WEP Warwick Evans Publishing

International Journal of Computer Science and Information Technology

ISSN: 3005-9682 (Print), ISSN: 3005-7140 (Online) | Volume 4, Number 2, Year 2024
DOI: https://doi.org/10.62051/ijcsit.v4n2.04
Journal homepage: https://wepub.org/index.php/IJCSIT/index

Recognition and Processing Strategies for colloquial and literary readings in Machine Translation: Taking Southern Fujian Dialect as an Example

Jinming Liu

Maynooth International Engineering College, Fuzhou University, Fuzhou, Fujian 350108, China
JINMING.LIU.2022@MUMAIL.IE

ABSTRACT

This study addresses the challenges of Neural Machine Translation (NMT) in handling textual and colloquial discrepancies in Chinese dialects, using Southern Min in Fujian as a case. It highlights the current state of NMT technology and explores sub word representation methods (e.g., Byte Pair Encoding) to mitigate issues with low-frequency and dialectal words. Dynamic attention mechanisms are discussed for their role in recognizing context-specific differences in reading and writing. Transfer learning and fine-tuning of pre-trained models are introduced as optimization strategies, alongside adaptive learning adjustments like dynamic learning rates, to enhance model flexibility and precision in complex linguistic scenarios. This paper offers practical approaches and theoretical insights to improve NMT's performance and adaptability in managing dialectical misreadings.

KEYWORDS

Neural Machine Translation (NMT); Southern Fujian Dialect; Chinese dialects

1. INTRODUCTION

In the realm of Natural Language Processing (NMT), the nuanced challenge of handling Colloquial and Literary Readings (CLR) in Chinese dialects, particularly in Southern Fujian dialect, presents both obstacles and opportunities. This dialect, known for its rich variations in spoken versus written language, exposes the limitations of NMT systems in accurately translating texts that contain these distinct linguistic features.

To address this, we'll delve into the evolution of NMT, highlighting the transformative role of attention mechanisms in enhancing context-aware translations. We'll scrutinize the current shortcomings of NMT when faced with dialect-specific phenomena like CLR, exploring how recent advancements in multilingual translation, low-resource, and zero-resource learning might offer indirect solutions or inspire new approaches. Key strategies include leveraging multi-task learning frameworks combined with monolingual and limited bilingual data to boost model versatility. Adaptive learning techniques and regularization methods will be examined for their potential to guide models in recognizing and differentiating CLR instances. Optimizing attention mechanisms to precisely manage the discrepancies between textual and colloquial readings during translation will also be a focal point.

Furthermore, this review will investigate tactics like data augmentation, noise injection, and hybrid modeling to bolster model accuracy and resilience against CLR misinterpretations in Southern Fujian dialect, aiming to refine NMT capabilities in navigating complex dialectal landscapes.

2. RELATED WORK

2.1. A summary of the phenomenon of CLR in Southern Fujian Dialect

The phenomenon of CLR in Southern Fujian Dialect, as a prominent language feature in Chinese dialects, refers to the phenomenon of different pronunciations of the same Chinese character in different contexts based on its formal use or oral habits. This phenomenon not only reflects the dynamism and historical sedimentation of language, but also deeply reflects the close connection between dialects, cultural inheritance, and social changes [1].

2.1.1. Definition and Characteristics

CLR, in short, refers to two or more pronunciations of the same word in a specific dialect due to different usage situations or contexts. In Southern Fujian dialect, there is a particularly rich and systematic phenomenon of textual and colloquial pronunciation, involving almost all phonetic categories. For example, in Xiamen dialect, more than half of the syllables have textual and colloquial pronunciation, which is usually related to written language or traditional Chinese pronunciation, while colloquial pronunciation is closely related to daily oral habits. This phenomenon is not only reflected in the differences in pronunciation and intonation, but also often accompanied by changes in vocabulary color, that is, the division of labor between different pronunciations in vocabulary color and pragmatic environment, reflecting the selection and distribution of components in the interaction between language systems [2].

