When is Char Better Than Subword: A Systematic Study of Segmentation Algorithms for Neural Machine Translation

Jiahuan Li [*] **Yutong Shen** [*] **Shujian Huang** [†] **Xinyu Dai** **Jiajun Chen**
National Key Laboratory for Novel Software Technology, Nanjing University, China
{lijh,shenyt}@smail.nju.edu.cn,
{huangsj,daixinyu,chenjj}@nju.edu.cn

Abstract

Subword segmentation algorithms have been a *de facto* choice when building neural machine translation systems. However, most of them need to learn a segmentation model based on some heuristics, which may produce suboptimal segmentation. This can be problematic in some scenarios when the target language has rich morphological changes or there is not enough data for learning compact composition rules. Translating at fully character level has the potential to alleviate the issue, but empirical performances of character-based models has not been fully explored. In this paper, we present an in-depth comparison between character-based and subword-based NMT systems under three settings: translating to typologically diverse languages, training with low resource, and adapting to unseen domains. Experimental results show strong competitiveness of character-based models. Further analyses show that compared to subword-based models, character-based models are better at handling morphological phenomena, generating rare and unknown words, and more suitable for transferring to unseen domains.

1 Introduction

Neural machine translation (NMT) has achieved great success in recent years. Modern NMT systems typically operate on subword level, using segmentation algorithms such as *byte pair encoding* (BPE) (Sennrich et al., 2016) or Morfessor (Creutz and Lagus, 2002). Compared to word-level models, subword segmentation helps overcome the out-of-vocabulary (OOV) problem and make better use of morphological information in the surface form.

Despite their empirical effectiveness, subword algorithms may produce improper segmentation due to their data-dependent nature. NMT models are typically robust to such errors when trained on large corpora or the target language is regular in morphological changes, like French or German. However, the problem will arise when such conditions are not met, i.e. there is not enough data for learning compact composition rules or the target language is morphologically rich and complex.

An alternative segmentation choice is to use fully character-level (CHAR) models (Lee et al., 2017; Cherry et al., 2018; Gupta et al., 2019; Gao et al., 2020; Banar et al., 2020), which has the potential to alleviate above issues. CHAR does not need to *learn* any segmentation rules and keeps all available information in the surface form, avoiding the risk of information loss due to improper segmentation. What is more, the main pain point of CHAR that it takes too long to train is less obvious in above settings since there is not as much data as in the rich resource setting. However, there has not been a comprehensive study in these settings.

In this paper, we conduct a systematic comparison between CHAR and other subword algorithms, e.g. BPE and Morfessor. Experiments show strong competitiveness of CHAR under three settings: translating to typologically diverse languages (Section 2), training with low resource (Section 3), and adapting to distant domains (Section 4). Further analyses show that compared to subword algorithms, the benefits of CHAR mainly come from better capture of the morphological phenomena, better generation of rare and unknown words, and better translation of domain-specific words.

2 Translation Across Typologically Diverse Languages

Human languages are known to exhibit diverse morphological phenomena, which could serve as a principle to classify languages into different morphological categories, such as *fusional*, *agglutinative*, *introflexive* and *isolating*. While previous

[*] Equal contribution
[†] Corresponding author

Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Short Papers), pages 543–549
August 1–6, 2021. ©2021 Association for Computational Linguistics

		Word	Char	BPE	Morf.
F.	Fr	39.1/.580	40.1/.589	**41.2/.597**	39.6/.592
	Ro	31.1/.487	**33.9/.526**	32.9/.517	30.6/.517
A.	Fi	21.9/.412	**23.5/.487**	22.3/.472	21.7/.466
	Tr	19.8/.396	**22.8/.456**	21.1/.440	16.9/.437
In.	Hi	14.0/.262	**15.6/.290**	14.8/.285	14.8/.276
	Ar	22.5/.451	**24.7/.491**	23.9/.481	23.5/.481
Is.	Vi	21.6/.374	**22.5/.385**	22.2/.381	21.1/.373
	Ml	22.9/.324	**25.0**/.349	24.3/.347	24.1/**.356**

Table 1: BLEU/chrF3 scores of systems translating from English to languages of different morphological categories, using different segmentation algorithms. Best score in each line is shown in bold.

works only focus on performances of character-level models when translating to fusional and agglutinative languages (Gupta et al., 2019; Libovický and Fraser, 2020), we conduct a comprehensive study covering all four morphological categories.

