

# Neural Machine Translation with Byte-Level Subwords

Changhan Wang<sup>†</sup>, Kyunghyun Cho<sup>†‡\*</sup> and Jiatao Gu<sup>†</sup>

<sup>†</sup> Facebook AI Research; <sup>‡</sup> New York University; <sup>\*</sup> CIFAR Global Scholar  
{changhan, kyunghyuncho, jgu}@fb.com

## Abstract

Almost all existing machine translation models are built on top of character-based vocabularies: characters, subwords or words. Rare characters from noisy text or character-rich languages such as Japanese and Chinese however can unnecessarily take up vocabulary slots and limit its compactness. Representing text at the level of bytes and using the 256 byte set as vocabulary is a potential solution to this issue. High computational cost has however prevented it from being widely deployed or used in practice. In this paper, we investigate byte-level subwords, specifically byte-level BPE (BBPE), which is compact than character vocabulary and has no out-of-vocabulary tokens, but is more efficient than using pure bytes only is. We claim that contextualizing BBPE embeddings is necessary, which can be implemented by a convolutional or recurrent layer. Our experiments show that BBPE has comparable performance to BPE while its size is only 1/8 of that for BPE. In the multilingual setting, BBPE maximizes vocabulary sharing across many languages and achieves better translation quality. Moreover, we show that BBPE enables transferring models between languages with non-overlapping character sets.

## Introduction

It has become a standard practice to build a vocabulary in neural machine translation (NMT) (Bahdanau, Cho, and Bengio 2014; Sutskever, Vinyals, and Le 2014) using byte-pair encoding (BPE) (Sennrich, Haddow, and Birch 2015). In this practice, we notice that BPE is used at the level of characters rather than at the level of bytes, which is more common in data compression. We suspect this is because text is often represented naturally as a sequence of characters, although it has recently been noticed that byte representation of text has its own advantages, such as compactness (up to 256 possible values) and being agnostic to languages.

In this paper, we look into byte-level “subwords” that are used to tokenize text into variable-length byte  $n$ -grams, as opposed to character-level subwords in which we represent text as a sequence of character  $n$ -grams. We specifically focus on byte-level BPE (BBPE), examining compact BBPE vocabularies in both bilingual and multilingual settings as

Copyright © 2020, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

片 | 手 の | 拍 手 | の | 音  
片 | 手 E3 81 | AE | 拍 | 手 E3 81 | AE | E9 9F | B3

Figure 1: BPE (upper) and BBPE (lower) tokenization of a Japanese sentence. Bytes (from partial characters) are represented by hexadecimal digits.

well as in a novel setup of transfer learning to a new language with a non-overlapping character set.

## Byte Level Text Representation

**Encoding Byte-Level Representation** We consider UTF-8 encoding of text, which encodes each Unicode character into 1 to 4 bytes. This allows us to model a sentence as a sequence of bytes instead of characters. While there are 138K Unicode characters covering over 150 languages, we represent a sentence in any language as a sequence of UTF-8 bytes (248 out of 256 possible bytes).

A byte sequence representation of text is often much longer (up to 4x) than a character sequence representation, which makes it computationally demanding to use bytes as they are. As an alternative, we consider segmenting a byte sequence into variable-length  $n$ -grams (byte-level “subwords”). Specifically, we learn BPE vocabulary on the byte-level representation which extends UTF-8 byte set with byte  $n$ -grams. We denote this type of vocabulary as B(yte-level)BPE in the rest of the paper. Figure 1 shows an example of BBPE tokenization.

