EVERY BIT COUNTS: RESOURCE-EFFICIENT LEARNING FOR ENDANGERED LANGUAGE TRANSLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Despite rapid advances in large language models (LLMs), endangered languages remain severely underserved due to their limited representation in pretraining corpora and the scarcity of standardized linguistic resources. This challenge is particularly acute for machine translation, where existing approaches typically address resource integration, data augmentation, and learning optimization in isolation, lacking a systematic framework that connects them under extreme resource scarcity. To the best of our knowledge, we present the first unified framework that connects these three stages through model confidence as a shared signal, using Manchu translation as a case study. Rather than applying confidence uniformly, we adapt its role to the objective of each stage. First, **D**ictionary-guided **K**nowledge **C**orrection (DKC) leverages bilingual lexical resources to correct high-confidence errors inherited from pretraining. Second, **R**everse **L**exical **M**apping **A**ugmentation (RLMA) expands limited supervision by promoting diversity across confidence levels. Third, **C**onfidence-**A**ware **P**rogressive **S**election (CAPS) prioritizes low-confidence examples that remain insufficiently learned and progressively introduces them according to the evolving model state. On Manchu translation, our framework improves BLEU from 16.40 to 26.38. The results demonstrate the complementary roles of confidence in correcting unreliable knowledge, diversifying augmented supervision, and prioritizing unresolved examples, providing a unified approach to exploiting scarce linguistic resources for endangered language translation.

1 INTRODUCTION

Neural machine translation (NMT) is highly dependent on large-scale parallel corpora, whereas extremely low-resource languages often face insufficient parallel data, limited linguistic tools, and inadequate pretrained model coverage (Haddow et al., 2022). Although back-translation (Sennrich et al., 2016), cross-lingual transfer and multilingual pretraining (Xue et al., 2021; Aji et al., 2020; Liu et al., 2020b), and data augmentation (Fadaee et al., 2017) can expand supervision, they nevertheless depend on additional monolingual corpora, related high-resource languages, or large-scale pretraining resources that are difficult to obtain in extremely low-resource settings. Manchu provides a representative case of extremely low-resource translation, and previous studies have explored the use of external resources such as monolingual corpora, word embeddings, and language models to improve translation performance (Seo et al., 2023; Lee et al., 2024; Pei et al., 2025; Wang et al., 2026).

Bilingual dictionaries provide a relatively accessible source of supervision through explicit and low-cost lexical-level cross-lingual correspondences. Prior work has used dictionaries for auxiliary data augmentation (Jones et al., 2023; Zheng et al., 2024) or aligned lexical replacement in multilingual pretraining (Lin et al., 2020), but commonly relies on source-side tokenization or word alignment and does not organize dictionary knowledge and parallel data through staged coordination. Active learning organizes training data using model-dependent criteria (Settles, 2009; Pendas et al., 2023; Hu & Neubig, 2021; Azeemi et al., 2025; Xu et al., 2020; Mohiuddin et al., 2022; Zhao et al., 2020; Liu et al., 2020a; Zhou et al., 2020), but conventional methods require human or oracle annotation. Overall, existing studies provide limited guidance on how dictionary knowledge, augmented bilingual examples, and model-aware data selection can be coordinated under extremely limited parallel supervision.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

In response to the situation described above, an efficient learning framework for extremely low-resource languages is constructed by combining dictionary resources, parallel example sentences, and training strategies, with confidence serving as the guiding signal. The main contributions are as follows:

1. Dictionary-guided Knowledge Correction (DKC) uses the pretrained model's confidence to identify lexical samples requiring further learning and applies dictionary-based supervision to correct potentially unreliable cross-lingual knowledge, without treating high-confidence predictions as reliable translations.
2. Reverse Lexical Mapping Augmentation (RLMA) constructs additional bilingual training examples through reverse lexical retrieval and synchronized source-target replacement, without requiring Manchu tokenization or explicit word alignment.
3. Confidence-Aware Progressive Selection (CAPS) introduces a confidence-aware progressive strategy for organizing the augmented examples, allowing the same augmentation budget to be utilized according to the evolving model state.

