# GTCOM and DLUT's Neural Machine Translation Systems for WMT24

**Hao Zong**[1] **Chao Bei**[2] **Conghu Yuan**[2]
**Wentao Chen**[2] **Huan Liu**[2] **Degen Huang**[1*]
[1]Dalian University of Technology
[2]Global Tone Communication Technology Co., Ltd.
zonghao@mail.dlut.edu.cn
{beichao, yuanconghu, chenwentao and liuhuan}@gtcom.com.cn
huangdg@dlut.edu.cn

## Abstract

This paper presents the submission from Global Tone Communication Co., Ltd. and Dalian University of Technology for the WMT24 shared general Machine Translation (MT) task at the Conference on Empirical Methods in Natural Language Processing (EMNLP). Our participation encompasses two language pairs: English to Japanese and Japanese to Chinese. The systems are developed without particular constraints or requirements, facilitating extensive research in machine translation. We emphasize back-translation, utilize multilingual translation models, and apply fine-tuning strategies to improve performance. Additionally, we integrate both human-generated and machine-generated data to fine-tune our models, leading to enhanced translation accuracy. The automatic evaluation results indicate that our system ranks first in terms of BLEU score for the Japanese to Chinese translation.

## 1 Introduction

In this study, we employ fairseq (Ott et al., 2019) as our development framework and adopt the transformer (Vaswani et al., 2017) as the main architecture. The primary ranking index for the submitted systems is BLEU (Papineni et al., 2002), which also serves as the evaluation metric for our translation system via sacreBLEU[1], consistent with our methodology from the previous year.

For data preprocessing, we conduct punctuation normalization, tokenization, and Byte Pair Encoding (BPE) (Sennrich et al., 2015) across all languages involved. Furthermore, we applied a truecase model for English, tailored to the specific linguistic features of each language. Regarding tokenization, we utilize Jieba[2] for Chinese, Mecab[3] for Japanese, and the Moses tokenizer.perl (Koehn et al., 2007) for English. Additionally, we incorporate knowledge-based rules along with a language model to cleanse parallel data, monolingual data, and synthetic data.

For the multilingual translation model, we consolidate all languages into a single model and enhance it with an English to Chinese parallel corpus to enrich the language information.

The remainder of this paper is structured as follows: Section 2 discusses the translation task and provides dataset statistics. Section 3 describes our baseline systems and introduces the proposed multilingual translation model. The data selection methodology is elaborated in Section 4. Section 5 presents experiments conducted on all translation directions, addressing data filtering, model architectures, back-translation, joint training strategies, adaptations of the multilingual model, fine-tuning, data selection, and ensemble decoding. Section 6 analyzes the results, offering insights into the efficacy of various techniques. Finally, Section 7 concludes the paper.

## 2 Task Description

This task focuses on bilingual text translation, with the provided data elaborated in Table 1, which includes both parallel and monolingual data. For the English-Japanese directions, the primary sources of parallel data include WikiMatrix (Schwenk et al., 2019), CCAligned (Rozis and Skadiņš, 2017), JESC (Pryzant et al., 2017), JParaCrawl v3.0 (Morishita et al., 2022), LinguaTools-WikiTitles (Tiedemann, 2012), News Commentary v16, and XLEnt (Tiedemann, 2012). For the Japanese-Chinese direction, the main parallel data is sourced from CCAligned, JParaCrawl, LinguaTools-WikiTitles, News Commentary v16, WikiMatrix, and XLEnt. Monolingual data comprises News Crawl (Kocmi et al., 2022) in English, Japanese, and Chinese; News Commentary in English, Japanese, and Chinese; and Europarl v10 in English. We uti-

[*] Corresponding Author
[1] https://github.com/mjpost/sacrebleu
[2] https://github.com/fxsjy/jieba
[3] https://github.com/taku910/mecab

*Proceedings of the Ninth Conference on Machine Translation (WMT)*, pages 227–231
November 15–16, 2024. ©2024 Association for Computational Linguistics

| Language | Number of Sentences |
|---|---|
| en-ja parallel data | 85.2M |
| ja-zh parallel data | 14.4M |
| en monolingual data | 168M |
| ja monolingual data | 22.8M |
| zh monolingual data | 23.9M |
| en-ja development set | 1000 |
| ja-zh development set | 1012 |

Table 1: Task Description

lized the provided development set from newstest2020 for English-Japanese and the FLoRes101 (NLLB Team, 2022) dataset for Japanese-Chinese.

