

AkibaNLP-TUT: Improving Low-Resource Machine Translation via Language-Specific and Length-Adaptive Noise Injection

Takumi Kachi **Tomoyoshi Akiba**
Toyohashi University of Technology
{kachi.takumi.kd, akiba.tomoyoshi.tk}@tut.jp

Hajime Tsukada
Aichi Sangyo University
tsukada@asu.ac.jp

Abstract

We present our system for the WMT 2026 Low-Resource Indic Language Translation Shared Task, leveraging a language-specific word-level noise injection method. Unlike the previous approach, which required an external low-resource language (LRL) monolingual corpus, our method constructs the LRL vocabulary and word-frequency statistics directly from source-side LRL text available in the test set. The resulting vocabulary is used to introduce language-specific noise into the source side of a closely related high-resource language (HRL) parallel corpus. We further investigate two hyperparameters inherited from the previous method: the maximum edit distance K used for candidate word selection and the noise ratio. Specifically, we replace the fixed maximum edit-distance constraint with a dynamic constraint determined by the character length of each source word, and evaluate six noise ratios ranging from 5% to 30%. In a single-run exploratory analysis of Assamese→English translation, the proposed dynamic constraint obtains higher test-set BLEU than the fixed constraint ($K = 5$) at all evaluated noise ratios, with the highest observed test-set BLEU at a 15% noise ratio. When the noise ratio is selected on the development set, the proposed method selects 15%, compared with 20% for the fixed constraint, and achieves higher test-set BLEU under these respective selections. We additionally report our official submission results for the WMT 2026 Low-Resource Indic Language Translation Shared Task across three translation directions (Assamese, Manipuri, and Bodo into English), and discuss the conditions under which language-specific noise injection is effective.

1 Introduction

There are approximately 7,000 languages in the world, but only a small subset of high-resource languages (HRLs) have sufficiently developed parallel

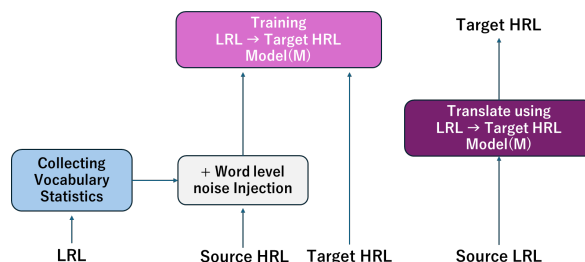


Figure 1: Overview of our proposed method: Language-Specific Word-Level Noise Injection.

corpora for machine translation (MT). For these languages, research leveraging multilingual pre-trained language models such as XLM-R (Conneau et al., 2020), mT5 (Xue et al., 2021), and No Language Left Behind (NLLB) (NLLB Team, 2022) has significantly advanced multilingual machine translation. Such efforts have enabled the learning of shared cross-lingual embedding spaces, thereby facilitating cross-lingual transfer.

In contrast, many languages are low-resource languages (LRLs), for which neither parallel nor sufficient monolingual resources are available. Although monolingual data is generally easier to collect than parallel data and has been widely utilized in techniques such as back-translation (BT) (Sennrich et al., 2016a) and continued pretraining of multilingual language models, many endangered or less-digitized languages still lack even large-scale monolingual corpora. This limitation motivates the development of methods that can improve LRL translation using only limited textual resources available during translation tasks.

Hamada et al. (2025) proposed a language-specific word-level noise injection method that utilizes lexical information from an LRL monolingual corpus. Instead of applying random character substitutions, the method replaces words in a closely related HRL parallel corpus with LRL words selected according to edit distance and word

frequency, thereby creating language-specific noisy training data. Experimental results demonstrated that this approach outperformed both conventional training without noise injection and character-level augmentation methods for several low-resource Indic language translation tasks.

Although the previous method achieved promising results, we found that the use of word-frequency information extracted from an LRL monolingual corpus is not always essential for effective noise generation. In particular, when the vocabulary is constructed from only the source-side sentences of the target LRL data, language-specific noise can still be generated without requiring an additional monolingual corpus. This observation suggests that language-specific word-level noise injection can be applied in more constrained low-resource scenarios where only a small amount of target-language text is available.

Furthermore, two important hyperparameters of the previous method were selected empirically. First, the maximum edit distance K used for candidate word selection was fixed to 5 regardless of the length of the source word. While this constraint prevents excessively different replacements, it imposes the same maximum perturbation magnitude on words of substantially different lengths. As a result, it may be overly restrictive for long words, for which larger edit distances can still represent moderate relative perturbations. Second, the noise ratio was fixed at 10%, and the influence of different noise levels was not investigated. Consequently, it remains unclear whether these parameter settings are optimal for language-specific word-level noise injection.

In this paper, we investigate these two limitations through a resource-efficient extension of language-specific word-level noise injection. We construct the LRL vocabulary and frequency statistics from source-side LRL text available in the translation task, rather than requiring an external monolingual corpus. We then replace the fixed maximum edit distance with a dynamic constraint determined by source-word character length and evaluate six noise ratios from 5% to 30%. We analyze these factors on Assamese→English translation and subsequently report the official WMT 2026 shared-task results.

