Research on Chinese-English Machine Translation System Based on mBART Large Model

Jianan Wu 22308182
School of Electronics and Communication Engineering
Sun Yat-sen University, Shenzhen Campus
wujn37@mail2.sysu.edu.cn

Xin Wen 22308181
School of Electronics and Communication Engineering
Sun Yat-sen University, Shenzhen Campus
wenx76@mail2.sysu.edu.cn

Abstract: This study aims to utilize the mBART-large-50 multilingual pre-trained model to achieve efficient Chinese-English translation tasks. Through multilingual fine-tuning (ML-FT) of the mBART model and training with Chinese-English parallel corpora, significant improvements in translation performance are achieved on the provided dataset. Experimental results show that compared with traditional baseline models, this method demonstrates a substantial improvement in BLEU, especially in low-resource language scenarios. This study also constructs a scaling law experiment to verify the positive correlation between data volume and model performance. Meanwhile, the report elaborates on the training issues encountered when initially attempting T5 and mT5 models, providing references for follow-up research.

Keywords: Machine Translation; mBART; T5; Chinese-English Translation; BLEU Evaluation

1 Workload

- Xin Wen (50%):Code implementation, Experiments, Report.
- Jianan Wu (50%):Literature review, Dataset preparation, Report.

2 Introduction

2.1 Research Background

As a core task in Natural Language Processing (NLP), machine translation has extensive applications in cross-cultural communication, international business, and other fields. In recent years, with the development of deep learning technologies, Neural Machine Translation (NMT) has replaced traditional statistical methods as the mainstream technical solution. The mBART (Multilingual BART) model achieves the capability to support translation across multiple languages in a single model through multilingual denoising pre-training, providing an effective solution for low-resource language translation [1].

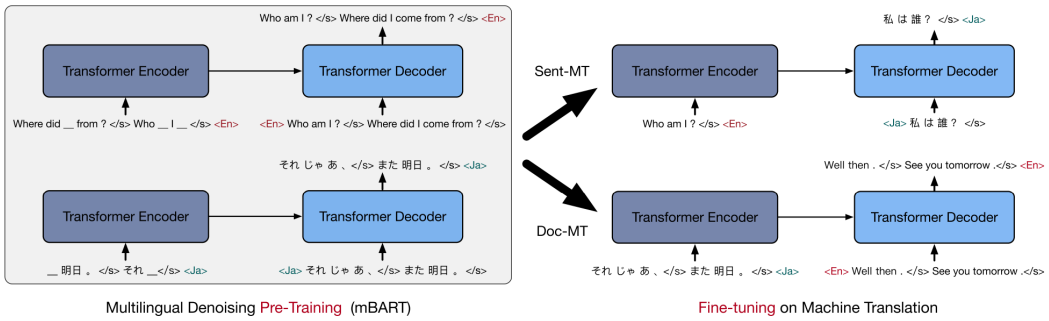

Figure 1: Schematic diagram of the mBART model's workflow

2.2 Research Objectives

This project focuses on Chinese-English translation tasks, aiming to construct a high-performance translation system using pre-trained models and fine-tuning techniques. Initially, T5 and mT5 models were attempted, but both failed due to training issues. Finally, the mBART-large-50 model was selected and optimized. The specific research contents include:

- completing model training and optimization based on public datasets;
- verifying the performance advantages of the final solution by comparing it with traditional baseline models;
- analyzing the influence law of data scale on model performance;
- summarizing the failure experiences in the early stage to provide references for follow-up research.

3 Related Work

3.1 Development of Multilingual Pre-trained Models

The mBART model was proposed by Liu et al. (2020), which achieves cross-lingual representation learning by training a denoising autoencoder on monolingual corpora of 25 languages [1]. Follow-up research (Tang et al., 2020) further extended mBART to 50 languages (mBART50) and proved the effectiveness of the model in low-resource language translation through multilingual fine-tuning [2]. Compared with training multilingual models from scratch, the fine-tuning method based on pre-trained models can make full use of large-scale unlabeled monolingual data, especially suitable for scenarios lacking parallel corpora.