## 3 Bilingual Baseline Model and Multilingual Translation Model

To establish a robust baseline for comparison with the multilingual model, we utilize the transformer_wmt_en_de as our bilingual baseline model, consisting of 24 encoder layers and 24 decoder layers. The multilingual translation model is designed to closely resemble the GTCOM2023 (Zong, 2023) model, referred to as the X to X model. To achieve superior translation quality, we include the English-Chinese parallel corpus as the primary auxiliary language pair to enhance linguistic information. We train a single multilingual model that encompasses all translation directions while applying joint Byte Pair Encoding (BPE) separately for all languages.

## 4 Data Selection

Similar to the last year, we use source test sets to train a text classification model based on RoBERTa (Liu et al., 2019). Specifically, we treat the in-domain test set as positive examples and select an equivalent amount of sentence pairs from the out-of-domain test set as negative examples. We fine-tune RoBERTa on this labeled dataset to develop a binary classifier capable of effectively distinguishing between in-domain and out-of-domain data. This classifier aids in selecting domain-specific training data from the general training corpus, with the chosen in-domain training data subsequently used to fine-tune the multilingual neural machine translation model.

Additionally, we also use prompt learning to explore an alternative data selection method. We develop a prompt template and leverage the generative capabilities of Meta-Llama-3-8B-Instruct [4] to create a domain classifier using loRA (Hu et al., 2021). The prompt template mirrors that used in GTCOM2023 from the last year, shows in Table 2. Specifically, we extract 800 sentences from the development set which belong to the news, social, e-commerce, or conversation domains. We manually select 200 sentences from the training set that do not match these domains or are of inferior quality, categorizing them as "other." We then utilize these 1,000 labeled examples to fine-tune the Meta-Llama-3-8B-Instruct model in loRA. The resulting prompt-based classifier effectively differentiates between domains in the training data. Sentences predicted as "News," "Social," "E-commerce," and "Conversation" are classified as in-domain data, while those labeled as "Other" are considered out-of-domain data.

## 5 Experiment

This section outlines the step-by-step experiments we conducted, with the entire workflow depicted in Figure 1.

- **Data Filtering:** The data filtering techniques largely replicate those utilized last year, incorporating human rules, language models, and repetition cleaning.
- **Baseline:** Our baseline is constructed using the transformer big architecture, which comprises 24 encoder layers and 24 decoder layers.
- **Back-translation:** We employ the best translation model to translate target sentences back to the source side, cleaning synthetic data using a language model. This process includes translating each language pair featured in the multilingual translation model. We combine the cleaned back-translation data with parallel sentences and train the multilingual translation model accordingly.
- **Joint Training:** We repeat the back-translation step using the optimal model until no further improvements are observed.
- **Multilingual Translation Model:** A single model is trained for all translation directions, with each direction utilizing joint BPE and a

