

Finding Better Subword Segmentation for Neural Machine Translation

Yingting Wu^{1,2} and Hai Zhao^{1,2}(✉)

¹ Department of Computer Science and Engineering, Shanghai Jiao Tong University

² Key Laboratory of Shanghai Education Commission for Intelligent Interaction and Cognitive Engineering, Shanghai Jiao Tong University, Shanghai, 200240, China
wuyingting@sjtu.edu.cn, zhaohai@cs.sjtu.edu.cn

Abstract. For different language pairs, word-level neural machine translation (NMT) models with a fixed-size vocabulary suffer from the same problem of representing out-of-vocabulary (OOV) words. The common practice usually replaces all these rare or unknown words with a $\langle \text{UNK} \rangle$ token, which limits the translation performance to some extent. Most of recent work handled such a problem by splitting words into characters or other specially extracted subword units to enable open-vocabulary translation. Byte pair encoding (BPE) is one of the successful attempts that has been shown extremely competitive by providing effective subword segmentation for NMT systems. In this paper, we extend the BPE style segmentation to a general unsupervised framework with three statistical measures: frequency (FRQ), accessor variety (AV) and description length gain (DLG). We test our approach on two translation tasks: German to English and Chinese to English. The experimental results show that AV and DLG enhanced systems outperform the FRQ baseline in the frequency weighted schemes at different significant levels.

Keywords: Neural machine translation · Subword segmentation.

1 Introduction

Neural Machine Translation [3,14,23] has achieved remarkable performance in recent years. The NMT system is commonly based on a word-level model with a restricted vocabulary of most frequent words. All the infrequent or unseen words are simply replaced with a special $\langle \text{UNK} \rangle$ token at the cost of decreasing translation accuracy. In order to better handle the problems of OOV words, recent work proposed the character-based [9,11,17] or word-character hybrid models [19]. Considering that character is not a sufficiently good minimal unit to compose word representation, it is of great value to find a more meaningful representation unit between word and character, i.e., subword.

The benefits of using variable-length subword representation in NMT are two-fold. First, it can effectively decrease the vocabulary size of the whole training set by encoding rare words with sequences of subword units. Second, it is also possible to translate or generate unseen words at inference. Currently, BPE

[13] has been applied commonly in NMT called subword NMT [22] that enables open-vocabulary translation. It is a frequency-based data compression algorithm that most of the high-frequency words are maintained and the infrequent ones are segmented into subwords. As the widely used BPE segmentation only uses frequency as its measure and does not rely on any morphological or linguistic knowledge, it can be regarded as a computationally motivated method. Some other work segmented words based on morphological information to obtain the representation of OOV words [2,5], which can be called linguistically motivated method. However, the availability of morphological resources for different languages cannot be always guaranteed. Thus, this work will focus on the general computationally motivated subword segmentation for NMT.

The development of Chinese word segmentation has played an important role in the natural language processing tasks [6,7,26]. Zhao and Kit [25] proposed a general unsupervised word segmentation framework for Chinese which consists of two collocative components, decoding algorithm and substring goodness measure. Their work shows that the substring frequency (FRQ) is not the best goodness measure for Chinese word segmentation. Since Chinese is an ideographic language without any explicit word boundaries, we can similarly adapt the reported better measures to the subword segmentation for alphabetic languages such as English by considering their words are written in character sequences without explicit character boundaries. Thus, we regard the BPE style segmentation as a general decoding algorithm, and enable it to work with another two advanced goodness measures, accessor variety (AV) [12] and description length gain (DLG) [16] in the hope of further enhancing the current NMT. We compare the effects of different measures with experiments on the German-English and Chinese-English translation tasks. The evaluations on the test sets prove that the extended BPE-style segmentation methods with the frequency weighted AV and DLG improve the FRQ baseline at different significant levels.

2 Neural Machine Translation

In this paper, we closely follow the neural machine translation model proposed by Bahdanau *et al.* [3], which is mainly based on a neural encoder-decoder network with attention mechanism [3,20].

