Dictionary-Injected Pre-training for Traditional Mongolian–Chinese Machine Translation

Zhiqiang Zhang, Yonghong Tian (✉), and Yehui Dang

College of Intelligent Science and Technology (College of Cyberspace Security), Inner Mongolia University of Technology, Hohhot 010080, China
tyh@imut.edu.cn

Abstract. In recent years, pretrained models have achieved remarkable progress in the field of natural language processing and have demonstrated superior performance across a variety of machine translation tasks. However, for the low-resource language pair of Traditional Mongolian and Chinese, the transfer capability of mainstream multilingual pretrained models is substantially limited, primarily because Traditional Mongolian is not covered in their pretraining corpora. To address this issue, this paper proposes a dictionary-injected pretraining method. Specifically, a constructed Mongolian–Chinese bilingual dictionary is used to perform lexical replacement on Chinese monolingual corpora, thereby generating mixed texts containing Traditional Mongolian. The model is then pretrained with a BART-style denoising autoencoding objective, enabling it to recover the original Chinese sentences under cross-lingual perturbations. In this way, the proposed method enhances the model's cross-lingual semantic understanding and improves Mongolian–Chinese machine translation performance. Experimental results show that, on the test sets for both Mongolian-to-Chinese and Chinese-to-Mongolian translation, the proposed method outperforms a strong BART baseline by 2.3 and 2.4 BLEU points, respectively, thereby confirming its effectiveness for Mongolian – Chinese machine translation tasks.

Keywords: Mongolian–Chinese bilingual dictionary; Mongolian–Chinese machine translation; pretrained models.

1 Introduction

In recent years, with the rapid development of deep learning technologies, neural machine translation has achieved remarkable progress on many high-resource language pair [1-3]. In terms of translation performance, neural machine translation (NMT) has comprehensively surpassed traditional phrase-based statistical machine translation models [4], and has become the standard implementation paradigm for machine translation service systems in industry [5]. Some researchers have even argued that, for certain domains and language pairs, the performance of NMT has approached or even exceeded that of human translators [6]. However, the strong performance of current NMT systems is largely predicated on the availability of large-scale, high-quality

parallel corpora. Constrained by factors such as limited market size and the high cost of data annotation, the quality of Mongolian–Chinese machine translation still lags far behind that of mainstream language pairs such as Chinese–English.

Under conditions of limited parallel data, monolingual data are generally more abundant and easier to collect. Consequently, researchers have naturally explored various methods for effectively leveraging both target-side and source-side monolingual data within different NMT frameworks. Among these approaches, back-translation is the simplest and most direct [7]. In this method, an initial reverse translation model is first trained in a supervised manner, and target-side monolingual data are then translated to construct additional synthetic parallel data for training the forward translation model. Back-translation can not only improve translation performance in low-resource scenarios, but also alleviate issues such as domain adaptation [8]. Over the past few years, inspired by advances in computer vision [9], self-supervised pre-training has sparked a surge of research interest in the field of natural language processing, driven jointly by massive amounts of unlabeled textual data, advanced distributed optimization techniques, powerful sequence learning models, and high-performance computing hardware [10-12]. Pre-trained models enable researchers to avoid training large and computationally expensive models from scratch; instead, they can directly adapt existing pre-trained models to downstream tasks through task-specific fine-tuning and thereby achieve strong performance. Among the many pre-trained models that have been proposed, representative examples include the masked language model BERT [10], the autoregressive language model GPT [13], the permutation language model XLNet, and the denoising autoencoding model BART. Among these models, BERT and XLNet are built upon the Transformer encoder and are capable of learning bidirectional representations of linguistic sequences, making them primarily suitable for semantic understanding tasks. GPT, by contrast, adopts the Transformer decoder and generates the target sequence token by token in an autoregressive manner, conditioned on the previously generated context and the current input. BART can be regarded as a more generalized pre-training framework that integrates the strengths of both BERT and GPT. Unlike either BERT or GPT, BART follows a sequence-to-sequence paradigm: various noise transformations are first applied to the input sequence on the encoder side, and the decoder is then trained to reconstruct the original sequence. Using denoising autoencoding as its optimization objective, BART jointly pre-trains both the encoder and the decoder, which makes it particularly well suited to encoder – decoder modeling tasks such as machine translation and question answering. BART was originally pre-trained for a single language, namely English. The subsequently proposed mBART [14] extends the BART training paradigm to multilingual settings for multilingual pre-training. Likewise, M2M-100 [15], which also adopts the BART-style training objective, further broadens the range of supported languages and enables many-to-many translation among 100 languages. For a low- to medium-resource language such as Mongolian, multilingual pre-training is a highly attractive paradigm, since, in addition to supporting multilingual translation, such large-scale pre-trained models can also facilitate the transfer of general semantic knowledge. However, neither mBART nor M2M-100 was trained on Mongolian data. Therefore, this paper aims to explore effective approaches for training

a BART-style Mongolian-Chinese pre-trained model, thereby laying a research foundation for subsequent Mongolian-Chinese machine translation studies.