2.1 Experiment Setup

Dataset We consider the translation from English to eight target languages representing four morphological categories, i.e. French (**Fr**) and Romanian (**Ro**) for *fusional*, Finnish (**Fi**) and Turkish (**Tr**) for *agglutinative*, Hebrew (**He**) and Arabic (**Ar**) for *introflexive*, and Vietnamese (**Vi**) and Malaysian (**Ml**) for *isolating*. We use OPUS-100 corpus[1] (Tiedemann, 2012), which consists of 1M parallel sentences for each language pair.

Model and Hyperparameters We use the Transformer architecture (Vaswani et al., 2017) throughout all experiments. To ensure results' reliability , we run an exhaustive search of hyperparameters including batch size and learning rate. Detailed hyperparameters can be found in Appendix A.

2.2 Results

The results are listed in Table 1. We can see that CHAR outperforms other algorithms in 7 out of 8 languages in terms of BLEU (Papineni et al., 2002) and chrF3 (Popović, 2015), showing strong competitiveness of CHAR's ability across languages. The only exception is the En-Fr language pair, which are known to be quite similar and is beneficial for BPE to learn a joint segmentation model.

It is intuitive that BPE and Morfessor cannot outperform CHAR on introflexive languages (Hi, Ar). Introflexive languages follows non-concatenative morphology (McCarthy, 1981), i.e. grammatical information is conveyed by directly modifying the root words. This makes it hard for linear segmenting methods such as BPE and Morfessor to work well. This finding is also consistent with previous research on other tasks (Zhu et al., 2019).

For isolating languages (Vi, Ml), there are rare morphological phenomena indicating grammatical relations, so segmentation algorithms do not greatly affect the performance. We can see that the two open-vocabulary segmentation algorithms (CHAR, BPE) show comparable performances.

Surprisingly, even for highly agglutinative languages such as Finnish and Turkish, which has very regular morphological changes by adding affixes or suffixes, CHAR still achieves better performance.

	Word	Char	BPE	Morf.
Comp. adj.	55.6	**70.8**	63.0	60.0
Det. poss.	49.6	**83.0**	78.0	78.4
Pron. hum	60.6	**67.0**	66.2	66.6
Local case	36.6	**61.8**	50.6	47.6
Pron. gender	73.6	76.6	**79.0**	**79.0**
Verb neg	96.6	97.2	**98.4**	98.0
Preposition	33.8	**69.2**	60.2	64.2
Future tense	51.4	43.8	**53.8**	50.8
Past tense	83.2	**91.8**	87.4	90.8
Pron. plural	74.6	**79.2**	77.4	75.2
Noun plural	48.8	**76.0**	62.8	60.8
Det. definite	38.4	38.8	40.8	**44.8**
Named Ent.	9.2	**70.4**	66.4	30.2
Number	65.4	**96.6**	91.2	77.8

Table 2: Performance of different segmentation algorithms on the MorphEval En-Fi benchmark. Each row represents a kind of morphological phenomenon.

2.3 Analysis on MorphEval

To understand where the advantages of CHAR model come from, we take Finnish as an example and evaluate the morphological competence of different models using MorphEval test suites (Burlot et al., 2018). MorphEval generates pairs of source sentences that differ by one kind of morphological phenomena, and assesses a MT system's ability by computing the percentage of its generated target sentences that convey as the source sentences. Higher accuracy means the model is more sensitive to the current morphological phenomenon.