The encoder is a bidirectional recurrent neural network (RNN) with gated recurrent unit (GRU) or long short-term memory (LSTM) unit. The forward RNN reads a source sentence $x = (x_1, \dots, x_m)$ from left to right and calculates a forward sequence of hidden states $(\vec{h}_1, \dots, \vec{h}_m)$. Similarly, the backward RNN reads the source sentence inversely and learns a backward sequence of hidden states $(\overleftarrow{h}_1, \dots, \overleftarrow{h}_m)$. The hidden states of both directions are concatenated to obtain the annotation vector $h_i = [\vec{h}_i, \overleftarrow{h}_i]^T$ for each word in the source sentence.

The decoder is a forward RNN initialized with the final state of the encoder. In the decoding phase, a target sentence $y = (y_1, \dots, y_n)$ is generated step by step. The conditional translation probability can be formulated as follows.

$$p(y_j | y_{<j}, x) = q(y_{j-1}, s_j, c_j),$$

where s_j and c_j denote the decoding state and the source context at the j -th time step respectively. Here, $q(\cdot)$ is the softmax layer and $y_{<j} = (y_1, \dots, y_{j-1})$. The context vector c_j is calculated as a weighted sum of the source annotations according to attention mechanism, where $c_j = \sum_{i=1}^m \alpha_{ji} h_i$. The alignment model α_{ji} defines the probability that how well y_j is aligned to x_i , which can be a single layer feed-forward neural network.

3 Unsupervised Subword Segmentation

Extracting substrings inside a word can be regarded as a segmentation process over character sequences, which is similar to splitting words over Chinese character sequences (e.g., Chinese sentences). Therefore, we borrow the idea from Zhao and Kit [25] who proposed a generalized framework for unsupervised Chinese word segmentation including two collocative modules, a decoding algorithm and an alternative goodness measure. They use a top-down decoding method that starts from the sentence level and then searches for the best segmentation for a particular text according to the given goodness function. Their empirical assessment on Chinese used the frequency of substrings as baseline among all goodness measures, which indicates that frequency is not an optimal segmentation criterion for Chinese. Motivated by this, we propose a combined framework of the extended BPE-style segmentation and some advanced substring measures for better subword representation in NMT.

3.1 Generalized BPE Segmentation

The BPE used in subword NMT is a bottom-up method, where the initial state of each word is a sequence of single characters and consecutive substring pairs with highest frequency are merged iteratively to compose subwords. In this section, we introduce a generalized BPE segmentation framework with arbitrary goodness measures besides frequency. The details are described in Algorithm 1.³

Given the corresponding segmentation measure, if a substring pair ('a', 'b') has the highest goodness score among all the candidate pairs, every occurrence of this pair will be replaced by a new symbol 'ab'. The merge operations are performed on the whole training corpus and the vocabulary size can be controlled by the number of merge operations N . The final vocabulary size is approximately equal to merge times plus the number of character types in the corpus. According to the merge list of substring pairs learned from training data, we can further apply the merge operation to the development set and test set. Each subword except the end of word is attached to a special token such as "@@" for the sake of restoring the segmented words after translation.

³ The source code has been released at <https://github.com/Lindsay125/gbpe>.

Algorithm 1 Generalized BPE segmentation

Input: the training corpus D , merge times N , goodness measure g **Output:** the segmented text D' , merge list V

- 1: The training corpus D is initialized as a set of character sequences in which every word is split into a sequence of characters, and the merge list V is set empty.
 - 2: Given the current segmentation state on D , calculate the goodness scores of all the distinct successive substring pairs according to the goodness measure g ;
 - 3: Search for the highest scored consecutive pair, add it to V and merge all the occurrences of such pairs on D ;
 - 4: If the merge times reaches N , the algorithm ends and returns D' and V .
 - 5: Otherwise, go to Step 2.
-

3.2 Goodness Measures

Given a word W with n characters, its segmentation S can be denoted as a subword sequence $s_1 s_2 \cdots s_m$ ($m \leq n$). For each merge iteration, every subword s_i (the concatenation of a consecutive pair) is assigned a goodness score $g(s_i)$ for how likely it is an independently translatable item within the whole word. In this study, we examine three types of goodness measures.

