Results of the WAT/WMT 2024 Shared Task on Patent Translation

Shohei Higashiyama
National Institute of Information and Communications Technology, Japan
shohei.higashiyama@nict.go.jp

Abstract

This paper presents the results of the patent translation shared task at the 11th Workshop on Asian Translation and 9th Conference on Machine Translation. Two teams participated in this task, and their submitted translation results for one or more of the six language directions were automatically and manually evaluated. The evaluation results demonstrate the strong performance of large language model-based systems from both participants.

1 Introduction

The patent translation task using the JPO Patent Corpus has been held under the Workshop on Asian Translation (WAT) in 2015–2023 (Nakazawa et al., 2023) and under the Conference on Machine Translation (WMT) this year.[1] Due to the high demand for patent translation, this task has attracted many participants particularly in the early WAT workshops: a total of 30 teams over the past 10 years as in Table 1.

This year, two teams participated in this task; one participant submitted translation results for two language directions, and the other for six out of six language directions, that is, Chinese↔Japanese, Korean↔Japanese, and English↔Japanese. Both teams employed large language model (LLM)-based systems, and the submitted translation results were evaluated by both automatic and human evaluation metrics. In this paper, we describe the evaluation dataset and procedure, and report the evaluation results for the submitted outputs.

2 Dataset

The JPO Patent Corpus (JPC)[2] was constructed by the Japan Patent Office (JPO) in collaboration with National Institute of Information and Communications Technology (NICT). The corpus consists of Chinese-Japanese (zh-ja), Korean-Japanese (ko-ja), and English-Japanese (en-ja) parallel sentences of patent descriptions. Most sentences were extracted from documents with one of four International Patent Classification sections: chemistry, electricity, mechanical engineering, and physics. As shown in Table 2, the dataset for each language pair consists of training, development, development-test, and multiple test sets. These datasets were constructed from patent families using automatic sentence alignment (Utiyama and Isahara, 2007), except for the test-N4 set where target sentences were manual translated from the source sentences.

A characteristic of the corpus is the use of fixed training and test datasets over the years, which allows for the comparison of new systems with past systems. The possible issue of data leakage is minimized: the data is provided only to applicants who have committed to participating in each annual workshop, and participants are required to delete the data after the workshop concludes.

Year	# of teams
2015	8
2016	6
2017	4
2018	2
2019	3
2020	2
2021	3
2022	0
2023	0
2024	2
Total	30

Table 1: The number of participant teams for the patent task over the years.

[1] Similarly to other WAT shared tasks, this task is organized by WAT organizers but is held under WMT this year due to the collaboration between the workshop and conference.

[2] https://lotus.kuee.kyoto-u.ac.jp/WAT/patent/

Proceedings of the Ninth Conference on Machine Translation (WMT), pages 118–123
November 15–16, 2024. ©2024 Association for Computational Linguistics

Set	# of Sentences zh-ja	 ko-ja	 en-ja	Published Years	Introduced Event
Train	1,000,000	1,000,000	1,000,000	2011–2013	WAT 2015–2016
Dev	2,000	2,000	2,000	2011–2013	WAT 2015–2016
DevTest	2,000	2,000	2,000	2011–2013	WAT 2015–2016
Test-N1	2,000	2,000	2,000	2011–2013	WAT 2015–2016
Test-N2	3,000	–	3,000	2016–2017	WAT 2018
Test-N3	204	230	668	2016–2017	WAT 2018
Test-N4	5,000	5,000	5,000	2019–2020	WAT 2022
Test-2022	10,204	7,230	10,668	2011–2020	WAT 2022

Table 2: Statistics of the JPO Corpus. The published years column represents the years for the source sentences. The introduced event column indicates the events for which each dataset was first introduced.

