Neural Machine Translation for Agglutinative Languages via Data Rejuvenation

Chen Zhao[1], Yatu Ji[1], Qing-Dao-Er-Ji Ren[1], Nier Wu[1], Lei Shi[2]
Fu Liu[1], YePai Jia[1]
[1]Inner Mongolia University of Technology, China
[2]Inner Mongolia Finance and Economics University, China
{20231800117,mljyt,renqingln,wunier04,20241800133,20231800142}@imut.edu.cn
{shilei}@imufe.edu.cn

Abstract

In Recent years, advances in Neural Machine Translation (NMT) heavily rely on large-scale parallel corpora. Within the context of China's Belt and Road Initiative, there is increasing demand for improving translation quality from agglutinative languages (e.g., Mongolian, Arabic) to Chinese. However, the translation scenarios for agglutinative languages (which form words by concatenating morphemes with clear boundaries) face significant challenges including data sparsity, quality imbalance, and inactive sample proliferation due to their morphological complexity and syntactic flexibility. This study presents a systematic analysis of data distribution characteristics in agglutinative languages and proposes a dual-module framework combining fine-grained inactive sample identification with target-side rejuvenation. Our framework first establishes a multi-dimensional evaluation system to accurately identify samples exhibiting low-frequency morphological interference or long-range word order mismatches. Subsequently, the target-side rejuvenation mechanism generates diversified noise-resistant translations through iterative optimization of sample contribution weights. Experimental results on four low-resource agglutinative language tasks demonstrate significant performance improvements (BLEU +2.1–3.4) across mainstream NMT architectures. Architecture-agnostic validation further confirms the framework's generalizability.

1 Introduction

Neural Machine Translation (NMT) depends heavily on large-scale training data (Koehn and Knowles, 2017), yet issues like data noise and complex patterns hinder effective training. Though methods such as curriculum learning (Edunov et al., 2020), data diversification (Nguyen et al., 2020), and denoising (Wang et al., 2018) improve data quality, they fail to tackle *inactive samples*—instances that contribute little or negatively to model performance. These samples, often affected by morphological complexity or word-order mismatches, are especially problematic in agglutinative-to-Chinese translation tasks (Yatu et al., 2024; Ji et al., 2019). The structural gap between SOV agglutinative languages and SVO Chinese limits sentence-level confidence metrics (Kumar and Sarawagi, 2019) in detecting unstable translations.

To address this challenge, we propose a data rejuvenation framework for agglutinative language translation, specifically handling: (1) low-frequency morpheme interference (e.g., Mongolian suffix -) through multi-dimensional metrics, and (2) SOV-to-SVO mismatches (e.g., Uyghur object-fronting) via target-side augmentation.

Specifically, we train a target-side data augmentation model on active samples as the regenerator to relabel inactive samples, thereby obtaining regenerated samples. First, multi-dimensional metrics (e.g., sentence probability mean, standard deviation, and token-level extremal probabilities) are designed to identify inactive samples with low-frequency morphology or word-order mismatches. Second, a target-side augmentation mechanism based on latent space modeling generates diverse translations to mitigate data sparsity and word-order distortion. Finally, active and regenerated samples are jointly trained (Guo et al., 2024). Experiments on Mongolian–Chinese, Uyghur–Chinese, and Arabic–Chinese tasks show consistent improvements across LSTM (Domhan, 2018), Transformer (Vaswani et al., 2017), and DynamicConv (Wu et al., 2019; Gehring et al., 2017) architectures.

2 Related Work

Inactive Samples. Inactive samples refer to training instances with minimal or negative contributions to model performance, primarily due to in-

effective feature encoding. This phenomenon is observed in both computer vision (e.g., 10% redundancy in CIFAR-10/ImageNet (Krizhevsky et al., 2009; Deng et al., 2009)) and NMT (Jiao et al., 2020). However, agglutinative languages (Mongolian, Arabic) pose unique challenges in Chinese translation: rich morphology (complex affixation) and free word order (SOV structure) induce distinctive inactive patterns like low-frequency morphological interference and long-range syntax mismatches. Traditional single-metric approaches (e.g., sentence-level probability) fail to capture these fine-grained features (Pan et al., 2020), motivating our multi-dimensional evaluation system integrating sentence probability statistics (mean/std) and token-level confidence extremes.