2 Related Work

Representative methods for low-resource machine translation include back-translation (Sennrich et al., 2016), as well as transfer learning and multilingual pre-training (Xue et al., 2021; Aji et al., 2020; Liu et al., 2020b). However, these approaches typically depend on additional monolingual resources or language coverage during pretraining. By contrast, bilingual dictionaries provide direct lexical-level cross-lingual supervision. Previous studies have demonstrated the effectiveness of lexical knowledge for low-resource translation through rare-word replacement (Fadaee et al., 2017), dictionary-based training and lexical replacement (Seo et al., 2023; Jones et al., 2023; Zheng et al., 2024), and cross-lingual aligned substitution (Lin et al., 2020). Recent research has also applied dictionaries to large language models to enhance translation performance (Lu et al., 2024; Pei et al., 2025).

Active learning reduces annotation cost by selecting unlabeled examples that are most informative to the current model (Settles, 2009). Pendas et al. study active learning for low-resource Mapudungun translation (Pendas et al., 2023); Hu and Neubig investigate phrase-level active selection (Hu & Neubig, 2021); Zhao et al. combine uncertainty with word-frequency information for the selection of NMT data (Zhao et al., 2020); and Azeemi et al. further integrate uncertainty and diversity in hybrid active learning (Azeemi et al., 2025). Additionally, some related studies have adjusted the order of introduction of training data based on loss changes, sample difficulty, or model prediction status (Xu et al., 2020; Mohiuddin et al., 2022; Liu et al., 2020a; Zhou et al., 2020; Liang et al., 2021), and have demonstrated that model-state-driven data organization can improve NMT training performance under limited data. Unlike the aforementioned methods, the proposed framework achieves competitive translation performance using only a bilingual dictionary and a minimal amount of parallel corpus, without utilizing any external data.

3 Method

3.1 Overall Framework

Due to the scarcity of large-scale Manchu–Mandarin parallel corpora, models relying on fixed subword vocabularies may encounter difficulties in representing rare words and handling language-specific segmentation. We therefore adopt ByT5-base (Xue et al., 2022) as the backbone translation model. ByT5 operates on UTF-8 byte sequences without requiring a language-specific subword vocabulary, providing a consistent representation for dictionary-based lexical supervision and sentence-level translation learning.

As illustrated in Fig. 1, the proposed framework consists of three stages. First, DKC uses the pretrained model's confidence to identify high-confidence lexical samples for further learning. Since high confidence does not necessarily indicate translation correctness when the pretrained model has limited Manchu knowledge, DKC uses the corresponding dictionary translations, rather than the model's generated outputs, as supervision to update its cross-lingual lexical knowledge before

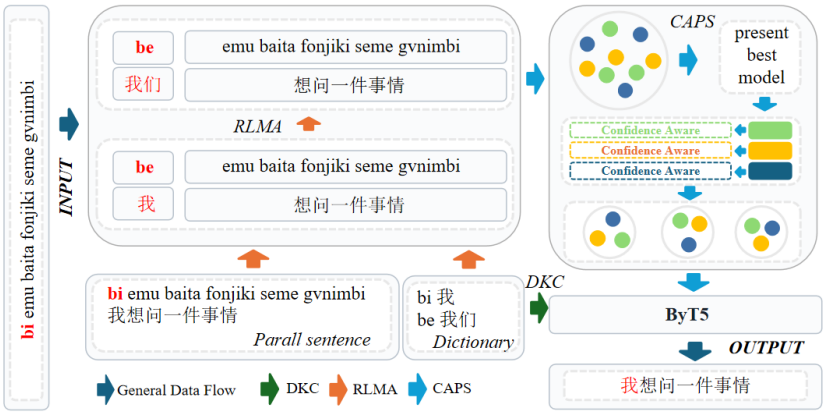

Figure 1: Overview of the proposed framework.

sentence-level training. Second, RLMA constructs additional bilingual training examples through reverse dictionary lookup and synchronized source–target replacement. Third, CAPS organizes the augmented examples according to the evolving model confidence. During sample selection, the reference translations of the augmented examples are withheld from confidence estimation; the selected examples are then incorporated into supervised training in three rounds, progressing from lower-confidence to higher-confidence samples.

