Large Pre-trained Language Models with Multilingual Prompt for Japanese Natural Language Tasks

Haiyue Song[1,2] Raj Dabre[2] Chenhui Chu[1] Sadao Kurohashi[1]
[1]Kyoto University [2]NICT
{song,chu,kuro}@nlp.ist.i.kyoto-u.ac.jp raj.dabre@nict.go.jp

Abstract

Pre-trained Language Models (PLMs) with in-context learning have achieved impressive performance on various English Natural Language Understanding (NLU) and generation tasks. However, applying PLMs to languages other than English is still a challenge. This is because the training data used for pre-training contains a huge percentage of English data and a significantly lower percentage of data in other languages. To alleviate this problem, we propose a multilingual prompt approach, where we provide the input in the target language as well as in English, the latter of which is obtained by Neural Machine Translation (NMT). We experimented on six Japanese datasets and achieved SOTA performance in two of them.

1 Introduction

In recent years, the world knowledge from the huge-scale training data and the generalization ability from the huge-scale model have enabled the PLMs to show promising performance on a wide range of Natural Language Processing (NLP) tasks [1, 2, 3]. This data-driven approach has especially shown promising results on English tasks [4]. It even shows better than the fine-tuning methods with a few-shot in-context learning setting [5]. Although the high generalization ability enables the PLMs to deal with NLP tasks in other languages, a performance gap exists between them and the same tasks in English. This phenomenon is largely due to the lack of training data in the target language [3, 6].

To alleviate the data distribution mismatch problem between the training data and testing data, increasing the percentage of non-English data in the training set is a trivial yet efficient approach. However, the disadvantages include the high cost of data collection and cleaning, the data scarcity for low-resource languages, and the drop in performance of

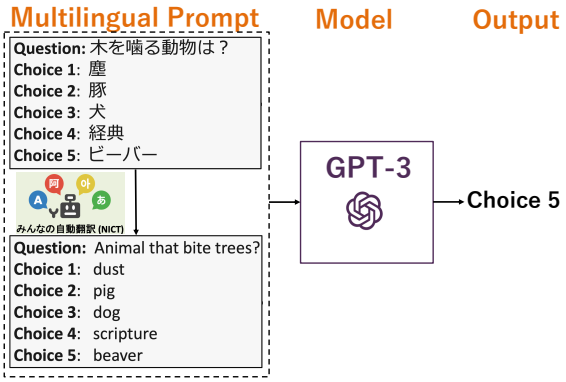

Figure 1: Overview of the proposed method.

English tasks [6]. Having English play the role of the intermediary is another effective way. This includes translating the non-English inputs into English [7], translating the English training data to the target language [8], or showing a few English examples as context to adjust the domain [4].

In this work, we solve the data distribution mismatch problem in PLMs with a novel multilingual prompt approach (See Figure 1). We first translate the input in the target language into English and feed them into PLMs. In this way, the prompt provides both accurate information in the original input and information in English that can be processed by the model without language mismatch.

We conducted experiments on the JGLUE benchmark that contains 5 language understanding tasks [9] and on the KWDLC dataset [10] for the Japanese word segmentation task. Compared with fine-tuning methods based on BERT variants, the proposed multilingual prompt approach based on a GPT-3 model achieves SOTA performance on the sentiment classification and Japanese word segmentation datasets and comparable performance on most of the rest datasets. Ablation studies show the importance of using multilingual rather than Japanese only or English only inputs. Analysis with examples further reveals how the English references help.

