



Title	Utilizing BERT with Auxiliary Sentences Generation to Improve Accuracy of Japanese Aspect-based Sentiment Analysis Task
Author(s)	Zhang, Yiyang; Takeshita, Masashi; Rzepka, Rafal et al.
Relation	Human Language Technologies as a Challenge for Computer Science and Linguistics - 2023
Citation	Human Language Technologies as a Challenge for Computer Science and Linguistics, 2023, 348-352
Issue Date	2024-02-06
Doc URL	https://hdl.handle.net/2115/95944
Rights(URL)	https://creativecommons.org/licenses/by-nc-nd/4.0/
Type	conference paper
File Information	Proc. Hum. Lang. Technol. Chall. Comput. Sci. Linguist. 2023_348-352.pdf



Utilizing BERT with Auxiliary Sentences Generation to Improve Accuracy of Japanese Aspect-based Sentiment Analysis Task

Yiyang Zhang¹, Masashi Takeshita¹, Rafal Rzepka², Kenji Araki²

¹Graduate School of Information Science and Technology, Hokkaido University

²Faculty of Information Science and Technology, Hokkaido University

Email: {yiyang.zhang, takeshita.masashi, rzepka, araki}@ist.hokudai.ac.jp

Abstract

Aspect-based Sentiment Analysis (ABSA) is an sentiment analysis task to determine the polarity of opinions regarding a particular aspect of a text. ABSA is a fine-grained sentiment analysis that can handle more detailed and useful information, hence it continues to attract attention in the NLP field. Although many studies have focused on English as the target language, there are still few studies focused on Japanese. Therefore, in this study, we aim to improve the accuracy of the Aspect Extraction (AE) subtask and the Aspect Sentiment Classification (ASC) subtask of the Japanese ABSA task. We improved the method proposed for the English ABSA task, which is to automatically generate auxiliary sentences and then fine-tune the BERT with sentence pairs that combine target sentences and the generated auxiliary sentences, and applied it to the Japanese ABSA task. As a result of the performance evaluation experiment based on a Japanese ABSA dataset called chABSA, compared to the results of the previous study, the F1-score improved by 7.54 points in the AE subtask and the accuracy improved by 3.1 points in the ASC subtask, confirming the effectiveness of the proposed method.

Keywords: Aspect-Based Sentiment Analysis, Aspect Extraction, Aspect Category Detection, Aspect Sentiment Classification

1. Introduction

Aspect-Based Sentiment Analysis (ABSA) is the task that determines the sentiment polarity towards a specific aspect in a sentence. For example, in the sentence “The sushi at this restaurant is good, but the ramen is bad”, there are two aspects of the object “this restaurant”: “sushi” and “ramen”. Since the speaker evaluates “sushi” as “delicious,” he has a “positive” feeling toward the aspect of “sushi,” while he evaluates “ramen” as “bad,” he has a “negative” feeling toward the aspect of “ramen. ABSA is expected to be used in many situations because it can perform sentiment analysis at a finer granularity than document-level sentiment analysis and sentence-level sentiment analysis, and can provide a relatively large amount of information. For example, if a company wants to analyze how consumers feel about each aspect of a product that the company has released, it is necessary to conduct an aspect-level sentiment analysis. The ABSA task consists of two main subtasks: the first is Aspect Extraction (AE), which extracts aspects contained in sentences. The second task is Aspect Sentiment Classification (ASC), which identifies the sentiment polarity of each aspect. In this paper, we focus on both of these two subtasks in ABSA task.

While there have been many studies on ABSA with English as the target language, to the best of authors knowledge there are only three studies on ABSA with Japanese as the target language, and it can be expected to achieve higher accuracy by using better methods.

This study aims to achieve a higher accuracy in Aspect Extraction (AE) subtask and Aspect Sentiment Classification (ASC) subtask than previous ABSA studies using Japanese as the target language. We improved the method Sun et al. (2019) proposed for English, which automatically generates auxiliary sentences and fine-tuning the BERT model

with sentence pairs combining the target sentences and the auxiliary sentences, and applied it to Japanese. We conducted an experiment based on a Japanese ABSA dataset called chABSA. Eventually, in both AE and ASC subtasks of ABSA, we achieved better results than previous studies based on the same dataset. We also studied how to apply this method to the ASC subtask. This paper is based on our findings described in technical reports written in Japanese (Zhang et al., 2022) (Zhang et al., 2023).