Throughout the framework, confidence is measured using the average autoregressive log probability of the model-generated translation. Given a Manchu source sequence x, the model receives the source sequence with the task prefix and generates a Mandarin translation $\hat{Y} = (\hat{y}_1, \ldots, \hat{y}_T)$ using greedy decoding (`num_beams`=1). The confidence score is defined as

$$\mathrm{Conf}_\theta(x) = \frac{1}{T} \sum_{t=1}^{T} \log P_\theta(\hat{y}_t \mid \hat{y}_{<t}, x), \tag{1}$$

where θ denotes the model parameters at the stage when confidence is evaluated, $\hat{y}_{<t}$ is the previously generated target prefix, and T is the effective generated sequence length, excluding padding positions. The score is non-positive, with higher values indicating greater confidence in the model's own prediction and lower values indicating greater uncertainty. Importantly, this confidence score measures the model's certainty rather than translation correctness. It can be computed without accessing reference translations during sample selection, allowing the same scoring formulation to support confidence-guided lexical learning and the subsequent organization of augmented examples.

3.2 DKC

Bilingual dictionaries provide lexical-level cross-lingual correspondences that are particularly valuable for endangered languages with limited parallel corpora. However, pretrained multilingual models may have insufficient knowledge of Manchu and can assign high confidence to translations that are not necessarily correct. Rather than treating such predictions as reliable supervision, Dictionary-guided Knowledge Correction (DKC) uses model confidence to prioritize lexical samples for further learning and employs their dictionary-provided translations to update the model's cross-lingual knowledge.

Given a bilingual dictionary $\mathcal{D}_{\mathrm{dict}} = \{(X_i, Y_i)\}_{i=1}^{N}$, where X_i is a Manchu lexical entry and Y_i is its Mandarin translation, we compute the confidence $\mathrm{Conf}_{\theta_0}(X_i)$ using the pretrained ByT5 model with parameters θ_0, following Eq. 1. High-confidence entries are selected for dictionary-guided supervision:

$$\mathcal{D}_{\mathrm{DKC}} = \mathrm{Select}_{\mathrm{high}}\left(\mathcal{D}_{\mathrm{dict}}, \{\mathrm{Conf}_{\theta_0}(X_i)\}_{i=1}^{N}\right), \tag{2}$$

where $\mathrm{Select}_{\mathrm{high}}$ denotes the confidence-based selection operation. Importantly, high confidence is used to prioritize samples for supervised learning rather than to establish the correctness of the

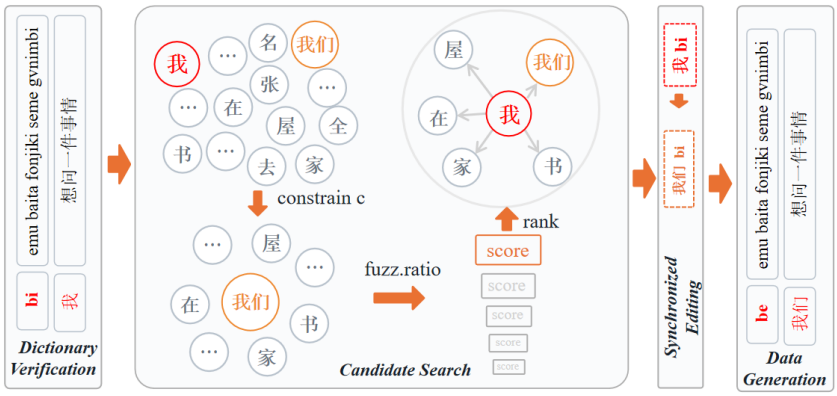

Figure 2: Illustration of RLMA.

model-generated translations. DKC does not require explicit error detection or use the generated translations as pseudo-labels.

For each selected dictionary pair (X, Y), ByT5 autoregressively models the conditional probability of the reference Mandarin translation as

$$P_\theta(Y \mid X) = \prod_{t=1}^{n} P_\theta(y_t \mid y_{<t}, X), \tag{3}$$

where $Y = (y_1, \ldots, y_n)$, $y_{<t} = (y_1, \ldots, y_{t-1})$, and θ denotes the model parameters. Both dictionary-guided supervision and subsequent sentence-level fine-tuning minimize the mean negative log-likelihood of the reference target sequence:

$$\mathcal{L}(X, Y; \theta) = -\frac{1}{n} \sum_{t=1}^{n} \log P_\theta(y_t \mid y_{<t}, X). \tag{4}$$

After DKC, the resulting model parameters are used to initialize sentence-level fine-tuning on the available Manchu–Mandarin parallel corpus. This allows the model to learn sentence-level contextual relationships and translation patterns based on its updated lexical knowledge.

