in Section 2.2. The strict criteria filtered out these slogans in the first step of the corpus construction, but some of them actually presented antitheses. For each pair of antonyms in a slogan, we obtain two training instances, one antonym replaced with the `[MASK]` token, and vice versa. In this way, we inject the lexical knowledge of antonyms into BERT.

[1] Unfortunately, we cannot release the corpus to the public because we do not own the copyrights of the slogans.

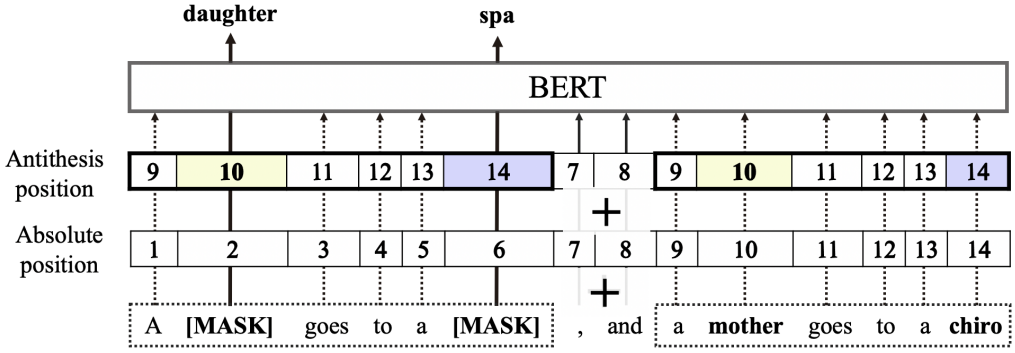

Figure 1: Outline of the proposed method. This is an example of predicting "daughter" and "spa" for the input text "A [MASK] goes to a [MASK], and a mother goes to a chiro." The phrases, "A [MASK] goes to a [MASK]" and "a mother goes to a chiro" consist of an antithesis structure.

3 Experiments and Results

3.1 Experimental settings

Dataset We split the 7,457 slogans in the antithesis corpus into training, development, and test data, and subsequently converted them into fill-in-the-blank instances that contained the [MASK] tokens. Each text yielded two masked instances because an antithesis structure has two contrastive phrases. In this manner, we obtained 11,922 training, 1,496 development, and 1,496 test instances. We also used 3,788 training instances from the pseudo-supervision data. In addition to them, we created a subset of the test data ("word level" hereafter) for a fair comparison with the human baseline. The above test data contained instances wherein the [MASK] token is split into subword units, which human subjects cannot fill in. Additionally, our fill-in-the-blank problem is complex for general cloud workers to solve because it require an understanding of the context. Therefore, we created the simple "word level" test set based on the following criteria: (1) a single word is selected as a masked token per test instance, (2) the selected word is not split into subwords, and (3) the part of speech is either a noun, verb, adjective, or adjective-verb. Furthermore, we randomly selected only one test instance per slogan to simplify the crowd-sourcing process. This process created 529 word level test instances, which is smaller than the entire test set.

Baselines We used two baselines, dictionary-lookup and pre-trained BERT without fine-tuning. The dictionary-lookup baseline examines whether the gold word of each blank is registered in the dictionary (Sanseido Editorial Office, 2017) as an antonym of any word in the corresponding phrase[2]. A pre-trained BERT without fine-tuning is evaluated to investigate the ability of the model to predict antonyms in a context without specialized training. We used BERT pre-trained with the Japanese Wikipedia[3]. To assess the difficulty of this task, we asked three human subjects to guess at most five possible words to fill the blanks in the test set.

We do not employ a non-contextual baseline other than the dictionary-lookup because the dataset has annotations of antithesis structures only at the segment level (phrase-to-phrase alignment) but not at the word level (word-to-word alignment).

Evaluation metrics We used top-1 and top-10 accuracy values as the measures for the correctness of model predictions. Because human subjects could not always come up with five answers for a blank, we used the top-1 and top-n accuracy values, wherein the number of n varied depending on the number of human responses in each instance.

