Findings of the WAT 2025 Shared Task on Japanese–English Article-level News Translation

Naoto Shirai, Kazutaka Kinugawa, Hitoshi Ito, Mino Hideya and Yoshihiko Kawai
NHK Science and Technology Research Laboratories
1-10-11 Kinuta, Setagaya-ku, Tokyo, Japan
{shirai.n-hk, kinugawa.k-jg, itou.h-ce, mino.h-gq, kawai.y-lk}@nhk.or.jp

Abstract

We present the preliminary findings of the WAT 2025 shared task on document-level translation from Japanese to English in the news domain [1]. This task focuses on translating full articles with particular attention to whether translation models can learn to produce expressions and stylistic features typical of English news writing, with the aim to generate outputs that resemble original English news articles. The task consists of three translation styles: (1) literal translation, (2) news-style translation, based on English articles edited to match Japanese content, and (3) finalized translation, the primary goal of this shared task. Only one team participated and submitted a system to a single subtask. All tasks were evaluated automatically, and one task was also evaluated manually to compare the submission with the baseline.

1 Introduction

Recent advances in large language models (LLMs) have shown strong potential to improve document-level translation (Wang et al., 2023). Several methods have been proposed to ensure document-level consistency, including context-aware prompting (Cui et al., 2024), fine-tuning (Wu et al., 2024), and agent based approaches (Wang et al., 2025).

In the domain of news translation, translators may need to move beyond fidelity to the source and consider the needs of the target readership (Schäffner, 2012). In practice, this can involve adapting context and expressions to improve clarity for local audiences. In our previous study (Nakazawa et al., 2020; Kinugawa et al., 2024), we constructed sentence-level and document-level news translation data using Japanese and English articles. Building on this foundation, the current shared task explores translation quality at the article level.

In this study, we set up an evaluation using Japanese and English articles published by a Japanese news agency to assess how well existing language models can produce translations that reflect consideration for the target reader. Our task consists of three translation styles: (1) literal translation, (2) news-style translation that preserves the content of Japanese articles while translating them into natural English, and (3) finalized translation, which refers to translation into an original English article. For the third task, which is the main focus of the shared task, we received a submission from one participating team. We evaluated both the baseline and submitted outputs to understand current challenges. The baseline model demonstrated improved performance in document-level BLEU (Papineni et al., 2002) scores when fine-tuned on each dataset. In the third task, a model fine-tuned and optimized with direct preference optimization (DPO) (Rafailov et al., 2023) by the submitting team achieved the highest performance in document-level BLEU scores. This suggests that using original translations as preference data may help to produce translations that are better aligned with reader expectations. However, human evaluation ranked GPT-4o (OpenAI et al., 2024) highest overall, which indicates that challenges remain in translating long articles with deep contextual understanding.

2 Task and Dataset

2.1 Task

This shared task focuses on evaluating the performance of document-level translation models in producing translations that reflect such reader-oriented adaptations when translating Japanese news articles into English. To achieve this, we defined three subtasks:

- **Task 1 Literal Translation**: Literal translation from the Japanese article.

[1] We hosted a shared task titled "Japanese → English: Article-level News Translation Tasks": https://lotus.kuee.kyoto-u.ac.jp/WAT/jiji-corpus/2025/.

Proceedings of the Twelfth Workshop on Asian Translation (WAT 2025), pages 93–97
December 24, 2025 ©2025 Association for Computational Linguistics

Table 1: Dataset statistics: number of articles ($|D| = 377$), tokens ($|T|$), and sentences ($|S|$), with per-article averages. We consider each headline as a single sentence for each article.

Data Name	$\|T\|$	$\|S\|$	$\|T\|/\|D\|$	$\|S\|/\|D\|$
Original Japanese Article	142,353	4,682	377.59	12.42
Original English Article	129,553	4,475	343.64	11.87
Literal English Translation	137,321	4,747	364.25	12.59
News-style English Translation	144,211	4,888	382.52	12.97

- **Task 2 News-style Translation**: Translation into natural English that preserves the content of the Japanese article.
- **Task 3 Finalized Translation**: Translation into the original English article, which serves as the main objective of this shared task.

Task 2 focuses on producing natural English while preserving the content of the Japanese article, whereas Task 3 aims to match the original English article written by the news agency. For Tasks 2 and 3, we would produce translations that include the dateline (e.g., location and date at the beginning of the article), in line with Jiji Press's English news writing style.

2.2 Dataset Construction

We constructed a dataset consisting of 377 Japanese–English article pairs published by Jiji Press [2] in 2024, each covering the same event. For each Japanese article, we provided two types of English translations: a literal version and news-style version. The dataset included:

- **Original Japanese Article**: Japanese article published by the news agency.
- **Original English Article**: English articles by the news agencies written for an international readership. Task 3 Reference Translation.
- **Literal English Translation**: Translation prioritizing lexical and syntactic fidelity to the original Japanese articles by the translator. Task 1 Reference Translation.
- **News-style English Translation**: Edited translation that reflects English news writing conventions while preserving the content and reporting intent of the Japanese original by the translator. Task 2 Reference Translation.

