Data Augmentation by Shuffling Phrases in Recognizing Textual Entailment

Kyosuke Takahagi and **Hiroyuki Shinnou**
Ibaraki University
{22nm730l,hiroyuki.shinnou.0828}@vc.ibaraki.ac.jp

Abstract

Data augmentation is a technique that aims to improve machine learning performance by increasing the number of training data. One method of data augmentation for Japanese sentences is to shuffle the order of phrases that compose a sentence while preserving the dependency relationships. This method has proven effective in improving the performance of text classification, especially when the training data is limited. In this study, we aimed to improve the performance of Recognizing Textual Entailment (RTE) using this method. RTE is recognized as a crucial technique for advancing natural language processing and is applied across various fields, including question-answering and machine translation. In the experiments, we addressed the RTE task using JSICK, a Japanese dataset, and pre-trained models, BERT and RoBERTa. The experimental results demonstrated that augmenting the training data with this method improved the performance of the models.

1 Introduction

Data augmentation is a technique that increases the number of training data by generating slightly modified copies of the existing data. When training a machine learning model, its performance can be improved through the utilization of data augmentation to increase the training data.

When performing data augmentation, it is crucial to generate natural data that does not negatively impact model training. When using data augmentation in supervised learning, it is typical to modify only the data without changing the labels. Consequently, the modified data must remain consistent with the labels of the original data. However, text data used in natural language processing (NLP) has a more intricate structure compared to image and audio data. As a result, data augmentation in NLP is prone to generating unnatural data when modifying the data, which may lead to inconsistencies with labels. Thus, effectively performing data augmentation proves challenging in NLP.

Despite these challenges, numerous effective data augmentation methods have been developed in NLP. We proposed the method named "data augmentation by shuffling phrases in a Japanese sentence"(Kyosuke Takahagi, 2022). The method involves modifying data by shuffling the order of phrases in a Japanese sentence. By considering the dependency relationship between phrases during the shuffling process, natural sentences that preserve the meaning of the original sentence can be generated (see Figure 1).

While the method has demonstrated effectiveness in text classification, its performance in other tasks remains unknown. In this study, we apply the method to Recognizing Textual Entailment (RTE). RTE is a task aimed at determining whether the relationship between two input sentences is entailment, contradiction, or neutral. It plays a crucial role in enabling computers to achieve natural language understanding. In this study, we examine whether augmenting the training dataset with the method improves the performance of an RTE model. The RTE task was addressed using JSICK(YANAKA and MINESHIMA, 2021), a Japanese dataset for RTE and Semantic Textual Similarity (STS), and pre-trained models, such as BERT (Bidirectional Encoder Representations from Transformers)(Devlin et al., 2019).

2 Related Work

Data augmentation has been extensively studied in the field of computer vision (CV). Various methods exist, including geometric transformations like image inversion and cropping, color space transformations, and image mixing(Shorten and Khoshgoftaar, 2019). On the other hand, research on data augmentation in NLP is often conducted secondarily after that in CV, and the number of studies on data augmentation is not as extensive as that

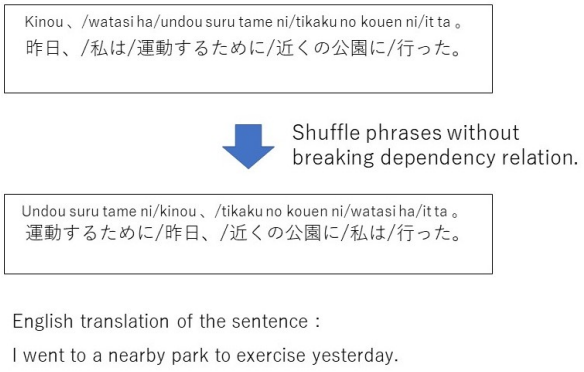

Figure 1: Example of shuffling the order of phrases in a Japanese sentence without breaking the dependency relationship.

of CV either. This is because the discrete nature of language and its complex semantic and syntactic structures make it difficult to modify text data while preserving labels. However, several effective methods have also been proposed in NLP that allow modification while preserving labels(Feng et al., 2021)(Chen et al., 2023).

