

A Semantic Uncertainty Sampling Strategy for Back-Translation in Low-Resources Neural Machine Translation

Anonymous ACL submission

Abstract

Back-translation has been proven effective in enhancing the performance of Neural Machine Translation (NMT), with its core mechanism relying on synthesizing parallel corpora to strengthen model training. However, while traditional back-translation methods alleviate the data scarcity in low-resource machine translation, their dependence on random sampling strategies ignores the semantic quality of monolingual data. This results in the contamination of model training through the inclusion of substantial low-quality samples in the generated corpora. To mitigate noise interference, additional training iterations or model scaling are required, significantly increasing computational costs. To address this challenge, this study proposes a Semantic Uncertainty Sampling strategy, which prioritizes sentences with higher semantic uncertainty as training samples by computationally evaluating the complexity of unannotated monolingual data. Experiments were conducted on three typical low-resource agglutinative language pairs: Mongolian-Chinese, Uyghur-Chinese, and Korean-Chinese. Results demonstrate an average BLEU score improvement of +1.7 on test sets across all three translation tasks, confirming the methods effectiveness in enhancing translation accuracy and fluency. This approach provides a novel pathway for the efficient utilization of unannotated data in low-resource language scenarios.

1 Introduction

The heavy reliance of NMT on large-scale parallel corpora significantly constrains performance improvement for low-resource languages (particularly minority languages), due to the difficulty in constructing high-quality bilingual datasets. In contrast, monolingual data has become a research focus given its accessibility, and methods leveraging monolingual resources to optimize model performance have been widely applied in low-resource

scenarios (Edunov et al., 2018; Xu et al., 2022; Had-dow et al., 2022; Ranathunga et al., 2023). Among these approaches, back-translation as a representative semi-supervised method breaks through the constraints of manual annotation by reversely generating pseudo-parallel data. It has been validated as a core strategy for enhancing translation quality (Sennrich et al., 2016a; Poncelas et al., 2018) and has become standard practice in building large-scale NMT systems due to its practicality (Siddhant et al., 2020; Huang et al., 2021).

Nevertheless, conventional back-translation implementations typically employ unfiltered monolingual corpora. While capitalizing on data abundance, this practice inevitably incorporates syntactically simplistic or semantically homogeneous sentences a dual detriment that not only squanders computational resources but also introduces noise that undermines models' capacity to capture sophisticated linguistic patterns. Although recent studies (Edunov et al., 2018) have attempted to enhance output diversity through optimized beam search strategies (Meister et al., 2020), these methods remain insufficient in mitigating the inherent noise from semantically redundant training instances. This limitation manifests as constrained model generalization capabilities, exposing critical gaps in proactive quality screening mechanisms for corpus curation.

To address these issues, this study proposes a semantic uncertainty back-translation sampling strategy. By identifying monolingual sentences with high semantic uncertainty and leveraging them for back-translation, this method efficiently improves model performance and mitigates the scarcity of low-resource corpora. Large-scale experiments demonstrate that the proposed uncertainty-based sampling strategy for self-training significantly outperforms random sampling. Extensive analysis of the generated outputs validates our claims and contributes to existing research in the following ways:

Demonstrates the necessity of semantic uncertainty sampling for back-translation.

Proposes a semantic uncertainty-aware back-translation sampling strategy, empirically validated for feasibility in low-resource language scenarios.

Transfers semantic information from the target language to the source language in low-resource settings, reducing the translation models reliance on parallel corpora.

2 Related Work

The development of data augmentation techniques for low-resource neural machine translation has seen researchers continuously overcome the bottleneck of parallel corpora through multi-dimensional innovations. Back-translation has been extensively explored: The monolingual data back-translation paradigm pioneered by Sennrich et al. (Sennrich et al., 2016b) established the foundation for pseudo-data generation. Subsequently, Daimeng et al. (Wei et al., 2023) introduced text style transfer technology (TST BT) to align generated data more closely with natural language distribution characteristics. Concurrently, Jiao et al. (Jiao et al., 2021) proposed a self-training strategy based on uncertainty probability from bilingual dictionaries, enhancing the model’s predictive capability for low-frequency words by filtering high-uncertainty monolingual sentences. Wei et al. (Wei et al., 2022) proposed an adjacency semantic space modeling framework, which dynamically partitions semantic boundaries and selects high-quality samples through a Gaussian mixture cyclic chain algorithm, achieving systematic optimization.