2 Related Work

Several studies have investigated the effects of introducing stochastic noise into neural machine

translation training to improve model robustness and generalization. Representative approaches include token masking and random word replacement. Wang et al. (2018) introduced SwitchOut, a data augmentation method that randomly replaces words in parallel corpora to increase the diversity of training data. Although these approaches improve data diversity and model robustness, they do not explicitly exploit the linguistic characteristics of low-resource languages (LRLs).

Back-Translation (BT) (Sennrich et al., 2016a) is one of the most widely used approaches for leveraging monolingual corpora in neural machine translation. BT generates pseudo-parallel data by translating target-side monolingual sentences into the source language, enabling the effective utilization of monolingual resources when parallel corpora are limited. Several extensions, including Iterative Back-Translation (Morita et al., 2018; Hoang et al., 2018; Zhang et al., 2018), have further improved translation quality by repeatedly generating pseudo-parallel data and updating translation models.

For extremely low-resource languages, CharSpan (Maurya et al., 2024) proposed a character-level noise injection method that augments the parallel corpus of a closely related high-resource language by exploiting lexical similarity. By introducing character-level perturbations, CharSpan improves cross-lingual transfer without requiring parallel data for the target LRL. However, the method does not utilize lexical frequency information obtained from monolingual corpora and therefore cannot model language-specific vocabulary distributions.

Hamada et al. (2025) addressed this limitation by proposing a language-specific word-level noise injection method that selects replacement words according to edit distance and word frequency extracted from an LRL monolingual corpus. Experimental results demonstrated that the proposed method outperformed both conventional training without noise injection and CharSpan on several low-resource Indic language translation tasks. However, the maximum edit distance and the noise ratio were determined empirically, and the influence of these hyperparameters on translation performance was not investigated.

In this work, we focus on these two hyperparameters and perform a systematic analysis of their impact on translation performance. Specifically, we investigate a dynamic maximum edit-distance constraint based on word length and evaluate multi-

HRL(Bn):	এমন কথা কিন্তু তিনি আপনাকে কোনওদিনও বলেননি ।
Eng:	But he never told me.
HRL(Bn) + Noise:	এমন কথা সিদ্ধু তিনি আপনাকে কোনোদিনে বলেননি ।

Figure 2: Example of word-level noise injection for Bengali (HRL). The original Bengali sentence and its English translation are shown at the top. In the noisy version (bottom), Bengali words are replaced with words selected from an Assamese vocabulary list.

ple noise ratios to identify more effective parameter settings for language-specific word-level noise injection.

3 Proposed Method

In this section, we describe a language-specific word-level noise injection method for improving LRL→En translation by augmenting the parallel corpus of a closely related high-resource language (HRL). The proposed method follows Hamada et al. (2025), in which language-specific word-level noise is generated by replacing words in the HRL corpus with words from the source-side LRL. The replacement process is based on edit distance and frequency-weighted candidate selection. In this study, we investigate the effects of different maximum edit-distance constraints and noise ratios, while all other components of the noise injection procedure remain identical to those of the previous method.

Figure 1 presents an overview of the proposed framework, and Figure 2 illustrates an example of the noise injection process.

The following sections describe the language-specific word replacement procedure and the modifications introduced in this study for determining the maximum edit distance and the noise ratio.

3.1 Language-Specific Word Replacement

The language-specific word replacement procedure follows Hamada et al. (2025). Since the objective of this study is to investigate the effects of the maximum edit-distance constraint and the noise ratio, the candidate selection and frequency-weighted replacement mechanism remain unchanged from Hamada et al. (2025); only the maximum edit-distance constraint and the noise ratio are modified. The modifications introduced in this work are described in Sections 3.3 and 3.4.

First, given a source sentence x , a word x_i is randomly selected for replacement. Next, an edit

distance d is sampled according to the geometric distribution shown in Equation 1. Following Hamada et al. (2025), we fix the success probability to $p = 0.5$, while the upper bound K of the sampled edit distance (denoted as $d_{\max}$ in the previous work) is set to either 5 (previous method) or the source-word length (proposed method).

$$d \sim P(d) \propto p(1-p)^{d-1}, \quad (d = 1, 2, \dots, K) \quad (1)$$

The parameters K and p determine the distribution of the noise magnitude. Based on the sampled edit distance, a candidate set $V(d, x_i)$ is constructed by extracting all words in the LRL vocabulary V_{LRL} whose Levenshtein distance from x_i equals d :

$$V(d, x_i) = \{w \mid w \in V_{\text{LRL}}, ED(w, x_i) = d\} \quad (2)$$

where $ED(\cdot, \cdot)$ denotes the Levenshtein edit distance.

After sampling the edit distance d , the replacement word w' is sampled from the candidate set $V(d, x_i)$ according to the relative frequency of each candidate word:

$$w' \sim P(w' \mid d, x_i) = \frac{f(w')}{\sum_{u \in V(d, x_i)} f(u)} \quad (3)$$

where $f(w')$ denotes the frequency of candidate word w' calculated from the Assamese source-side text used for constructing the vocabulary.

