Length-aware Byte Pair Encoding for Mitigating Over-segmentation in Korean Machine Translation

Jungseob Lee[1*], Hyeonseok Moon[1*], Seungjun Lee[1], Chanjun Park[2†],
Sugyeong Eo[1], Hyunwoong Ko[3], Jaehyung Seo[1], Seungyoon Lee[1], Heuiseok Lim[1†]
[1]Korea University, [2]Upstage AI, [3]Kakao Brain
{omanma1928, glee889, dzzy6505, djtnrud,
seojae777, dltmddbs100, limhseok}@korea.ac.kr[1]
chanjun.park@upstage.ai[2], kevin.brain@kakaobrain.com[3]

Abstract

Byte Pair Encoding is an effective approach in machine translation across several languages. However, our analysis indicates that BPE is prone to over-segmentation in the morphologically rich language, Korean, which can erode word semantics and lead to semantic confusion during training. This semantic confusion, stemming from over-segmentation, ultimately contributes to a degradation of overall translation quality. To address this issue, we introduce Length-aware Subword Vocabulary Construction (LeVoC), a novel approach strategically incorporating longer words into the vocabulary. By utilizing an external monolingual Korean corpus, LeVoC extracts and integrates long words, effectively preserving morphological information and reducing semantic confusion. Our experiments demonstrate that LeVoC not only significantly outperforms BPE, but also can be applied to and surpass current state-of-the-art morpheme-aware subword tokenization methods. We provide evidence that the difficulty in translating sentences with long words in Korean is associated with morphological compositionality, and LeVoC's ability to reduce semantic confusion during training leads to improved translation quality.

1 Introduction

The utilization of subword-unit segmentation methods has become a prevailing trend in neural machine translation (NMT), with widespread acceptance among researchers (Farhad et al., 2021; Barrault et al., 2020). Among the various methods available, Byte Pair Encoding (BPE) (Sennrich et al., 2016) has emerged as a popular choice in numerous NMT studies due to its language-agnostic applicability and its ability to strike a balance between compression efficiency and computational complexity (Vaswani et al., 2017; Ng et al., 2019; Zhou et al., 2022; Dankers et al., 2022; Zhang et al., 2022; Zhang and Feng, 2022; He et al., 2022).

	Original Word	**Segmented Subwords**
Example 1.	우두머리 (Chief Leader)	우두 + 머리 (Null meaning) + (Head)
Example 2.	큰코다치다 (Pay Dearly)	큰 + 코 + 다치다 (Big) + (Nose) + (Hurt)
Example 3.	바람개비 (Windmill)	바람 + 개비 (Wind or Affair) + (Whirligig)

Table 1: Examples of over-tokenization in Korean.

Although BPE is effective, it often results in over-segmentation when applied to Korean, an agglutinative and morphologically rich language (Keren et al., 2022). This issue primarily stems from the language-agnostic and frequency-based characteristics of BPE, which can inadvertently lead to excessive segmentation of words.

Concretely, we identified three potential instances of over-segmentation, as presented in Table 1. **The first** example is when the subword created via segmentation loses its meaning. For example, when the Korean word "우두머리 (Chief Leader)" is segmented into "우두 + 머리", "우두" becomes a word that does not have any meaning by itself. This results in creating a subword of null meaning that cannot be applied in any context, such as a prefix or suffix. **In the second** example, the word's meaning may not be preserved even when the individual meanings of subwords are combined. For instance, the meaning of "큰코다치다 (pay dearly)" changes when it is interpreted as a combination of the subwords "큰 (big)" + "코 (nose)" + "다치다 (hurt)." This demonstrates that the word's meaning cannot be fully preserved simply by segmenting it into interpretable units. **The third** example pertains to the increase in semantic ambiguity of subwords caused by word segmentation. For instance, consider the compound word "바람개비 (Windmill)." This word can be segmented as "바람 (wind or affair) + 개비 (whirligig)" since it is a

[*]Equal Contribution [†] Corresponding Author

Findings of the Association for Computational Linguistics: ACL 2024, pages 2287–2303
August 11-16, 2024 ©2024 Association for Computational Linguistics

compound word. However, the subword “바람” itself can have multiple meanings, such as "wind" or "affair," which results in ambiguity after segmentation.

Over-segmentation not only distorts the meaning of sentences but also introduces semantic noise during the training phase, which can significantly degrade the translation quality (Keren et al., 2022). In our investigation, we have observed that this semantic noise is further exacerbated in domain shift scenarios, where the model is challenged with out-of-domain test sets. This accumulation of semantic confusion is particularly detrimental in translation tasks where the fidelity of sentence semantics is paramount. Our findings underscore the critical need for addressing over-segmentation to ensure robustness against domain shifts and to maintain the integrity of translated content.

To address this challenge, we propose **LeVoC** (**Le**ngth-aware Subword **Vo**cabulary **C**onstruction), a novel approach specifically designed to mitigate the over-segmentation issue in Korean. By prioritizing morphologically rich long words during tokenization, LeVoC effectively alleviates over-segmentation, paving the way for more accurate and meaningful translations. To implement LeVoC, we first create a large-scale vocabulary using BPE and then build an auxiliary vocabulary set consisting of morphologically rich long words derived from it. We then construct a new tokenizer by incorporating the auxiliary vocabulary into the small-sized BPE vocabulary. More specifically, LeVoC preserves the existing tokenizer's capacity to overcome the issue of out-of-vocabulary (OOV) words, and the integration of long words from the auxiliary vocabulary additionally relieves the tendency for over-segmentation, leading to a more coherent and accurate representation. Through our experiments, we demonstrate that LeVoC mitigates over-segmentation issues and reduces semantic confusion in Korean NMT tasks, thereby enhancing translation performance.

We demonstrate that the NMT model trained using LeVoC effectively preserves the morphological structure of the sentences, resulting in superior performance on out-of-domain data that were not present during training. To the best of our knowledge, our method is the first to enhance translation performance by addressing over-segmentation issues stemming from the linguistic characteristics of Korean in an unsupervised manner. We publicly release LeVoC to support further studies[1].

2 Related Works

The constraints of BPE (Sennrich et al., 2016), especially when applied to agglutinative languages with rich morphological features, have been extensively reported. These languages, characterized by the fusion of numerous morphemes into individual words, pose a significant obstacle to the subword tokenization approach, which is fundamentally based on frequency or likelihood of occurrence (Bostrom and Durrett, 2020; Nzeyimana and Rubungo, 2022).

Several studies have attempted to address these limitations. For instance, some studies have applied morphological segmentation prior to subword tokenization (Park et al., 2019, 2020), while others have utilized Part-of-Speech (POS) tagging as additional tag features (Chimalamarri and Sitaram, 2021). Hofmann et al. (2022) proposed a method to preserve the morphological structure by recursively exploring the longest sub-string.

In addition, some studies have explored alternative approaches to subword tokenization. Clark et al. (2022) introduced a tokenization-free deep encoder that allows parameters to be shared by hashable code points without assigning embeddings for each character. Aguilar et al. (2021) proposed a method to handle noises robustly by learning and generating the subword embeddings through a pre-trained language model without any restrictions on the vocabulary.

While numerous studies have explored subword tokenization, the issue of over-segmentation specific to Korean language processing remains largely unexplored. LeVoC mitigates over-segmentation through a focus on vocabulary construction particularly in the context of the Korean language.

3 Proposed Method

3.1 Length-aware Subword Vocabulary Construction

Motivation In agglutinative languages, words are composed of several interdependent morphemes. This characteristic can be seen as a positive aspect in that it allows for constructing a vocabulary that covers most of the words in the dictionary using only a small number of subwords. This is the ideal scenario, wherein a small-sized vocabulary

[1] The code for LeVoC is publicly available at `https://github.com/js-lee-AI/LeVoC_BPE`

can be built by considering morphological factors and selecting only frequently-used subwords from the corresponding corpus. However, even with a subword that covers the entire vocabulary, the original meaning is not always preserved when segmented. This is due to the over-segmentation problem, in which the word's meaning is distorted, as previously mentioned in Table 1.

