Should we find another model?: Improving Neural Machine Translation Performance with ONE-Piece Tokenization Method without Model Modification

Chanjun Park[1†], **Sugyeong Eo**[1†], **Hyeonseok Moon**[1†], **Heuiseok Lim**[1*]
[1]Korea University, South Korea
{bcj1210, djtnrud, glee889, limhseok}@korea.ac.kr

Abstract

Most of the recent natural language processing (NLP) studies are based on the pretrain-finetuning approach (PFA). However for small and medium-sized industries with insufficient hardware, there are many limitations in servicing latest PFA based NLP application software, due to slow speed and insufficient memory. Since these approaches generally require large amounts of data, it is much more difficult to service with PFA especially for low-resource languages. We propose a new tokenization method, ONE-Piece, to address this limitation. ONE-Piece combines morphologically-aware subword tokenization and vocabulary communicating method, which has not been carefully considered before. Our proposed method can also be utilized without modifying the model structure. We experiment by applying ONE-Piece to Korean, a morphologically-rich and low-resource language. We revealed that ONE-Piece with vanilla transformer model can achieve comparable performance to the current Korean-English machine translation state-of-the-art model.

1 Introduction

Recent studies using pretrain-finetuning approach (PFA) technique have achieved state-of-the-art (SOTA) performance in many natural language processing (NLP) tasks and are becoming the latest trend (Devlin et al., 2018; Yang et al., 2019; Radford et al., 2019; Brown et al., 2020; Liu et al., 2019; Clark et al., 2020). To utilize the PFA, a large amount of pre-training data and a system with sufficient computing power are required. For example, T5 (Raffel et al., 2019) was trained with 11 B parameters and 1 T tokens in order to get SOTA performance, and for GPT3 (Brown et al., 2020), 170 B parameters were required to train a model to demonstrate the best performance.

The trend of model research based on PFA raises two problems. First, it is hard to expect a similar performance for the low-resource setting. This is because most studies based on the PFA technique rely on large amounts of data (Zoph et al., 2016). But for low-resource languages, it is difficult to provide the comparable amount of data required by recent papers. Second, it is necessary to overturn the existing model and pre-train a new model from scratch to create a PFA-based model that follows the latest research trends.

Since the PFA-based model requires many parameters, companies without adequate server or graphic processing unit (GPU) environments may have many difficulties in configuring the service environment and utilizing the latest model (Park et al., 2020b). Therefore, new approaches are required to ensure high performance for low-resource languages and companies lacking extensive server and GPU environments.

To solve this problem, many researches are being conducted on the way of improving the performance of NLP application software without changing the model through data pre and post-processing, typically in machine translation (Pal et al., 2016; Currey et al., 2017; Banerjee and Bhattacharyya, 2018; Koehn et al., 2018; Kudo, 2018; Park et al., 2020b). Reflecting this trend, we conducted a study on an optimized tokenization that can improve the performance of neural machine translation (NMT) without changing the model.

We propose two perspectives for optimized tokenization. First, we analyze the limitations of byte pair encoding (BPE) (Sennrich et al., 2015) and sentencepiece (Kudo and Richardson, 2018), which can easily be applied to various languages. Due to its language-agnostic characteristic, these methods are currently used as the defaults in language model research and existing tokenization methods. However, there are 7,111 languages around the world. More than 50 million people speak 25 languages

[†] All authors contributed equally.
[*] Corresponding author.

Proceedings of NAACL HLT 2021: IndustryTrack Papers, pages 97–104
June 6–11, 2021. ©2021 Association for Computational Linguistics

as their mother tongue that have various morphological characteristics such as isolating language, agglutinative language, and fusional language. Considering this, it seems hard to assert that applying sentencepiece and BPE always produce the best performance.

Second, we focus on the problem that there is not enough discussion about the corpus used in tokenizer training. Several studies that applied BPE and sentencepiece use a merged bilingual corpus, that combines two language corpora into one, when training its tokenizer (Song et al., 2019; Liu et al., 2020). However in these studies, merged bilingual corpus is utilized without sufficient comparative analysis.

In this study, tokenization methods which leveraging merged bilingual corpora and separate bilingual corpora are denoted as Vocabulary Communicating (VC) and Vocabulary Separating (VS), respectively. We denote VC and VS as vocabulary methods and compare the performance of each method in NMT. In other words, we further figure out the optimal tokenization method through comparative experiments on various tokenization methods.

