Modeling Multi-granularity Segmentation for Rare Words in Neural Machine Translation

Anonymous ACL submission

Abstract

Segmenting rare words into subwords has become a commonly used and effective way to alleviate the open vocabulary problem in Neural Machine Translation (NMT). The existing dominant segmentation methods either give rare words a single segmentation or a fixed segmentation, which leads to a lack of morphological diversity in representing words. For rare words, we first obtain segmentation with different granularities through Byte Pair Encoding (BPE) and BPE-Dropout, and then propose BPEATT model to dynamically mix the BPE subwords and BPE-Dropout subwords, which enhances the encoder's ability to represent rich morphological information. Experiments on six translation benchmarks of different scales show that our proposed method significantly outperforms the baseline model and has obvious advantages over related methods [1].

1 Introduction

The vocabulary plays a crucial role in the Neural Machine Translation (NMT) models (Bahdanau et al., 2015; Luong et al., 2015; Wu et al., 2016; Vaswani et al., 2017a). However, the open vocabulary problem has always puzzled the MT community and seriously affects the quality and readability of the machine translation. To deal with his problem, researchers have proposed many segmentation methods to handle the rare words [2], either performing on the space-separated words (Arppe et al., 2005; Bahdanau et al., 2014), or dividing words into characters (Kim et al., 2016; Lee et al., 2017), or into subwords (Schuster and Nakajima, 2012; Sennrich et al., 2016; Kudo and Richardson, 2018). However, most of segmentation methods give rare words a single and fixed segmentation result, which leads to a lack of morphological diversity in the representation of words. To make the model learn the compositionality of words and be robust to segmentation errors, there are two main lines of research: either generating word segmentations dynamically, or combining word segmentations with different granularities.

Subword Regularization (Kudo, 2018) trained the NMT model with multiple subword segmentations, which are probabilistically sampled by a pretrained unigram language model (ULM) during training. But it requires a pretrained ULM, which increases the cost of training the model. BPE-Dropout (Provilkov et al., 2020) randomly disturbed the segmentation procedure of the standard BPE, leading to diverse segmentations for rare words. However, BPE-dropout has a high probability of losing the original BPE segmentation information of words during training. He and Haffari (2020) proposed Dynamic Programming Encoding (DPE) to utilize segmentation methods with different granularities on source and target sentences: the source sentence is segmented using BPE-Dropout, and the target sentence is segmented using the DPE algorithm mixing characters and subwords. But it is limited to languages and cannot be applied to Asian languages, such as Chinese.

There are also researchers who mix multiple segmentation methods with different granularities. Wu et al. (2020) leveraged the mixed representations from different tokenization approaches for sequence generation tasks where the two different approaches are from the frequency-based and language-model-based methods. Zhang and Li (2020) proposed a multiple-grained method AMBERT to take both sequences of words (fine-grained segmentation) and sequences of phrases (coarse-grained segmentation) as inputs and improved natural language understanding (NLU) tasks. Generally, the previous methods either did

[1] The code will be released upon publication

[2] Generally speaking, rare words should be words whose word frequency is statistically lower than a certain threshold. But we claim that if a word is segmented into characters or subwords, we define the word as a rare word, otherwise it is a common word.

not consider multiple granularities, or output a specific segmentation fixedly for rare words.

We propose BPEATT model to mix the BPE segmentation and BPE-Dropout segmentation dynamically. For each rare word, we first obtain BPE subwords and BPE-Dropout subwords, and then make attention across all subwords, so that each rare word contains morphological information with different segmentation granularity. BPEATT is simple and effective. For each rare word, our method considers both BPE segmentation and dynamic random segmentation, and does not require pretrained model and on-the-fly operations. Experiments show that BPEATT outperforms BPE, BPE-Dropout, DPE on three datasets with different scales and the model of Wu et al. (2020) on low-resource datasets.

2 Methodology

We introduce BPEATT model, a fusion method of mixed granularity segmentation. Note that our algorithm relies on the native BPE segmentation and BPE-Dropout segmentation.