[4] https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct

| | |
|---|---|
| Instructions | Please determine the domain to which the given sentence belongs based on the following criteria.<br>1. Sentence Correctness: If the sentence is incomplete, incoherent, or grammatically incorrect, label it as "Other" domain. If the sentence is complete, fluent, and grammatically correct, proceed to the next step.<br>2. Domain Identification: Analyze the content of the sentence to identify the possible domain it belongs to. Consider the following domains: News, Social, E-commerce, Conversation, and Other. If the sentence shows clear indications of being from a specific domain, label it accordingly, otherwise label it as "Other" domain.<br>Please label the sentence with the appropriate domain:<br>- If the sentence is from the News domain, label it as "News".<br>- If the sentence is from the Social domain, label it as "Social".<br>- If the sentence is from the E-commerce domain, label it as "E-commerce".<br>- If the sentence is from the Conversation domain, label it as "Conversation".<br>- If the sentence does not fit any specific domain or is incorrect, label it as "Other". |
| Sentence | Sunday Best: Enter 1880s New York in HBO's "The Gilded Age" |
| Domain | News |

Table 2: Prompt Template.

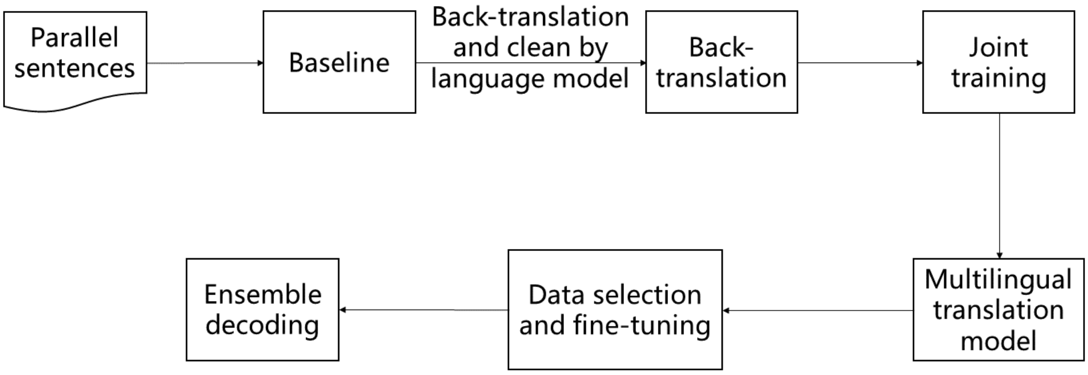
Parallel sentences
Baseline
Back-translation and clean by language model
Back-translation
Joint training
Multilingual translation model
Data selection and fine-tuning
Ensemble decoding

Figure 1: The work flow of GTCOM machine translation competition systems

shared vocabulary. The multilingual translation model consists of 24 encoder layers and 24 decoder layers, employing the transformer big architecture.

- **Fine-tuning:** The multilingual translation model is fine-tuned for each direction and bi-direction separately. For instance, we fine-tuned en2ja and ja2en on the multilingual translation model and fine-tuned en2ja on the multilingual translation model for English to Japanese separately.

- **Data Selection:** The model described in the Data Selection section is employed to choose a domain-specific training dataset, which is then fine-tuned on the multilingual translation model.

- **Ensemble Decoding:** We utilize the GMSE Algorithm (Deng et al., 2018) to select models, aiming for optimal performance.

## 6 Results and Analysis

Table 3 displays the BLEU scores evaluated on the development set for English to Japanese and Japanese to Chinese. As indicated in the table, back-translation remains the most effective data augmentation technique for enhancing translation quality from a data perspective. The multilingual translation model also demonstrates significant improvements across all translation directions. As shown in Table 4, our prompt learning strategy is

| Model | en2ja | ja2zh |
|---|---|---|
| Baseline | 26.36 | 15.07 |
| + Back-translation | 27.26 | 20.75 |
| Multilingual Translation Model | 26.50 | 15.20 |
| + Back-translation | 27.40 | 21.24 |
| + Bilingual Fine-tuning | 27.51 | 21.34 |
| + Single Fine-tuning | 27.22 | 20.98 |
| Ensemble Decoding | 27.95 | 22.21 |

Table 3: BLEU scores for English to Japanese and Japanese to Chinese. Values are calculated based on word counts.

| Direction | BLEU | BLEU with DS |
|---|---|---|
| en-ja | 39.2 | 39.7 |
| ja-zh | 32.9 | 32.3 |

Table 4: The final online automatic evaluation BLEU with/without prompt learning in data selection.

still able to improve the BLEU score on the direction of English to Japanese, but there was some decline in the Japanese-to-Chinese direction.

