Constraint-Augmented Mongolian-Chinese Neural Machine Translation Based on Dynamic Feedback Alignment (Student Abstract)

Shuting Dai[1], Yatu Ji[1]*, Yanli Wang[1], Lei Shi[2], Qing-Dao-Er-Ji Ren[1], Nier Wu[1], Na Liu[1]

[1]Inner Mongolia University of Technology, China
[2]Inner Mongolia University of Finance and Economics, China
1243003558@qq.com, {MLjyt, 20221800125, renqingln, wunier04, csnaliu}@imut.edu.cn, shilei@imufe.edu.cn

Abstract

The scarcity of parallel corpora for Mongolian and Chinese constrains the performance of Mongolian-Chinese neural machine translation (NMT), particularly manifesting in inadequate accuracy in translating specialized terminology. To address this limitation, this study adopts a lexically constrained augmentation strategy that constructs pseudo-source sentences by appending Chinese constraint words to Mongolian source texts, while enforcing the inclusion of these constraints in the output to improve translation accuracy. However, this approach presents two inherent drawbacks: processing pseudo-sentences with a single encoder tends to induce semantic interference, while the introduced constraint words may exacerbate alignment errors during decoding. To overcome these limitations, this paper propose a **C**onstraint-**A**ugmented Mongolian-Chinese **NMT** method (CANMT) based on dynamic feedback alignment. The method employs a dual-encoder architecture to isolate bilingual representations, coupled with a dynamic feedback alignment module that progressively reduces alignment errors through iterative refinement, thereby enhancing overall translation performance.

Introduction

Mongolian-Chinese NMT faces challenges due to the scarcity of parallel corpora, which impedes the model's capacity to fully acquire bilingual linguistic patterns. This limitation not only restricts overall translation performance but also leads to inadequate accuracy in translating domain-specific terminology. To alleviate this issue, inspired by the work of Jon (Jon et al. 2021), this paper constructs pseudo-source sentences by appending Chinese constraint words to Mongolian source sentences. This approach simultaneously achieves effective expansion of training data and enhances terminology translation accuracy by leveraging target-side lexical prior information provided by the constraint words.

However, this method exhibits significant drawbacks: on one hand, Mongolian and Chinese languages tend to cause feature interference in the shared encoding space, resulting in one language dominating the other; on the other hand, conventional attention mechanisms lack explicit alignment constraints (Lu et al. 2022), making it difficult for the model to accurately establish relationships between constraint words and corresponding source language segments, while potentially introducing erroneous alignment biases during decoding.

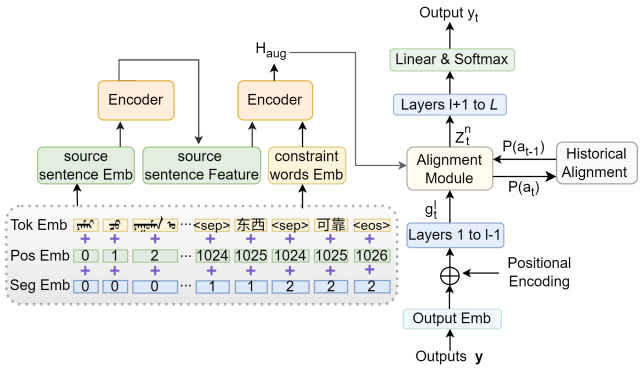

Figure 1: Architecture Diagram of Constraint-Augmented Model Based on Dynamic Feedback Alignment

To address these limitations, this paper proposes CANMT: a dual-encoder architecture is designed to separately process source sentences and constraint words, reducing cross-lingual feature interference; simultaneously, a dynamic feedback alignment module module is incorporated into the decoder to dynamically strengthen the correspondence between constraint words and source language during decoding, thereby improving both terminology translation accuracy and overall translation quality. The overall architecture of the model is shown in Figure 1.

Methodology

Dual-Encoder Constraint Integration. The words required to appear in the target translation are defined as constraint words, based on which a pseudo-source sentence is constructed: $\hat{X} = [x, \langle sep \rangle, c_1, \langle sep \rangle, ..., \langle sep \rangle, c_N, \langle eos \rangle]$. Here, x denotes the source sentence, c_i represents the constraint words, and $\langle sep \rangle$ is used to achieve structured separation between the source sentence and the constraint words. During the encoding phase, the model aggregates the word embeddings, position embeddings, and segment embeddings of each token in the pseudo-source sentence. The segment

*Corresponding author.

embeddings assist the model in distinguishing whether a token originates from the source sentence or the constraint word sequence. Subsequently, the dual encoders process this information to generate a joint vector representation: $H_{aug} = ENCODE\left(E'_{aug}\right)_{joint}$, where E'_{aug} is the concatenated vector of the Mongolian source sentence and the constraint words.H_{aug} provides sufficient contextual information for the decoder, ultimately guiding the model to generate translations that include the constraint words while maintaining semantic coherence and grammatical correctness.