2.1 Definitions

We represent a sentence in source language as $X=(x_1, \cdots, x_i, \cdots, x_j, \cdots, x_{|X|})$ where x_i means the i^{th} word in X. Then BPE and BPE-Dropout segmentations of x_i are denoted by functions BPE(x_i) and BPE-Dropout($x_i, \boldsymbol{p}$) respectively. We discard the merging operation with the probability $\boldsymbol{p}$ for BPE-Dropout.

Algorithm 1 Generating MGS

Input: The source sentence X
Output: MGS
  MGS $\leftarrow$ NULL String
  for $i = 1, 2, ..., |X|$ **do**
    $B \leftarrow$ BPE(x_i)
    if B equals x_i **then**
      MGS += (x_i).
    else
      $C \leftarrow$ BPE-Dropout(x_i, $\boldsymbol{p}$)
      MGS += (B, C)
    end if
  end for

2.2 Mixed Granularity Sequence

We elaborate on how to generate the Mixed Granularity Sequence (MGS) for X. Assume that x_i is a rare word and x_j is a common word, the sequence of subwords generated by the BPE model is $B=\left(x_i^{b_1}, \cdots, x_i^{b_k}, \cdots, x_i^{b_{|B|}}\right)$ where $x_i^{b_k}$ denotes the k^{th} subword of B, and the sequence of subwords produced by the BPE-Dropout model is $C=\left(x_i^{c_1}, \cdots, x_i^{c_k}, \cdots, x_i^{c_{|C|}}\right)$ where $x_i^{c_k}$ denotes the k^{th} subword of C. Then x_i will be replaced by the combination of B and C, and the MGS can be denoted as $(x_1, \cdots, B, C, \cdots, x_j, \cdots, x_{|X|})$. The procedure is described in Algorithm 1.

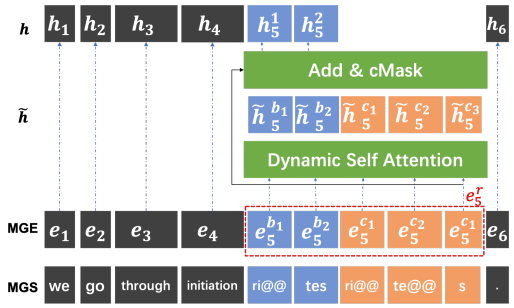


Figure 1: An illustration of the BPEATT architecture. MGS represents the input sequence and the 5^{th} word is a rare word. Black box, blue box and orange box represent common words, BPE subwords, BPE-Dropout subwords respectively. Note that the example sequence here only contains one rare word. In fact, a sequence may contain multiple rare words. In this case, DSA shares the weight among rare words. We ignore the residuals in the figure.

2.3 Architecture

To make the encoder get more morphological information from the subword units, we utilize the Self-Attention with Dynamic Mask (denoted as Dynamic Self-Attention) to fuse information as depicted in Figure 1.

The embedding sequence of Mixed Granularity is called Mixed Granularity Embedding (noted as MGE). For the rare word x_i, we represent the embedding sequence of all subword units as e_i^r

$$e_i^r = \left(e_i^{b_1}, \cdots, e_i^{b_{|B|}}, e_i^{c_1}, \cdots, e_i^{c_{|C|}}\right) \quad (1)$$

Due to the different number of subwords in different rare words, we propose to use Dynamic Self-Attention (abbreviated as DSA). DSA is actually the same as the self-attention mechanism in Vaswani et al. (2017b), except that it only acts on the subword units of rare words, which means e_i^r is used as Q, K and V respectively. The fusion representation sequence $\tilde{h}_i$ of the rare word x_i is calculated by $\tilde{h}_i = \text{DSA}\left(e_i^r, e_i^r, e_i^r\right)$. Our extra experiments show that DSA may cause a slight loss of information in the original BPE subword units, so residual connection is conducted to compensate for this[3]. Finally, we leverage Add & cMask

[3]We also tried to stack more layers in DSA, but under the

operations to ignore the representations of BPE-Dropout subwords and produce the final output $H = (h_1, \cdots, h_i, \cdots)$ as shown in Figure 1.

$$h_i = \begin{cases} e_i & x_i \notin \boldsymbol{R} \\ (e_i^{b_1} \oplus \tilde{h}_i^{b_1}, \cdots, e_i^{b_{|B|}} \oplus \tilde{h}_i^{b_{|B|}}) & x_i \in \boldsymbol{R} \end{cases} \quad (2)$$

where $\boldsymbol{R}$ is the set of rare words extracted from the training set. $\oplus$ means the element-wise addition between two vectors. We then feed H into the original encoder of Transformer, and the subsequent model is exactly the same as Vaswani et al. (2017b).