Frequency of Substring The *frequency of substring* (FRQ) serves as the baseline in the Chinese word segmentation system of Zhao and Kit [25]. Its basic idea is to compare the frequency of two partially overlapped character n -grams and then the shorter substring with lower or equal frequency is discarded as a redundant candidate. We define the corresponding goodness score as $g_{FRQ}(s_i)$, which is the count of word types in the vocabulary that contain s_i . For efficiency, only those substrings that occur over once are considered in the candidate list.

Accessor Variety The *accessor variety* (AV) proposed by Feng *et al.* [12] is a measure to evaluate how likely a substring can be a relatively independent word, which is reported to be good at addressing low frequency words. Given a particular substring, the main idea of AV is that if the type of successive tokens with respect to the corresponding substring increases, it is more likely to be at boundary. Formally, AV is defined as the minimum of $L_{av}(s_i)$ and $R_{av}(s_i)$.

$$g_{AV}(s_i) = AV(s_i) = \min \{L_{av}(s_i), R_{av}(s_i)\},$$

where the left AV $L_{av}(s_i)$ is the number of distinct predecessor tokens of s_i and the right AV $R_{av}(s_i)$ records the number of s_i 's different successor tokens.

Description Length Gain The *description length* of a corpus X is defined as the Shannon-Fano code length for the corpus [15]. It can be calculated as

$$DL(X) = -|X| \sum_{x \in V_X} \hat{p}(x) \log \hat{p}(x) = - \sum_{x \in V_X} c(x) \log \frac{c(x)}{|X|},$$

where V_X is the token vocabulary of X , $c(x)$ is the count of token x in X , $\hat{p}(x) = c(x)/|X|$ and $|X|$ represents the total token count in X .

Kit and Wilks [16] proposed to use the *description length gain* (DLG) for lexical acquisition on word boundary prediction tasks and showed the effectiveness of this goodness measure. The DLG of a particular substring s_i in X is then defined as the description length change while we substitute s_i with an index r and take a note of this operation at the end of X , i.e.,

$$g_{DLG}(s_i) = DLG(s_i) = DL(X) - DL(X[r \rightarrow s_i] \oplus s_i),$$

where $X[r \rightarrow s_i]$ is the resultant corpus by replacing all occurrences of s_i with r throughout X and \oplus represents the concatenation of two strings with a delimiter.

Frequency Weighted Schemes Note that parallel corpus for NMT typically holds a large vocabulary with noise, all of these statistical goodness measures may be biased by the data noise. Especially for FRQ and AV, which are only based on type statistics, their calculation will be misled by too many word types even though each word type has low frequency. Considering that high frequency words usually correspond to those reliable ones with regular forms, we also introduce three frequency weighted variants,⁴

$$g'_{FRQ}(s_i) = g_{FRQ}(s_i) \sum_{\forall w, s_i \in w} f(w),$$

$$g'_{AV}(s_i) = g_{AV}(s_i) \sum_{\forall w, s_i \in w} f(w), \quad g'_{DLG}(s_i) = g_{DLG}(s_i) \sum_{\forall w, s_i \in w} f(w),$$

where $f(w)$ is the frequency of word w in the corpus. Note that here our FRQ'-BPE will slightly differ from the BPE in [22] whose goodness score is directly computed through counting s_i in the entire corpus.

4 Experiments

4.1 Setup

Our experiments will be performed on two typical language pairs, German to English and Chinese to English. The translation quality is evaluated by the 4-gram case-sensitive BLEU [21] and we use sign test [10] to test the statistical significance of our results.

For German-English task, the evaluations are based on data from TED talks corpora of the IWSLT 2014 evaluation campaign [8]. We tokenize all the training data with the script of Moses⁵ and remove sentences longer than 50 words. The training set comprises of about 153K sentence pairs with 100K German words

⁴ Though DLG is already frequency weighted as its definition, the proposed extra frequency weight is empirically verified effective from our preliminary experiments.

⁵ <https://github.com/moses-smt/mosesdecoder/blob/master/scripts/tokenizer>

and 50K English words. The development set consists of 6969 sentence pairs which is randomly extracted from training data. The test set is a concatenation of dev2010, dev2012, tst2010, tst2011 and tst2012 which results in 6750 sentence pairs. As German and English share the similar alphabet and have a large overlap of vocabulary, the learning of subwords is performed on the union of the source and target corpora as suggested by Sennrich *et al.* [22].