If the sampled edit distance yields an empty candidate set, another edit distance is sampled and the candidate search is retried, up to 10 attempts for each source word. If no valid replacement candidate is found within these attempts, the original word is retained and no characters are counted toward the injected noise.

To quantify how frequently empty candidate sets and unsuccessful replacements occur, we additionally analyzed the replacement process on one training split. Table 2 reports the proportion of candidate-search attempts that yielded an empty candidate set and the proportion of replacement operations that ultimately failed after up to ten attempts. Empty candidate sets occurred frequently, accounting for 64.36% of candidate-search attempts under $K = 5$ and 63.68% under $K = |x_i|$. However, repeated sampling of the edit distance

Corpora	Lang.	Usage	# Sent.
Samanantar	En-Bn	Train	4,302,288
Official WMT 2026 Shared Task training data	En-As	Train	53,000
		Dev	1,000
WMT 2025 Shared Task test set	En-As	Test	2,000
	As	Noise Inj.	2,000

Table 1: Overview of the datasets used in the hyperparameter analysis (En = English, Bn = Bengali, As = Assamese). The datasets used for the official WMT 2026 shared task submissions are summarized separately in Section 5.4

Statistic	$K = 5$	$K = x_i $
Empty-candidate rate (%)	64.36	63.68
Final failure rate (%)	7.90	7.26

Table 2: Replacement failure statistics under the fixed and dynamic maximum edit-distance constraints on one training split. The empty-candidate rate is computed over candidate-search attempts, whereas the final failure rate is computed over replacement operations that either succeed or fail after up to ten attempts.

allowed most replacement operations to eventually find a valid candidate. The final failure rate was 7.90% for $K = 5$ and 7.26% for $K = |x_i|$.

The procedure is repeated while valid word replacements can be found and the remaining character budget allows another replacement. Because replacements are performed at the word level and are subject to the remaining noise budget, the realized noise ratio may be slightly lower than the specified ratio.

3.2 Utilizing Source-Side Test Data

In Hamada et al. (2025), the LRL vocabulary V_{LRL} and the word frequency $f(w)$ used in Equation 3 were computed from a dedicated LRL monolingual corpus collected independently of the parallel training and test data. In this work, we instead construct V_{LRL} and $f(w)$ directly from the Assamese source-side sentences of the publicly available WMT 2025 Low-Resource Indic Language Translation Shared Task test set (Pakray et al., 2025), consisting of approximately 2,000 sentences, without relying on any additional monolingual corpus. This allows language-specific word-level noise injection to be applied even in more constrained low-resource scenarios where the only available LRL text is the small amount distributed with the shared task itself.

3.3 Dynamic Maximum Edit Distance

In the previous method, the maximum edit distance K in Equation 1 was fixed to 5 for all source words. This imposes the same upper bound on words of

different lengths. For long words, in particular, the fixed bound limits the maximum possible perturbation magnitude even when larger edit distances may still represent relatively moderate changes to the word.

To account for this difference, we dynamically determine the maximum edit distance according to the character length of the selected source word. Specifically, for a source word x_i , the maximum edit distance is defined as

$$K = |x_i| \quad (4)$$

where $|x_i|$ denotes the number of characters in the source word.

The edit distance is sampled from the same geometric distribution as in Equation 1, but with this word-specific upper bound. Importantly, the proposed method does not modify the underlying geometric sampling mechanism and therefore does not directly increase the probability of selecting a large edit distance for longer words. Instead, it changes only the maximum admissible distance: short words cannot be assigned an edit distance greater than their character length, while long words can admit larger edit distances. The purpose of this modification is therefore to make the maximum perturbation magnitude dependent on source-word length rather than to explicitly favor longer-word replacement. To verify how the dynamic constraint affects the realized edit distances, we further analyzed the successful word replacements in one training split. Table 3 summarizes the average realized edit distance and the proportion of replacements with $d > 5$ for different original word-length ranges.

Under the proposed constraint, the average realized edit distance increases with the length of the original word, from 1.44 for words of length 1–3 to 3.53 for words of length 11 or more. More importantly, replacements with $d > 5$ account for 1.96% of all successful replacements under $K = |x_i|$, whereas such replacements are impossible under the fixed $K = 5$ constraint. The proportion of re-

Original Word Length	Avg. d	$d > 5$ (%)
1–3	1.44	0.00
4–6	2.08	0.82
7–10	2.88	7.01
11+	3.53	20.91

Table 3: Realized edit-distance statistics for successful replacements under the proposed dynamic constraint ($K = |x_i|$), grouped by the character length of the original word. The analysis uses one training split.

placements with $d > 5$ also increases with word length, from 0.82% for words of length 4–6 to 7.01% for words of length 7–10 and 20.91% for words of length 11 or more. In contrast, under the fixed $K = 5$ setting, the average edit distance for words of length 11 or more is 2.90, compared with 3.53 under $K = |x_i|$. These results show that the dynamic constraint allows larger edit distances to be realized for longer words, while retaining the same underlying geometric sampling mechanism.

