Word, Subword or Character? An Empirical Study of Granularity in Chinese-English NMT

Yining Wang[†], Long Zhou[†], Jiajun Zhang[†], Chengqing Zong[†‡]

[†]University of Chinese Academy of Sciences
National Laboratory of Pattern Recognition, CASIA
[‡]CAS Center for Excellence in Brain Science and Intelligence Technology
{yining.wang,long.zhou,jjzhang,cqzong}@nlpr.ia.ac.cn

Abstract. Neural machine translation (NMT), a new approach to machine translation, has been proved to outperform conventional statistical machine translation (SMT) across a variety of language pairs. Translation is an open-vocabulary problem, but most existing NMT systems operate with a fixed vocabulary, which causes the incapability of translating rare words. This problem can be alleviated by using different translation granularities, such as character, subword and hybrid word-character. Translation involving Chinese is one of the most difficult tasks in machine translation, however, to the best of our knowledge, there has not been any other work exploring which translation granularity is most suitable for Chinese in NMT. In this paper, we conduct an extensive comparison using Chinese-English NMT as a case study. Furthermore, we discuss the advantages and disadvantages of various translation granularities in detail. Our experiments show that subword model performs best for Chinese-to-English translation while hybrid word-character model is most suitable for English-to-Chinese translation.

1 Introduction

Neural machine translation (NMT) proposed by Kalchbrenner and Blunsom [9] and Sutskever et al. [22] has achieved significant progress in recent years. Unlike traditional statistical machine translation(SMT) [11, 3, 26] which contains multiple separately tuned components, NMT builds an end-to-end framework to model the entire translation process. For several language pairs, NMT has already achieved better translation performance than SMT [25, 8].

Conventional NMT system limits the vocabulary to a modest-sized vocabulary in both sides and words out of vocabulary are replaced by a special **UNK** symbol. However, the process of training and decoding is often conducted on an open vocabulary, in which an obvious problem is that NMT model is incapable of translating rare words. In particular, if a source word is outside the source vocabulary or its translation is outside the target vocabulary, the model is unable to generate proper translation for this word during decoding. Both Sutskever et al. [22] and Bahdanau et al. [1] have observed that sentences with many out-of-vocabulary words tend to be translated much more poorly than sentences mainly containing frequent words.

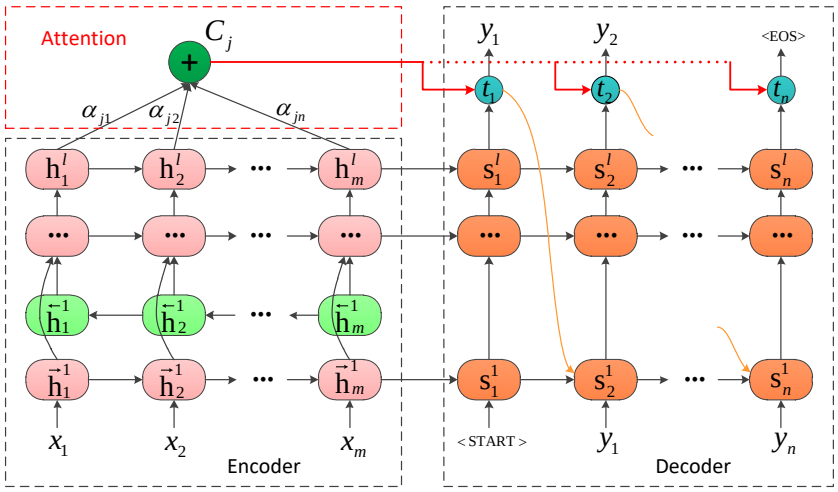

Fig. 1. The architecture of neural machine translation model.

To address this problem, many researchers propose a broad category of approaches by employing different translation granularities. Most of these are below the word level, e.g. characters [4], hybrid word-characters [13, 25], and more intelligent subwords [20, 25]. Besides, pioneering studies [25, 8] demonstrate that translation tasks involving Chinese are some of the most difficult problems in NMT systems. However, there is no study that shows which translation granularity is suitable for Chinese-to-English and English-to-Chinese translation tasks.

In this work, we make an empirical comparison of different translation granularities for bidirectional English-Chinese translation tasks. In addition, we analyze the impact of these strategies on the translation results in detail. We demonstrate that subword model is the best translation granularity for Chinese-to-English NMT while hybrid word-character model is most suitable for English-to-Chinese translation. Our experiment shows that all subword methods are not bounded by the vocabulary size. Furthermore, we carry out the experiments that employ different translation granularities of source side and target side. The translation result shows that when the source granularity is hybrid word-character level and the target sentences are split into subword level by BPE method, it can achieve the best translation performance for Chinese-to-English translation task. To the best of our knowledge, this is the first work on an empirical comparison of various translation granularities for bidirectional Chinese-English translations.

