Natural Language Generation with Transformer Self-Translation: A Case Study of Multilingual Translation Models

Yewei Zhang[1], *Yatao Mu*[1*], *Jiawei Zhang*[2]

[1] *Vocational & Technical College, Inner Mongolia Agricultural University, China*
[2] *Chi Feng No. 2 Middle School, China*
15547153143@163.com, yeweikiki@aliyun.com, jinhecao93@sina.com

Abstract

In the context of large-scale deep learning models, natural language generation systems are required to handle multiple tasks concurrently. However, the addition of a new task typically necessitates retraining the model from scratch with both the original and new data, resulting in considerable time and resource consumption. Moreover, as the number of supported tasks grows, models with a fixed number of parameters may face capacity limitations, which can degrade overall performance. To address this challenge, this study introduces a Transformer-based self-back-translation approach for natural language generation, termed TransNMT, using multilingual translation as a case study. This method modularizes the model to enable dynamic scalability, effectively mitigating the capacity constraints posed by fixed parameters. Furthermore, a self-back-translation mechanism is designed for the TransNMT model, consisting of both forward and backward translation, which refines the model's performance internally while reducing external noise. This approach allows the model to perform well, particularly in low-resource translation tasks. Experimental results demonstrate significant improvements in BLEU scores across four low-resource and three high-resource language datasets, with the highest improvement reaching 2.7 BLEU points in one of the low-resource languages.

Keywords: Natural language processing, Natural language generation, Multilingual translation

1 Introduction

In recent years, Neural Machine Translation (NMT) has achieved significant improvements in translation quality when processing tens of millions of sentences. However, the translation quality of NMT is not satisfactory when dealing with low-resource languages with limited training data. To address the scarcity of training data for low-resource languages, two main solutions exist: one is to adopt data augmentation strategies by back-translating existing data and synthesizing parallel data to expand the training corpus; the other is to utilize multilingual NMT models that support translation between multiple languages, adjusting these models to enhance the translation quality of low-resource languages.

The data augmentation method expands the training data by back-translating existing low-resource data and combining it with original data to create new parallel data, without directly collecting new data. However, this method does not fully explore the deep semantic information of low-resource data, and the augmented data often contains a certain level of noise, which can adversely affect the model training process and reduce model performance.

Multilingual NMT models typically employ multi-task learning, training on data containing multiple translation language pairs. Previous research has found that multilingual NMT can generally improve translation quality between low-resource language pairs. Multilingual NMT even possesses the ability to translate between language pairs not included in the training data, known as "zero-shot" translation capability, which is highly valuable in practical applications since collecting translation data for all language pairs is often challenging. The current mainstream multilingual NMT model structure generally adopts the classic Sequence-to-sequence (Seq2Seq) model. However, as the number of languages supported by the Seq2Seq model continues to increase, translation performance may decline due to capacity bottlenecks in model parameters. Furthermore, since model parameters are shared across all languages, adding support for a new language requires retraining the model using all language data, which can be time-consuming and computationally expensive.

To address this issue, this research proposes a natural language generation method based on Transformer self-back-translation, named TransNMT, leveraging multilingual NMT as its foundation. This method not only maintains the model's performance on existing tasks but also further enhances its translation quality on low-resource translation tasks.

The main contributions of this paper are as follows:

- An M2TAB module based on the Transformer attention mechanism is proposed for multilingual NMT models. This module enables the NMT model to form language-agnostic interlingual representations, thereby enhancing the zero-shot translation capability of the NMT model and facilitating translation tasks for low-resource languages.

*Corresponding Author: Yatao Mu; Email: yeweikiki@aliyun.com

DOI: https://doi.org/10.70003/160792642025112606005

- The M2TAB module is equipped with a Mixture of Expert (MoE) module to support dynamic expansion and alleviate capacity bottlenecks. By modularizing M2TAB, the maintainability of the improved NMT model is enhanced.
- A self-back-translation mechanism for both forward and backward translation is defined in the TransNMT model, eliminating the need for manually designed data augmentation schemes. This internal optimization avoids external noise and further improves the model's translation quality on low-resource translation tasks.

