UTSK25 at WAT2025 Patent Claims Translation/Evaluation Task

Haruto Azami, Zhang Yin, Futo Kajita, Nobuyori Nishimura, Takehito Utsuro
Deg. Prog. Sys.&Inf. Eng., Grad. Sch. Sci.&Tech., University of Tsukuba

Abstract

This paper presents the submission of UTSK25 for the English–Japanese and Japanese–English at the WAT2025 Patent Claims Translation/Evaluation Task. We use a single translation model for both translation directions, built from a large language model through monolingual and bilingual continual pretraining and bilingual supervised fine-tuning. We finally generate translations via prompt engineering to reduce omissions and hallucinations.

1 Introduction

We describe our UTSK25 translation system for the WAT2025 English–Japanese (En–Ja) and Japanese–English (Ja–En) Patent Claims Translation/Evaluation Task. Our translation model is trained on a pretraining large language model (LLM), rinna/llama-3-youko-8b[1]. We combine two training stages (Kondo et al., 2024; Azami et al., 2025) to train a single model for both directions: continual pretraining (CPT) (Ke et al., 2023) and supervised fine-tuning (SFT) (Zhang et al., 2024). After training the single translation model, we generate translations with prompt engineering techniques designed to mitigate omissions and hallucinations. The following sections show the details of our system.

2 Approaches

2.1 Training

Continual pretraining Continual pretraining (CPT) extends the training of LLMs by further optimizing the causal language modeling objective on new monolingual corpora (Ke et al., 2023). The goal is to optimize the model parameters θ by minimizing the negative log-likelihood $\mathcal{L}_{\text{CPT}}$ over a corpus $\mathcal{D}_{\text{CPT}}$. Given a corpus $\mathcal{D}_{\text{CPT}} := \{\mathbf{y}_i\}_{i=1}^{|\mathcal{D}_{\text{CPT}}|}$ composed of token sequences $\mathbf{y} = (y_1, \ldots, y_{|\mathbf{y}|})$ from the vocabulary $\mathcal{V}$ (where $\mathbf{y} \in \mathcal{V}^*$), the loss is defined as:

$$\underset{\theta}{\operatorname{argmin}} \sum_{\mathbf{y} \in \mathcal{D}_{\text{CPT}}} \mathcal{L}_{\text{CPT}}(\mathbf{y}; \theta), \tag{1}$$

$$\mathcal{L}_{\text{CPT}}(\mathbf{y}; \theta) := -\sum_{t=1}^{|\mathbf{y}|} \log p_\theta(y_t | \mathbf{y}_{<t}). \tag{2}$$

This objective trains the model to predict the next token y_t given its history $\mathbf{y}_{<t}$. For efficiency, practical implementations often limit the context to a fixed-size window c, using $\mathbf{y}_{[t-c,t)} := (y_{t-c}, \ldots y_{t-1})$ as the condition instead of the full sequence $\mathbf{y}_{<t}$. This formulation is identical to the standard pretraining objective for causal LMs.

Supervised fine-tuning Supervised fine-tuning (SFT) optimizes pretrained model parameters θ for downstream tasks using a labeled dataset (Zhang et al., 2024). This dataset, $\mathcal{D}_{\text{SFT}} := \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{|\mathcal{D}_{\text{SFT}}|} \subset \mathcal{V}^* \times \mathcal{V}^*$, contains pairs of an input $\mathbf{x}$ and its corresponding ground-truth output $\mathbf{y}$. The optimization objective is to minimize the negative log-likelihood $\mathcal{L}_{\text{SFT}}$ over all pairs in $\mathcal{D}_{\text{SFT}}$:

$$\underset{\theta}{\operatorname{argmin}} \sum_{(\mathbf{x},\mathbf{y}) \in \mathcal{D}_{\text{SFT}}} \mathcal{L}_{\text{SFT}}(\mathbf{x}, \mathbf{y}; \theta), \tag{3}$$

$$\mathcal{L}_{\text{SFT}}(\mathbf{x}, \mathbf{y}; \theta) := -\log p_\theta(\mathbf{y} \mid \mathbf{x}). \tag{4}$$

This process steers the model to generate outputs conditioned on the input that are consistent with the human-annotated targets.

2.2 prompt engineering

We generate translations with prompt engineering techniques designed to mitigate omissions and hallucinations only for the En–Ja translation.

3 Submission System

We train the En–Ja and Ja–En single translation model from a pretrained LLM, llama3-youko-8b.

[1] https://huggingface.co/rinna/llama-3-youko-8b

Proceedings of the Twelfth Workshop on Asian Translation (WAT 2025), pages 30–34
December 24, 2025 ©2025 Association for Computational Linguistics

Submission	En–Ja Prompt Used
System 1 (Primary)	Prompt 2
System 2 (Not Primary)	Prompt 3
System 3 (Not Primary)	Prompt 1

Table 1: Submitted systems. All systems use the identical bilingual (En–Ja/Ja–En) model trained with CPT and SFT, differing only in the prompt used for the En–Ja direction.