3.4 Noise Ratio

In the previous method, the Noise Ratio was fixed at 10% for all experiments. While this setting was shown to be effective, the influence of different noise ratios on translation performance was not investigated.

To analyze the effect of the amount of injected noise, we construct multiple noise-injected training corpora with different noise ratios. Specifically, we evaluate six noise ratios: 5%, 10%, 15%, 20%, 25%, and 30%.

The noise ratio specifies the target proportion of source-side characters to be modified in each sentence. Specifically, for a sentence containing N source-side characters and a noise ratio r , the replacement process sets an upper budget of $\lfloor Nr \rfloor$ changed characters. A word replacement with edit distance d is accepted only when adding d changed characters does not exceed this budget. Therefore, the realized noise ratio may be slightly lower than the specified value because replacements are performed at the word level.

By evaluating multiple noise ratios under the same training conditions, we characterize how the amount of injected language-specific noise affects translation performance.

4 Experimental Setup

4.1 Datasets

Table 1 summarizes the datasets used in this study. For the main training corpus, we use the

Bengali–English portion of the Samanantar parallel corpus (Ramesh et al., 2022), which contains 4,302,288 parallel sentence pairs. Since Bengali is a high-resource language closely related to Assamese, this corpus serves as the source parallel data for language-specific noise injection.

For the development set, we use the last 1,000 sentence pairs of the WMT 2026 Low-Resource Indic Language Translation Shared Task Assamese–English training data. Translation performance for the hyperparameter analysis is evaluated on the publicly available WMT 2025 Assamese–English test set (Pakray et al., 2025), consisting of 2,000 sentence pairs.

For the hyperparameter analysis in Sections 5.1 and 5.2, the Assamese source side of the WMT 2025 test set is used solely as monolingual text for constructing the LRL vocabulary and word-frequency statistics used in noise injection. The English references are used only for evaluation and are never used for training, noise construction, or model selection. Thus, the hyperparameter analysis explicitly uses source-side test text as an available monolingual resource, while the target-side references remain unused except for evaluation.

Separately, to investigate the effect of genuine LRL parallel data, Section 5.3 additionally uses the official WMT 2026 Assamese–English training data. Of the 54,000 available sentence pairs, the first 53,000 are used for training and the last 1,000 are reserved for development. These parallel data are incorporated into the training corpus for the additional parallel-data experiments and are not used as the monolingual vocabulary source in the main hyperparameter analysis.

The official WMT 2026 submission experiments use the task-provided training and test data described in Section 5.4; their data usage therefore differs from the WMT 2025-based hyperparameter analysis.

4.2 Data Preprocessing

The preprocessing pipeline is identical to that used in Hamada et al. (2025). For the English data, we first applied Unicode normalization (NFKC), followed by tokenization using Sacremoses. We then applied truecasing to reduce case variation at the beginning of sentences and in proper nouns. Finally, we learned and applied Byte Pair Encoding (BPE) (Sennrich et al., 2016b; Gage, 1994) using subword-nmt with 16,000 merge operations.

For the Bengali and Assamese data, we applied

the same preprocessing steps as for English, including Unicode normalization (NFKC), tokenization with Sacremoses, and BPE with 16,000 merge operations. Since these languages do not distinguish between uppercase and lowercase characters, truecasing was not applied.

4.3 Experimental Settings

All models were trained using the Transformer architecture implemented in Fairseq (Ott et al., 2019). Both the encoder and decoder consisted of six layers with eight attention heads. Optimization was performed using Adam with $\beta_1 = 0.9$ and $\beta_2 = 0.98$, and the initial learning rate was set to 0.0005 with the inverse square-root learning rate scheduler. The maximum number of tokens per batch was set to 4,000, and the learning rate was warmed up for 4,000 updates.

Label-smoothed cross-entropy with a smoothing factor of 0.1 was used as the training criterion. The dropout rate and attention dropout rate were both set to 0.1, and weight decay was set to 0.0001. Gradient clipping was disabled during training. Layer normalization was applied before each encoder and decoder block following the pre-layer normalization Transformer configuration. Early stopping was applied with a patience of five validation evaluations.

Unless otherwise stated, all experimental settings were kept identical across different edit-distance constraints and noise ratios.

4.4 Evaluation Metrics

Translation quality was evaluated using three automatic evaluation metrics: BLEU (Papineni et al., 2002), chrF++ (Popović, 2017), and Translation Edit Rate (TER) (Snover et al., 2006).

BLEU was computed using the fairseq-score toolkit with the reference translations provided in the WMT 2025 Low-Resource Indic Language Translation Shared Task test set.

We also report chrF++ and TER. chrF++ evaluates translation quality based on both character- and word-level n-gram overlap and is particularly suitable for morphologically rich languages, while TER measures the minimum number of edit operations required to transform the system output into the reference translation. These complementary metrics provide additional perspectives on translation quality beyond BLEU.

5 Results

5.1 Comparison with Previous Methods

Table 4 compares the proposed method with the baseline model without language-specific noise injection and the previous language-specific word-level noise injection method. Unlike the previous comparison at a fixed noise ratio, the noise ratio for each method is first selected using the development set and then evaluated on the test set.