As shown in Table 2, CHAR performs the best in 10 out of 14 tests. Among these 10 tests, in comparative adjectives, possessive determiner, local postposition case, preposition case, plural nouns, CHAR surpasses other algorithms notably by at least 5% accuracy. This indicates CHAR's strong ability to capture the fine-grained morphological

[1] http://data.statmt.org/opus-100-corpus/v1.0/supervised/

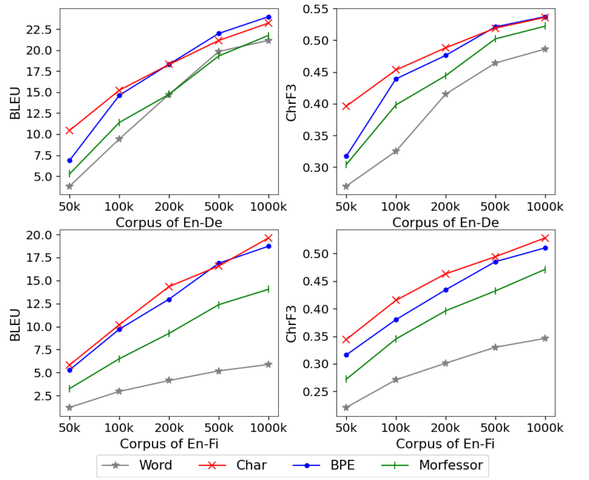


Figure 1: BLEU and chrF3 score curves using different amount of parallel data for training En-De (above) and En-Fi (below) translation systems.

phenomena, which is crucial for MT models when translating into morphologically rich languages.

Interestingly, three of four morphological phenomena on which CHAR falls behind are so-called *stability* features (Burlot et al., 2018), which are expressed differently in the source language but should be expressed identically in the target language[2]. The disadvantage of CHAR in this kind of phenomena shows CHAR-based model may be less robust to lexical changes to source-side changes, and the reason needs to be further researched.

3 Translation with Low Resource

Subword algorithms help alleviate the OOV problem. However, most of them are based on heuristics and may produce wrong segmentation. While this problem is not so evident when there is enough data to learn robust composition rules, in low-resource setting it could be a different story and their effectiveness should be examined. While for CHAR, pure character sequences can directly provide all the information to the model for learning the composition rules. Therefore a prudent choice of segmentation should be studied in this setting.

3.1 Experiment Setup

We perform evaluation on WMT14 En-De[3] and WMT17 En-Fi[4] dataset. Datasets of size 50k, 100k, 200k, 500k and 1000k are subsampled from the original training dataset and serve as training data of different resource conditions. For validation and test, we use the original development and test split.

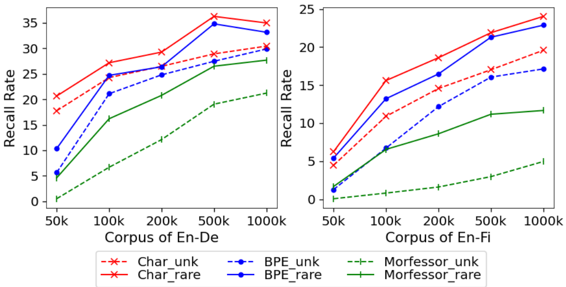


Figure 2: Recall rates of unknown and rare words generated by systems based on different tokenizers models. Words appearing no more than 5 times in the training set are considered as rare words.

Previous works (Sennrich and Zhang, 2019; Nguyen and Chiang, 2017) show that in low resource settings the evaluation results can be sensitive to model size (e.g. hidden dimension, layer number) and the number of BPE merges k, so we run an additional search of hidden dimension, layer number and k, and report the best results in this section. See Appendix A for details.

3.2 Results

We evaluate models with BLEU and chrF3. The results are showed in Figure 1. In general, the performances of CHAR and BPE are on par, and are better than Word and Morfessor. In different data conditions, the results varies.