Synonym replacement is a data augmentation technique commonly used in NLP. It involves replacing certain words in a sentence with their synonymous counterparts. These synonyms are identified based on similarities in a pre-defined dictionary or word embedding space. Additionally, techniques for locally modifying a sentence, such as deleting some words or swapping two words, can also alter the sentence while preserving its original meaning. Wei and Zou (2019) proposed Easy Data Augmentation (EDA), which combines these simple operations. Data Augmentation using back translation has also been extensively studied(Xia et al., 2019)(Chen et al., 2020). Back translation refers to the process of translating a sentence into another language and then back into the original language. It has proven to be an effective method for various tasks, especially in the domain of machine translation. Guo et al. (2019) applied Mixup(Zhang et al., 2018), a technique commonly used in the field of CV, and proposed two methods: one that mixes embedded representations of words (wordMixup) and another that mixes embedded representations of sentences (senMixup). In recent years, data augmentation using Large Language Models (LLMs) such as BERT and GPT-3 has also been extensively studied(Sahu et al., 2022)(Dai et al., 2023).

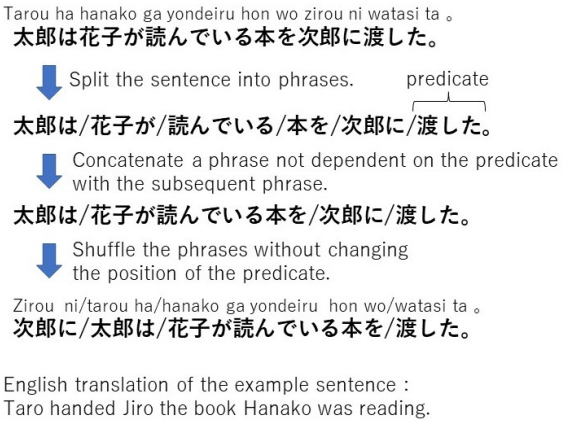

Figure 2: Process of modifying the example sentence.

3 Methods

3.1 Data Augmentation by Shuffling Phrases

In this study, we employ data augmentation method we proposed in the paper (Kyosuke Takahagi, 2022). The method involves reordering the position of a phrase in Japanese sentences while preserving the dependency relationships. The procedure is as follows:

1. Divide the sentence into phrases.
2. Connect phrases that do not modify the predicate (excluding the predicate itself) to the subsequent phrases.
3. Shuffle the order of phrases after concatenation (ensuring that the predicate position remains at the end of the sentence).

Figure 2 illustrates the process of performing the modification on the example sentence. To divide sentences into phrases and identify the dependency of each phrase, CaboCha, a Japanese dependency parser based on Support Vector Machines, is utilized. CaboCha splits Japanese sentences into phrases and provides detailed information about each phrase. The output result of the dependency parsing for the example sentence using CaboCha is depicted in Figure 3.

3.2 Application and effect on text classification

The data in a text classification dataset consists of text and labels representing the categories of that text. When using the method to augment the data, the text is only modified without changing the labels (by shuffling the phrases in each sentence that makes up the text).

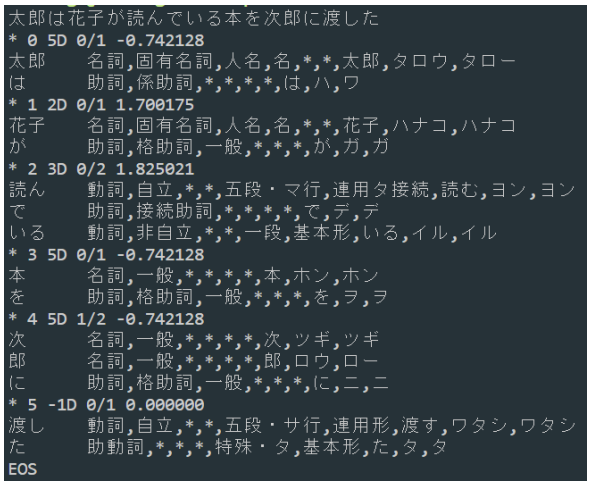

Figure 3: The result of parsing the example sentence by CaboCha

Table 1: Statistics on the dataset used for text classification.

Label	Category	Train	Val	Test
0	Dokujo Tsushin	10	100	100
1	IT Life Hack	10	100	100
2	Kaden Channel	10	100	100
3	livedoor HOMME	10	100	100
4	MOVIE ENTER	10	100	100
5	Peachy	10	100	100
6	Smax	10	100	100
7	Sports Watch	10	100	100
8	Topic News	10	100	100
	Total	90	900	900

In the paper (Kyosuke Takahagi, 2022), we conducted experiments to evaluate whether the method improves the classification performance in text classification using BERT. The experiments were verified using a small dataset created from the livedoor news corpus[1]. The dataset statistics are shown in Table 1. In the experiments, data augmentation is applied to the training data in this dataset, doubling its size by generating one augmented training data for each original training data.