On the development set, the previous method achieved its highest BLEU score (8.81) at a noise ratio of 20%, whereas the proposed method achieved its best BLEU score (8.66) at 15%. These selected configurations were subsequently evaluated on the test set.

Although the previous method achieved a slightly higher BLEU score on the development set, its selected configuration obtained only 6.71 BLEU on the test set. In contrast, the proposed method achieved 7.65 BLEU on the test set using the noise ratio selected on the development set. This single-run result shows that the development-selected configuration of the proposed method achieves higher test-set BLEU than the corresponding development-selected configuration of the previous method (7.65 vs. 6.71). However, whether this difference reflects more robust generalization cannot be established from a single run.

Compared with the baseline model without noise injection (2.68 BLEU), both methods substantially improve translation quality, while the proposed method achieves the highest test-set performance after development-set-based model selection.

5.2 Effect of the Noise Ratio

Table 5 summarizes the effect of different noise ratios on BLEU for both the previous method and the proposed method. Note that this table reports both development- and test-set scores for all six noise ratios purely for exploratory analysis; the noise ratio actually used for model selection is chosen on the development set alone, as summarized in Table 4 and Section 5.1. Each value corresponds to a single training run, as noise injection and model training involve stochastic components (see Limitations).

For the proposed method, translation performance improves from 5% to 15%, where the highest observed test-set BLEU of 7.65 is obtained. At noise ratios above 15%, the observed test-set BLEU scores decrease, reaching 6.68 at 30%. Because each configuration corresponds to a single

Model	Dev BLEU	Test BLEU
w/o noise	3.04	2.68
Previous ($K = 5$, selected on dev: 20%)	8.81	6.71
Proposed (dynamic K , selected on dev: 15%)	8.66	7.65

Table 4: Comparison using the best noise ratio selected on the development set for each method. The previous method selected 20% on the development set, whereas the proposed method selected 15%. Test BLEU is reported using the corresponding selected configuration.

Noise Ratio	$K = 5$		$K = x_i $	
	Dev	Test	Dev	Test
5%	8.60	7.10	7.93	7.42
10%	8.20	7.38	8.48	7.51
15%	8.56	7.08	8.66	7.65
20%	8.81	6.71	8.04	7.46
25%	7.86	6.17	8.59	6.92
30%	8.23	6.44	8.16	6.68

Table 5: Development- and test-set BLEU scores across different noise ratios for the previous method (fixed $K = 5$) and the proposed method (dynamic $K = |x_i|$). Bold configurations (20% for $K = 5$, 15% for $K = |x_i|$) correspond to the dev-selected settings summarized in Table 4.

training run, however, these observations should be interpreted as an empirical trend in this experiment rather than evidence of a monotonic relationship between noise ratio and translation quality.

For the previous method, the best observed test-set BLEU score is obtained at 10% (7.38), while the scores at higher noise ratios fluctuate rather than following a monotonic pattern. In particular, BLEU increases from 6.17 at 25% to 6.44 at 30%. Since each configuration was evaluated with a single training run, this variation may reflect stochastic training and noise-generation effects, and we do not interpret it as evidence of a systematic relationship between the fixed edit-distance constraint and noise ratio.

The proposed method obtains higher test-set BLEU than the previous method at all six evaluated noise ratios. The largest difference is observed at 15%, where BLEU increases from 7.08 to 7.65. The proposed method achieves its highest observed test-set BLEU at 15%, while the previous method achieves its highest observed test-set BLEU at 10% (7.38).

These results suggest that the maximum edit-distance constraint and noise ratio can affect translation performance in this experimental setting. However, because each configuration corresponds to a single training run, the observed differences and trends should be regarded as exploratory rather than conclusive.

5.3 Effect of Additional Parallel Data

The hyperparameter analysis in Sections 5.1 and 5.2 does not incorporate the official WMT 2026 Assamese–English parallel training data into model training. Instead, the Assamese source side of the WMT 2025 test set is used only as monolingual text for constructing the LRL vocabulary and word-frequency statistics for noise injection. To examine how the proposed method behaves when genuine LRL parallel training data is available, we additionally incorporate the official WMT 2026 Assamese–English training data into the training corpus.

The official WMT 2026 Assamese–English data contains 54,000 sentence pairs, of which the first 53,000 are used for training and the last 1,000 are reserved for development. For the proposed method, the noise-injected Bengali–English corpus is concatenated with the 53,000 Assamese–English training sentence pairs, and the resulting corpus is used to train the model from scratch. The no-noise baseline is trained analogously by concatenating the original Bengali–English corpus with the same 53,000 Assamese–English training sentence pairs. Thus, the “+ 2026 train” configurations in Table 6 do not represent fine-tuning or continued training; they denote training from scratch on the concatenated corpus.

To provide an additional baseline, we also train a model using only the 53,000 WMT 2026 Assamese–English training sentence pairs, with the remaining 1,000 sentence pairs used as the development set. This model does not use the Bengali–English Samanantar corpus or language-specific noise injection. All models in this experiment are evaluated on the same 2,000-sentence WMT 2025 Assamese–English test set used in the preceding experiments.

