



HOKKAIDO UNIVERSITY

Title	Development of Japanese WSC273 Winograd Schema Challenge Dataset and Comparison between Japanese and English BERT Baselines
Author(s)	Hashimoto, Ryo; 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, 91-95
Issue Date	2024-02-06
Doc URL	https://hdl.handle.net/2115/95943
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_91-95.pdf



Development of Japanese WSC273 Winograd Schema Challenge Dataset and Comparison between Japanese and English BERT Baselines

¹Ryo Hashimoto, ²Masashi Takeshita, ³Rafal Rzepka, ³Kenji Araki

¹School of Engineering, Hokkaido University

²Graduate School of Information Science and Technology, Hokkaido University

³Faculty of Information Science and Technology, Hokkaido University

{Ryo79676, takeshita.masashi.68}@gmail.com

{rzepka, araki}@ist.hokudai.ac.jp

Abstract

Winograd Schema Challenge (Levesque et al., 2012) has become a popular benchmark for natural language understanding and commonsense reasoning research for English language, and there are many related studies, but few studies have dealt with a similar benchmark in Japanese. In this study, we use the Winograd Schema Challenge dataset (WSC273) translated into Japanese (WSC273-ja) and propose a method to improve the commonsense reasoning ability of Japanese language models. After conducting the same experiment in English and comparing the results, we discuss the differences in commonsense reasoning ability between Japanese and English language models. Specifically, we first provide a baseline by evaluating a pre-trained language model using WSC273-ja in a zero-shot test. Next, we fine-tune the language models using WSCR-ja, a Japanese dataset of pronoun disambiguation problems similar to WSC273-ja. This WSCR-ja is a simpler but larger dataset than WSC273-ja. The model is then tested to see how well it can correctly answer the original WSC273-ja, which consists of more complex questions. The results are used to analyze the differences and similarities between questions with low correct response rates in Japanese and English, as well as to identify future issues to be addressed.

Keywords: Winograd Schema Challenge, Dataset creation, Japanese language

1. Introduction

In recent years, the emergence of pre-trained models with self-supervised learning on large-scale unlabelled datasets has dramatically improved model performance and achieved remarkable results in a variety of machine learning tasks. In the field of natural language processing, many pre-trained language models, including GPT (Radford et al., 2018) and BERT (Devlin et al., 2018), have emerged and continued to update the state-of-the-art in various tasks over the past few years. For example, in WSC273, a commonsense reasoning task, the accuracy was around 60% before the pre-trained language models were introduced, but in the current state-of-the-art (Sakaguchi et al., 2021), the accuracy is over 90%, which is almost the same level as that of humans. However, research on this topic is most often limited to English, because the lack of a complete Japanese WSC273 dataset, no study using WSC in this language has been conducted. In this paper, we evaluate the commonsense reasoning ability of the Japanese language BERT model by constructing a baseline using WSC273 translated into Japanese (hereafter referred to as WSC273-ja) and verifying whether the model can achieve higher accuracy on WSC273-ja by fine-tuning with a similar dataset in Japanese. As a result, it was confirmed that fine-tuning with similar data is also effective in the Japanese model. The structure of this experiment is described below. First, we evaluate the commonsense reasoning ability of BERT, which was pre-trained on a Japanese corpus, using WSC273-ja in a zero-shot test. Next, the same model is fine-tuned using the Japanese translation of WSCR (Shibata et al., 2015) (hereafter referred to as

WSCR-ja), and we test whether the scores improved when the original WSC273-ja is used for evaluation. The results show that Japanese scores are lower than English scores in the same experiment, both at baseline and after fine-tuning. At baseline, the accuracy in Japanese is 0.567, while in English it is 0.619. In the Japanese language model, the accuracy between baseline and after fine-tuning is improved from 0.567 to 0.578. These results confirm that fine-tuning with similar dataset is effective in improving commonsense reasoning ability of the Japanese BERT.