Data Manipulation. Existing methods fall into two categories: 1) Data purification/augmentation (Gao et al., 2024) including denoising (Wang et al., 2018) and forward translation (Nguyen et al., 2020; Jin, 2024; Li et al., 2022); 2) Sample weighting via self-paced learning (easy samples), hard example mining, or curriculum learning. While effective for general NMT, these approaches inadequately address agglutinative-specific issues. For instance, Jiao et al.'s (Jiao et al., 2020) forward translation method introduces word order errors during SOV-to-SVO conversion (Luo et al., 2024), amplifying translation noise. Our innovation lies in target-side data augmentation through latent space posterior distribution modeling, generating multiple noise-resistant translation variants to mitigate single-annotation dependency.

Low-Resource Utilization. Recent advances leverage knowledge distillation and corpus refinement: Ding et al. (Ding et al., 2021, 2022) propose bidirectional distillation to enhance low-frequency word alignment, while Briakou et al. (Briakou and Carpuat, 2022) employ semantic equivalence classifiers for noise filtering. These methods synergistically complement our sample activation framework—bidirectional distillation expands lexical coverage, corpus refinement ensures data purity, and our multi-metric evaluation optimizes sample utility weights—collectively enhancing NMT robustness for agglutinative languages.

3 Methodology

This chapter presents the architecture of the data rejuvenation framework for agglutinative languages (Figure 1). The Identification Module implementing multi-metric evaluation (sentence-level probability, standard deviation, min/max token probabilities) to detect inactive samples through fine-grained analysis of translation behaviors under complex morphological and syntactic structures; 2) Activation Module employing target-side data augmentation to generate diverse translations, thereby enhancing low-contribution samples' utility. The regenerated samples are combined with original active data to train the final NMT model.

3.1 Identification Model

Current NMT approaches predominantly rely on single metrics (e.g., sentence-level probability) to evaluate sample activity. However, this paradigm exhibits critical limitations in low-resource language pairs with significant grammatical divergence like agglutinative-to-Chinese translation. Firstly, sentence-level metrics fail to account for: (1) low-frequency token impacts (e.g., their probabilities are masked by high-frequency counterparts), (2) long-range dependencies, (3) complex syntactic structures—all crucial for capturing grammatical relationships and semantic coherence (Mohamed and Al-Azani, 2025; Shaalan et al., 2019; Refai et al., 2023). Additionally, the coarse-grained nature of sentence-level metrics lacks token-wise translation quality assessment, impairing both model training efficacy and inactive sample identification.

To address these deficiencies, we propose a multi-metric evaluation framework that comprehensively analyzes training samples through four dimensions:

Sentence-level probability (p_{sent_mean}): The trained Neural Machine Translation (NMT) model evaluates the generation relationship between source and target sentences by computing the sentence-level probability $p(y|x)$, which represents the confidence of generating target sentence y given source sentence x. Specifically, this probability is derived by calculating the conditional probability $p(y_t|x, y_{<t})$ at each time step, where T is the length of the target sentence, y_t denotes the t-th word in the target sentence, x is the source sentence, and $y_{<t}$ represents the first $t-1$ target words. This computation indicates that the model progressively assesses the conditional probability of each word during target sentence generation, ultimately determining the overall sentence probability. A low sentence-level probability for a training sample suggests poor translation quality, weak alignment with

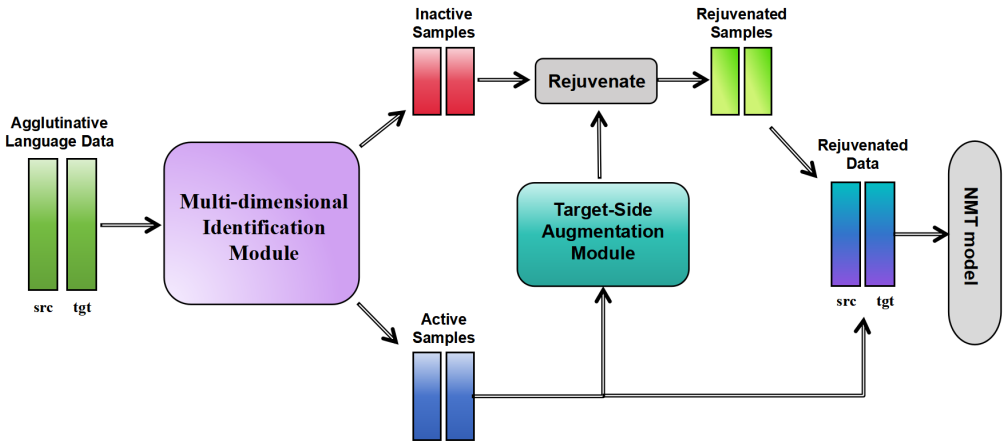

Figure 1: The framework of data rejuvenation. The inactive samples are identified from the original training data, reconstructed through a rejuvenation model, and then combined with active samples for NMT model training.