Our study reveals that subword tokenization methods such as BPE, which construct vocabularies in a language-agnostic manner, are particularly susceptible to over-segmentation issues. We attribute this to the fact that the existing subword tokenization methods generally rely on the statistical analysis of occurrence frequencies and do not consider *long words*[2]. In the case of BPE or SentencePiece (Kudo and Richardson, 2018), increasing the vocabulary size does not solve the over-segmentation problem and leads to delayed parameter updates in the embedding layer (Cherry et al., 2018; Ding et al., 2019).

To address this limitation, we propose intentionally increasing the vocabulary composition ratio for long words. Specifically, we hypothesize that over-segmentation issues must be considered to achieve high translation performance in Korean after addressing the OOV problem by increasing the number of BPE merge operations.

Methodology We denote $V_{\{\}}$ as the vocabulary of the BPE model $BPE_{\{\}}$. The LeVoC method is designed to augment the proportion of long tokens in the vocabulary. Rather than employing the BPE model derived from a parallel corpus P, LeVoC employs a BPE_{LeVoC} that exhibits a higher proportion of long tokens. The vocabulary V_{LeVoC} is formulated by merging the vocabularies obtained from the BPE model and the long token set. When the target vocabulary size for BPE_{LeVoC} is s, we train the BPE_{small} with a vocabulary size of $s(1-r)$, where r is a hyperparameter set between 0 and 1 to denote the proportion of added long tokens in V_{LeVoC}. The pseudocode for constructing LeVoC is illustrated in Figure 1.

Step 1. Employ a large monolingual Korean corpus, M, capable of extracting long tokens. Then divide M into two corpora: M_{Ext} for extracting long tokens and M_{Sel} for selecting tokens to be included in V_{LeVoC}.

Step 2. By utilizing M_{Ext}, train BPE_{Ext} with a vocabulary size of s_{Ext} and BPE_{small} with a vocabulary size of $s(1-r)$.

```
# P - Parallel corpus for training, comprising Korean
  text segmented by spaces
# M - Monolingual corpus used for extracting and long
  tokens
# s - Target vocabulary size for the Byte Pair Encoding
  model
# r - Hyperparameter set between 0 and 1 to represent
  the proportion of long tokens in the final vocab

def length_aware_subword_vocab_const(P, M, s, r):
  # Step 1: Split the monolingual corpus into two
    halves
  M_ext, M_sel = RAND_HALF_SPLIT(M)
  # Step 2: Train two BPE models with different vocab
    sizes
  BPE_small = BPE(s*(1-r), M)
  BPE_ext = BPE(1024k, M_ext)
  # Step 3: Extract words from BPE_ext vocab that have
    length ≥ 4
  BOW_ext = [w | w in BPE_ext.vocab if len(w) >= 4]
  # Step 4: Initialize LeVoC with BPE_small vocab
  V_LeVoC = BPE_small.vocab.copy()
  # Step 5: Add words from BOW_ext to LeVoC until it
    reaches target size
  while len(V_LeVoC) < s do
    # Select the word with the highest frequency in
      M_sel
    w = max(BOW_ext, key=λx: M_sel.count(x))
    # If the word is in P and not in LeVoC, add it to
      LeVoC
    if w ∈ P and w ∉ V_LeVoC then
      V_LeVoC.append(w)

  return V_LeVoC
```

Figure 1: Pseudocode for the Length-Aware Subword Vocabulary Construction method.

Step 3. Among the subwords included in V_{Ext}, extract lengthy tokens which character length is longer than L to create a bag of subwords, BOW_{Ext}. L is priorly set by a hyperparameter.

Step 4, 5. V_{LeVoC} is formulated by adding frequently occurring long tokens from BOW_{Ext} to V_{small}. Specifically, we add tokens in BOW_{Ext} sequentially until the number of added tokens reaches $s * r$. We prioritize subwords with high token frequency in M_{Sel} but exclude subwords not present in P.

Through these procedures, we can formulate V_{LeVoC} with a deliberately increased distribution for long tokens.

3.2 LeVoC Hyperparameter Setting

We put forth a series of recommendations for hyperparameter configurations. **(i)** When constructing LeVoC for generic domain NMT models, it is imperative to ensure that the number of words included in the context set M is sufficiently large. This is because, during the NMT training process, the model's performance may decline if the long tokens added from BOW_{Ext} are not included or are under-represented in the NMT training corpus.

[2] In this study, we define "long word" or "long token" as a word which character length is longer than 4.

Vocab Size	# of tokens by length					% of length less than 4
	1	2	3	4	5+	
4k	1,890	1,678	341	82	5	97.8
8k	2,068	4,046	1,425	396	61	94.3
16k	2,259	7,736	4,259	1,449	293	89.1
32k	2,419	13,120	10,527	4,695	1,235	81.5
64k	2,516	20,893	23,082	12,817	4,668	72.7

Table 2: Token length of BPE vocabulary established by 18GiB monolingual Korean corpus.

This can result in less updated word embeddings in the corresponding embedding layer (Luong et al.; Schick and Schütze, 2020; Yu et al., 2022). We suggest setting the context set to contain more than 0.2 million words, and in our experiments, we utilize a context set of approximately 1 billion words to increase diversity by including a wide range of words. **(ii)** We recommend setting the value of the s_{Ext} to be sufficiently large to ensure that various long tokens are covered. We found that increasing the value of s_{Ext} allows for extracting various tokens that maintain the morphological structure in our pilot study. In our experiments, we set the value of s_{Ext} to 1024k. **(iii)** The minimum length of long tokens, L, is recommended to be set to 4 or 5 for the Korean language. Table 2 shows that in vocabularies with fewer than 64k tokens, up to 98% of token lengths are less than 4. To capture a diverse range of long tokens that preserve the morphological structure, we set L to 4 in our primary experiments, and 5 in our case study (in Appendix A). **(iv)** We found that either a value of r of 0.2 or 0.4 effectively maintains the performance of the original BPE while adequately incorporating long tokens. In all our experiments, we selected the higher performing value of r between these two.

4 Experimental Setup

4.1 Dataset Details

We employed the AI-Hub[3] Korean-English parallel corpus to assess machine translation performance. The AI-Hub dataset, comprising approximately 1.6 million sentences, ensures high quality through human verification. We allocated substantial weight to the validation and test sets. Specifically, we randomly selected approximately 1.3 million samples from the AI-Hub data for the training set, 250,000 samples for the validation set, and the remaining for the test set. Detailed information on the dataset is provided in Appendix B.

Moreover, our experimental results on vocabulary corpus selection indicated that in the baseline BPE, a larger corpus (18 GiB) delivers superior performance compared to the only use of the AI-Hub corpus. As a result, we integrated a larger corpus into all experiments, including the baselines, to ensure a consistent and thorough evaluation. Additional information on corpus selection can be found in Appendix C.

4.2 Experimental Design

We utilized the BPE method as a baseline in our machine translation and morphological coverage experiments. We evaluated the performance of the models trained with the combination of English vocabulary of size 16k and Korean vocabulary of sizes 16k, 32k, and 64k.

In assessing machine translation performance, we train a Transformer (Vaswani et al., 2017) model from scratch and compared the translation performance of our proposed LeVoC-BPE with that of BPE. We also trained all machine translation models using Fairseq (Ott et al., 2019) with identical hyperparameters. The specific values of these hyperparameters are provided in Appendix D.

To evaluate morphological coverage, we extract morphologically rich words of four or more characters from both out-of-domain and in-domain corpora using the morpheme segmenter mecab-ko[4]. These words were subsequently used as an evaluation set for morpheme segmentation.

4.3 Evaluation Details

We assessed the performance of each NMT model using the SacreBLEU score (Papineni et al., 2002; Post, 2018), and chrF^{++} (Popović, 2015) metrics, with a beam size of 5 for all cases. The models were evaluated on the FLoRes dev-test sets (NLLB Team, 2022)[5] for out-of-domain comparison, and the divided AI-Hub test sets, for in-domain comparison. To evaluate Korean sentences, we segmented them into morpheme units, as their performance can be unduly underestimated depending on postpositions, suffixes, and prefixes (Park et al., 2020; Bandyopadhyay et al., 2021). For this purpose, we employed the segmenter, mecab-ko.

To explore the correlation between enhancements in machine translation performance and the morphological quality of long tokens, we report the

[3] https://aihub.or.kr/

[4] https://bitbucket.org/eunjeon/mecab-ko-dic/

[5] We utilized this dataset under a CC-BY-SA 4.0 license.