This experiment was conducted with a fixed noise ratio of 10%, prior to the systematic noise-ratio analysis presented in Sections 5.1 and 5.2. Consequently, 10% may not be the optimal ratio when additional parallel data is available, and we did not revisit this choice after the later analysis.

Model	BLEU
w/o noise(Bengali)	2.68
Proposed (10%)	7.51
WMT 2026 train only(Assamese)	0.78
w/o noise + 2026 train	15.12
Proposed (10%) + 2026 train	14.96

Table 6: Comparison of models trained with the Bengali–English base corpus, the WMT 2026 Assamese–English parallel training data alone, or their combination. The noise-injected configurations use the proposed method with a fixed 10% noise ratio. The “+ 2026 train” configurations are trained from scratch on the concatenation of the Bengali–English corpus and the WMT 2026 Assamese–English training data.

We discuss this limitation further below.

Table 6 shows that training only on the 53,000 WMT 2026 Assamese–English parallel training sentence pairs results in a BLEU score of 0.78 on the WMT 2025 test set. In contrast, adding the same parallel data to the Bengali–English base corpus substantially improves translation performance, raising BLEU from 2.68 to 15.12 for the no-noise baseline and from 7.51 to 14.96 for the proposed method.

Notably, once the WMT 2026 parallel training data is added to the Bengali–English base corpus, the proposed noise-injection method does not provide an additional gain under the fixed 10% noise ratio used in this experiment. The proposed method achieves 14.96 BLEU, slightly below the corresponding no-noise baseline at 15.12 BLEU. Thus, in this particular training configuration, the benefit obtained from adding genuine Assamese–English parallel data is substantially larger than the additional effect of noise injection.

However, these results should not be interpreted as evidence that noise injection is ineffective when parallel data is available. The experiment evaluates only a single noise ratio and a single training configuration, and the 10% noise ratio was selected before the systematic analysis in Sections 5.1 and 5.2. We therefore leave a systematic investigation of the interaction between parallel-data quantity, training strategy, and noise ratio for future work.

5.4 Task Results

According to the official evaluation results released by the shared task organizers (Pakray et al., 2026), we submitted systems for three translation directions: Assamese→English (one primary and two contrastive systems), Bodo→English (one primary and one contrastive system), and

Manipuri→English (one primary system). For Assamese and Manipuri, the Bengali–English portion of the Samanantar corpus served as the base HRL parallel data, with the last 1,000 sentence pairs of the respective WMT 2026 shared-task training data used as the development set; for Bodo, the Hindi–English portion of Samanantar was used as the base corpus in the same manner. All submitted systems employed the proposed dynamic maximum edit-distance constraint ($K = |x_i|$) together with a noise ratio of 10%. However, the systematic comparison of different noise ratios presented in Sections 5.1 and 5.2 was conducted after the shared-task submission. Therefore, the submitted systems represent only one configuration of the proposed method evaluated in this paper.

For Assamese→English, the official evaluation (Pakray et al., 2026) shows that our primary system, which used the source-side sentences of the official WMT 2026 test set as a monolingual corpus for language-specific noise injection together with the official parallel training data, achieved a BLEU score of 11.01 and a chrF++ score of 39.18. The contrastive2 system, which used the source side of the official WMT 2026 training data as the monolingual corpus for noise injection instead of the test-set source sentences, performed comparably (BLEU 10.47, chrF++ 39.81), suggesting that the source of the monolingual text used for noise generation had only a small effect once official parallel training data was also included. In contrast, the contrastive1 system, which used only the source-side sentences of the official WMT 2026 test set as the monolingual corpus for noise injection without incorporating the official parallel training data, scored substantially lower (BLEU 6.06, chrF++ 31.98, TER 122.98), indicating that the addition of official parallel training data played a larger role in overall performance than the specific choice of noise source.

A similar pattern was observed for Bodo→English: the primary system, which likewise used the source-side sentences of the official WMT 2026 test set as a monolingual corpus for language-specific noise injection together with the official parallel training data, achieved a BLEU score of 12.19, whereas the contrastive1 system without the official training data scored markedly lower (BLEU 1.41, TER 135.09). This is consistent with the Assamese results and further suggests that, for both directions, the presence of official parallel training data was the important

Corpora	Language	Usage	# Sentences
Samanantar	Bn2En	Train	4,302,288
	Hi2En	Train	5,062,852
Official WMT 2026 training data	As2En	Train / Noise Injection	53,000
		Dev	1,000
	Mni2En	Train / Noise Injection	22,687
		Dev	1,000
	Bodo2En	Train	14,215
		Dev	1,000
Official WMT 2026 Shared Task test set	Assamese	Noise Injection	1,000
	Bodo	Noise Injection	1,000

Table 7: Overview of the corpora used for our WMT 2026 IndicMT shared task submissions, including the base HRL parallel corpora (Samanantar), the official WMT 2026 training/validation data for each LRL direction, and the WMT 2026 test-set source sentences used for noise generation.