the source sentence, and low model confidence, thereby contributing minimally to model performance.

$$P_{sent\_mean} = \frac{1}{T}\sum_{t=1}^{T} p(y_t|x, y_{<t}) \quad (1)$$

Sentence Probability Standard Deviation (p_{sent_std}): The trained NMT model computes the standard deviation P_{sent_std} of sentence probabilities, where P_{sent_mean} is the mean of sequence conditional probabilities and T is the sequence length. By calculating the square root of the mean squared deviation between each time step's conditional probability $p(y_t|x, y_{<t})$ and the mean P_{sent_mean}, we obtain P_{sent_std}, which measures the fluctuation degree of generation probabilities. A high P_{sent_std} indicates significant confidence volatility during target sentence generation, suggesting inconsistent translation quality. Consequently, such samples are less effective for model improvement and may be classified as low-contribution examples.

$$P_{sent\_std} = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(p(y_t|x, y_{<t}) - P_{sent\_mean})^2} \quad (2)$$

Minimum Token Probability (P_{tok_min}): Represents the lowest token-level confidence in generating target sentence y from source sentence x. Intuitively, a low P_{tok_min} indicates that certain tokens in the example are unlikely during generation, potentially providing insufficient information to enhance translation performance. Here, $p(y_t|x, y_{<t})$ denotes the probability of generating the t-th token in the target sentence given the source sentence x:

$$P_{tok\_min} = \min_t p(y_t|x, y_{<t}) \quad (3)$$

Maximum Token Probability (P_{tok_max}): Represents the highest confidence level for a single token during target sentence generation. A high P_{tok_max} indicates strong model confidence in generating a specific token:

$$P_{tok\_max} = \max_t p(y_t|x, y_{<t}) \quad (4)$$

Composite score: The composite score for each sample is computed through a weighted combination of four metrics:

$$\text{CompositeScore} = \alpha \cdot P_{sent\_mean} + \beta \cdot \frac{1}{P_{sent\_std}+\epsilon} + \gamma \cdot P_{tok\_min} + \delta \cdot \log P_{tok\_max}$$

where α, β, γ, and δ are weighting coefficients optimized via grid search (empirically set to 0.4, 0.3, 0.2, and 0.1 respectively), with $\epsilon = 1 \times 10^{-5}$ preventing division by zero. The inverse relationship with P_{sent_std} explicitly penalizes high-variance samples.

Samples are then ranked by their composite scores, and those below the threshold τ are identified as *inactive*. These typically exhibit:

- Significant probability fluctuations (high P_{sent_std})
- Extremely low token probabilities (P_{tok_min})
- Overconfident predictions (high P_{tok_max})

Such samples are prioritized for rejuvenation during optimization.

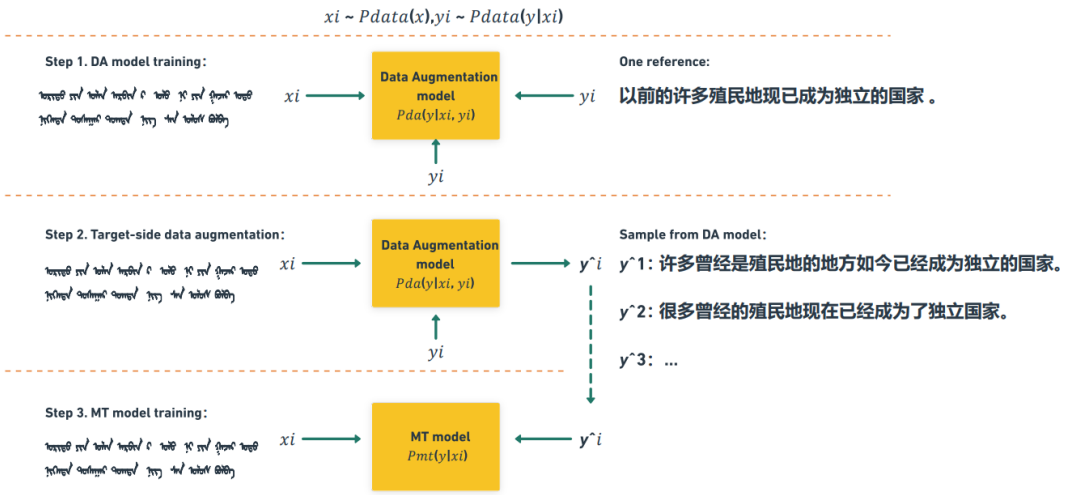

Figure 2: The illustration of Target-Side Data rejuvenation: The rejuvenation model estimates translation distributions and samples data, optimizing MT model training through an intermediate latent variable.