2. Related work

Early ABSA research focused primarily on feature engineering (Wagner et al., 2014) (Kiritchenko et al., 2014) and deep learning techniques (Nguyen and Shirai, 2015) (Wang et al., 2016) (Tang et al., 2016) (Wang et al., 2017) (Ma et al., 2018) (Liu et al., 2018) suitable for ABSA task. Recently, pre-trained language models like ELMo (Peters et al., 2018), GPT (Radford et al., 2018), BERT (Devlin et al., 2019) have shown the ability to reduce the complexity of feature engineering. In particular, since BERT was pre-trained in the task of next sentence prediction, it performs well in tasks that require an understanding sentence-sentence relationships, such as the QA and NLI. Thus, Sun et al. (2019) proposed a method that utilizes this advantage of BERT and achieved new state-of-the-art results on English ABSA datasets.

In the studies using Japanese as the target language, Akai and Atsumi (2019) applied a neural network model using the self-attention mechanism used in a previous study using English as the target language to Japanese, and conducted experiments on sentences in the KNB corpus (Hashimoto et al., 2011) with emotion tags for reputation information. The accuracy of the sentiment classification reached 85%. Miura et al. (2020) proposed a self-attention neural network

model incorporating BERT, which consists of an aspect category classification network and an aspect sentiment analysis network. Experimental results show that the model achieves up to 85.6% accuracy in the AE subtask and up to 70.68% of F1-score in the ASC subtask.

However, there are two main problems in the aforementioned Miura et al. (2020) study using Japanese as the target language. In the Miura et al. (2020)'s method, each sentence is input directly into BERT for fine-tuning, which does not take advantage of the feature of BERT that allows it to perform better on sentence pair classification problems such as QA and NLI tasks. The second is the way to evaluate the performance of the model in the AE subtask. In the AE subtask of the Miura et al. (2020)'s study, the performance was evaluated by changing the dropout rate on test data, but we think it is more appropriate to find the dropout rate that achieves the best accuracy on validation data and evaluate the model performance using that dropout rate. We hypothesize that this method of performance evaluation is more appropriate.

In order to solve the above problems, this study aims to find a more suitable fine-tuning method for BERT, and evaluate the performance of the model in the AE subtask using a more appropriate performance evaluation method, and eventually achieve better results in the aspect category detection task, i.e., the AE subtask, and the ASC subtask achieved by Miura et al. (2020).

3. Propose method

In this study, we apply the method proposed by Sun et al. (2019), which automatically generates auxiliary sentences and then fine-tunes BERT with sentence pairs generated by combining the target sentence and the auxiliary sentence, which we think is a more suitable fine-tuning method for BERT, to Japanese. We improved the method of constructing auxiliary sentences in order to find a more suitable fine-tuning method for BERT than the original method, and because of some characteristics of Japanese, we also studied how to apply this method for the ASC subtask.

The following sections describe the BERT model, auxiliary sentence generation methods, and the BERT-pair model, which is fine-tuned by sentence pairs.

3.1. BERT

In this study, we conduct experiments using the pre-trained Japanese BERT-base model¹ provided by Tohoku University, which is pre-trained on the Japanese version of Wikipedia.

3.2. Methods of auxiliary sentences generation

3.2.1. Aspect Extraction (AE) subtask

The methods for generating auxiliary sentences in AE subtasks is shown in **Table 1**. Each of the auxiliary sentence generation methods is described below in detail.

- QA: In the Question Answering method, auxiliary sentences are generated in the form of questions, and

the system is asked to answer those questions. The specific method is to generate auxiliary sentences by combining the aspect category and the sentence “が含まれていますか” (Does it contain [aspect category]).

- NLI: Natural Language Inference method generates auxiliary sentences using only aspect categories.