2 Neural Machine Translation

Our models are based on an encoder-decoder architecture with attention mechanism proposed by Luong et al. [14], which utilizes stacked LSTM layers for both

encoder and decoder as illustrated in Figure 1. In this section, we make a review of NMT framework.

First, the NMT encodes the source sentence $X = (x_1, x_2, ..., x_m)$ into a sequence of context vector representation $C = (h_1, h_2, ..., h_m)$. Then, the NMT decodes from the context vector representation C and generates target translation $Y = (y_1, y_2, ..., y_n)$ one word each time by maximizing the probability of $p(y_j|y_{<j}, C)$. Next, We review the encoder and decoder frameworks briefly.

Encoder: The context vector representation $C = (h_1^l, h_2^l, ..., h_m^l)$ is generated by the encoder using l stacked LSTM layers. Bi-directional connections are used for the bottom encoder layer, and h_i^1 is a concatenation vector as shown in Eq. (1):

$$h_i^1 = \begin{bmatrix} \overrightarrow{h}_i^1 \\ \overleftarrow{h}_i^1 \end{bmatrix} = \begin{bmatrix} LSTM(\overrightarrow{h}_{i-1}^1, x_i) \\ LSTM(\overleftarrow{h}_{i-1}^1, x_i) \end{bmatrix} \tag{1}$$

All other encoder layers are unidirectional, and h_i^k is calculated as follows:

$$h_i^k = LSTM(h_{i-1}^k, h_i^{k-1}) \tag{2}$$

Decoder: The conditional probability $p(y_j|y_{<j}, C)$ is formulated as

$$p(y_j|Y_{<j}, C) = p(y_j|Y_{<j}, c_j) = softmax(W_s t_j) \tag{3}$$

Specifically, we employ a simple concatenation layer to produce an attentional hidden state t_j:

$$t_j = tanh(W_c[s_j^l; c_j] + b) = tanh(W_c^1 s_j^l + W_c^2 c_j + b) \tag{4}$$

where s_j^l denotes the target hidden state at the top layer of a stacking LSTM. The attention model calculates c_j as the weighted sum of the source-side context vector representation, just as illustrated in the upper left corner of Figure 1.

$$c_j = \sum_{i=1}^{m} ATT(s_j^l, h_i^l) \cdot h_i^l = \sum_{i=1}^{m} \alpha_{ji} h_i^l \tag{5}$$

where α_{ji} is a normalized item calculated as follows:

$$\alpha_{ji} = \frac{exp(h_i^{l^T} \cdot s_j^l)}{\sum_{i'} exp(h_{i'}^{l^T} \cdot s_j^l)} \tag{6}$$

s_j^k is computed by using the following formula:

$$s_j^k = LSTM(s_{j-1}^k, s_j^{k-1}) \tag{7}$$

If $k = 1$, s_j^1 will be calculated by combining t_{j-1} as feed input [14]:

$$s_j^1 = LSTM(s_{j-1}^1, y_{j-1}, t_{j-1}) \tag{8}$$

Sentence:	龙年新春，繁花似锦的深圳处处洋溢着欢乐祥和的气氛。
Word:	龙年 新春， 繁花似锦 的 深圳 处处 洋溢 着 欢乐 祥和 的 气氛。
Character:	龙 年 新 春， 繁 花 似 锦 的 深 圳 处 处 洋 溢 着 欢 乐 祥 和 的 气 氛。
Hybrid :	<B>龙 <E>年 新春， <B>繁 <M>花 <M>似 <E>锦 的 深圳 处处 洋溢 着 欢乐 祥和 的 气氛。
BPE:	龙年 新春 ， 繁花@@ 似@@ 锦 的 深圳 处处 洋溢 着 欢乐 祥和 的 气氛 。
Wordpiece:	▁龙 年 ▁新春 ▁， ▁繁花 似 锦 ▁的 ▁深圳 ▁处处 ▁洋 溢 ▁着 ▁欢乐 ▁祥 和 ▁的 ▁气氛 ▁。

Fig. 2. An example of different translation granularities

Given the bilingual training data $D = \{(X^{(z)}, Y^{(z)})\}_{z=1}^{Z}$, all parameters of the attention-based NMT are optimized to maximize the following conditional log-likelihood:

$$L(\theta) = \frac{1}{Z}\sum_{z=1}^{Z}\sum_{j=1}^{n} logp(y_j^{(z)}|y_{<j}^{(z)}, X^{(z)}, \theta) \tag{9}$$

3 Description of Different Translation Granularities

We revisit how the source and target sentences (X and Y) are represented in NMT. For the source side of any given training corpus, we scan through the whole corpus to build a vocabulary V_x of unique tokens. A source sentence $X = (x_1, x_2, ..., x_m)$ is then built as a sequence of the integer indices. The target sentence is similarly transformed into a target sequence of integer indices.