According to our preliminary experiments and subjective judgment, we selected the combinations of training methods and prompts.

We show the system overview in Table 1.

3.1 Continual Pretraining

We perform bilingual CPT for our translation model. For CPT, we use a subset of the JParaPat dataset (Nagata et al., 2025).Table 3 summarizes the data statistics for CPT.

We filter this subset to remove entries where the English side contains "(canceled.)". The CPT corpus is balanced, containing 50% English-to-Japanese (En–Ja) and 50% Japanese-to-English (Ja–En) examples.

The CPT hyperparameters are listed in Table 2.

3.2 Supervised Fine-tuning

Following CPT, we conduct supervised fine-tuning (SFT). For SFT, we use the 2020 patent claims data from JParaPat (Nagata et al., 2025).

While the original dataset consists of line-by-line parallel data, some patent claims span multiple lines. To address this, we first construct claim-level pairs by segmenting the Japanese text at kuten (。) and the English text at periods (.). We also filter out pairs containing "(canceled.)" on the English side, similar to the CPT data preparation.

From this processed dataset, we then sample our final training data, selecting only pairs with LaBSE (Feng et al., 2022) embedding similarity scores between 0.9 and 0.95. Table 3 summarizes the SFT data statistics. The final SFT corpus is also balanced, with 50% En–Ja and 50% Ja–En pairs.

The SFT hyperparameters are also listed in Table 2.

3.3 Prompt Engineering

We use one prompt for Ja–En translation and three distinct prompts for En–Ja translation.

Hyperparameter	CPT	SFT
Optimizer	AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.95$) (Loshchilov and Hutter, 2019)	
Learning rate	2.5×10^{-5}	1×10^{-6}
Scheduler	cosine	inverse square root
Warmup ratio	1%	1%
Weight decay	0.1	0.1
Gradient clip	1.0	1.0
Epoch	1	3
Batch size	1,024 chunks	64 sentence pairs
Chunk size	2,048 tokens	N/A
Accelerator	DeepSpeed ZeRO-2 (Rasley et al., 2020)	
Precision	bfloat16	bfloat16

Table 2: Hyperparameters of CPT and SFT.

Usage	Time Period	Data Type	Sentence Pairs	English Words
CPT	2016~2019, 2020(non-claims)	train	97,491,362	3.09B
		dev	10,000	317K
SFT	2020	train	30,000	824K
		dev	3,000	88.9K

Table 3: Usage and Details of Patent Parallel Data

Ja–En Prompt

The prompt used for Ja–En translation is as follows:

Ja–En Prompt

これを日本語から英語に翻訳してください。
ただし文頭に関係のない数字を出さないようにしてください。:
日本語: {japanese_text}
英語:

The English translation of the above prompt is: "Translate this from Japanese to English. However, do not start the sentence with an irrelevant number."

En–Ja Prompts

The three distinct prompts used for En–Ja translation are shown below.

En–Ja Prompt 1 : Not Primary

Translate this from English to Japanese:
English: {English_text}
Japanese:

SFT Configuration		En–Ja(Ref 2)		Ja–En			
Data Construction	Data Filtering	BLEU	COMET	Ref 1		Ref 2	
				BLEU	COMET	BLEU	COMET
line-by-line	length-based	48.0	89.63	59.4	84.59	65.3	84.96
line-by-line	LaBSE and length-based	49.4	89.59	58.5	84.65	65.3	85.13
claim-level	**LaBSE-based**	**49.3**	**89.41**	**63.0**	**85.10**	**70.4**	**85.62**

(a) Comparison of SFT Data Preparation Strategies (All SFT models are initialized from the CPT model.)

Training Configuration		En–Ja (Ref 2)		Ja–En			
CPT	SFT	BLEU	COMET	Ref 1		Ref 2	
				BLEU	COMET	BLEU	COMET
✗	✓	24.5	87.67	16.8	75.54	20.0	75.97
✓	✓	**49.3**	**89.41**	**63.0**	**85.10**	**70.4**	**85.62**

(b) Comparison of SFT-only vs CPT+SFT.

Table 4: Automatic Evaluation Results on the WAT2025 Development Sets (Underlined configuration denotes the one used in our submission system.)

En–Ja Prompt 2 : Primary

Translate from English to Japanese.
Keep all meanings. Do not skip or invent anything.
English: {English_text}
Japanese:

En–Ja Prompt 3 : Not Primary

Translate this from English to Japanese.
Do not include anything unrelated to the input.
English: {English_text}
Japanese:

4 Experiments

4.1 Ablation study of training methods

We investigate the effects of each training method.