2.1.2. influence factor

The formation and evolution of vernacular reading in Southern Fujian Chinese has been significantly influenced by historical, geographic, and socio-cultural factors. Historically, the differentiation of Chinese language led to divergences between classical and spoken language, which became pronounced in Southern Fujian due to the assimilation of different language systems over time. Geographically, Southern Fujian's strategic location as a crossroads for linguistic exchange fostered the borrowing and integration of language components, contributing to the complexity and diversity of reading and writing patterns.

At the socio-cultural level, the education system, political changes, and economic development have all played roles in shaping reading practices. Classical pronunciations were systematically preserved during the Ming and Qing dynasties due to the imperial examination system's emphasis on classical Chinese. However, with the modern promotion of Mandarin and educational reforms, the use of literary pronunciations has declined, giving way to a growing prominence of colloquial pronunciations. These historical, geographic, and socio-cultural dynamics have collectively contributed to the rich tapestry of vernacular reading in Southern Fujian Chinese [1].

2.1.3. Example analysis

Taking Xiamen dialect as an example, the phenomenon of different pronunciations between the written and spoken languages almost covers all sound categories in terms of phonetic distribution. For example, the difference between the pronunciations of Ge Yun wen and Bai reflects the interaction between different language levels in the historical development process. For example, the pronunciation of certain specific characters such as "ma" in Xiamen dialect directly reflects the remnants of early ancient Chinese phonetics, while the pronunciation of characters mostly corresponds to the phonetic system of middle ancient Chinese. This not only demonstrates the hierarchical nature of different pronunciations between text and vernacular in phonetics, but also reflects the uneven development of language systems [2].

In summary, the phenomenon of vernacular reading in Southern Fujian Chinese is a complex one that integrates historical, regional, and socio-cultural factors. It occupies an important position in linguistic research, not only of great significance for language change and dialect research, but also

poses new challenges and research directions for fields such as machine translation and natural language processing [1, 2].

2.2. The Challenge of Identifying Phenomenon of CLR in Machine Translation

2.2.1. Dataset limitations

The lack of fully annotated textual and colloquial reading datasets is a key factor limiting the training effectiveness of machine translation models. Due to the complexity of specific dialects involved in the phenomenon of textual and colloquial misreading, and the varying manifestations within different dialect regions and even within the same dialect, it is difficult to construct a comprehensive and accurate dataset for annotating textual and colloquial misreading. This data scarcity not only affects the model's ability to learn the patterns of text and vernacular reading, but also limits the model's ability to accurately identify and process the phenomenon of text and vernacular reading [3]. In practice, research on multilingual neural machine translation often relies on a large amount of parallel data, but there is a serious shortage of datasets for specific language phenomena such as text vernacular reading, which limits the model's generalization ability and translation quality [4].

2.2.2. Model capability

The current neural machine translation (NMT) model, although performing well in handling high resource language pairs, has significant limitations in dealing with language phenomena such as text vernacular reading. Attention mechanism is an important component in NMT models, which can help the model focus on key information in the input sequence [5]. However, traditional attention mechanisms may not be sufficient to accurately capture contextual dependencies and semantic differences in cross reading, especially when cross reading involves not only phonological changes, but also differences in lexical color and pragmatic background. As proposed in "Neural Machine Translation by Jointly Learning to Align and Translate", the model needs to have the ability to automatically find the source language sentence parts related to the target word without explicit segmentation, which puts higher requirements on the recognition and processing of text misreading [6].

2.2.3. Evaluation and optimization

Evaluating the accuracy of recognizing textual and colloquial errors in machine translation is a complex task that requires a combination of linguistic expertise and modern evaluation techniques. Traditional automatic evaluation metrics such as BLEU may not accurately reflect the subtle differences in text and vernacular reading processing, so more refined evaluation methods are needed, such as introducing manual evaluation or using more complex evaluation metrics such as TER (Translation Edit Rate) or chrF, to better measure the fit between translation output and source language text and vernacular reading differences [7]. In terms of optimization strategies, exploring the use of transfer learning, semi supervised learning, and unsupervised learning techniques, combined with monolingual data and limited bilingual corpus, can be an effective way to improve the model's ability to handle textual and colloquial reading errors.

