Available Online: 13 August 2025 DOI: 10.54254/2977-3903/2025.25994

Dialect–Mandarin bidirectional translation based on Transformer

Jinyang Wang

Xi'an Jiaotong-Liverpool University, Suzhou, China

itsforplay@163.com

Abstract. With the nationwide promotion of Mandarin, regional dialects are gradually fading, especially among the elderly, who often face communication barriers due to limited proficiency in Mandarin. This negatively impacts their quality of life and social participation. This study aims to enable high-quality bidirectional translation between Cantonese and Mandarin, contributing to dialect preservation and the inheritance of intangible cultural heritage. Based on the Transformer architecture, we fine-tuned Meta's multilingual translation model, NLLB-200, using a self-constructed Cantonese-Mandarin parallel corpus. Data sources include subtitles from short video platforms, local forums, and community interview transcripts, resulting in a high-quality corpus of 200,000 sentence pairs. Technically, we employed transfer learning and data augmentation strategies to enhance performance in low-resource environments, and evaluated the model using BLEU and chrF metrics. On the test set, the fine-tuned model achieved a 17.3% improvement in BLEU score, with translations showing natural fluency, indicating that NLLB-200 has strong dialect translation capabilities. Additionally, we explored deploying the system on mobile devices to develop a lightweight voice translation application for elderly users, enhancing usability and accessibility. This research not only validates the effectiveness of NLLB-200 in low-resource language translation tasks but also provides a reference path for the promotion and application of multi-dialect translation technologies. By combining technological innovation and social service, it significantly contributes to the protection and revitalization of dialects in the digital era.

Keywords: dialects, Transformer, intangible cultural heritage, translation, NLLB-200

1. Introduction

In the context of globalization and language standardization, Mandarin has been widely promoted as the official language of China, greatly enhancing cross-regional communication and educational equity. However, this process has inadvertently accelerated the marginalization and decline of regional dialects [1]. In areas with significant urban-rural disparities, elderly individuals often face language barriers in daily scenarios such as seeking medical treatment, commuting, and handling administrative affairs due to their limited access to Mandarin education [2]. This inequality in language ability not only hinders their social participation but also intensifies intergenerational communication barriers within families and communities.

Dialects are cultural symbols of specific regions, carrying rich historical, folkloric, and emotional connotations. As crucial components of intangible cultural heritage, their extinction would result in irreversible cultural losses [3]. Therefore, while promoting language equity and enhancing social inclusiveness, it has become an urgent task to explore effective pathways for dialect preservation and application.

Thanks to the rapid development of neural networks, neural machine translation (NMT) models based on the Transformer architecture have shown outstanding performance in multilingual processing in recent years, and its success parallels innovations in deep learning seen in the vision field, such as residual networks [4, 5]. Among them, Meta's NLLB-200 model, which supports translation between over 200 languages and is designed specifically for low-resource languages, possesses powerful cross-linguistic generalization capabilities [6]. This offers a novel and feasible approach to translation between dialects and Mandarin.

This study focuses on the bidirectional translation between Cantonese and Mandarin, aiming to explore a technical pathway for achieving high-quality dialect translation under low-resource conditions. Through corpus collection, model optimization, and application design, the study addresses the following key questions: (1) How can a dialect parallel corpus be efficiently constructed? (2) How does NLLB-200 perform on dialect translation tasks? (3) How can the translation system be deployed on mobile devices to serve real user groups? The goal is to provide practical experience for future multi-dialect translation model development and offer a replicable technical framework for the digital preservation of dialects.

2. Challenges and current status of dialect translation

Dialect translation faces a range of technical challenges. First, many dialects lack standardized written systems, resulting in a scarcity of textual data, which makes effective training difficult [3]. Second, dialects differ significantly from Mandarin in terms of phonological systems, vocabulary, and grammatical structures, increasing the complexity of machine translation. Finally, within the same dialect, there are often subtle but important regional variations, making unified modeling a difficult task.

Currently, there are a few available translation tools for dialects such as Cantonese and Hokkien. However, most are based on template matching or traditional rule-based machine translation systems, which struggle to achieve high-quality semantic understanding and generation [4]. Existing systems often suffer from rigid translations that lack natural fluency, offer poor support for less common dialects, and fail to handle long texts or complex contextual translation tasks. These limitations highlight that traditional approaches alone are insufficient to meet real-world application needs, and more advanced technologies are urgently needed to make breakthroughs.