All the experiments are made on a Korean dataset, which is a representative of low-resource and morphologically rich language (MRL). In particular, we propose ONE-Piece that combines the VC method and morphological segmentation followed by sentencepiece. Through comparative experiments with tokenization methods currently used in NLP research, such as BPE and sentencepiece, we revealed that ONE-Piece can encourage the optimal performance in Korean-English machine translation. The contributions of our study are as follows:

- We proposed a new subword tokenization method, ONE-Piece, which leveraging morphological segmentation and vocabulary communicating method. Through ONE-Piece, we can obtain better performance than the existing tokenization methods such as BPE and sentencepiece.
- Based on linguistic analysis, we showed that constructing corpus for training tokenizer is an important factor that has a critical influence on machine translation performance.
- We presented a new viewpoint for pre-processing that can improve translation performance without modifying model structure. Our proposal considered industrial service and demonstrated high speed and performance without using PFA.

2 Proposed Method

This study proposes an optimal tokenization method for improving machine translation performance from the viewpoints of morphological segmentation and vocabulary method. We derive an optimal tokenization method for Korean-English machine translation by conducting a case study that combines the morphological segmentation and vocabulary method.

2.1 Morphologically-Aware SentencePiece

Korean is classified as an agglutinative language according to its type of morphemes. Due to the nature of agglutinative languages, one word can comprises substantive (noun/pronoun/numeral) followed by postposition, or the stem followed by the ending. Table 1 shows the result of tokenizing Korean sentences through BPE (Sennrich et al., 2015), sentencepiece (Kudo and Richardson, 2018), and morphological segmentation using MeCab-ko.

In the case of BPE and sentencepiece, the postpositions '가 (ga), 는 (neun), 를 (leul), 의 (ui), 인 (in)' have not been properly separated from the substantives. This failures in separating the postpositions from the substantives can lead to mistranslation of entities and grammartically incorrect translation. Generally, the postposition indicates the grammatical relationship to the substantive and plays an important role in organizing the meaning of words. Therefore, miss-separating the postpositions can lead to the incorrect translation of the whole sentence, and misunderstanding of the semantic relationship.

Also, in the case of BPE and sentencepiece, the entities (red-common noun, blue-proper noun) are over-tokenized. Both methods tokenize sentences based on frequency and probability without considering linguistic characteristics. This can lead to inappropriate segmentation between substantives and postpositions, or between stems and endings. These problems can be alleviated by employing morphological segmentation. In this study, we quantitatively analyze the effect of morphological segmentation in NMT, and propose the optimal method of leveraging it by combining sentencepiece.

2.2 Why MeCab-ko?

We use Konlpy (Park and Cho, 2014) for morphological segmentation of Korean sentences. Konlpy

Target Sentence	BPE	sentencepiece	MeCab-ko
The number of diagnoses started to soar, just as Lorna and Judith predicted, indeed hoped, that it would	진단/ 숫자는/ 급증@@/했고/ **로@@/나**와/ **주@@/디@@/스**가/ 예상@@/했고/ **진@@/실**로/ 그들이/ 바랬@@/던/ 것처럼	_진단/_숫자는/_급증/했고/**_로/나**와/**_주/디/스**가/_예상/했고/**_진/실**로/_그들이/_바/랬/던/_것처럼’	진단/ 숫자/는/ 급증/했/고/ **로나(NNP)**/와/ **주디스(NNP)**/가/ 예상/했/고/ **진실**로/ 그/들/이/ 바랬/던/ 것/처럼
Instead of blaming parents for causing autism, Asperger framed it as a lifelong, polygenetic disability	**자폐@@/증**을/ 부모의/ 탓@@/으로/ 돌리는/ 대신/ **아스@@/퍼@@/거**는/ 그것을/ 장기적인/ 다@@/**기@@/원**의/ 장애@@/로	**_자폐/증**을/_부모의/_탓/으로/_돌리는/_대신/**_아스/퍼/거**는/_그것을/_장기적인/_다/**기/원**의/_장애/로	**자폐증**/을/ 부모/의/ 탓/으로/ 돌리/는/ 대신/ **아스퍼거(NNP)**/는/ 그것/을/ 장기/적/인/ 다/**기원**/의/ 장애/로

Table 1: Comparison of BPE, sentencepiece and MeCab-ko segmentation results.

is an open-source Korean morphological analyzer package which provides 6 morphological analyzers: MeCab-ko, Kkma, Komoran, Hannanum, Okt, and Twitter. In this study, we select an analyzer that shows the best performance among them by experimenting morphological analysis for up to 1 M characters. In particular, since inference speed is a very important factor in the industry field, we focused on the time required for morphological analysis. The inference time required for each analyzer is shown in Figure 1.