2.3. Existing Research and Technical Solutions

2.3.1. Data augmentation

The data augmentation strategy aims to creatively utilize existing resources, expand the diversity of training data, and address complex language phenomena such as text and vernacular reading. For example, using bilingual or multilingual parallel corpora is a common means of enhancement. Neural Machine Translation by Jointly Learning to Align and Translate proposes a model for joint learning alignment and translation, which allows the model to automatically search for source sentence parts related to the target word without explicit segmentation, helping the model better understand and

handle the phenomenon of textual misreading [6]. The utilization of monolingual corpora is also crucial, as in "Neural Machine Translation of Rare Words with Subword Units", which introduces the use of sub word units to encode rare and unknown words. This method improves the model's ability to handle low-frequency and unregistered words, indirectly helping to deal with vocabulary phenomena unique to reading comprehension [8].

2.3.2. Model innovation

To adapt to the unique feature of reading different texts, researchers have designed specific model architectures and learning strategies. The enhanced attention mechanism is proposed in "Learning Phrase Representations Using RNN Encoder Decoder for Statistical Machine Translation". Through a recursive neural network encoder decoder framework, the model can better capture the dependency relationship between the source and target languages, which is particularly important in dealing with text misreading, as it helps the model understand the different pronunciations of the same word in different contexts [9]. The adaptive learning strategy is reflected in Google's Neural Machine Translation System. The system introduced in the article can learn complex mappings between the source language and the target language through large-scale neural network models, providing a powerful infrastructure for handling language variants such as cross reading [10].

2.3.3. Post editing and evaluation

The use of post editing and automation tools is crucial for improving translation quality. Exploring the Limits of Transfer Learning with a Unified Text to Text Transformer emphasizes the importance of fine-tuning the model after training, which can be seen as a post editing strategy. By fine-tuning data in specific fields, the model can better adapt to specific language phenomena such as text to vernacular reading [11]. In terms of evaluation, A Comprehensive Survey of Multilingual Neural Machine Translation points out that in addition to traditional BLEU scores, it is necessary to develop specific evaluation indicators and tools for reading comprehension, which helps to accurately measure the model's performance in dealing with such complex language phenomena [5].

3. STRATEGIES FOR NEURAL MACHINE TRANSLATION TO HANDLE THE PHENOMENON OF CLR

3.1. Dual Encoder Decoder Architecture Optimization

3.1.1. Optimized dual encoder design

The architecture's innovation is centered around its dual path separation mechanism, deploying specialized encoders for each path. One encoder is geared toward standard language (text reading), focusing on capturing the structural attributes of conventional grammar and vocabulary. The other is dedicated to colloquial or dialect variants (vernacular reading), concentrating on extracting the variant information of oral expression and dialect characteristics. The vector representations emitted by these two encoders are imbued with rich linguistic information reflective of their respective styles, setting the stage for subsequent integration and translation.

3.1.2. Advanced algorithms and operational optimization

Advanced fusion mechanisms, including dynamic weighted fusion and dual stream interaction, are utilized to seamlessly merge text and vernacular features, transcending simple concatenation. This enhancement ensures the model optimally utilizes information from both reading styles, enriching contextual understanding and yielding smoother, more accurate translations. An adaptive attention mechanism is integrated to dynamically balance the contribution of text and vernacular encoders during decoding. This allows the model to selectively emphasize formal or informal aspects according to the translation context, focusing on the most relevant source material components for precise output.

Training incorporates deep multitasking, expanding beyond core translation to include text/vernacular classification and semantic annotation. This broadens the model's stylistic awareness, enhances generalization, and reduces overfitting risks. Coupled with an adaptive learning rate and regularization, the model maintains robustness and consistently high translation quality across diverse text and vernacular scenarios.

3.2. Optimization Application of Sub word Representation

3.2.1. Efficient Sub word Unit Generation and Processing

The optimization practice of Byte Pair Encoding (BPE) involves generating subwords through the iterative merging of high-frequency character pairs [6]. To refine efficiency and flexibility, an iterative optimization strategy is employed, such as instituting a frequency threshold to dynamically recalibrate the merging strategy or taking into account grammatical structure information during the merging process to ensure the generated subwords adhere more closely to language rules. Moreover, for specific domains or linguistic traits, a tailored initialization character set can be established, which precludes domain terms or specific characters to expedite convergence and elevate translation quality.