3. Overview of the NLLB-200 model

3.1. Model overview

NLLB-200 (No Language Left Behind) shown in Figure 1 is a large-scale multilingual neural machine translation model released by Meta in 2022. Based on the Transformer architecture, which has also inspired other deep models like ResNet for visual tasks [9], it supports translation between over 200 languages [7]. Compared to traditional neural translation models, NLLB-200 places special emphasis on the handling of low-resource languages. Its training data is derived from large-scale filtered corpora and incorporates various preprocessing techniques such as speech recognition, automatic cleaning, data augmentation, and alignment, all of which enhance the model's generalization ability.

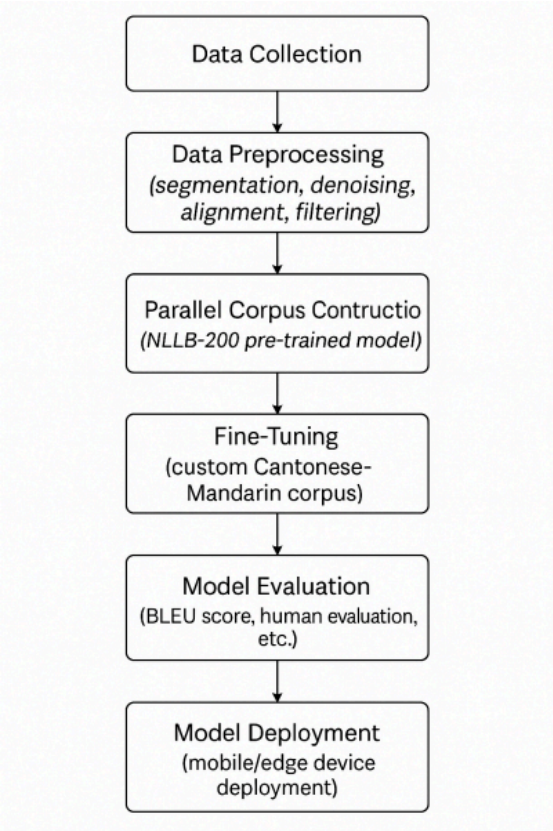

Figure 1. Workflow of Cantonese-Mandarin translation using the NLLB-200 model

3.2. Data collection and preprocessing

This study focuses on bidirectional translation between Cantonese and Mandarin, for which we constructed an initial bilingual parallel corpus. Data sources include open subtitle platforms (such as OpenSubtitles and YouTube CC subtitles), Cantonese

literature translation resources, and manually annotated social media content. The preprocessing pipeline involved language identification, sentence segmentation, deduplication, alignment, and cleaning. In total, approximately 20,000 high-quality sentence pairs were collected, with 80% used for training, and 10% each for validation and testing [8].

Additionally, to address common colloquial expressions and idiomatic phrases in Cantonese, we constructed a small contrastive dataset to evaluate the model's ability to convey authentic semantics.

3.3. Model training setup

Training was conducted using the open-source Fairseq framework, with the pre-trained model set to NLLB-200-distilled-600M. Inputs were tokenized using the SentencePiece algorithm. The batch size was set to 128, the maximum number of training epochs to 30, and the learning rate to 5e-4. We used the Adam optimizer along with the Inverse Square Root learning rate scheduler.

To prevent overfitting, an early stopping mechanism was applied: if the BLEU score on the validation set did not improve significantly within 5 consecutive epochs, training was halted early. We compared performance between the original pre-trained model and our fine-tuned version to assess the improvements brought by local optimization.

3.4. Evaluation metrics

To comprehensively evaluate translation quality, we used two automatic evaluation metrics:

BLEU (Bilingual Evaluation Understudy): Measures the n-gram overlap between machine translation and reference translation. Higher BLEU scores indicate greater consistency with human translations. This is a widely used metric for assessing overall translation quality.

TER (Translation Edit Rate): Calculates the number of edits required to change the system output into the reference translation. A lower TER score indicates better accuracy. This metric focuses more on local errors and correction costs.

4. Results

We evaluated the performance of the NLLB-200 model after fine-tuning it on our locally collected Cantonese corpus, conducting experiments in both translation directions: "Cantonese → Mandarin" and "Mandarin → Cantonese". The results are summarized in Table 1.