3 Experiments

3.1 Configurations

Datasets. We conduct experiments on 4 WMT translation benchmarks which are WMT09 En→Hu, WMT14 En→De, WMT14 De→En and WMT20 Zh→En respectively, and 3 low-resource IWSLT14 translation benchmarks, which are En→De, En→Ro and En→Pt-br respectively. Case-sensitive 4-gram BLEU (Papineni et al., 2002) is calculated by SacreBLEU (Post, 2018)[4]. The details of datasets and settings are described in Appendix.

Systems. We train the baseline on datasets segmented using the standard BPE model (Sennrich et al., 2016). Both BPE-Dropout (Provilkov et al., 2020) and DPE (He and Haffari, 2020) reuse the vocabularies of baseline. We apply BPE-Dropout on the source and target sides with the dropout probability $\boldsymbol{p} = 0.1$. For DPE, we follow all settings in He and Haffari (2020). For the low-resource datasets, we train the model of Wu et al. (2020) (denoted as SGMR) by using the open source code.

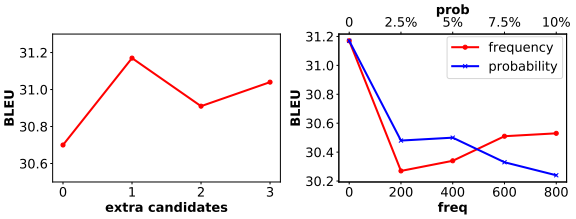
3.2 Effect of Multi-granularity Candidates

To explore whether richer subwords information can benefit BPEATT model, we randomly select 39,402 sentence-pairs from the training set as validation set to explore the influence of multi-granularity segmentation from two perspectives.

Number of Multi-granularity Candidates. As shown in Figure 2(a), more candidates cannot bring further improvement to the model, but still brings benefits to the baseline. Based on this, we only use one BPE-Dropout segmentation candidate.

base setting, the performance is roughly the same as using only one layer.

[4] BLEU+case.mixed+numrefs.1+smooth.exp+tok.13a+version.1.5.1



(a) (b)

Figure 2: Experiments on the WMT14 En→De validation set we randomly sample from the training set. (a) The horizontal axis represents the number of the BPE-Dropout segmentation candidates for one rare word. 0 indicates the baseline. (b) The red curve means that common words with a word frequency lower than **freq** are considered as rare words and segmented by BPE-Dropout; The blue curve indicates that common words are treated as rare words with the probability of **prob**.

Ratio of Rare Words in Training Set. We try to increase the number of words in the training set to be segmented to see if the model can enrich the morphological representation of the words. In addition to original rare words, for one thing, we treat a word whose frequency is lower than the threshold **freq** as a new rare word; and for another, we randomly select the new rare words with a probability **prob** from all words. The two strategies for increasing rare words did not improve BPEATT model, so we do not increase the rare words in the training set, the results are showed in Figure 2(b).

3.3 Main Results

Table 1 shows the results on WMT datasets. It can be seen that our proposed BPEATT consistently performs better than standard BPE, BPE-Dropout and DPE. BPEATT outperforms the standard BPE by 0.55, 0.88 and 0.62 BLEU points on WMT09 En→Hu, WMT14 En→De and De→En test sets respectively. On the WMT20 Zh→En test set BPEATT improves the baseline by 0.66 BLEU points on average. Besides, on the WMT datasets BPEATT gains about 0.41 and 0.57 BLEU score over DPE and BPE-Dropout on average. We have also reproduced BPE-Dropout and DPE on IWSLT14 low-resource datasets. As shown in Table 2, BPEATT outperforms the baseline system and all related methods on the En→De and En→Ro test sets. On the En→Pt-br test set, it is also superior to all other methods, except for BPE-Dropout.

