World Journal of Advanced Research and Reviews
eISSN: 2581-9615 CODEN (USA): WJARAI
Cross Ref DOI: 10.30574/wjarr
Journal homepage: https://wjarr.com/

(RESEARCH ARTICLE)

Check for updates

# Korean Subword vocabulary optimization by removing compositional words in neural machine translation

Kim Ryonghyok [1,*], Kim Kwanghyok [1], An Songil [2], Ryang Cholho [1] and Choe Jinhyok [1]

[1] Department of Artificial Intelligence, Artificial Intelligence Technology Institute, Kim Il Sung University, Pyongyang, Democratic People's Republic of Korea.
[2] Department of Foreign Language, Kim Il Sung University, Pyongyang, Democratic People's Republic of Korea.

World Journal of Advanced Research and Reviews, 2026, 29(03), 1008-1015

Publication history: Received on 17 January 2026; revised on 25 February 2026; accepted on 27 February 2026

Article DOI: https://doi.org/10.30574/wjarr.2026.29.3.0477

## Abstract

Byte Pair Encoding (BPE) is widely recognized as an effective approach for machine translation across multiple languages. However, in morphologically rich languages such as Korean, BPE can lead to excessive segmentation, which harms word semantics and creates semantic confusion during the training. This semantic confusion ultimately leads to an overall degradation in translation quality. Subword segmentation is an effective solution to the vocabulary problem in neural machine translation. This paper proposes a method to optimize the Korean subword vocabulary for neural machine translation, based on the fact that a Korean subword vocabulary created with the BPE training algorithm contains many compositional subwords. The optimized Korean subword vocabulary demonstrates experimentally stabilized translation performance by maintaining a balanced distribution while removing unnecessary compositional subwords.

**Keywords:** Korean Translation; NMT; Subword Vocabulary; BPE Learning Algorithm; Vocabulary Optimization

## 1. Introduction

The vocabulary problem has been recognized as one of the most important research topics since the early stages of research on neural machine translation, which has become the mainstream of machine translation research. [1] To address this problem, approaches based on subword-, character-, and byte-level vocabularies have been proposed. From an academic perspective, however, it is difficult to draw a definitive conclusion as to which of these units—subwords, characters, or bytes—is the most suitable translation unit for neural machine translation. In practical machine translation system development, subword-level translation models are generally acknowledged to outperform character- and byte-level models. [2,8]

Although small-scale experiments have shown that character-level translation models may have the potential to surpass subword-level models, many challenges remain in practical implementations, including the substantial increase in the length of segmented input sentences. [3]

Moreover, the fundamental unit of meaning is the word (or morpheme), and the modeling of the translation process—whose objective is to transform meaning from one language to another—naturally relies on units that carry semantic content. [4]

* Corresponding author: Kim Ryonghyok

Regardless of whether subwords, characters, or bytes are used as a means of addressing the vocabulary problem, a common issue arises: how to determine an appropriate vocabulary size. [5,6,9] In previous studies, this issue has typically been addressed not through theoretical analysis or explicit selection criteria, but through empirical choices based on experimental evaluations of the final neural machine translation performance.

In this paper, we propose a method for optimizing a Korean subword vocabulary from the perspective of neural machine translation, motivated by the observation that a Korean subword vocabulary constructed using the BPE training algorithm contains a large number of compositional subwords. Translation experiments conducted on the same training data confirm that the proposed Korean subword vocabulary update method consistently improves the accuracy of the translation model.

## 2. The role of subword segmentation in NMT

### 2.1. NMT model as a class classifier

Neural machine translation (NMT) can be defined as the task of converting a source sentence- $x = x_1 x_2 \dots x_m$ to the target sentence- $y = y_1 y_2 \dots y_n$. There exist several variants of the neural machine translation model, including RNN-based, CNN-based, and the Transformer-based architectures. Despite their structural differences, all of these models share a common objective: to maximize the conditional probability of the target sentence given the source sentence.

$$P(y|x) = \prod_{t=1}^{n} P(y_t|y_{<t}, x_{1:m})$$

for pairs $(x, y)$of source and target language sentences in parallel training data.

