Rakuten's Participation in WMT 2024 Patent Translation Task

Ohnmar Htun and **Alberto Poncelas**
Rakuten Institute of Technology
Rakuten Group, Inc.
{ohnmar.htun alberto.poncelas}@rakuten.com

Abstract

This paper introduces our machine translation system (team *sakura*), developed for the 2024 WMT Patent Translation Task. Our system focuses on translations between Japanese-English, Japanese-Korean, and Japanese-Chinese. As large language models have shown good results for various natural language processing tasks, we have adopted the *RakutenAI-7B-chat* model, which has demonstrated effectiveness in English and Japanese. We fine-tune this model with patent-domain parallel texts and translate using multiple prompts.

1 Introduction

Machine Translation (MT) systems are becoming increasingly important in the translation industry. While generic MT models are good at translating common phrases into everyday language, they often struggle with specialized domains unless they have been specifically tuned for those areas. Patent documents are an example of this specialized content.

The patent translation shared task[1] at Conference on Machine Translation (WMT) 2024 aims to bring together Natural Language Processing (NLP) researchers to assess and explore innovative methods for translating patents, specifically between Japanese (Ja) and English (En), Korean (Ko) or Chinese (Zh), and vice versa.

Recently, significant advancements have been made in the field of NLP due to the development of Large Language Models (LLMs). Unlike encoder-decoder models, which are typically created to perform a single task such as machine translation, LLMs are designed for multiple NLP purposes. As a result, LLMs are often pre-trained on larger and more diverse texts, which helps improve the model's language understanding. In our work, we propose using an LLM fine-tuned with parallel data to perform accurate translations in the patent domain.

An LLM that has been specifically adapted to multiple NLP tasks in both English and Japanese is the *RakutenAI-7B* (Rakuten Group, Inc. et al., 2024) model. It has been pre-trained on a large volume of data, and its tokenizer has been optimized for the character-per-token rate in Japanese, making it ideal for complex tasks such as Japanese translation.

We participated in the patent translation (*sakura* team) shared task. In our proposal, we fine-tune the LLM with patent-domain bilingual data to build a multilingual model that achieves high-quality translations in multiple language directions. In addition, we produce translations using multiple prompts to further boost performance.

2 Related Work

LLM has been explored in the patent industry for tasks such as claim generation (Jiang et al., 2024), Question-Answer, or Classification (Bai et al., 2024).

Regarding the patent translation, previous participants in the JPO shared task have explored various methodologies, including training encoder-decoder models as suggested by Park and Lee (2021), utilizing Transformer-based NMT model (Vaswani et al., 2017) with ensemble decoding (Susanto et al., 2019) and adapting pre-trained models such as BART (Lewis et al., 2020) mBART (Liu et al., 2020) with patent-specific data (Wang and Htun, 2020; Kim and Komachi, 2021).

3 Task Description

The shared task consists of translating a set of sentences from patent publications in the En ↔ Ja, Ko ↔ Ja and Zh ↔ Ja language directions. The

[1] https://www2.statmt.org/wmt24/patent-task.html

Proceedings of the Ninth Conference on Machine Translation, pages 643–646

text belongs to the domains of Chemistry, Electricity, Mechanical Engineering or Physics.

These sentences, are organized as different test set according to the year the patents were published:

- *test-n1*: Published between 2011 and 2013 (same test sets used in the past years).
- *test-n2*: Published between 2016 and 2017 (not available for Ko-Ja).
- *test-n3*: Published between 2016 and 2017 (but target sentences were manually created by translating source sentences).
- *test-n4*: Published between 2019 and 2020.
- *test-2022*: The union of the previous n1 to n4 sets.

The *test-n1* to *test-n4* vary in size from 2K to 5K sentences depending on the language. The only exception is *test-n3*, which was created manually and contains between 200 and 700 sentences. The total size of these tests, i.e. *test-2022*, ranges from 7K to 10K sentences.

3.1 Evaluation

In order to determine the performance of our model, we submit the translation of the test sets mentioned above. The results of the different tasks are published in https://lotus.kuee.kyoto-u.ac.jp/WAT/evaluation/index.html.

The translations are tokenized using Juman (Kurohashi and Kawahara, 2009), KyTea[2], Mecab (Kudo, 2005) or Moses tokenizer[3]. The website presents multiple evaluation metrics for evaluating the translation. In this paper we present only the BLEU (Papineni et al., 2002) scores. The other metrics are correlated with BLEU. We refer to their website for the rest of the metrics.

3.2 Training Data

The organizers of the shared task also provide the JPO Patent Corpus (JPC) for training. This is a dataset built by the Japan Patent Office[4] consisting of sets of 1M parallel sentences for each language pair (English-Japanese, Chinese-Japanese and Korean-Japanese).