4 Analysis and Discuss

4.1 Robustness Across Domains

In order to figure out how BPEATT performs when the data distribution on the test set is quit differ-

Model	WMT09	WMT14		WMT20 Zh→En			
	En→Hu	En→De	De→En	newstest2018	newstest2019	newstest2020	Avg.
BPE	12.91*	27.50†	30.80*	23.37	24.15†	25.29†	24.27
BPE-Dropout	12.94*	28.01	30.83*	23.38	24.19†	25.35*	24.31
DPE	13.38	28.15	31.10	23.42	24.25†	25.39*	24.35
BPEATT	**13.46**	**28.38**	**31.42**	**23.74**	**25.12**	**26.02**	**24.96**

Table 1: Overview of BLEU scores on WMT datasets. Bold indicates the best BLEU score. "†" and "*" indicate statistically significent difference with $p < 0.01$ and $p < 0.05$ from BPEATT respectively computed via Collins et al. (2005).

Model	IWSLT14		
	En→De	En→Ro	En→Pt-br
BPE	28.49	28.55	39.48
BPE-Dropout	28.43	28.77	**39.86**
DPE	28.60	28.73	39.59
SGMR	28.65	28.42	39.55
BPEATT	**28.65**	**28.97**	39.74

Table 2: The results of five systems on the IWSLT14 datasets. Bold indicates the best BLEU score.

ent from that on the training set. We use the three methods of BPE, BPE-Dropout and BPEATT to train three models on the WMT14 En→De training set in the news domain, and then use these three models to respectively translate two test sets in the biomedical domain and one test set in the IT domain. The results are listed in Table 3. Compared with the baseline and related methods, the benefits of BPEATT on the test set in the biomedical domain are much greater than the benefits of the test set in the news domain (such as 2.17>0.88 and 1.14>0.88) and the IT domain (such as 4.03>0.88).

Question about why BPEATT gains more when the difference in data distribution is relatively large may be asked. We counted the proportion of words segmented by BPE and BPE-Dropout in the test set, and found that the proportions in the three test sets of newstest2014, NHS 24 and Cochrane are about 7.8%, 4.5% and 8.7%, respectively. We roughly guess that the more rare words on the test set, our model can capture more diverse morphological information of rare words, which is more useful for the model to understand and translate rare words.

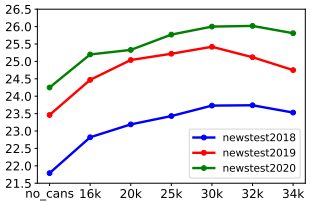


Figure 3: Comparison results of different BPE operations acting on the test set. no_cans means no extra BPE-Dropout segmentation candidates in test set.

Model	Test Sets		
	Cochrane	NHS 24	Batch3
BPE	29.15	29.37	32.47
BPE-Dropout	$30.24^{\uparrow 1.09}$	$29.41^{\uparrow 0.04}$	$32.59^{\uparrow 0.12}$
BPEATT	$31.32^{\uparrow 2.17}$	$30.54^{\uparrow 1.14}$	$36.50^{\uparrow 4.03}$

Table 3: The performance of the three models trained on the WMT14 En→De training set in the news domain on the test set in the biomedical and IT domain. The score after the arrow ↑ indicates the increase in BLEU compared to standard BPE.

4.2 Impact of the Number of BPE Operations

According to our experience, the larger the number of BPE operation, the larger the number of words to be segmented, and the finer the granularity of word segmentation. So we try to control the number of rare words in the test set by adjusting the number of the BPE operation.

We train 6 BPE models with the number of BPE operations of 16k, 20k, 25k, 30k, 32k and 34k. 32k BPE model and the corresponding BPE-dropout strategy are employed to process the training set. Then we all BPE models and corresponding BPE-dropout strategies to process the test set. Note that the number of BPE-dropout operation equals that of BPE operation. Interestingly, too many or too few rare words on the test set will damage our model, the best effect is achieved only when the number of BPE operations applied to the test set are roughly the same as that applied to the training set. For BPEATT, it is not that the more words that are segmented, the greater the benefit.