BBPE symbols can be partial characters shared by different characters or the combination of complete and partial characters. This arbitrariness may necessitate incorporating a larger context surrounding each symbol for disambiguation and learning the character boundaries. In this work, we base our experiments on Transformer (Vaswani et al. 2017) models. We propose to use either a depth-wise convolutional layer (Kaiser, Gomez, and Chollet 2017) or a bidirectional recurrent layer with gated recurrent units (Cho et al. 2014, GRU,) to contextualize BBPE embeddings before feeding them into the model:

$$\mathbf{x}_{ctx\_emb} = \text{DepthWiseConv}(\mathbf{x}_{emb})$$

or

$$\mathbf{x}_{ctx\_emb} = \text{BiGRU}(\mathbf{x}_{emb})$$

**Decoding with Byte-Level Subwords** While any sentence can be represented as a byte sequence, the converse is, however, not necessarily true in that there are byte sequences that do not translate to valid character sequences. Empirically, we find that invalid outputs from trained models are very rare. We do not observe any in the experiments described below (note that one of them does have a large test set of 165K examples). And a common error pattern in half-trained models is redundant repeating bytes. In our system, we try to recover as many Unicode characters as possible from this error pattern efficiently in linear time. The algorithm is as follows: For a given byte sequence  $\{B\}_{k=1}^N$ , we denote the maximum number of characters that we can recover from it as  $f(k)$ . Then  $f(k)$  has optimal substructure and can be solved by dynamic programming:

$$f(k) = \max_{t=1,2,3,4} \{f(k-t) + g(k-t+1, k)\} \quad (1)$$

where  $g(i, j) = 1$  if  $\{B\}_{k=i}^j$  corresponds to a valid character, otherwise 0. When  $f(k)$  is calculated recursively, we also record the selections at each position  $k$  so that we can recover the solution through backtracking. The design of UTF-8 encoding ensures the uniqueness of this recovery process: for a character UTF-8 encoded with multiple bytes, its trailing bytes will not make a valid UTF-8 encoded character. Then the best selection in Eq. 1 is unique and so is the final solution.

## Experimental Settings

**Datasets** We run experiments on three bilingual corpora as well as a many-to-English multilingual dataset:

- English-German (En-De): we replicate the same setting of (Vaswani et al. 2017) which uses WMT 2014<sup>1</sup> data (newstest13 for validation and newstest14 for testing)
- Japanese-English (Ja-En): we follow (Michel and Neubig 2018) and concatenate KFTT<sup>2</sup> (Neubig 2011), TED<sup>3</sup> (Cettolo, Girardi, and Federico 2012) and JESC<sup>4</sup> (Pryzant et al. 2017) to construct training, validation and test sets.
- Sinhala-English (Si-En): we use the data from FLoRes (Guzmán et al. 2019).
- Many-to-English (X-En): we adopt the TED Talks corpus compiled by (Ye et al. 2018), which includes parallel data for 59 languages. For our experiments, we use English as target and the other 58 languages as source. We sample 22K examples from the 135K development set for validation.

<sup>1</sup><http://statmt.org/wmt14/translation-task.html>

<sup>2</sup><http://www.phontron.com/kfft>

<sup>3</sup><https://wit3.fbk.eu/mt.php?release=2017-01-trnted>

<sup>4</sup><https://nlp.stanford.edu/projects/jesc>

Table 1 shows an overview statistics of these datasets. We learn (B)BPE vocabularies jointly on source and target sentences using SentencePiece (Kudo and Richardson 2018).

	En-De	Ja-En	Si-En	X-En
Train	4.5M	3.5M	405K	5.1M
Dev	3K	4K	3K	22K*
Test	3K	12K	3K	165K

Table 1: Dataset statistics in number of sentences. \* Sub-sampled from the full 135K development set.

Model	$N$	$d_{model}$	$d_{ff}$	$h$	$P_{drop}$	Params
$T_{flores}$	5	512	2048	2	0.4	38M
$T_{base}$	6	512	2048	8	0.1	44M
$T_{big}$	6	1024	4096	16	0.3	180M

Table 2: Transformer models used in the experiments (using the notations in Vaswani et al. 2017).

**Models and Learning** We use Fairseq (Ott et al. 2019) to train Transformers (Vaswani et al. 2017) with the same learning rate schedule in the original paper. All model configurations are listed in table 2. We set attention and ReLU dropout to 0.1, except Si-En for which we use 0.2. We use 0.2 residual dropout for  $T_{base}$  models in X-En. We use a kernel size of 5 and a padding of 2 on both sides for all convolutional layers.