3.2.2. Integration and Optimization of Advanced Sub word Representation in NMT

Deep optimization in open vocabulary translation through subword models employs advanced compression and deep learning techniques to dynamically manage sub word lists, enhancing efficiency and accuracy in generating and translating novel or infrequent words. Meta grammar modeling allows adaptive sub word adjustments during translation, boosting vocabulary creation.

Translation models' adaptability and generalization are augmented by sub word models, which facilitate learning of vocabulary construction, transliteration, transcription patterns, and morphological shifts. This is particularly beneficial for inflection-rich languages and texts with many proper nouns, improving translation quality and versatility through informed sub word boundary use. Rare word processing is intelligently refined by optimizing sub word models. Beyond basic sub word combinations, context-sensitive weighting and dynamic selection based on historical frequencies ensure translations better align with intended meanings and reduce ambiguities.

Dynamic attention mechanisms, integrated within sub word models, enable specialized attention heads to discern fine distinctions between standard and vernacular readings in source texts. Multi-head attention during decoding synthesizes these features, accurately conveying the original context and style, thereby enhancing translation naturalness and precision.

3.3. Tuning of Attention Mechanism and Adaptability to the Phenomenon of CLR

3.3.1. Fine processing of text vernacular reading data

Collect a corpus covering a wide range of instances of textual and colloquial reading, and meticulously annotate the textual and colloquial reading forms in each instance to ensure that the model can access a rich variety of variant samples.

Based on the phonetic and semantic features of textual and colloquial reading, the data is subdivided to provide multi-dimensional input information for the model, such as the hierarchical division of tone and intonation, as well as the distinction of lexical meaning.