Table 2 shows the evaluation results obtained using the test data. The evaluation metric is accuracy. As shown in Table 2, the model trained on the augmented training data achived a 1.7% higher accuracy compared to the model trained on the original training data. These results demonstrate the effectiveness of the method for text classification.

[1] https://www.rondhuit.com/download.html#ldcc

Table 2: Evaluation result of the method in text classification using BERT(%).

	Original	Augmented	Difference
Accuracy	73.4	75.1	+1.7

3.3 Application on RTE

The data in an RTE dataset generally consists of three parts: a premise sentence, a hypothesis sentence, and a label indicating the relationship between the two sentences. In this study, we generate three types of data for each original data: data obtained by modifying only the premise sentences, data obtained by modifying only the hypothesis sentences, and data obtained by modifying both the premise sentences and the hypothesis sentences. In all cases, the label remains unchanged.

4 Experiment Setup

4.1 Dataset

This study uses JSICK, a Japanese dataset for RTE and STS. JSICK is derived from the SICK (Sentences Involving Compositional Knowledge) dataset, which was originally in English(Marelli et al., 2014). However, it has been manually translated into Japanese, and the correct labels have been re-annotated to suit the Japanese context.

The definition of entailment relations in JSICK complies with the definition in the original SICK dataset. For each pair (T, H) consisting of a presuppositional sentence T and a hypothetical sentence H, it is assigned one of three labels: "entailment," "contradiction," or "neutral." Entailment refers to a relation in which sentence H is always true when sentence T is true. Contradiction is a relation in which sentence H is always false when sentence T is true. Lastly, neutral is a relation in which sentence H cannot be determined to be true or false when sentence T is true.

JSICK is pre-divided into training and test data, and 10% of the training data is reserved for validation purposes. Table 3 shows an example of the data contained in JSICK, while Table 4 shows the distribution of each data type. On average, there are 13.2 words per sentence, and the vocabulary consists of 2,432 unique words. Additionally, "neutral" accounts for about 60% of the entailment relation labels.

Table 3: Examples of data included in JSICK.

Pair of sentences(premise: T, hypothesis: H)	Entailment relation
T: A young woman is playing guitar. H: A girl is playing guitar.	Entailment
T: A tiger is walking out of a cage. H: A tiger is walking around inside a cage.	Contradiction
T: A man is cutting potatoes. H: A man is cutting tomatoes.	Neutral

Table 4: Number of each data in JSICK.

	Train	Validation	Test	Total
Entailment	991 (22.0%)	100 (20.0%)	1,088 (22.1%)	2,179 (22.0%)
Contradiction	748 (16.6%)	75 (15.0%)	797 (16.2%)	1,620 (16.3%)
Neutral	2761 (61.4%)	325 (65.0%)	3,042 (61.7%)	6,128 (61.7%)
Total	4,500	500	4,927	9,927

Table 5: Number of training data before and after the augmentation.

	Before	After
Entailment	991 (22.0%)	3,600 (22.1%)
Contradiction	748 (16.6%)	2,739 (16.8%)
Neutral	2761 (61.4%)	9,981 (61.2%)
Total	4,500	16,320

4.2 Augmentation of Training Data

Data augmentation is applied to the JSICK training dataset using the way described in Section 3.3. Certain sentences in the JSICK dataset are too short to be shuffled, which may lead to the generation of a sentence identical to the original sentence during the modification process. If the generated training data duplicates already existing training data, it is not used as training data. Table 5 shows the counts of training data before and after the augmentation process.

4.3 Models

In this experiment, we utilize BERT and RoBERTa, both of which are pre-trained models, as RTE models. Specifically, we employ "cl-tohoku/bert-base-japanese-v2"[2] (Tohoku-BERT), published by Inui Lab, Tohoku University, and "rinna/japanese-roberta-base"[3] (rinna-RoBERTa), published by rinna Co., Ltd.

[2] https://huggingface.co/cl-tohoku/bert-base-japanese-v2

[3] https://huggingface.co/rinna/japanese-roberta-base

To ensure consistency, we unify the hyperparameters for both models. The batch size is set to 8, and we employ stochastic gradient descent (SGD) as the optimization algorithm with a learning rate of 0.01. Additionally, we adopt the cross-entropy loss function. The number of epochs is determined using early stopping.