For neural machine translation of low-resource language pairs, researchers address challenges of corpus scarcity and morphological complexity through multi-dimensional technological innovations. In Mongolian-Chinese translation, Ji et al. (Ji et al., 2019) enhanced model robustness by injecting Mongolian morphological noise via an adversarial training framework. Zhang’s team (Zhang et al., 2023) optimized document-level context modeling through dual encoders with dynamic caching mechanisms. Sun et al. (Sun et al., 2021) combined back-translation with a dual-learning framework, achieving a 22% improvement in translation robustness. In Uyghur-Chinese translation, Feng et al. (Feng et al., 2023) designed an ensemble pruning algorithm based on back-translation to balance resource consump-

tion and performance, while Yan et al. (Yan et al., 2024) improved Uyghur-to-Chinese translation performance by leveraging zero-resource transfer learning in multilingual translation models. For Korean-Chinese translation, Li et al. (Li et al., 2023) proposed the LW-Transformer model incorporating pre-normalization and localized self-attention mechanisms, which significantly improved Sino-Korean machine translation performance. These approaches synergistically advanced the practical application of low-resource translation technology through multi-level system collaboration.

At the foundational architecture level, the evolution of cross-lingual pretraining models has injected new momentum into low-resource language research. Although general models like XLM-R (Conneau et al., 2020) excel in multilingual tasks, their support for Chinese minority languages remains limited. The CINO model (Yang et al., 2022), through secondary pretraining on corpora of Tibetan, Mongolian (Uyghur script) and Uyghur, achieved a 13% Macro-F1 improvement over baselines, providing critical infrastructure for low-resource language studies. These advancements jointly enhance the robustness and domain adaptability of translation models in resource-constrained scenarios.

The performance enhancement of low-resource NMT remains constrained by three factors: agglutinative morphological structures, free word-order characteristics, and scarce parallel corpora. While existing methods demonstrate commendable results in specific domains, two critical limitations persist: (1) Traditional data filtering strategies fail to effectively capture the semantic complexity of low-resource languages; (2) Current evaluation systems lack fine-grained quantitative analysis of translations. To address these issues, this study proposes semantic uncertainty sampling, which optimizes training sample selection through dynamic evaluation of uncertainty distributions in source-target semantic spaces, while employing multiple evaluation metrics to comprehensively assess model performance.

3 Methodology

As proposed by Zhou et al. (Zhou et al., 2019), the complexity of parallel corpora can be quantified by aggregating the translation uncertainty across all source sentences. Formally, for a source sentence x ,

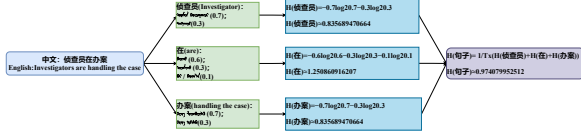


Figure 1: Graph of semantic uncertainty computation

the translation uncertainty of its selected translation y can be formulated as the conditional entropy:

$$\mathcal{H}(Y|X=x) = - \sum_{y \in \mathcal{Y}} p(y|x) \log p(y|x) \quad (1)$$

$$\approx \sum_{t=1}^{T_x} \mathcal{H}(y|x = x_t) \quad (2)$$

Here, T_x denotes the length of the source sentence, X and Y represent the sets of source-language and target-language sentences, respectively. x and y denote specific instances of source and target sentences, while x_t and y_t correspond to words in the source and target vocabularies. x_t indicates the segmented sub-unit. Generally, a high $\mathcal{H}(Y|X=x)$ (conditional entropy) implies a greater number of plausible translation candidates for the source sentence X . Equation (2) estimates the translation uncertainty of a source sentence by aggregating all potential translation candidates from a parallel corpus. However, this method cannot be directly applied to monolingual sentences due to the absence of corresponding translation candidates.

To address this limitation, Jiao et al. (Jiao et al., 2021) utilized authentic parallel corpora to estimate the target word distribution $P(y|x)$ conditioned on each source word x . This distribution is then employed to quantify the translation uncertainty of monolingual instances. Furthermore, the process incorporates bilingual dictionaries as reference knowledge to measure the uncertainty of monolingual sentences.

Although Jiao’s method provides a partial solution, it still has limitations. In our experiments, the lack of sufficient parallel corpora makes obtaining precise translation probabilities extremely difficult, directly resulting in the loss of critical information during computation. These factors collectively constrain the effectiveness of improving translation quality through alignment methods alone.