2 Related Work

2.1 Neural Machine Translation

Kalchbrenner et al. [1] were the first to introduce neural network models into the field of machine translation, utilizing Recurrent Neural Networks (RNNs) for generating translations and integrating Convolutional Neural Networks (CNNs) to optimize source text processing, thereby opening up a new avenue for machine translation. Sutskever et al. [2] constructed an RNN-based encoder-decoder architecture through the sequence-to-sequence (seq2seq) learning framework, where the encoder encodes the source text and the decoder generates the translation. This end-to-end structure simplifies the complex processes of traditional translation systems and serves as the foundation for Neural Machine Translation (NMT). However, RNN models are prone to gradient vanishing/exploding issues when dealing with long sentences, affecting translation quality. To address this issue, Sak et al. [3] proposed Long Short-Term Memory (LSTM) networks, while Cho et al. [4] further streamlined the structure into Gated Recurrent Units (GRUs), effectively mitigating gradient challenges. RNN models based on the encoder-decoder architecture have continued to attract in-depth research and improvements due to their superior performance. Building upon this foundation, Bahdanau et al. [5] innovatively introduced the attention mechanism into neural machine translation, presenting the RNNSearch model. This model enables the decoder to focus on critical parts of the source text, effectively handling long-distance semantic dependencies.

As research progressed, more neural network structures were incorporated into machine translation. In 2015, Meng et al. [6] integrated CNNs into statistical machine translation, while Gehring et al. [7] designed a fully convolutional encoder-decoder architecture in 2017, enabling parallel encoding of source text, significantly enhancing translation efficiency and quality. In the same year, Vaswani et al. [8] introduced the Transformer model, leveraging its unique self-attention mechanism and "multi-head" attention calculation to achieve parallel processing and efficient learning of information. Its translation performance far surpasses RNN and CNN models, making it the current mainstream translation model.

2.2 Low-Resource Multilingual Translation

Ha et al. [9] were the first to propose constructing a multilingual neural machine translation (NMT) model within a unified framework, utilizing manual tags on source language corpora to indicate target languages, thereby reducing model complexity. Subsequent research showed that training a single NMT model directly on multilingual corpora, without additional tagging, can naturally adapt to multiple language pairs for translation [10]. Given NMT's heavy reliance on data, translation quality in low-resource settings is inherently limited. To address the scarcity of corpora, two primary strategies are employed: data augmentation and multilingual modeling. Data augmentation involves techniques such as synonym replacement, reordering, and back-translation to increase training data. For example, Hinton et al. [11] treated monolingual corpora as bilingual data with missing parallel sentences to construct new corpora, while back-translation leverages existing NMT models for bidirectional translation to expand parallel corpora [12]. Li et al. [13] generated pseudo-parallel data through back-translation to synthesize new corpora and applied filtering methods. Artetxe et al. [14] employed iterative back-translation until translation quality ceased to improve. Although back-translation can enhance translation quality, especially in low-resource scenarios [15-16], it adds a preprocessing step that consumes resources and may amplify errors between independent models, degrading the quality of training data.

Multilingual modeling approaches aim to improve translation quality for low-resource languages by adjusting model structures and parameters. Gu et al. [17] applied meta-learning algorithms to view low-resource translation as a meta-learning problem, enhancing model adaptability. Gu et al. [18] proposed a multi-resource boosting strategy, sharing vocabulary and sentence representations from multiple source languages to the target language, facilitating low-resource language learning. Kocmi et al. [19] implemented transfer learning to transfer knowledge from high-resource to low-resource language models. Kong et al. [20] designed a multilingual deep encoder method to share lexical information, improving word representation learning. Xia et al. [21] leveraged machine translation to induce multilingual training data from abundant English data, expanding the scale and diversity of training. Qin et al. [22] fine-tuned pre-trained BERT models to align multilingual representations through mixed contextual information, enhancing cross-lingual translation capabilities. Singh et al. [23] incorporated cross-lingual features from similar languages into multilingual models, specifically targeting the improvement of low-resource language translation quality.

Despite these methods significantly improving translation quality in low-resource environments, model adjustments may lead to overfitting issues. Therefore, achieving cross-lingual semantic generalization to further enhance low-resource language translation quality remains an urgent research topic requiring intensive investigation.