3.3 RLMA

Reliable source-side lexical replacement is challenging for Manchu due to the lack of dedicated tokenization and word-alignment tools. RLMA addresses this limitation through reverse dictionary lookup and synchronized bilingual replacement. As illustrated in Fig. 2, the procedure consists of four steps: dictionary validation, candidate search, synchronized editing, and augmented data generation.

Before augmentation, the original parallel sentences are ranked according to model confidence, and the top 50% are selected as augmentation candidates. For each selected sentence, each Mandarin word has a 50% probability of entering the replacement procedure. An augmented example is accepted when the model confidence differs from that of the original sentence by 10%–30%. If no valid augmentation is obtained, the replacement procedure is repeated up to 10 times. When all attempts fail, the current sentence is discarded from augmentation, and the highest-confidence sentence from the remaining candidate pool is selected. This process continues until the number of newly added examples reaches 50% of the original training set.

Dictionary validation. Let the Manchu–Mandarin dictionary be

$$\mathcal{D} = \{(m_j, c_j)\}_{j=1}^{N_\mathcal{D}}, \tag{5}$$

where m_j is a Manchu term or phrase, c_j is one of its Mandarin interpretations, and $N_\mathcal{D}$ is the number of valid bilingual correspondences. Multiple Mandarin interpretations of the same Manchu entry are represented as separate dictionary pairs.

Given an original parallel example (x, y), RLMA first tokenizes the Mandarin target sentence y using `jieba`. For a Mandarin word c, its Manchu correspondences are retrieved from the dictionary and validated against the source sentence:

$$V(c, x) = \{m \mid (m, c) \in \mathcal{D},\ m \subseteq_{\text{str}} x\}, \tag{6}$$

where $\subseteq_{\text{str}}$ denotes substring occurrence. Only words with $V(c, x) \neq \varnothing$ proceed to candidate replacement. The corresponding Manchu term m_c is selected from the validated set, avoiding replacement of a Manchu term that does not occur in the original source sentence.

Candidate search. For each eligible Mandarin word c, RLMA searches the dictionary for alternative Mandarin interpretations with similar character lengths:

$$\mathcal{C}(c) = \left\{c' \in \mathcal{C}_{\mathcal{D}} \;\middle|\; c' \neq c,\ \big||c'| - |c|\big| \leq 2\right\}, \tag{7}$$

where $\mathcal{C}_{\mathcal{D}}$ denotes the set of Mandarin interpretations in the dictionary. Candidate similarity is measured using `fuzz.ratio` from RapidFuzz, denoted by $R(c, c') \in [0, 100]$. RLMA retains the 30 candidates with the highest character similarity and reranks them using a score $S(c, c')$ that incorporates a length-difference penalty. Only candidates satisfying $S(c, c') \geq 45$ are eligible for replacement. The selected Mandarin candidate is

$$c^* = \underset{c' \in \text{Top30}_R(\mathcal{C}(c))}{\arg\max} S(c, c'). \tag{8}$$

Synchronized editing. After selecting c^*, its corresponding Manchu term m^* is retrieved from the dictionary. RLMA then replaces the validated Manchu term and the original Mandarin word synchronously:

$$(x', y') = \left(\text{ReplaceFirst}(x, m_c, m^*), \text{ReplaceFirst}(y, c, c^*)\right), \tag{9}$$

where (x', y') denotes the augmented parallel example. This procedure uses dictionary correspondences and source-side substring validation without requiring Manchu tokenization or explicit word alignment.

Augmented data generation. The accepted augmented examples form the candidate pool

$$\mathcal{P}_0 = \{(x'_i, y'_i)\}_{i=1}^{M}, \tag{10}$$

where M is the number of augmented examples. The generated Mandarin targets y'_i serve as training targets after sample selection, while confidence estimation for the subsequent CAPS stage is performed without accessing these targets.

3.4 CAPS

CAPS progressively incorporates RLMA-generated examples into translation training according to the evolving confidence of the model. Unlike conventional uncertainty-based active learning, which selects unlabeled examples for annotation, CAPS operates on an existing pool of augmented bilingual examples. Their Mandarin targets are withheld during sample selection and are used only after the corresponding examples have been selected for supervised training. The model is first trained on the original parallel corpus. Subsequently, CAPS performs three rounds of confidence-aware selection, with budgets of 33.3%, 33.3%, and 33.4% of the initial augmented candidate pool, respectively.