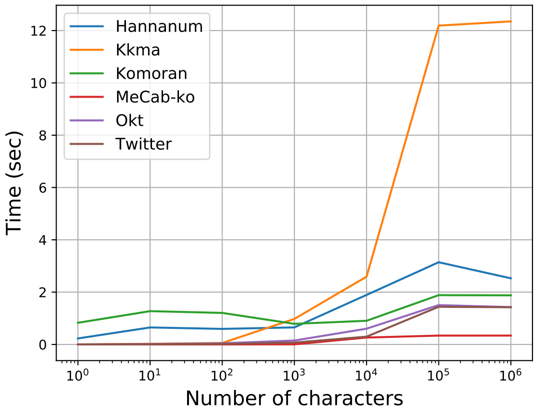

Figure 1: Inference time of morphological analyzer

As shown in Figure 1, MeCab-ko shows the best results compared to other morphological analyzers. It takes 0.3353 secs in processing 1 M characters. Additionally, through experiments on different number of characters, we can see that MeCab-ko conducts analysis of the input sequence at a stable speed despite the exponential increase in the number of characters. For these reasons, we adopt MeCab-ko by its high processing speed and stability in character length.

2.3 Vocabulary Communicating Method

The VC method has been used in several PFA-based models. In MASS (Song et al., 2019), a 60K vocabulary was extracted by composing the source and target language into a merged bilingual corpus. In mBART (Liu et al., 2020), the CC25 corpus was composed of a total of 25 languages extracted from CommonCrawl (CC) (Lample and Conneau, 2019; Wenzek et al., 2019) and used for unified vocabulary extraction. When using the VC method in mBART, there is a generalization effect for unseen languages. However, this effect has not been sufficiently discussed for languages that do not share an alphabet, and no quantitative basis for a generalization effect has been proposed. In this study, we conducted probing for this approach through quantitative analysis.

In practical cases, source and target languages often communicate to each other; source language is contained in target sentences, and vice versa. In the case of our training data, approximately 6.9% of source sentences contains English tokens. For instance, domain specific terms such as "Host IP" can not be replaced by Korean token and constitute Korean sentences in its original form.

For the case of VS method, each language only contributes to the processing of corresponding language corpus, and different tokenizers are applied to the source and target sentences. If a vocabulary is extracted according to the VS method, source language dictionary is composed by reflecting only small fraction of the target languages, which is con-

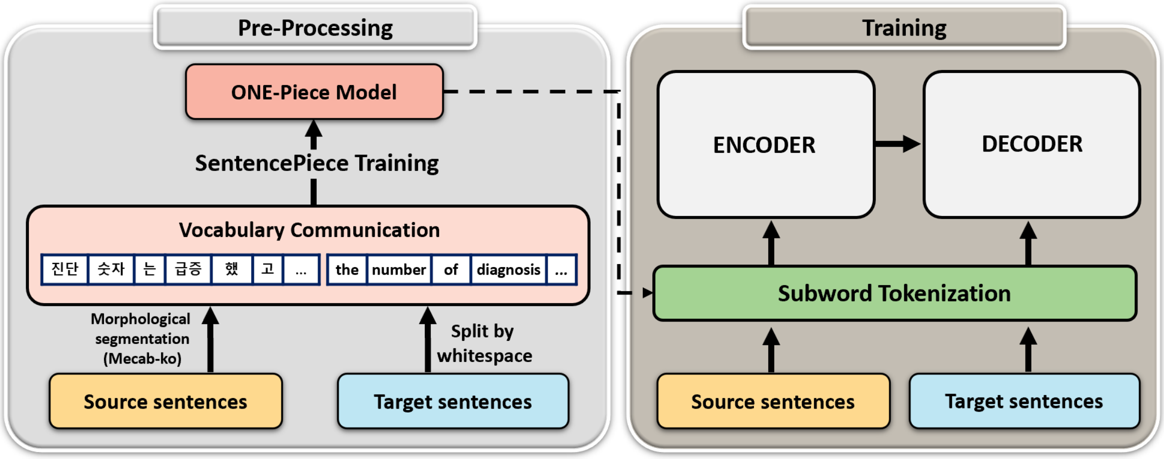

Figure 2: Overall Architecture of NMT training process using ONE-Piece model

tained in source sentences. In this case, target language token, which is not contained in source language dictionary but contained in target language dictionary, is treated as unknown.