3 Evaluation Procedure

3.1 Automatic Evaluation

Task participants were required to submit translation results via the WAT Submission site.[3] For the results submitted with the "publish" checkbox selected, automatic evaluation scores were calculated and displayed in the WAT Evaluation site.[4] As the automatic evaluation metrics, we used BLEU (Papineni et al., 2002) with `multi-bleu.perl` in the Moses toolkit (Koehn et al., 2007) version 2.1.1[5] and RIBES (Isozaki et al., 2010) with `RIBES.py` version 1.02.4.[6]

Prior to calculating scores, reference sentences and output translation sentences were tokenized with the tokenization tools for each language: Juman 7.0 (Kurohashi et al., 1994), KyTea 0.4.6 (Neubig et al., 2011) with the full SVM model[7] and MeCab 0.996 (Kudo et al., 2004) with IPA dictionary 2.7.0[8] for Japanese, KyTea 0.4.6 with the full SVM Model (MSR model) and Stanford Word Segmenter (Tseng, 2005) version 2014-06-16 with the CTB and PKU models[9] for Chinese, mecab-ko[10] for Korean, and `tokenizer.perl`[11] in the Moses toolkit for English. The detailed procedures are shown on the WAT Evaluation site.[12]

5	All important information is transmitted correctly. (100%)
4	Almost all important information is transmitted correctly. (80%–)
3	More than half of important information is transmitted correctly. (50%–)
2	Some of important information is transmitted correctly. (20%–)
1	Almost all important information is NOT transmitted correctly. (–20%)

Table 3: Ratings and their descriptions in the JPO adequacy criterion.

3.2 Human Evaluation

We conducted human evaluation for selected translation results based on the JPO adequacy evaluation criterion, which is originally defined by JPO to assess the quality of translated patent documents. For this evaluation, we used the test-N3 set for each language direction for the following reasons: (1) parallel sentences have been manually aligned (translations were manually created from the original sentences), and (2) both participants submitted results for this test set.

The evaluation was performed by two annotators (translation experts) for each system as follows. First, 200 sentences for evaluation were randomly selected from the test-N3 set in advance (the same 200 sentences were used for all systems). (2) The 200 pairs of the source sentences and translated sentences by the system were shown to each annotator, and the ratings between 1 and 5 were assigned to each sentence by the annotator as shown in Table 3.

[3] https://lotus.kuee.kyoto-u.ac.jp/WAT/submission/index.php

[4] https://lotus.kuee.kyoto-u.ac.jp/WAT/evaluation/index.html

[5] https://github.com/moses-smt/mosesdecoder/tree/RELEASE-2.1.1

[6] http://www.kecl.ntt.co.jp/icl/lirg/ribes/index.html

[7] http://www.phontron.com/kytea/model.html

[8] http://code.google.com/p/mecab/downloads/detail?name=mecab-ipadic-2.7.0-20070801.tar.gz

[9] http://nlp.stanford.edu/software/segmenter.shtml

[10] https://bitbucket.org/eunjeon/mecab-ko/

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

[12] http://lotus.kuee.kyoto-u.ac.jp/WAT/evaluation/index.html

Parameter	Value
encoder_type	brnn
brnn_merge	concat
src_seq_length	150
tgt_seq_length	150
src_vocab_size	100,000
tgt_vocab_size	100,000
src_words_min_frequency	1
tgt_words_min_frequency	1

Table 4: The configuration used for the baseline model. For other parameters, tge default values were used.

4 Baseline System

The organizers built a baseline system, a recurrent neural network (RNN) encoder-decoder model with attention (Bahdanau et al., 2014) using OpenNMT (Klein et al., 2017) with the configuration shown in Table 4 and the same tokenizers for automatic evaluation explained in §3.1. This baseline system uses the old neural machine translation (NMT) model built for WAT 2018 and serves as a weak baseline for comparison. However, as shown in §6, many past participants have adopted Transformer-based systems, allowing for the performance comparison with Transformer models (Vaswani et al., 2017) for recent participants.

5 Participant Systems

Two teams participated in the patent translation task: GenAI (Yonsei University) and sakura (Rakuten Institute of Technology). The details on the submitted systems are as follows.

sakura used an LLM-based system fine-tuned with simple translation prompt on the JPC training set for the corresponding language pair. As their backbone model, they adopted RakutenAI-7B-chat,[13] which had been pretrained on English and Japanese texts.