From the perspective of language generation, a neural machine translation model consisting of an encoder and a decoder can be regarded as a class classifier. At each decoding step, it outputs a token belonging to the target-language vocabulary based on a feature vector derived from the source sentence and the previously generated target tokens. Concretely, the word classifier takes as input the hidden state vector produced by the autoregressive framework and outputs a word from the target vocabulary.

In general, a class classifier $C$ learns from a training example set

$$\text{TS} = \{(x_i, y_i) | i = \overline{1, N}, x_i \in X, y_i \in Y\}$$

where the objective is to learn a classification function that maps vectors in the input feature space X to elements of the output class set Y.

In the context of neural machine translation, the training example set TS corresponds to the collection of classification instances that the model must learn as a class classifier. Consequently, the size of $TS$ is equal to the total number of target-language tokens obtained after subword segmentation of the training corpus.

Let us divide the TS into class-specific subsets:

$$TS = TS_1 \cup TS_2 \cup \dots \cup TS_{|Y|}$$

$$\forall i, j \in \{1, \dots, |Y|\}, \ , i \neq j \Rightarrow \ TS_i \cap TS_j = \emptyset$$

$$\forall i \in \{1, \dots, |Y|\}, TS_i = \{(x_j, y_j) | (x_j, y_j) \in TS, y_j = c_i\}$$

It is widely recognized that when the frequency distribution of class-specific training examples is highly imbalanced, a class classifier trained on such data tends to perform better on classes with higher frequencies. Naturally, if the input feature vectors are well clustered by class, effective training may still be possible even under imbalanced class distributions. Nevertheless, it is generally assumed that class classifier training becomes more stable and reliable when the number of training examples per class is uniformly distributed.

Therefore, it is useful to compare the frequency distribution of class-specific training examples in a given training dataset against a uniform distribution. To this end, we define the following metric:

$$d(TS) = \sum_{i=1}^{|Y|} \left| \frac{|TS_i|}{N} - \frac{1}{|Y|} \right| \qquad (1)$$

This metric is determined by the frequency distribution of subword occurrences in the target-language subword vocabulary of the neural machine translation model. Although it does not constitute an absolute evaluation criterion, it provides a meaningful comparative measure for analyzing and contrasting different target-language subword vocabularies.

### 2.2. Role of subword segmentation

The manner in which a neural machine translation (NMT) model translates a given source-language word is not determined solely by the number of training instances in which that word appears. Rather, it is influenced by the diversity of contextual environments in which the word occurs and by the variability of its corresponding translations within source-language sentences. Nevertheless, it is evident that, as the frequency of a source-language word in the parallel training corpus increases, the number of aligned translation examples also increases, thereby improving the likelihood that the NMT model will translate the word correctly.

Subword segmentation decomposes low-frequency words into sequences of subword units with significantly higher occurrence frequencies. As a result, it reshapes the frequency distribution of training examples in a way that is more favorable for the training of neural machine translation models, which can be viewed as large-scale class classifiers. From this perspective, it does not need to strictly necessary for subword segmentation to conform to linguistically motivated morphological boundaries. For a class classifier, the critical factor is the frequency distribution of class-specific training examples; whether the partitioning of a word with respect to linguistic morphology does not, by itself, determine these distributional properties.

In this sense, previously proposed character-level and even byte-level segmentation methods are grounded in the same fundamental assumption. However, the widely held belief that linguistically motivated subword segmentation inherently leads to desirable class-frequency distributions is, in many cases, an intuitively accepted yet insufficiently justified convention.

A crucial aspect of analyzing the impact of subword segmentation on the translation accuracy of NMT models lies in the model's ability to translate source-language words or expressions that are absent from the training data. In summary, while subword segmentation does provide certain advantages for training translation models from the standpoint of class classification, such advantages materialize only when the translations of individual subword units can be effectively recombined to produce an accurate translation of the original word. This is particularly relevant in cases such as the translation of native English words into Korean, where appropriate register and compositional reconstruction are required.