Direction	Vocabulary Size	Method	Out-domain		In-domain	
			BLEU	**chrF^{++}**	**BLEU**	**chrF^{++}**
En2Ko	16k	BPE	20.05	30.15	37.67	46.14
		LeVoC-BPE	**21.63**	**31.60**	**38.01**	**46.43**
	32k	BPE	19.35	29.49	37.40	45.84
		LeVoC-BPE	**21.48**	**31.24**	**38.24**	**45.91**
	64k	BPE	19.34	29.38	37.22	45.51
		LeVoC-BPE	**20.83**	**30.73**	**37.91**	**45.73**
Ko2En	16k	BPE	21.73	49.43	38.62	61.90
		LeVoC-BPE	**22.18**	**49.94**	**38.82**	**62.05**
	32k	BPE	20.74	48.98	38.30	61.71
		LeVoC-BPE	**21.69**	**49.40**	**38.52**	**61.84**
	64k	BPE	20.43	48.38	38.02	61.73
		LeVoC-BPE	**21.49**	**49.20**	**38.30**	**61.90**

Table 3: Comparison of the performances of LeVoC-BPE and baseline BPE on in-domain data from a randomly split test set from AI-Hub and out-of-domain data from the FloRes test set, which was not included in the training data.

average subword count and utilize full-match (Hofmann et al., 2022). The average subword count, indicative of a token's morphological information, is computed as the mean number of subwords encompassing all morpheme elements with 100% coverage. On the other hand, full-match denotes the accuracy of subwords that preserve morphological words intact, thereby facilitating the verification of tokenization accuracy. According to the definition of full-match by Hofmann et al. (2022), it occurs when the tokenization precisely aligns with the gold segmentation. This measurement is particularly suitable for assessing the mitigation of over-segmentation of long words, thereby providing a robust framework for evaluating our hypothesis regarding the performance.

The results regarding the precision, recall, and F1 scores for the morpheme boundary of tokenizers can be found in Appendix E. Briefly, LeVoC exhibits a higher recall and F1-score compared to BPE when given the same vocabulary size, indicating its superior performance in detecting reference boundaries.

5 Experimental Results

LeVoC with BPE Table 3 presents the evaluation results of LeVoC-BPE in comparison with the BPE. The experimental results are consistent with the findings of Ding et al. (2019), who observed that smaller vocabulary sizes lead to improved performance. The proposed LeVoC models demonstrate superior performance to all the baselines in both Ko2En and En2Ko tasks. Notably, in the case of En2Ko, LeVoC exhibited an improvement of up to 2.13 BLEU and 1.75 chrF^{++} on the out-of-domain test set compared to the baselines. Additionally, LeVoC achieves an improvement of up to 0.84 BLEU and 0.29 chrF^{++} on the in-domain test set.

LeVoC-BPE outperforms in the target language. We conjecture that this is caused by some long tokens, which are seldom observed in the machine translation training dataset, resulting in inadequately trained embeddings for these tokens (Schick and Schütze, 2020). The Transformer (Vaswani et al., 2017) architecture can effectively decode context-appropriate tokens by suppressing the generation of infrequently trained long Korean tokens in En2Ko. However, in Ko2En, where rare tokens are invariably encoded with the corresponding embedding vector, parameters for the infrequent words may not be adequately trained. This can cause the corresponding parameters to resemble a random embedding and fail to accurately reflect contextual information.

Moreover, we observe that LeVoC demonstrates superior translation performance on both out-of-domain and in-domain test sets. While in-domain performance is less affected by over-segmentation and the resulting semantic confusion, we observe that the impact becomes significantly more pronounced in an out-of-domain context. This finding aligns with our motivation, which emphasizes that over-segmentation during the training phase can systematically degrade the performance of translation models, particularly compromising reliability when faced with complex or out-of-domain texts. Given that LeVoCs are trained using identical hyperparameters, vocabulary size, and dataset as BPE, the results imply that LeVoC can rectify the issue of over-segmentation in the Korean language and be generalized for other datasets. Furthermore, an analysis of LeVoC's performance with downscaled training data sizes is presented in Section 6.3. In summary, it is observed that as the quantity of train-

Method	Out-domain		In-domain	
	Full Match (%)	Avg Subword	Full Match (%)	Avg Subword
BPE - 16k	11.05	**2.89**	13.52	**2.60**
LeVoC-BPE - 16k	**14.74**$_{(+3.69)}$	2.97$_{(+0.08)}$	**18.23**$_{(+4.71)}$	2.65$_{(+0.05)}$
BPE - 32k	20.96	**2.46**	25.15	**2.19**
LeVoC-BPE - 32k	**24.19**$_{(+3.23)}$	2.55$_{(+0.09)}$	**29.24**$_{(+4.09)}$	2.25$_{(+0.06)}$
BPE - 64k	36.63	**2.07**	41.82	1.87
LeVoC-BPE - 64k	**40.78**$_{(+4.15)}$	2.08$_{(+0.01)}$	**46.54**$_{(+4.72)}$	**1.85**$_{(-0.02)}$

Table 4: Experimental results regarding the preservation of morphological information. In the case of vocabularies of the same size, LeVoC-BPE has a significantly higher ‘Full Match’ rate than the baseline BPE, but it records slightly poorer performance at ‘Average Subword Count.’

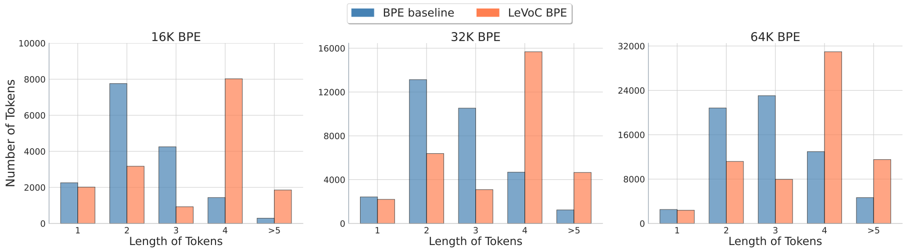

Figure 2: Token length distribution for vocabulary lists. The baseline BPE has a low ratio of long words and is heavily skewed toward short words. The LeVoC algorithm intentionally increases the ratio of long words (those with more than four characters).

ing data decreases, the performance gains provided by LeVoC diminish. This observation supports our initial premise that over-segmentation introduces accumulated noise into translation performance, which has a more pronounced effect in scenarios with larger training datasets.

The comparative analysis of the performances, using only the training corpus without any external corpus, is presented in Appendix F. Succinctly, barring the 64k vocabulary size, the performance of LeVoC trained with an external corpus surpasses that of LeVoC trained only with the training corpus. In addition, a comparison between Unigram (Kudo, 2018) and LeVoC-Unigram, with further experimental results, is provided in Section 6.5.

Preservation of Morphological Information Table 4 presents the experimental results regarding the preservation of morphological information. For vocabularies of equivalent size, LeVoC-BPE exhibits a substantial improvement in full-match scores compared to the baseline BPE method. However, LeVoC-BPE displays a minor decrease in terms of average subword count performance.

Our findings indicate that LeVoC can augment the morphological quality compared to the BPE with an equivalent vocabulary size. A high full-match rate signifies the ability to encode morphological information directly via the added long tokens. Despite achieving superior full-match performance, LeVoC exhibits a slightly lower performance of average subword count than BPE. The results suggest that even LeVoC cannot incorporate all the long tokens encountered during the training or inference phase. Nevertheless, LeVoC can segment the corresponding tokens at a finer granularity without the OOV problem.

The distribution of the token lengths for BPE and LeVoC-BPE vocabularies are depicted in Figure 2. As illustrated, LeVoC exhibits a higher composition ratio in long tokens, while BPE displays a skewed distribution towards shorter tokens. These distributions may serve as a crucial factor in achieving higher morphological full-match rates.