Let the original parallel training set and the initial augmented candidate pool be

$$\mathcal{S}_0 = \{(x_i, y_i)\}_{i=1}^{N}, \qquad \mathcal{P}_0 = \{(x'_i, y'_i)\}_{i=1}^{M}, \tag{11}$$

where N and M denote the numbers of original and augmented parallel examples, respectively. At round $r \in \{1, 2, 3\}$, the model parameters θ^*_{r-1} obtained from the preceding training stage are used to generate translations for the Manchu source sequences in the remaining candidate pool $\mathcal{P}_{r-1}$. The confidence of each candidate is computed using Eq. 1:

$$c_i^{(r)} = \text{Conf}_{\theta^*_{r-1}}(x'_i). \tag{12}$$

The score depends only on the source sequence and the model-generated translation; the augmented Mandarin target y'_i is not used for confidence estimation. Since Eq. 1 defines confidence as the average log probability, lower scores indicate greater model uncertainty.

CAPS selects the B_r lowest-confidence examples from the remaining candidate pool:

$$\mathcal{Q}_r = \underset{\substack{\mathcal{Q}\subseteq\mathcal{P}_{r-1}\\|\mathcal{Q}|=B_r}}{\arg\min} \sum_{(x_i',y_i')\in\mathcal{Q}} c_i^{(r)}, \tag{13}$$

where B_r is the selection budget for round r. The selected examples are incorporated into the cumulative training set, while the remaining candidate pool is updated as

$$\mathcal{S}_r = \mathcal{S}_{r-1} \cup \mathcal{Q}_r, \qquad \mathcal{P}_r = \mathcal{P}_{r-1} \setminus \mathcal{Q}_r. \tag{14}$$

The model is then further trained on $\mathcal{S}_r$, using the augmented Mandarin targets as supervision. After each round, confidence is recomputed for the remaining candidates using the updated model, allowing subsequent selection to reflect changes in the model's translation knowledge. Across the three rounds, all augmented examples are progressively incorporated into training, while their introduction order is determined by the model's evolving confidence.

4 Experiments

4.1 Datasets

Dictionary. The dictionary data are collected from an online Manchu-Mandarin dictionary (`http://anakv.anakv.com/dict/`). After normalization, 15445 Manchu-Mandarin dictionary records are retained, including Manchu words and a small number of phrases. Because of lexical polysemy, a single Manchu entry may correspond to multiple Mandarin definitions. The dictionary is split 9:1 into training and validation sets for lexical-level pretraining.

Parallel sentences. The parallel examples are derived from materials such as everyday Manchu language, dialogue texts, and related books available on the internet. After normalization, 1409 Manchu to Mandarin parallel sentence pairs are obtained and split 7:1:2 into training, validation, and test sets. Unless otherwise stated, the terms training set, validation set, and test set below refer to this parallel-sentence dataset.

4.2 Setup

Parameter settings. All experiments use ByT5-base as the backbone model unless otherwise specified, and the reported results are averaged over three runs. During sentence-level training, the maximum source and target sequence lengths are both set to 512. We use a learning rate of 1×10^{-5}, a per-GPU batch size of 4, and 4 gradient accumulation steps. Training is conducted for up to 50 epochs, with early stopping after eight consecutive validation evaluations without improvement. The checkpoint with the lowest validation loss is retained.

Evaluation metrics. We evaluate translation quality using BLEU, chrF, ROUGE-L, and translation edit rate (TER). Higher BLEU, chrF, and ROUGE-L scores indicate greater agreement with the reference translations, whereas lower TER indicates fewer edits required to match the references. These complementary metrics provide a broader assessment of translation quality than any single metric.

4.3 Results

4.3.1 Effectiveness of DKC

We first compare ByT5-base (Xue et al., 2022) with two multilingual sequence-to-sequence models, mT5 (Xue et al., 2021) and NLLB-200 distilled 600M (Costa-jussà et al., 2024), to establish the baseline performance on Manchu–Mandarin translation. Unlike the subword-based tokenization used by mT5 and NLLB-200, ByT5 directly processes UTF-8 byte sequences without requiring a language-specific subword vocabulary. We then evaluate the effect of introducing DKC into ByT5. The results are reported in Table 1.