Table 1: Method of auxiliary sentences generation in AE subtask

Method	Auxiliary sentence	Output
QA	(aspect category) が含まれていますか (Does it contain [aspect category])	Yes/No
NLI	(aspect category)	Yes/No

3.2.2. Aspect Sentiment Classification (ASC) subtask

The auxiliary sentence generation methods in the ASC subtask are shown in **Table 2**. Each of the auxiliary sentence generation methods is described below.

- QA-M: M stands for Multi, meaning that the classifications are not binary but multiple. This method generates auxiliary sentences in the form of questions, and then let the system to answer those questions. Specifically, the method is to generate auxiliary sentences by combining an aspect, i.e., the object whose sentiment polarity is to be judged, with the sentence “についてどう思いますか” (What do you think of [aspect]?).
- QA-M-Improve: Our improved auxiliary sentence construction method based on QA-M. The original QA-M auxiliary sentence construction method of Sun et al. (2019) sets the question to “What do you think of” (についてどう思いますか), but we believe that there is a more appropriate question sentence, and we assume that “に対する感情極性は何ですか” (What is the sentiment polarity towards) is a direct question about the sentiment polarity, and therefore, the BERT model can be guided to obtain more correct results and higher accuracy can be achieved.
- NLI-M: NLI means Natural Language Inference, and the meaning of M is same as QA-M. In this method, auxiliary sentences are generated using only the aspect.
- QA-B: Here, B stands for Binary, meaning that the class to be classified is binary (“Yes” and “No”). The method generates three auxiliary sentences by combining the aspect with three phrases, “に対する感情極性は肯定的” (the polarity of [aspect] is positive), “に対する感情極性は否定的” (the polarity of [aspect] is negative), and “に対する感情極性は中立的” (the polarity of [aspect] is neutral) respectively. When validating with test data, the sentence with the highest probability of “Yes” among the three sentences is selected, and the polarity indicated by that sentence is the polarity determined by the model. When expressing sentiment polarity, there are two types of expressions in Japanese: Kanji and Katakana. Therefore, in order to verify whether the Kanji expression: “肯定的” (positive), “否定的” (negative), “中立的” (neu-

¹<https://huggingface.co/cl-tohoku/bert-base-japanese-whole-word-masking>

tral) or the Katakana expression: “ポジティブ” (positive), “ネガティブ” (negative), “ニュートラル” (neutral) can construct auxiliary sentences that can achieve higher accuracy, we conducted an experiment by generating auxiliary sentences expressing sentiment polarity in each of them.

- NLI-B: This method generates three auxiliary sentences by combining the aspect and three phrases “肯定的” (positive), “否定的” (negative), and “中立的” (neutral) with a hyphen, respectively. Everything else is the same as QA-B.

Table 2: Method of auxiliary sentences generation in ASC subtask

Method	Auxiliary sentence	Output
QA-M	(aspect) についてどう思いますか (What do you think of [aspect])	Pos/Neg/Neu
QA-M-Improve	(aspect) に対する感情極性は何ですか (What is the sentiment polarity towards [aspect])	Pos/Neg/Neu
NLI-M	(aspect)	Pos/Neg/Neu
QA-B	• (aspect) に対する感情極性は肯定的/ポジティブだ (the polarity of [aspect] is positive) • (aspect) に対する感情極性は否定的/ネガティブだ (the polarity of [aspect] is negative) • (aspect) に対する感情極性は中立的/ニュートラルだ (the polarity of [aspect] is neutral)	Yes/No
NLI-B	• (aspect)-肯定的/ポジティブ ([aspect]-positive) • (aspect)-否定的/ネガティブ ([aspect]-negative) • (aspect)-中立的/ニュートラル ([aspect]-neutral)	Yes/No

3.3. BERT-pair model

After generating auxiliary sentences, for each target sentence, we combine the target sentence and the token “[SEP]” and the auxiliary sentence as a new sentence, and use the new sentences to fine-tune the BERT. The resulting BERT is then named with the name “BERT-pair” combined with the names of the respective auxiliary sentence generation methods.

4. Evaluation

This section describes the dataset, hyperparameters, evaluation index, and experimental results for the experiment in the AE subtask and the experiment in the ASC subtask.