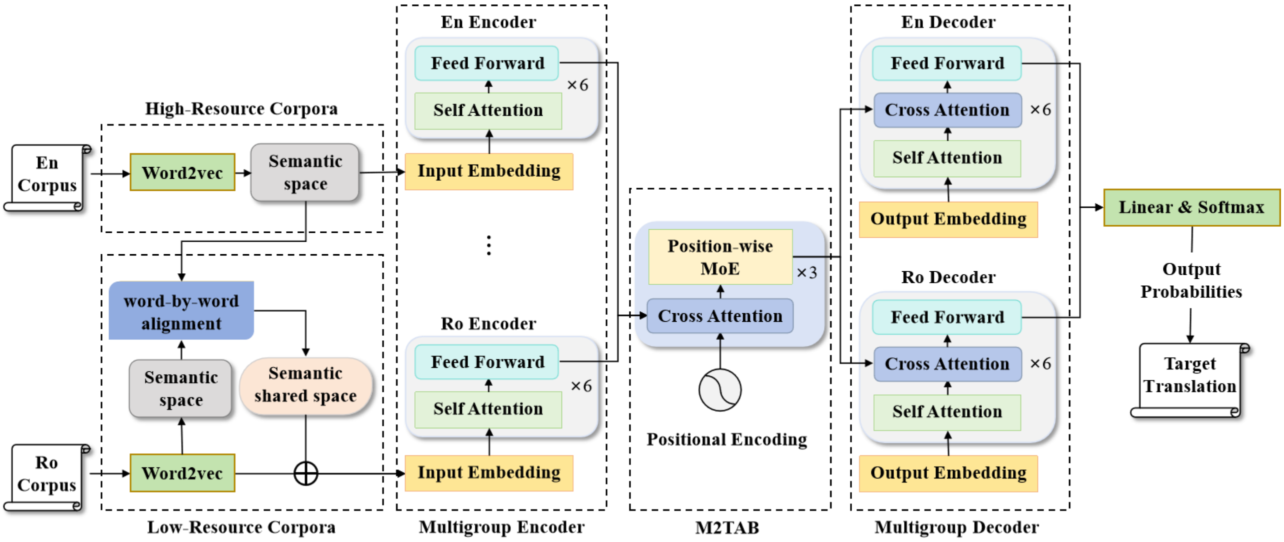

Figure 1. The overall structure of the TransNMT model

3 Method

3.1 Overall Framework

The architecture of the TransNMT model proposed in this study is illustrated in Figure 1. During the model training phase, we utilize the word2vec algorithm to conduct word embedding training for corpora of both high-resource and low-resource languages, aiming to obtain semantic space representations for their respective vocabularies. For the semantic space of high-resource languages, we directly employ them as pre-trained word embeddings and input them into the corresponding language's encoder. In contrast, for the semantic space of low-resource languages, we integrate it into the semantic space of high-resource languages through word-by-word alignment techniques, creating a shared semantic space encompassing both high and low-resource languages. Within this shared semantic space, we derive representations for low-resource words based on the representations of high-resource words, and these derived representations are then used as word embeddings and input into the encoders of their respective languages.

Subsequently, through the forward translation module of TransNMT, we obtain the forward translation results, which are then used as input for training the backward translation module of TransNMT. During the training process, we leverage the loss from backward translation to optimize the forward translation. This self-back-translation mechanism enables the predicted sentences generated by forward translation to learn potential additions or omissions in the source sentences during training, thereby achieving data augmentation effects.

Compared to traditional sequence-to-sequence (Seq2Seq) models, this model adopts a more flexible design by assigning independent encoders and decoders to each language. This design allows us to freely combine translation modules based on the source and target languages, enabling translation tasks in all directions. Furthermore, since each language has its own independent parameters within the TransNMT model, the model does not encounter capacity bottlenecks when supporting more languages. At the same time, adding support for new languages merely requires adding the corresponding language modules and training them, without affecting existing language modules.

To address the potential decline in zero-shot translation capability caused by independent parameters for each language, the TransNMT model integrates the M2TAB module based on the Transformer attention mechanism within its translation modules. This module serves as a bridge, connecting the encoders and decoders of different languages through attention mechanisms, enabling parameter sharing. This design facilitates the formation of language-agnostic interlingual representations within the TransNMT model, thereby enhancing its zero-shot translation capability and enabling it to handle translation tasks for low-resource and minority languages.