Therefore, this paper employs multilingual models to directly estimate word-level translation distributions. By introducing semantic similarity to refine translation probabilities, we use the model to

generate vectorized representations of the source word x and candidate target word y . The formula is extended as:

$$\mathcal{H}_{\text{sem}}(x) = - \frac{1}{T_x} \sum_{t=1}^{T_x} \sum_{i=1}^{y_i} p_{\text{sem}}(y_i | x_t) \log p_{\text{sem}}(y_i | x_t) \quad (3)$$

Here, H_{sem} denotes the semantic uncertainty on the source-language sentence x . For each source word x , the semantic similarity of the target word y is transformed into a probability:

$$p_{\text{sem}}(y_i | x_t) = \frac{s(x_t, y_i)}{\sum_{y' \in \mathcal{Y}} s(x_t, y')} \quad (4)$$

Where $s(x_t, y_i)$ denotes the semantic similarity score between the source term x_t and target term y_i , $\mathcal{Y}$ represents semantically similar lexical items in the candidate targets. $\sum_{y' \in \mathcal{Y}} s(x_t, y')$ indicates the summation of semantic similarity scores over all candidate target terms $s(x_t, y_i)$, used for normalization.

$$p = \frac{[\alpha \cdot H_{\text{sem}}(x)]^\beta}{\sum_{\mathcal{M}_x} [\alpha \cdot H_{\text{sem}}(x)]^\beta}, \quad (5)$$

$$\alpha = \begin{cases} 1, & H_{\text{sem}}(x) \leq H_{\text{max}} \\ \max \left(\frac{2H_{\text{max}}}{H_{\text{sem}}(x)} - 1, 0 \right), & \text{else} \end{cases} \quad (6)$$

As shown in Figure 1, a cross-lingual model was used to calculate the semantic similarity between the Chinese sentence “侦查员在办案” (lit. “investigators are handling the case”) and its Mongolian counterpart. The process first involved detailed tokenization of the sentences, followed by entropy calculations for individual words to quantify internal uncertainty. The total sentence information entropy was approximately 0.974, indicating that the original sentence possesses a certain level of complexity and uncertainty.

According to the uncertainty measurement for monolingual data specified in Formula (3), the uncertainty-aware self-training sampling strategy prioritizes sampling sentences with relatively higher uncertainty. To ensure data diversity and mitigate the risk of dominance by overly uncertain sentences, we sample monolingual sentences based on an uncertainty distribution that penalizes peak uncertainty. Specifically, given the number of sentences to sample, the sampling probability is

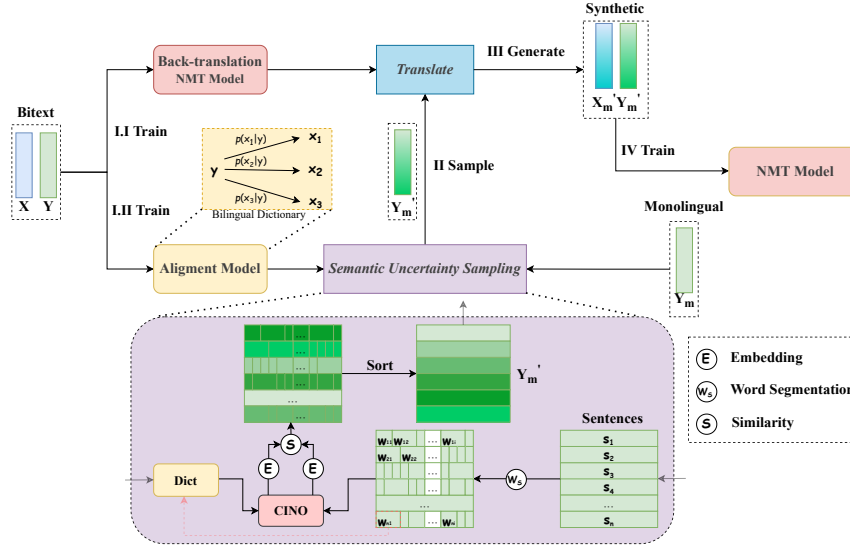
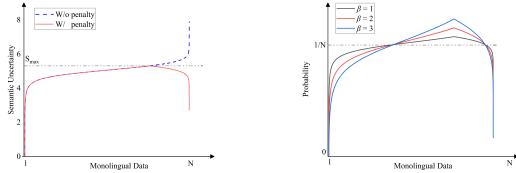


Figure 3: Semantic uncertainty sampling structure diagram: The proposed framework for self-training sampling based on semantic uncertainty is illustrated in the figure. The yellow and purple sections represent the methods integrated into the standard self-training framework. “Bitext”, “Monolingual” and “Synthetic” denote authentic parallel data, monolingual data, and synthetic parallel data, respectively.

controlled by configuring two hyperparameters as follows:

Here, α penalizes excessively high uncertainties surpassing the maximum uncertainty threshold U_{max} , while parameter β adjusts the distribution such that larger β values allocate more probability mass to sentences with higher uncertainty.