## 3. A Korean subword dictionary optimization method for neural machine translation

### 3.1. Characteristics of Korean Subword Dictionaries

One notable property of the Korean subword vocabulary constructed via the BPE learning algorithm is the prevalence of compositional subwords

**Table 1** Frequency of occurrence of several Korean sub-words

| Korean subword | FoO | Korean subword | FoO | Korean subword | FoO |
|---|---|---|---|---|---|
| 국가 | 45668 | 규칙 | 45108 | 기관 | 61584 |
| 국가. | 3445 | 규칙. | 3060 | 기관. | 8266 |
| 국가가. | 1262 | 규칙과. | 911 | 기관과. | 1943 |

| | | | | | |
|---|---|---|---|---|---|
| 국가계획 | 1186 | 규칙들에· | 830 | 기관들· | 707 |
| 국가적인· | 2235 | 규칙 | 45108 | 기관을· | 3029 |
| | | 규칙· | 3060 | 기관의· | 5530 |
| | | | | 기관이· | 2112 |

Since the 《국가계획》(National Planning) is a composite word of two subwords, 《국가》(National) and 《계획》(Planning), it is a subject that the NMT model can be successfully translated using the 《국가》(National) and 《계획》(Planning) subwords without having to register it in the subword dictionary.

In addition, the sub-words such as《국가는·》,《국가로·》,《국가를·》,《국가에·》,《국가에서·》,《국가와·》, and《국가의·》 listed in the table are also included in the compound word, as they are created by adding the sub-words《국가》(National) to the vomits that perform grammatical functions.

Therefore, it is quite reasonable to assume that the new subword dictionary created by removing these subword from the current subword dictionary will not negatively affect the training of the neural machine translation model.

### 3.2. A Korean subword dictionary optimization method with Composite Word Elimination

The core idea of our subword vocabulary optimization method is to remove a subword token XY from the subword vocabulary when the following conditions are satisfied: the tokens XY, X, and Y all exist in the vocabulary, and XY is a compositional word formed by combining X and Y.

Korean is a language characterized by a wide variety of compound formations involving two or more lexical elements. Typical examples include compound nouns and combinations of nouns with postpositions or other functional morphemes. Such compound forms can be detected automatically, since a prefix–string relationship exists between the compound word and its constituent elements.

Based on Equation (1), we analyze how the removal of such compositional subwords affects the distance between the frequency distribution induced by the resulting subword vocabulary and the uniform distribution. For the sake of simplicity, consider the case in which a subword class $c_k$is composed of two subword classes $c_p$and $c_q$. If $c_k$is removed from the subword vocabulary, all training instances previously assigned to $c_k$will be redistributed to $c_p$and $c_q$, as the compound subword will be segmented into its two constituent subwords. Consequently, the class frequencies of $c_p$and $c_q$will increase by the occurrence frequency of $c_k$, thereby altering the overall class-frequency distribution in the direction of a more balanced distribution. Then the difference between the subword frequency distribution and the uniform distribution of the subword word dictionary containing subword $c_k$ and the subword dictionary containing no $c_k$ will be

$$d_1(TS) = \sum_{i=1}^{|\mathrm{Y}|} \left| \frac{|TS_i|}{N} - \frac{1}{|\mathrm{Y}|} \right| \qquad (2)$$

$$d_2(TS') = \sum_{j=1}^{|\mathrm{Y}|-1} \left| \frac{|TS'_j|}{N'} - \frac{1}{|\mathrm{Y}|-1} \right| \qquad (3)$$

Here, TS and $TS'$ are the training examples set as class classifiers when using a subword dictionary, containing or not a subword $c_k$, and N and $N'$ are the number of training examples.

So,

$$d_1(TS) - d_2(TS') = \left|\frac{|TS_k|}{N} - \frac{1}{|\mathrm{Y}|}\right| + \left|\frac{|TS_p|}{N} - \frac{1}{|\mathrm{Y}|}\right| + \left|\frac{|TS_q|}{N} - \frac{1}{|\mathrm{Y}|}\right| - \left(\left|\frac{|TS_k| + |TS_p|}{N'} - \frac{1}{|\mathrm{Y}| - 1}\right| + \left|\frac{|TS_k| + |TS_q|}{N'} - \frac{1}{|\mathrm{Y}| - 1}\right|\right) \quad (4)$$

$$N' = N + 2|TS_k| - |TS_k| = N + |TS_k| \quad (5)$$

For simplicity of the expansion, let us denote $|TS_k| = C_k, |TS_p| = C_p, |TS_q| = C_q$.