4.4 Evaluation

The performance of the trained RTE model is evaluated based on the accuracy of the test data. Additionally, we measure the accuracy for the specific classes of "entailment," "contradiction," and "neutral" in the test data.

For evaluation, we compare the performance of models trained on un-augmented training data with models trained on augmented training data. Five models are constructed for each scenario, and the average accuracy of the five models is used for evaluation.

5 Experimental result

In the experiments, we utilize two types of training data sets: JSICK's 4,500 training data and the augmented training data described in section 4.2. Models are separately fine-tuned on the two training data sets. By comparing the performance of the two models thus fine-tuned, we verified whether the method proposed in the paper(Kyosuke Takahagi, 2022) is effective for RTE. The details of the experiments are described in section 4.

Table 6 shows the results. When utilizing rinna-RoBERTa as the model, the application of data

augmentation to the training data resulted in a performance improvement of 0.8%. When examining the performance for each label, the performance for contradiction data increased by 3.4%. However, the performance for neutral data only showed a marginal increase. When utilizing Tohoku-BERT as the model, the application of data augmentation to the training data resulted in a performance improvement of 0.5%. When examining the performance for each label, the performance for contradiction data increased by 1.8%. However, the performance for entailment data showed a slight decrease.

6 Discussion

6.1 Factors that Improved the Performance of Contradiction Data

By augmenting the training data with the method proposed in the paper (Kyosuke Takahagi, 2022), the model's performance to recognize contradiction data showed a significant improvement compared to entailment and neutral data (see Section 5). This improvement may be attributed to the limited presence of contradiction data in the original training dataset. Generally, data augmentation techniques involving simple transformations, such as the method, tend to be more effective when the amount of available training data is limited. Conversely, when an adequate amount of training data is available for learning, data augmentation has little impact on improving model performance(Wei and Zou, 2019).

The JSICK dataset used in this study has the lowest proportion of contradiction data. Moreover, when examining the performance of the model trained with the original training data, the performance for contradiction data is the lowest among the three labels. In other words, the original training data did not provide sufficient contradiction data to effectively address the RTE task. As a result, the model's performance in recognizing contradiction relations was significantly improved by utilizing the method and increasing the amount of contradiction data.

6.2 Comparison with Other Methods

The method can generate natural sentences that preserve the meaning of the original sentence by modifying Japanese sentences in a manner that maintains dependency relations. However, it remains unclear whether this feature improves the performance of the RTE model. In this section, we compare the method with methods that shuffle the order of sentence components without considering the dependency relation. Through this comparison, we validate whether generating natural sentences that preserve the meaning of the original sentence contributes to improving the model's performance.

The two methods used for comparison with the method are a method that shuffles phrases without considering the dependency relations (random phrase shuffling) and a method that shuffles words (random word shuffling). Random phrase shuffling involves shuffling the order of phrases, except for predicates, after a sentence is segmented into phrases. On the other hand, random word shuffling involves shuffling the order of words after a sentence is segmented into words. Random phrase shuffling uses CaboCha for phrase segmentation, while random word shuffling uses MeCab(KUDO, 2005), a morphological analysis engine, for word segmentation. The evaluation of these two methods follows the process described in Section 4. However, only RoBERTa is employed as the RTE model. The experimental results are shown in Table 7, along with the results from Section 5.

As a result, the method achieved the highest performance, while word shuffling yielded lower performance compared to no augmentation. When examining the performance for each label, the method achieved the highest performance for both entailment and contradiction. However, for neutral labels, random shuffling outperformed the method. These findings suggest that data augmentation through sentence modification is more effective in improving model performance when generating natural sentences that preserve the meaning of the original sentences.

7 Conclusion

In this study, we validated the effectiveness of the method proposed in the paper (Kyosuke Takahagi, 2022), which involves shuffling phrases in Japanese sentences while preserving the dependency relations, for RTE. The experiment included the construction of two types of models: one utilizing the training data augmented by the method, and the other using the un-augmented training data. By comparing the performance of these two models, we confirmed whether the method improves the model's performance. RoBERTa and BERT were employed as models.

Table 6: Comparison of model performance with and without data augmentation(%). The evaluation metric is accuracy.