GenAI used an LLM-based system fine-tuned on only 1,000 sentences from the JPC training set. Their backbone model is Mistral-Nemo-Instruct-2407 (12B),[14] which had been pretrained on multilingual texts. During both fine-tuning and testing, their system identified domain-specific terms in each input source sentence by matching them with their bilingual terminology dictionary, and then generated the translation based on prompt that required the use of the specified term pairs.

6 Evaluation Results

6.1 Main Results

For the same reasons mentioned in §3.2, we only present the results for the test-N3 set; results for other test sets can be found at the WAT Evaluation site.[15] Table 5, 6, 7, 8, 9, and 10 show the performance of evaluated system for each language direction (systems with “$*$r” indicate they used external resources). The tables present the automatic and human evaluation scores of the two participants' systems (one system per participant, selected based on the BLEU score), as well as the organizer's baseline and the best participant systems from previous years. The model type columns indicate whether the system employed statistical machine translation (SMT), RNN-based NMT, or Transformer (TF)-based NMT, and whether it corresponds to a decoder-only model (Dec) or an encoder-decoder model (EncDec). The BLEU/RIBES scores for the translation tasks into Japanese and Chinese represent the average BLEU/RIBES scores based on three different tokenizers.[16] The JPO adequacy scores (Adeq) represent the average of the scores assigned by two annotators.

We observed the following findings. (1) Unsurprisingly, both participants' systems as well as all previous best systems outperformed the baseline for all language directions in terms of automatic metrics. (2) The LLM-based systems by the two participants achieved strong results in terms of automatic metrics; GenAI's system outperformed the previous systems for ko→ja and ja→ko and sakura's system outperformed the previous systems for ja→ko and ja→en. However, the previous systems maintained the highest scores for zh→ja, ja→zh, and en→ja. (3) Both participants' systems achieved high adequacy scores of over 4. However, importantly, a system with a higher automatic evaluation score did not necessarily achieved a higher human evaluation score. Specifically, sakura's system yielded lower automatic evaluation scores than GenAI's system (e.g., BLEU of 52.77 vs. 67.10 for ja→ko and 68.00 vs. 70.60 for ja→ko), but

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

[14] https://huggingface.co/mistralai/Mistral-Nemo-Instruct-2407

[15] https://lotus.kuee.kyoto-u.ac.jp/WAT/evaluation/index.html

[16] Three tokenizers indicate Juman, KyTea, and MeCab for Japanese, and KyTea and Stanford Word Segmenter (CTB and PKU models) for Chinese.

System	Model Type	BLEU	RIBES	Adeq
GenAI best	TF Dec	67.10	0.9225	4.66
2018 best	SMT	54.63	0.9056	–
2019 best	TF EncDec	54.42	0.9012	–
2020 best	TF EncDec	53.77	0.9044	–
2021 best^{*r}	TF EncDec	53.48	0.9014	–
sakura best	TF Dec	52.77	0.8982	4.67
Baseline	RNN EncDec	52.65	0.8975	–

Table 5: Results on the ko→ja test-N3 set.

System	Model Type	BLEU	RIBES	Adeq
GenAI best	TF Dec	70.60	0.9391	4.39
sakura best	TF Dec	68.00	0.9268	4.76
2021 best	TF EncDec	66.25	0.9252	–
2019 best	TF EncDec	65.74	0.9228	–
2020 best	TF EncDec	64.30	0.9223	–
Baseline	RNN EncDec	62.43	0.9153	–

Table 6: Results on the ja→ko test-N3 set.

System	Model Type	BLEU	RIBES	Adeq
2020 best	TF EncDec	40.51	0.7568	–
2019 best	TF EncDec	24.96	0.7639	–
2018 best	TF EncDec	24.87	0.7492	–
2021 best^{*r}	TF EncDec	22.67	0.7716	–
sakura best	TF Dec	20.83	0.7615	4.24
Baseline	RNN EncDec	17.28	0.7322	–

Table 7: Results on the zh→ja test-N3 set.