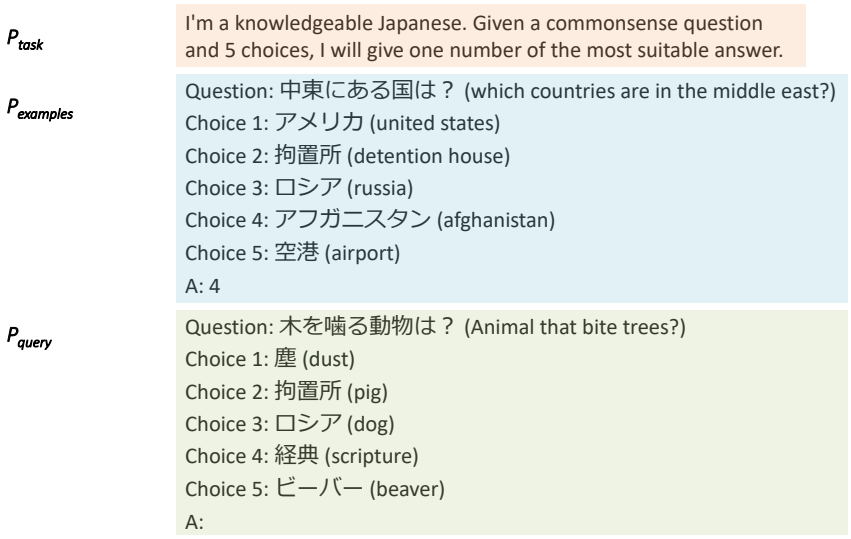

Figure 2: Prompt format with multilingual examples and the query.

2 Related Works

As the scales of PLMs increase (e.g., BERT series [1, 11], T5 model [2], and the GPT series [12, 4]), few-shot in-context learning without back-propagation to update the model parameters becomes possible [13]. The in-context learning approach makes it extremely practical, where the model can be applied to entirely new tasks requiring only a minimal amount of annotation.

To solve the non-English NLP tasks, multilingual PLMs approaches (e.g., mBERT [1], mT5 [14], and XGLM [6]) attempt to collect more balanced data from more than one hundrad languages and solve tasks in language other than English directly. Other attempts to alleviate the data distribution mismatch problem of PLMs pre-trained on unbalanced data include fine-tuning based methods [15, 16], translation-based methods [7, 8], and zero-shot methods [4], which require less computation. Multilingual input is also a common method in multilingual MT [17, 18].

3 Methods

We first briefly define the notations of the task, prompt, and output. We then explain the proposed multilingual prompt with a few-shot in-context learning method.

3.1 Notations

We define the parameter in the PLM as θ. For each task T, we define the dataset as D. The prompt of the PLM is defined as P, where $P = [P_{task}, P_{examples}, P_{query}]$ is a concatenation of three parts as presented in Figure 2.

- **The task explanation part** P_{task} provides the information of the task and the format of input and output of each data examples. It varies with the task.
- **The example part** $P_{examples}$ provides N examples. Each example E_i consists of two parts, the question part Q_i and the correct answer A_i.
- **The query part** P_{query} provides the question q at the inference time without the answer.

We call a setting zero-shot if the $P_{examples}$ is empty ($N = 0$), few-shot if N is a small number. We keep $N < 10$ in all few-shot experiments. The output is defined as y; for different tasks T, the set of possible output values differs.

3.2 Multilingual Prompt

We keep the target language Japanese across the experiments. For all the texts t_{Ja} in $P_{examples}$ and P_{query}, we apply a high-quality machine translation tool Textra[1)] to translate it into a English text t_{En}. We then combine t_{Ja} and t_{En} to form a multilingual text t_{Mix}. We then replace t_{Ja} with t_{Mix} to generate the multilingual $P_{examples}$ and P_{query} as shown in Figure 2.

In the zero-shot multilingual prompt setting, the output of the PLM $y = f(P|\theta)$ tends to be multilingual, and we only keep the Japanese part as a prediction. In the few-shot setting, the output will be in the correct format following the $P_{examples}$. Note that we use English for P_{task} across all the experiments.

1) textra.nict.go.jp

Table 1: Results on the JGLUE benchmark. **Bold** represents the best performance except for Human. Blue represents the best setting among the proposed methods.