4.1. Dataset

In this study, we use a dataset called chABSA². chABSA dataset is a dataset based on securities reports of listed companies (fiscal year 2016), and contains the sentiment polarity information of “Positive”, “Negative”, and “Neutral” for each aspect included in a sentence. Another ABSA dataset for Japanese is Rakuten Travel: Review Aspect and Sentiment Tagged Corpus (Nakayama et al., 2022) by Rakuten, but it is not used in this study because it is not publicly available for free.

Similar to Miura et al. (2020)’s study, this study deals with 15 aspect categories in this dataset, consisting of combinations of three entities (company, business, product) and five attribute (sales, profit, amount, price, cost) (e.g., company#sales, business#amount). However, since the number of aspects belonging to the aspect category “company#price” is zero, the actual number of aspect categories is 14, and the total amount of sentences is 1,077, the total number of aspects is 2,079.

4.1.1. Aspect Extraction (AE) subtask

Since the proposed method expands each sentence to 14 sentences depending on the number of aspect categories handled, the total number of data in the actual data set is 15,078 (14 times 1,077). First, this dataset is divided into training, validation, and test data in a 7:1:2 ratio for experimentation. After adjusting the dropout rate to achieve the highest accuracy on the validation data, the validation data is put into the training data for experimentation. Thus, the final experiment divides this data set into 12,062 training data and 3,016 validation data at a ratio of 4:1.

4.1.2. Aspect Sentiment Classification (ASC) subtask

In the proposed auxiliary sentence generation methods QA-M, QA-M-Improve, and NLI-M, one auxiliary sentence is generated for each aspect, resulting in 2,079 pieces of data, which are divided into 1,663 training data and 416 test data in a 4:1 ratio for the experiments. In contrast, the proposed auxiliary sentence generation methods QA-B and NLI-B generate three auxiliary sentences for each aspect, resulting in 6,237 sentences in total, which are divided into 4,989 training data and 1,248 test data at a ratio of 4:1 for the experiments.

4.2. Hyperparameters

4.2.1. Aspect Extraction (AE) subtask

In the AE subtask experiments, the hyperparameters were set as shown in Table 3.

Here, the dropout rate was adjusted using the validation data. Since the highest F1-score was obtained in the validation data when the dropout rate was set to 0 for all models, we set the dropout rate to 0.

4.2.2. Aspect Sentiment Classification (ASC) subtask

In the experiments in the ASC subtask, the hyperparameters were set almost the same as in the AE subtask except

²<https://github.com/chakki-works/chABSA-dataset>

Table 3: Hyperparameters in AE subtask

Hyperparameter	Value
Learning rate	2e-5
Epoch	3
Batch size	16
Dropout rate	0
random_state (pandas.DataFrame.sample)	4,040
Random seed (Pytorch)	2,020

for two differences. First, we set the dropout rate to 0.1. Second, we set the batch to 16 for the M-type models, i.e. the models whose output is multi, and to 48 (three times 16) for the B-type models, i.e. the models whose output is binary. The reason for this is that the auxiliary sentences generation methods with “B” generate three sentences for each sentence and each aspect contained in that sentence, corresponding to the three sentiment polarities, and we want BERT to learn these three sentences as one coherent unit.

4.3. Evaluation index

As in the study by Miura et al. (2020), in the AE subtask, we use the F1-score, which is the harmonic mean index of the precision and recall, and in the ASC subtask, we use the accuracy as the evaluation index.

4.4. Experimental results

4.4.1. Aspect Extraction (AE) subtask

The experimental results in the AE subtask are shown in **Table 4**. “BERT-pair-NLI”, fine-tuned with the auxiliary sentences generated by the NLI auxiliary sentence generation method, outperformed all the results achieved by the models of the existing methods and exceeded all the other proposed models, obtaining the highest precision, recall, and F1-score.

Table 4: Experimental results in the AE subtask. Underlined numbers indicate better results than in previous studies; bold numbers indicate the best results.