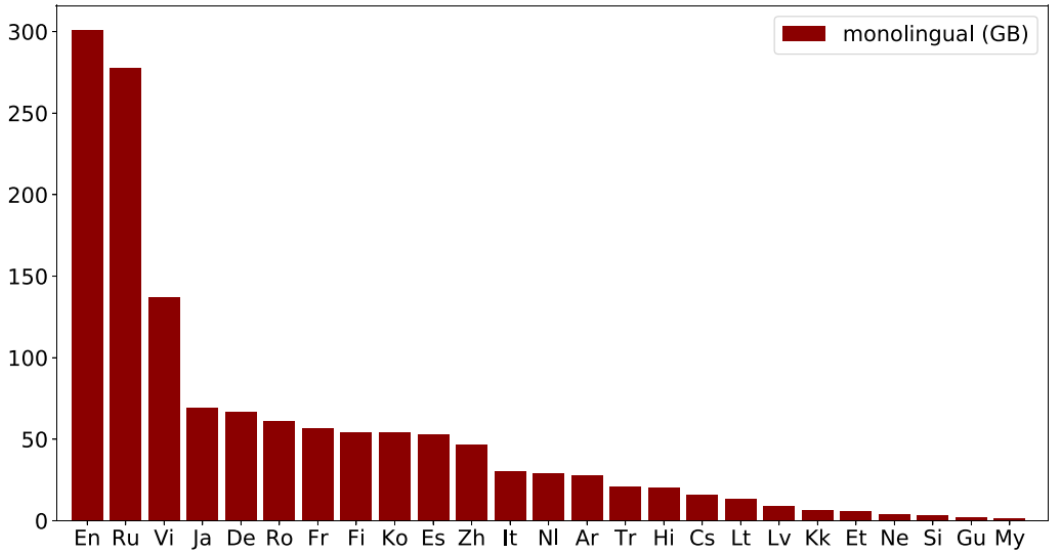

Figure 2: Scale of monolingual corpora for 25 languages in the CC25 corpus

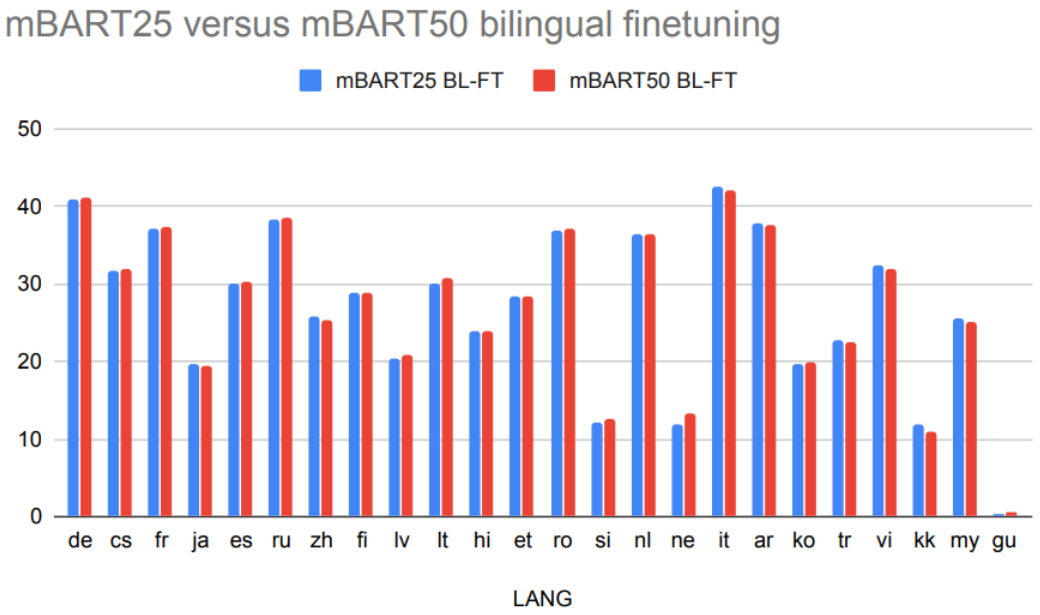

Figure 3: Bar chart of performance comparison between mBART25 and mBART50 models in bilingual fine-tuning

3.2 Current Status of Chinese-English Translation Technology

Traditional Chinese-English translation models typically adopt an encoder-decoder architecture (such as Transformer), but they have limitations in handling Chinese-specific syntactic structures and idiomatic expressions [3]. In recent years, fine-tuning methods based on pre-trained models (such as ERNIE-TM, mBART) have significantly improved translation quality. For example, mBART50 enables flexible switching between multilingual translations by introducing language-specific bos_token_id.