**Inference and Evaluation** We set beam width to 4 for En-De and 5 for the other and use the best checkpoint by validation loss to generate the predictions. We calculate case-sensitive tokenized BLEU (Papineni et al. 2002) as the metrics using sacreBLEU (Post 2018).

## Results and Analysis

### Qualitative Comparison: BPE vs. BBPE

**Symbol Frequency Distribution** Since the construction of BBPE vocabulary starts from UTF-8 byte set, it has the flexibility of decomposing rare characters into byte  $n$ -grams from the vocabulary instead of including them directly. This frees vocabulary slots for other frequent symbols. Figure 2 compares the symbol frequency distribution of BPE and BBPE. We can see that BBPE symbols are more evenly distributed than BPE ones, even when the latter has already been much more evenly distributed than pure characters. By setting different BBPE vocabulary sizes, we can control the level of rare character decomposition and symbol sharing across different characters. Table 3 shows the ratio of BBPE tokens with partial characters. We can see that large portion of rare characters are decomposed on Ja-En and X-En, which has a large character set of 8K and 11K, respectively. Figure 5 provides an example from Ja-En tokenized with different BBPE vocabularies, where we can see how tokens look like as the tokenization granularity goes from fine to coarse.

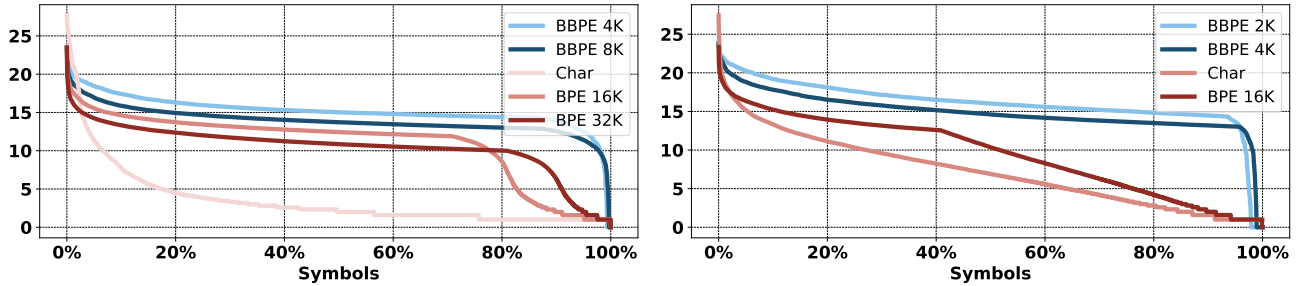


Figure 2: Symbol frequencies (in log2 scale) for En-De (left) and X-En (right) vocabularies. BBPE enables a more consistent distribution of vocabulary across frequencies.

BBPE	2K	4K	8K	16K	32K
En-De	4.3%	4.9%	5.5%	6.1%	6.5%
Ja-En	46.0%	47.6%	49.4%	51.2%	34.8%
X-En	36.8%	39.1%	41.3%	43.6%	23.0%

Table 3: Ratio of BBPE tokens with partial characters.

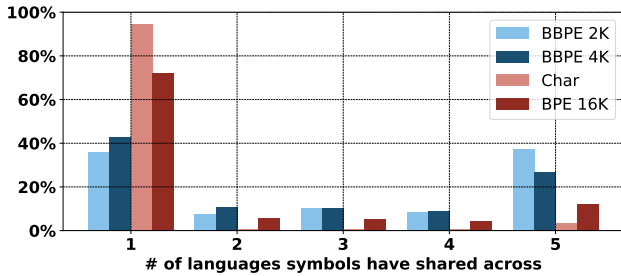


Figure 3: The numbers of languages symbols have shared across Ar, He, Ru, Ko and It (from X-En). Note that these languages have mutually different character sets.

**Cross-Lingual Sharing** In the multilingual setting, symbol sharing also happens across different languages despite the different writing systems. This allows maximizing parameter sharing not only for the model part but also the vocabulary part in a universal model. Figure 3 illustrates the level of BBPE symbol sharing across the top 5 languages (by number of train examples) in X-En whose writing systems are different from each other.