3.2 Results

Table 1 reports the accuracy of the prediction of blanks on the test data. The proposed method achieved 29.3% top-1 and 53.8% top-10 accuracies measured for all instances, and 30.4% top-1 and 49.1% top-n accuracies measured at the word level. The pre-trained BERT without fine-tuning obtained much lower accuracies than those of the proposed method. This indicates that a general masked language model was insufficient to predict antonyms even if presented in the context of the antithesis structure.

[2]This baseline presents the upper bound of the performance of dictionary-lookup because it knows the gold words.

[3]https://github.com/cl-tohoku/bert-japanese

	All		Word level	
	Acc@1	Acc@10	Acc@1	Acc@n
dictionary-lookup	-	-	9.6	-
pre-trained BERT (w/o fine-tuning)	15.0	40.9	15.7	39.1
fine-tuned BERT (default masking)	24.4	51.4	25.0	44.4
- default masking + contrastive masking	28.8	52.6	27.4	47.4
+ antithesis positional encodings	28.7	53.5	27.4	48.0
+ pseudo-supervision data	**29.3**	**53.8**	**30.4**	**49.1**
human (lowest)	-	-	31.5	52.3
human (highest)	-	-	34.5	59.1
human (votes from three subjects)	-	-	51.8	66.6

Table 1: Accuracy values of antonym prediction.

	Contrastiveness	Naturalness
human	94	90
method	88	85

Table 2: Number of contrastive and natural instances (out of 100) judged by a human.

Fine-tuning BERT on the supervision data boosted the performance, especially for top-1 predictions (+9.4 and +9.3 points). Contrastive masking improved all the accuracies, and antithesis positional encodings improved the top-10 and top-n accuracies in particular (+1.2 to +3.0 points for the former and +0.9 and +0.6 points for the latter). Moreover, we confirmed that the pseudo-supervision data improved the accuracy, especially for top-1 predictions (+0.6 and +3.0 points). The fact that these proposed methods contribute to the performance shows the importance of fine-tuning BERT with a special focus on antonym prediction.

The baseline of dictionary-lookup obtained 9.6% top-1 accuracy measured at the word level. We found that the low coverage of the dictionary was the leading cause: the dictionary had entries for only 39.3% of antonyms in the test data.

Table 1 also illustrates the results of human subjects who had the lowest and highest accuracy when solving the fill-in-the-blank task. The accuracy values of all the human subjects were better than those of the proposed method, although the performance of each human subject varied. However, even the best-performing human subject could achieve 34.5% top-1 and 59.1% top-n accuracies, which justifies the difficulty of this task. Conversely, with the most lenient evaluation in which we regard a prediction as correct if any of the three human subjects provided the right answer, the top-1 accuracy was 51.8%, and the top-n accuracy was 66.6%. The performance increase in top-1 implies that multiple words are acceptable for the blanks in the test data, and the characteristic is considered the reason why even human subjects cannot achieve high accuracy values. To investigate such cases with multiple possible correct words, we conducted a subjective evaluation of the quality of the answers of the human subjects and the proposed method (with the participation of another human subject). For this analysis, we used 100 instances sampled at random from the test data for which the answers of both human subjects and methods did not match the correct word. We chose an answer from three subjects at random for each instance because we had multiple answers from three subjects. Table 2 reports the number of predictions from the human subjects and the proposed method for which the manual evaluation recognized contrastiveness and naturalness (fluency). The results reveal that more than 85% of both the answers of the human subjects and the predicted words of the proposed method are appropriate as antonyms.

To summarize, we found that the automatic evaluation using a single correct answer (accuracy) underestimated the context-aware antonym prediction performance because there could be multiple acceptable answers. However, the subjective evaluation revealed that the predictions of the proposed method were satisfactory in terms of contrastiveness (as an antonym) and naturalness in a context.

3.3 Analysis

We list the predictions by the baseline (BERT without fine-tuning) and the proposed method, and the answers by each human subject in Table 3.