3.2 Semantic Shared Space

Under the framework of Multilingual Neural Machine Translation (NMT), accurately modeling the complex semantic relationships between vocabulary items across different languages represents a central challenge in enhancing system performance. In conventional approaches, a comprehensive vocabulary is constructed by multilingual NMT models, integrating lexical items from all source languages. However, this integration strategy does not naturally facilitate the sharing of language-specific vocabularies within a unified embedding space, particularly in scenarios where data-rich and data-scarce languages coexist. For languages with abundant data, learning their word embeddings is relatively straightforward and effective; conversely, languages with extreme data scarcity struggle to form high-quality lexical representations due to insufficient training samples, posing a notable performance bottleneck.

This paper aims to adopt a semantic space-sharing strategy, leveraging the corpora of resource-rich languages to learn representations for vocabulary items in resource-poor languages. By mapping low-resource vocabulary into high-resource corpora, a shared word embedding space is achieved, thereby transferring knowledge from high-resource languages to low-resource ones.

To accomplish this goal, we first preprocess the monolingual corpora of each low-resource language using the word2vec technique, extracting initial monolingual embedding sets for each language (i.e., constructing independent monolingual semantic spaces). Subsequently, we introduce automated word-by-word alignment technology, which aligns words between bilingual sentences based on similarity measures. Through this step, the word embeddings of low-resource languages are effectively transformed and mapped into a shared semantic framework grounded in high-resource languages, enabling cross-lingual knowledge sharing and transfer. This process not only promotes the effective migration of knowledge from high-resource languages to low-resource environments but also enhances the model's ability to comprehend and process low-resource vocabulary.

Figure 2 shows examples of aligning the English semantic space, which is taken as the high-resource corpus semantic space, with the German semantic space and Romanian semantic space respectively.

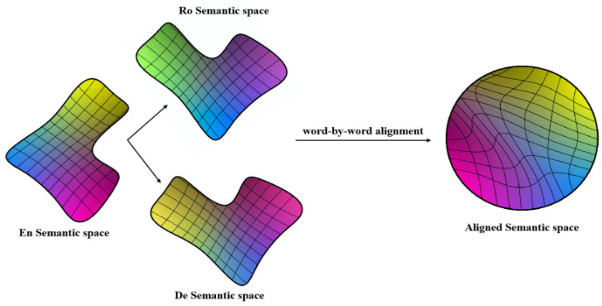

Figure 2. Schematic diagram of semantic space alignment

3.3 Translation Module of TransNMT

The TransNMT model's translation module assigns independent encoders and decoders to each language, allowing for flexible combinations based on the source and target language pair, enabling multi-directional translation tasks. Each encoder consists of stacked feedforward and self-attention sub-layers, while each decoder features stacked feedforward, multi-head attention, and self-attention sub-layers. Furthermore, TransNMT incorporates a distinct neural network module, M2TAB, which mirrors the Transformer architecture and acts as a bridge between language encoders and decoders via the attention mechanism.

Within the translation module, the encoder and decoder for the i-th language are denoted as Enc_i and Dec_i, respectively. Given a pair of sentences (x_i, y_j), representing a translation from source language i to target language j, where i,j∈{1,..., K}, and K is the total number of supported languages. The TransNMT model is trained by maximizing the likelihood estimation on the training set $D_{i,j}$ for all available language pairs in the set S. The objective of maximizing the likelihood estimation, denoted as L, is formally defined as follows:

$$L(\theta)=\sum_{\substack{(x_i,y_j)\in D_{i,j},\\ (i,j)\in S}} \log p\left(y_j|x_i;\theta\right) \quad \textbf{(1)}$$

In which, the probability $p(y_j| x_i)$ is modeled as:

$$p\left(y_j|x_i\right)=Dec_j\left(\text{M2TAB}\left(Enc_i\left(x_i\right)\right)\right) \quad \textbf{(2)}$$

In the formula, M2TAB(·) represents the M2TAB module proposed in this paper.