**Impact on Sequence Lengths** Compared to BPE, BBPE symbols are generally finer-grained with shorter byte-level lengths, which results in longer tokenized sequences as well as longer training and inference time. BBPE, however, is optimized for compression-based objective (the same as BPE), and is still more efficient than character vocabulary. Table 4 lists the average lengths of training sentences tokenized with different vocabularies. We can observe that sentences tokenized with BBPE have significantly shorter lengths than the character ones, even when the BBPE vocabulary is much smaller (for example only 1/5 of character set size on X-En). Another observation is that source-target length ratio

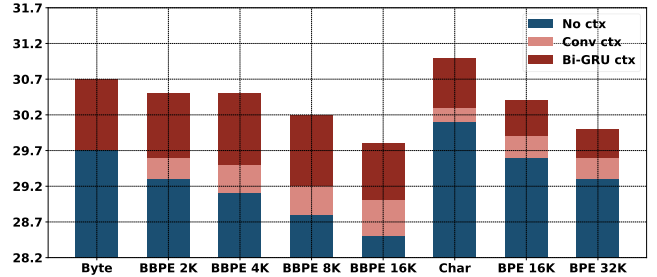


Figure 4: X-En validation BLEU for models without contextualization, with local contextualization (depth-wise convolution) and with long-range contextualization (Bi-GRU). The y axis starts from 28.2 to focus on the gain portions and facilitate comparison across different vocabularies.

for BBPE tends to be much larger when source character set and target character set have very different sizes (for example 11K for X-En source side and 0.1K for the target side). And this situation becomes more severe when BBPE vocabulary size increases. In this case, alignments may be more difficult to learn during model training, since target tokens need attentions on multiple source tokens more often.

### Importance of Contextualization

We compare three different ways of contextualizing token embeddings; none, 1-layer convolution and 1-layer bi-GRU, on X-En with  $T_{base}$  model. We observe from Figure 4 that all kinds of vocabularies can benefit from embedding contextualization. Performance gains are more significant on fine-grained vocabularies: byte, character and BBPE. For BBPE, long-range contextual information from Bi-GRU brings over 4% gain on validation BLEU in all the cases. Encoding context in the token embeddings reduces the difficulties of learning attentions on multiple source tokens and makes model training easier. In the following experiments, we contextualize BBPE with Bi-GRU by default. We denote (B)BPE with Bi-GRU as “(B)BPE <size>+” and the one without contextualization as “(B)BPE <size>”. And we similarly define “Byte+” and “Char+”.

		Byte	BBPE								Char			BPE		
		256	1K	2K	3K	4K	8K	11K	16K	32K	3K	8K	11K	8K	16K	32K
En-De	Source	143	57	48	43	41	36		33	30	143			40	33	31
	Target	160	64	55	50	48	42		38	35	157			43	36	32
Ja-En	Source	55	28	26		24	23		21	21		19			12	10
	Target	53	23	20		17	15		15	13		52			15	14
X-En	Source	126	77	70		65	62	60	59	57			89		40	32
	Target	103	49	43		37	33	32	30	27			103		35	30

Table 4: Average lengths of training sentences tokenized with different vocabularies.

### BBPE on Noisy Character Sets

The En-De training set has quite a few noisy sentence pairs often containing a few non-latin alphabets due to misalignment and code-switched sentences. This leads to a 3.4K character set, while in contrast, English and German both have less than 30 alphabets. Since BPE includes all characters, those rare characters will waste quite a lot of BPE vocabulary slots. For comparison, we try with small BBPE 2K and 4K vocabulary where rare characters are excluded. We find that their performance are comparable to the BPE 32K baseline while having smaller model capacity (see table 5).

		Test BLEU	Params
$T_{base}$	Byte+	26.59	45M
	BBPE 2K+	26.98	47M
	BBPE 4K+	27.08	47M
	Char+	26.73	47M
	BPE 32K	27.31	61M
	BPE 32K+	<b>27.41</b>	62M
	BPE 37K*	27.3	65M
	$T_{big}$	Byte+	26.94
BBPE 2K+		<b>28.78</b>	183M
BBPE 4K+		28.27	185M
Char+		27.24	185M
BPE 32K		28.36	210M
BPE 32K+		28.77	215M
BPE 37K*		28.4	213M

Table 5: En-De test BLEU. \* (Vaswani et al. 2017).