When the antonymy was easy to understand,

Example (A)		Example (B)	
別れの曲だったのに、[MASK]の曲になった。 It was the farewell song, but became the [MASK] song.		地球の環境より、まず[MASK]の環境。 Put the environment of the [MASK], before the environment of the earth.	
correct answer: 出会い encounter		correct answer: 心 mind	
baseline	別れ, 最後, 今, 人生 farewell, last, present, life's	baseline	宇宙, 水, 地球, 太陽, 植物 universe, water, earth, sun, plants
proposed	出会い, 憧れ, 最高, 始まり encounter, longing, best, beginning	proposed	家族, 私, 周り, トイレ, 家 family, of myself, vicinity, restroom, house
human 1	出会い, 再会, 初恋, 永遠 encounter, reunion, first love, eternal	human 1	自宅, 自分, 部屋, 職場 one's home, of myself, room, workplace
human 2	出会い, 始まり, 邂逅 (unexpected) encounter, beginning	human 2	自分, 私, 周辺, 室内, 家内 of myself, vicinity, room, one's wife
human 3	出会い encounter	human 3	国, 家庭, 町, 周り country, family, town, vicinity

Table 3: Examples of prediction of methods and answers of human subjects. Owing to space limitations, we removed some duplicated words, which are synonyms (e.g., beginning and start) or the same word in different character types (e.g., Hiragana and Kanji in Japanese).

such as "farewell–encounter" in Example (A), both the proposed method and the human subjects could output the correct word as the first candidate. Compared to the baseline, the proposed method could focus on the contrastiveness between the phrases "the farewell song–the encounter song."

In Example (B), the correct word was not predicted or answered, but their outputs were the antonyms. The output word should be contrasted with the word "earth" in terms of its scale and degree of familiarity. In this respect, both the prediction of the proposed method and the answers of the human subjects satisfied the semantic contrast and naturalness as sentences because they were lined up with words that mainly referred to objects and people around them, and all of them satisfied the semantic contrast and naturalness as sentences. However, the correct answer was "mind," which has "physical and mental contrasts" in addition to perspectives of scale and familiarity. To deal with these cases, it is necessary to clarify from what perspective the two words are contrasted.

Some cases were difficult to predict both by the proposed method and human subjects. Such instances require prior knowledge and imaginations about objects mentioned in the text (advertisement targets in case of slogans), for example, "From lightness within [MASK] to lightness almost weightless," where the gold answer is "tolerance" for a *glass* product. It requires additional input information about the target of the sentence to deal with such cases.

4 Conclusion

In this study, we addressed the task of predicting antonyms within a context. We proposed methods for adapting BERT to antonym prediction, such as domain adaptation using an antithesis corpus, contrastive masking, antithesis positional encodings, and pseudo-supervision data collection. The proposed method achieved 29.3% top-1 and 53.8% top-10 accuracies on the test data. Although these values seem low, an automatic evaluation based on a single correct word underestimates the performance because multiple valid words can fill in the blanks. The subjective evaluation revealed that more than 85% of the words predicted by the proposed method were appropriate as antonyms. Our proposed task and method will be useful in many real-world applications that use contrastive expressions. Although we used Japanese text in this study, it can be applied to any language as far as the annotated data is available. In the future, we will extend the proposed method to generate text with antithesis, and explore the fill-in-the-blanks problem setting for other semantic relations.

References

Mitsuaki Aota, Akira Akiyama, Hideki Azuma, et al. 2007. *Saishinyaku copy bible (in Japanese) (English translation:The brand new slogan bible)*. Sendenkaigi Co., Ltd.

Oren Barkan, Avi Caciularu, and Ido Dagan. 2020. Within-between lexical relation classification. In

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 3521–3527, Online. Association for Computational Linguistics.

Jose Camacho-Collados, Claudio Delli Bovi, Luis Espinosa-Anke, Sergio Oramas, Tommaso Pasini, Enrico Santus, Vered Shwartz, Roberto Navigli, and Horacio Saggion. 2018. SemEval-2018 task 9: Hypernym discovery. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 712–724, New Orleans, Louisiana. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

John Heritage and David Greatbatch. 1986. Generating applause: A study of rhetoric and response at party political conferences. *American journal of sociology*, 92(1):110–157.