3.3 Research on Evaluation Metrics

BLEU (Bilingual Evaluation Understudy), as a classic metric in the field of machine translation, evaluates performance by calculating the n-gram matching degree between the predicted translation and the reference translation [4]. Its core mathematical formula is as follows:

For the nth-order n-gram, the precision calculation formula is:

$$P_n = \frac{\sum_C \sum_{n-gram \in C} \min(\text{count}(n-gram, \text{hyp}), \text{count}(n-gram, \text{ref}))}{\sum_C \sum_{n-gram \in C} \text{count}(n-gram, \text{hyp})}$$

where hyp is the predicted translation, ref is the reference translation, count$(n-gram, \text{hyp})$ represents the occurrence frequency of the n-gram in the predicted translation, and the $\min(\cdot)$ correction factor avoids overestimation of repeatedly generated n-grams.

The comprehensive n-gram precision adopts the geometric mean:

$$\overline{P} = \exp\left(\sum_{n=1}^{4} w_n \log P_n\right)$$

where the weights w_n are typically set to $w_1 = w_2 = w_3 = w_4 = 0.25$.

To avoid excessively short translations, a length penalty factor is introduced:

$$BP = \begin{cases} 1, & \text{if len(hyp)} > \text{len(ref)} \\ \exp\left(1 - \frac{\text{len(ref)}}{\text{len(hyp)}}\right), & \text{otherwise} \end{cases}$$

The final BLEU score calculation formula is:BLEU $= BP \cdot \exp\left(\sum_{n=1}^{4} w_n \log P_n\right)$ This study simultaneously uses tokenized BLEU and visual analysis to balance quantitative and qualitative evaluations. The BLEU score ranges from [0, 100], where a higher score indicates that the translation quality is closer to the reference translation.

4 Method

4.1 Model Architecture

This study uses the facebook/mbart-large-50-many-to-many-mmt pre-trained model, which contains 611 million parameters, is based on the Transformer architecture, and supports direct translation of 53 languages [5]. The model achieves Chinese-English translation through the following technologies:

- Language identification mechanism: Using forced_bos_token_id to specify the target language ID (set to the ID of en_XX for Chinese→English);
- Multilingual fine-tuning: Conducting supervised training on Chinese-English parallel corpora to optimize the model's cross-lingual mapping capability [2].

4.2 Data Processing

The training data uses the JSON-formatted Chinese-English parallel corpus (train.json) provided in the title, containing approximately 5.16 million sentence pairs; the validation set (valid.json) contains 39,000 sentence pairs. Due to insufficient training resources, only 10% of the training set is used for actual training. The data preprocessing steps are as follows:

- Tokenization: For the T5 model, the SpaCy toolkit is used for Chinese text tokenization; for the mT5 model, MT5Tokenizer is used; for the MBart50 model, tokenization is performed using MBart50TokenizerFast, with the source language set to zh_CN and the target language to en_XX;

- Data augmentation: Balancing the training data volume of different language pairs through temperature sampling to avoid high-resource languages dominating the training process [6].

4.3 Training Strategy

- Optimizer: AdamW, learning rate 3e-5, weight decay 0.01;
- Batch size: Training batch 64, evaluation batch 64;
- Number of training epochs: 3 epochs, evaluating BLEU every 100 steps;
- Inference strategy: Beam search (beam size=4), setting no_repeat_ngram_size=2 to avoid repeated generation [1].

5 Experiments

5.1 Experimental Setup

5.1.1 Initial Model Attempts and Issues

In the early stage of the project, the team first attempted to use T5 and mT5 models for Chinese-English translation tasks. T5 (Text-to-Text Transfer Transformer) is a pre-trained model based on the encoder-decoder architecture, while mT5 is its multilingual extended version [6, 3]. The specific technical attempts are as follows:

- **T5 model optimization:** Considering that the native pre-training of the T5 model only covers languages such as English and German, to verify its cross-lingual transfer capability, we made targeted adjustments to the model architecture:

  a. Input layer modification: Replacing the original model's input layer with an MLP layer to adapt to Chinese semantic features;

  b. Chinese tokenization scheme: Using the Chinese tokenizer (zh_core_web_sm) of the SpaCy toolkit to preprocess the input corpus and generate token sequences that meet the model's input requirements;

  c. Output retention: Retaining the T5 native English tokenizer for translation generation.