Model	Dropout	Precision	Recall	F1-Score
Miura et al. (2020)	0	0.7530	0.6531	0.6985
	0.2	0.7545	0.6667	0.7068
	0.5	0.7313	0.6737	0.7009
BERT-pair-QA	0	<u>0.8094</u>	<u>0.7282</u>	<u>0.7666</u>
BERT-pair-NLI	0	<u>0.8321</u>	<u>0.7379</u>	<u>0.7822</u>

4.4.2. Aspect Sentiment Classification (ASC) subtask

The experimental results in the ASC subtask are shown in **Table 5**. The “BERT-pair-QA-M-Improve” model, which we fine-tuned with auxiliary sentences generated by QA-M-Improve, an improved version of QA-M, outperformed the existing methods and achieved the highest accuracy among all methods.

The M-type models, i.e. the models whose output is multi, obtained a overall higher accuracy than the B-type models, i.e. the models whose output is binary. The models using Kanji expression obtained a higher accuracy than the models using Katakana expression.

Table 5: Experimental results in the ASC subtask. Katakana means expressing sentiment polarity in Katakana expression. Kanji means expressing sentiment polarity in Kanji expression. Underlined numbers indicate better results than in previous studies; bold numbers indicate the best results.

Model	Accuracy
Miura et al. (2020)	85.6
BERT-pair-QA-M	<u>87.0</u>
BERT-pair-QA-M-Improve	<u>88.7</u>
BERT-pair-NLI-M	<u>87.5</u>
BERT-pair-QA-B (Katakana)	<u>82.2</u>
BERT-pair-QA-B (Kanji)	82.9
BERT-pair-NLI-B (Katakana)	81.7
BERT-pair-NLI-B (Kanji)	82.0

5. Discussion

This section discusses the experimental results of the AE and ASC subtasks, respectively.

5.1. Aspect Extraction (AE) subtask

The experimental results confirm that the proposed method achieves better accuracy than the existing methods. The first reason is that the proposed method expands each sentence according to the number of aspect categories in the dataset (14 in this study), which expands the dataset, increases the number of sentences used for training, and increases the training data for the model. The second reason is that BERT has been pre-trained on the “next sentence prediction” task, which captures the sentence-sentence relationships better and thus performs better on sentence pair classification problems such as the QA and NLI tasks.

The reason why “BERT-pair-NLI” obtained higher F1-scores than “BERT-pair-QA” is still unknown and we will investigate this further in the future.

5.2. Aspect Sentiment Classification (ASC) subtask

The experimental results confirm that the proposed method achieves better accuracy than the existing methods. The reason for this is the same as in the AE subtask, and the only difference is that the proposed method expands each sentence according to the number of aspects in the dataset. Furthermore, the results showed that the models using the Kanji expression (“肯定的”, “否定的”, and “中立的”) achieved higher accuracy and higher F1-score than the models using the Katakana expression (“ポジティブ”, “ネガティブ”, and “ニュートラル”) when expressing sentiment polarity in auxiliary sentences. The reason for this is that Japanese BERT was pre-trained on the Japanese version of Wikipedia, and in this corpus, the Kanji expression are more often used to express sentiment polarity than the Katakana expression, therefore the model captures the meaning better with Kanji expression than Katakana expression when expressing sentiment polarity.

The reason the “BERT-pair-QA-M-Improve” model

achieved better results than the “BERT-pair-QA-M” model is probably that the auxiliary sentence the former model used asked a more direct question about the sentiment polarity, and thus induced BERT to yield more correct results. These results confirm that higher accuracy can be obtained by constructing more appropriate auxiliary sentences.

6. Conclusion

In this study, we improved the method originally proposed for English language, which automatically generates auxiliary sentences and fine-tuned the BERT with sentence pairs that combine target sentences and auxiliary sentences, to the aspect category detection subtask, i.e., the AE subtask, and the ASC subtask of the Japanese ABSA task. Experimental results on the chABSA dataset showed that the F1-score in the AE subtask improved by 7.54 points over the previous study, and the accuracy in the ASC subtask improved by 3.1 points over the previous study, confirming the effectiveness of the proposed method. Furthermore, in the experiment on the ASC subtask, the result that the accuracy of the “BERT-pair-QA-M-Improve” model is 1.7 points higher than the “BERT-pair-QA-M” model indicates that higher accuracy can be achieved if more appropriate auxiliary sentences are constructed. In the future, we plan to search for better methods to further improve the accuracy of the Japanese ABSA task.