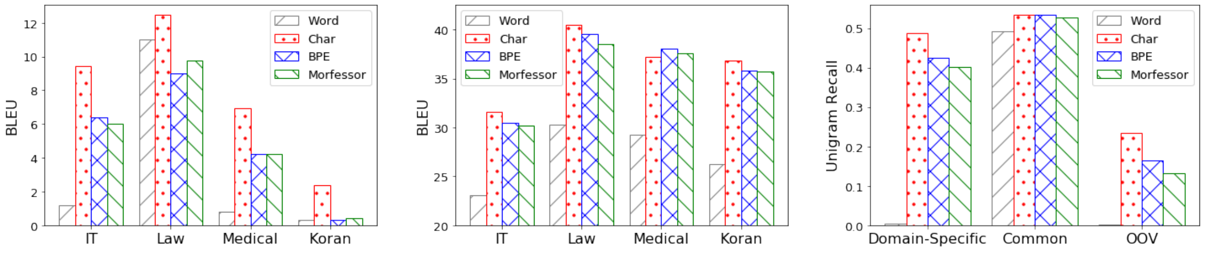
medium-resource When there are plenty resources, e.g. 500k and 1000k, the performance of CHAR and BPE are comparable but different for different language pairs. For En-Fi, CHAR is better than BPE. It is because morphological changes in Finnish are quite complex. More fine-grained segmentation like CHAR is needed to learn corresponding rules. Conversely, German's morphological changes are so regular that BPE can learn most of merging rules, making them performing better.

low-resource When the corpus size is 50k to 200k, CHAR performs the best among four segmentation methods. BPE and Morfessor usually regard frequently occurring words as single tokens, many of which contain rich morphological information. This, together with the improper segmentation problem, prevents NMT models from learning correct composition rules, damaging the model's generalization ability on rare and unknown words. In low resource setting this problem would be more se-

[2] For example, English uses *he/she* to convey the masculine/feminine contrast, but Finnish uses the same pronoun *hän* regardless of the gender of the antecedent.

[3] http://www.statmt.org/wmt14/translation-task.html

[4] http://statmt.org/wmt17/translation-task.html



(a) Average OOD BLEU (No Adapt). (b) Average OOD BLEU (Finetune). (c) Recall of different word types.

Figure 3: Domain robustness of translation systems based on different segmentation algorithms.

vere, since there are much more rare and unknown words but not enough data for learning compact composition rules.

Compared with subwords, character-based models learn combinations directly from character sequences. Not limited to fixed char sequence patterns in subwords, more words with different morphological changes can be generated through CHAR. Therefore, CHAR can learn more correct composition rules than subword-based model, leading to better translation of rare and unknown words.

3.3 Analysis on Rare and Unknown Words

To further support the above analysis, we evaluate the translation quality of rare and unknown words by calculating their recall rates. The results are showed in Figure 2. We can see that CHAR has achieved the highest recall rates of rare and unknown words. Although, as the resource increases, the gap between CHAR and BPE is shrinking gradually, the results can still prove that CHAR can capture more morpheme information, performing better at generating rare and unknown words.

4 Translation Across Distant Domains

Domain robustness (Müller et al., 2020), which refers to models' generalization ability on unseen domains, is important for NMT applications. However, subword algorithms need to learn segmentation rules from a given corpus, which may be domain-specific. When applied to a new domain, they may improperly segment target-domain specific words, hurting the domain robustness. In contrast, CHAR does not suffer from the issue. In this section, we investigate how different segmentation algorithms affect NMT models' domain robustness.

4.1 Experiment Setup

We use the same corpora as (Koehn and Knowles, 2017), which is a De-En dataset covering subsets of four domains: *Law*, *Medical*, *IT* and *Koran*. Following Koehn and Knowles (2017), each time we train a source domain model on one of four subsets and report results on test sets of the other three domain. We experiment in two settings: **No Adapt** and **Finetune**. The first one involves no target domain data, while the latter uses randomly sampled 100k sentence pairs from target domain data to finetune the source domain model.

4.2 Results

We report the average out-of-domain (OOD) BLEU scores of NMT systems based on different segmentation algorithms in Figure 3a and Figure 3b. As can be seen from the figure, CHAR surpasses other algorithms in almost all settings, except when finetuning from Medical to others. This illustrates the suitability of CHAR for domain robustness, especially when there is no enough data for adaptation.