6 Case Studies

6.1 Impact on Morphologically Rich Setup

We propose that the successful translation of long words contributes significantly to the enhanced performance of LeVoC, primarily due to the distribution of token length within the same vocabulary size. To validate this, we present two scenarios: (1) Evaluation of the morphologically rich setting (MORPH-RICH), where all sentences contain a minimum of five words, each with at least four char-

Vocab Size	Method	MORPH-RICH BLEU	MORPH-IDEAL BLEU
16k	BPE	19.38	23.59
	LeVoC-BPE	**21.09**$_{(+1.71)}$	**29.93**$_{(+6.34)}$
32k	BPE	18.84	22.27
	LeVoC-BPE	**20.62**$_{(+1.78)}$	**25.20**$_{(+2.93)}$
64k	BPE	19.34	25.43
	LeVoC-BPE	**20.23**$_{(+0.88)}$	**26.80**$_{(+1.37)}$

Table 5: English to Korean translation performances for selected sentences from the out-of-domain test set.

acters, and (2) Evaluation of the morphologically ideal setting (MORPH-IDEAL), where each sentence includes at least five long words, each with a minimum of four characters that are found within the LeVoC vocabulary. The experimental results are displayed in Table 5. It is crucial to note that while preserving morphological information is a significant factor in LeVoC's high MT performance, other factors may also contribute to preserving morphological information.

Our findings indicate that LeVoC surpasses BPE in both settings, particularly demonstrating superior performance[6] in the MORPH-IDEAL setting. We deduce that the preservation of morphological information contributes to the high machine translation performance of LeVoC in these settings. In our prior experiments in Section 5, we discovered that LeVoC-BPE exhibits higher morphological full-match performance than BPE, suggesting that LeVoC-BPE can preserve morphological information within each sentence more effectively than BPE. This advantage may significantly enhance the performance of LeVoC in the MORPH-RICH and MORPH-IDEAL settings, as evidenced by the wider performance gap between LeVoc-BPE and BPE in these settings.

6.2 Morpheme-aware Supervised Method Setup

Using a morpheme segmenter is one of the key solutions for enhancing morpheme coverage. Morpheme segmenters are perceived as effective tools for integrating linguistic information. In practical terms, incorporating a morpheme segmenter into a BPE tokenizer can significantly improve Korean MT systems and is currently considered the state-of-the-art tokenizer (Park et al., 2020, 2021). In this section, we apply LeVoC to a morpheme-based BPE tokenizer (Morphs-BPE) and investigate whether we can achieve additional performance improvement. The Morphs-BPE model used in our experiments is a Korean-specific state-of-the-art tokenizer that applies morpheme-based segmentation followed by BPE training (Park et al., 2020). We apply our proposed method, LeVoC, to the Morphs-BPE tokenizer, resulting in the LeVoC-Morphs-BPE. The translation performances of this tokenizer are outlined in Table 6.

Vocab Size	Method	En2Ko BLEU	Ko2En BLEU	Full Match (%)
16K	Morphs-BPE	21.67	21.74	24.42
	LeVoC-Morphs-BPE	**21.98**	**21.97**	**32.94**
32K	Morphs-BPE	22.27	21.59	46.77
	LeVoC-Morphs-BPE	**22.32**	**22.11**	**58.52**
64K	Morphs-BPE	20.94	21.67	70.27
	LeVoC-Morphs-BPE	**21.50**	**21.91**	**75.11**

Table 6: Performance Comparison of LeVoC-Morphs-BPE with Morpheme-Based BPE. LeVoC shows better performance compared to the morpheme-BPE, albeit with small gains.

Experimental results indicate that LeVoC improves all BLEU and full-match scores of Morphs-BPE cases, with a maximum enhancement of 0.56 in BLEU score and 11.75 in full-match performance. This demonstrates the high applicability of LeVoC and suggests that intentionally increasing the composition ratio of long tokens in a BPE vocabulary provides a significant advantage in Korean MT models. Furthermore, these results confirm that even a morpheme-aware BPE utilizing a morpheme analyzer can still face the issue of over-segmentation. The token length distributions of both tokenizers are described in Appendix G.

Data Ratio	BPE	LeVoC-BPE
1/40 (32k)	**12.67**	12.51
1/20 (65k)	13.95	**14.31**
1/10 (130k)	15.60	**16.24**
1/1 (1.3M)	20.05	**21.63**

Table 7: Out-of-domain BLEU scores for varying training data sizes with a vocabulary size of 16k in English-to-Korean translation tasks.

6.3 Impact of Training Data Size

Table 7 shows the impact of training data size on the performance of the LeVoC method. It was ob-

[6] A direct comparison between LeVoC-BPE and Morph-BPE may not yield accurate results. LeVoC-BPE does not perform BPE at the morpheme level, unlike Morph-BPE which employs the same morpheme splitter, mecab-ko, for BPE at the morpheme level. Comparing these two methods could lead to an imbalanced and potentially unfair setup.

Source	Vocab Size	Method	Out-domain BLEU	In-domain BLEU
German	32k	BPE	28.79	30.98
German	32k	LeVoC	**29.19**	**31.04**
German	64k	BPE	29.00	**30.91**
German	64k	LeVoC	**29.23**	30.77

Table 8: Performance comparison between BPE and LeVoC for German.

served that as the volume of training data decreases, the performance gains afforded by LeVoC diminish. This trend may be attributed to two primary factors: (i) there may not be enough training corpus for the longer tokens added to LeVoC to be trained with appropriate embedding values, and (ii) a reduction in semantic confusion due to over-segmentation as the quantity of training data is scaled down. These observations lend support to our initial premise that over-segmentation exerts a more substantial influence on translation performance in scenarios where the training dataset is larger. The results indicate that while LeVoC can improve translation performance, its benefits are more pronounced with larger amounts of training data, where over-segmentation is more likely to cause semantic confusion.

6.4 LeVoC to Another Language

To investigate the impact of another language on LeVoC, we extended our experiments to include the German language, known for its morphological complexity. Utilizing a subset of the WMT19 En-De News Translation Task dataset[7], we maintained a consistent data setup with 1.3 million training pairs and 250,000 validation pairs, mirroring the Korean data configuration.

While the application of LeVoC to Korean showed consistent performance improvements over BPE across both vocabulary sizes and domains, the same approach applied to German, as shown in Table 8, yielded only marginal improvements. This suggests that the effectiveness of the LeVoC method may vary depending on the language.

6.5 LeVoC in Unigram Setup

Table 9 presents the outcomes of our initial examination on the performance comparison between Unigram (Kudo, 2018) and LeVoC-Unigram in English to Korean translation. The table elucidates the enhancements achieved by LeVoC when applied to both BPE and Unigram tokenization methods. It is observed that LeVoC consistently improves performance in out-of-domain settings for both tokenization methods, albeit with a smaller margin of improvement compared to BPE. In contrast, in in-domain settings, LeVoC-Unigram exhibits lower performance than Unigram alone.

Vocabulary Size	Method	Out-domain BLEU	In-domain BLEU
16k	BPE	20.05	37.67
	Unigram	20.18	37.96
	LeVoC-BPE	**21.63**	**38.01**
	LeVoC-Unigram	20.61	37.80
32k	BPE	19.35	37.40
	Unigram	19.80	37.93
	LeVoC-BPE	**21.48**	**38.24**
	LeVoC-Unigram	20.12	37.49
64k	BPE	19.34	37.22
	Unigram	19.89	37.42
	LeVoC-BPE	**20.83**	**37.91**
	LeVoC-Unigram	19.98	37.36

Table 9: Comparative analysis of LeVoC enhancements across Unigram method in English to Korean Translation.

7 Qualitative Analysis

Segmentation and Translation Fidelity In Table 10, we provide an empirical case demonstrating the adverse effects of BPE over-segmentation on the Korean-to-English translation task and how LeVoC ameliorates these issues. The BPE method dismantles the word "상대적으로 (relatively)" into subwords that lose their meaning, while LeVoC preserves the morphological integrity of the original word, resulting in a more accurate translation. This instance highlights the necessity of incorporating morphological considerations during segmentation to retain the intended meaning and enhance translation quality.

However, the implications of over-segmentation are not limited to isolated translation errors; it represents a systemic issue within NMT that introduces semantic noise across the training dataset. This noise impairs the model's ability to learn and generalize language patterns, leading to a potential decline in performance, especially with complex or non-standard texts. Initial training stages may exhibit frequent word-level mistranslations due to over-segmentation, but these errors become less common as training progresses. Nonetheless, persistent semantic noise continues to affect the model's performance adversely.