Christopher Hidey and Kathy McKeown. 2019. Fixed that for you: Generating contrastive claims with semantic edits. 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 1756–1767, Minneapolis, Minnesota. Association for Computational Linguistics.

Hristo Katrandjiev, Ivo Velinov, and Kalina Radova. 2016. Usage of rhetorical figures in advertising slogans. *Trakia Journal of Sciences*, 14(03):267–274.

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, Barcelona, Spain. Association for Computational Linguistics.

Geoffrey Leech. 1976. *Semantics*. Penguin Books.

Juncen Li, Robin Jia, He He, and Percy Liang. 2018. Delete, retrieve, generate: a simple approach to sentiment and style transfer. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1865–1874, New Orleans, Louisiana. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Takashi Nakahata. 2008. *Honto no koto wo iu to, yoku, shikarareru. katsu copy no zenbu (in Japanese) (English translation : We are often scolded when we say the truth. All of the slogans for winning.)*. Sendenkaigi Co., Ltd.

Laura Rimell, Amandla Mabona, Luana Bulat, and Douwe Kiela. 2017. Learning to negate adjectives with bilinear models. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 71–78.

Sanseido Editorial Office, editor. 2017. *Hantaigo Tairitsugo Dictionary (in Japanese) (English translation : Antonym Dictionary)*. Sanseido Co.,Ltd.

Sendenkaigi Award Committee. 2003–2018. *SKAT.2–SKAT.17*. Sendenkaigi Co., Ltd.

Vered Shwartz and Ido Dagan. 2016. Path-based vs. distributional information in recognizing lexical semantic relations. In *Proceedings of the 5th Workshop on Cognitive Aspects of the Lexicon (CogALex - V)*, pages 24–29, Osaka, Japan. The COLING 2016 Organizing Committee.

Masakazu Taniyama. 2007. *Koukoku copy tte kou kakunnda! dokuhon (in Japanese) (English translation : This is how you write advertising slogans! A textbook)*. Sendenkaigi Co., Ltd.

Satoshi Umeda. 2016. *"Kotoba ni dekiru" ha buki ni naru. (in Japanese) (English translation:The ability to "put into words" is a weapon.)*. Nikkei Publishing Inc.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi 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 Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Rui Yan, Cheng-Te Li, Xiaohua Hu, and Ming Zhang. 2016. Chinese couplet generation with neural network structures. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2347–2357.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems*, volume 32, pages 5753–5763. Curran Associates, Inc.

A Construction of an antithesis corpus

As described in Section 2.2, the annotation process for constructing the antithesis corpus was divided into the following two steps: (1) identification of candidate antitheses and (2) annotation of the span of the antithesis. In step (1), we assigned each slogan to five workers to determine whether the slogan contained an antithesis. If more than three workers determined that the slogan contained an antithesis, we would consider it as a candidate antithesis. Thus, we succeeded in extracting 9,720 slogans that contained antitheses. In step (2), we selected two workers with high-annotation quality and asked them to annotate each antithesis with its span, for example, "[A lean body] leads [a bold life]."

B Model architectures and implementation details

We used $\mathrm{BERT}_{\mathrm{BASE}}$, which has 12 layers, 768 hidden states, 12 heads, and 110M parameters for all the experiments. During the pre-training, the whole word masking was enabled. We used Mecab (Kudo et al., 2004) as the tokenizer.

Our implementation, including the code for evaluation, was based on Huggingface Transformers (Wolf et al., 2020). In fine-tuning BERT with the AdamW (Loshchilov and Hutter, 2019) optimizer, we set a batch size of 8, a maximum sequence length of 50, and the remaining parameters were set to the default values. The experiments were run on servers with an Nvidia Tesla P100 GPU. The total number of epochs for the fine-tuning was 6, determined by the accuracy of development data.