Dataset	MARC-Ja	JSTS		JNLI	JSQuAD		JCommonsenseQA
Metrics	Acc	Pearson	Spearman	Acc	EM	F1	Acc
Baselines:							
Tohoku BERT large	0.955	0.913	0.872	0.900	0.880	0.946	0.816
NICT BERT base	0.958	0.910	0.871	0.902	0.897	0.947	0.823
XLM RoBERTa large	0.964	0.918	0.884	0.919	-	-	0.840
Waseda RoBERTa large (s128)	0.954	**0.930**	**0.896**	0.924	0.884	0.940	**0.907**
Waseda RoBERTa large (s512)	0.961	0.926	0.892	**0.926**	**0.918**	**0.963**	0.891
Proposed:							
Zero-shot Japanese	0.969	0.817	0.730	0.503	0.813	-	0.847
Zero-shot English	0.947	0.646	0.515	0.369	0.383	-	0.788
Zero-shot Multilingual	0.969	0.841	0.784	0.480	0.808	-	0.861
Few-shot Japanese	0.974	0.817	0.808	0.572	0.866	0.941	0.898
Few-shot English	0.969	0.836	0.819	0.495	0.484[2)]	-	0.822
Few-shot Multilingual	**0.975**	0.834	0.824	0.625	0.859	0.942	0.885
Human	0.989	0.899	0.861	0.925	0.871	0.944	0.986

Table 2: Results on the KWDLC dataset for the Japanese word segmentation task. **Bold** means the best performance.

Metrics	Precision	Recall	F1
Baselines:			
CRF+BERT [19]	0.618	0.653	0.635
Proposed:			
Few-shot Japanese	0.884	**0.868**	0.874
Few-shot English	0.530	0.504	0.511
Few-shot Multilingual	**0.886**	0.867	**0.875**
References:			
MeCab [20]	0.802	0.840	0.819
Kytea [21]	0.654	0.769	0.705

4 Experimental Settings

4.1 Datasets

We use five datasets from the JGLUE benchmark [9], including 1) MARC-ja, a text classification dataset, 2) JSTS, the Japanese version of the semantic textual similarity dataset, 3) JNLI, the Japanese version of the natural language inference dataset, 4) JSQuAD, the Japanese version of reading comprehension dataset, and 5) JCommonsenseQA, the Japanese version multiple-choice commonsense Question Answering (QA) dataset.

Additionally, we add a Japanese word segmentation task to verify whether the English translation will still help for the task that requires Japanese syntax information processing ability. We test on the KWDLC dataset [10], which is segmented by Jumanpp [22] and revised by experts.

The evaluation metrics keep the same with the previous work [9, 19], using accuracy for multi-choice tasks, Pearson and Spearman for the ranking task, Exact Match (EM) and F1 for the QA task, precision, recall, and F1 for the word segmentation task.

4.2 Model Settings

PLM We run experiments on a publicly available pretrained language model GPT-3 Codex (175B)[3)] with *temperature* as 0, *top p* as 1, *frequency penalty* as 0, *presence penalty* as 0, and *max tokens* as 200.

MT Model We use a publicly available MT tool TexTra[4)] with the general Japanese to English model.

Few-shot We manually select $N < 10$ examples from the train set that cover all types of output labels. For the MARC-Ja dataset, we use 2 examples with positive labels and 2 with negative labels. For the JSTS dataset, we use 10 examples with labels from 0.0 to 5.0 in a roughly uniform distribution. We use 6 examples for JNLI, 5 examples for JSQuAD, 5 for JCommonsenseQA, and 7 for KWDLC. For the JSQuAD dataset, in the case that there are multiple correct labels, we show the first one in $P_{examples}$.

2) Compared with the translated answers.

3) beta.openai.com/docs/models/codex

4) mt-auto-minhon-mlt.ucri.jgn-x.jp

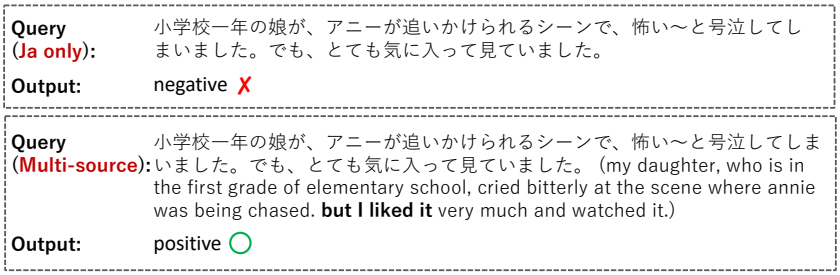

Figure 3: An example from the MARC-Ja dataset where multilingual prompt helps.