Driven by the rapid advancement of artificial intelligence technologies and the growing demand for cross-lingual communication, Mongolian－Chinese machine translation has achieved substantial progress. These methods include data augmentation for translation improvement, unsupervised or semi-supervised approaches [16], reinforcement learning [17], adversarial learning , transfer learning [18], non-autoregressive Mongolian-Chinese machine translation [19], and training strategies assisted by pre-trained models [20]. In addition, there have also been some related studies on Mongolian pre-trained language models, such as CINO, a pre-trained language model for Chinese ethnic minority languages.

A bilingual dictionary can not only offer word-level translation correspondences, but can also be used to construct mixed-language texts through lexical substitution, thereby allowing the model to directly encounter target-language vocabulary during the pre-training stage and enhancing its cross-lingual semantic alignment capability. Furthermore, inspired by the observation that the use of mixed language in cross-lingual communication often improves communicative efficiency [21], this paper proposes a dictionary-injection-based pre-training method. Specifically, a Mongolian-Chinese bilingual dictionary is first constructed, and the dictionary is then used to partially replace words in Chinese monolingual corpora so as to generate mixed texts containing traditional Mongolian lexical items. Subsequently, the model is pre-trained with a BART-style denoising autoencoding objective, requiring it to reconstruct the original Chinese sentences under cross-lingual perturbations. In this way, the model learns lexical mapping relationships between Mongolian and Chinese and improves downstream translation performance. An example of dictionary-based data injection is shown in Table 1.

Table 1. Example of dictionary-based data injection

Type	The content of the sentence
Original sentence	新时代，中国人民安居乐业，城市建设越来越好，特别是内蒙古自治区的首府呼和浩特
Mixed sentence	ᠰᠢᠨ᠎ᠡ ᠴᠠᠭ ᠦᠶ᠎ᠡ，中国 ᠠᠷᠠᠳ ᠲᠦᠮᠡᠨ 安居乐业，ᠬᠣᠲᠠ 建设越来越好，特别是 ᠥᠪᠦᠷ ᠮᠣᠩᠭᠣᠯ ᠥᠪᠡᠷᠲᠡᠭᠡᠨ ᠵᠠᠰᠠᠬᠤ ᠣᠷᠣᠨ 的首府 ᠬᠥᠬᠡᠬᠣᠲᠠ

The replaced Chinese words can be regarded as a form of noise injected into the original text. This dictionary-based noise injection scheme not only perturbs the input sequence, but also provides the model with broad opportunities to learn bilingual knowledge associations. In essence, it encourages the model to perform decoding with denoising as the objective while jointly learning cross-lingual semantic alignment and sequential contextual information, thereby acquiring cross-lingual joint representations. This, in turn, facilitates effective fine-tuning under conditions of limited parallel data. In addition, the dictionary-injection-based pre-training method enables the model to

acquire domain-specific translation knowledge from bilingual dictionaries, thus offering a feasible and low-cost solution for domain adaptation in machine translation.In summary, the main contributions of this paper are as follows:This paper proposes a novel pretraining method for Mongolian–Chinese translation, namely the Pretraining Method of Dictionary Injection (PMDI). The method uses an existing multilingual pretrained model for parameter initialization, introduces Mongolian vocabulary into Chinese corpora to generate mixed texts, and performs pretraining with a BART-style denoising autoencoding objective. In this way, the model is encouraged to reconstruct the original Chinese sentences under cross-lingual perturbations, thereby improving its modeling capability for Traditional Mongolian.A Mongolian–Chinese bilingual dictionary is designed and used for mixed-corpus construction. The constructed bilingual dictionary is employed to replace Chinese words with Mongolian words, supporting the generation of mixed corpora as input for pretraining, although the dictionary itself is not emphasized as the primary contribution of this work.The effectiveness of the proposed method is validated on Mongolian–Chinese translation tasks. By applying the dictionary-injected pretraining method to a Mongolian–Chinese translation model, the proposed approach significantly improves BLEU scores over baseline models, demonstrating its practical value in low-resource language scenarios.