The VC method can alleviate this problem. As previously mentioned, the VC method construct a merged corpus and the vocabulary extracted from this merged corpus is identically applied to the source and target sentences. By using VC method, the source and target language can interact within the same vocabulary and are mutually referenceable. Therefore, the source and target language can interact within the same vocabulary and are mutually referenceable. This can lead to full understanding of target language tokens in source sentences and vice verssa.

2.4 ONE-Piece

ONE-Piece is a subword tokenization method that utilizes morphological analysis and the VC method. By applying morphological analysis, characteristics of an agglutinative language, that a single word can comprises multiple morphemes, can be considered. Then by following sentencepiece, applying VC method, can alleviate the out of vocabulary (OOV) problem.

The ONE-piece can be obtained by following processes. First, from a parallel corpus P, which is consist of source sentences $S = \{S_i\}_{i=1}^{N}$ and target sentences $T = \{T_i\}_{i=1}^{N}$, merged corpus M is created. More specifically, this procedures can be described as follows:

$$\begin{aligned} S_i &= \{s_i^j\}_{j=1}^{n_i} \\ T_i &= \{t_i^j\}_{j=1}^{m_i} \end{aligned} \tag{1}$$

s_i^j denote j^{th} word of source sentence S_i, which is segmented by whitespace, and n_i indicate the word length of S_i. Similarly, t_i^j denote j^{th} word, and m_i indicate the word length of target sentence T_i, which is segmented by whitespace.

We apply morphological analyzer to agglutinative language. In this paper, source sentences is re-segmented by morpheme-units, through morphological analyer. This can be denoted as equation (2).

$$Seg_i = MA(S_i) = \{seg_i^j\}_{j=1}^{k_i} \tag{2}$$

MA indicates morphological analyzer for source language. By MA, morpheme-unit-segmented sentence Seg_i is generated from source sentence S_i. k_i denotes morpheme-token length of Seg_i. Since a word comprises one or more morphemes, k_i is always equal to or greater than j_i. Then by combining all the Seg_i and T_i into one, merged corpus M is generated as equation (3).

$$M = [T_1, \ldots, T_N, Seg_1, \ldots, Seg_N] \tag{3}$$

M is composed of both source language and target language. As M is created, we can generate ONE-piece by training sentencepiece model by M.

Figure 2 is an overall architecture that describes the process of training NMT model by leveraging ONE-Piece. For Korean sentences in the source part, morphological segmentation is performed with MeCab-ko, and English sentences corresponding to the target side are segmented by whitespace. After combining source sentences and target sentences, we train sentencepiece model by using them. In this process, ONE-Piece model is created. Through ONE-Piece, input sentences are segmented into subwords and fed into the encoder and decoder for training NMT model.

3 Experiments

3.1 Dataset and Experimental Setting

We utilized Korean-English parallel corpora from 3 different data sources for our dataset: the AI Hub Korean-English parallel corpus[1], OpenSubtitles[2], and the IWSLT-17 TED corpus (Cettolo et al., 2017). We constructed 2.7 M sentence pairs from these data sources. For better NMT performance, we applied parallel corpus filtering to our corpus and construct 2.2 M sentence pairs for training. We applied the same filtering method as Park et al. (2020a). We randomly selected 5,000 sentence pairs from our training data for validation and used IWSLT-16 and IWSLT-17 test sets, which is consist of 1,143 and 1,429 sentence pairs, for performance evaluation.

Since our ultimate purpose is to check whether the performance of the NMT model can be improved only by the subword tokenization method without changing the model, we adopt vanilla transformer as our baseline. The performance evaluation of translation results was conducted based on the BLEU score (Papineni et al., 2002). To measure the score, we adopted multi-bleu.perl script[3] in Moses.

3.2 Experimental Results

3.2.1 Verification of the Effectiveness of the VC Method

In this section, we experimentally compare and verify the performance of Korean-English machine translation using VC and VS methods. By applying each method to BPE and sentencepiece, we investigate the impact of the vocabulary method in the performance of NMT. The experimental results are shown in Table 2.

In sentencepiece, the VC method outperforms the VS method by 1.34 BLEU score on the IWSLT-16 test set and 0.99 BLEU score on the IWSLT-17 test set. Conversely for BPE, the VS method outperforms the VC method by 2.78 BLEU score on the IWSLT-16 test set and 2.42 BLEU score on the IWSLT-17 test set. There are some cases where the VS method yields a more superior performance than the VC method, depending on the tokenization algorithm. In other words, the VC method does not show consistently superior performance to the VS method.