### BBPE on Character-Rich Languages

Languages using logographic writing systems, such as Chinese and Japanese, can have over 50K characters, while only a small portion of them are frequently used. Our Ja-En dataset has a set of 7.9K characters, where 99.99% tokens in the training set are covered by the top 2.4K characters. With this observation, we experiment with BBPE 4K which is roughly 50% of the character set size. We find that BBPE is comparable to BPE and even outperforms BPE when using larger  $T_{big}$  model (see table 6).

### BBPE on Many-to-En Translation

Our many-to-En dataset contains 58 languages (parallelly to English) and 10.8K characters from different writing sys-

		KFTT	TED	JESC	All
# of train samples		440K	223K	2.8M	3.5M
# of test samples		1.2K	8.5K	2K	11.7K
Michel et.al. (2018)		20.77	13.25	18.00	-
$T_{base}$	Byte+	23.12	15.14	15.69	16.27
	BBPE 4K+	<b>24.15</b>	15.59	16.10	16.80
	Char+	23.67	15.26	15.68	16.43
	BPE 16K+	23.63	<b>16.15</b>	<b>16.18</b>	<b>17.19</b>
$T_{big}$	Byte+	23.68	16.08	16.29	17.46
	BBPE 4K+	23.88	<b>19.0</b>	<b>17.93</b>	<b>19.58</b>
	Char+	23.71	16.69	17.01	18.33
	BPE 16K+	<b>24.08</b>	18.34	17.89	19.14

Table 6: Ja-En test BLEU scores.

tems, between which characters are not necessarily shared. The characters, however, share byte  $n$ -grams. We experiment with BBPE 2K and 4K that have 12.5% and 25% size of the baseline BPE vocabulary. As shown in Table 7, both of them beat the BPE baseline on overall BLEU as well as on most of the languages both with high and low resources (Note that the test set is as large as 165K and even small gaps in BLEU may suggest significant difference). We also notice that byte model and character model perform significantly better than all BPE and BBPE models in this multilingual setting. This might be because that BBPE and BPE suffer from imbalanced source and target sentence lengths as well as various token granularities in multilingual parallel sentences (sources in different languages and granularities into same targets). Nonetheless, BBPE is still the most practical solution since it makes a good balance between performance (better BLEU than BPE) and speed (much shorter tokenized sentences than characters and bytes).

### Transfer Learning on Unseen Characters

Because BBPE contains all UTF-8 bytes and has no out-of-vocabulary tokens, BBPE-based models can be transferred between languages with non-overlapping character sets. In comparison, it is impossible to do so with character-based vocabularies without replacing the vocabulary and re-training embeddings from scratch. Our Si-En dataset has 77 Sinhala scripts that are disjoint with the X-En character set. We experiment transferring a pretrained (on X-En) BBPE 4K  $T_{flores}$  model to this dataset while reusing the original

		Ar	De	He	It	Az	Be	Gl	Sk	All	Params
# of train examples		213K	167K	211K	203K	5.9K	4.5K	10K	61K	5.1M	
# of test examples		6K	4.5K	5.5K	5.6K	0.9K	0.7K	1K	2.4K	165K	
Aharoni et al. 19		25.93	28.87	30.19	32.42						
Neubig & Hu 18						11.7	18.3	29.1	28.3		
$T_{base}$	Byte+	31.13	35.98	36.77	38.36	14.64	<b>25.12</b>	35.12	33.08	30.38	45M
	Char+	<b>31.52</b>	<b>36.73</b>	<b>36.85</b>	<b>38.62</b>	<b>15.40</b>	24.90	<b>35.44</b>	<b>33.31</b>	<b>30.75</b>	51M
$T_{base}$	BBPE 2K+	<b>30.79</b>	<b>35.53</b>	<b>36.27</b>	<b>37.82</b>	13.64	24.70	<b>34.17</b>	<b>32.83</b>	<b>29.91</b>	46M
	BBPE 4K+	30.64	34.93	36.07	37.62	<b>13.76</b>	<b>24.84</b>	33.90	32.12	29.74	47M
	BPE 16K	29.70	34.35	34.47	37.02	13.28	24.61	33.55	31.72	29.00	53M
	BPE 16K+	30.20	34.97	35.55	37.49	12.65	23.66	33.95	32.16	29.62	54M
	BPE 32K	29.02	34.08	34.18	36.63	12.56	22.48	32.33	31.26	28.81	61M
	BPE 32K+	29.87	34.64	35.26	37.43	12.35	22.05	33.62	31.61	29.43	62M