2. Background

2.1. Winograd Schema Challenge

The Winograd Schema Challenge is a pronoun disambiguation problem proposed by Levesque et al. (2012) as a more practical alternative to the Turing Test. The Turing Test is a method using dialogue system, but its high adaptability and flexibility made it possible to deceive the evaluator through various deceptions and tricks. To eliminate these undesirable effects, it is more effective to impose a task that is easy for humans but difficult for machines. The specific task of the Winograd Schema Challenge (WSC) is a reference resolution, in which two sentences, each with a few words different, are paired as shown in **Ex 1** below. Each of these sentences contains an anaphora and two candidate antecedents, and the system is asked to correctly identify the antecedent to which the anaphora corresponds. These questions are designed in such a way that human can easily identify the antecedent, but are difficult for systems using only selectional restrictions or statistical methods to answer correctly.

Ex 1. *John couldn't see the stage with Billy in front of him because he is so [short/tall].
Who is so [short/tall]?
Answers: John/Billy.*

Specifically, the WS must satisfy the following requirements. (Levesque et al., 2012)

1. Containing two antecedent candidates in a sentence
2. A pronoun or possessive adjective is used as the referent for one of the antecedents, but it is grammatically appropriate to use it for the other antecedent
3. The question is to determine the referent of the antecedent
4. It contains a word called, "special word", which, when replaced by another word called "alternate word", changes the antecedent to which the antecedent refers, although the sentence still makes sense.

In the case of **Ex 1**, the special word is "short" and the alternate word is "tall", and in each case, the antecedent pointed to by the anaphora "he" changes between "John" and "Billy". The fourth requirement makes the two sentence pairs in WSC statistically very similar to each other in terms of context, but the correct antecedents are different from each other. This prevents the system from immediately exploiting statistical features even if it has access to a large corpus. The key point of WSC is that it is extremely unlikely that there are statistical or other features of special or alternate words that can be reversed from one answer to another. In a WSC constructed in this way, background knowledge that does not appear in the context is needed to understand what is happening and to select an answer. Levesque et al. state that bringing this background knowledge is what thinking is all about.

The original Winograd Schema Challenge consisted of 137 pairs of sentences and a few sentences added later, and consisted of 284 questions, one for each pair of sentences, as shown in **Ex 1**. 273 of these questions (after excluding 11 of exceptional form) are widely used as the WSC273 evaluation set, which is also used in this study.

2.2. BERT

BERT uses a Masked Language Model (MLM) pre-training objective inspired by the cloze task (Taylor, 1953), which masks some tokens of the input by replacing them with [MASK] tokens and predicts the masked words based on context alone. In addition to MLM, BERT also uses Next Sentence Prediction (NSP), which jointly pre-trains text pairs, for pre-training purposes. The two unsupervised learning methods are used to pre-train each other. After pre-training with each of these two unsupervised learning methods, the model initialized with the parameters pre-trained here is fine-tuned with labeled data for downstream tasks. These fine-tuned models are initialized with the same pre-trained parameters, but after fine-tuning they become distinct models with different parameters for each task. In addition to the pre-training task, model size also has a significant impact on performance. Using the same hyperparameters and training procedure but varying the number of layers, hidden units, and attention heads, larger models

achieve higher accuracy.