3.2 Rejuvenation Model

In current NMT tasks, traditional optimization methods primarily rely on forward and backward translation, which expands training data by generating new source or target translations. However, these approaches exhibit limitations in low-resource agglutinative language translation: 1) Forward translation heavily depends on source language word order and syntax, often causing semantic drift when processing free-word-order agglutinative languages, thereby reducing data effectiveness; 2) Backward translation increases target-side samples but lacks diversity, especially in capturing long-range dependencies, complex syntactic structures, and low-frequency vocabulary, failing to effectively model source-target alignment. Consequently, generated samples inadequately improve model learning on inactive samples. To address these issues, we employ target-side data augmentation for inactive sample rejuvenation. This method models the posterior distribution of target sentences to generate diverse potential translations, smoothing the training data distribution. Figure 2 illustrates an example of target-side data augmentation for Mongolians.

The core of target-side data augmentation lies in modeling the posterior distribution $P_{da}(y|x_i, y_j)$ of target sentences. Given source sentence x_i and target sentence y_i, we introduce latent variable z, decomposing the posterior as:

$$P_{da}(y|x_i, y_j) = \sum_{z \in Z_i} P_\phi(y|x_i, z) P_a(z|y_i) \quad (5)$$

The Z_i is the latent space; $P_\phi(y|x_i, z)$ represents the conditional translation distribution, modeling target sentence generation from x_i and z; $P_a(z|y_i)$ denotes the latent variable distribution given y_i, describing the likelihood of generating z from y_i.

After posterior modeling, the augmentation process samples latent variables to generate diverse target translations, enhancing data variety and model generalization. Specifically, for each x_i, we first sample $\{z_j\}$ from $P_a(z|y_i)$, where each z_j represents a semantic feature guiding diverse translation generation. Then, we generate potential translations y_j by maximizing the translation probability:

$$y_j = \arg\max_y P_\phi(y|x_i, z_j) \quad (6)$$

The final augmented set is:

$$\hat{y}_i = \left\{ \arg\max_y P_\phi(y|x_i, z_j) | z_j \sim P_a(z|y_i) \right\}_{j=1}^{M} \quad (7)$$

This set of potential translations not only exhibits formal diversity but also maintains semantic consistency guided by the posterior distribution. This augmentation process significantly expands the possible target translations for each source sentence, thereby enhancing both the diversity and quality of the data.

4 EXPERIMENT

4.1 Experimental Setup

The experimental data in this paper is sourced from in-house Mongolian-Chinese parallel corpora

and publicly available Arabic-Chinese and Korean-Chinese datasets. The Mongolian-Chinese corpus consists of 500K sentence pairs, covering dialogues, government documents, news texts, and CCMT data, with 400K pairs selected for training. Additionally, we utilize two public corpora—OpenSubtitles v2024 and MultiCCAligned v1.1—to construct Arabic and Korean datasets. OpenSubtitles v2024 contains movie and TV subtitles, focusing on colloquial and multi-domain coverage, while MultiCCAligned v1.1 is derived from automatically aligned multilingual web content, offering diverse domains and large-scale data. Approximately 300K sentence pairs from each dataset are used for Arabic-Chinese and Korean-Chinese training. For each language pair, 5K sentence pairs are reserved for validation and 5K for testing. All data undergoes tokenization and BPE processing, with results reported using BLEU.

We implement the proposed data rejuvenation framework on representative NMT architectures:

- **LSTM**: Integrated within the Transformer framework.
- **Transformer**: Pure attention-based architecture.
- **DynamicConv**: Lightweight dynamic convolutional architecture.

All models are implemented using Fairseq (Ott et al., 2019). Training configurations:

- LSTM: 300K steps with 32K tokens/batch (4096 $\times$ 8)
- Transformer: 300K (BASE)/1M (BIG) steps with 32K tokens/batch
- DynamicConv: 1M steps with 57K tokens/batch (3584 $\times$ 16)

Finally, this study conducts experimental investigations using DynamicConv on the identification module (§3.1) and activation module (§3.2), followed by reporting translation performance across diverse model architectures and language pairs.

4.2 Inactive Examples

This section validates the rationality and consistency of the identified inactive samples through a series of experiments.