For Chinese-English task, we use the News Commentary v12 dataset from the news translation shared task of WMT 2017 [4] which consists of 227K parallel sentence pairs. The development set and test set are randomly selected from the whole corpora with 3000 sentence pairs respectively, and the remaining part is served as the training set. Since Chinese and English have totally different character vocabularies, the subword (word for Chinese) learning is performed on the source and target corpora separately and we also try different learning strategies of segmentation on both sides.

During training, we use LSTM units for both encoder and decoder. Each direction of the LSTM encoder and decoder are with 256 dimensions. The word embedding and the attention size are both set to 256. The model is trained using Adam optimizer with the initial learning rate of 0.001. The batch size is set to 32 for German-English task and 64 for Chinese-English task. We train each model for 40 epochs and halve the learning rate every 10 epochs. The training set is reshuffled at the beginning of each epoch. During decoding, greedy search and beam search with size 10 are both performed to optimize the performance.

Table 1. BLEU scores with normal AV and DLG on German-English test set

Method	Merge Times		
	10K	20K	30K
FRQ-BPE	27.55	27.94	28.29
AV-BPE	27.42	27.96	27.60
DLG-BPE	20.95	20.83	21.03

4.2 Results

Table 1 reports the translation performance after applying FRQ-BPE, DLG-BPE and AV-BPE segmentation on the German-English task with a beam size of 10. From Table 1, we notice that FRQ-BPE and AV-BPE outperform DLG-BPE at different merge times (vocabulary size) with a large margin and FRQ-BPE behaves the best among the three measures. Thus, we consider combining the corpus-level word frequency with the three goodness measures to further enhance the performance as we discussed in Section 3.2.

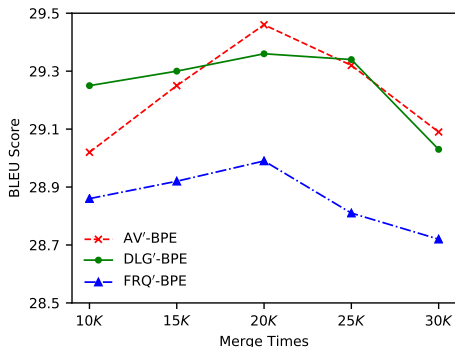
Table 2 reports BLEU scores with different frequency weighted goodness measures and corresponding merge times on the German-English test set.⁶ For

⁶ “++” indicates that the corresponding BLEU is significantly better than the best score of FRQ’-BPE at the significant level $p < 0.01$, “+”: $p < 0.05$.

Table 2. BLEU scores with frequency-weighted schemes on German-English test set

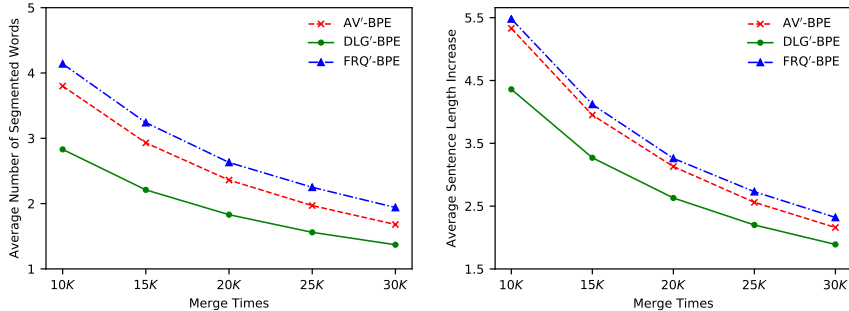
N	FRQ'-BPE		AV'-BPE		DLG'-BPE	
	Greedy	Beam	Greedy	Beam	Greedy	Beam
10K	27.36	28.86	27.72	29.02	27.56	29.25
15K	27.42	28.92	27.56	29.25	27.75	29.30
20K	27.51	28.99	27.65	29.46⁺⁺	27.76⁺	29.36
25K	27.22	28.81	27.69	29.32	27.72	29.34
30K	27.06	28.72	27.52	29.09	27.46	29.03

greedy search, both DLG'-BPE and AV'-BPE achieve better performance than FRQ'-BPE and the best score is obtained with DLG'-BPE of 20K merge operations. For beam search, AV'-BPE and DLG'-BPE outperform FRQ'-BPE with improvement of 0.47 and 0.37 BLEU points when $N = 20K$ respectively. Figure 1 plots the variation curves of BLEU scores with different goodness measures against the merge times N when the beam size is 10. From Figure 1, we observe that the peak value of BLEU with different criteria appears at the 20K merge times. In general, AV'-BPE and DLG'-BPE are comparable with each other and both superior to FRQ'-BPE with a margin.