5 Conclusion

We propose BPEATT to act on rare words to alleviate the open vocabulary problem. Comparative experiments show that our method is significantly better than related word segmentation methods. Analytical experiments verify that BPEATT can learn more morphological information of rare words, and the the benefits are more significant in test scenarios with greater differences in data distribution.

References

Antti Arppe, Lauri Carlson, Krister Lindén, Jussi Piitulainen, Mickael Suominen, Martti Vainio, Hanna Westerlund, and Anssi Yli-Jyrä. 2005. Inquiries into words, constraints and contexts.

Dzmitry Bahdanau, Kyung Hyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. 3rd International Conference on Learning Representations, ICLR 2015 ; Conference date: 07-05-2015 Through 09-05-2015.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *3rd International Conference on Learning Representations, ICLR 2015*.

M. Cettolo, J. Niehues, S. Stüker, L. Bentivogli, and Marcello Federico. 2015. Report on the 11 th iwslt evaluation campaign , iwslt 2014.

Michael Collins, Philipp Koehn, and Ivona Kučerová. 2005. Clause restructuring for statistical machine translation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 531–540, Ann Arbor, Michigan. Association for Computational Linguistics.

Xuanli He and Gholamreza Haffari. 2020. Dynamic programming encoding for subword segmentation in neural machine translation. pages 3042–3051.

Yoon Kim, Yacine Jernite, David Sontag, and Alexander Rush. 2016. Character-aware neural language models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

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.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. 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.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. pages 1412–1421.

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 on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Gabriel Pereyra, George Tucker, Jan Chorowski, Lukasz Kaiser, and Geoffrey E. Hinton. 2017. Regularizing neural networks by penalizing confident output distributions. *CoRR*, abs/1701.06548.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. 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.

Mike Schuster and Kaisuke Nakajima. 2012. Japanese and korean voice search. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5149–5152. IEEE.

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.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, undefinedukasz Kaiser, and Illia Polosukhin. 2017a. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, undefinedukasz Kaiser, and Illia Polosukhin. 2017b. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Process-*

ing Systems, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Lijun Wu, Shufang Xie, Yingce Xia, Yang Fan, Jian-Huang Lai, Tao Qin, and Tieyan Liu. 2020. Sequence generation with mixed representations. In *International Conference on Machine Learning*, pages 10388–10398. PMLR.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144.

Xinsong Zhang and Hang Li. 2020. Ambert: A pre-trained language model with multi-grained tokenization. *arXiv preprint arXiv:2008.11869*.

A Appendix

A.1 Dataset

We conduct our experiments on small-scale, medium-scale and large-scale datasets and explore the generalization of BPEATT for different languages. Table 4 details the size of corpus. For WMT20 Chinese→English (denoted as ZH→EN), the validation set is newstest2017 and the testsets are newstest2018, newstest2019 and newstest2020. We use newstest2013 and newstest2014 as the validation set and test set for WMT14 English→German (denoted as En→De). To compare with BPE-Dropout we use lowercased dataset for En→De translation. For small-scale tasks, we have WMT09 English→Hungarian (denoted as EN→HU) and lowercased IWSLT-2014 datasets (Cettolo et al., 2015) for English→German, English→Romanian (denoted as EN→RO) and English→Portuguese-Brazil(denoted as EN→Pt-Br). For EN→HU, newstest2008 and newstest2009 are used as validation set and test set respectively. For IWSLT-2014 datasets, we seperate 7k language pairs from the training set as validset and concatenate the dev2010, tst2010, tst2011, tst2012 for the test set. The Biomedical test sets are from HimL test sets 2017 [5]. And the IT test set called Batch3 is from WMT16 IT-Domain. Noted that, the test set of BPEATT need to have the same preprocessing operation as training set.