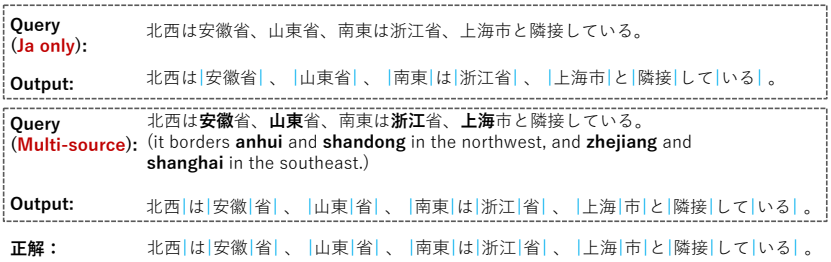

Figure 4: An example from the KWDLC dataset where multilingual prompt helps.

Baselines For the JGLUE benchmark, we compare with the fully supervised methods fine-tuned on Tohoku BERT,[5] NICT BERT base,[6] Waseda RoBERTa,[7] and XLM RoBERTa models [23]. For the KWDLC dataset, we compare it with a previous unsupervised method based on BERT [19]. We added existing tools MeCab [20] with ipadic dictionary and KyTea [21] as references. Note that the KWDLC is originally segmented by Jumanpp [22] therefore the F1 score using Jumanpp is near 1.0.

5 Experimental Results

5.1 Main Results

Tables 1 and 2 show the results of the JGLUE benchmark and KWDLC dataset correspondingly. The proposed methods show SOTA performance on the text classification dataset and Japanese word segmentation dataset. The performance is also comparable to the fine-tuned methods on two QA datasets. However, we found that the performance is worse on the tasks JNLI and JSTS that require reasoning ability. We can observe that the fine-tuned methods outperform humans on these two tasks.

The few-shot methods gave higher scores than the zero-shot ones on all the tasks, showing the effectiveness of in-context learning. We found the improvement from both better output format and adaptation to the task domain.

The multilingual prompt outperforms the Japanese or English only prompt on most tasks. We assume that it is due to less distribution mismatch between the testing and training data.

5.2 Case Analysis

We gave two cases where the multilingual prompt leads to correct prediction. Figure 3 provides a text classification example, where the object is implicit and omitted in the Japanese text. However, in the translated text, the phrase *but I liked it* is completed, and the object information becomes explicit, which helps the model to give the correct prediction. Figure 4 illustrates another case from the word segmentation task. The names of the province (Anhui, Shandong. . .) are given as separate English words in the translation, which provides word boundary information and helps to segment Japanese words correctly.

6 Conclusion and Future Work

In this work, we proposed a multilingual prompt approach to better apply the PLMs to non-English tasks. The prompt contains the original input in Japanese and the translated input in English, which alleviates the data distribution mismatch problem. We experimented on six Japanese datasets and achieved SOTA performance on two of them. Future work includes adding Chain-of-though (CoT) [5], a cross-lingual version of the few-shot retrieval method [24], and a web-searching method [25].

5) huggingface.co/cl-tohoku/bert-base-japanese-v2
6) alaginrc.nict.go.jp/nict-bert/index.html
7) huggingface.co/nlp-waseda/roberta-base-japanese

Acknowledgement

This work was supported by JSPS KAKENHI Grant Number JP21J23124. Part of the work was done during an internship at NICT.

References

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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)**, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[2] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. **Journal of Machine Learning Research**, Vol. 21, No. 140, pp. 1–67, 2020.

[3] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020.

[4] Genta Indra Winata, Andrea Madotto, Zhaojiang Lin, Rosanne Liu, Jason Yosinski, and Pascale Fung. Language models are few-shot multilingual learners. In **Proceedings of the 1st Workshop on Multilingual Representation Learning**, pp. 1–15, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[5] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models, 2022.