In the TransNMT model integrated with M2TAB, there is no direct connection between the encoder and the decoder; instead, they individually compute attention scores with M2TAB. As shown in Figure 1, for the encoders of each language, M2TAB acts as a decoder, where each position of M2TAB computes attention scores with all positions of the encoder's output sequence. Similarly, for the decoders of each language, M2TAB takes on the role of an encoder, with each position of the decoder computing attention scores with all positions of M2TAB's output sequence. The M2TAB module is formally defined as follows:

$$H^l_{M2TAB}=FFN\left(\text{MoE}\left(Q,K,V\right)\right) \quad \textbf{(3)}$$

$$Q=H^{l-1}_{M2TAB}\in\mathbb{R}^{d\times r} \quad \textbf{(4)}$$

$$K,V=H_{Enc_i}\in\mathbb{R}^{d\times n} \quad \textbf{(5)}$$

In the formula, H^l_{M2TAB} represents the hidden state of the l-th layer ($l \in [1, L]$) of M2TAB. H^L_{M2TAB}, being the top-level output hidden state, is used to calculate the attention scores between M2TAB and the decoder. H^0_{M2TAB} is the input representation of M2TAB, and this work adopts position encoding similar to that in Transformer, with options for both learnable and fixed position encodings. $Attn(\cdot)$ denotes the multi-head attention module, and MoE(·) represents the position-wise MoE sublayer. The query matrix Q comes from the output of the preceding stacked layers of M2TAB, while the key matrix K and value matrix V are derived from the output sequence representation H_{enc_i} of the i-th language encoder. d is the hidden layer size, n is the length of the encoder's input sequence, and r represents the sequence length of M2TAB.

3.4 M2TAB Module

As shown in Figure 3, the M2TAB module is composed of stacked multi-head attention sublayers and MoE sublayers. It serves as a bridge connecting various language encoders and decoders through the attention mechanism, enabling parameter sharing and assisting the TransNMT model in forming a language-agnostic interlingual representation.

Given an input token x, $E_i(x)$ represents the output of the i-th expert network. Then, the output y of the Mixture of Experts (MoE) module can be expressed as:

$$y = \sum_{i=1}^{n} softmax\left(x * W_g\right)_i E_i\left(x\right) \tag{6}$$

Where $W_g \in \mathbb{R}^{d \times n}$ is a learnable weight matrix, and the softmax function is used to distribute the weights among the expert networks when processing the input tokens. M2TAB equipped with MoE can be dynamically extended during the process of incremental learning by increasing the number of expert networks.

As shown in Figure 3, during the incremental training phase, an additional expert network $E_{n+1}(x)$ is added to the MoE layer for the newly introduced language, and the dimensionality of the gating network is increased by one to accommodate the expanded MoE. The output y of the expanded MoE module is updated as follows:

$$y = \sum_{i=1}^{n+1} softmax\left(x * W'_g\right)_i E_i\left(x\right) \tag{7}$$

Where $W'_g \in \mathbb{R}^{d \times (n+1)}$ is the weight matrix after expansion.

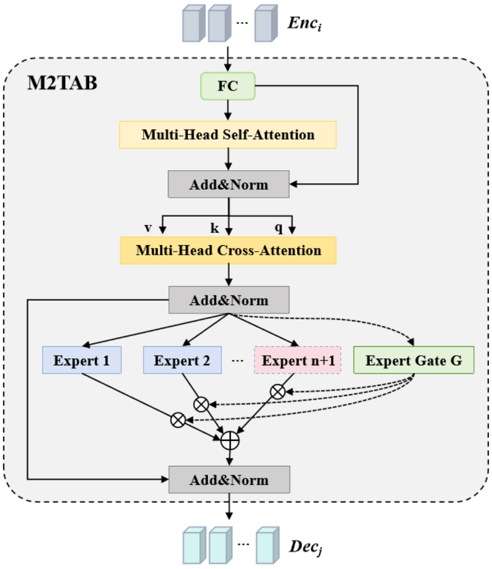

Figure 3. Schematic diagram of M2TAB module

3.5 Self-Back-Translation

In this paper, two TransNMT models are defined for forward and backward translation, where the output from the forward translation serves as input for the backward translation. The optimized backward translation result is then used as input for the source language.