Since $\frac{1}{|\mathrm{Y}|-1} - \frac{1}{|\mathrm{Y}|} \approx 10^{-5}$ for $|\mathrm{Y}| = 32\mathrm{K}$, we have

$$d_1(TS) - d_2(TS') \approx \left|\frac{C_k}{N} - \frac{1}{|\mathrm{Y}|}\right| + \left|\frac{C_p}{N} - \frac{1}{|\mathrm{Y}|}\right| + \left|\frac{C_q}{N} - \frac{1}{|\mathrm{Y}|}\right| - \left(\left|\frac{C_k + C_p}{N'} - \frac{1}{|\mathrm{Y}|}\right| + \left|\frac{C_k + C_q}{N'} - \frac{1}{|\mathrm{Y}|}\right|\right) \quad (6)$$

Let us evaluate (6) if the frequency of occurrence of the $c_p$ and $c_q$ are all less than the frequency of occurrence corresponding to a uniform distribution, with and without $c_k$ in the subword dictionary.

$$\frac{C_k}{N}, \frac{C_p}{N}, \frac{C_q}{N}, \frac{C_p + C_k}{N}, \frac{C_q + C_k}{N} < \frac{1}{|\mathrm{Y}|} \Rightarrow$$

$$d_1(TS) - d_2(TS') \approx \left(\frac{1}{|\mathrm{Y}|} - \frac{C_k}{N}\right) + \left(\frac{1}{|\mathrm{Y}|} - \frac{C_p}{N}\right) + \left(\frac{1}{|\mathrm{Y}|} - \frac{C_p}{N}\right) - \left(\frac{1}{|\mathrm{Y}|} - \frac{C_k + C_p}{N'} + \frac{1}{|\mathrm{Y}|} - \frac{C_k + C_q}{N'}\right)$$
$$= \frac{1}{|\mathrm{Y}|} - \frac{C_k + C_p + C_q}{N} + \frac{2C_k + C_p + C_q}{N'} \quad (7)$$

Using (5), (7) becomes $d_1(TS) - d_2(TS') \approx \frac{1}{|\mathrm{Y}|} - \frac{C'}{N} + \frac{C' + C_k}{N + C_k} > \frac{1}{|\mathrm{Y}|}$.

We consider $\forall N, C_k, C' > 0, \ \frac{C'}{N} < \frac{C' + C_k}{N + C_k}$.

That is, if the occurrence frequencies of the three subwords $c_k$, $c_p$, and $c_q$ are all lower than the frequencies implied by a uniform distribution, then the subword vocabulary obtained by deleting $c_k$ is closer to the uniform distribution than the original subword vocabulary.

Conversely, if the occurrence frequencies of the constituent subwords $c_p$ and $c_q$ are both higher than those corresponding to a uniform distribution, the original subword vocabulary is closer to the uniform distribution than the vocabulary in which the compound subword $c_k$ has been removed.

$$\frac{C_k}{N}, \frac{C_p}{N}, \frac{C_q}{N} > \frac{1}{|\mathrm{Y}|} \Rightarrow$$

$$d_1(TS) - d_2(TS') \approx \left(\frac{C_k}{N} - \frac{1}{|\mathrm{Y}|}\right) + \left(\frac{C_p}{N} - \frac{1}{|\mathrm{Y}|}\right) + \left(\frac{C_q}{N} - \frac{1}{|\mathrm{Y}|}\right) - \left[\left(\frac{C_k + C_p}{N'} - \frac{1}{|\mathrm{Y}|}\right) + \left(\frac{C_k + C_q}{N'} - \frac{1}{|\mathrm{Y}|}\right)\right]$$
$$= \frac{C_k + C_p + C_q}{N} - \frac{2C_k + C_p + C_q}{N'} - \frac{1}{|\mathrm{Y}|} \quad (8)$$

$$d_1(TS) - d_2(TS') \approx \frac{C'}{N} - \frac{C' + C_k}{N + C_k} - \frac{1}{|\mathrm{Y}|} < -\frac{1}{|\mathrm{Y}|}$$

In addition, depending on the respective occurrence frequencies of the subwords $c_p$, $c_q$, and $c_k$, several distinct cases may arise. By removing the subword $c_k$from the subword vocabulary, it becomes possible to determine—on a case-by-case basis—whether the resulting subword frequency distribution moves closer to the uniform distribution or diverges further from it. This assessment can be carried out by directly evaluating the change in the distribution distance defined in Equation (1).