Tokenization Method	IWSLT-16 (BLEU)	IWSLT-17 (BLEU)
VC SP	21.63	19.11
VS SP	20.29	18.12
VC BPE	17.47	15.42
VS BPE	20.25	17.84

Table 2: Korean-English NMT results applying different vocabulary method in BPE and sentencepiece. SP refers to sentencepiece.

Currently, many studies have employed the VC method based tokenizer as a default choice, regardless of the tokenization algorithm. From this experiment, we revealed that the current default option may not be the optimal choice depending on the selection of the tokenization algorithm. We further show that selecting vocabulary method is an important factor that significantly affects machine translation performance. This indicates that the vocabulary method must be considered when adopting a tokenization algorithm to ensure the optimal machine translation performance.

3.2.2 Verification of the Effectiveness of Morphological Segmentation

In this section, we verify the impact of the morphological segmentation. We experimented two tokenization methods using MeCab-ko in Korean corpus. The first method is to segment by morpheme units, and the second method is to add sentencepiece after this process, as first suggested by Park et al. (2019). Whereas Park et al. (2019) used VS method based tokenizers in all of their experiments, we utilized VS method based tokenizers for this experiment. Our results are shown in Table 3.

Tokenization Method	IWSLT-16 (BLEU)	IWSLT-17 (BLEU)
VS SP	20.29	18.12
VS MeCab-ko	19.61	17.08
VS MeCab-ko+SP	19.78	17.49

Table 3: Korean-English NMT results using MeCab-ko. All experiments are implemented using the VS method. sentencepiece is denoted as SP.

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

[2] http://opus.nlpl.eu/OpenSubtitles-v2018.php

[3] https://github.com/moses-smt/mosesdecoder/blob/master/scripts/generic/multi-bleu.perl

Applying sentencepiece after morphological segmentation demonstrates better performance in both the IWSLT-16 and IWSLT-17 test sets compared to the MeCab-ko based segmentation without sentencepiece. However, our results show that applying morphological segmentation for tokenizer training yields overall performance degradation in both test sets. This is contrary to the experimental results of Park et al. (2019), which claim that morphological analysis consistently improves machine translation performance. The main difference between our experiment and Park et al. (2019) is the vocabulary method. From these results, we can infer that the effect of applying morphological segmentation on NMT is relatively different depending on the vocabulary method. This indicates that prior to applying morphological segmentation, the vocabulary method must be considered to get improved NMT performance.

3.2.3 Verification of the ONE-Piece

ONE-Piece differs from existing tokenizers in that it utilizes VC method and the morphological segmentation followed by sentencepiece. In this section, we verify the effectiveness of ONE-Piece by comparing NMT performance using various preprocessing strategies based on the VC method. The results are shown in Table 4.

Tokenization Method	IWSLT-16 (BLEU)	IWSLT-17 (BLEU)
VC Word	7.98	7.16
VC Character	16.39	17.06
VC BPE	17.47	15.42
VC sentencepiece	21.63	19.11
ONE-Piece (ours)	24.95	22.58

Table 4: Korean-English NMT results of different tokenization algorithms. All the experiments are implemented using the VC method.

Compared to the VC-based tokenizer, ONE-Piece produces at least 3.32 BLEU score superior translation performance. This result suggests that further improvement can be made by applying ONE-Piece to other existing sentencepiece-based NMT models.

In sections 3.2.1 and 3.2.2, we revealed that vocabulary method and morphological segmentation significantly affect the NMT performance, but neither of these consistently improve the NMT performance by themselves. However as shown in table 4, by properly combining these two factors, we can derive mutual supplementation effect which lead to a meaningful improvement in the translation performance. This can be viewed as the new criteria for constructing corpus for training tokenizer.

3.2.4 Comparison with Existing Studies

We compare the performance of vanilla transformer model applying ONE-Piece with the performance of mBART(Liu et al., 2020). mBART was trained with 610 M params and 5.6 B tokens from the CC corpus. mBART utilized morpheme based segmentation using MeCab-Ko in the Korean corpus and applied sentencepiece in the English corpus, which is the same tokenization method as VS MeCab-ko in Table 3.

	mBART	MeCab-ko	ONE-Piece
IWSLT-17 (BLEU)	24.6	17.08	22.58
model parameter	610M	32M	32M

Table 5: Comparison of proposed ONE-Piece model with mBART.