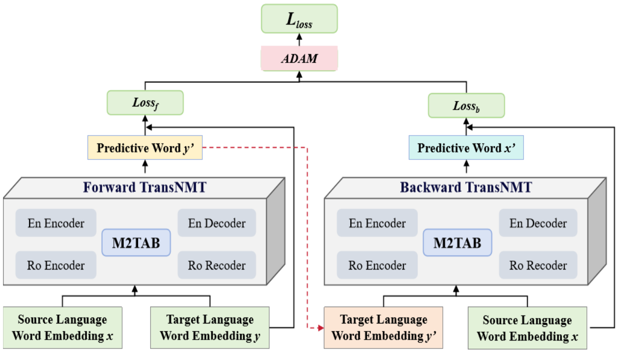

Figure 4. Schematic diagram of the self-back-translation mechanism of TransNMT

By jointly optimizing the parameters of both TransNMT models, the system identifies and corrects deficiencies in the forward translation results, thereby enriching the training data for forward translation. This automatic back-translation approach eliminates the need for manually designed data augmentation strategies, internally optimizing the process to reduce external noise and allowing the model to perform effectively in low-resource translation tasks. Figure 4 illustrates how TransNMT integrates with back-translation.

The loss function for the forward translation TransNMT is defined as $Loss_f$, and the loss function for the backward translation TransNMT is defined as $Loss_b$. Both utilize the cross-entropy loss function. During the training process, the weighted sum of these two losses is optimized using *ADAM*, a stochastic gradient descent method for optimizing stochastic objective functions based on first-order gradients. The total loss L_{loss} is calculated as follows:

$$L_{loss} = ADAM\left(Loss_f + \lambda Loss_b\right) \tag{8}$$

4 Experiment

4.1 Dataset

The experimental data in this paper is a hybrid combination of parallel corpora created from the Europarl Parallel Corpus and the TED Talks corpus. It comprises parallel corpora from four low-resource languages (LRL), namely Romanian (Ro), Azerbaijani (Aze), Belarusian (Bel), and Galician (Glg), to English (En), as well as the semantic spaces of three high-resource languages (HRL), German (De), Finnish (Fi), and French (Fr), which are jointly trained to assist in verifying the effectiveness of low-resource translation. The statistical information of the parallel corpora is presented in Table 1.

Table 1. Parallel corpus statistics

LRL	Train	HRL	Train
Ro	6.0k	De	182k
Aze	5.97k	Fi	103k
Bel	4.51k	Fr	185k
Glg	10.0k	-	-

4.2 Ablation Study

To verify the effectiveness of introducing the M2TAB module and the self-back-translation mechanism, as well as their impact on model performance, ablation experiments were conducted. The experimental results are presented in Table 2.

Table 2. Results of ablation experiments

M2TAB	SBT	Ro	Aze	Bel	Glg
–	–	15.7	8.9	12.8	25.7
√	–	21.8	11.6	16.9	27.6
–	√	22.3	12.6	15.2	28.5
√	√	**30.5**	**15.2**	**20.7**	**33.1**

(**Note:** "√" indicates add, "–" indicates no add)

4.3 Comparative Experiment

The TransNMT model was compared with various classical neural machine translation models, including Multi-NMT RNN, Multi-NMT TR, Word-SDE, and Multi-NMT TR_SBT. This comparative analysis aimed to evaluate the performance and effectiveness of the TransNMT model in multi-language translation tasks. The detailed experimental results of this comparison are summarized in Table 3, highlighting the advancements achieved by the TransNMT approach.

Table 3. The results of various models in four languages under low-resource conditions

Model	BLEU			
	Ro	Aze	Bel	Glg
Multi-NMT RNN	11.4	7.8	10.4	24.9
Multi-NMT TR	16.5	9.2	12.5	26.7
Word-SDE	27.6	11.3	17.7	29.5
Multi-NMT TR_ SBT	29.4	13.5	19.4	31.8
TransNMT(Ours)	**30.5**	**15.2**	**20.7**	**33.1**

In order to further validate the effectiveness of the semantic shared space approach, this paper expands from the initial English semantic space as the base semantic space to an English-centered semantic space, with the joint space of German (De), Finnish (Fi), and French (Fr) serving as auxiliary semantic spaces. The TransNMT model is trained on these four low-resource languages. The experimental results are shown in Table 4. The results show that the addition of auxiliary semantic spaces has improved the experimental results to a certain extent, demonstrating the effectiveness of the semantic shared space approach.