Without DKC, ByT5 achieves a BLEU score of 16.40, compared with 14.37 for NLLB-200 distilled 600M and 4.11 for mT5. ByT5 also obtains higher chrF and ROUGE-L scores and lower TER under the evaluated settings. These results support the use of ByT5 as the backbone model in our

Table 1: Baseline comparison and the effect of DKC.

Model Setting	BLEU↑	chrF↑	ROUGE-L↑	TER↓
NLLB-200 distilled 600M	14.37	13.66	42.98	96.82
mT5	4.11	5.15	21.75	99.65
ByT5	16.40	15.06	46.96	93.11
ByT5+DKC	**22.25**	**19.59**	**58.97**	**92.93**

Table 2: Comparison of data augmentation strategies using DKC-initialized ByT5.

Method	BLEU↑	chrF↑	ROUGE-L↑	TER↓
ByT5+DKC	22.25	19.59	58.97	92.93
ByT5+DKC+RMP	23.55	20.63	59.35	92.23
ByT5+DKC+KHSR	22.63	20.19	58.51	94.00
ByT5+DKC+MLA	24.93	**21.93**	**60.03**	91.87
ByT5+DKC+RLMA	**25.00**	**21.93**	59.75	**91.52**

experiments, although the differences cannot be attributed to tokenization alone because the models also differ in architecture, pretraining data, and training objectives.

Introducing DKC improves BLEU from 16.40 to 22.25, chrF from 15.06 to 19.59, and ROUGE-L from 46.96 to 58.97, while TER decreases slightly from 93.11 to 92.93. These results demonstrate that dictionary-guided training improves the evaluated ByT5 model's translation performance under limited parallel supervision. DKC uses the dictionary-provided translations to strengthen lexical-level cross-lingual learning before sentence-level fine-tuning, providing an updated model initialization for subsequent training. We therefore use ByT5 initialized through DKC as the starting point for the remaining experiments.

4.3.2 Effectiveness of RLMA

We evaluate RLMA against three data augmentation strategies using the DKC-initialized ByT5 model: Random Mask Perturbation (RMP), K-fold Hard Sample Reuse (KHSR), and Monolingual Lexical Augmentation (MLA). RMP randomly selects original training sentences and replaces each Manchu word with `MASK` at a probability of 50%, while leaving the Mandarin target unchanged. KHSR partitions the training data into five folds and iteratively uses four folds for pseudo-training and one for pseudo-validation. In each iteration, 10% of the samples from the pseudo-training and pseudo-validation subsets are selected as hard examples for reuse. MLA follows the same Mandarin-side candidate retrieval and replacement procedure as RLMA but leaves the Manchu source unchanged, serving as an ablation of synchronized bilingual replacement. RLMA instead retrieves the corresponding Manchu term through the bilingual dictionary and synchronously updates both sides of the parallel example. Each augmentation strategy adds examples equivalent to 50% of the original training set. The results are reported in Table 2.

Compared with the DKC baseline, RLMA improves BLEU from 22.25 to 25.00, chrF from 19.59 to 21.93, and ROUGE-L from 58.97 to 59.75, while reducing TER from 92.93 to 91.52. RMP yields smaller improvements across all four metrics, whereas KHSR provides a modest BLEU gain but decreases ROUGE-L and increases TER. These results indicate that the evaluated lexical augmentation strategies provide greater benefits than hard-sample reuse under the current experimental setting.

MLA provides a more direct comparison for examining synchronized bilingual replacement because it uses the same Mandarin-side candidate retrieval and replacement procedure as RLMA while keeping the Manchu source unchanged. Relative to MLA, RLMA increases BLEU by 0.07 and reduces TER by 0.35, achieves the same chrF score, and obtains a lower ROUGE-L score (59.75 versus 60.03). Thus, synchronized bilingual replacement yields small improvements in BLEU and TER under the evaluated setting, but does not consistently outperform target-side-only augmentation across all metrics. By updating the corresponding Manchu and Mandarin terms together, RLMA aims to preserve dictionary-level lexical correspondence in augmented parallel examples, although the present results do not isolate the contribution of this mechanism from the confidence-based augmentation and filtering procedures.