2 Method

2.1 NMT

Neural Machine Translation (NMT) directly models the conditional probability of translating a source-language sentence into a target-language sentence through an encoder–decoder framework. Given a source sentence$X = (x_1, x_2, \dots, x_m)$ and target sentence$Y = (y_1, y_2, \dots, y_n)$, its objective can be formulated as follows:

$$P(Y|X) = \prod_{t=1}^{n} P(y_t|y_{<t}, X; \theta) \quad (1)$$

In this equation, $y_{<t}$ denotes the target tokens generated prior to the current time step, and represents the model parameters. Currently, mainstream Neural Machine Translation (NMT) models predominantly adopt the Transformer architecture. By leveraging self-attention mechanisms to model long-range dependencies, these models have achieved remarkable performance across various translation tasks. However, for the traditional Mongolian-Chinese language pair, which is characterized as a low-resource scenario, standard NMT models still struggle to achieve ideal performance. This is primarily due to the scarcity of parallel corpora and the difficulties associated with cross-lingual word alignment.

2.2 Pretrained model mBART

mBART is a multilingual sequence-to-sequence pretrained model whose core idea is to construct corrupted inputs from large-scale monolingual corpora and recover the original text through a denoising autoencoding task, thereby learning general contextual representation and generation capabilities. Let the original sentence be X, and the

corrupted input be $\tilde{X}$. Its pretraining objective can be formulated as follows:

$$L_{pre} = -\sum_{t=1}^{n} log\, P(X_t | x_{<t}, \tilde{X}; \quad (2)$$

Here, $\tilde{X}$ denotes the input sequence corrupted by a noise function. During pretraining, mBART adopts several noising strategies, including word masking, sentence order permutation, document rotation, word deletion, and text span infilling, as illustrated in Fig. 1. The goal of these noising operations is to disrupt the surface structure of the original sentence and thus encourage the model to learn more robust semantic representations. In downstream machine translation tasks, mBART is typically first pretrained on large-scale multilingual monolingual corpora and then fine-tuned on parallel corpora, which helps improve performance in low-resource translation scenarios. However, mainstream multilingual pretrained models generally do not provide sufficient coverage of Traditional Mongolian, which limits their cross-lingual transfer ability in Mongolian–Chinese translation tasks. Based on this, the present study extends the mBART denoising pretraining framework by introducing bilingual-dictionary-based cross-lingual lexical replacement noise, enabling the model to explicitly learn lexical mapping relationships between Traditional Mongolian and Chinese during the pretraining stage.

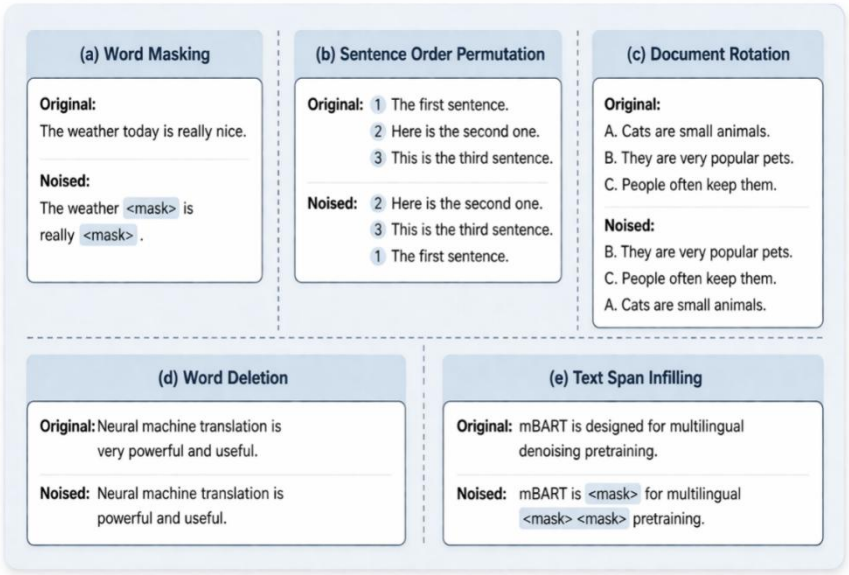

Fig. 1. Schematic diagram of five noise addition methods for mBART

2.3 Bilingual Dictionary-Based Pretraining Method

To address the problem that mainstream multilingual pretrained models do not cover Traditional Mongolian, resulting in insufficient transfer capability for low-resource

Mongolian–Chinese translation tasks, this paper proposes a dictionary-injected pretraining method. Building upon the mBART denoising pretraining framework, the proposed method introduces a Mongolian–Chinese bilingual dictionary to construct cross-lingual mixed texts, enabling the model to explicitly learn lexical mapping relationships between Traditional Mongolian and Chinese during the pretraining stage, thereby improving downstream performance in both Mongolian-to-Chinese and Chinese-to-Mongolian translation. However, unlike BART, PMDI adopts dictionary injection as the primary cross-lingual perturbation mechanism to replace conventional random noising strategies. Dictionary injection not only serves to corrupt the original input, but also encourages the encoder to learn cross-lingual semantic alignment information, thereby enabling the model to acquire more effective cross-lingual joint representations.