(a) Monolingual uncertainty probability distribution graph. (b) Sampling probability distribution graph.

Figure 2: Comparison of uncertainty and probability distributions.

As shown in Fig.2, the model’s performance under different monolingual data scales is demonstrated. When applying the penalty term (W/ penalty) with $\beta=3$, the model exhibits lower semantic uncertainty and higher probability increase rate under small data volumes; however, performance degradation occurs with increasing data due to over-regularization. In contrast, when $\beta=1$, the model effectively balances generalization capability and uncertainty control through gradual probability variations and stable regularization strength.

This indicates that the β value not only affects model stability on small datasets, but also determines its overfitting risk and performance on large datasets, highlighting β ’s pivotal role in regulating penalty term intensity.

The back-translation process involves several key steps: first, training a reverse NMT model on real parallel data; second, aligning words in the alignment model, computing semantic similarity, and sampling monolingual sentences based on semantic uncertainty; third, translating the sampled monolingual sentences using the reverse NMT model to generate synthetic parallel data; and finally, training a new NMT model on the combined synthetic and real parallel data. Figure 3 illustrates the framework of our semantic uncertainty-based sampling approach.

4 Experiments

4.1 Dataset

In this experiment, the research group utilized a Mongolian-Chinese NMT corpus comprising 1.2 million sentence pairs. The corpus spans multiple domains: 300k CCMT evaluation benchmarks, 200k government documents, 300k legal statutes, 50k historical archives, 100k specialized articles, daily conversational texts and other fields.. Additionally, the test set incorporates a challenging and representative 50k bilingual legal question-answer dataset (Zhaomuerlige and Wang, 2024).Through-

out the experiment, all corpora were tokenized using the Moses scripts. Sentences with lengths between 1 and 1000 tokens were retained from the original corpus. Subsequently, BPE (Sennrich et al., 2016b) with 40K merge operations was applied to enhance vocabulary representation efficiency and flexibility.

The monolingual Chinese corpus used for sampling tasks was sourced from the WMT2024 news dataset (Kocmi et al., 2024), which contains over 5 million sentences crawled in 2023.

To validate the generalizability and adaptability of the semantic uncertainty sampling method across low-resource language translation tasks, this study extended experiments beyond Mongolian-to-Chinese to include Korean-to-Chinese and Uyghur-to-Chinese translation tasks. This cross-lingual design ensures consistent performance across diverse language pairs.

For the Korean-to-Chinese task, the CCAIlgnd dataset (El-Kishky et al., 2020) was employed, containing approximately 1.02 million parallel sentences. In the Uyghur-to-Chinese task, a dataset with 600,000 parallel sentences was utilized.

4.2 Model

This study employs a standard TRANSFORMER architecture (Vaswani et al., 2017) as the core framework, comprising 6-layer stacked encoder modules and 6-layer symmetrical decoder modules. The implementation specifies a word embedding dimension of 512, with the feed-forward network hidden layer dimension expanded to 2048. Each attention sublayer incorporates 8 parallel attention heads. The system was deeply customized through the Fairseq (v0.10.2) open-source framework (Ott et al., 2019), strictly adhering to the TRANSFORMER_BASE parameter configuration scheme proposed by Vaswani et al. (Vaswani et al., 2017) (2017). Deployed on an NVIDIA GeForce RTX 3090 GPU (24GB VRAM) using PyTorch 1.9, the single-GPU training environment employed a mixed-precision training strategy to optimize VRAM utilization. Validation was performed after each epoch, with the best-performing intermediate model on the validation set retained as the final model.

4.3 Evaluation Metrics

Within our research framework, to ensure experimental objectivity and reliability while providing a solid reference for subsequent studies, we se-

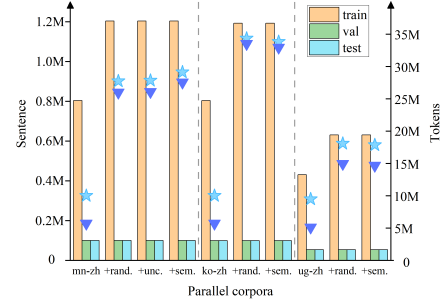


Figure 4: Parallel corpus diagram: The scale of the corpora used in the experiments is shown in the figure. The three sections separated by dashed lines represent the mn-zh, ko-zh, and ug-zh parallel corpora, respectively. The bar charts represent the number of sentences, while the (pentagram) and (triangle) markers denote the number of tokens in the training sets.

lected multiple evaluation metrics to quantify machine translation system performance. Specifically, we employ the sacreBLEU (Post, 2018) tool to compute BLEU (Bilingual Evaluation Understudy) (Papineni et al., 2002) scores as the primary metric, supplemented by CHRF (Character n-gram F-score) (Popović, 2015) and TER (Translation Edit Rate) (Snover et al., 2006).