**Fig. 1.** The curves of BLEU scores with different merge times and segmentation criteria on German-English test set (beam size=10).

We also calculate and plot the average number of segmented words in each sentence and average sentence length increase after subword segmentation with different criteria on German-English task in Figure 2. Figure 2(a) shows that, with the same merge times, AV'-BPE and DLG'-BPE maintain more original word forms than FRQ'-BPE within each sentence. The sentence length is an important factor during training as too long sequences will slow down the training and NMT shows a weak performance on long sentences. Figure 2(b) also shows that both DLG'-BPE and AV'-BPE can better control the sequence length than FRQ'-BPE. For both Figure 2(a) and 2(b), the curve of AV'-BPE with best

performance is between that of DLG'-BPE and FRQ'-BPE, which means that it is crucial to find the proper granularity for segmentation.



(a) Average number of segmented words in each sentence. (b) Average sentence length increase.

Fig. 2. Statistics of words and sentence after segmentation.

Table 3. BLEU scores on Chinese-English test set (beam size=10)

Source	Target		
	FRQ'-BPE	AV'-BPE	DLG'-BPE
FRQ'-BPE	20.07	20.27	20.31
AV'-BPE	19.96	20.23	20.13
DLG'-BPE	20.70	20.45	20.72⁺⁺

In Table 3, we report the BLEU scores with different segmentation measures on the test set of Chinese-English task. Here, the subword segmentation learning is performed separately on the source and target sides with 30K merge operations which is tuned the same way as German-English task. We also try different goodness measures on each side at the same time. It can be found that DLG'-BPE on the both sides significantly outperforms FRQ'-BPE with 0.65 BLEU points. Figure 3 shows the translation performance comparison of different goodness measures when the source segmentation is fixed. It can be seen that DLG'-BPE does much better at the source segmentation for Chinese than the other two, while AV'-BPE and FRQ'-BPE are on a par with each other.

4.3 Translation Examples

Table 4 shows some translation examples with different segmentation criteria on the test sets of German-English and Chinese-English tasks. We use the “|” token to mark the splitting points within the words (sentences for Chinese).

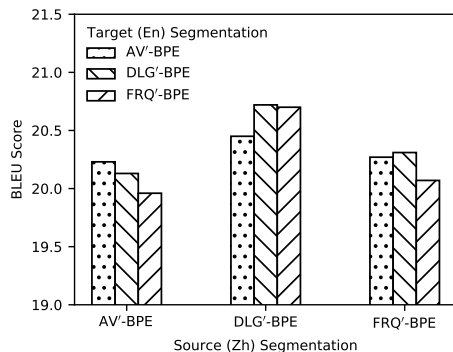


Fig. 3. The bar chart of BLEU scores with different segmentation criteria on Chinese-English test set (beam size=10).

In the first case, AV'-BPE performs best by translating “*erstaunlich wissbegierige*” to an acceptable alternative “*astonishingly inventive*” compared to the ground truth “*astonishingly inquisitive*”. DLG'-BPE omits the translation of “*wissbegierige*” probably as the segmentation of this word is more confusing than that of the other two methods. FRQ'-BPE behaves not badly in the interpretation of “*erstaunlich*” but generates a non-existing word “*inwalkable*”.

In the second case, DLG'-BPE shows the better translation quality than the other two. The critical difference in splitting phrase “在多大程度上” (*by what a degree, how much*) mostly results in the diverse translation behaviors. Only DLG'-BPE correctly translates it to the correct meaning “*how much*”, while FRQ'-BPE and AV'-BPE give the undesirable answers of “*how else*” and “*what kind of*” respectively.