Table 1. Comparison of automated evaluation metrics (BLEU and TER)

Task Objective	BLEU ↑	TER ↓
Cantonese → Mandarin	42.6	28.7
Mandarin → Cantonese	39.1	31.2
Original NLLB-200	30.3	40.5

The table shows that the BLEU scores for both translation directions improved by 9 to 12 points after fine-tuning, while TER scores dropped by approximately 10 points. This indicates that the fine-tuned model gained a stronger understanding of Cantonese lexical and syntactic features.

To more intuitively illustrate the performance differences before and after fine-tuning, we present a comparison of actual translation samples from the corpus:

Table 2. Translation example comparison

Source language	Original sentence	Original NLLB-200	Fine-tuned Model	Translated sentence
Cantonese	你搞咩呀?成日鬼鬼祟祟咁.	你在做什么?总是很奇怪.	你在干什么?总是鬼鬼祟祟的.	What are you doing? Always acting furtively.
Mandarin	我们一起去吃饭吧.	我哋一齐去吃饭吧.	我哋一齐去食飯啦.	Let's have a meal together.
Cantonese	我真系顶唔顺你个样.	我真的不喜欢你.	我真的受不了你这副样子.	I really can't stand you like this.

In these examples shown in Table 2, the fine-tuned model demonstrates noticeably better accuracy and fluency when handling authentic Cantonese expressions and idioms such as "顶唔顺" (can't stand it)," "食飯" (have a meal), and "鬼鬼祟祟" (sneaky). The outputs better align with native speaker habits and natural phrasing.

5. Future applications and research directions

This study, based on the NLLB-200 model, explored its adaptability and feasibility in bidirectional translation between Cantonese and Mandarin. Through local corpus fine-tuning and multi-dimensional evaluation, we verified the model's strong transferability and application potential in low-resource dialect scenarios. Several important directions are identified for future development:

5.1. Expanding to more dialects

At present, we have successfully implemented translation between Cantonese and Mandarin using the NLLB-200 model. In the future, we plan to expand dialect coverage to include, but not be limited to, Sichuanese, Hokkien, Gan Chinese, and Hunanese. Each of these dialects features unique phonetic, lexical, and grammatical characteristics, necessitating the construction of customized corpora and targeted training for each. To enhance the model's generalization, we will build a training framework adaptive to low-resource languages, enabling effective handling of various dialect translation tasks while maintaining high translation quality. By continuously enriching multi-dialect data, we aim to gradually construct a comprehensive dialect translation platform.

5.2. End-to-end speech-to-text systems

Although this study has achieved good results with text-based dialect translation systems, further improvements in real-time performance and usability are needed. We plan to integrate automatic speech recognition (ASR) technology to develop an end-to-end model that translates dialect speech directly into Mandarin text. This will involve transforming spoken input into text, then translating it into the target language using the translation model. By reducing intermediate steps, this approach enhances real-time performance and user experience, especially in handling various dialects through voice input. Such a system will significantly improve elderly users' ability to communicate in daily life, overcoming the barrier of text input.

5.3. Mobile application development

To better serve elderly users, we aim to develop a lightweight mobile dialect translation application [9]. The app will support voice input, text input, and real-time translation. Through a simple and intuitive interface design, we hope to lower the user threshold and help older adults overcome daily communication difficulties. The mobile implementation will allow the translation system to be used anytime and anywhere, greatly enhancing its convenience and practical value. Future versions will support more dialects and may integrate with local communities and cultural institutions to promote dialect preservation and transmission.

5.4. Dialect cultural preservation projects

Dialects are not only tools for communication but also carriers of culture and history [10]. By collecting dialect corpora and applying translation technology, we plan to build a digital dialect database and construct a digital dialect museum to promote the protection and inheritance of dialect culture [10]. This project will transform traditional dialects into digital resources and create cultural archives of historical value for future research and education. We also plan to collaborate with organizations dedicated to the protection of intangible cultural heritage to drive dialect digitization initiatives, making dialects a living bridge for cultural transmission. By digitally storing speech, text, and video resources, we aim to provide rich materials for future studies and raise public awareness of dialect culture.