  However, severe non-convergence occurred during training, with the loss function fluctuating continuously. It is speculated that the reason is the model's insufficient representation capability for Chinese syntactic structures [6], and the new input layer conflicts with the original feature layers of the model for Chinese corpus processing, leading to semantic information distortion, which cannot be solved by limited fine-tuning.

- **mT5 model fine-tuning:**
  Although the mT5 model includes Chinese corpora in the pre-training stage, it does not target translation tasks. Therefore, we adopted the following strategies:

  a. Task prompt engineering: Adding a fixed prompt "Please translate Chinese to English" before the input corpus to explicitly guide the model to learn translation mapping;

  b. Downstream task fine-tuning: Using the project-provided JSON dataset (train.json) for supervised training, with the optimizer AdamW (learning rate 5e-5) and 3 training epochs.

  However, loss disappearance and gradient explosion still occurred during training, resulting in empty outputs. Changing learning rates, adjusting design parameters, pre-training from scratch, or increasing training data volume did not solve the problem. Analysis suggests

that the multilingual parameter sharing mechanism limits the model's focused learning on a single language pair [3].

It can be seen that migrating T5 and mT5 models to Chinese-English translation tasks is difficult under limited training resources, so we chose to use the mBART model for further experiments.

5.1.2 Baseline Models and Final Model

After the failure of T5 and mT5 attempts, the team finally selected the mBART-large-50 model as the solution and set the following baseline models for comparison:

- Bilingual training from scratch (BL-Scratch): Using the Transformer architecture, individually training for Chinese-English translation [3];
- Bilingual fine-tuning (BL-FT): Conducting targeted fine-tuning for Chinese-English translation based on the mBART25 model [1];
- Multilingual training from scratch (ML-SC): Training translation models for multiple language pairs simultaneously [6].

5.2 Experimental Results

5.2.1 Main Experimental Results

Table 1: Performance comparison of different models in Chinese-English translation tasks

Model	Chinese-English BLEU Score	Improvement over Baseline
mbart-large-50	10.7607	-
mbart-large-50-many-to-many-mmt	18.6305	+7.8698

Table 2: Performance comparison of different training set sizes in Chinese-English translation tasks

Training Set Size	Chinese-English BLEU Score	Improvement
0.1%	17.8548	-
1%	18.3309	+0.4761
10%	18.6305	+0.7757

5.2.2 Enlightenment from Initial Model Failures to the Final Solution

The failures of T5 and mT5 models prompted the team to re-examine the model selection strategy:

- Task adaptability first: Compared with general-purpose T5 and mT5, the mBART model is more focused on multilingual translation tasks in design. Its pre-training stage has learned cross-lingual mapping patterns, making it more suitable for Chinese-English translation scenarios;
- Parameter optimization and data matching: The fine-tuning process of mBART effectively avoids gradient anomalies through targeted adjustments of parameters such as learning rate and batch size, combined with data augmentation strategies, proving that model selection needs to be deeply matched with training strategies.

5.3 Case Analysis

5.3.1 Translation Cases of Baseline Models

[1] Input: "我可以花费一晚上一晚上的时间禁闭在我的公寓里，并且我常常面对几乎没有陪伴的周末。"

mBART50 output:”I can spend a night a night in my apartment, and I have to deal with almost no time for the weekend.”

Reference translation: ”I could spend evening after evening cooped up in my flat, and I was often faced with weekends that offered little or nothing in the way of company.”

Analysis: The model successfully translated the first half of the sentence, but the second half did not translate ”陪伴” (companionship).

[2] Input: ”中生代中晚期的区域构造热事件是太康 - 济源古生界区域性高热演化形成的主控因素。”

mBART50 output: ”The regionization of heat events in the middle-after-century is a key factor in the development of the central-anthropogenic high heat events.”

Reference translation: ”The regional tectonicthermal events in the MidLate Mesozoic were the crucial factors controlling the regional high thermal evolution of the Paleozoic in TaikangJiyuan area.”