As shown in Table 5, when the same tokenization method used in mBART was applied to the baseline model, the performance was 7.52 BLEU lower than that of mBART. However, by applying ONE-Piece to the baseline model, the performance difference narrowed to a 2.02 BLEU score. This shows that applying ONE-Piece enables the vanilla transformer model to have similar performance to the SOTA model. Although the baseline model using ONE-Piece did not exceed the performance of mBART, it is a notable result considering that the number of parameters required by the baseline model is 32 M, approximately 5% of the number of parameters compared to mBART.

The significance of this experiment is that simply by changing the tokenization method, a model with a small number of parameters can achieve a similar performance to SOTA model, which is trained with a more advanced algorithm and larger number of parameters.

4 Conclusion

In this study, we proposed a new tokenization method called ONE-Piece. This can provide the best performance in Korean-English machine translation compared with other tokenization methods.

Our results quantitatively confirmed the effect of the vocabulary method and morphological segmentation on NMT performance. Furthermore, we experimentally proved that the VC method and morphological segmentation cannot consistently improve the performance of NMT by themselves. Our results showed that significant and consistent performance improvement can only be achieved in NMT if they are properly used together. By using ONE-Piece, the vanilla transformer model shows comparable translation performance to the mBART. Accordingly, we expect that companies that have difficulties using the latest PFA-based model, due to an inadequate server environment, will be able to utilize our proposed model to provide sufficiently good performance.

Acknowledgments

This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2018-0-01405) supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation), Institute for Information & communications Technology Planning & Evaluation (IITP), grant funded by the Korean government (MSIT) (No. 2020-0-00368, A Neural-Symbolic Model for Knowledge Acquisition and Inference Techniques) and MSIT(Ministry of Science and ICT), Korea, under the ICT Creative Consilience program(IITP-2021-2020-0-01819) supervised by the IITP(Institute for Information & communications Technology Planning Evaluation).

References

Tamali Banerjee and Pushpak Bhattacharyya. 2018. Meaningless yet meaningful: Morphology grounded subword-level nmt. In *Proceedings of the Second Workshop on Subword/Character LEvel Models*, pages 55–60.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Mauro Cettolo, Marcello Federico, Luisa Bentivogli, Niehues Jan, Stüker Sebastian, Sudoh Katsuitho, Yoshino Koichiro, and Federmann Christian. 2017. Overview of the iwslt 2017 evaluation campaign. In *International Workshop on Spoken Language Translation*, pages 2–14.

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

Anna Currey, Antonio Valerio Miceli-Barone, and Kenneth Heafield. 2017. Copied monolingual data improves low-resource neural machine translation. In *Proceedings of the Second Conference on Machine Translation*, pages 148–156.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Philipp Koehn, Huda Khayrallah, Kenneth Heafield, and Mikel L Forcada. 2018. Findings of the wmt 2018 shared task on parallel corpus filtering. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 726–739.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. *arXiv preprint arXiv:1804.10959*.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808.06226*.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *arXiv preprint arXiv:2001.08210*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Santanu Pal, Sudip Kumar Naskar, Mihaela Vela, and Josef van Genabith. 2016. A neural network based approach to automatic post-editing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 281–286.

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, Gyeongmin Kim, and HeuiSeok Lim. 2019. Parallel corpus filtering and korean optimized subword tokenization for machine translation. In *The 31st Annual Conference on Human Cognitive Language Technology*, pages 221–224.

Chanjun Park, Yeonsu Lee, Chanhee Lee, and Heuiseok Lim. 2020a. Quality, not quantity? : Effect of parallel corpus quantity and quality on neural machine translation. In *The 32st Annual Conference on Human Cognitive Language Technology*, pages 363–368.

Chanjun Park, Yeongwook Yang, Kinam Park, and Heuiseok Lim. 2020b. Decoding strategies for improving low-resource machine translation. *Electronics*, 9(10):1562.

Eunjeong L. Park and Sungzoon Cho. 2014. Konlpy: Korean natural language processing in python. In *Proceedings of the 26th Annual Conference on Human Cognitive Language Technology*, Chuncheon, Korea.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2019. Mass: Masked sequence to sequence pre-training for language generation. *arXiv preprint arXiv:1905.02450*.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2019. Ccnet: Extracting high quality monolingual datasets from web crawl data. *arXiv preprint arXiv:1911.00359*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems*, pages 5753–5763.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. *arXiv preprint arXiv:1604.02201*.