4.2.1 Rationality of Multi-Dimensional Evaluation

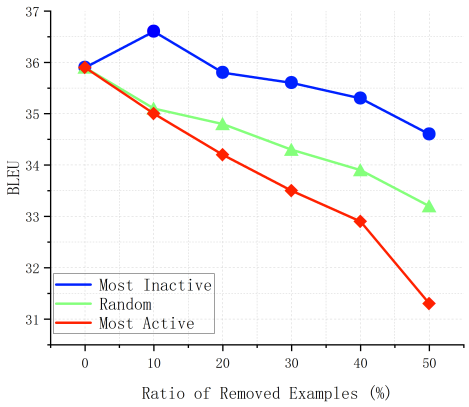

Figure 3: Translation Performance of NMT Models Trained on Data with Least Active Samples Removed: Results are compared with models trained on the most active samples and randomly sampled examples.

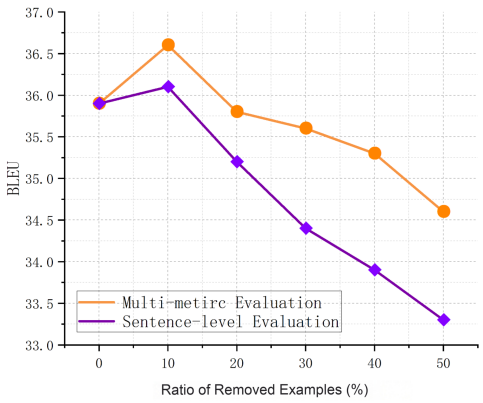

Figure 4: Comparison of the impact degree on translation performance between inactive samples identified using a multi-dimensional evaluation system and those identified solely by sentence-level probability.

This experiment validates the rationality of inactive sample identification by analyzing their impact on translation performance. Theoretically, removing inactive samples lacking effective information should not significantly affect model performance. Based on this hypothesis, we remove the lowest probability samples (most inactive) and evaluate NMT models trained on the remaining data. Figure 3 demonstrates the impact of removing the most inactive samples from the Mongolian-Chinese parallel corpus identified by our multi-dimensional evaluation system. Overall, translation performance gradually declines with an increased removal ratio. However, compared to random removal, inactive

sample removal shows milder performance degradation, while active sample removal causes the most significant deterioration. Notably, removing 10% of the most inactive samples slightly improves performance, aligning with findings in computer vision datasets.

Furthermore, we compare inactive samples identified by sentence-level probability methods and our multi-dimensional evaluation system. As shown in Figure 4, the multi-dimensional system demonstrates a smaller performance impact and slower decline rates under identical removal ratios, proving its superior rationality in inactive sample identification.

4.2.2 Validation of Inactive Sample Overlap Rate

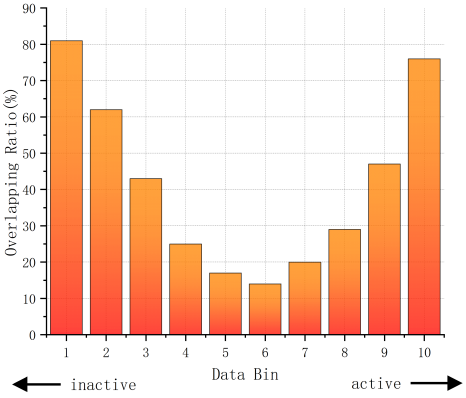

Figure 5: Overlap Ratio of Sample Activity Levels Identified by the Multi-Dimensional Evaluation System Across Model Variants

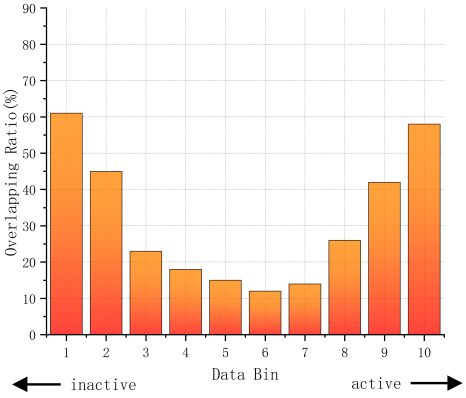

Figure 6: Overlap Ratio of Sample Activity Levels Identified by Sentence-Level Probability Across Model Variants

Since the identification of inactive samples relies on trained NMT models, a critical question arises: Are these identified inactive samples model-dependent? In other words, do different NMT models mark distinct portions of training data as inactive? To address this, we perform data binning and compute the proportion of samples shared among LSTM, Transformer, and DynamicConv models. A higher shared proportion indicates greater consistency across models, suggesting that these samples are not influenced by specific model architectures.