System	Model Type	BLEU	RIBES	Adeq
2020 best	TF EncDec	44.34	0.8340	–
2021 best^{*r}	TF EncDec	31.09	0.8550	–
2019 best	TF EncDec	29.82	0.8390	–
sakura best	TF EncDec	26.60	0.8245	4.33
2018 best	TF EncDec	24.66	0.8261	–
Baseline	RNN EncDec	23.68	0.7886	–

Table 8: Results on the ja→zh test-N3 set.

System	Model Type	BLEU	RIBES	Adeq
2019 best^{*r}	TF Enc-Dec	55.32	0.8827	–
sakura best	TF Dec	53.93	0.8803	4.44
2021 best^{*r}	TF Enc-Dec	53.34	0.8753	–
2018 best^{*r}	SMT	52.07	0.8643	–
2020 best	TF Enc-Dec	50.95	0.8665	–
Baseline	RNN Enc-Dec	46.39	0.8438	–

Table 9: Results on the en→ja test-N3 set.

System	Model Type	BLEU	RIBES	Adeq
sakura best	TF Dec	43.20	0.8505	4.08
2019 best^{*r}	TF Enc-Dec	41.37	0.8499	–
2021 best^{*r}	TF Enc-Dec	40.73	0.8546	–
2020 best	TF Enc-Dec	39.94	0.8413	–
Baseline	RNN Enc-Dec	35.01	0.8230	–

Table 10: Results on the ja→en test-N3 set.

achieved similar or better adequacy scores (4.67 vs. 4.66 for ja→ko and 4.76 vs. 4.39 for ja→ko). This result highlights the need for using a variety of evaluation metrics, such as neural-based metrics, which have been demonstrated to correlate well with human judgement (Freitag et al., 2023).

6.2 Detailed Human Evaluation Results

Table 11 shows the detailed results of the JPO adequacy evaluation for a total of eight participant systems, which were selected from among the same participant's systems based on the BLEU score. The "Adequacy Score" column represents the average of ratings assigned to 200 sentences by each annotator for the Annotator="A"/"B" rows and the average and standard deviation of the average score by the two annotators (A and B) for the Annotator="Both" row, which is shown as the adequacy score (Adeq) in Table 5–10.

We observed the following findings. First, most sentences were assigned scores over 4 (75% or more sentences for each translation result, except for sakura's ja-en result evaluated by Annotator B). This indicates that there were many high-quality translation overall, but more accurate systems have room for development, considering that the translations with a score lower than 5 account for more than 20–50% in most cases of annotator-level evaluation results.

Second, the difference of sentence-level scores between two annotators ("Diff Score") was 0 or 1 in most cases, and there were only nine sentences with the difference score of 2 over all translation results. As a result, the adequacy scores between two annotators were close in many cases, but relatively large standard deviation (close to or greater than 0.2) was observed in three cases, i.e., sakura ja-zh, GenAI ja-ko, and sakura ja-en results. In the latter cases, there were somewhat many mismatches; each translation result included over 100 sentences with a score difference of 1 from the two annotators and/or a few sentences with a score difference of 2.

For the nine sentences with a score difference of 2, we conducted a meta-review by a third evaluator, distinct from the two annotators. We found that which annotator provided the more appropriate rating varied depending on the example. In some examples, one annotator overlooked a mistranslation and assigned a higher rating. In other examples, there were no mistranslations, but one annotator still assigned a lower rating. Additionally, in cases