4.4 Experimental Results and Analysis

As shown in Figure 4, this chart illustrates the distribution of sentence counts and word numbers across different parallel corpora. For the Mongolian-Chinese (mn-zh), Korean-Chinese (ko-zh), and Uyghur-Chinese (ug-zh) parallel corpora, the training sets exhibit varying degrees of expansion in both sentence counts and word numbers after applying three augmentation methods: random sampling, uncertainty-aware sampling, and semantic uncertainty-aware sampling. For instance, the mn-zh corpus increased its training sentences from approximately 0.8 million to around 1.2 million through these augmentation methods, with word counts correspondingly rising from 1.1 million to about 2.8 million. Overall, the chart clearly demonstrates the scale variations of different corpora across datasets.

This study employed the experimental configuration described in Section 4.2, using the TRANSFORMER_BASE model as the base architecture. It was compared against several sampling methods: Baseline, Random Sampling, Uncertainty Sampling, and our proposed Semantic Uncertainty Sampling. The experiments aimed to evaluate the impact of different sampling strategies on machine

System	Model	BLEU-4		sacreBLEU		chrF		TER	
(Zhang et al., 2024)	BITEXT	—	—	32.73	—	—	—	—	—
	+Easy Data Augmentation	—	—	33.15	—	—	—	—	—
	+Back Translation	—	—	33.57	—	—	—	—	—
	+ Iterative Back-Translation	—	—	34.55	—	—	—	—	—
(Wei and Ren, 2024)	BITEXT	—	—	32.48	—	—	—	—	—
	+Methods(Dropout)	—	—	33.93	—	—	—	—	—
	+Methods(Swap)	—	—	35.16	—	—	—	—	—
	+Methods(Replacement)	—	—	35.27	—	—	—	—	—
		16w	Law	16w	Law	16w	Law	16w	Law
<i>This Work</i>	BITEXT(mn-zh)	27.87	15.48	33.8	23.6	31.1	22.0	64.1	66.0
	+40w randomSamp	31.34	22.17	35.4	29.3	32.8	26.7	60.3	59.1
	+40w UncSamp	31.12	22.64	35.2	29.7	32.7	27.0	60.3	58.8
	+40w SemUncSamp(ours)	31.48	22.38	35.8	31.1	33.3	28.4	59.5	57.7

Table 1: Model Performance Scores on 16w and Law Domains: “16w” represents a test set of 160,000 sentences selected from the original Mongolian-Chinese parallel corpus, strictly independent of the training and validation sets; “law” denotes the legal Q&A dataset(Zhou et al., 2019). Lower TER indicates better performance.

translation performance for the Mongolian dataset.

According to Table 1, the model performance comparison among three research teams in machine translation tasks is evaluated using BLEU-4, sacreBLEU, chrF and TER metrics. Zhang(Zhang et al., 2024) employed progressive data augmentation techniques (e.g., iterative back-translation) on the BITEXT model to enhance sacreBLEU from 32.73 to 34.55. Wei(Wei and Ren, 2024) achieved the highest sacreBLEU score of 35.27 among compared methods through regularization-based model improvements using replacement strategies. Our experiments on general domain (16w) and legal domain (Law) datasets revealed insufficient domain adaptability of the baseline model, manifesting in a BLEU-4 of merely 15.48 and a TER as high as 66.0 for Law dataset. The proposed Semantic-UncSamp method optimized sampling strategies to achieve comprehensive optimal performance on Law dataset with sacreBLEU 31.1, chrF 28.4 and TER 57.7, demonstrating dual improvements in fluency and accuracy for specialized domain translation, particularly validating its effectiveness in vertical fields like legal translation. Furthermore, Uncertainty Sampling (UncSamp) elevated BLEU-4 to 22.64 on Law dataset, indicating the superiority of flexible sampling strategies over conventional data augmentation methods. Collectively, our work demonstrates that focused optimization of sampling strategies can more significantly enhance translation performance compared to traditional data augmentation approaches, effectively balancing semantic diversity enhancement with noise re-

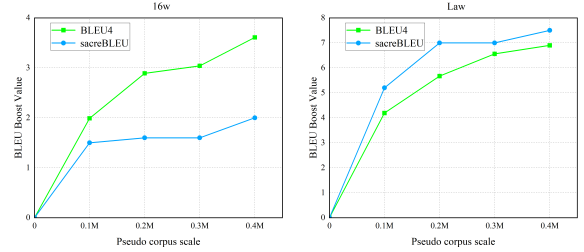


Figure 5: The Impact of Different Scales of Pseudo Corpora in Mixed Corpora on Translation Results

duction, thereby providing an optimized direction for machine translation model refinement.