Setup To validate our SFT data preparation strategy, we conduct a comparative study on different configurations, as detailed in Table 4a. All SFT models are initialized from the same CPT model. To separately analyze the contribution of CPT itself, we additionally report a comparison between models trained with SFT only and those trained with CPT followed by SFT. The results are summarized in Table 4b. Specifically, we investigate the impact of data construction and the corresponding filtering methods:

- **Data Construction:** We compare models trained on the original **line-by-line** data against the **claim-level** data used in our submission system.
- **Data Filtering:** We apply filtering strategies appropriate for each construction method. For the **line-by-line** data, which includes many short segments, we test a **length-based** filter and a combination of **LaBSE and length-based** filters. For our **claim-level** data, where sentences are already concatenated and sufficiently long, we apply only the **LaBSE-based** filter (Sub.).

All models are trained with the same hyperparameters as our submission system, as described in Section 4.1, unless otherwise noted.

For En-Ja translation, we used prompt 2, as described in Section 3.3.

Results The results of the automatic evaluation on the WAT2025 Patent Claims Translation/Evaluation Tasks development sets are presented in Table 4. Although two references are publicly available for both En–Ja and Ja–En, only reference 2 is used for the En–Ja evaluation due to omissions found in reference 1, while both references are reported for Ja–En. As shown in Table 4a, our submission configuration—claim-level data construction combined with LaBSE-based filtering—achieves the best performance across both Ja–En reference sets (Ref 1: 63.0 BLEU, Ref 2: 70.4 BLEU) and also maintains competitive performance in En–Ja. Furthermore, Table 4b demonstrates that CPT+SFT yields substantial improvements over SFT-only training in all evaluation settings, confirming the effectiveness of CPT as a pretraining stage.

5 Conclusion

We built our system for the WAT2025 Patent Claim Translation/Evaluation Task. Our model was trained with the combinations of CPT and SFT, initializing from a pretrained LLM (rinna/llama-3-youko-8b). To mitigate omissions and hallucinations, we generated translations via prompt engineering, especially for the En–Ja direction.

In our experiments, we observed that the SFT data preparation strategy is a critical factor for patent translation. We demonstrated that our submission's approach—using **claim-level** data construction and **LaBSE-based** filtering—yielded the best performance, particularly in the Ja–En direction. This highlights the importance of aligning SFT data with the logical structure of patent claims, rather than using simple line-by-line data.

Nevertheless, as patent claims often contain complex dependencies, eliminating omissions and hallucinations remains a challenge. We hope to further improve the adequacy and robustness of patent claim translation in future work.

Limitations

Small Development Set Size Although we conducted a comparative analysis in our ablation study (Section 4.1), the development sets provided by the task organizers are small. Therefore, it remains uncertain whether the strong performance of our submission configuration will generalize robustly across all types of patent claims.

Data Construction Imperfections Our claim-level data construction method relies on automatic segmentation using end-of-sentence symbols (Section 3.2). However, exceptions to these rules exist, which may lead to some data pairs having broken parallel relationships. Although we employed LaBSE-based filtering to mitigate this issue, it is not guaranteed that this filtering process successfully eliminated all such misaligned pairs from the SFT dataset.

Acknowledgements

This work partially used computational resources of Pegasus provided by the Multidisciplinary Cooperative Research Program in the Center for Computational Sciences, University of Tsukuba.

Author Contributions

Haruto Azami applied CPT and SFT,conducted translation experiments as described in Section 4.1 and other preliminary experiments,and selected the submission system.
Zhang Yin filtered SFT data using LaBSE embedding similarity scores.
Futo Kajita checked translation results to select the submission system.
Nobuyori Nishimura checked translation results to select the submission system.
Takehito Utsuro built and managed our team.

References

Haruto Azami, Minato Kondo, Takehito Utsuro, and Masaaki Nagata. 2025. Patent claim translation via continual pre-training of large language models with parallel data. In Proceedings of Machine Translation Summit XX: Volume 1, pages 300–314, Geneva, Switzerland. European Association for Machine Translation.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Zixuan Ke, Yijia Shao, Haowei Lin, Tatsuya Konishi, Gyuhak Kim, and Bing Liu. 2023. Continual pre-training of language models. In The Eleventh International Conference on Learning Representations.

Minato Kondo, Takehito Utsuro, and Masaaki Nagata. 2024. Enhancing translation accuracy of large language models through continual pre-training on parallel data. In Proceedings of the 21st International Conference on Spoken Language Translation (IWSLT 2024), pages 203–220, Bangkok, Thailand (in-person and online). Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In International Conference on Learning Representations.

Masaaki Nagata, Katsuki Chousa, and Norihito Yasuda. 2025. Japarapat: A large-scale japanese-english parallel patent application corpus.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD '20, page 3505–3506, New York, NY, USA. Association for Computing Machinery.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, and Guoyin Wang. 2024. Instruction tuning for large language models: A survey.