LeVoC addresses this issue by prioritizing including longer, morphologically rich subwords into the vocabulary, which supports the learning process by providing a clearer semantic signal.

[7] http://www.statmt.org/wmt19/

Method	**Korean Segmentation**
BPE 16k	페 / 르 / 시아 / 어는 / **상 (Null meaning)** / **대적 (hostility)** / **으로 (-ly)** / 쉽 / 고 / 대부분 / 규칙적 / 인 / 문 / 법을 / 가지고 / 있습니다 / .
LeVoC-BPE 16k	페르 / 시아 / 어는 / **상대적으로 (relatively)** / 쉽 /고 / 대부분 / 규칙적인 / 문 / 법 / 을 가지고 / 있습니다 / .
Target English: Persian has a relatively easy and mostly regular grammar	
BPE 16k Translation: Persian is a easy and mostly regular grammar.	
LeVoC-BPE 16k Translation: Persian is a relatively easy and mostly regular grammar.	

Table 10: Comparison of Ko2En translation results for out-of-domain samples. Each encoded token is separated by a forward slash (‘/’).

Source	노동자의 권리가 보장되고 일하는 사람들이 자기 **직장에서** **(In workplace)** 고용불안 없이 안전하게 일할 수 있는 사회가 됐으면 좋겠다. (I want society to be where workers rights are guaranteed and those who work can work safely **in their workplaces** without anxiety.)
BPE 16k	_노동 / 자의 / _권 / 리가 / _보장 / 되고 / _일하는 / _사람들이 / _자기 / **_직 (Null meaning)** / **장에서 (In the place)** / _고용 / 불 / 안 / _없이 / _안전하게 / _일할 / _수 / _있는 / _사회가 / _됐 / 으면 / _좋겠다 / .
LeVoC 16k	_노동자의 / _권 / 리가 / _보장 / 되고 / _일하는 / _사람들이 / _자기 / **_직장에서** **(In workplaces)** / _고용 / 불 / 안 / _없이 / _안전하게 / _일할 / _수 / _있는 / _사회가 / _됐 / 으면 / _좋겠다 / .
Source	뉴욕에서 쉽게 접할 수 없는 한국 민화의 아름다움 (Beauty)을 많은 뉴욕 시민들이 함께 즐기고 **알아보는** **(Recognizing)** 모습이 뜻깊었다. (It was meaningful to see many New Yorkers enjoying and **recognizing** the beauty of Korean folk paintings that were not easilyaccessible in New York.)
BPE 16k	_뉴욕 / 에서 / _쉽게 / _접할 / _수 / _없는 / _한국 / _민 / 화의 / _아름다 / 움을 / _많은 / _뉴욕 /_시민들이 / _함께 / _즐기고 / **_알아 (Know)** / **보는 (Looking)** / _모습이 / _뜻 / 깊 / 었다 / .
LeVoC 16k	_뉴욕에서 / _쉽게 / _접할 / _수 / _없는 / _한국 / _민 / 화의 / _아름다움을 / _많은 / _뉴욕 / _시민들이 / _함께 / _즐 /기고 / **_알아보는** **(Recognizing)** / _모습이 / _뜻 / 깊 / 었다 / .
Source	세 번째 통합-학습 관점에서는 **의사결정** **(Decision-making)** 과정에서 새로운 아이디어를 더 많이 활용하게 되고, 업무 효율이 높아진다. (In the third integrated-learning perspective / new ideas are used more in the **decision-making** process / and work efficiency is enhanced.)
BPE 16k	_세 / _번째 / _통합 / - / 학습 / _관 / 점 / 에서는 / **_의사 (Doctor, Opinion)** / **결정 (Decision)** / _과정에서 / _새로운 / _아이디어를 /_더 / _많이 / _활용 / 하게 / _되고 / , / _업무 / _효율 / 이 / _높아 / 진다 / .
LeVoC 16k	_세 / _번째 / _통합 / - / 학습 / _관 / 점 / 에서는 / **_의사결정** **(Decision-making)** / _과정에서 / _새로운 / _아이디어를 / _더 / _많이 /_활용 / 하게 / _되고 / , / _업무 / _효율 / 이 / _높아진다 / .

Table 11: Encoded sentences of 16k BPE and LeVoC. ‘Source’ is a cherry-picked sentence from the in-domain test set. Each encoded token is separated by a forward slash (‘/’).

Morphological Structure Preservation As depicted in Table 11, the results of BPE and LeVoC encoding examples for the in-domain test set are presented. Rows 1-3, 4-6, and 7-9 of the table correspond to cases 1, 2, and 3 outlined in Table 1, respectively. In the first case (rows 1-3), BPE decomposes the word “직장에서 (In workplace)” into “직 (Null meaning)” + “장에서 (In the place)”, resulting in the loss of the word’s meaning. In the second case (rows 4-6), BPE segments “알아보는 (Recognizing)” into two other words with almost irrelevant meanings as “알아 (Know)” + “보는 (Looking).” Similarly, in the last case (rows 7-9), “의사결정 (Decision-making)” is segmented into “의사 (Doctor, Opinion)”, which has ambiguous meanings, and “결정 (Decision).” Conversely, LeVoC includes these words in its vocabulary and encodes them as “직장에서 (In workplace)”, “알아보는 (Recognizing)”, and “의사결정 (Decision-making)”, without disassembling the words. Therefore, LeVoC can maintain a more accurate morphological structure compared to BPE. In Appendix H, we further present a qualitative analysis of extensive examples of generative results in English-to-Korean translation.

8 Conclusion

In this paper, we focus on the over-segmentation issue in the Korean MT models that erodes morphological information in the text. We discovered that conventional BPE tokenizers, while effectively alleviating the OOV problem, are prone to over-segmentation. We found that this semantic confusion, which stems from over-segmentation, ultimately degrades the overall translation quality. To address this issue, we propose LeVoC, a simple and effective method for mitigating over-segmentation by preserving the morphological structure of long tokens. We found that LeVoC can notably enhance the MT performance and preserve morphological information, especially for morphologically rich and ideal settings. LeVoC can be seamlessly integrated with a morpheme analyzer designed for the Korean language to maintain morphological integrity. Our comprehensive evaluation revealed that addressing over-segmentation via manipulating token length distribution in a BPE vocabulary improves MT performance substantially. We hope that the findings from our study will inspire further research in natural language processing by considering Korean’s unique characteristics.

Limitations

The LeVoC methodology, while promising, has certain limitations that should be acknowledged.

Focus on Korean Over-Segmentation The focus of this study was specifically on addressing the over-segmentation issue in Korean language, which is particularly pronounced when using the BPE due to the unique morphological richness of the language. This focus on BPE-driven over-segmentation in Korean is the primary reason why our experiments were not extended to other subword methods such as Unigram (Kudo, 2018) or Wordpiece (Schuster and Nakajima, 2012).

Language-Dependent Efficacy Our generalization study, as shown in Section 6.4, highlighted another significant limitation of the LeVoC methodology. While applying LeVoC to Korean consistently yielded performance improvements over BPE across different vocabulary sizes and domains, the same method when applied to German only resulted in marginal improvements. This limitation was also observed in our additional generalization experiments with other languages such as Mongolian and Japanese, which are not included in our experiments. These findings suggest that the efficacy of the LeVoC method may be language-dependent and potentially influenced by specific linguistic characteristics, such as the setting of the hyperparameter L. The discrepancy in performance improvement underscores the importance of considering the unique linguistic characteristics of each language when developing tokenization strategies. Further research is needed to understand how LeVoC can be effectively adapted to other morphologically rich languages.

Vocabulary Size Variation Due to computational resource constraints, we were unable to conduct experiments across a diverse range of vocabulary sizes. The impact of adding morphological words to the vocabulary with varying amounts of external data and different vocabulary sizes remains unexplored.

Domain-Specific Experiments Lastly, our experiments were primarily confined to the domain of neural machine translation. The applicability and effectiveness of the LeVoC methodology in other areas such as language modeling and language-agnostic segmentation have not been investigated and remain potential avenues for future research.