Table 7: X-En test BLEU on all 58 languages, top-4 (Ar, De, He, It) and bottom-4 (Az, Be, Gl, Sk) languages by number of training samples. Note that the test set is very large (165K) and even small gaps in BLEU may suggest significant difference.

		Train	Finetune	BLEU
$T_{flores}$	BBPE 5K*	Si-En	-	7.2
	BBPE 4K+	Si-En	-	7.1
$T_{flores}$	BBPE 4K+	X-En	-	0.3
	BBPE 4K+	X-En	enc	8.3
	BBPE 4K+	X-En	enc, dec	8.1
	BBPE 4K+	X-En	embed, enc	<b>9.0</b>
	BBPE 4K+	X-En	all	8.6

Table 8: Transferring pretrained X-En model to Si-En. BBPE 4K is learned on X-En. \* (Guzmán et al. 2019).

vocabulary. As shown in table 8, the transferred model gains 0.9-1.8 BLEU points compared to the baselines, suggesting the generality of pretrained BBPE embeddings and its ability to adapt to different languages with unseen characters. This transfer learning paradigm is free from the limitation of out-of-vocabulary tokens and can be very generic. We just show the extreme case of totally unseen character set, but the pre-trained model may also be transferred to any languages and datasets to improve performance or warm-start model training to save time.

## Related Work

**Subword Vocabularies** Previous works have shown that finer-grained vocabularies consistently outperforms word-level vocabularies in many settings, for example, vocabularies based on morpheme segmentation (Nießen and Ney 2000; Luong, Socher, and Manning 2013), byte-pair encoding (Sennrich, Haddow, and Birch 2015) and vocabularies from unigram language model (Kudo 2018). Our byte-level subword vocabularies are based on byte-pair encoding, while we use bytes as the basic units to compose subwords.

**Character Vocabulary** Existing works also explored pure character vocabulary for machine translation. (Kim et al. 2016) proposed building word representations from character ones; (Chung, Cho, and Bengio 2016) removed the re-

striction of word boundaries and directly learned decoding in character level; (Lee, Cho, and Hofmann 2017) further extended it to a fully character-level model in a multilingual setting; (Cherry et al. 2018) showed that character-level models generally outperforms subword-level ones given enough model capacity.

**Byte-Level Vocabularies** The closest work to ours is the byte-level BPE vocabulary used in GPT-2, a large-scale English language model (Radford et al. 2019). They however rely heavily on hard-coded merging rules and have not conducted any analysis on how their byte-level BPE impacts the quality of language modeling. A vocabulary consisting purely of bytes has previously been used in several natural language processing tasks: part-of-speech tagging and named entity recognition (Gillick et al. 2016), translation (Costa-Juss, Escolano, and Fonollosa 2017), machine reading (Kenter, Jones, and Hewlett 2018) and speech recognition (Li et al. 2019).

**Transformer with Convolution or RNN** There are evidences for performance gains from combining Transformer with convolutional or recurrent layers in the area of NMT (Chen et al. 2018), speech recognition (Li et al. 2019; Mohamed, Okhonko, and Zettlemoyer 2019) and language modeling (Chenguang Wang 2019).