[2] https://www.jiji.com/

Hyperparameter	**Value**
Optimizer	adamw_torch
Learning rate	5e-5
Weight decay	0.01
LR scheduler	cosine
Warmup steps	20
Micro batch size	1
Gradient accumulation steps	1
Epochs	5

Table 2: Hyperparameters used for the SFT models of each task.

The literal and news-style translations were newly created for this task. Translators were instructed to maintain either strict fidelity or stylistic adaptation, depending on the target version. The dataset was split randomly into three subsets: 227 articles for training data, 50 for development data, and 100 for test data. Statistical information [3] for the dataset is shown in Table 1. In addition to the newly constructed data, we also distributed the Jiji2020 dataset [4], which we proposed previously.

3 Approach

3.1 Baseline Models

As baseline systems, we used GPT-4o [5] and Qwen3-8B [6] (Yang et al., 2025), a multilingual model with strong performance in Japanese. For Qwen3-8B, we evaluated both zero-shot inference and supervised fine-tuning (SFT) using 227 training pairs aligned with the expected outputs for each task. The hyperparameters and prompts used in each setting are shown in Tables 2 and 3, respectively. This experiment used four NVIDIA A100 GPUs.

[3] We used SpaCy: https://spacy.io/
[4] https://lotus.kuee.kyoto-u.ac.jp/WAT/jiji-corpus/2020/
[5] Version "gpt-4o-2024-11-20" provided by Azure OpenAI.
[6] https://huggingface.co/Qwen/Qwen3-8B

Table 3: Prompts used for each task.

Task 1 Literal Translation
Translate the following Japanese news article into English. The output should consist of a headline, followed by a newline, then a body. Do not use extra line breaks or markdown symbols.
+ [Original Japanese Article]
Task 2 News-style Translation
Translate and edit the following Japanese news article into English. The output should consist of a headline, followed by a newline, then a body starting with an appropriate dateline (e.g., "Tokyo, Jan. 1 (Jiji Press)–"). Rephrase and restructure the article. Do not use extra line breaks or markdown symbols.
+ [Original Japanese Article]
Task 3 Finalized Translation
Translate and edit the following Japanese news article into English. The output should consist of a headline, followed by a newline, then a body starting with an appropriate dateline (e.g., "Tokyo, Jan. 1 (Jiji Press)–"). Rephrase and restructure the article, adjusting the amount of information as needed to match English news style. Do not use extra line breaks or markdown symbols.
+ [Original Japanese Article]

3.2 Submission: NHK-system for Task 3

NHK-system (Mino et al., 2025) is the only submitted model for Task 3. It was trained with SFT and further optimized using DPO with Low-Rank Adaptation (LoRA) (Hu et al., 2021). In this setup, translations resembling literal or news-style outputs were considered as negative examples, whereas Original English Article were preferred. This approach aimed to improve alignment with English news writing. The system was implemented using the Qwen3-8B model.

Task 1: Literal Translation		
Model	**Method**	**d-BLEU**
GPT-4o	Zero-shot	24.87
Qwen3-8B	Zero-shot	22.46
	SFT	**27.45**
Task 2: News-style Translation		
Model	**Method**	**d-BLEU**
GPT-4o	Zero-shot	17.40
Qwen3-8B	Zero-shot	17.15
	SFT	**21.74**

Table 4: Results of Task 1 (top) and Task 2 (bottom).

4 Evaluation

4.1 Automatic Evaluation

We evaluated all tasks using document-level BLEU (d-BLEU) (Liu et al., 2020), which is based on n-gram matches across the whole document [7].

4.2 Human Evaluation

For Task 3, which received system submissions, we additionally conducted human evaluation. Two evaluators were assigned to each criterion. The model's outputs were compared through blind pairwise evaluation based on the perspectives of Adequacy and Fluency. Each submitted translation was compared with the baseline outputs, and assessed as a win, tie, or loss. The final score for each system was computed as the average of these outcomes across all comparisons.

[7]We used SacreBLEU (Post, 2018): `https://github.com/mjpost/sacrebleu`.

5 Result

5.1 Results of Tasks 1 and 2

Table 4 shows the results of the automatic evaluation of the baseline for literal translation and news-style translation. Results indicate that learning from the corresponding parallel data improved d-BLEU scores by over 4.5 points. Furthermore, GPT-4o achieved higher scores than Qwen3-8B's zero-shot model.

system	model	method	d-BLEU
Baseline	GPT-4o	Zero-shot	13.33
	Qwen3-8B	Zero-shot	14.09
		SFT	19.54
NHK-system	Qwen3-8B	SFT and DPO	**22.72**

Table 5: Results of Task 3 (finalized translation).

NHK-system	Win	Tie	Lose
vs GPT-4o	5.5 / 13.5	27 / 38.5	**67.5 / 48**
vs Qwen3-8B Zero-shot	14.5 / 19	**43 / 42**	42.5 / 39
vs Qwen3-8B SFT	**47** / 22	40 / **51.5**	13 / 26.5

Table 6: Human evaluation results (Adequacy/Fluency) showing Win/Tie/Lose ratios against different baselines.