Table 4. Results of the impact of auxiliary languages on the translation quality of the model

De	Fi	Fr	Low resource language			
			Ro	Aze	Bel	Glg
–	–	–	30.5	15.2	20.7	33.1
√	–	–	31.2	15.9	21.2	33.6
√	√	–	32.1	16.8	22.1	34.5
√	√	√	**33.2**	**17.7**	**23.2**	**35.3**

(**Note:** "√" indicates add, "–" indicates no add)

5 Conclusion

This study introduces a natural language generation approach termed TransNMT, which is grounded in multilingual neural machine translation (NMT) and leverages a Transformer-based self-back-translation mechanism. This approach addresses the challenge of learning lexical representations in low-resource settings. In the translation process, low-resource lexical items are represented by their high-resource counterparts, effectively utilizing high-resource corpora to improve the translation quality of low-resource content. Additionally, by embedding the self-back-translation structure within TransNMT, the issue of data scarcity in low-resource translation is further mitigated.

To address potential declines in zero-shot translation performance caused by language-specific parameters, the TransNMT model incorporates the M2TAB module, which is based on the Transformer attention mechanism. This module serves as a bridge between the encoders and decoders of different languages by enabling parameter sharing through the attention mechanism. This design allows the TransNMT model to develop language-agnostic interlingual representations, thereby improving zero-shot translation capabilities and facilitating translation tasks for minority and low-resource languages. Experimental results indicate that, compared to other baseline models, the proposed approach demonstrates superior performance in low-resource translation scenarios.

References

[1] N. Kalchbrenner, P. Blunsom, Recurrent Continuous Translation Models, *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, Seattle, Washington, USA, 2013, pp. 1700-1709. https://aclanthology.org/D13-1176/

[2] I. Sutskever, O. Vinyals, Q. V. Le, Sequence to sequence learning with neural networks, *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, Cambridge, MA, USA, 2014, pp. 3104-3112.

[3] H. Sak, A. Senior, F. Beaufays, Long Short-Term Memory Based Recurrent Neural Network Architectures for Large Vocabulary Speech Recognition, *arXiv preprint*, arXiv:1402.1128, February, 2014.
https://arxiv.org/abs/1402.1128

[4] K. Cho, B. V. Merrienboer, C. Gülçehre, D. Bahdanau, F. Bougares, H. Schwenk, Y. Bengio, Learning phrase representations using RNN encoder-decoder for statistical machine translation, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar, 2014, pp. 1724-1734.
https://doi.org/10.3115/v1/D14-1179

[5] D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, *3rd International Conference on Learning Representations*, San Diego, CA, US, 2015.

[6] F. Meng, Z. Lu, M. Wang, H. Li, W. Jiang, Q. Liu, Encoding source language with convolutional neural network for machine translation, *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, Beijing, China, 2015, pp. 20-30.
https://doi.org/10.3115/v1/P15-1003

[7] J. Gehring, M. Auli, D. Grangie, Y. Dauphin, A Convolutional Encoder Model for Neural Machine Translation, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, Vancouver, Canada, 2017, pp. 123-135.
https://doi.org/10.18653/v1/P17-1012

[8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, *Proceedings of the 31st International Conference on Neural Information Processing Systems*, Long Beach, California, USA, 2017, pp. 6000-6010.

[9] T. L. Ha, J. Niehues, A. Waibel, Toward Multilingual Neural Machine Translation with Universal Encoder and Decoder, *Proceedings of the 13th International Conference on Spoken Language Translation*, Seattle, Washington D.C, 2016.
https://aclanthology.org/2016.iwslt-1.6/

[10] B. Zoph, K. Knight, Multi-Source Neural Translation, *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, San Diego, California, 2016, pp. 30-34.
https://doi.org/10.18653/v1/N16-1004

[11] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, R. R. Salakhutdinov, Improving neural networks by preventing co-adaptation of feature detectors, *arXiv preprint*, arXiv:1207.0580, July, 2012.
https://arxiv.org/abs/1207.0580