Model	Train data	All	Entailment	Contradiction	Neutral
rinna-RoBERTa	Original	89.0	85.4	77.9	93.2
	Augmented	89.8	86.0	81.4	93.4
	Difference	+0.8	+0.6	+3.4	+0.2
Tohuku-BERT	Original	88.5	85.3	81.2	91.6
	Augmented	89.0	84.9	83.0	92.1
	Difference	+0.5	-0.4	+1.8	+0.5

Table 7: Comparison of model performance when training data is augmented using each method (%). Models are RoBERTa. The evaluation metric is accuracy.

Augmentation method	All	Entailment	Contradiction	Neutral
No augmentation	89.0	85.4	77.9	93.2
The method	**89.8**	**86.0**	**81.4**	93.4
Random phrase shuffling	89.5	85.0	80.4	**93.5**
Random word shuffling	88.8	83.1	80.9	93.0

As a result, we confirmed that the method improved the model performance for both RoBERTa and BERT. We also evaluated the performance of the models for each label and observed that the performance for contradiction data was significantly improved.

In the future, our goal is to further refine the method to achieve more effective data augmentation techniques. Additionally, we aim to validate the effectiveness of the method for tasks other than text classification and RTE.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Number JP23K11212 and the NINJAL Collaborative Research Projects.

References

Jiaao Chen, Derek Tam, Colin Raffel, Mohit Bansal, and Diyi Yang. 2023. An Empirical Survey of Data Augmentation for Limited Data Learning in NLP. *Transactions of the Association for Computational Linguistics*, 11:191–211.

Jiaao Chen, Yuwei Wu, and Diyi Yang. 2020. Semi-supervised models via data augmentation for classifying interactive affective responses. In *AffCon@AAAI*.

Haixing Dai, Zheng Liu, Wenxiong Liao, Xiaoke Huang, Zihao Wu, Lin Zhao, Wei Liu, Ninghao Liu, Sheng Li, Dajiang Zhu, Hongmin Cai, Quanzheng Li, Dinggang Shen, Tianming Liu, and Xiang Li. 2023. Chataug: Leveraging chatgpt for text data augmentation. *ArXiv*, abs/2302.13007.

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.

Steven Y. Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. 2021. A survey of data augmentation approaches for NLP. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 968–988, Online. Association for Computational Linguistics.

Hongyu Guo, Yongyi Mao, and Richong Zhang. 2019. Augmenting data with mixup for sentence classification: An empirical study. *ArXiv*, abs/1905.08941.

T. KUDO. 2005. Mecab : Yet another part-of-speech and morphological analyzer. *http://mecab.sourceforge.net/*.

Hiroyuki Shinnou Kyosuke Takahagi. 2022. Data augmentation by shuffling phrases in a japanese sentence. In *IPSJ SIG Technical Report*, volume 2022-NL-252, pages 1 – 7.

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.

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. 2014. A SICK cure for the evaluation of compositional distributional semantic models. In

Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), pages 216–223, Reykjavik, Iceland. European Language Resources Association (ELRA).

Gaurav Sahu, Pau Rodriguez, Issam Laradji, Parmida Atighehchian, David Vazquez, and Dzmitry Bahdanau. 2022. Data augmentation for intent classification with off-the-shelf large language models. In *Proceedings of the 4th Workshop on NLP for Conversational AI*, pages 47–57, Dublin, Ireland. Association for Computational Linguistics.

Connor Shorten and Taghi M. Khoshgoftaar. 2019. A survey on image data augmentation for deep learning. *J. Big Data*, 6:60.

Yuji Matsumoto Taku Kudo. 2002. Japanese dependency analysis using cascaded chunking. In *CoNLL 2002: Proceedings of the 6th Conference on Natural Language Learning 2002 (COLING 2002 Post-Conference Workshops)*, pages 63–69.

Jason Wei and Kai Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6382–6388, Hong Kong, China. Association for Computational Linguistics.

Mengzhou Xia, Xiang Kong, Antonios Anastasopoulos, and Graham Neubig. 2019. Generalized data augmentation for low-resource translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5786–5796, Florence, Italy. Association for Computational Linguistics.

Hitomi YANAKA and Koji MINESHIMA. 2021. Jsick: Japanese sentences involving compositional knowledge dataset. *Proceedings of the Annual Conference of JSAI*, JSAI2021:4J3GS6f02–4J3GS6f02.

Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. 2018. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*.