Figure 5 demonstrates the impact of back-translation data scale on model performance in Mongolian-to-Chinese translation. Initially (pseudo-corpus scale=0), the model achieves baseline values of 15.48 BLEU4 and 23.6 sacreBLEU. With pseudo-corpus expansion, performance improves significantly: at 0.1M scale, both metrics show rapid enhancement, indicating that minimal back-translation data suffices for improvement; when reaching 0.4M scale, BLEU4 rises to approximately 22.38 with concurrent sacreBLEU growth. Overall, model capability strengthens continuously with pseudo-corpus enlargement, confirming back-translation’s effectiveness in low-resource scenarios most pronounced in small-data conditions, while performance gains gradually stabilize yet maintain steady progression with increased data volume.

To validate whether different data partitioning strategies affect model performance, Table 2 presents control experiments using three indepen-

Metric	Uyghur-to-Chinese			Korean-to-Chinese		
	BITEXT	+40w RandSamp	+40w SemUncSamp(ours)	BITEXT	+40w RandSamp	+40w SemUncSamp(ours)
BIEU-4	29.54	30.9	31.08	36.80	37.08	37.16
Precision 1-gram	60.9	62.2	62.2	59.7	60.2	60.9
Precision 2-gram	35.0	36.6	36.7	41.7	41.9	42.8
Precision 3-gram	23.4	24.8	24.9	34.2	34.2	35.3
Precision 4-gram	16.7	17.8	18.0	29.5	29.5	30.5
sacreBLEU	35.7	36.9	37.6	40.2	40.9	41.4
chrF	32.8	33.9	34.9	45.0	45.4	46.0
TER	58.4	57.3	56.4	57.1	56.6	55.7

Table 4: The comparative results of various model evaluation metrics on Uyghur-to-Chinese and Korean-to-Chinese translation datasets. Notably, lower TER score indicates superior model performance.

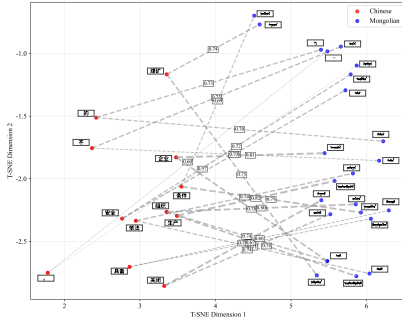


Figure 8: Chinese vs. Mongolian Text Embeddings in t-SNE Space

diverse bilingual alignment patterns.

This study employs t-SNE technique to perform dimensionality reduction visualization on Chinese-Mongolian bilingual word embedding spaces, generating a 2D mapping atlas (Figure 8) that reveals cross-lingual semantic alignment characteristics. Results indicate significant clustering between Chinese (red) and Mongolian (blue) lexical items in low-dimensional space, encompassing cross-lingual mappings of both domain-specific terms and high-frequency lexical items. The semantically correlated networks connected by gray dashed lines (annotated with confidence levels) further quantitatively validate cross-lingual lexical similarities, providing intuitive evidence for machine translation model evaluation.

This study conducted supplementary comparative experiments targeting Korean-to-Chinese and Uyghur-to-Chinese translation tasks to further validate the performance of the proposed sampling strategy across different language pairs.

Experimental results (Table 4) demonstrate the superiority of the semantic uncertainty-aware sampling strategy in Uyghur-Chinese and Korean-Chinese translation tasks. The method effectively improves translation quality even in lin-

guistically divergent contexts, such as those involving substantial syntactic and lexical disparities. For the Uyghur-Chinese task, the approach outperforms baseline models across all metrics (BLEU, chrF, and TER). In the Korean-Chinese task, leveraging 400k semantically uncertain training instances achieves state-of-the-art performance, including a BLEU4 score of 37.16 and optimal sacreBLEU/chrF values. These findings confirm the strategy’s capability to model cross-lingual semantic correspondences, significantly enhancing translation robustness in morphosyntactically distinct language pairs.