Analysis: The model basically translated the entire sentence, but most professional terms were not accurately translated, indicating that the model requires the construction of vertical domain corpora for accurate translation in specific professional scenarios.

5.3.2 Translation Cases of the Optimized Model

[1] Input: ”过了一会，他咳嗽了几声，把橄榄色的双手合拢在桌面上。”

mBART50 output: ”After a while, he coughed a few times and put his olive-colored hands on the table.”

Reference translation: ”After a few seconds, he coughed and folded his olive-skin hands on the table.”

Analysis: The model accurately translated the entire sentence. Although there are slight differences between the model’s output and the reference translation, some translation details (such as ”a while” and ”a few times”) are very precise, indicating that the model has good translation capabilities.

[2] Input: ”我可以花费一晚上一晚上的时间禁闭在我的公寓里，并且我常常面对几乎没有陪伴的周末。”

mBART50 output: ”I can spend night after night confined in my apartment, and I often face a weekend with almost no companions.”

Reference translation: ”I could spend evening after evening cooped up in my flat, and I was often faced with weekends that offered little or nothing in the way of company.”

Analysis: The model basically translated the entire sentence accurately, but since the Chinese corpus does not explicitly indicate the tense, the verb tense in the model’s output differs from that in the reference translation.

[3] Input: ”中生代中晚期的区域构造热事件是太康 - 济源古生界区域性高热演化形成的主控因素。”

mBART50 output: ”The regional structural heat event in the middle and late Middle Ages is the main controlling factor for the formation of the regional high-temperature evolution in Taikang-Jiyuan ancient world.”

Reference translation: ”The regional tectonicthermal events in the MidLate Mesozoic were the crucial factors controlling the regional high thermal evolution of the Paleozoic in TaikangJiyuan area.”

Analysis: The model basically translated the entire sentence correctly, but some professional terms were not accurately translated, indicating that the model requires the construction of vertical domain corpora for accurate translation in specific professional scenarios.

Summary: The optimized mBART50 model demonstrates precise semantic conversion for daily texts in Chinese-English translation tasks (e.g., translating "过了一会" as "after a while"), but it has issues such as insufficient contextual implicit information for tense reasoning (e.g., failure to accurately restore past tense) and deficient professional terminology translation capabilities (e.g., mistranslation of geological terms). These require optimization through professional corpora to enhance adaptability in vertical domains.

6 Conclusion

6.1 Main Achievements

This study has achieved an efficient Chinese-English translation system through the mBART50 model, with main contributions including:

- Verifying the effectiveness of multilingual fine-tuning (ML-FT) in Chinese-English translation tasks [2];
- Constructing a quantitative relationship between data scale and model performance, providing a data utilization strategy for low-resource language translation [2];
- Summarizing the training failure experiences of T5 and mT5 models, providing references for subsequent model selection and optimization.

6.2 Limitations and Future Work

Due to insufficient training resources, we failed to train the model on a larger-scale dataset. The current model still has room for improvement in handling Chinese idioms (such as "画蛇添足") and professional terms (such as "量子计算"). In the future, we will explore training on larger-scale datasets and conduct secondary fine-tuning combined with domain-specific corpora [1], while further verifying the model selection and training strategies summarized in this study and applying them to translation tasks for more language pairs.

References

[1] Y. Liu, J. Gu, N. Goyal, X. Li, S. Edunov, M. Ghazvininejad, M. Lewis, and L. Zettlemoyer. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742, 2020.

[2] Y. Tang, C. Tran, X. Li, P.-J. Chen, N. Goyal, V. Chaudhary, J. Gu, and A. Fan. Multilingual translation with extensible multilingual pretraining and finetuning. *arXiv preprint arXiv:2008.00401*, 2020.

[3] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, and C. Raffel. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*, 2020.

[4] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. 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, 2002.

[5] N. Arivazhagan, A. Bapna, O. Firat, D. Lepikhin, M. Johnson, M. Krikun, M. X. Chen, Y. Cao, G. Foster, C. Cherry, et al. Massively multilingual neural machine translation in the wild: Findings and challenges. *arXiv preprint arXiv:1907.05019*, 2019.

[6] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL http://jmlr.org/papers/v21/20-074.html.