5.2 Results of Task 3

Table 5 shows the results of the automatic evaluation for Task 3 of the baseline and NHK-system. The submitted system achieved the highest d-BLEU score against all baselines, outperforming the SFT-only baseline by 3.18 points. This suggests that DPO may help models better align with news-specific style and terminology.

Table 6 shows the human evaluation results for Task 3. This table indicates whether the submitted system outperformed (defined as a win) each baseline. The submitted model achieved scores higher than the SFT-only baseline in Adequacy and obtained comparable results in Fluency. These results also demonstrate the effectiveness of DPO. However, zero-shot models such as GPT-4o and Qwen3-8B significantly outperformed the submitted system. Notably, although GPT-4o achieved lower d-BLEU scores than fine-tuning models, it excelled at producing translations tailored to the target audience. These findings highlight the importance of multifaceted evaluation in document-level translation, especially human evaluation.

6 Conclusion

This paper reports preliminary findings from the Japanese–English news article translation task at WAT2025. The task was designed to evaluate document-level translation capabilities through three subtasks: literal translation, news-style translation, and finalized translation, focusing on whether LLMs can produce translations that resemble articles intended for English-speaking readers. SFT improved performance by approximately 4.5 document-level BLEU points in the literal and news-style subtasks. For the finalized translation, applying DPO in addition to SFT achieved a 3.18-point BLEU improvement over an SFT-only model. Human evaluation indicated that GPT-4o outperformed the baseline, thereby highlighting that improvements in BLEU did not consistently align with human assessments, particularly in adequacy and fluency. Overall, the findings indicate potential benefits of LLM tuning and, in specific cases, DPO for improving certain aspects of translation accuracy, while raising open questions about evaluation criteria and alignment with human assessments in news translation. In future work, we will investigate learning strategies and evaluation frameworks that better capture the requirements of document-level news translation.

Limitations

This study has several limitations. First, the experiments used Japanese–English news articles from a single news agency, restricting the findings to this dataset. Second, the evaluation relied on a single reference, which restricted the ability to capture diverse valid translations and may have biased evaluation metrics toward particular stylistic choices. Third, we used BLEU as an automated evaluation method, but it may not be an appropriate substitute for human evaluation (Mathur et al., 2020). In addition, human evaluation was conducted only in a pairwise manner, so absolute evaluations are also needed. Further exploration is required to understand the relationship between human and automatic evaluations, and to establish appropriate criteria for document-level translation assessment.

Ethical Statements

This study used news articles that were originally published by Jiji Press, Ltd. To protect privacy, all personal names were anonymized through pseudonymization, except for those of public figures.

Acknowledgments

We are deeply grateful to Hidehiro Asaka and Takayuki Kawakami for providing the valuable data used in this research. These research results were obtained from commissioned research (No. 225) by the National Institute of Information and Communications Technology (NICT), Japan. We thank Edanz (`https://jp.edanz.com/home`) for editing a draft of this manuscript.

References

Menglong Cui, Jiangcun Du, Shaolin Zhu, and Deyi Xiong. 2024. Efficiently exploring large language models for document-level machine translation with in-context learning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 10885–10897, Bangkok, Thailand. Association for Computational Linguistics.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Kazutaka Kinugawa, Hideya Mino, Isao Goto, and Naoto Shirai. 2024. Findings of the WMT 2024 shared task on non-repetitive translation. In *Proceedings of the Ninth Conference on Machine Translation*, pages 715–727, Miami, Florida, USA. Association for Computational Linguistics.

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.

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2020. Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4984–4997, Online. Association for Computational Linguistics.

Hideya Mino, Rei Endo Endo, and Yoshihiko Kawai. 2025. NHK submission to WAT 2025: Leveraging preference optimization for article-level news translation tasks. In *Proceedings of the 12th Workshop on Asian Translation*, Mumbai, India.

Toshiaki Nakazawa, Hideki Nakayama, Chenchen Ding, Raj Dabre, Shohei Higashiyama, Hideya Mino, Isao Goto, Win Pa Pa, Anoop Kunchukuttan, Shantipriya Parida, Ondřej Bojar, and Sadao Kurohashi. 2020. Overview of the 7th workshop on Asian translation. In *Proceedings of the 7th Workshop on Asian Translation*, pages 1–44, Suzhou, China. Association for Computational Linguistics.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, and 401 others. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

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. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, pages 53728–53741. Curran Associates, Inc.

Christina Schäffner. 2012. Rethinking transediting. *Meta*, 57(4):866–883.

Longyue Wang, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu. 2023. Document-level machine translation with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16646–16661, Singapore. Association for Computational Linguistics.

Yutong Wang, Jiali Zeng, Xuebo Liu, Derek Wong, Fandong Meng, Jie Zhou, and Min Zhang. 2025. Delta: An online document-level translation agent based on multi-level memory. In *International Conference on Representation Learning*, volume 2025, pages 15708–15731.

Minghao Wu, Thuy-Trang Vu, Lizhen Qu, George Foster, and Gholamreza Haffari. 2024. Adapting large language models for document-level machine translation. *Preprint*, arXiv:2401.06468.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.