Following Wang et al. (Jiao et al., 2020), we partition the data into 10 equal deciles (each containing 10% of training samples). Figure 5 presents results from the multi-dimensional evaluation method across three model architectures. For inactive samples (first decile), the overlap ratio consistently exceeds 80% across architectures, with highly active samples (tenth decile) also showing strong consistency. This high consistency suggests that inactive sample identification depends more on data distribution than specific model architectures. Figure 6 compares results from sentence-level probability methods across the same architectures. The overlap ratios for the least and most active samples are 60% and 57%, respectively, significantly lower than those from the multi-dimensional method. This indicates poorer identification performance, greater susceptibility to model architecture, and reduced stability.

4.3 Activation of Inactive Samples

This section first evaluates all samples using the identification model's multi-metric assessment, computing composite scores. The lowest-scoring $R\%$ (Ratio) samples are marked as inactive, and the impact of activating varying proportions of inactive samples on translation performance is analyzed. Experimental results demonstrate that activating inactive samples consistently outperforms the non-activated control, validating the effectiveness and necessity of data activation. As shown in Figure 7, BLEU scores exhibit a declining trend with increasing $R\%$ values. This trend is expected, as some relatively higher-scoring samples still contribute to the NMT model, and their rejuvenation may degrade translation quality. Therefore, in subsequent experiments, the lowest-scoring 10% of samples are treated as inactive.

4.4 Main Result

This section presents experimental results of the Data Rejuvenation method on four agglutinative-to-Chinese translation tasks: Mongolian-Chinese (mn-

Model	mn-zh	ug-zh	ko-zh	ar-zh
Existing NMT Systems				
LSTM	26.82	27.10	24.43	28.17
Transformer-Base	27.34	28.21	30.45	33.35
Transformer-Big	31.78	33.41	31.42	35.14
Transformer + CSGAN	34.81	32.64	31.84	35.64
DynamicConv	33.25	32.32	31.69	37.28
GCN	30.41	30.23	31.52	32.45
GCN+att	31.62	32.34	31.95	33.74
Our NMT Systems (with Data Rejuvenation)				
LSTM + Agglutinative Language Data Rejuvenation	28.74↑ (+1.92)	29.26↑ (+2.16)	27.13↑ (+2.70)	30.18↑ (+2.01)
Transformer-Base + Agglutinative Language Data Rejuvenation	30.65↑ (+3.31)	31.52↑ (+31.1)	32.58↑ (+2.13)	36.84↑ (+3.49)
Transformer-Big + Agglutinative Language Data Rejuvenation	35.54↑ (+3.76)	34.91↑ (+1.50)	34.53↑ (+3.7)	39.81↑ (+4.67)
DynamicConv + Agglutinative Language Data Rejuvenation	36.58↑ (+3.33)	35.20↑ (+2.88)	34.22↑ (+2.53)	40.54↑ (+3.26)

Table 1: Evaluation of translation performance (BLEU scores) across model architectures and language pairs. "↑": indicates statistically significant improvement over the corresponding baseline.

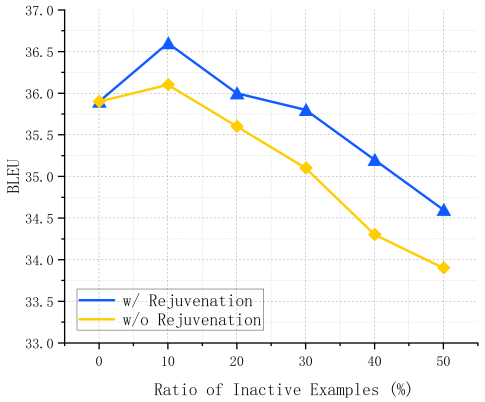

Figure 7: Effect of Activating different proportions of inactive samples on translation performance.

zh) (Qing-dao-er ji et al., 2020), Uyghur-Chinese (ug-zh) (Wang et al., 2019; Xu et al., 2021), Korean-Chinese (ko-zh), and Arabic-Chinese (ar-zh). As shown in Table 1, Data Rejuvenation consistently outperforms baseline models across LSTM, Transformer, and DynamicConv architectures.

For Mongolian-Chinese (mn-zh), the LSTM model improves from 26.8 to 28.7 BLEU (+1.9), Transformer-Base from 27.3 to 30.6 (+3.3), Transformer-Big from 31.7 to 35.5 (+3.8), and DynamicConv from 33.2 to 36.5 (+3.3). Similar improvements are observed in other language pairs: DynamicConv achieves 37.8 BLEU (+3.0) for Uyghur-Chinese, Transformer-Big reaches 36.7 (+4.4) for Korean-Chinese, and DynamicConv attains 40.5 (+3.3) for Arabic-Chinese.