4.3 Analysis on Different Types of Words

To understand the advantages of CHAR, we take the setting of finetuning from *IT* to *Medical* as an example and analyze performances on different types of words. Specifically, we divide words in the test set into three types: **(1) Domain-specific** words occur only in the target domain training data; **(2) Common** words occur in both the source and target domain training data; **(3) OOV** words *do not* occur in both training data.

The result can be seen in Figure 3c. CHAR achieves better performance on OOV words, which is consistent with findings in Section 3. While performances of CHAR and subword-based algorithms are on par on common words, CHAR outperforms the others by a large margin on domain-specific words. This suggests that the advantage of CHAR mainly comes from the correct translation of domain-specific and OOV words, which may be segmented improperly by subword algorithms.

	Word	Char	BPE	Morf.	BPE-D
No adapting	11.03	**12.46**	9.02	9.74	11.11
Finetune	30.26	**40.53**	39.53	38.49	40.26

Table 3: Average OOD BLEU of models based on different subword algorithms when adapting from *Law* to other domains. BPE-D: BPE-dropout (Provilkov et al., 2020)

4.4 Comparison with Advanced Segmentation Algorithms

Although we focus on deterministic segmentation algorithms in this paper, there are more advanced ones such as BPE-dropout (Provilkov et al., 2020) and subword regularization (Kudo, 2018), which produce multiple segmentation candidates when training and show improved performance. Therefore, we also conduct experiments comparing CHAR with BPE-dropout in terms of domain adaptation performance. We take the setting of adapting from *Law* to other domains and report results in Table 3. As can be seen, although BPE-dropout surpasses BPE by a large margin, CHAR still achieves the best performance, which again shows the superiority of CHAR.

5 Related Work

Character-level neural machine translation has received growing attention in recent years. Lee et al. (2017) first propose a fully character-level NMT model based on recurrent encoder-decoder architecture and convolutional layers, which shows a promising results. Gao et al. (2020) propose to incorporate convolution layers in the more advanced Transformer architecture and show their model can learn more robust character-level alignments.

However, translating at character level may incur significant computational overhead. Therefore, later works on character-level NMT (Cherry et al., 2018; Banar et al., 2020) mainly focus on reducing computation cost of them. Cherry et al. (2018) show that by employing source sequence compression techniques, the quality and efficiency of character-based models can be properly balanced. Banar et al. (2020) share the same idea as Cherry et al. (2018) but build their models using Transformer architecture. Our work differs from theirs in that we aim to analyze the performance of existing models instead of exploring novel architectures.

There are also several researches on comparison between CHAR and other subword algorithms (Durrani et al., 2019; Gupta et al., 2019). Durrani et al. (2019) compare character-based models and subword-based models in terms of representation quality, and find that representation learned by the former are more suitable for modeling morphology, and more robust to noisy input. Gupta et al. (2019) investigate the performance of different segmentation algorithms when using Transformer architecture, and find that character-based models can achieve better performance when translating noisy text or text from a different domain. Our finds are consistent with them, yet we conduct a more large-scale and in-depth analysis by covering language pairs from more language families and explaining where the advantage of character-based models comes from.

6 Conclusion

We conduct a comprehensive study and show advantages of CHAR over subword algorithms in three settings: translating to typologically diverse languages, translating with low resource, and adapting to distant domains. Note that although we have tried our best to take as much language pairs as possible into consideration, there are certainly a lot of languages remaining uncovered in this paper. However, we believe our experimental results can serve as an evidence of character-based NMT models' strong competitiveness. We hope more attention will be drawn to them, including exploring their more benefits and reducing the possibly higher computation cost in practice.

Acknowledgments

We would like to thank the anonymous reviewers for their insightful comments. Shujian Huang is the corresponding author. This work is supported by the National Key R&D Program of China (No. 2019QY1806), National Science Foundation of China (No. 61772261, U1836221). This work is also partially supported by the research funding from ZTE Corporation.

References

Nikolay Banar, Walter Daelemans, and Mike Kestemont. 2020. Character-level transformer-based neural machine translation. *arXiv preprint arXiv:2005.11239*.