First, a Mongolian–Chinese bilingual dictionary is constructed on the basis of existing Mongolian–Chinese parallel corpora and word alignment results, denoted as:

$$D = \{(c_i, m_i)\}_{i=1}^{K} \tag{3}$$

Here, c_i denotes a Chinese lexical item, m_i denotes its corresponding Traditional Mongolian lexical item, and K represents the size of the dictionary. This bilingual dictionary provides reliable cross-lingual replacement resources for subsequent mixed-text construction.

On this basis, the present study performs partial lexical replacement on Chinese monolingual corpora using the bilingual dictionary, thereby generating Chinese–Mongolian mixed texts containing Traditional Mongolian vocabulary. Let the original Chinese sentence be $X = (x_1, x_2, \dots, x_n)$. For each word x_i in the sentence, if it exists in the bilingual dictionary D, it is replaced with its corresponding Traditional Mongolian equivalent $d(x_i)$ with probability p, thereby yielding the mixed text $\tilde{X}$. The replacement process can be formulated as follows:

$$\tilde{x}_i = \begin{cases} d(x_i), & x_i \in D \text{且} r < p \\ x_i, otherwise \end{cases} \tag{4}$$

Here, r is a random variable, p denotes the replacement probability, and $d(x_i)$ represents the Traditional Mongolian translation of x_i in the bilingual dictionary. Unlike conventional random masking noise, this replacement strategy introduces target-language words with explicit semantic correspondences, thereby providing the model with explicit cross-lingual alignment information.

After the mixed text is generated, this paper adopts a BART-style denoising autoencoding task for pretraining, taking the mixed text (\tilde{X}) as the input and the original Chinese sentence (X) as the reconstruction target, so that the model learns to recover the original sentence under cross-lingual perturbations. The corresponding training objective is defined as follows:

$$L_{dict} = -\sum_{t=1}^{n} logP\left(x_t \middle| x_{<t}, \tilde{X}; \theta\right) \tag{5}$$

In this process, the model is required not only to understand the contextual semantics of the sentence, but also to establish mapping relationships between Traditional Mongolian words and Chinese words, thereby enhancing its cross-lingual representation capability. After the dictionary-injected pretraining stage is completed, the resulting model parameters are used to initialize the downstream translation model, which is then fine-tuned in a supervised manner on Mongolian–Chinese parallel corpora to verify the effectiveness of the proposed method on both Mongolian-to-Chinese and Chinese-to-Mongolian translation tasks.Overall, the core idea of the proposed method lies in introducing Traditional Mongolian vocabulary into the pretraining stage in advance through a bilingual dictionary, so that the model can learn Mongolian–Chinese lexical correspondences while reconstructing the original text. In this way, the proposed approach compensates for the insufficient coverage of Traditional Mongolian in mainstream pretrained models and improves the performance of low-resource Mongolian–Chinese machine translation. The specific pre-training and fine-tuning processes are illustrated in Fig. 2 and Fig. 3.

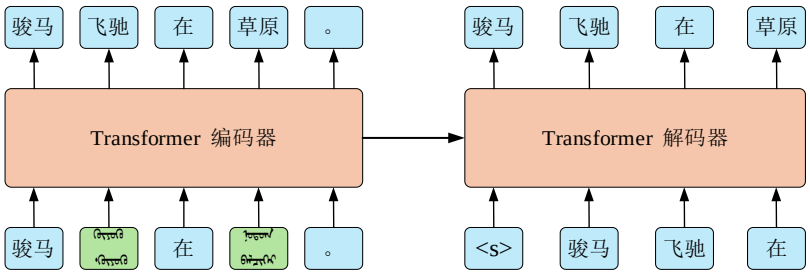

Fig. 2. Pre-training process

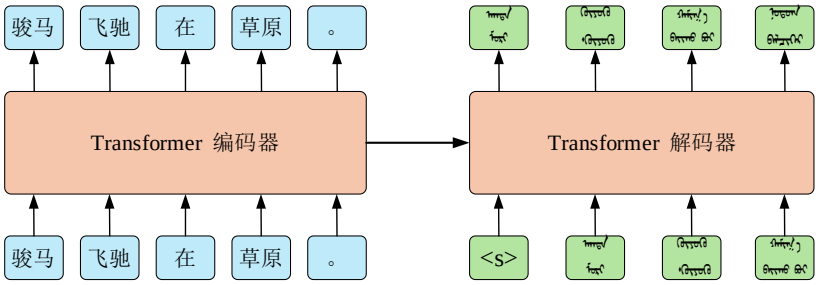

Fig. 3. Fine-tuning process

3 Experiment

3.1 Data Setting

Dataset. The bilingual data used in this study are derived from an existing Traditional