Direction	System	BLEU	chrF++	TER↓
As→En	primary	11.01	39.18	99.69
	contrastive1	6.06	31.98	122.98
	contrastive2	10.47	39.81	99.71
Bodo→En	primary	12.19	40.30	90.33
	contrastive1	1.41	17.87	135.09
Mni→En	primary	0.50	18.09	292.36

Table 8: Official evaluation results of our submitted systems (Akiba-NLP-TUT) on the WMT 2026 IndicMT Shared Task test sets, as released by the task organizers.

factor in system performance at submission time.

For Manipuri→English, our primary system, which used the source side of the official WMT 2026 Manipuri training data as the monolingual corpus for language-specific noise injection, obtained a BLEU score of only 0.50 and a TER of 292.36, indicating a substantial breakdown in output quality. Given the typological distance between Manipuri (Sino-Tibetan) and Bengali discussed in the Limitations section, we suspect that the noise-injection procedure may have been less effective at producing useful training signal for this language pair, though we did not conduct a separate diagnostic experiment to confirm this at submission time.

Overall, the submitted systems correspond to a single configuration of the proposed method, employing the dynamic maximum edit-distance constraint together with a noise ratio of 10%. The broader comparison across different noise ratios presented in Sections 5.1 and 5.2 was conducted after the shared-task submission and therefore serves as a subsequent analysis of the proposed method.

6 Conclusion

In this paper, we extended the language-specific word-level noise injection method proposed by Hamada et al. (2025) by eliminating the need for

an external low-resource language (LRL) monolingual corpus and by investigating two important hyperparameters: the maximum edit distance K used for candidate word selection and the noise ratio. Specifically, we construct the LRL vocabulary and word-frequency statistics directly from the source-side LRL text available in the translation task, propose a dynamic maximum edit-distance constraint that adapts to the character length of each source word, and evaluate translation performance across six noise ratios.

Experimental results on Assamese→English translation showed that the proposed dynamic edit-distance constraint achieved higher test-set BLEU than the fixed constraint at all six evaluated noise ratios in our single-run analysis. The proposed method achieved its highest observed test-set BLEU at a 15% noise ratio, while the fixed constraint achieved its highest observed test-set BLEU at 10%. When noise ratios were selected using the development set, the proposed method selected 15% and achieved 7.65 BLEU on the test set, whereas the fixed constraint selected 20% and achieved 6.71 BLEU. These results suggest that the maximum edit-distance constraint and noise ratio can affect translation performance in this noise-injection framework, while the single-run design prevents us from drawing firm conclusions about statistical significance or generalization across runs.

When the 53,000 WMT 2026 Assamese–English training sentence pairs were added to the Bengali–English base corpus, BLEU increased substantially for both the no-noise and proposed configurations, while the proposed method with a fixed 10% noise ratio did not outperform the corresponding no-noise baseline. This result highlights the need to further investigate how noise injection interacts

with the amount of genuine LRL parallel data.

We additionally reported our official submission results for the WMT 2026 Low-Resource Indic Language Translation Shared Task, covering the Assamese→English, Manipuri→English, and Bodo→English directions. The official results varied substantially across language pairs: the proposed system achieved substantially higher BLEU for Assamese and Bodo than for Manipuri. For Assamese, the comparison between the primary and contrastive2 systems also showed that using different source-side text for noise generation produced relatively similar performance when the official parallel training data was included. These results suggest that the effectiveness of language-specific noise injection may depend on the relationship between the LRL and the HRL used for noise generation.

Future work includes verifying the proposed hyperparameter settings on additional LRL-HRL pairs with varying degrees of similarity, evaluating the method under multiple random seeds to assess the stability of the observed trends, and extending the analysis to translation directions beyond LRL→English.

Limitations

First, both the proposed and previous noise-injection methods involve inherent randomness: the edit distance sampled from the geometric distribution and the selection of replacement words both vary across runs, so a given noise ratio does not deterministically yield a fixed level of perturbation. In this study, each configuration was evaluated with a single training run due to computational constraints. As a result, some of the fluctuations observed across noise ratios and between the two methods—such as the non-monotonic behavior of the fixed-distance method between 25% and 30%—may partly reflect run-to-run variance rather than a stable underlying trend. Repeating each configuration with multiple random seeds and reporting averaged scores would allow for a more reliable characterization of the effect of noise ratio and edit-distance strategy on translation performance.

Second, the effectiveness of the proposed method appears to depend strongly on the lexical and typological similarity between the target LRL and the HRL used for noise injection. In this study, Bengali served as the HRL for both Assamese (Indo-Aryan) and Manipuri (Sino-Tibetan). While

the method yielded consistent improvements for Assamese, gains for Manipuri were considerably smaller both in this work and in our team’s prior submission (Hamada et al., 2025), and our WMT 2026 primary submission for Manipuri→English underperformed substantially relative to other participating systems. This suggests that the current approach is best suited to LRL-HRL pairs with a reasonable degree of lexical closeness, and that selecting an appropriately related HRL is a prerequisite for the method to be effective, rather than a detail that can be addressed after the fact through hyperparameter tuning alone.