## Conclusion

We proposed BBPE which builds a byte-level subword vocabulary for machine translation. It results in a much more compact vocabulary than character-based ones do without the loss of performance. In multilingual settings, the former often outperforms the latter. BBPE does not have any out-of-vocabulary tokens, allowing us to transfer a model using BBPE between languages with non-overlapping vocabularies. This transfer learning paradigm is actually very generic and can be applied to any languages and datasets for performance gain or training acceleration. With the same vocabulary size, BBPE segments sentences into shorter sequences

Original	質問して__証明と証拠を求めましょう	Ask__questions,__demand__proof,__demand__evidence.	
Byte	E8 B3 AA E5 95 8F E3 81 97 E3 81 A6 E2 96 81 E8 A8 BC E6 98 E6 E3 81 A8 E8 A8 BC E6 8B A0 E3 82 92 E6 B1 82 E3 82 81 E3 81 BE E3 81 97 E3 82 87 E3 81 86	41 73 6B E2 96 81 71 75 65 73 74 69 6F 6E 73 2C E2 96 81 64 65 6D 61 6E 64 E2 96 81 70 72 6F 66 2C E2 96 81 64 65 6D 61 6E 64 E2 96 81 65 76 69 64 65 6E 63 65 2E	
BBPE	1K	E8 B3 AA E595 8F しE381 A6 __E8 A8 BC 明 E381 A8 E8 A8 BC E6 8B A0 をE6 B1 82 めE381 BE しょう	A s k __qu est ions , __dem and __pro of , __dem and __ev id ence .
	2K	E8 B3 AA 問 しE381 A6 __E8 A8BC 明 E381 A8 E8 A8BC E68B A0 を E6 B1 82 めE381 BE しょう	A s k __qu est ion s , __dem and __pro of , __dem and __e vid ence .
	4K	E8 B3 AA 問 しE381 A6 __E8 A8BC 明E381 A8 E8 A8BC 抛 をE6 B1 82 めE381 BE しょう	A s k __qu est ions , __dem and __pro of , __dem and __ev id ence .
	8K	E8 B3 AA問 しE381 A6 __E8 A8BC 明E381 A8 E8 A8BC 抛 をE6 B1 82 めE381 BE しょう	A s k __questions , __demand __pro of , __demand __evidence .
	16K	E8 B3 AA問 しE381 A6 __E8 A8BC 明E381 A8 E8 A8BC 抛 をE6 B1 82 めE381 BE しょう	A s k __questions , __demand __proof , __demand __evidence .
	32K	E8 B3 AA問 しE381 A6 __E8 A8BC 明E381 A8 E8 A8BC 抛 をE6 B1 82 めE381 BE しょう	A s k __questions , __demand __proof , __demand __evidence .
CHAR	質問して__証明と証拠を求めましょう	A s k __questions , __demand __proof , __demand __evidence .	
BPE	16K	質問して__証明と証拠を求めましょう	A s k __questions , __demand __pro of , __demand __evidence .
	32K	質問して__証明と証拠を求めましょう	A s k __questions , __demand __proof , __demand __evidence .

Figure 5: An example from Ja-En tokenized with different vocabularies. Raw spaces are replaced by underscores and spaces are used to split tokens. We can observe how tokens look like as the tokenization granularity goes from fine to coarse: Byte (256) → BBPE (1K, 2K, 4K, 8K) → Char (8K) → BBPE (16K, 32K) → BPE (16K, 32K).

than character-based methods do, leading to faster training and inference. Our future work includes: eliminating source-target sentence length imbalance; evaluating BBPE in one-to-many and many-to-many translation settings; exploring better segmentation algorithms for byte-level subwords.