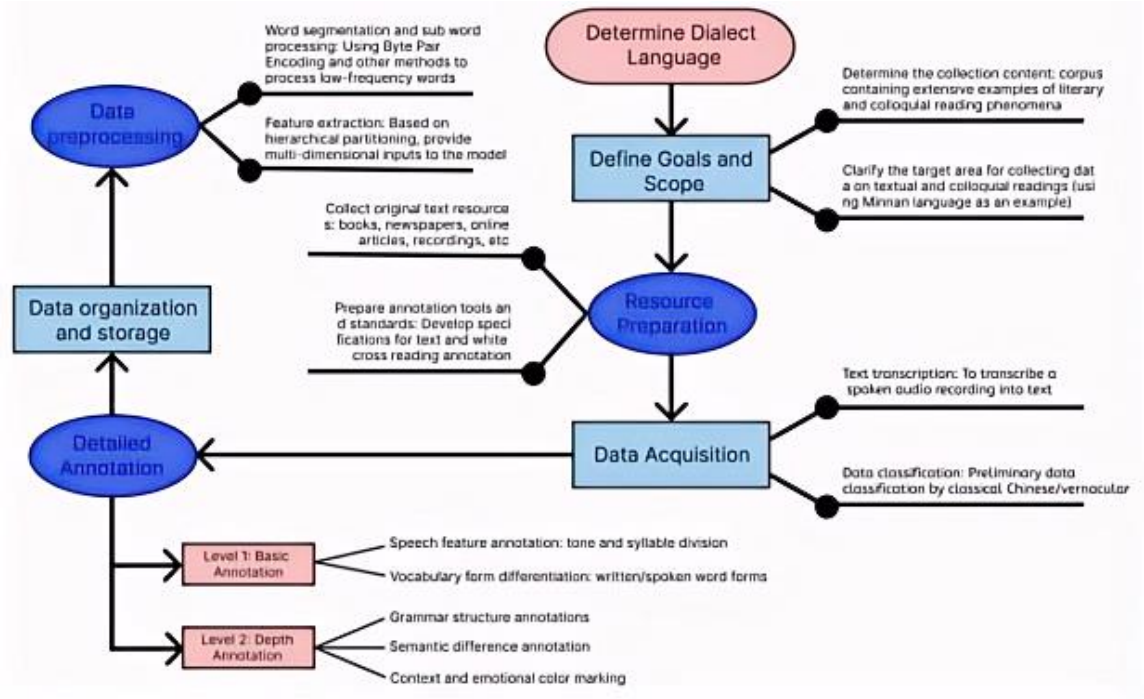

Figure 1. Data collection and annotation process for the phenomenon of CLR

3.3.2. Multidimensional attention mechanism design

Build a neural network with multi head attention, each focusing on different dimensions of information (such as grammatical structure, semantic content, and pronunciation features) to capture the complexity of textual and colloquial reading.

Based on the characteristics of textual and colloquial reading, a customized attention weight calculation method is designed. For example, additive attention mechanism is used to emphasize the matching of syntactic structures, while multiplicative attention is used to capture semantic similarity. Here is a matching design:

Additive attention mechanism: used to emphasize the matching of grammatical structures. This mechanism allocates attention weights by calculating the similarity of the grammatical relationship between each word in the source sentence and the current generated word. It is suitable for capturing structural features of sentences, such as subject verb object relationships.

Multiplicative attention mechanism: used to capture semantic similarity. This mechanism allocates attention weights by calculating the dot product of semantic vectors between each word in the source sentence and the current generated word, which is suitable for identifying semantic correlations.

Assuming we have a hidden state matrix $H = [h_1, h_2, ..., h_n]$ for a source language sentence, where h_i is the hidden state vector for the i-th word. On the target language side, the current hidden state of the generated word is s_t.

Let W_a and v_a be two learnable parameter matrices and vectors, respectively. The additive attention function can be defined as:

$$e_i = v_a^T \tanh(W_a[s_t;h_i]) \quad (1)$$

The multiplicative attention mechanism can directly calculate the dot product of s_t and h_i, and then obtain attention weights through softmax normalization. The formula is as follows:

$$e_i = s_t^T h_i \quad (2)$$

In practice, we can linearly combine the outputs of additive and multiplicative attention mechanisms, or use gating mechanisms to control the contributions of the two attention mechanisms. For example, a gating vector g can be defined to determine the mixing ratio of two mechanisms:

$$a_i = \sigma(g)\cdot\text{softmax}(e_i^{add}) + (1 - \sigma(g))\cdot\text{softmax}(e_i^{mult}) \quad (3)$$

Among them, σ is the sigmoid function, and e_i^{add} and e_i^{mult} represent the normalized scores of additive and multiplicative attention mechanisms, respectively.

3.3.3. Dynamic adaptive weight control

Initialize weights for each attention head and implement a dynamic adjustment mechanism tied to training progression. This ensures the model progressively learns optimal attention allocation for text and vernacular reading contexts. Through online learning strategies, utilizing translation quality feedback (such as BLEU scores) to dynamically adjust attention parameters, the model gradually optimizes its processing strategy for text vernacular reading phenomenon during the training process [7].