Lang	Team	Data ID	Annotator	Adequacy Score	Distribution of Ratings					Diff Score		
				(Avg. ± SD)	1	2	3	4	5	0	1	2
zh-ja	sakura	7302	A	4.24	4	4	24	76	92			
			B	4.24	2	6	26	74	92			
			Both	4.24 ± 0						130	70	0
ja-zh	sakura	7257	A	4.50	2	5	17	43	133			
			B	4.15	7	10	30	52	101			
			Both	4.33 ± 0.18						120	80	2
ko-ja	sakura	7311	A	4.79	1	1	7	21	170			
			B	4.55	2	0	9	65	124			
			Both	4.67 ± 0.12						137	63	0
ko-ja	GenAI	7180	A	4.84	0	0	1	37	162			
			B	4.51	0	0	0	99	101			
			Both	4.66 ± 0.15						124	76	0
ja-ko	sakura	7224	A	4.64	0	4	4	52	140			
			B	4.87	0	1	4	15	180			
			Both	4.76 ± 0.12						148	52	0
ja-ko	GenAI	7267	A	4.16	0	7	38	71	84			
			B	4.61	0	0	9	60	131			
			Both	4.39 ± 0.23						98	102	0
en-ja	sakura	7278	A	4.49	0	4	15	61	120			
			B	4.40	0	3	35	41	121			
			Both	4.44 ± 0.04						123	73	0
ja-en	sakura	7309	A	3.83	2	22	59	43	74			
			B	4.33	1	5	26	64	104			
			Both	4.08 ± 0.25						79	144	7

Table 11: Detailed results of the JPO adequacy evaluation for the test-N3 set. The "Distribution of Ratings" column shows the number of sentences with each rating of 1–5. The "Diff Score" represents the number of sentences with each difference score, which means the difference of ratings between two annotators.

where the translation contained garbled characters, one annotator assigned a lower rating.

7 Conclusion

This paper summarizes the results of the WAT/WMT 2024 shared task on patent translation. The patent translation task using the JPO Patent Corpus has been held for ten years, and this will be the last time.[17] We believe that extensive development efforts by task participants over the past 10 years have contributed to advance machine translation technologies for the patent domain.

References

Dzmitry Bahdanau, KyungHyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv:1409.0473*.

Markus Freitag, Nitika Mathur, Chi-kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Tom Kocmi, Frederic Blain, Daniel Deutsch, Craig Stewart, Chrysoula Zerva, Sheila Castilho, Alon Lavie, and George Foster. 2023. Results of WMT23 metrics shared task: Metrics might be guilty but references are not innocent. In *Proceedings of the Eighth Conference on Machine Translation*, pages 578–628, Singapore. Association for Computational Linguistics.

Hideki Isozaki, Tsutomu Hirao, Kevin Duh, Katsuhito Sudoh, and Hajime Tsukada. 2010. Automatic evaluation of translation quality for distant language pairs. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 944–952, Stroudsburg, PA, USA. Association for Computational Linguistics.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72. Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session*.

[17] The JPO Patent Copurs will be provided to applicants via the ALAGIN forum (`https://www.alagin.jp/index-e.html`) for future research.

Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. 2004. Applying conditional random fields to Japanese morphological analysis. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 230–237, Barcelona, Spain. Association for Computational Linguistics.

Sadao Kurohashi, Toshihisa Nakamura, Yuji Matsumoto, and Makoto Nagao. 1994. Improvements of Japanese morphological analyzer JUMAN. In *Proceedings of The International Workshop on Sharable Natural Language*, pages 22–28.

Toshiaki Nakazawa, Kazutaka Kinugawa, Hideya Mino, Isao Goto, Raj Dabre, Shohei Higashiyama, Shantipriya Parida, Makoto Morishita, Ondřej Bojar, Akiko Eriguchi, Yusuke Oda, Chenhui Chu, and Sadao Kurohashi. 2023. Overview of the 10th workshop on Asian translation. In *Proceedings of the 10th Workshop on Asian Translation*, pages 1–28, Macau SAR, China. Asia-Pacific Association for Machine Translation.

Graham Neubig, Yosuke Nakata, and Shinsuke Mori. 2011. Pointwise prediction for robust, adaptable Japanese morphological analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2*, HLT '11, pages 529–533, Stroudsburg, PA, USA. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*, pages 311–318.

Huihsin Tseng. 2005. A conditional random field word segmenter. In *In Fourth SIGHAN Workshop on Chinese Language Processing*.

Masao Utiyama and Hitoshi Isahara. 2007. A Japanese-English patent parallel corpus. In *Proceedings of Machine Translation Summit XI: Papers*, Copenhagen, Denmark.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.