Ethics Statement

In our study, we utilized datasets in which potentially identifiable sentences present in the wikitext[8] had been removed by participants. Additionally, we utilized the official datasets that effectively mitigated concerns regarding aggregation privacy, thus ensuring that the data did not contain any personally identifiable information. Consequently, our work does not incorporate any potentially harmful sentences. However, it is crucial to acknowledge that, as with any general translation system, potential biases, such as those related to gender, may still be present (Stanovsky et al., 2019).

Acknowledgments

This work was supported by Institute for Information & communications Technology Promotion(IITP) grant funded by the Korea government(MSIT) (RS-2024-00398115, Research on the reliability and coherence of outcomes produced by Generative AI). This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(NRF-2021R1A6A1A03045425). This work was supported by Institute for Information & communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (No. 2022-0-00369, (Part 4) Development of AI Technology to support Expert Decision-making that can Explain the Reasons/Grounds for Judgment Results based on Expert Knowledge).

References

Gustavo Aguilar, Bryan McCann, Tong Niu, Nazneen Rajani, Nitish Shirish Keskar, and Thamar Solorio. 2021. Char2subword: Extending the subword embedding space using robust character compositionality. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1640–1651.

Saptarashmi Bandyopadhyay, Tasnim Kabir, Zizhen Lian, and Marine Carpuat. 2021. The university of maryland, college park submission to large-scale multilingual shared task at wmt 2021. In *Proceedings of the Sixth Conference on Machine Translation*, pages 383–386.

Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi,

[8] https://dumps.wikimedia.org/kowiki/

Philipp Koehn, Chi-kiu Lo, Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos Zampieri. 2020. Findings of the 2020 conference on machine translation (WMT20). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55, Online. Association for Computational Linguistics.

Kaj Bostrom and Greg Durrett. 2020. Byte pair encoding is suboptimal for language model pretraining. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4617–4624.

Colin Cherry, George Foster, Ankur Bapna, Orhan Firat, and Wolfgang Macherey. 2018. Revisiting character-based neural machine translation with capacity and compression. In *EMNLP*.

Santwana Chimalamarri and Dinkar Sitaram. 2021. Linguistically enhanced word segmentation for better neural machine translation of low resource agglutinative languages. *International Journal of Speech Technology*, 24(4):1047–1053.

Jonathan H Clark, Dan Garrette, Iulia Turc, and John Wieting. 2022. Canine: Pre-training an efficient tokenization-free encoder for language representation. *Transactions of the Association for Computational Linguistics*, 10:73–91.

Verna Dankers, Elia Bruni, and Dieuwke Hupkes. 2022. The paradox of the compositionality of natural language: A neural machine translation case study. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4154–4175.

Shuoyang Ding, Adithya Renduchintala, and Kevin Duh. 2019. A call for prudent choice of subword merge operations in neural machine translation. In *Proceedings of Machine Translation Summit XVII: Research Track*, pages 204–213.

Akhbardeh Farhad, Arkhangorodsky Arkady, Biesialska Magdalena, Bojar Ondřej, Chatterjee Rajen, Chaudhary Vishrav, Marta R Costa-jussa, España-Bonet Cristina, Fan Angela, Federmann Christian, et al. 2021. Findings of the 2021 conference on machine translation (wmt21). In *Proceedings of the Sixth Conference on Machine Translation*, pages 1–88. Association for Computational Linguistics.

Zhiwei He, Xing Wang, Rui Wang, Shuming Shi, and Zhaopeng Tu. 2022. Bridging the data gap between training and inference for unsupervised neural machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6611–6623.

Valentin Hofmann, Hinrich Schütze, and Janet Pierrehumbert. 2022. An embarrassingly simple method to mitigate undesirable properties of pretrained language model tokenizers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 385–393.

Omri Keren, Tal Avinari, Reut Tsarfaty, and Omer Levy. 2022. Breaking character: Are subwords good enough for mrls after all? *arXiv preprint arXiv:2204.04748*.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71.

Minh-Thang Luong, Ilya Sutskever, Quoc V Le, Oriol Vinyals, and Wojciech Zaremba. Addressing the rare word problem in neural machine translation.

Nathan Ng, Kyra Yee, Alexei Baevski, Myle Ott, Michael Auli, and Sergey Edunov. 2019. Facebook fair's wmt19 news translation task submission. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 314–319.

James Cross Onur Çelebi Maha Elbayad Kenneth Heafield Kevin Heffernan Elahe Kalbassi Janice Lam Daniel Licht Jean Maillard Anna Sun Skyler Wang Guillaume Wenzek Al Youngblood Bapi Akula Loic Barrault Gabriel Mejia Gonzalez Prangthip Hansanti John Hoffman Semarley Jarrett Kaushik Ram Sadagopan Dirk Rowe Shannon Spruit Chau Tran Pierre Andrews Necip Fazil Ayan Shruti Bhosale Sergey Edunov Angela Fan Cynthia Gao Vedanuj Goswami Francisco Guzmán Philipp Koehn Alexandre Mourachko Christophe Ropers Safiyyah Saleem Holger Schwenk Jeff Wang NLLB Team, Marta R. Costa-jussà. 2022. No language left behind: Scaling human-centered machine translation.

Antoine Nzeyimana and Andre Niyongabo Rubungo. 2022. Kinyabert: a morphology-aware kinyarwanda language model. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5347–5363.

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.

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.

Chanjun Park, Sugyeong Eo, Hyeonseok Moon, and Heui-Seok Lim. 2021. Should we find another model?: Improving neural machine translation performance with one-piece tokenization method without model modification. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Papers*, pages 97–104.

Chanjun Park, Heuiseok Lim, et al. 2019. Parallel corpus filtering and korean-optimized subword tokenization for machine translation. In *Annual Conference on Human and Language Technology*, pages 221–224. Human and Language Technology.

Kyubyong Park, Joohong Lee, Seongbo Jang, and Dawoon Jung. 2020. An empirical study of tokenization strategies for various korean nlp tasks. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 133–142.

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.

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.

Timo Schick and Hinrich Schütze. 2020. Rare words: A major problem for contextualized embeddings and how to fix it by attentive mimicking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8766–8774.

Mike Schuster and Kaisuke Nakajima. 2012. Japanese and korean voice search. In *2012 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5149–5152. IEEE.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725.

Gabriel Stanovsky, Noah A Smith, and Luke Zettlemoyer. 2019. Evaluating gender bias in machine translation. *arXiv preprint arXiv:1906.00591*.

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

Sangwon Yu, Jongyoon Song, Heeseung Kim, Seongmin Lee, Woo-Jong Ryu, and Sungroh Yoon. 2022. Rare tokens degenerate all tokens: Improving neural text generation via adaptive gradient gating for rare token embeddings. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 29–45.

Biao Zhang, Ankur Bapna, Melvin Johnson, Ali Dabirmoghaddam, Naveen Arivazhagan, and Orhan Firat. 2022. Multilingual document-level translation enables zero-shot transfer from sentences to documents. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4176–4192.

Shaolei Zhang and Yang Feng. 2022. Reducing position bias in simultaneous machine translation with length-aware framework. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6775–6788.

Chulun Zhou, Fandong Meng, Jie Zhou, Min Zhang, Hongji Wang, and Jinsong Su. 2022. Confidence based bidirectional global context aware training framework for neural machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2878–2889.

A Observations on Minimum Subword Length L

Vocab Size	L	Out-domain BLEU	In-domain BLEU	Full Match (%)
16k	4	**21.63**	**38.01**	**14.74**
	5	21.21	37.84	8.75
32k	4	**21.48**	**38.24**	**24.19**
	5	20.22	37.69	20.05
64k	4	**20.83**	**37.91**	**40.78**
	5	19.83	37.34	34.79

Table 12: Results of the ablation study for the minimum length. For all reported vocabulary sizes, a minimum length of 4 yields the highest BLEU score and full-match rate.

In this section, we experiment on the minimum subword length L that determines the standard for the "long token." We perform the experiments on the out-of-domain test set to inspect more general standards, and the results are presented in Table 12.

In the English-to-Korean translation task, we found that a minimum subword length of $L = 4$ outperforms $L = 5$ in all the cases. While $L = 5$ enables the extraction of more morphologically rich subwords, it limits the inclusion of words of length 4 in the vocabulary, resulting in a lower full-match rate. Additionally, using $L = 5$ necessitates a larger corpus to maintain the desired vocabulary size. Based on these results, we conclude that a minimum subword length of $L = 4$ is appropriate for use in LeVoC.