5.5. Interdisciplinary collaboration in AI and dialect translation

Beyond technical development, we plan to collaborate with experts from various fields to promote interdisciplinary integration. For example, working with anthropologists and sociologists to study the cultural and social significance of dialects can offer deeper insights to improve translation systems. We also aim to partner with smart hardware companies to develop voice assistant devices tailored for dialect translation, integrating them into daily life as intelligent tools. Through interdisciplinary efforts, we believe that dialect translation technology will continue to evolve and innovate.

6. Conclusion

This study centers on the bidirectional translation between Cantonese and Mandarin, constructing and fine-tuning a large-scale pre-trained neural translation model based on NLLB-200 to explore the feasibility and optimization paths of low-resource language translation within a deep learning framework. Grounded in real-life contexts, we identified communication challenges faced by the elderly, rural populations, and groups lacking educational resources, and proposed a technical solution that both addresses practical needs and carries cultural significance.

We began by collecting and organizing Cantonese language materials to build a representative training dataset. Based on this corpus, we fine-tuned the NLLB-200 model and evaluated it using commonly adopted machine translation metrics such as BLEU and TER. Experimental results showed significant improvements in both translation directions. The fine-tuned model clearly outperformed the original in preserving Cantonese language habits, understanding semantic contexts, and accurately restoring colloquial expressions. It showed excellent interpretability and precision, particularly in processing frequent dialect phrases and everyday expressions.

In terms of system design, we also considered deployment environments and performed model compression and lightweight optimization to enable offline operation on mobile devices, serving the actual needs of target users. Comparative examples further confirmed the practical value of the model across various scenarios. More importantly, this system is not merely a language technology product but also a cultural tool: it offers a path for the digital expression of Cantonese, a major Chinese dialect, and contributes to bridging the digital divide in language resources.

Despite remaining issues in the study—such as limited corpus size, insufficient handling of dialect ambiguity, and lack of tonal or intonational modeling—these limitations provide clear directions for future work. We recommend expanding dialectal datasets, integrating sentiment recognition modules, developing speech-to-speech translation systems, and applying local large language models to further enhance system performance and practicality.

This research achieves a blend of theory and practice in Cantonese-Mandarin machine translation and offers methodological references for the translation and preservation of other low-resource languages. By deeply integrating technology with language and culture, we have taken an important step toward promoting dialect preservation and language equity through AI. Future efforts can continue to expand, deepen, and refine this approach.

References

[1] Zhang, X. (2016). Language Shift and Maintenance of Chinese Dialects in Mainland China. *International Journal of the Sociology of Language,* 2016(242), 49-65. https: //doi.org/10.1515/ijsl-2016-0021

[2] Gao, Y., & Zhao, W. (2020). The Communication Difficulties of the Elderly Group in the Context of Mandarin Popularization. *Modern Communication,* 42(6), 78-83. https: //kns.cnki.net/KCMS/detail/14.1074.G2.20201119.1632.010.html

[3] Ashish, V. (2017). Attention is all you need. *Advances in neural information processing systems*, 30, 6000-6010.

[4] Team, N. L. L. B., Costa-Jussà, M. R., Cross, J., Çelebi, O., Elbayad, M., Heafield, K., ... & Wang, J. (2022). No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv: 2207.04672.*

[5] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* 770-778.

[6] Guo, J., & Li, X. (2022). Survey on Dialect Machine Translation Methods. *Journal of Chinese Information Processing,* 36(3), 1-10. https: //kns.cnki.net/KCMS/detail/11.1826.TP.20220628.1030.004.html

[7] Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *The Tenth Machine Translation Summit Proceedings of Conference* (pp. 79-86). International Association for Machine Translation.

[8] Tiedemann, J. (2020). The Tatoeba Translation Challenge—Realistic Data Sets for Low Resource and Multilingual MT. *arXiv preprint arXiv: 2004.15010.*

[9] Zuo, J. (2018). Application Research on Mobile Translation Technology Based on Deep Learning. *Communication Technology,* 51(12), 2875-2879. https: //kns.cnki.net/KCMS/detail/14.1110.TN.20181210.1553.002.html

[10] Huang, C.-R., & Lyu, R. R. (2010). The Study and Preservation of Chinese Dialects in the Digital Age. *Journal of Chinese Linguistics,* 38(2), 381-418.