Dynamic Feedback Alignment Module. To address the issue of insufficient constraint word representation in traditional constrained translation, this paper introduces a dynamic feedback alignment module to optimize alignment performance. First, based on Alignment Error Rate (AER) score evaluation, alignment information is extracted using a bidirectional Mongolian-Chinese model. The module insertion layer is determined via the formula: $l_{b,x\to y}, l_{b,y\to x} = \arg\min_{i,j} AER\left(A^i_{x\to y}, A^j_{y\to x}\right)$, where $l_{b,x\to y}$ denotes the layer used for module construction, $l_{b,y\to x}$ represents the gold alignment reference layer, $A^i_{x\to y}$ is the alignment result from the i-th layer of the Mongolian-Chinese model, and $A^j_{y\to x}$ is the alignment result from the j-th layer of the Chinese-Mongolian model.

During decoding, a query vector is constructed as: $q^n_t = [g^l_t, e(y_t), P(a_{t-1})]W^n_Q$, where g^l_t is the decoder state at layer l, $e(y_t)$ is the embedding of the current target word, and $P(a_{t-1})$ is the historical alignment distribution. This query vector is combined with the key matrix $K^n = HW^n_K$ generated by the dual encoders. Through multi-head attention, the alignment probability distribution is computed as: $P(a_t|x, y_{\leq t}) = \frac{1}{\eta}\sum_{n=1}^{\eta} softmax\left(\frac{q^n_t (K^n)^T}{\sqrt{d}}\right)$. The value vectors $V^n = HW^n_V$ are then weighted and fused using this distribution to generate the contextual representation: $Z^n_t = P(a^n_t|x, y_{\leq t})V^n$, which guides the translation generation process. This module enhances the representation of constraint words, thereby improving both translation quality and word alignment precision.

Experiments

Dataset. The dataset employed in this study consists of 1.2 million copyrighted Mongolian-Chinese parallel sentence pairs, encompassing daily dialogues, government documents, and official evaluation data from CCMT. Among these, one million sentence pairs were augmented with additional constraint markers during training, while the remaining 200,000 pairs were retained in their original form to ensure the model's generalization capability in unconstrained scenarios.

Results. This study utilizes two key evaluation metrics: AER, which quantifies word alignment accuracy, and BLEU, which assesses overall translation quality. As shown in Table 1, all reported results are averaged over multiple runs. Our proposed method demonstrates superior performance compared to other constraint-augmented baseline models. Specifically, the CANMT model achieves a remarkable 23.22% reduction in AER over the Transformer baseline, exhibiting the best lexical alignment accuracy among all compared systems. In terms of BLEU score, CANMT also shows a significant advantage, outperforming the Transformer baseline by 1.29 points and surpassing all other constraint-based baselines.

Model	BLEU	AER
Transformer	29.70	60.74%
Code-Switch	30.28	—
Factor Encoder	30.62	39.85%
Add-Align-Head	—	39.53%
EAM	—	38.92%
Post-Align	—	38.41%
CANMT (our)	30.99	37.52%

Table 1: Comparison Table of BLEU Scores and AER Between CANMT and Other Constraint-Augmented Models.

Conclusion

The proposed CANMT method effectively mitigates cross-lingual semantic interference through its dual-encoder architecture that separately processes source sentences and constraint words. Simultaneously, the incorporated dynamic feedback alignment module significantly reduces lexical alignment errors. Experimental results demonstrate that CANMT achieves a 1.29 BLEU point improvement over the Transformer baseline, along with a substantial 23.22% reduction in AER. These findings conclusively validate the effectiveness of our approach for constraint-augmented Mongolian-Chinese NMT.

Acknowledgments

This study was supported by the National Natural Science Foundation of China (Grant Nos. 62206138 and 62466044); the Inner Mongolia Natural Science Foundation (Grant Nos. 2024MS06009, 2024MS06017, and 2024QN06021); the Research Program of Science and Technology at Universities of Inner Mongolia Autonomous Region (Grant No. NJZZ23081); the Basic Research Expenses of Inner Mongolia (Grant No. ZTY2025072); and the Science and Technology Plan Project of Inner Mongolia Autonomous Region (Grant Nos. 2025YFHH0083 and 2025YFHH0115).

References

Jon, J.; Aires, J. P.; Varis, D.; and Bojar, O. 2021. End-to-End Lexically Constrained Machine Translation for Morphologically Rich Languages. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, 4019–4033.

Lu, Y.; Zhang, J.; Zeng, J.; Wu, S.; and Zong, C. 2022. Attention Analysis and Calibration for Transformer in Natural Language Generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30: 1927–1938.