B Dataset

BPE Corpus The corpus employed in our experiments, referred to as the BPE corpus, comprises the following sources: Wikitext[9] (with potentially identifying sentences removed by study participants), ModuCorpus[10] (including messenger, written, colloquial, and web data), AI-Hub news data, AI-Hub MT corpus[11] (repeated five times), and a validation corpus. The total raw data amounts to 18.5 GB. By removing date or time notations (e.g., `(20191115 09:48)`) and repeated words (e.g., `blah blah`, `....`), we obtained a corpus of 17.9 GB. The statistics of this corpus used to construct the vocabularies are presented in Table 13.

# of Character	7,228,413,470
# of Words	1,659,449,657
# of Sentences	46,970,388
Avg of Word Length per Sentence	153.89

Table 13: Details of the BPE corpus.

AI-Hub The AI-Hub dataset comprises a total of 1.6 million sentences across five domains: news (0.8 million), websites (0.1 million), culture (0.1 million), ordinance (0.1 million), and spoken language (0.5 million). This is a Ko↔En parallel dataset, with an error rate of less than 1%. We divide the samples into train, validation, and test sets to analyze the experimental results while maintaining the domain ratios. The number of sentence pairs in the separated samples is presented in Table 14.

[9] https://dumps.wikimedia.org/kowiki/. We utilized this dataset under a CC BY-SA 3.0 license.

[10] https://corpus.korean.go.kr/

[11] https://www.aihub.or.kr/

	# of Sentence Pairs
Train	1,281,934
Validation	256,387
Test	64,097

Table 14: The number of AI-Hub sentence pairs used in experiments.

C Selection of Vocabulary Corpus

Table 15 shows the result of our pilot study that trains the transformer-based Ko2En NMT model using the vocabulary constructed through the large corpus (18GB) and AI-Hub. All V_{small}, V_{Ext}, and baseline BPE are trained through a large corpus for a fair highest performance comparison.

Vocab Size	Corpus Size	AI-Hub BLEU
16k	18GB	**38.62**
	174MB	37.99
32k	18GB	**38.30**
	174MB	37.80
64k	18GB	**38.02**
	174MB	37.69

Table 15: Korean-to-English machine translation performances trained with our 18 GB corpus and AI-Hub Korean monolingual corpus.

D MT setting details

We adopt the `fairseq` framework (Ott et al., 2019) for training the NMT model. To ensure a fair comparison, we train a Transformer-base NMT architecture from scratch using the same hyperparameters[12]. We train a MT model for each setting and the corresponding training arguments are noted in Figure 3.

```
--fp16
--fp16-init-scale 4096
--arch transformer
--optimizer adam --adam-betas ‘(0.9, 0.98)’
--clip-norm 1.0
--lr 5e-4 --lr-scheduler inverse_sqrt
--warmup-updates 4000
--dropout 0.1 –weight-decay 0.0001  --task translation
--criterion label_smoothed_cross_entropy
--label-smoothing 0.1
--max-tokens 4096
--update-freq 4
--best-checkpoint-metric bleu
--maximize-best-checkpoint-metric
--max-update 100000
--activation-fn gelu
--warmup-updates 4000  --share-decoder-input-output-embed
```

Figure 3: `fairseq` training arguments for our NMT experimental setting.

[12] We trained each model for approximately 30 hours with two NVIDIA RTX 6000 GPUs.

E Morpheme Boundary Analysis with LeVoC and BPE

Vocabulary Size	Method	Precision	Recall	F1
16k	BPE	**86.8**	69.9	77.4
	LeVoC	86.4	**70.7**	**77.8**
32k	BPE	**89.0**	66.7	76.3
	LeVoC	88.7	**67.2**	**76.5**
64k	BPE	**91.3**	63.7	75.0
	LeVoC	91.1	**64.1**	**75.3**

Table 16: Result of the Korean morpheme boundary experiments, which compares LeVoC and BPE on out-of-domain data using mecab-ko generated reference boundaries. The table shows the Precision, Recall, and F1-score for each method with the same vocabulary size. LeVoC has lower precision but higher recall and F1-score compared to BPE.

We utilized mecab-ko to generate reference boundaries from out-of-domain data, excluding all English words. In all cases, LeVoC exhibits lower precision but higher recall and F1-score compared to BPE when given the same vocabulary size. The diminished precision implies that LeVoC identifies boundaries even where actual morpheme boundaries do not exist, resulting in increased segmentation. The elevated recall signifies that LeVoC outperforms BPE in detecting reference boundaries, indicating that it does not overlook many genuine morpheme boundaries.

F Evaluating LeVoC with Only Training Corpus

In certain situations, an additional corpus may not be available. In such cases, we evaluate the performance of the MT model by configuring LeVoC using only the training corpus, i.e., using P to build BPE_{LeVoC}. For s_{Ext}, we adopt 512k instead of 1024k due to the training corpus size. Under this condition, V_{Ext} can not be significantly increased unless the training corpus is extensive, meaning that the number of long tokens that can be extracted may be limited. The results are presented in Table 17.

The performance of the LeVoC model trained using only the training corpus is generally similar to that of the LeVoC model trained using the additional large corpus, with the latter exhibiting slightly better performance.

Vocabulary Size	Method	BLEU
16k	LeVoC-BPE	**38.01**
	LeVoC-BPE*	37.64
32k	LeVoC-BPE	**38.24**
	LeVoC-BPE*	37.50
64k	LeVoC-BPE	37.91
	LeVoC-BPE*	**38.09**

Table 17: English-to-Korean performances of the LeVoC-BPE trained solely on the training corpus, evaluated on the in-domain test set. * denotes vocabulary that is only trained on the training dataset.

G Token Length Distribution of Morpheme Analyzer-based Vocabulary

We intentionally upscale the distribution of long subwords (with more than four character lengths) when constructing the LeVoC-BPE and LeVoC-Morphs-BPE vocabularies. Figure 4 explains the token length distribution of the vocabulary's lexical list generated by the morpheme analyzer-based approach.

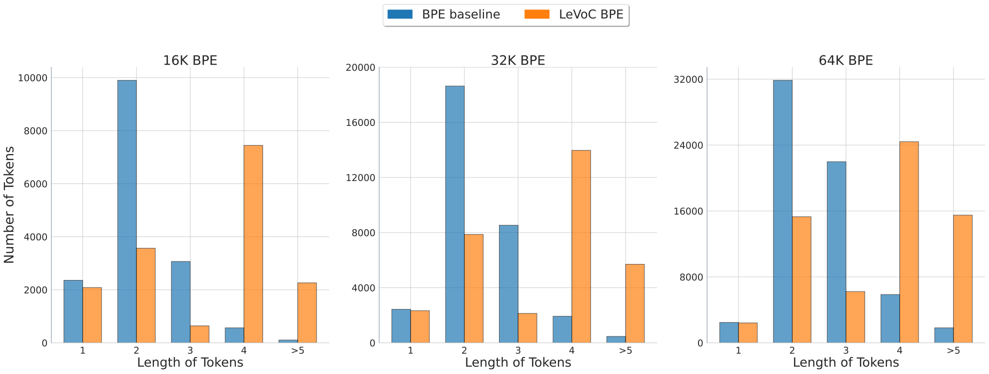

Figure 4: Distribution of token length for each LeVoC-morpheme BPE and morpheme-aware BPE vocabulary list.

H Examples of English-to-Korean Translation

For more extensive qualitative analysis, we provide inference examples of our NMT models. We compare the Korean-to-English BPE model to LeVoC-BPE by sampling low BLEU examples in the out-of-domain test set. The compared results are shown in Table 18. LeVoC-BPE decodes long words into long tokens, while BPE segments them into tokens of relatively constant length.