These results demonstrate the effectiveness and generalization capability of Data Rejuvenation across agglutinative languages. Notably, these improvements are achieved without additional data or significant model modifications, highlighting its practicality in resource-constrained scenarios.

4.5 Comparative Experiment

Training Data	BLEU	Δ
Raw Data	32.3	-
- 10% mul_Inactive Examples	35.58	+3.28
+ Rejuvenated Examples	36.47	+4.17
- 10% mul_Inactive Examples	35.58	+3.28
+ Forward Translation	34.1	+1.8
- 10% sent_Inactive Examples	33.6	+1.3
+ Rejuvenated Examples	34.87	+2.57
- 10% sent_Inactive Examples	33.6	+1.3
+ Forward Translation	33.2	+0.9

Table 2: A comparison is made between different methods of identifying and activating low-contribution samples and their resulting impact on the final NMT model training outcomes.

This section designs a comparative experiment to evaluate the combined effects of different inactive sample identification and activation methods in Mongolian-Chinese translation. We analyze their impact on final NMT model training and explore the role of two distinct models in data optimization. Experimental results (Table 2) show that: 1) sentence-level probability identification combined with target-side data augmentation improves translation quality; 2) multi-dimensional evaluation paired with forward translation also enhances model training. However, our proposed method—combining multi-dimensional evaluation with target-side data augmentation for inactive sample activation—achieves the best overall perfor-

mance. This demonstrates that our approach significantly improves inactive sample activation quality in Mongolian-Chinese translation, establishing a solid foundation for low-resource language data optimization.

5 Conclusion

This study proposes a data rejuvenation method for agglutinative language-to-Chinese NMT, combining multi-dimensional evaluation for precise inactive sample identification with target-side data augmentation for rejuvenation. Experiments show significant performance improvements across NMT architectures (LSTM, Transformer, DynamicConv) and language pairs (Mongolian-Chinese, Uyghur-Chinese, Korean-Chinese, Arabic-Chinese), while enhancing model stability and generalization. Compared to sentence-level probability methods, our approach better captures local confidence fluctuations in agglutinative translation and mitigates forward-translation instability. The framework optimizes data distribution without additional training data, offering a universal solution for low-resource scenarios. Future work will explore deep feature learning for inactive sample identification and extend applications to more agglutinative languages.

6 Limitation

Threshold Dependency: The evaluation system uses heuristic thresholds (e.g., τ) to detect inactive samples. While empirically validated, these thresholds may need manual tuning for different languages/datasets. Automating their selection (e.g., via reinforcement learning) could improve adaptability in low-resource settings.

Computational Cost: The target-side rejuvenation mechanism increases training overhead. Decomposition reduces memory usage, but latent space modeling and iterative sampling slow down inference, especially for morphologically complex sentences. Future work may employ lightweight latent representations or parallelized sampling to optimize efficiency.

Language Coverage: Experiments are limited to agglutinative languages (e.g., Mongolian, Uyghur) with SOV-to-SVO divergence. Generalizing to typologically diverse languages (e.g., polysynthetic Inuktitut) may require adjustments for unique morphological or alignment features.

Acknowledgements

This study is supported by the National Natural Science Foundation of China (62206138, 62466044), Inner Mongolia Natural Science Foundation (2024MS06009, 2024MS06017, 2024QN06021), Research Program of Science and Technology at Universities of Inner Mongolia Autonomous Region (NJZZ23081), Inner Mongolia Basic research expenses (ZTY2025072), and Science Research Foundation of Inner Mongolia University of Technology (BS2021079).

References

Eleftheria Briakou and Marine Carpuat. 2022. Can synthetic translations improve bitext quality? In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4753–4766.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255.

Liang Ding, Longyue Wang, Xuebo Liu, Derek F Wong, Dacheng Tao, and Zhaopeng Tu. 2021. Rejuvenating low-frequency words: Making the most of parallel data in non-autoregressive translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3431–3441.

Liang Ding, Longyue Wang, Shuming Shi, Dacheng Tao, and Zhaopeng Tu. 2022. Redistributing low-frequency words: Making the most of monolingual data in non-autoregressive translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2417–2426.

Tobias Domhan. 2018. How much attention do you need? a granular analysis of neural machine translation architectures. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1799–1808.