5 Related Work

Previous work resorted to various techniques to deal with the representation of rare words for NMT. Recently, character-level and other subword-based models have become increasingly popular and achieved great performance for different language pairs. The character-level models often rely on convolutional or recurrent neural networks to encode or decode the character sequences. For instance, Costa-jussà and Fonollosa [11] employed convolutional and highway layers upon characters to form the word embeddings. Ling *et al.* [18] used a bidirectional LSTM to combine character embeddings to word embeddings and generated the target word character by character. Similarly, Ataman and Federico [1] proposed to improve the quality of source representations of rare words by augmenting its embedding layer with a bi-RNN, which can learn compositional input representations at different levels of granularity. Luong and Manning [19] proposed a word-character hybrid model that translates mostly at the word level and consults the character information for rare words with an additional deep character-based LSTM. Different from above methods which still consider

Table 4. German-English and Chinese-English translation examples.

Source	...und wozu wir als erstaunlich wissbegierige Spezies fähig sind .
Reference	...and all that we can be as an astonishingly inquisitive species .
AV'-BPE	...und wo zu wir als erstaunlich wiss beg ier ige spezies fähig sindand what we are , as an aston ish ingly inven tive species .
DLG'-BPE	...und wo zu wir als erstaunlich wis sbe gi erige spezies fähig sindand what we do as an ast ound ing species .
FRQ'-BPE	...und wo zu wir als erstaun lich wiss beg ier ige spezies fähig sindand what we are as an amazing ly in walk able species .
Source	我们应当在多大程度上担心社会不平等呢？
Reference	How much should we worry about inequality ?
AV'-BPE	我们 应当 在多大程度上 担心 社会 不平等 呢？ What kind of social inequality we should be ?
DLG'-BPE	我们 应当 在 多大程度上 担心 社会 不平等 呢？ How much should we worry about social inequality ?
FRQ'-BPE	我们 应当 在 多 大程度 上 担心 社会 不 平等 呢？ How else should we worry about social inequality ?

word boundaries, Chung *et al.* [9] introduced a character-level decoder without explicit segmentation while the encoder is still at subword-level. Further, Lee *et al.* [17] proposed the fully character-level NMT model that maps a source character sequence to a target one without any explicit segmentation for both encoder and decoder. At present, the most popular subword-level method is independent of the training procedure and can be regarded as a preprocessing step such as BPE [22] and the wordpieces [24]. The main advantage of this approach is that the original architecture of the word-level model can be reused without increasing much complexity of training or decoding. Here, we focus on the subword segmentation improvement by introducing better substring measures.

6 Conclusion

In this paper, we introduce a generalized subword segmentation framework to enable open-vocabulary translation in NMT. Specifically, we extend the frequency only based BPE segmentation to a general bottom-up case that is capable of incorporating other advanced substring goodness measures, AV and DLG. We empirically compare the effects of different goodness measure schemes on the IWSLT14 German-English and WMT17 News Commentary Chinese-English translation tasks. The experimental results show that frequency weighted DLG'-BPE and AV'-BPE achieves stably better performance than FRQ'-BPE at different significant levels. In general, AV'-BPE shows best performance on the German-English task and DLG'-BPE behaves better on the Chinese-English task. The difference on the best subword segmentation strategies indicates that the choice may be sensitive to specific language pairs to some extent, which deserves further exploration in the future.

Acknowledgments. This paper was partially supported by National Key Research and Development Program of China (No. 2017YFB0304100), National Natural Science Foundation of China (No. 61672343 and No. 61733011), Key Project of National Society Science Foundation of China (No. 15-ZDA041), The Art and Science Interdisciplinary Funds of Shanghai Jiao Tong University (No. 14JCRZ04).