3. Related Work

In recent years, as in case of other NLP tasks, methods using pre-trained, fine-tuned language models have dramatically improved the accuracy of WSC. However, WSC273 has only 273 examples. It is too small to fine-tune pre-trained model. WSCR (Rahman and Ng, 2012) was originally proposed to improve tasks that cannot be answered correctly by simple statistical methods such as WSC and consists of 941 sentence pairs. Each sentence is divided into a first half and a second half by a conjunction. The first half contains multiple candidate antecedents, and the second half contains an anaphora that refers to one of the candidate antecedents. Two sentences with the same first half, different second halves, and different antecedents to which the antecedent refers are paired. The dataset is about seven times larger than WSC273, but it is not strictly the same as WSC in that it does not necessarily require background knowledge not represented in the input sentences, and the conditions are slightly relaxed. In addition, Rahman and Ng's study (2012) did not evaluate WSC273 itself. Kocijan et al. (2019) were the first to use WSCR for fine-tuning BERT. They showed that fine-tuning BERT on the this dataset and a dataset generated from Wikipedia can robustly improve performance. Kocijan et al. updated the then state-of-the-art by fine-tuning BERT with WSCR and achieved an accuracy of 72.5%. The WSCR dataset used in this study consists of a training set of 1,316 sentences and a test set of 564 sentences, totaling 1,880 sentences, excluding duplications with the WSC273 dataset. The model that has achieved the highest accuracy in this task as of 2023 is the one proposed by Sakaguchi et al. (2021). This is a pre-trained language model based on BERT called RoBERTa (Liu et al., 2019), fine-tuned on a dataset called WinoGrande, which recorded 90.1% accuracy when evaluated using WSC273. So far, we have shown that fine-tuning of pre-trained language models with similar data is effective for WSC, and that state-of-the-art is reaching the human level. However, these are only studies on the original WSC273, and all the experiments were conducted in English. In our study, we conduct fine-tuning using WSCR-ja and the Japanese version of BERT, which was pre-trained by the Tohoku University¹, and evaluate WSC273-ja to verify whether the above architecture is also effective for Japanese, and compare the differences between languages.

4. Our Approach

4.1. Construction of Japanese WSC dataset

To begin the experiment, we first constructed a Japanese version of the WSC dataset. Originally, WSC273 was translated into Japanese by Language Media Laboratory at Hokkaido University (WSC273-ja), but due to a few grammatical errors and shortcomings, we made some corrections and additions following the original WSC273. We

¹<https://www.nlp.ecei.tohoku.ac.jp/news-release/3284/>

name the dataset WSC273-ja to match its English equivalent. In addition, to support input in BERT, we replaced the anaphora with [MASK] and reformatted it as shown in Ex 2.

Ex 2. Sentence:

じゃがいもの入った袋が、小麦粉の入った袋の上にあるので、最初に [MASK] を動かさなければならない。

The sack of potatoes had been placed above the bag of flour, so [MASK] had to be moved first.

Candidates : じゃがいもの入った袋, 小麦粉の入った袋

The sack of potatoes, The bag of flour

Answer : じゃがいもの入った袋 *The sack of potatoes*

4.2. Baseline construction by zero-shot

Next, a baseline is constructed by evaluating zero shots without fine-tuning. The model is BERT-large, which is published by Tohoku University. It follows the architecture of the English version of BERT-large, with 24 layers, 1024 hidden size, and 16 attention heads. The number of parameters is also 340M, as in the original BERT-large, but when comparing them, it should be noted that the size of the corpus used for pre-training is smaller in the Japanese version than in the English version. We experiment with the Japanese version of the corpus without changing the corpus size to match that of the English version. Furthermore, we tackle the WSC273-ja task with a model class called MASKedLM (MLM) in Japanese BERT as described above. This model class is used in the pre-training phase to predict the probability of [MASK] words. In this case, the process of replacing the anaphora with [MASK] was performed in advance. The candidate with the higher probability is taken as the predicted answer. We compared them with the correct answer and calculated its accuracy, which is used as the baseline. During tokenization, there may be multiple tokens for a candidate answer. In this case, [MASK] tokens are added to equal the number of tokens. Then, the harmonic mean of the probability of each of these multiple tokens being in [MASK] is taken as the probability of the candidate answer and compared with the probability of the other candidate answer. Similar experiments were conducted using the original WSC273 and BERT-large, as well as constructing an English version of the baseline.