Mongolian–Chinese parallel corpus maintained by our laboratory, containing approximately 350,000 sentence pairs. The corpus covers a wide range of text types, including government work reports, news texts, legal and official documents, daily expressions, novel excerpts, as well as online dialogues and chat data. This broad domain diversity enables the corpus to provide relatively rich linguistic phenomena and translation examples for model training. To enable the subsequent model to learn bilingual correspondence knowledge from the corpus more effectively, the raw parallel data were first cleaned and normalized. For the divided datasets, punctuation normalization was first conducted to reduce vocabulary redundancy caused by inconsistent punctuation forms. Abnormal sentences consisting entirely of digits or containing excessive meaningless symbols were then removed to reduce the interference of noisy data in model training. In addition, to control the quality of training samples and improve training efficiency, sentence length filtering was further applied to discard excessively long sentences while retaining parallel sentence pairs with a relatively reasonable length distribution. After the above preprocessing steps, a high-quality Mongolian–Chinese bilingual dataset was obtained for subsequent experiments. The preprocessing methods for monolingual data and parallel data are consistent. Both involve word segmentation followed by sub-word learning. After preprocessing steps such as regular denoising and deduplication, the final retained sizes of Mongolian and Chinese monolingual data are 5.4M and 4.3M, respectively. The final data scale is shown in Table 2.

Table 2. Parallel data and monolingual scale

Data type	Parallel data/sentence pairs	Monolingual data/sentence pairs
Training set	350K	5.3M/4.2M
Validation set	3.5K	52K/41K
Test set	3.5K	52K/41K
Total	357K	5.4M/4.3M

Bilingual Dictionary. The bilingual dictionary is a key resource in the proposed method and is primarily used to perform lexical replacement on Chinese monolingual corpora in order to construct Chinese–Mongolian mixed texts containing Traditional Mongolian vocabulary. Specifically, initial word pairs are first extracted on the basis of word alignment results from Mongolian–Chinese parallel corpora. Subsequently, Fast Text is used to train Chinese and Mongolian monolingual word embeddings, and the cross-lingual word embedding spaces are aligned through methods such as iterative normalization and adversarial mapping. Finally, nearest-neighbor retrieval is performed using the CSLS similarity metric to generate a large-scale Mongolian–Chinese bilingual dictionary that includes both common words and nouns. The construction process is shown in Fig. 4. This construction strategy makes it possible to obtain relatively high-quality cross-lingual lexical correspondences under limited manually annotated resources, thereby providing support for dictionary injection in the subsequent pretraining stage.During dictionary usage, this study mainly selects lexical

items with relatively clear translation correspondences as replacement candidates, so as to ensure that the resulting mixed texts preserve the main semantic structure of the original sentences while introducing effective cross-lingual lexical information. s also pointed out in the research proposal, the bilingual dictionary, as a source of prior translation knowledge, can bring the vector representations of aligned words across the two languages closer together and enhance the model's cross-lingual transfer and semantic representation ability through dictionary injection. The final Mongolian–Chinese bilingual dictionary constructed in this study contains 283,713 entries, mainly including high-frequency common words, frequently used nouns, and some domain-specific vocabulary.