Tasks	Training set	Validation set	Test set
WMT20 Zh→En	40M	2k	4k/2k/2k
WMT14 En↔De	4M	3k	3k
WMT09 En↔Hu	0.6M	2k	2.5k
IWSLT14 En→De	0.18M	7k	6.7k
IWSLT14 En→Ro	0.18M	7k	5.5k
IWSLT14 En→Pt-br	0.18M	7k	5.3k

Table 4: Sizes of the datasets

A.2 Setting

32k BPE operations are applied for WMT09 En→Hu, WMT14 En→De and De→En, 37k for WMT20 Zh→En. But for IWSLT14 datasets, we leverage 10k BPE operations. We preprocess all the datasets with the Moses toolkits[6]. The joint vocabulary sizes of the WMT' EN-DE and IWLST datasets are 32k and 10k respectively. The dropout rate (Srivastava et al., 2014) is 0.3 for IWSLT14 datasets and WMT14 En→De and 0.1 for others' datasets. The model is optimized with Adam (Kingma and Ba, 2015) (0.9, 0.98). Label smoothing (Pereyra et al., 2017) is also used with weight 0.1.

For WMT's datasets, we leverage Transformer-base following Vaswani et al. (2017b) and for IWSLT's datasets, all the models are based on transformer_iwslt_de_en.

We use beam search (Sutskever et al., 2014) with beam size 5 and the length penalty with 0.6 for WMT datasets and 1.0 for IWSLT datasets.

A.3 Case Study

Table 5 demonstrates the effects of BPEATT on two sentences. For the example 1, there are *Sàdǐngdǐng*, *yīnyuèfēnggé* two rare words in the sentence, the baseline model translates them to "its" and "music style" respectively. However BPEATT translate the rare words correctly. In the example 2, the underline sentence *Déguó qiáomín kèláiménsī · sāixīshuō* has many rare word peices appearing simultaneously in one sentence which causes heavy confusion. On the contrary, BPEATT has the ability to handle the missing translation caused by rare words. The relationship between context and single-tokenization can contribute to missing translation. However, fusion with multi-granularity segmentation contains more sub-word information. Therefore BPEATT is expert in fusing information and achieves better performance to the simultaneous appearance of rare words.

[5] https://www.himl.eu/test-sets

[6] https://github.com/moses-smt/mosesdecoder

Source-1	Sàdǐngdǐng píngjiè dútède yīnyuèfēnggé jiāméng huánqiú chàngpiàn
BPE [BPE-Dropout]	Sàdǐng@@ dǐng [Sà@@ dǐng@@ dǐng] píngjiè dútède yīnyuè@@ fēnggé [yīnyuè@@ fēng@@ gé] jiāméng huánqiú chàngpiàn
Reference-1	Sa Dingding, with her unique musical style, joined Universal Music Group
BPE-1	With its unique music style , he joined Universal
BPE-Dropout-1	With his unique musical style, he joined Universal Records.
BPEATT-1	**Sa Dingding** joined Universal Records in a unique **musical style**
Source-2	Déguó qiáomín kèláiménsī · sāixīshuō : " rúguǒ wǒnéng zài nàer chénggōng, wǒ yǐhòu zài rènhédìfāng dōunéng chénggōng " tā yǐnyòng deshì gēshǒu fúlánkè xīnàtèlā gēsòng niǔyuēshì de gēcí
BPE [BPE-Dropout]	Déguó qiáomín kèlái@@ ménsī [kè@@ lái@@ mén@@ sī] · sāixī@@ shuō [sāi@@ xī@@ shuō] : " rúguǒ wǒnéng zài nàer chénggōng, wǒ yǐhòu zài rènhédìfāng dōunéng chénggōng " tā yǐnyòng deshì gēshǒu fúlánkè xī@@ nàtèlā [xī@@ nà@@ tèlā] gēsòng niǔyuēshì de gēcí
Reference-2	The German expatriate Clemens Sage said: "If I can succeed there, I can succeed anywhere." He quoted a song written by singer Frank Sinatra to praise New York City.
BPE-2	"If I can do it there, I can do it anywhere," he said, quoting singer Frank Sinatra's lyrics to the praises of New York City .
BPE-Dropout-2	"If I can succeed there, I can do it anywhere," he said, quoting singer Frank Sinatra's lyrics to the praises of New York City .
BPEATT-2	"If I can succeed there, I can succeed anywhere in the future," said German diaspora **Clemens Sage**, who quoted singer Frank **Sinatra** as praising New York City lyrics.

Table 5: Two examples of BPEATT translation. Red font indicates the rare words. Bold font indicates output by the proposed BPEATT model. The wave indicates the missing translation parts of the BPE/BPE-Dropout model's output.