## 4. Experimental and evaluation

We experimentally evaluate the effectiveness of the previously proposed method for optimizing a Korean subword vocabulary through compositional subword removal by training an English–Korean neural machine translation model.

### 4.1. Experimental Objectives and Hypotheses

The objective of this experiment is to verify whether the Korean subword vocabulary optimization method, based on compositional subword removal, has a positive impact on the training process and translation accuracy of a neural machine translation model.

Accordingly, the following hypotheses are formulated:

- Hypothesis 1: The subword vocabulary updated by removing compositional subwords makes the subword frequency distribution more closely approximate a uniform distribution.
- Hypothesis 2: This improvement in the distribution stabilizes the training of the neural machine translation model, which functions as a classifier, and consequently improves its final translation accuracy.

### 4.2. Training Data and Model Configuration

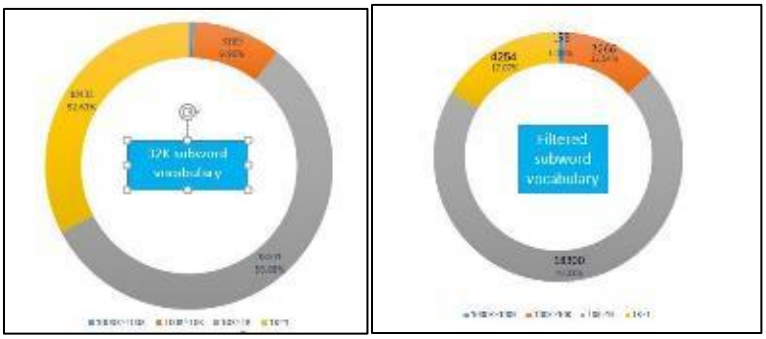
For training the English–Korean neural machine translation model, publicly available Korean–English parallel corpora from the OPUS repository, specifically CCAligned and OpenSubtitles, we used as training data. For subword vocabulary training, a monolingual corpus was constructed from the Korean Wikipedia dump. The English–Korean parallel texts were used for model training, and the translation model adopted a standard Transformer architecture. The sizes of the development and test sets were 3,000 and 2,000 sentences, respectively.

Model training was conducted for approximately 15 epochs using an NVIDIA GeForce RTX-3060. The dropout rate was set to 0.1, and both the encoder and decoder consisted of 6 layers each, with a hidden dimension of 512. The Adam optimizer was used with parameters $\beta_1 = 0.9$, $\beta_2 = 0.98$, and $\varepsilon = 10^{-9}$, and the number of warm-up steps was set to 12,000.

### 4.3. Training results

In the 32K-scale Korean subword vocabulary, the number of tokens automatically identified as compositional subwords was 5,982, and as a result, the updated Korean subword vocabulary was reduced by approximately 18.7% compared to the original subword vocabulary. That is, the only difference between the two models compared in the experiment lies in the method of subword vocabulary construction.

Below, we present the distributions of the original 32K subword vocabulary and the updated subword vocabulary.




(Left: before updating, Right: after updating)

**Figure 2** Distributional characteristics of the subword vocabulary according to frequency ranges

When compositional subwords contained in the Korean subword vocabulary are removed, the distance between the frequency distribution of training instances and the uniform distribution decreases from 1,919.09 to 1,559.4.

In the vocabulary before updating, a large number of specific compositional subwords appear with very low frequencies, resulting in a pronounced long-tail phenomenon.

In the updated vocabulary, as these rare compositional subwords are removed, the proportion of subwords in the medium-frequency range increases relatively.

This improvement experimentally supports Hypothesis 1, which states that the distributional characteristics of training instances for the neural machine translation model, acting as a classifier, become more closely approximated to a uniform distribution.