Table 3: Comparison of strategies for incorporating RLMA-generated examples into training.

Method	BLEU↑	chrF↑	ROUGE-L↑	TER↓
ByT5+DKC+RLMA	25.00	21.93	59.75	91.52
ByT5+DKC+RLMA+Random	25.92	22.79	60.81	91.66
ByT5+DKC+RLMA+CAPS	**26.38**	**22.87**	**61.61**	**91.17**

4.3.3 Effectiveness of CAPS

We evaluate how the RLMA-generated examples are incorporated into training by comparing three strategies: direct training with all augmented examples, random progressive selection, and CAPS. Both progressive strategies introduce the augmented examples over three rounds, with selection budgets of 33.3%, 33.3%, and 33.4% of the initial candidate pool, respectively. Each round continues training from the best checkpoint of the preceding round. The two progressive strategies ultimately use the same augmented examples and differ in their selection order: Random introduces examples in a random order, whereas CAPS prioritizes examples with lower confidence under the current model. The results are reported in Table 3.

Compared with direct training on all RLMA-generated examples, random progressive selection improves BLEU by 0.92, chrF by 0.86, and ROUGE-L by 1.06, although TER increases slightly from 91.52 to 91.66. This suggests that progressively introducing augmented examples can benefit translation performance under the evaluated training configuration. CAPS further improves BLEU from 25.92 to 26.38, chrF from 22.79 to 22.87, and ROUGE-L from 60.81 to 61.61, while reducing TER from 91.66 to 91.17. Since both progressive strategies use the same augmented examples and selection budgets, these results suggest that confidence-guided ordering provides additional benefits beyond random progressive selection in the evaluated setting.

5 Conclusion

This paper presents a resource-efficient framework for extremely low-resource Manchu-to-Mandarin translation, consisting of DKC, RLMA, and CAPS. All three use confidence as the measurement criterion. First, DKC injects vocabulary-level cross-lingual knowledge into ByT5 to provide a solid confidence foundation. Then, RLMA uses confidence to construct weakly labeled augmented data. Finally, in the CAPS process, confidence is combined with model-aware training.

DKC improves BLEU from 16.40 to 22.25, RLMA further improves it to 25.00, and after CAPS it reaches 26.38. Experimental results show that the proposed resource-efficient learning framework can achieve better results under very limited data conditions, providing a new research direction for endangered language translation.

AI Use Statement

Generative AI tools were used to assist with language editing, improving clarity, and refining the presentation of the manuscript. The authors reviewed and revised the AI-assisted text and take full responsibility for the paper's content. The authors independently developed the methods, experiments, and scientific claims.

Ethics statement

All materials used in this article are publicly available online resources.

Reproducibility statement

If the paper is accepted, we are willing to make relevant materials public.

Author Contributions

If the paper is accepted, the specific author contributions will be disclosed.

ACKNOWLEDGMENTS

If the paper is accepted, all acknowledgments such as funding will be supplemented.

REFERENCES

Alham Fikri Aji, Nikolay Bogoychev, Kenneth Heafield, et al. In neural machine translation, what does transfer learning transfer? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7701–7710, 2020. doi: 10.18653/v1/2020.acl-main.688.

Abdul Hameed Azeemi, Ihsan Ayyub Qazi, and Agha Ali Raza. To label or not to label: Hybrid active learning for neural machine translation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 3071–3082, 2025. URL https://aclanthology.org/2025.coling-main.206/.

Marta R. Costa-jussà, James Cross, Onur Çelebi, et al. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841–846, 2024. doi: 10.1038/s41586-024-07335-x.

Marzieh Fadaee, Arianna Bisazza, and Christof Monz. Data augmentation for low-resource neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 567–573, 2017. doi: 10.18653/v1/P17-2090.

Barry Haddow, Rachel Bawden, Antonio Valerio Miceli Barone, et al. Survey of low-resource machine translation. *Computational Linguistics*, 48(3):673–732, 2022. doi: 10.1162/coli_a_00446.

Junjie Hu and Graham Neubig. Phrase-level active learning for neural machine translation. In *Proceedings of the Sixth Conference on Machine Translation*, pp. 1087–1099, 2021. URL https://aclanthology.org/2021.wmt-1.117/.