References

- Akai, Ryuichi and Masayasu Atsumi, 2019. Application of aspect-based sentiment analysis using self-attention mechanism to Japanese sentences. In *The 33rd Annual Conference of the Japanese Society for Artificial Intelligence*, 2019.
- Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*.
- Hashimoto, Chikara, Sadao Kurohashi, Daisuke Kawahara, Keiji Shinzato, and Masaaki Nagata, 2011. Construction of a blog corpus with syntactic, anaphoric, and sentiment annotations. *Journal of Natural Language Processing*, 18(2):175–201.
- Kiritchenko, Svetlana, Xiaodan Zhu, Colin Cherry, and Saif Mohammad, 2014. Nrc-canada-2014: Detecting aspects and sentiment in customer reviews. In *Proceedings of the 8th international workshop on SemEval*.
- Liu, Fei, Trevor Cohn, and Timothy Baldwin, 2018. Recurrent entity networks with delayed memory update for targeted aspect-based sentiment analysis. In *Proceedings of NAACL-HLT*.
- Ma, Yukun, Haiyun Peng, and Erik Cambria, 2018. Targeted aspect-based sentiment analysis via embedding commonsense knowledge into an attentive lstm. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.
- Miura, Yoshihide, Ryuichi Akai, and Masayasu Atsumi, 2020. Self-attention neural network for sentiment analysis of multiple aspects in sentences. In *The 34th Annual Conference of the Japanese Society for Artificial Intelligence*, 2020.
- Nakayama, Yuki, Koji Murakami, Gautam Kumar, Sudha Bhingardive, and Ikuko Hardaway, 2022. A large-scale Japanese dataset for aspect-based sentiment analysis. In *Proceedings of the 13th LREC*.
- Nguyen, Thien Hai and Kiyooki Shirai, 2015. Phrasernn: Phrase recursive neural network for aspect-based sentiment analysis. In *Proceedings of the 2015 EMNLP*.
- Peters, Matthew E., Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer, 2018. Deep contextualized word representations. In *Proceedings of NAACL-HLT*.
- Radford, Alec, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al., 2018. Improving language understanding by generative pre-training.
- Sun, Chi, Luyao Huang, and Xipeng Qiu, 2019. Utilizing bert for aspect-based sentiment analysis via constructing auxiliary sentence. In *Proceedings of NAACL-HLT*.
- Tang, Duyu, Bing Qin, and Ting Liu, 2016. Aspect level sentiment classification with deep memory network. In *Proceedings of the 2016 EMNLP*.
- Wagner, Joachim, Piyush Arora, Santiago Cortes, Utsab Barman, Dasha Bogdanova, Jennifer Foster, and Lamia Tounsi, 2014. Dcu: Aspect-based polarity classification for semeval task 4. In *SemEval@ COLING*.
- Wang, Bo, Maria Liakata, Arkaitz Zubiaga, and Rob Procter, 2017. Tdparse: Multi-target-specific sentiment recognition on twitter. In *Proceedings of the 15th EACL: Volume 1, Long Papers*.
- Wang, Yequan, Minlie Huang, Xiaoyan Zhu, and Li Zhao, 2016. Attention-based lstm for aspect-level sentiment classification. In *Proceedings of the 2016 EMNLP*.
- Zhang, Yiyang, Rafal Rzepka, and Kenji Araki, 2022. Using bert and auxiliary sentences generation to improve accuracy of Japanese aspect-based sentiment analysis. In *70th Special Interest Group on Language Sense Processing Engineering (SIG-LSE)*. The Japanese Society for Artificial Intelligence.
- Zhang, Yiyang, Masashi Takeshita, Rafal Rzepka, and Kenji Araki, 2023. Utilizing bert with auxiliary sentences generation to improve accuracy of aspect category detection in Japanese aspect-based sentiment analysis. In *NLP2023*. The Association for Natural Language Processing.