[12] R. Sennrich, B. Haddow, A. Birch, Improving Neural Machine Translation Models with Monolingual Data, *54th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics (ACL)*, Berlin, Germany, 2016, pp. 86-96.
https://doi.org/10.18653/v1/P16-1009

[13] Y. Li, X. Li, Y. Yang, R. Dong, A diverse data augmentation strategy for low-resource neural machine translation, *Information*, Vol. 11, No. 5, Article No. 255, May, 2020.
https://doi.org/10.3390/info11050255

[14] M. Artetxe, G. Labaka, N. Casas, E. Agirre, Do all roads lead to Rome? Understanding the role of initialization in iterative back-translation, *Knowledge-Based Systems*, Vol. 206, Article No. 106401, October, 2020.
https://doi.org/10.1016/j.knosys.2020.106401

[15] A. Poncelas, D. Shterionov, A. Way, G. M. D. B. Wenniger, P. Passban, Investigating Backtranslation in Neural Machine Translation, *Proceedings of the 21st Annual Conference of the European Association for Machine Translation*, Alicante, Spain, 2018, pp. 249-258.
https://aclanthology.org/2018.eamt-main.25/

[16] A. Karakanta, J. Dehdari, J. Genabith, Neural machine translation for low-resource languages without parallel corpora, *Machine Translation*, Vol. 32, No. 1-2, pp. 167-189, June, 2018.
https://doi.org/10.1007/s10590-017-9203-5

[17] J. Gu, Y. Wang, Y. Chen, V. O. K. Li, K. Cho, Meta-learning for low-resource neural machine translation, *2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium, 2018, pp. 3622-3631.
https://doi.org/10.18653/v1/D18-1398

[18] J. Gu, H. Hassan, J. Devlin, V. O. K. Li, Universal Neural Machine Translation for Extremely Low Resource Languages, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, New Orleans, Louisiana, 2018, pp. 344-354.
https://doi.org/10.18653/v1/N18-1032

[19] T. Kocmi, O. Bojar, Trivial Transfer Learning for Low-Resource Neural Machine Translation, *Proceedings of the Third Conference on Machine Translation: Research Papers*, Brussels, Belgium, 2018, pp. 244-252.
https://doi.org/10.18653/v1/W18-6325

[20] X. Kong, A. Renduchintala, J. Cross, Y. Tang, J. Gu, X. Li, Multilingual Neural Machine Translation with Deep Encoder and Multiple Shallow Decoders, *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, Online, 2021, pp. 1613-1624.
https://doi.org/10.18653/v1/2021.eacl-main.138

[21] M. Xia, E. Monti, Multilingual Neural Semantic Parsing for Low-Resourced Languages, *Proceedings of* SEM 2021: The Tenth Joint Conference on Lexical and Computational Semantics*, Online, 2021, pp. 185-194.
https://doi.org/10.18653/v1/2021.starsem-1.17

[22] L. Qin, M. Ni, Y. Zhang, W. Che, CoSDA-ML: multi-lingual code-switching data augmentation for zero-shot cross-lingual NLP, *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, Yokohama, Japan, 2021, pp. 3853-3860.

[23] S. M. Singh, T. D. Singh, An empirical study of low-resource neural machine translation of manipuri in multilingual settings, *Neural Computing and Applications*, Vol. 34, No. 17, pp. 14823-14844, September, 2022.
https://doi.org/10.1007/s00521-022-07337-8

Biographies

Yewei Zhang, associate professor, master supervisor, holds a master's degree in English education from Flinders University in Australia. Currently, she serves as the deputy director of the International Cooperation and Exchange Office of the Vocational and Technical College of Inner Mongolia Agricultural University, in charge of international exchange and cooperation. Her research interests mainly include but are not limited to natural language generation, neural machine translation, and multimodal learning.

Yatao Mu, graduated from the College of Foreign Languages of Inner Mongolia Agricultural University. Currently, he serves as the director of the Party and Government Office. His research interests mainly include but are not limited to natural language processing, multimodal learning, and machine learning.

Jiawei Zhang, graduated from the English translation major of Inner Mongolia University. Her research interests mainly include, but are not limited to natural language processing, multimodal learning, and pattern recognition.