References

1. Ataman, D., Federico, M.: Compositional representation of morphologically-rich input for neural machine translation. arXiv preprint arXiv:1805.02036 (2018)
2. Ataman, D., Negri, M., Turchi, M., Federico, M.: Linguistically motivated vocabulary reduction for neural machine translation from turkish to english. *The Prague Bulletin of Mathematical Linguistics* **108**(1), 331–342 (2017)
3. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. In: *Proceedings of 3rd International Conference on Learning Representations* (2015)
4. 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.: Findings of the 2017 conference on machine translation. In: *Proceedings of the 2nd Conference on Machine Translation, Volume 2: Shared Task Papers*. pp. 169–214 (2017)
5. Botha, J., Blunsom, P.: Compositional morphology for word representations and language modelling. In: *International Conference on Machine Learning*. pp. 1899–1907 (2014)
6. Cai, D., Zhao, H.: Neural word segmentation learning for chinese. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*. pp. 409–420 (2016)
7. Cai, D., Zhao, H., Zhang, Z., Xin, Y., Wu, Y., Huang, F.: Fast and accurate neural word segmentation for chinese. In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*. pp. 608–615 (2017)
8. Cettolo, M., Niehues, J., Stüker, S., Bentivogli, L., Federico, M.: Report on the 11th iwslt evaluation campaign. In: *the 11th International Workshop on Spoken Language Translation*. Lake Tahoe, USA (2014)
9. Chung, J., Cho, K., Bengio, Y.: A character-level decoder without explicit segmentation for neural machine translation. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. pp. 1693–1703. Berlin, Germany (2016)
10. Collins, M., Koehn, P., Kučerová, I.: Clause restructuring for statistical machine translation. In: *Proceedings of the 43rd annual meeting on association for computational linguistics*. pp. 531–540 (2005)
11. Costa-jussà, M.R., Fonollosa, J.A.R.: Character-based neural machine translation. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. pp. 357–361. Berlin, Germany (2016)
12. Feng, H., Chen, K., Deng, X., Zheng, W.: Accessor variety criteria for chinese word extraction. *Computational Linguistics* **30**(1), 75–93 (2004)
13. Gage, P.: A new algorithm for data compression. *The C Users Journal* **12**(2), 23–38 (1994)

14. Kalchbrenner, N., Blunsom, P.: Recurrent continuous translation models. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. pp. 1700–1709. Seattle, Washington, USA (2013)
15. Kit, C.: A goodness measure for phrase learning via compression with the mdl principle. In: Proceedings of the ESSLLI Student session. pp. 175–187 (1998)
16. Kit, C., Wilks, Y.: Unsupervised learning of word boundary with description length gain. In: Proceedings of the 3rd Conference on Computational Natural Language Learning. pp. 1–6 (1999)
17. Lee, J., Cho, K., Hofmann, T.: Fully character-level neural machine translation without explicit segmentation. *Transactions of the Association for Computational Linguistics* **5**, 365–378 (2017)
18. Ling, W., Trancoso, I., Dyer, C., Black, A.W.: Character-based neural machine translation. CoRR [abs/1511.04586](https://arxiv.org/abs/1511.04586) (2015)
19. Luong, M.T., Manning, C.D.: Achieving open vocabulary neural machine translation with hybrid word-character models. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 1054–1063. Berlin, Germany (2016)
20. Luong, T., Pham, H., Manning, C.D.: Effective approaches to attention-based neural machine translation. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. pp. 1412–1421. Lisbon, Portugal (2015)
21. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of 40th Annual Meeting of the Association for Computational Linguistics. pp. 311–318 (2002)
22. Sennrich, R., Haddow, B., Birch, A.: Neural machine translation of rare words with subword units. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. pp. 1715–1725. Berlin, Germany (2016)
23. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. In: *Advances in Neural Information Processing Systems* **27**, pp. 3104–3112. Curran Associates, Inc. (2014)
24. Wu, Y., Schuster, M., Chen, Z., Le, Q.V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., et al.: Google’s neural machine translation system: Bridging the gap between human and machine translation. arXiv preprint [arXiv:1609.08144](https://arxiv.org/abs/1609.08144) (2016)
25. Zhao, H., Kit, C.: An empirical comparison of goodness measures for unsupervised chinese word segmentation with a unified framework. In: Proceedings of the 3rd International Joint Conference on Natural Language Processing. pp. 9–16 (2008)
26. Zhao, H., Utiyama, M., Sumita, E., Lu, B.L.: An empirical study on word segmentation for chinese machine translation. In: Gelbukh, A. (ed.) *Computational Linguistics and Intelligent Text Processing*. pp. 248–263. Springer Berlin Heidelberg, Berlin, Heidelberg (2013)