Franck Burlot, Yves Scherrer, Vinit Ravishankar, Ondřej Bojar, Stig-Arne Grönroos, Maarit Koponen, Tommi Nieminen, and François Yvon. 2018. The wmt'18 morpheval test suites for english-czech, english-german, english-finnish and turkish-english.

In *3rd Conference on Machine Translation (WMT 18)*, volume 2, pages 550–564.

Colin Cherry, George Foster, Ankur Bapna, Orhan Firat, and Wolfgang Macherey. 2018. Revisiting character-based neural machine translation with capacity and compression. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4295–4305, Brussels, Belgium. Association for Computational Linguistics.

Mathias Creutz and Krista Lagus. 2002. Unsupervised discovery of morphemes. In *Proceedings of the ACL-02 Workshop on Morphological and Phonological Learning*, pages 21–30. Association for Computational Linguistics.

Nadir Durrani, Fahim Dalvi, Hassan Sajjad, Yonatan Belinkov, and Preslav Nakov. 2019. One size does not fit all: Comparing NMT representations of different granularities. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1504–1516, Minneapolis, Minnesota. Association for Computational Linguistics.

Yingqiang Gao, Nikola I. Nikolov, Yuhuang Hu, and Richard H.R. Hahnloser. 2020. Character-level translation with self-attention. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1591–1604, Online. Association for Computational Linguistics.

Rohit Gupta, Laurent Besacier, Marc Dymetman, and Matthias Gallé. 2019. Character-based NMT with transformer. *CoRR*, abs/1911.04997.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver. Association for Computational Linguistics.

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, Melbourne, Australia. Association for Computational Linguistics.

Jason Lee, Kyunghyun Cho, and Thomas Hofmann. 2017. Fully character-level neural machine translation without explicit segmentation. *Transactions of the Association for Computational Linguistics*, 5:365–378.

Jindřich Libovický and Alexander Fraser. 2020. Towards reasonably-sized character-level transformer NMT by finetuning subword systems. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2572–2579, Online. Association for Computational Linguistics.

John J McCarthy. 1981. A prosodic theory of nonconcatenative morphology. *Linguistic inquiry*, 12(3):373–418.

Mathias Müller, Annette Rios, and Rico Sennrich. 2020. Domain robustness in neural machine translation. In *Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 151–164, Virtual. Association for Machine Translation in the Americas.

Toan Q. Nguyen and David Chiang. 2017. Transfer learning across low-resource, related languages for neural machine translation. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 296–301, Taipei, Taiwan. Asian Federation of Natural Language Processing.

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.

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, Lisbon, Portugal. Association for Computational Linguistics.

Ivan Provilkov, Dmitrii Emelianenko, and Elena Voita. 2020. BPE-dropout: Simple and effective subword regularization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1882–1892, Online. Association for Computational Linguistics.

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, Berlin, Germany. Association for Computational Linguistics.

Rico Sennrich and Biao Zhang. 2019. Revisiting low-resource neural machine translation: A case study. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 211–221, Florence, Italy. Association for Computational Linguistics.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 6000–6010.

Yi Zhu, Ivan Vulić, and Anna Korhonen. 2019. A systematic study of leveraging subword information for learning word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 912–932, Minneapolis, Minnesota. Association for Computational Linguistics.

A Hyperparameters

We conduct a grid search of hyperparameters for the training of Transformer models, including batch size (tokens per batch) and learning rate. For batch size, the searching range is $\{4096, 8192, 16384, 32768\}$. For learning rate, the searching range is $\{5e-5, 1e-4, 5e-4, 1e-3\}$.

Besides, we also experiment with diffrent model size and number of bpe merges k in the low resource settings(50k, 100k, 200k). The searching range of k is $\{2000, 10000\}$. We consider four kinds of model size, i.e. *tiny*, *mini*, *small* and *base*, which differ in their hidden size and transformer layers. The details can be found in Table 4.

	hidden size	layer
tiny	128	2
mini	256	4
small	512	4
base	512	6

Table 4: Detailed hyperparameters for different model sizes.