Third, the hyperparameter analysis is centered on a single LRL–English direction, Assamese→English, and therefore does not establish that the observed preference for the dynamic constraint or a particular noise ratio transfers to other language pairs. In addition, we do not compare the proposed length-based constraint with alternative length-dependent schedules, so the results do not establish that $K = |x_i|$ is uniquely preferable among possible dynamic constraints.

Fourth, the effect of the noise ratio was systematically evaluated only in the setting without the additional WMT 2026 Assamese–English parallel training data. The experiment with approximately 53,000 parallel sentence pairs used a fixed 10% noise ratio and trained the model from scratch on the concatenation of the noise-injected Bengali–English corpus and the Assamese–English training data. Thus, the results do not determine whether a different noise ratio or an alternative training strategy, such as sequential fine-tuning, would make noise injection beneficial when genuine LRL parallel data is available.

Fifth, although the proposed method constructs its vocabulary and frequency statistics from source-side LRL text available in the translation task, the main hyperparameter analysis does not provide a fully controlled comparison against an independently collected LRL monolingual corpus using otherwise identical settings. The official Assamese submission provides a comparison between different source texts for noise generation, but that comparison also changes the overall shared-task training setting. A controlled vocabulary-source ablation would therefore be useful for isolating this factor.

Finally, we do not compare our relatively small Fairseq Transformer models with contemporary large multilingual or Indic-focused language mod-

els. Such a comparison would be informative for contextualizing the absolute translation quality, but differences in model scale, pretraining data, finetuning data, and evaluation settings would make a direct comparison difficult to interpret. We therefore leave a controlled comparison with contemporary multilingual models for future work.

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 ACL*, pages 8440–8451. Association for Computational Linguistics.
- Philip Gage. 1994. A new algorithm for data compression. *C Users Journal*, 12(2):23–38.
- Shoki Hamada, Tomoyosi Akiba, and Hajime Tsukada. 2025. [AkibaNLP-TUT: Injecting language-specific word-level noise for low-resource language translation](#). In *Proceedings of the Tenth Conference on Machine Translation*, pages 1259–1264, Suzhou, China. Association for Computational Linguistics.
- Vu Cong Duy Hoang, Philipp Koehn, Gholamreza Haffari, and Trevor Cohn. 2018. Iterative back-translation for neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 18–24.
- Kaushal Maurya, Rahul Kejriwal, Maunendra Desarkar, and Anoop Kunchukuttan. 2024. CharSpan: Utilizing lexical similarity to enable zero-shot machine translation for extremely low-resource languages. In *Proceedings of EACL 2024*, pages 294–310.
- Tomohiro Morita, Tomoyosi Akiba, and Hajime Tsukada. 2018. A study on unsupervised adaptation of neural machine translation with bidirectional back-translation. In *IPSJ SIG Technical Report*.
- NLLB Team. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.
- 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, Minneapolis, Minnesota. Association for Computational Linguistics.
- Partha Pakray, Reddi Mohana Krishna, Santanu Pal, Advaita Vetagiri, Sandeep Kumar Dash, Arnab Kumar Maji, Saralin A. Lyngdoh, Lenin Laitonjam, Anupam Jamatia, Koj Sambyo, Ajit Das, and Riyanka Manna. 2025. Findings of WMT 2025 shared task on low-resource Indic languages translation. In *Proceedings of the Tenth Conference on Machine Translation*, pages 532–553. Association for Computational Linguistics.
- Partha Pakray, Santanu Pal, Advaita Vetagiri, Kshetrimayum Boynao Singh, Sandeep Kumar Dash, Arnab Kumar Maji, Saralin A. Lyngdoh, Lenin Laitonjam, Anupam Jamatia, Ajit Das, Sunita Warjri, Uzzal Sharma, Koj Sambyo, and Riyanka Manna. 2026. Findings of WMT 2026 shared task on low-resource Indic languages translation. In *Proceedings of the Eleventh Workshop on Machine Translation*.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of ACL*, pages 311–318.
- Maja Popović. 2017. chrF++: Words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618.
- Gowtham Ramesh, Sumanth Doddapaneni, Aravindh Bheemaraj, Mayank Jobanputra, Vivek Raghavan, et al. 2022. Samanantar: The largest publicly available parallel corpora collection for 11 Indic languages. *Transactions of the Association for Computational Linguistics*, 10:145–162.
- 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*, pages 86–96.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 1715–1725.
- Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of AMTA*, pages 223–231.
- Xinyi Wang, Hieu Pham, Zihang Dai, and Graham Neubig. 2018. SwitchOut: An efficient data augmentation algorithm for neural machine translation. In *Proceedings of EMNLP*, pages 856–861.
- Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of NAACL-HLT*, pages 483–498. Association for Computational Linguistics.
- Zhirui Zhang, Shujie Liu, Mu Li, Ming Zhou, and En-hong Chen. 2018. Joint training for neural machine translation models with monolingual data. In *Proceedings of the AAAI Conference on Artificial Intelligence*.