## 7 Conclusion

This paper introduces the neural machine translation systems developed by GTCOM and DLUT for the WMT24 shared general MT task. We apply three primary techniques to enhance translation quality: back-translation, a multilingual translation model, and fine-tuning accompanied by data selection. Through these methods, we achieve notable improvements in automatic evaluation metrics, as illustrated in Table 5.

| Direction | BLEU | CometKiwi |
|---|---|---|
| en-ja | 39.7 | 0.697 |
| ja-zh | 32.9 | 0.586 |

Table 5: Final online automatic evaluation results.

## Acknowledgments

The authors express their sincere gratitude for the financial support provided by the National Key Research and Development Program of China (2020AAA0108005) and the Key Research and Development Program of Yunnan Province (Grant No. 202203AA080004). This work also receives substantial support from the 2030 Artificial Intelligence Research Institute of Global Tone Communication Technology Co., Ltd.[5]

## References

Yongchao Deng, Shanbo Cheng, Jun Lu, Kai Song, Jingang Wang, Shenglan Wu, Liang Yao, Guchun Zhang, Haibo Zhang, Pei Zhang, et al. 2018. Alibaba's neural machine translation systems for wmt18. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 368–376.

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. *arXiv preprint arXiv:2106.09685*.

Tom Kocmi, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Rebecca Knowles, Philipp Koehn, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Michal Novák, Martin Popel, and Maja Popović. 2022. Findings of the 2022 conference on machine translation (WMT22). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1–45, Abu Dhabi, United Arab Emirates (Hybrid). 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, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the association for computational linguistics companion volume proceedings of the demo and poster sessions*, pages 177–180.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Makoto Morishita, Katsuki Chousa, Jun Suzuki, and Masaaki Nagata. 2022. Jparacrawl v3. 0: A large-scale english-japanese parallel corpus. *arXiv preprint arXiv:2202.12607*.

James Cross Onur Çelebi Maha Elbayad Kenneth Heafield Kevin Heffernan Elahe Kalbassi Janice Lam Daniel Licht Jean Maillard Anna Sun Skyler Wang Guillaume Wenzek Al Youngblood Bapi Akula Loic Barrault Gabriel Mejia Gonzalez Prangthip Hansanti John Hoffman Semarley Jarrett Kaushik Ram Sadagopan Dirk Rowe Shannon Spruit Chau Tran Pierre Andrews Necip Fazil Ayan Shruti Bhosale Sergey Edunov Angela Fan Cynthia Gao Vedanuj

[5] https://www.gtcom.com.cn

Goswami Francisco Guzmán Philipp Koehn Alexandre Mourachko Christophe Ropers Safiyyah Saleem Holger Schwenk Jeff Wang NLLB Team, Marta R. Costa-jussà. 2022. No language left behind: Scaling human-centered machine translation.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

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.

Reid Pryzant, Yongjoo Chung, Dan Jurafsky, and Denny Britz. 2017. Jesc: Japanese-english subtitle corpus. *arXiv preprint arXiv:1710.10639*.

Roberts Rozis and Raivis Skadiņš. 2017. Tilde model-multilingual open data for eu languages. In *Proceedings of the 21st Nordic Conference on Computational Linguistics*, pages 263–265.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2019. Wikimatrix: Mining 135m parallel sentences in 1620 language pairs from wikipedia. *arXiv preprint arXiv:1907.05791*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Lrec*, volume 2012, pages 2214–2218. Citeseer.

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*, pages 5998–6008.

Hao Zong. 2023. Gtcom and dlut's neural machine translation systems for wmt23. In *Proceedings of the Eighth Conference on Machine Translation*, pages 192–197.