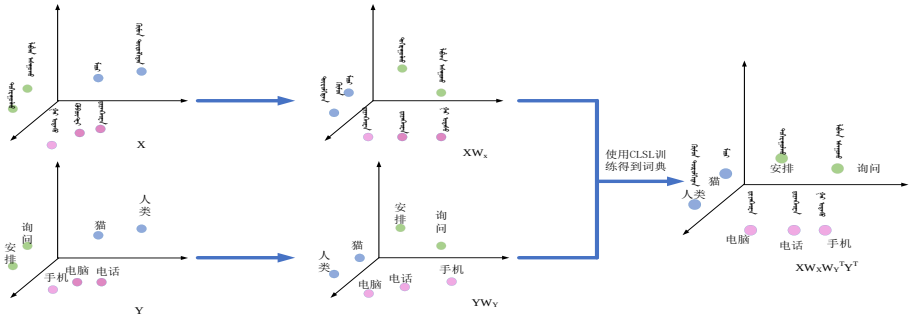

Fig. 4. Bilingual Dictionary construction process

3.2 Main Experimental Results

To evaluate the effectiveness of the proposed method, experiments are conducted on both Traditional Mongolian-to-Chinese (Mn→Zh) and Chinese-to-Traditional Mongolian (Zh→Mn) translation tasks. The proposed dictionary-injected pretraining model (PMDI) is compared with several baseline systems, including a strong BART baseline, Transformer, and a back-translation model. The results are summarized in Table 3, where PMDI consistently achieves superior performance in both translation directions. Specifically, PMDI improves the BLEU score by 2.3 points over the strong BART baseline on the Mn→Zh task, and by 2.4 points on the Zh→Mn task. These results demonstrate that the proposed dictionary-injection-based pretraining strategy effectively enhances the model's ability to capture cross-lingual lexical correspondences, thereby significantly improving translation quality in low-resource scenarios.

Further analysis shows that PMDI yields consistent performance gains in both directions, indicating that the proposed method is effective not only for Mongolian-to-Chinese translation but also generalizes well to the reverse direction. This can be attributed to the fact that conventional denoising pretraining primarily focuses on recovering monolingual context, whereas the proposed method introduces mixed Mongolian–Chinese text through bilingual dictionary substitution. As a result, the model is exposed to Traditional Mongolian tokens during pretraining and is encouraged

to explicitly learn lexical correspondences between Mongolian and Chinese when reconstructing the original sentence. In this way, the model not only preserves the contextual modeling capability of the original pretraining framework, but also strengthens cross-lingual semantic alignment. Moreover, as illustrated in Fig. 5 and Fig. 6, PMDI exhibits better convergence behavior in terms of validation BLEU and training loss.

Table 3. The BLEU scores of each model on the test set

Model	Mn→Zh	Zh→Mn
Transformer	27.3	27
Back Translation	28.1	27.3
BART	29.5	29.2
PMDI	31.8	31.6