[2] https://www.phontron.com/kytea/

[3] https://github.com/moses-smt/mosesdecoder/blob/RELEASE-2.1.1/scripts/tokenizer/tokenizer.perl

[4] https://www.jpo.go.jp/index.htm

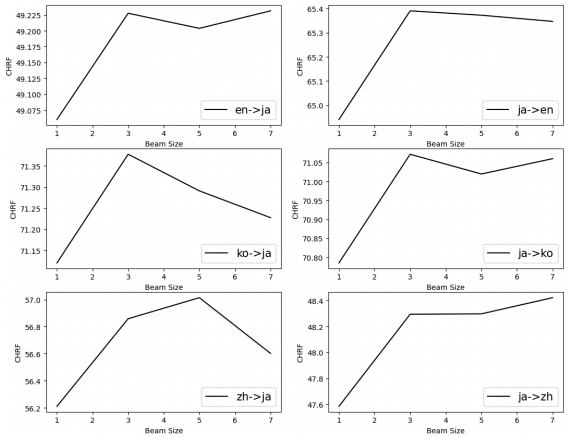

Figure 1: Performance of the fine-tuned model on the dev set using different beam sizes for decoding.

The data also include a dev set in the same domain with around 2K sentences each.

4 Experimental Settings

For our experiments, we fine-tune the *RakutenAI-7B-chat*[5] model, which has been optimized for the English and Japanese languages. However, it has not been explicitly adapted for other languages such as Korean and Chinese.

We use the JPO data described in Section 3.2 for this fine-tuning and do not incorporate any additional data other than what has been provided by the shared task organizers. The training process involves 200K steps with a batch size of 8. We fine-tune the model using the prompt "*Translate the following English text to Japanese:*" appending the sentence to be translated and replacing the source and target languages as needed for each language direction.

4.1 Influence of Beam Size

For decoding, we chose a beam size of three. While larger beam sizes involve considering more candidate translations, this does not always result in better performance. We tested our fine-tuned model with beam sizes of 1, 3, 5, and 7 on the development set. The results, measured using CHRF (Popović, 2015) metric, are shown in Figure 1. Although there is no single optimal beam size, our findings indicate that increasing the beam size beyond three does not lead to significant improvements and in some cases it may even degrade performance.

[5] https://huggingface.co/Rakuten/RakutenAI-7B-chat

Test	Direction	BLEU	Δ
test-2022	En → Ja	53.4	+4.5
	Ja → En	50.1	+5.6
test-n1	En → Ja	51.1	+5.8
	Ja → En	49.3	+5.2
test-n2	En → Ja	46.3	+5.7
	Ja → En	43.9	+6.2
test-n3	En → Ja	54.9	+7.4
	Ja → En	43.1	+8.1
test-n4	En → Ja	62.1	+1.6
	Ja → En	59.7	+4.8

Table 1: BLEU scores for Japanese-English translation (using Moses tokenizer for Ja → En and kytea for En → Ja). The column Δ indicates the difference between the scores of our model and those of the organizers.

Test	Direction	BLEU	Δ
test-2022	Zh → Ja	56.6	+5.5
	Ja → Zh	46.2	+1.5
test-n1	Zh → Ja	53.4	+6.7
	Ja → Zh	41.7	+2.6
test-n2	Zh → Ja	51.3	+5.3
	Ja → Zh	40.6	+1.5
test-n3	Zh → Ja	21.8	+4.0
	Ja → Zh	27.0	+3.2
test-n4	Zh → Ja	68.7	+3.7
	Ja → Zh	58.7	+1.2

Table 2: BLEU scores for Japanese-Chinese translation (using Kytea tokenizer). The column Δ indicates the difference between the scores of our model and those of the organizers.

Test	Direction	BLEU	Δ
test-2022	Ko → Ja	74.3	+0.4
	Ja → Ko	75.4	+2.6
test-n1	Ko → Ja	73.3	+1.6
	Ja → Ko	72.9	+2.2
test-n3	Ko → Ja	52.3	+0.3
	Ja → Ko	68.0	+5.6
test-n4	Ko → Ja	77.3	+0.2
	Ja → Ko	78.4	+3.7

Table 3: BLEU scores for Japanese-Korean translation (using Mecab tokenizer). The column Δ indicates the difference between the scores of our model and those of the organizers.

4.2 Influence of the Prompt

At decoding time, we perform multiple translations using different variations of the prompt. The prompts are the following:

- *Translate the following English text to Japanese* (same as training data)
- *Translate the following English sentence to Japanese:* (replace "text" with "sentence")
- *Translate the following text to Japanese:* (omit the source language)
- *Translate the text to Japanese:* (above prompt rephrased)
- *Translate the following English patent text to Japanese:* (explicitly indicate that is a patent text)

We use LASER (Heffernan et al., 2022) scores compared to the source to retrieve the best translation. Although all of them are similar, there are small nuances that can increase the quality by around 0.5-1 BLEU points.