5 Conclusion

In this work, addressing the dependency of back-translation tasks on high-quality data in NMT, this paper proposes a semantic uncertainty-based sampling strategy. By identifying and sampling monolingual data with higher semantic uncertainty, this method enhances the quality of training data in the back-translation process. Experimental results demonstrate that compared to traditional random sampling approaches, the semantic uncertainty-based sampling strategy achieves improved translation quality. It ensures that the data used in back-translation is both sufficient in quantity and higher in quality, enabling targeted resolution of the model’s weaknesses and blind spots.

6 Limitations

The experiment relies on advanced cross-lingual models; however, for low-resource languages, their training data volume is relatively limited, which may lead to insufficient generalization capabilities of the models. Consequently, how to enhance the performance of these models on specific low-resource languages has become a pressing issue to be addressed.

References

- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451. Association for Computational Linguistics.
- Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500. Association for Computational Linguistics.
- Ahmed El-Kishky, Vishrav Chaudhary, Francisco Guzmán, and Philipp Koehn. 2020. CCAIghed: A massive collection of cross-lingual web-document pairs. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 5960–5969. Association for Computational Linguistics.
- Xiao Feng, Ya-Ting Yang, Rui Dong, and Bo Ma. 2023. Uyghur and chinese machine translation system based on ensemble pruning. *Manufacturing Automation*, 45(2):69–73.
- Barry Haddow, Rachel Bawden, Antonio Valerio Miceli Barone, Jindich Helcl, and Alexandra Birch. 2022. Survey of low-resource machine translation. *Computational Linguistics*, 48.
- Guoping Huang, Lemaoy Liu, Xing Wang, Longyue Wang, Huayang Li, Zhaopeng Tu, Chengyan Huang, and Shuming Shi. 2021. Transmart: A practical interactive machine translation system. *Computing Research Repository*. ArXiv:2105.13072, Version 1.
- 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. Association for Computational Linguistics.
- Wenxiang Jiao, Xing Wang, Zhaopeng Tu, Shuming Shi, Michael Lyu, and Irwin King. 2021. Self-training sampling with monolingual data uncertainty for neural machine 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 2840–2850. Association for Computational Linguistics.
- Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, and 3 others. 2024. Findings of the WMT24 general machine translation shared task: The LLM era is here but MT is not solved yet. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1–46. Association for Computational Linguistics.
- Yongheng Li, Yahui Zhao, Guozhe Jin, Zhejun Jin, and Rongyi Cui. 2023. Local information fused transformer model for korean-chinese machine translation. In *2023 16th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, pages 1–6.
- Clara Meister, Tim Vieira, and Ryan Cotterell. 2020. Best-first beam search. *Transactions of the Association for Computational Linguistics*, 8:795–809.
- 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. Association for Computational Linguistics.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318. Association for Computational Linguistics.
- Alberto Poncelas, Dimitar Shterionov, Andy Way, Gideon Maillette de Buy Wenniger, and Peyman Passban. 2018. Investigating backtranslation in neural machine translation. In *Proceedings of the 21st Annual Conference of the European Association for Machine Translation*, pages 269–278.
- Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395. Association for Computational Linguistics.
- Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191. Association for Computational Linguistics.
- Surangika Ranathunga, En-Shiun Annie Lee, Marjana Prifti Skenduli, Ravi Shekhar, Mehreen Alam, and Rishemjit Kaur. 2023. Neural machine translation for low-resource languages: A survey. *ACM Comput. Surv.*, 55.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96. Association for Computational Linguistics.