4.3. Fine-tuning BERT with WSCR dataset

Finally, Japanese BERT is fine-tuned using the WSCR-ja dataset, and then evaluated using WSC273-ja. WSCR-ja follows Kocijan et al. (2019) and uses a training set of 1,316 sentences and a test set of 564 sentences, for a total of 1,880 sentences, like the original WSCR. The model used is the same as the one used for baseline construction in the previous section. However, in addition to MLM, a model class called MultipleChoice (MC) is used. This is a model class that predicts the probability of each candidate sentence following the previous sentence given a certain preamble and multiple candidate sentences. In this experiment, the problem sentences are separated before and after the [MASK] token, with the part before the appearance of

[MASK] as the pre-sentence and the latter part including [MASK] as the candidate sentence. Furthermore, two sentences are used as candidate sentences, with the [MASK] token in the candidate sentence replaced by each candidate answer. The probability of each of these two sentences following the pre-sentence is predicted, and fine-tuning is performed with the one with the larger probability as the predicted answer. The hyperparameters for fine-tuning are explored among a learning rate of 1e-5, batch size of {8, 16, 32, 64}, and number of epochs of {5, 10, 15, 30}. Like WSCR, MLM is also fine-tuned following Kocijan et al. (2019) with a learning rate of 5e-6, batch size of 64, and number of epochs of 30, and the loss function is varied as indicated in Eq 1 below. Note that c_1 and c_2 are candidate correct and incorrect answers, s is a training sentence, and $\mathbb{P}(c|s)$ represents their predicted probability. In this case, the hyperparameters α and β are 20 and 0.2, respectively.

Eq 1. $L = -\log \mathbb{P}(c_1|s) +$

$$\alpha \cdot \max(0, \log \mathbb{P}(c_2|s) - \log \mathbb{P}(c_1|s) + \beta)$$

As mentioned in the original BERT paper (Devlin et al., 2018), BERT-large tends to be unstable when fine-tuning on small datasets. Therefore, multiple experiments are conducted using random seeds, and the model with the best results is selected. In this study, each hyperparameter is fine-tuned three times, and the model with the highest accuracy on the WSCR-ja test set is selected as the proposed model among all models. As described in the previous section, we also conduct fine-tuning of the English version of the models using the original WSCR and BERT-large. These models are evaluated using WSC273-ja and WSC273, respectively. In addition, we construct WSCR-ja-small dataset, which excludes grammatically and culturally unnatural examples from WSCR-ja. The resulting WSCR-ja-small consists of 1,196 training sets and 530 test sets, for a total of 1,726 examples. Then, WSCR-small is constructed from the original WSCR, excluding the same examples that were excluded from WSCR-ja-small. By comparing the two, we analyze the impact of the above inappropriate examples on fine-tuning.

5. Evaluation

5.1. Baseline

The baseline accuracy was 0.567 in Japanese and 0.619 in English as shown in Table 1. This accuracy in English is identical to the one reported in the related study by Kocijan et al. (2019) described in Section 3.

	baseline	WSCR	WSC273
BERT-ja (MC)	0.567	0.719	0.575
BERT-ja (MLM)	0.567	0.728	0.578
BERT (MC)	0.619	0.820	0.688
BERT (MLM)	0.619	0.785	0.652

Table 1: Accuracy for each learning objective (WSCR-ja and WSCR)

	WSCR-small	WSC273
BERT-ja (MC)	0.700	0.578
BERT-ja (MLM)	0.718	0.578
BERT (MC)	0.811	0.673
BERT (MLM)	0.784	0.659

Table 2: Accuracy for each learning objective (WSCR-ja-small and WSCR-small)

BERT-ja	5	10	15	30
8	0.595	0.693	0.698	0.716
16	0.585	0.684	0.710	0.719
32	0.586	0.695	0.663	0.705
64	0.547	0.595	0.581	0.648
BERT	5	10	15	30
8	0.707	0.796	0.799	0.519
16	0.702	0.769	0.804	0.820
32	0.542	0.744	0.794	0.808
64	0.553	0.647	0.684	0.803

Table 3: Accuracy of the WSCR test set for each hyperparameter of MC model, where the first column represents the batch size and the column headers the number of epochs. Each value is selected as the highest among the experiments with three different seeds (learning rates are all 1e-5)