Sergey Edunov, Myle Ott, Marc'Aurelio Ranzato, and Michael Auli. 2020. On the evaluation of machine translation systems trained with back-translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2836–2846.

Yuan Gao, Feng Hou, and Ruili Wang. 2024. A novel two-step fine-tuning framework for transfer learning in low-resource neural machine translation. In *Findings of the Association for Computational Linguistics 2024*, pages 3214–3224.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. 2017. Convolutional sequence to sequence learning. In *International conference on machine learning*, pages 1243–1252.

Yanzhu Guo, Guokan Shang, Michalis Vazirgiannis, and Chloé Clavel. 2024. The curious decline of linguistic diversity: Training language models on synthetic text. In *Findings of the Association for Computational Linguistics*, pages 3589–3604.

Yatu Ji, Hongxu Hou, Chen Junjie, and Nier Wu. 2019. Improving mongolian-chinese neural machine translation with morphological noise. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 123–129.

Wenxiang Jiao, Xing Wang, Shilin He, Irwin King, Michael Lyu, and Zhaopeng Tu. 2020. Data rejuvenation: Exploiting inactive training examples for neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 2255–2266.

Bo Jin. 2024. Neural machine translation based on semantic word replacement. In *Proceedings of the 2024 International Conference on Generative Artificial Intelligence and Information Security*, pages 106–112.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39.

Alex Krizhevsky, Geoffrey Hinton, and 1 others. 2009. Learning multiple layers of features from tiny images. *Handbook of Systemic Autoimmune Diseases*, 1(4).

Aviral Kumar and Sunita Sarawagi. 2019. Calibration of encoder decoder models for neural machine translation. *arXiv e-prints*, pages 1895–1903.

Bohan Li, Yutai Hou, and Wanxiang Che. 2022. Data augmentation approaches in natural language processing: A survey. *Ai Open*, 3:71–90.

Jiaming Luo, Colin Cherry, and George Foster. 2024. To diverge or not to diverge: A morphosyntactic perspective on machine translation vs human translation. *Transactions of the Association for Computational Linguistics*, 12:355–371.

Mohanad Mohamed and Sadam Al-Azani. 2025. Enhancing arabic nlp tasks through character-level models and data augmentation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 2744–2757.

Xuan-Phi Nguyen, Shafiq Joty, Kui Wu, and Ai Ti Aw. 2020. Data diversification: A simple strategy for neural machine translation. *Advances in Neural Information Processing Systems*, 33:10018–10029.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53.

Yirong Pan, Xiao Li, Yating Yang, and Rui Dong. 2020. Multi-source neural model for machine translation of agglutinative language. *Future Internet*, 12(6):96.

Ren Qing-dao-er ji, Yi La Su, and Wan Wan Liu. 2020. Research on the lstm mongolian and chinese machine translation based on morpheme encoding. *Neural Computing and Applications*, 32:41–49.

Dania Refai, Saleh Abu-Soud, and Mohammad J Abdel-Rahman. 2023. Data augmentation using transformers and similarity measures for improving arabic text classification. *IEEE Access*, 11:132516–132531.

Khaled Shaalan, Sanjeera Siddiqui, Manar Alkhatib, and Azza Abdel Monem. 2019. Challenges in arabic natural language processing. In *Computational linguistics, speech and image processing for arabic language*, pages 59–83.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wei Wang, Taro Watanabe, Macduff Hughes, Tetsuji Nakagawa, and Ciprian Chelba. 2018. Denoising neural machine translation training with trusted data and online data selection. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 133–143.

YaJuan Wang, Xiao Li, YaTing Yang, Azmat Anwar, and Rui Dong. 2019. Research of uyghur-chinese machine translation system combination based on semantic information. In *Natural Language Processing and Chinese Computing: 8th CCF International Conference, NLPCC 2019, Dunhuang, China, October 9–14, 2019, Proceedings, Part II 8*, pages 497–507.

Felix Wu, Angela Fan, Alexei Baevski, Yann Dauphin, and Michael Auli. 2019. Pay less attention with lightweight and dynamic convolutions. In *International Conference on Learning Representations*, pages 14–20.

Zhiwang Xu, Huibin Qin, and Yongzhu Hua. 2021. Research on uyghur-chinese neural machine translation based on the transformer at multistrategy segmentation granularity. *Mobile Information Systems*, 2021(1):5744248.

Ji Yatu, Zhang Huinuan, Wu Nier, Lu Min, Shi Bao, and 1 others. 2024. A review of mongolian neural machine translation from the perspective of training. In *2024 International Joint Conference on Neural Networks*, pages 1–10.