5 Results

In this section we present the translation performance achieve by our model on the different language directions.

5.1 Japanese-English

First, Table 1 illustrates the performance of our model on English-Japanese translation. We observe that our model achieves the best results for this language pair compared to the model of the organizers, with an average improvement of 5 BLEU points for English-to-Japanese and 6 BLEU points for Japanese-to-English. Furthermore, it shows greater improvements in this pair compared to the other language pairs. This success can be attributed to the fact that our model was pre-trained on these two languages, benefiting from higher exposure.

5.2 Japanese-Chinese

Table 2 presents the results for Chinese-Japanese translation. Although improvements are observed across all test sets, there is a notable disparity between the language directions. While Zh → Ja shows an improvement of 5 BLEU points, for the reverse direction there is an improvement of 1.5 BLEU points.

5.3 Japanese-Korean

Lastly, in Table 3 we show the performance of the Japanese-Korean translation. For this language pair we achieve smaller improvements when compared to the baseline of the organizers.

6 Conclusion

In this paper, we described our MT model developed for the 2024 WMT Patent Translation Task, specifically for English-Japanese, Japanese-Korean, and Japanese-Chinese translations. The ranking system has evaluated participating teams every year from 2016 to 2024. Our model achieved first place in 20 out of the 28 tasks without using external data. Our approach involves fine-tuning the "*RakutenAI-7B-chat*" model using sentences from the patent domain and decoding with multiple prompts. Although this model was originally pre-trained only on English and Japanese data, fine-tuning with Korean and Chinese text has led to good translation performance, surpassing the models submitted in previous years for the same task.

References

Zilong Bai, Ruiji Zhang, Linqing Chen, Qijun Cai, Yuan Zhong, Cong Wang Yan Fang, Jie Fang, Jing Sun, Weikuan Wang, Lizhi Zhou, et al. 2024. PatentGPT: A Large Language Model for Intellectual Property. *arXiv preprint arXiv:2404.18255*.

Kevin Heffernan, Onur Çelebi, and Holger Schwenk. 2022. Bitext Mining Using Distilled Sentence Representations for Low-Resource Languages. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2101–2112, Abu Dhabi, United Arab Emirates.

Lekang Jiang, Caiqi Zhang, Pascal A Scherz, and Stephan Goetz. 2024. Can Large Language Models Generate High-quality Patent Claims? *Preprint*, arXiv:2406.19465.

Hwichan Kim and Mamoru Komachi. 2021. TMU NMT system with Japanese BART for the patent task of WAT 2021. In *Proceedings of the 8th Workshop on Asian Translation*, pages 133–137, Bangkok, Thailand.

Taku Kudo. 2005. Mecab: Yet another part-of-speech and morphological analyzer. *https://taku910.github.io/mecab/*.

Sadao Kurohashi and Daisuke Kawahara. 2009. Japanese Morphological Analysis System JUMAN 7.0 Users Manual. *http://nlp.ist.i.kyoto-u.ac.jp*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual Denoising Pre-training for Neural Machine Translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA.

Heesoo Park and Dongjun Lee. 2021. Bering Lab's Submissions on WAT 2021 Shared Task. In *Proceedings of the 8th Workshop on Asian Translation*, pages 141–145, Bangkok, Thailand.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal.

Rakuten Group, Inc., Aaron Levine, Connie Huang, Chenguang Wang, Eduardo Batista, Ewa Szymanska, Hongyi Ding, Hou Wei Chou, Jean-François Pessiot, Johanes Effendi, Justin Chiu, Kai Torben Ohlhus, Karan Chopra, Keiji Shinzato, Koji Murakami, Lee Xiong, Lei Chen, Maki Kubota, Maksim Tkachenko, Miroku Lee, Naoki Takahashi, Prathyusha Jwalapuram, Ryutaro Tatsushima, Saurabh Jain, Sunil Kumar Yadav, Ting Cai, Wei-Te Chen, Yandi Xia, Yuki Nakayama, and Yutaka Higashiyama. 2024. RakutenAI-7B: Extending Large Language Models for Japanese. *Preprint*, arXiv:2403.15484.

Raymond Hendy Susanto, Ohnmar Htun, and Liling Tan. 2019. Sarah's Participation in WAT 2019. In *Proceedings of the 6th Workshop on Asian Translation*, pages 152–158, Hong Kong, China.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. In *Advances in Neural Information Processing Systems*, volume 30.

Dongzhe Wang and Ohnmar Htun. 2020. Goku's Participation in WAT 2020. In *Proceedings of the 7th Workshop on Asian Translation*, pages 135–141, Suzhou, China.