### 4.4. Translation Accuracy Evaluation Results

Translation accuracy was evaluated using the BLEU metric for each translation direction. Table 1 presents the BLEU performance improvements obtained when using only the BPE vocabulary compared to using the optimized subword vocabulary (BPE-Optim) on the same test set.

**Table 2** When using the updated subword vocabulary (BPE-Optim), a consistent performance improvement of approximately +0.51 BLEU is observed.

| **Method** | **EN2KO BLEU** | **KO2EN BLEU** |
|---|---|---|
| BPE- 32K | 24.93 | 24.91 |
| BPE-Optim-32K | 25.44 | 25.37 |

Although the magnitude of this improvement may appear numerically modest, it can be regarded as a meaningful gain, given that it is achieved solely by changing the vocabulary construction method under conditions of large-scale parallel corpora and a standard Transformer model.

### 4.5. Interpretation of Results and Discussion

The experimental results allow for the following interpretations:

- First, the removal of compositional subwords effectively eliminates unnecessary rare classes without damaging the meaning units essential for translation.
- Second, this process helps stabilize the training of the neural machine translation model by reconstructing the target-language subword vocabulary into a more appropriately sized set of classes.
- Third, these effects suggest that the observed improvements are not due to the inherent superiority of subword units themselves, but rather to the optimization of the internal structure of the subword vocabulary.

These findings are significant in that they experimentally demonstrate that the quality of subword vocabulary construction is an important independent factor, moving beyond the ongoing debate over “subword versus character/byte” as translation unit.

## 5. Conclusions and future work

It can be concluded that the proposed method of subword lexical optimization through compositional word removal is applicable to NMT model of all translation languages with Korean language as the source or target language, and has some degree of effectiveness in improving translation accuracy.

This approach may have a greater impact on translation model learning in low-resource language environments, i.e. in the absence of training data.

In low-resource language environments, subword lexicon optimization methods are the main future research directions for improving the accuracy of neural machine translation.

## Compliance with ethical standards

*Disclosure of conflict of interest*

No conflict of interest to be disclosed.

## References


[1] Felix Stahlberg, Neural Machine Translation: A Review and Survey, 2020, arXiv:1912.02047v2

[2] Rohit Gupta, Laurent Besacier, Marc Dymetman, and Matthias Gallé. 2019. Character-based nmt with transformer. arXiv preprint arXiv:1911.04997

[3] Bojar, O., Chatterjee, R., Federmann, C., Graham, Y., Haddow, B., Huang, S., Huck, M., Koehn, P., Liu, Q., Logacheva, V., Monz, C., Negri, M., Post, M., Rubino, R., Specia, L., & Turchi, M. (2017). Findings of the 2017 conference on machine translation (WMT17). In Proceedings of the Second Conference on Machine Translation (pp. 169–214). Association for Computational Linguistics.

[4] Bojar, O., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Koehn, P., & Monz, C. (2018). Findings of the 2018 conference on machine translation (WMT18). In Proceedings of the Third Conference on Machine Translation: Shared Task Papers (pp. 272–303). Association for Computational Linguistics.

[5] Ximena Gutierrez-Vasques, Christian Bentz, Olga Sozinova, and Tanja Samardzic (2021), From characters to words: the turning point of BPE merges, Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume (pp. 3454-3468). Association for Computational Linguistics.

[6] Sennrich, R., Haddow, B., & Birch, A. (2016). *Neural machine translation of rare words with subword units*. ACL.

[7] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL 2019* (pp. 4171–4186).

[8] Jungseob Lee, Hyeonseok Moon, Seungjun Lee, Chanjun Park , Sugyeong Eo, Hyunwoong Ko, Jaehyung Seo, Seungyoon Lee, Heuiseok Lim, Length-aware. Byte Pair Encoding for Mitigating Over-segmentation in Korean Machine Translation. Findings of the Association for Computational Linguistics: ACL 2024, (pp. 2287–2303)

[9] Seungjun Lee, Jungseob Lee, Hyeonseok Moon, Chanjun Park, Jaehyung Seo, Sugyeong Eo, Seonmin Koo 1 and Heuiseok Lim. A Survey on Evaluation Metrics for Machine Translation. A Survey on Evaluation Metrics for Machine Translation. Mathematics 2023, 11 (pp. 1006).