681	Rico Sennrich, Barry Haddow, and Alexandra Birch.	<i>Human Language Technologies</i> , pages 419–430. As-	738
682	2016b. Neural machine translation of rare words with	sociation for Computational Linguistics.	739
683	subword units. In <i>Proceedings of the 54th Annual</i>		
684	<i>Meeting of the Association for Computational Lin-</i>	Ziyue Yan, Hongying Zan, Yifan Guo, and Hongfei Xu.	740
685	<i>guistics (Volume 1: Long Papers)</i> , pages 1715–1725.	2024. Transferring zero-shot multilingual chinese-	741
686	Association for Computational Linguistics.	chinese translation model for chinese minority lan-	742
		guage translation. In <i>2024 International Confer-</i>	743
687	Aditya Siddhant, Ankur Bapna, Yuan Cao, Orhan Firat,	<i>ence on Asian Language Processing (IALP)</i> , pages	744
688	Mia Chen, Sneha Kudugunta, Naveen Arivazhagan,	133–138.	745
689	and Yonghui Wu. 2020. Leveraging monolingual		
690	data with self-supervision for multilingual neural ma-	Ziqing Yang, Zihang Xu, Yiming Cui, Baoxin Wang,	746
691	chine translation. In <i>Proceedings of the 58th Annual</i>	Min Lin, Dayong Wu, and Zhigang Chen. 2022.	747
692	<i>Meeting of the Association for Computational Lin-</i>	CINO: A Chinese minority pre-trained language	748
693	<i>guistics</i> , pages 2827–2835. Association for Compu-	model. In <i>Proceedings of the 29th International</i>	749
694	tational Linguistics.	<i>Conference on Computational Linguistics</i> , pages	750
		3937–3949. International Committee on Computa-	751
695	Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea	tional Linguistics.	752
696	Micciulla, and John Makhoul. 2006. A study of trans-		
697	lation edit rate with targeted human annotation. In	Huinan Zhang, Yatu Ji, Nier Wu, and Min Lu. 2024. A	753
698	<i>Proceedings of the 7th Conference of the Association</i>	mongolianchinese neural machine translation method	754
699	<i>for Machine Translation in the Americas: Techni-</i>	based on semantic-context data augmentation. <i>Ap-</i>	755
700	<i>cal Papers</i> , pages 223–231. Association for Machine	<i>plied Sciences</i> , 14(8).	756
701	Translation in the Americas.		
		Junjin Zhang, Yonghong Tian, Zheyu Song, and Yufeng	757
702	Shuo Sun, Hongxu Hou, Nier Wu, Xin Chang, Xi-	Hao. 2023. Mongolian-chinese machine translation	758
703	aoning Jia, and Haoran Li. 2021. Iterative knowl-	based on text context information. In <i>2023 8th Inter-</i>	759
704	edge refinement-based dual learning for mongolian-	<i>national Conference on Intelligent Computing and</i>	760
705	chinese machine translation. <i>Journal of Xiamen Uni-</i>	<i>Signal Processing (ICSP)</i> , pages 1741–1744.	761
706	<i>versity (Natural Science)</i> , 60(4):687–692.		
		Chao Zhaomuerlige and Sirigulun Wang. 2024.	762
707	Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob	Chinese-mongolian bilingual legal domain question-	763
708	Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz	answering corpus dataset. <i>China Scientific Data</i> ,	764
709	Kaiser, and Illia Polosukhin. 2017. Attention is all	9(04):83–91.	765
710	you need. In <i>Advances in neural information pro-</i>		
711	<i>cessing systems</i> , pages 5998–6008.	Chunting Zhou, Graham Neubig, and Jiatao Gu.	766
		2019. Understanding knowledge distillation in	767
712	Daimeng Wei, Zhanglin Wu, Hengchao Shang, Zongyao	non-autoregressive machine translation. <i>CoRR</i> ,	768
713	Li, Minghan Wang, Jiaxin Guo, Xiaoyu Chen,	abs/1911.02727.	769
714	Zhengzhe Yu, and Hao Yang. 2023. Text style trans-		
715	fer back-translation. In <i>Proceedings of the 61st</i>		
716	<i>Annual Meeting of the Association for Computa-</i>		
717	<i>tional Linguistics (Volume 1: Long Papers)</i> , pages		
718	7944–7959. Association for Computational Linguis-		
719	tics.		
720	Xiangpeng Wei, Heng Yu, Yue Hu, Rongxiang Weng,		
721	Weihua Luo, and Rong Jin. 2022. Learning to gen-		
722	eralize to more: Continuous semantic augmentation		
723	for neural machine translation. In <i>Proceedings of the</i>		
724	<i>60th Annual Meeting of the Association for Compu-</i>		
725	<i>tational Linguistics (Volume 1: Long Papers)</i> , pages		
726	7930–7944. Association for Computational Linguis-		
727	tics.		
728	Xuerong Wei and Qing-Dao-Er-Ji Ren. 2024. A		
729	language-driven data augmentation method for		
730	mongolian-chinese neural machine translation. In		
731	<i>2024 International Conference on Asian Language</i>		
732	<i>Processing (IALP)</i> , pages 297–302.		
733	Jiahao Xu, Yubin Ruan, Wei Bi, Guoping Huang, Shum-		
734	ing Shi, Lihui Chen, and Lemao Liu. 2022. On syn-		
735	thetic data for back translation. In <i>Proceedings of</i>		
736	<i>the 2022 Conference of the North American Chap-</i>		
737	<i>ter of the Association for Computational Linguistics:</i>		