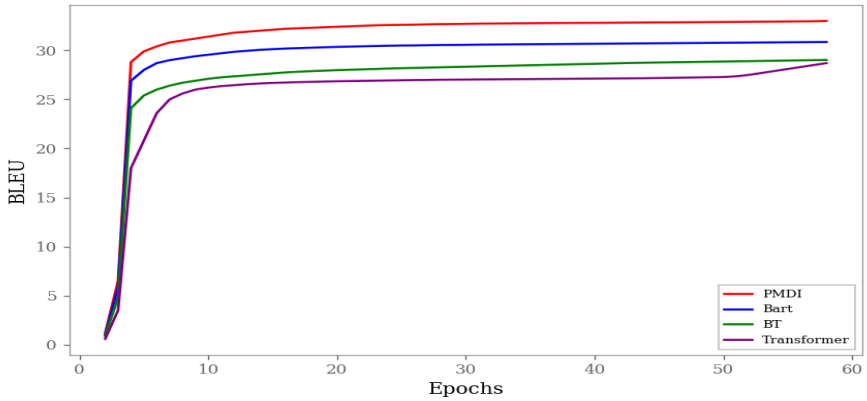

Fig. 5. The BLEU score variations of various models

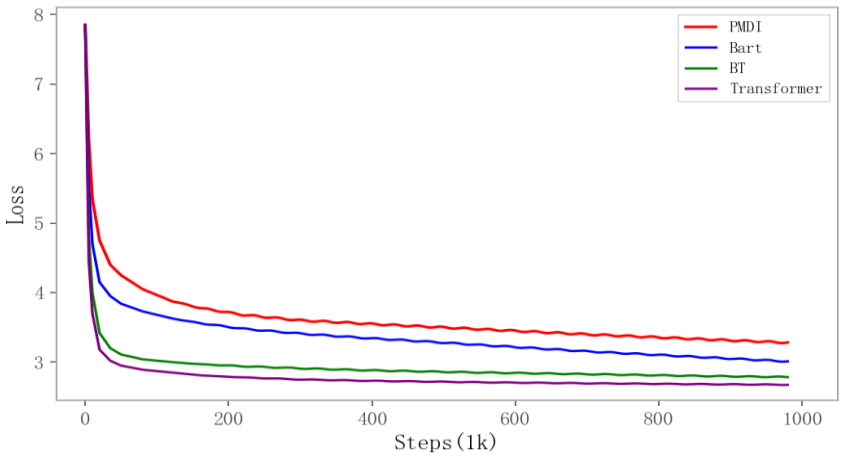

Fig. 6. Training loss variation of each model

To further verify the effectiveness of the proposed method in a concrete translation scenario, we select a representative example from the test set for qualitative analysis. The original Traditional Mongolian sentence in this example is "ᠳᠠᠷᠤᠭᠠᠨ ᠤ ᠮᠣᠩᠭᠣᠯ ᠡᠮᠡᠨᠡᠯᠢᠭ᠎ᠦᠨ ᠰᠡᠷᠢᠨ ᠵᠢᠨᠰᠦᠨ ᠤᠷᠤᠰᠬᠠᠯ ᠪᠣᠯᠤᠨ ᠰᠠᠯᠬᠢᠨ ᠤᠷᠤᠰᠬᠠᠯ ᠤᠨ ᠵᠢᠨᠰᠦᠨ ᠤᠷᠤᠰᠬᠠᠯ ᠤᠨ ᠲᠤᠷᠰᠢᠯᠲᠠ ᠵᠢᠨᠰᠦᠨ ᠤᠷᠤᠰᠬᠠᠯ ᠤᠨ ᠲᠤᠷᠰᠢᠯᠲᠠ", and its reference translation in Chinese is "蒙医药研究人员开展沙棘活性成分提取实验". The translations generated by different models for this example are presented in Table 4. In the test example, the PMDI model is able to accurately translate domain-specific terms such as "sea buckthorn active ingredient" and "extraction", and the generated translation is superior to those of the other baseline models in terms of semantic completeness and fluency. In contrast, the Transformer model is more prone to problems such as lexical drift and inappropriate collocations. Although the back-translation model improves translation quality to some extent, it still lacks accuracy in translating low-frequency words and fixed expressions. The strong BART baseline, on the other hand, exhibits occasional omissions or unnatural expressions in certain lexical items.

Table 4. Translation example from the test set

Model	Translation
Transformer	蒙医药研究人员研究活沙棘提取实验
Back Translation	蒙医药的研究人员正在开展活性成分的沙棘实验
BART	来自蒙医药的沙棘研究所的研究人员开展了沙棘活性成分的提取实验
PMDI	来自蒙医药的研究人员开展了沙棘活性成分提取实验

3.3 Ablation Study on the Dictionary Replacement Ratio

To investigate the effect of dictionary injection intensity on model performance, we further conduct an ablation study on the replacement ratio. Under the premise that all other experimental settings remain unchanged, only the dictionary replacement ratio during the pre-training stage is varied, with the values set to 10%, 20%, 30%, 40%, and 50%, respectively. Model performance is then evaluated on both the Traditional Mongolian-to-Chinese and Chinese-to-Traditional Mongolian translation tasks. The experimental results are presented in Table 5.

Table 5. BLEU scores under different dictionary replacement ratios

Replacement Ratio	Mn→Zh	Zh→Mn
10%	29.9	29.8
20%	31.0	31.4
30%	31.8	31.6
40%	30.7	30.5
50%	29.7	29.5

The experimental results show that the dictionary replacement ratio is an important factor affecting translation performance. When the replacement ratio is relatively low, the amount of cross-lingual lexical information introduced into the mixed corpus is insufficient, making it difficult for the model to fully learn the lexical mapping relationships between Traditional Mongolian and Chinese. As a result, the overall performance improvement remains limited. As the replacement ratio gradually increases, the model is exposed to more Traditional Mongolian words with explicit translation correspondences during pre-training, which strengthens its cross-lingual semantic alignment ability and consequently improves translation performance. However, when the replacement ratio becomes excessively high, a large number of lexical substitutions may disrupt the semantic coherence and contextual structure of the original sentences, thereby substantially increasing the difficulty of the denoising reconstruction task and weakening the effectiveness of pre-training. This indicates that dictionary injection must strike a balance between preserving contextual information and introducing cross-lingual perturbations, rather than simply benefiting from a higher replacement ratio. A comprehensive comparison of the results under different settings shows that a replacement ratio of 30% achieves the best overall performance. This ratio not only provides the model with sufficient cross-lingual lexical alignment signals, but also preserves the semantic structure of the original sentences to a large extent. Therefore, 30% is adopted as the default replacement ratio in the subsequent main experiments.