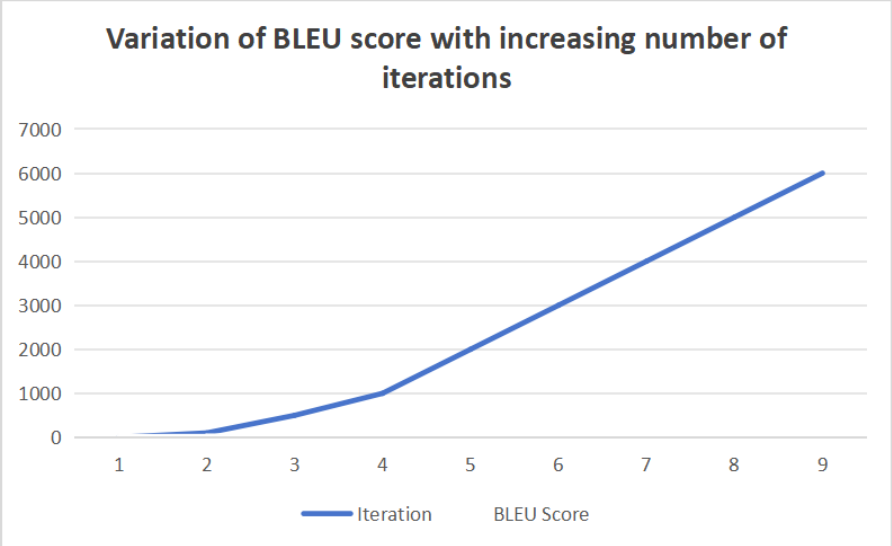

Figure 2. Simulated BLEU score changes with training iterations

3.3.4. Refined training and evaluation process

Reinforcement training strategy: Adopting reinforcement learning methods, the model directly optimizes the reward signal based on the accuracy of cross reading in translation tasks, promoting the model to learn more effective attention allocation strategies.

Careful monitoring and optimization: closely monitor the rationality and effectiveness of attention distribution during the training process, adjust the learning rate and regularization strategy in a timely manner based on the performance of the model on the development set, and prevent over fitting or under fitting.

3.3.5. Comprehensive evaluation and continuous iteration

By comprehensively using BLEU scores, manual reviews, and evaluation indicators specific to text and vernacular reading, ensure that the model can not only translate correctly but also maintain appropriate conversion of text and vernacular reading. Based on evaluation results, continuously fine-tuning model parameters or introducing transfer learning strategies, using pre trained models from other related tasks (such as speech recognition and text inclusion) to enhance the ability to understand text and vernacular reading, achieving higher quality translation.

3.3.6. Limitations of BLEU scores in evaluation

When paying attention to the phenomenon of textual and colloquial reading differences (CLR) in Minnan language, a minority language, there may be several significant differences in BLEU scores compared to training general language corpora:

BLEU might inadequately represent the authenticity and cultural appropriateness of translations for dialects like Minnan due to its pronunciation quirks and colloquial nuances. These elements, involving complex shifts in phonetics, lexicon, and context, typically elude BLEU's detection.

Universal language models often start with higher BLEU scores, leveraging vast parallel corpora that adhere to standardized linguistic norms. Conversely, Minnan-targeted models might need additional time and data to grasp cross-reading patterns, leading to lower initial BLEU scores.

When handling homophonous readings in Minnan, BLEU scores can be more volatile. This is particularly evident when incorporating new datasets or refining strategies (e.g., tweaking attention mechanisms), as the model adjusts to diverse reading styles.

Given these challenges, evaluating Minnan translations might necessitate integrating additional metrics such as Translation Error Rate (TER), chrF, and possibly human evaluations to more accurately gauge translation quality and proficiency in managing CLR phenomena.

3.4. Adaptive Learning Strategy Optimization

3.4.1. Deepening strategies for transfer learning and fine-tuning

WoThe working principle involves leveraging transfer learning with a robust, pre-trained base model from extensive multilingual or Standard Chinese corpora to capture comprehensive linguistic and semantic details. During fine-tuning, the focus shifts to domain-specific datasets and curated vernacular reading examples for precise supervised learning. The aim is to enable the model to accurately identify and appropriately process vernacular variations, akin to Google's NMT system's approach of optimizing for particular language pairs and integrating specialized instruction on diverse reading rules.

Algorithm update: During the fine-tuning process, a more flexible optimizer configuration is adopted, such as AdamW, which adds weight attenuation on top of Adam to help reduce model overfitting. In addition, adopting the Layer wise Adaptive Rate Scaling (LARS) strategy to adaptively adjust the learning rate based on the activation of each layer can further improve the learning efficiency of the model when dealing with delicate language differences such as text and vernacular reading.

3.4.2. Fine tuned application of reinforcement learning

Working principle: Incorporating reinforcement learning, the model refines translation strategies via environmental interaction and feedback, optimizing outputs. A reward mechanism boosts fidelity and fluency, guiding the model to adeptly handle textual and vernacular discrepancies through a seq2seq framework.

Innovation: Blending explorative and exploitative learning, a Curiosity Driven Learning mechanism spurs discovery of novel reading patterns while leveraging established rules. Algorithms like Proximal Policy Optimization (PPO) or Asynchronous Advantage Actor Critic (A3C) enhance policy update stability and learning efficacy, ideal for complex decision-making in text reading scenarios.