En (Source)	In the churchyard, there are interesting marble sculptures of doves over some tombs.
Ko (Target)	교회 뜰에는 몇몇 무덤 위에 비둘기의 흥미로운 대리석 조각들이 있다.
BPE 16k	교회 / 마 / 당 / 에는 일부 / 무덤 / 위에 / 비 / 둘 / 기 / 조각 / 들이 / 재미 /있는 /대리 / 석 / 조각 / 상 / 들이 / 있다 / . (In the church yard, there are doves sculptures and interesting marble sculptures over some tombs.)
LeVoC-BPE 16k	교회 / 마 / 당 / 에는 / 무 / 덤 / 위로 / 비 / 둘 / 기 / 들의 / 흥 / 미 / 로운 / 대리 / 석 / 조 / 각 / 들이 / 있다 / . (In the church yard, there are doves marble sculptures of doves over tombs.)
BPE 32k	교회 / 마 / 당 / 에는 / 일부 / 무 / 덤 / 위로 / 비 / 둘 / 기 / 들의 / 흥미로운 / 대리 / 석 / 조 / 형 / 물이 / 있다 / . (In the church yard, there are interesting marble sculptures of doves over some tombs.)
LeVoC-BPE 32k	교회 / 마 / 당 / 에는 무 / 덤 위 의 / 비 / 둘 / 기 조 각 / 이 / 재미 / 있 / 는데 / 요 / . (The sculpture of doves on the tombs is interesting in the church yard.)
BPE 64k	교회 / 마당 / 에는 / 일부 / 무덤 / 위에 / 비둘기 / 모양의 / 재미있는 / 대리 /석 / 조형물이 / 있다 / . (In the church yard , there are interesting marble sculptures in the shape of doves over some tombs.)
LeVoC-BPE 64k	교회 / 마 / 당 / 에는 / 일부 / 무 / 덤 / 위에 / 비 / 둘 / 기 / 조 / 각 / 이 / 재미있는 / 것이 / 있다 / . (In the church yard, there are a interesting sculptures over some tumbs.)
En (Source)	Since there was limited response to this tactic, Germany expected a similar response to its unrestricted submarine warfare.
Ko (Target)	이 전술에 대한 대응이 제한적이었기 때문에 독일은 무제한 잠수함전에서 비슷한 대응을 예상했습니다.
BPE 16k	이런 / 전술 / 에 / 대한 / 대응이 / 제한 / 적이 / 었던 / 만큼 / , / 독일 / 은 / 제한 / 없는 / 해 / 저 / 전 / 에도 / 비슷한 / 반응을 / 보일 / 것으로 / 예상했다 / . (As the response to these tactics was limited, Germany expected a similar reaction to the unrestricted submarine warfare.)
LeVoC-BPE 16k	이 / 전 술 / 에 / 대한 / 대응 / 이 / 제한 / 적이 / 었 / 기 / 때문에 / 독일 / 은 / 제한 / 없는 / 잠/ 수 / 함 / 전쟁 / 에도 / 유 / 사한 / 대응 / 을 / 예상했다 / . (Since there was limited response to this tactic, Germany expected a similary response to its unrestricted submarine warfare)
BPE 32k	이러한 / 전술 / 에 / 대한 / 대응 / 은 / 제한 / 적이 / 었기 / 때문에 / 독일은 / 제한 / 없는 / 잠수 / 함 / 전쟁 / 과 / 비슷한 / 반응을 / 예상했다 / . (Since the response to these tactics was limited, German expected a reaction similar to that of unrestricted submarine warfare.)
LeVoC-BPE 32k	이 / 전술 / 에 / 대한 / 대응 / 이 / 제한 / 적이 / 었 기 / 때문에 / 독일 / 은 / 무 / 제한 / 잠 / 수 / 함 / 전쟁 / 과 / 유 / 사한 / 대응 / 을 / 기대했다 / . (Since the response to this tactic was limited, Germany expected a similar response to its unrestricted submarine warfare.)
BPE 64k	독일은 / 이 / 같은 / 전술 / 대응이 / 제한 / 적이 / 었기 / 때문에 / 제한 / 되지 / 않은 / 잠수함 / 전 / 기와 / 유사한 / 대응을 / 기대했다 / . (Germany expected a response similar to unrestricted submarine warfare because this tactical response was limited.)
LeVoC-BPE 64k	이 / 전술 / 에 / 대한 / 대응 / 이 / 제한적이 / 었 / 기 / 때문에 / 독일 / 은 / 제한 / 되지 / 않은 / 잠수 / 함 / 전쟁 / 과 / 비슷한 / 대응 / 을 / 보 / 일 / 것으로 / 예상했다 / . (Since there was limited response to this tactic, Germany would expect a similar response to its unrestricted submarine warfare.)
En (Source)	Persian has a relatively easy and mostly regular grammar.
Ko (Target)	페르시아어는 상대적으로 쉽고 대부분 규칙적인 문법을 가지고 있습니다.
BPE 16k	페 / 르 / 시아 / 는 / 비교적 / 쉽고 / 대부분 / 규칙 / 적인 / 문 / 법을 / 가지고 / 있습니다 / . (Persia has relatively easy and mostly regular grammar.)
LeVoC-BPE 16k	페 / 르 / 시아 / 어는 / 문 / 법이 / 비교 / 적 / 쉽 / 고 / 대부분 / 규칙적인 / 문 / 법 / 입니다 / . (Persian is relatively easy and mostly regular grammar.)
BPE 32k	페르 / 시아 / 는 / 비교적 / 쉽고 / 규칙 / 적인 / 문 / 법을 / 가지고 / 있습니다 / . (Persia has relatively easy and regular grammar.)
LeVoC-BPE 32k	페 / 르 / 시아 / 어는 / 비교 / 적 / 쉽 / 고 / 규칙적인 / 문 / 법을 / 가지고 / 있습니다 / . (Persian has relatively easy and regular grammar.)
BPE 64k	페르시아 / 는 / 비교적 / 쉽고 / 일반적인 / 문 / 법을 / 가지고 / 있습니다 / . (Persia has relatively easy and general grammar.)
LeVoC-BPE 64k	페르시아 / 어는 / 비교적으로 / 쉽 / 고 / 대부분 / 규칙적인 / 문 / 법을 / 가지고 / 있습니다 / . (Persian has relatively easy and mostly regular grammar.)
En (Source)	With only eighteen medals available a day, a number of countries have failed to make the medal podium.
Ko (Target)	하루에 열여덟 개의 메달만 주어지기 때문에, 많은 국가가 메달 단상에 오르지 못했습니다.
BPE 16k	하루에 / 메달 / 이 / 18 / 개 / 뿐 / 인 / 가운데 / 메달 / 리스트 / 를 / 만들 / 지 / 못한 / 나라가 / 속 / 출 / 했다 . (Among only the 18 medals per day, a number of countries failed to make a medal list.)
LeVoC-BPE 16k	하루 / 에 / 메 / 달 / 을 / 18 / 개 / 밖에 / 쓸 / 수 / 없는 / 상황에서 / , / 수많은 / 국가 / 가 / 메 / 달 / 단 / 상에 / 오르 / 지 / 못했다 / . (In a situation where only 18 medals can be available per day, a large number of countries have failed to make the medal podium.)
BPE 32k	하루 / 메달 / 이 / 18 / 개 / 밖에 / 안 / 되는 / 상황에서 / 다수의 / 국가가 / 메달 / 단 / 상을 / 하지 / 못하고 / 있다 / . (In a situation of only 18 medals per day, a number of countries can't award the medal podium.)
LeVoC-BPE 32k	하루 / 에 / 18 / 개의 / 메 / 달 / 을 / 따 는 / 데 / 그치면서 / , / 다 수의 / 국가 / 가 / 메 / 달 / 단 / 상에 / 오르 / 지 못했습니다 / . (Winning only 18 medals in a day, a number of contries have failed to make medal podium.)
BPE 64k	단 / 하루 / 메달 / 18 / 개 / 만이 / 가능한 / 상황에서 / 수많은 / 나라가 / 메달 / 시상 / 대를 / 만들지 / 못했다 / . (With only 18 medals available a day, a number of countries have failed to establish a medal podium.)
LeVoC-BPE 64k	하루 / 열 / 여 / 덟 / 개 / 밖에 / 메 / 달 / 이 / 없는 / 상황에서 / 메 / 달 / 단 / 상에 / 오르 / 지 / 못한 / 나라 / 가 / 적지 / 않다 / . (With only eighteen medals available a day, there are a number of countries that have failed to make medal podium.)

Table 18: Translation examples of the out-of-domain test set generated by NMT models trained with LeVoC-BPE and baseline BPE. Each decoding step of translation results is separated by a '/'.