5.2. Adaptation by fine-tuning

The results of fine-tuning with WSCR or WSCR-small are shown in Tables 1 and 2, respectively. The values in the tables represent the accuracy when evaluating the test set and WSC273 dataset for each model. The accuracies of the test set at each hyperparameter when fine-tuning MC model with the WSCR training set are shown in Table 3. The top of Table 3 shows the results for Japanese BERT and the bottom for the original English BERT. The first column indicates the batch size, and column headers shows the number of epochs. These are the highest accuracy results from the three different seeding experiments. The model with the highest accuracy (0.719 in Japanese and 0.820 in English) is the one with batch size 16 and number of epochs 30 in both Japanese and English, and was selected as the proposed model for MC. As shown in Table 1, the result for WSC273 using the proposed model is 0.575 for Japanese, while for English it is 0.688. The MLM used the hyperparameters described in the previous section, and three experiments were conducted with different seeds, and the one with the highest accuracy on the test set was used as the proposed model. Fine-tuning with WSCR-small was performed for both MC and MLM using the same hyperparameters as the above proposed model. The results are shown in Table 2.

6. Discussion

The accuracy of the model at baseline and after fine-tuning is 0.567 and 0.578 for the Japanese model, respectively, while it is 0.619 and 0.688 for the English model. The results show that the accuracy of the Japanese model is lower than that of the English model both at baseline and after

fine-tuning. There is a large gap between the accuracies on the WSCR test set and on the WSC273 for both the Japanese (0.728 vs. 0.578) and English (0.820 vs. 0.688) models. In addition, although both are improved by fine-tuning, the accuracy of Japanese is not as good as the accuracy of English. In this section, we discuss the results from two viewpoints: the influence of the model and the influence of the dataset. Based on these discussion, we propose some perspectives for future research.

6.1. The influence of the model

As mentioned in Section 4, Japanese BERT was created following BERT, so the architecture and number of parameters are the same as the original BERT. However, the size of the corpus used for pre-training Japanese BERT is only about 1/4 of the original. We believe that this difference in corpus size is one of the reasons why Japanese BERT was lower than BERT in the final score. This is also mentioned in Shibata et al.’s study (2019), which shows that the larger the corpus size, the better the performance of the model. One reason for the difference in results between Japanese and English seems to be the pre-trained models themselves. It is not clear at this point whether this is due to the corpus size alone, or whether the pre-training of the language model in a language other than English has some other effects. The first step is to build and compare pre-trained models in English and Japanese, under several conditions including matched corpus size. Another possible reason for the difference in results between Japanese and English is linguistic differences. This is discussed in detail in the next subsection.

Next, we compare the model after fine-tuning with the existing studies: as already mentioned in Section 5, the baseline results obtained in this study for English BERT is the same as the value shown in the related study by Kocijan et al. (2019) We believe that this is a reasonable baseline for Japanese BERT, which was evaluated under the same conditions (except for the corpus size in the pre-training phase). However, the score after fine-tuning is only 0.688 for the MC model in this study, while Kocijan et al. recorded 0.714 using only WSCR. The same MLM model as Kocijan et al. only reaches a score of about 0.652. In this experiment, we used Hugging Face Transformer library, but it is not clear if this is the reason why we could not reproduce the results of Kocijan et al. It is necessary to continue to investigate the causes through, for example, the search for hyperparameters. On the other hand, in the Japanese model, a slight increase in accuracy was observed for both the test set and WSC273-ja dataset when MLM was used compared to MC. This suggests that MLM may be more compatible with the Japanese WSC than MC. Finally, as for the change in performance depending on the hyperparameters, the accuracy improves with the number of epochs in both the MC models, and the batch size of 16 appears to be most optimal.