Alex Jones, Isaac Caswell, Orhan Firat, et al. GATITOS: Using a new multilingual lexicon for low-resource machine translation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 371–405, 2023. doi: 10.18653/v1/2023.emnlp-main.26.

Sangah Lee, Sungjoo Byun, Jean Seo, et al. ManNER & ManPOS: Pioneering NLP for endangered manchu language. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, pp. 11030–11039, 2024. doi: 10.63317/3y253nmet9k9.

Chen Liang, Haoming Jiang, Xuebo Liu, et al. Token-wise curriculum learning for neural machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 3658–3670, 2021. doi: 10.18653/v1/2021.findings-emnlp.310.

Zehui Lin, Xiao Pan, Mingxuan Wang, et al. Pre-training multilingual neural machine translation by leveraging alignment information. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 2649–2663, 2020. doi: 10.18653/v1/2020.emnlp-main.210.

Xuebo Liu, Houtim Lai, Derek F. Wong, et al. Norm-based curriculum learning for neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 427–436, 2020a. doi: 10.18653/v1/2020.acl-main.41.

Yinhan Liu, Jiatao Gu, Naman Goyal, et al. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742, 2020b. doi: 10.1162/tacl_a_00343.

Hongyuan Lu, Haoyang Yang, Haoyun Huang, Dongdong Zhang, Wai Lam, and Furu Wei. Chain-of-dictionary prompting elicits translation in large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 958–976, 2024. doi: 10.18653/v1/2024.emnlp-main.55.

Tasnim Mohiuddin, Philipp Koehn, Vishrav Chaudhary, et al. Data selection curriculum for neural machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 1569–1582, 2022. doi: 10.18653/v1/2022.findings-emnlp.113.

Renhao Pei, Yihong Liu, Peiqin Lin, et al. Understanding in-context machine translation for low-resource languages: A case study on manchu. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8767–8788, 2025. doi: 10.18653/v1/2025.acl-long.429.

María Begoña Pendas, Andrés Carvallo, and Carlos Aspillaga. Neural machine translation through active learning on low-resource languages: The case of spanish to mapudungun. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pp. 6–11, 2023. doi: 10.18653/v1/2023.americasnlp-1.2.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 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)*, pp. 86–96, 2016. doi: 10.18653/v1/P16-1009.

Jean Seo, Sungjoo Byun, Minha Kang, et al. Mergen: The first manchu–korean machine translation model trained on augmented data. In *Proceedings of the 3rd Workshop on Multi-lingual Representation Learning*, pp. 118–124, 2023. doi: 10.18653/v1/2023.mrl-1.10.

Burr Settles. Active learning literature survey. Technical Report 1648, University of Wisconsin–Madison, 2009. URL `https://burrsettles.com/pub/settles.activelearning.pdf`.

Meiqi Wang, Xiaoxin Sun, Dongjie Wang, et al. ManCC: A task-anchored benchmark for manchu–classical chinese cross-lingual modeling. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 27271–27292, 2026. doi: 10.18653/v1/2026.findings-acl.1359.

Chen Xu, Bo Hu, Yujing Jiang, et al. Dynamic curriculum learning for low-resource neural machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 3977–3989, 2020. doi: 10.18653/v1/2020.coling-main.352.

Linting Xue, Noah Constant, Adam Roberts, et al. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 483–498, 2021. doi: 10.18653/v1/2021.naacl-main.41.

Linting Xue, Aditya Barua, Noah Constant, et al. ByT5: Towards a token-free future with pre-trained byte-to-byte models. *Transactions of the Association for Computational Linguistics*, 10: 291–306, 2022. doi: 10.1162/tacl_a_00461.

Yang Zhao, Haoran Zhang, Shuqiao Zhou, et al. Active learning approaches to enhancing neural machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 1796–1806, 2020. doi: 10.18653/v1/2020.findings-emnlp.162.

Francis Zheng, Edison Marrese-Taylor, and Yutaka Matsuo. Improving low-resource machine translation for formosan languages using bilingual lexical resources. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 11248–11259, 2024. doi: 10.18653/v1/2024.findings-acl.670.

Yikai Zhou, Baosong Yang, Derek F. Wong, Yu Wan, and Lidia S. Chao. Uncertainty-aware curriculum learning for neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 6934–6944, 2020. doi: 10.18653/v1/2020.acl-main.620.