## References

- Bahdanau, D.; Cho, K.; and Bengio, Y. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.
- Cettolo, M.; Girardi, C.; and Federico, M. 2012. Wit3: Web inventory of transcribed and translated talks. In *Conference of European Association for Machine Translation*, 261–268.
- Chen, M. X.; Firat, O.; Bapna, A.; Johnson, M.; Macherey, W.; Foster, G.; Jones, L.; Parmar, N.; Schuster, M.; Chen, Z.; et al. 2018. The best of both worlds: Combining recent advances in neural machine translation. *arXiv preprint arXiv:1804.09849*.
- Chenguang Wang, Mu Li, A. J. S. 2019. Language models with transformers. In *ArXiv e-prints*.
- Cherry, C.; Foster, G.; Bapna, A.; Firat, O.; and Macherey, W. 2018. Revisiting character-based neural machine translation with capacity and compression. *arXiv preprint arXiv:1808.09943*.
- Cho, K.; van Merriënboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; and Bengio, Y. 2014. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1724–1734.
- Chung, J.; Cho, K.; and Bengio, Y. 2016. A character-level decoder without explicit segmentation for neural machine translation. *arXiv preprint arXiv:1603.06147*.
- Costa-Juss, M. R.; Escolano, C.; and Fonollosa, J. A. 2017. Byte-based neural machine translation. In *Proceedings of the First Workshop on Subword and Character Level Models in NLP*.
- Gillick, D.; Brunk, C.; Vinyals, O.; and Subramanya, A. 2016. Multilingual language processing from bytes. In *Proceedings of NAACL-HLT*.
- Guzmán, F.; Chen, P.-J.; Ott, M.; Pino, J.; Lample, G.; Koehn, P.; Chaudhary, V.; and Ranzato, M. 2019. Two new evaluation datasets for low-resource machine translation: Nepali-english and sinhala-english.
- Kaiser, L.; Gomez, A. N.; and Chollet, F. 2017. Depthwise separable convolutions for neural machine translation. *arXiv preprint arXiv:1706.03059*.
- Kenter, T.; Jones, L.; and Hewlett, D. 2018. Byte-level machine reading across morphologically varied languages. In *Thirty-Second AAAI Conference on Artificial Intelligence*.
- Kim, Y.; Jernite, Y.; Sontag, D.; and Rush, A. M. 2016. Character-aware neural language models. In *Thirtieth AAAI Conference on Artificial Intelligence*.
- Kudo, T., and Richardson, J. 2018. Sentencepiece: A simple and language independent subword tokenizer and deto-

- kenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 66–71.
- Kudo, T. 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)*, 66–75.
- Lee, J.; Cho, K.; and Hofmann, T. 2017. Fully character-level neural machine translation without explicit segmentation. *Transactions of the Association for Computational Linguistics* 5:365–378.
- Li, B.; Zhang, Y.; Sainath, T.; Wu, Y.; and Chan, W. 2019. Bytes are all you need: End-to-end multilingual speech recognition and synthesis with bytes. In *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.
- Luong, T.; Socher, R.; and Manning, C. 2013. Better word representations with recursive neural networks for morphology. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, 104–113.
- Michel, P., and Neubig, G. 2018. Mntn: A testbed for machine translation of noisy text. *arXiv preprint arXiv:1809.00388*.
- Mohamed, A.; Okhonko, D.; and Zettlemoyer, L. 2019. Transformers with convolutional context for asr. *arXiv preprint arXiv:1904.11660*.
- Neubig, G. 2011. The Kyoto free translation task. <http://www.phontron.com/kfft>.
- Nießen, S., and Ney, H. 2000. Improving smt quality with morpho-syntactic analysis. In *Proceedings of the 18th conference on Computational linguistics-Volume 2*, 1081–1085. Association for Computational Linguistics.
- Ott, M.; Edunov, S.; Baevski, A.; Fan, A.; Gross, S.; Ng, N.; Grangier, D.; and Auli, M. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.
- Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, 311–318. Association for Computational Linguistics.
- Post, M. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, 186–191. Association for Computational Linguistics.
- Pryzant, R.; Chung, Y.; Jurafsky, D.; and Britz, D. 2017. JESC: Japanese-English Subtitle Corpus. *ArXiv e-prints*.
- Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; and Sutskever, I. 2019. Language models are unsupervised multitask learners.
- Sennrich, R.; Haddow, B.; and Birch, A. 2015. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*.
- Sutskever, I.; Vinyals, O.; and Le, Q. V. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, 3104–3112.
- Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, 5998–6008.
- Ye, Q.; Devendra, S.; Matthieu, F.; Sarguna, P.; and Graham, N. 2018. When and why are pre-trained word embeddings useful for neural machine translation. In *HLT-NAACL*.