4 Conclusion

This paper proposes a dictionary-injection-based pre-training method for Traditional Mongolian–Chinese machine translation, inspired by the observation that mixed-

language usage in bilingual communication can facilitate interaction. By injecting bilingual dictionary entries into large-scale Mongolian and Chinese monolingual corpora and adopting a BART-style denoising autoencoding objective, the proposed method enables the model to learn cross-lingual lexical correspondences during pre-training and then adapt effectively to downstream translation tasks through supervised fine-tuning. Experimental results show that the proposed method consistently outperforms a strong BART baseline, yielding BLEU improvements of 2.3 and 2.4 points on the Mongolian-to-Chinese and Chinese-to-Mongolian test sets, respectively. Furthermore, the ablation study on replacement ratio demonstrates that a moderate level of dictionary injection is most effective, suggesting that translation performance depends on a proper trade-off between contextual preservation and cross-lingual perturbation. Overall, the results verify the effectiveness of dictionary injection for low-resource Traditional Mongolian–Chinese machine translation. In future work, we will investigate larger-scale monolingual data, more precise dictionary construction and injection strategies, and more effective hybrid pre-training frameworks to further enhance translation quality.

Acknowledgment. This research is supported by National Natural Science Foundation of China (Grant No. 62466043).

References

1. Sutskever I, Vinyals O, Le Q V. Sequence to sequence learning with neural networks[C]//Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems. Quebec, Canada, 2014: 3104-3112.
2. Gehring J, Auli M, Grangier D, et al. Convolutional sequence to sequence learning[C]//Proceedings of the 34th International Conference on Machine Learning. Australia, 2017: 1243-1252.
3. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems. USA, 2017: 6000-6010.
4. BROWN P F, COCKE J, DELLA PIETRA S A, et al. A statistical approach to machine translation[J]. Computational Linguistics, 1990, 16(2): 79-85.
5. WU Y, SCHUSTER M, CHEN Z, et al. Google's neural machine translation system: Bridging the gap between human and machine translation[C]//Advances in Neural Information Processing Systems 27 (NIPS 2014). 2016.
6. HASSAN H, AUE A, CHEN C, et al. Achieving human parity on automatic Chinese to English news translation[J/OL]. arXiv preprint arXiv:1803.05567, 2018.
7. SENNRICH R, HADDOW B, BIRCH A. Improving neural machine translation models with monolingual data[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Berlin, Germany: Association for Computational Linguistics, 2016: 86-96.
8. KUMARI S, JAISWAL N, PATIDAR M, et al. Domain adaptation for NMT via filtered iterative back-translation[C]//Proceedings of the 2nd Workshop on Domain Adaptation for NLP. Kyiv, Ukraine: Association for Computational Linguistics, 2021: 263-271.
9. HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos, CA, USA: IEEE Computer Society, 2016: 770-778.

10. DEVLIN J, CHANG M W, LEE K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg, PA, USA: Association for Computational Linguistics, 2019: 4171-4186.
11. BROWN T B, MANN B, RYDER N, et al. Language models are few-shot learners[C]//Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020. Red Hook, NY, USA: Curran Associates, Inc., 2020: 1877-1901.
12. LIU Y, OTT M, GOYAL N, et al. RoBERTa: A robustly optimized BERT pretraining approach[EB/OL]. arXiv preprint arXiv:1907.11692, 2019.
13. RADFORD A, WU J, CHILD R, et al. Language models are unsupervised multitask learners[EB/OL].https://cdn.openai.com/better-language-models/language-models-are-unsupervised-multitask-learners.pdf, 2019[2022-04-15].
14. LIU Y, GU J, GOYAL N, et al. Multilingual denoising pre-training for neural machine translation[J]. Transactions of the Association for Computational Linguistics, 2020, 8: 726-742.
15. FAN A, BHOSALE S, SCHWENK H, et al. Beyond English-centric multilingual machine translation[J]. Journal of Machine Learning Research, 2021, 22(107): 1-48.
16. SUN S, HOU H X, WU N, et al. Dual learning Mongolian-Chinese machine translation based on iterative knowledge refining[J]. Journal of Xiamen University (Natural Science Edition), 2021, 60(4): 687-692.
17. BAIT G. Mongolian-Chinese Neural Network Machine Translation Based on Reinforcement Learning[D]. Hohhot: Inner Mongolia University, 2020.
18. WANG Y F, SU Y L, ZHAO Y P, et al. Mongolian-Chinese neural machine translation model based on parameter migration[J]. Computer Applications and Software, 2020, 37(9): 87-93.
19. GU J, BRADBURY J, XIONG C, et al. Non-autoregressive neural machine translation[J]. arXiv: 1711.02281, 2017.
20. WENG R, YU H, HUANG S, et al. Acquiring Knowledge from Pre-trained Model to Neural Machine Translation[C]//Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(06): 9266-9273.
21. MATRAS Y. Mixed languages: A functional-communicative approach[J]. Bilingualism: Language and Cognition, 2000, 3(2): 79-99.