[6] Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, Veselin Stoyanov, and Xian Li. Few-shot learning with multilingual language models, 2021.

[7] Tim Isbister, Fredrik Carlsson, and Magnus Sahlgren. Should we stop training more monolingual models, and simply use machine translation instead?, 2021.

[8] 吉越卓見, 河原大輔, 黒橋禎夫. 機械翻訳を用いた自然言語推論データセットの多言語化. 第 244 回自然言語処理研究会, 2020.7.3.

[9] Kentaro Kurihara, Daisuke Kawahara, and Tomohide Shibata. JGLUE: Japanese general language understanding evaluation. In **Proceedings of the Thirteenth Language Resources and Evaluation Conference**, pp. 2957–2966, Marseille, France, June 2022. European Language Resources Association.

[10] 萩行正嗣, 河原大輔, 黒橋禎夫. 多様な文書の書き始めに対する意味関係タグ付きコーパスの構築とその分析. 自然言語処理, Vol. 21, No. 2, pp. 213–247, 2014.

[11] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019.

[12] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

[13] Damai Dai, Yutao Sun, Li Dong, Yaru Hao, Zhifang Sui, and Furu Wei. Why can gpt learn in-context? language models secretly perform gradient descent as meta-optimizers, 2022.

[14] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mT5: A massively multilingual pre-trained text-to-text transformer. In **Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 483–498, Online, June 2021. Association for Computational Linguistics.

[15] Bo Zheng, Li Dong, Shaohan Huang, Wenhui Wang, Zewen Chi, Saksham Singhal, Wanxiang Che, Ting Liu, Xia Song, and Furu Wei. Consistency regularization for cross-lingual fine-tuning. In **Proceedings of the 59th Annual Meeting of the ACL and the 11th IJCNLP (Volume 1: Long Papers)**, pp. 3403–3417, Online, August 2021. Association for Computational Linguistics.

[16] Zhaojiang Lin, Andrea Madotto, Genta Indra Winata, Peng Xu, Feijun Jiang, Yuxiang Hu, Chen Shi, and Pascale Fung. Bitod: A bilingual multi-domain dataset for task-oriented dialogue modeling, 2021.

[17] Raj Dabre, Fabien Cromieres, and Sadao Kurohashi. Enabling multi-source neural machine translation by concatenating source sentences in multiple languages. In **Proceedings of Machine Translation Summit XVI: Research Track**, pp. 96–107, Nagoya Japan, September 18 – September 22 2017.

[18] Barret Zoph and Kevin Knight. Multi-source neural translation. In **Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 30–34, San Diego, California, June 2016. Association for Computational Linguistics.

[19] 和田有輝也, 村脇有吾, 黒橋禎夫. セミマルコフ crf 自己符号化器による教師なし単語分割. 言語処理学会 第 28 回年次大会, 浜松, 2022.3.16.

[20] Takumitsu Kudo. Mecab : Yet another part-of-speech and morphological analyzer. 2005.

[21] Graham Neubig, Yosuke Nakata, and Shinsuke Mori. Pointwise prediction for robust, adaptable Japanese morphological analysis. In **Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies**, pp. 529–533, Portland, Oregon, USA, June 2011. Association for Computational Linguistics.

[22] Arseny Tolmachev, Daisuke Kawahara, and Sadao Kurohashi. Juman++: A morphological analysis toolkit for scriptio continua. In **Proceedings of the 2018 Conference on EMNLP: System Demonstrations**, pp. 54–59, Brussels, Belgium, November 2018. Association for Computational Linguistics.

[23] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. **CoRR**, Vol. abs/1911.02116, , 2019.

[24] Ohad Rubin, Jonathan Herzig, and Jonathan Berant. Learning to retrieve prompts for in-context learning. In **Proceedings of the 2022 Conference of the NAACL:HLT**, pp. 2655–2671, Seattle, United States, July 2022. Association for Computational Linguistics.

[25] Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. Webgpt: Browser-assisted question-answering with human feedback, 2021.