3.4.3. Intelligent adjustment of dynamic learning rate strategy

For the training of reading comprehension, a more intelligent learning rate adjustment strategy, such as One Cycle Policy, is adopted. This strategy first rapidly increases the learning rate to accelerate model convergence, and then gradually decreases to make fine adjustments. In addition, the introduction of an adaptive learning rate scheduler, such as Cosine Annealing with Warm Restarts,

not only automatically adjusts the learning rate based on the training stage, but also restarts when learning stops, stimulating the model to jump out of local optima.

Based on the characteristics of text vernacular reading data, design a dynamic learning rate adjustment function, and adaptively adjust the learning rate according to the model's performance on text vernacular reading instances. For example, when the model encounters difficulties in dealing with complex cross reading situations, it can temporarily increase the learning rate to accelerate learning; After familiarizing the model with common text to vernacular conversion rules, the learning rate is correspondingly slowed down to consolidate the learning results.

4. CONCLUSION

This study addresses the challenges in Neural Machine Translation (NMT) concerning textual and colloquial discrepancies in Chinese dialects, notably Minnan. It highlights the role of subword representation techniques and dynamic attention mechanisms in enhancing translation accuracy for low-resource dialects. Transfer learning and fine-tuning strategies with large-scale, target-language-similar models prove effective, with comprehensive Minnan text data boosting performance. Standardizing Minnan script aids quality improvement, while leveraging Mandarin monolingual corpora for Minnan translation enhancement is a future research avenue.

Despite advancements, unresolved issues persist: efficiently utilizing sparse parallel and abundant monolingual data, constructing high-quality, wide-ranging datasets for textual-colloquial differentiation, refining evaluation metrics, and precisely handling long-distance dependencies in translation contexts. These challenges underscore the need for continued innovation in NMT's adaptability to complex linguistic phenomena, focusing on low-resource dialects like Minnan. Future studies aim to tackle these obstacles, advancing NMT's capabilities in dialect translation and low-resource language processing.

REFERENCES

[1] Lin, X.F., Wu, X.F. (2015). Xiamen, Zhangzhou, and Quanzhou Minnan Dialects in the Perspective of Cross Strait Exchange. Southeast Academic, (06), 244-245.

[2] Li, R.L. (1963). The colloquial and literary readings of Xiamen Dielact. Xiamen University(Social Science Edition), (02), 58-59, 95-96.

[3] Chu, C.H., & Wang, R. (2018). A Survey of Domain Adaptation for Neural Machine Translation. Proceedings of the 27th International Conference on Computational Linguistics, 1304-1319.

[4] Raj Dabre, Chu, C.H. & Kunchukuttan, A. (2020). A Comprehensive Survey of Multilingual Neural Machine Translation, 111: 22-23.

[5] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L. & Polosukhin, I.. (2017). Attention is All You Need. In Advances in Neural information Processing Systems (pp.5998-6008), 2.

[6] Bahdanau, D, Cho, K, & Bengio, Y. (2015). Neural Machine Translation by Jointly Learning to Align and Translate. Proceedings of the International Conference on Learning Representations (ICLR), 3-4.

[7] Gibadulin., Valeev, A., Khusainova., & Khan, A. (2020). A Survey of Methods to Leverage Monolingual Data in Low-resource Neural Machine Translation, 6-7.

[8] Sennrich, R., Haddow, B, & Birch, A. (2016). Neural Machine Translation of Rare Words with sub word Units. Proceedings of the 54thAnnual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 1715-1725.

[9] Cho, K, Van Merrienboer, B., Gulcehre, Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation, 2-3.

[10] Wu, Y.H., Schuster, M., Chen, Z.F., Le, Q.V., Norouzi, M., Macherey, W et al. (2016). Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation, 2-4.

[11] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y.Q., Li, W., Liu, P.J.. (2020). Exploring the limits of Transfer Learning with a Unified Text-to-Text Transformer, /ournal of Machine Learning Research, 21(140), 1-67.