6.2. The influence of the dataset

In the previous section, we looked at the difference between the Japanese and English results mainly in terms of the in-

fluence of the model, but in this section, we will discuss it in terms of the influence of the dataset. After analyzing both sets, we have not observed any particular differences due to differences between languages, however it should be noted that in Japanese subject is often omitted, which sometimes led to creating redundant sentences for masking. As shown above, there is a clear difference in performance improvement between the Japanese and English models after fine-tuning. At baseline, the accuracy of Japanese is 0.567 and that of English is 0.619, with a difference of 0.052. After fine-tuning, the accuracy of Japanese is 0.578 and that of English is 0.688, with a difference of 0.110. The difference is even larger than when comparing the two baselines. This may be due to the WSCR-ja training set. As noted in the original paper (Shibata et al., 2015), the English WSCR contains inappropriate examples. In addition to these examples, WSCR-ja is known to contain several examples that are inappropriate due to the translation of English into Japanese. These examples include words that are common in English texts but do not usually appear in Japanese texts due to cultural differences. Therefore, we created WSCR-small by removing those inappropriate examples (120 in the training data and 34 in the test data) and observed the change in adaptation to WSC, as shown in Table 2. As can be seen from the table, despite the reduction in the size of the training data set, the accuracy of WSC273 remains the same or even improves for the Japanese BERT. Combined with the decrease in accuracy in the test set, it appears that the removal of inappropriate examples from the WSCR data set has made the Japanese BERT more adaptable to the WSC273. This can be explained by the fact that the accuracy of the test set and WSC273 decreased in the English MC model. However, in the English MLM model, there was a slight improvement (from 0.652 to 0.659) in the accuracy of WSC273, suggesting that some of the removed cases were inappropriate before translation. In addition to these problems, there is also issue of WSC itself being vulnerable, as described by Sakaguchi et al. (2021) To begin with, WSC was designed so that it could not be solved by simple statistical methods, but as technology has advanced, it is no longer necessarily a problem that cannot be solved by statistical methods alone. However, as Levesque et al. noted in their original paper (2012), statistical methods have not been completely ruled out.

7. Conclusion

In this study, we constructed a baseline by evaluating the Japanese WSC dataset with zero-shot using BERT, which was pre-trained on a corpus in Japanese.

We also conducted fine-tuning of Japanese BERT with WSCR-ja, a large similar dataset of WSC translated into Japanese, and confirmed an improvement in accuracy from 0.567 to 0.578. This indicates that fine-tuning with Japanese WSCR is effective in improving the common sense reasoning ability of Japanese BERT, and that the Japanese model is not as well adapted to WSC as the English model. Furthermore, we demonstrated that the Japanese model is more adaptable to WSC by removing

inappropriate examples from the Japanese WSCR. Future work includes further cross-language comparisons using different models, testing whether background knowledge can be used more clearly by incorporating graphs, etc., and improving the Japanese dataset.

References

- Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Kocijan, Vid, Ana-Maria Cretu, Oana-Maria Camburu, Yordan Yordanov, and Thomas Lukasiewicz, 2019. A surprisingly robust trick for the winograd schema challenge. In Anna Korhonen and David Traum (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28 - August 2, 2019*. Association for Computational Linguistics.
- Levesque, Hector, Ernest Davis, and Leora Morgenstern, 2012. The Winograd Schema Challenge. In *Thirteenth international conference on the principles of knowledge representation and reasoning*.
- Liu, Yinhan, 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. *arXiv preprint arXiv:1907.11692*.
- Radford, Alec, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al., 2018. Improving language understanding by generative pre-training. *Technical Report. OpenAI*.
- Rahman, Altaf and Vincent Ng, 2012. Resolving complex cases of definite pronouns: the Winograd Schema Challenge. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*.
- Sakaguchi, Keisuke, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi, 2021. WinoGrande: An adversarial Winograd Schema Challenge at scale. *Communications of the ACM*, 64(9):99–106.
- Shibata, Tomohide, Daisuke Kawahara, and Sadao Kurohashi, 2019. Improved accuracy of Japanese parsing with BERT in Japanese. *Proceedings of the Twenty-fifth Annual Meeting of the Association for Natural Language Processing*:205–208.
- Shibata, Tomohide, Shotaro Kohama, and Sadao Kurohashi, 2015. Construction and analysis of the Japanese Winograd Schema Challenge in Japanese. *Proceedings of the Twenty-first Annual Meeting of the Association for Natural Language Processing*:493–496.
- Taylor, Wilson L, 1953. “Cloze procedure”: A new tool for measuring readability. *Journalism quarterly*, 30(4):415–433.