The property of NMT allows us great freedom in the choice of token units, and we can segment sentences in different ways. In this section, we will elaborate on four proposed approaches about the choice of translation granularities.

3.1 Character Level

This translation granularity is easy to implement. For this granularity, what we have to do is split the sentence into a sequence of characters. However, the character-level modeling on the English side is more challenging, as the network has to be able to deal with long and coherent sequence of characters. In this case, the number of characters is often 300∼1000 symbols long, where the size of the state space grows exponentially. Therefore, this is a great challenge for us to handle.

Besides, the alphabet of English is only consist of 26 letters, in which the vocabulary of English side is too small. Considering these facts, we only separate

the Chinese side sentences into characters rather than both sides. Figure 2 shows an example of this translation granularity for character level.

3.2 Hybrid Word-Characters Level

In regular word-based NMT, for all words outside the source vocabulary, one feeds the universal embedding representing **UNK** as input to the encoder. This is problematic because it discards valuable information about the source word. To address that, hybrid word-character approach will be adopted. In this part, we will introduce this granularity in detail.

Unlike in the conventional word model where out-of-vocabulary words are collapsed into a single **UNK** symbol, we convert these words into the sequence of constituent characters. Special prefixes are prepended to the characters. The purpose of the prefixes is to show the location of the characters in a word, and to distinguish them from normal in-vocabulary characters. There are three prefixes: ⟨**B**⟩, ⟨**M**⟩, and ⟨**E**⟩, indicating beginning of the word, middle of the word and end of the word, respectively. During decoding, the output may also contain sequences of special tokens. With the prefixes, it is trivial to reverse the tokenization to the original words as part of a post-processing step. Using this approach, in Figure 2, we can see the word "龙年" is segmented into "⟨B⟩龙 ⟨E⟩年", and the word "繁花似锦" is segmented into "⟨B⟩繁 ⟨M⟩花 ⟨M⟩似 ⟨E⟩锦".

3.3 Subword Level

Considering languages with productive word formation processes such as agglutination and compounding, translation models require mechanisms that segment the sentence below the word level (In this paper, we call this level of symbols as subword units). In this part, we will introduce the two different methods of translation granularity on subword level.

BPE Method Byte pair encoding (BPE) [5] is a compression algorithm. This simple data compression technique iteratively replaces the most frequent pair of bytes in a sequence with a single, unused byte. This compression method is first introduced into translation granularity by Sennrich et al. [20]. In this approach, instead of merging frequent pairs of bytes, characters or character sequences will be merged.

A detailed introduction of algorithm in learning BPE operations is showed in Sennrich et al. [20]. During decoding time, each word first split into sequences of characters, then learned operation will be applied to merge the characters into larger, known symbols. For BPE method, a special symbol is also needed to indicate the merging position. In Figure 2, the word "繁花似锦" is segmented into three subword units, and the first three units are appended a special suffix "@@". In decoding step, the translation results contain the special tokens as well. With these suffixes, we can recover the output easily.

WPM Method The wordpiece model (WPM) implementation is initially developed to solve a Japanese/Korean segmentation problem for the speech recognition system [18]. This approach is completely data-driven and guaranteed to generate a deterministic segmentation for any possible sequence of characters, which is similar to the above method.

The wordpiece model is generated using a data-driven approach to maximize the language-model likelihood of the training data, given an evolving word definition. The training method of WPM is described in more detail in Schuster and Nakajima [18]. As shown in Figure 2, a special symbol is only prepended at the beginning of the words. In this case, the words "龙年", "繁花似锦", "洋溢" and "祥和" are split into subwords, and the rest words remain the same except for a special prefix "_".

4 Experiments

4.1 Dataset

We perform all these translation granularities on the NIST bidirectional Chinese-English translation tasks. The evaluation metric is BLEU [17] as calculated by the `multi-bleu.perl` script.

Our training data consists of 2.09M sentence pairs extracted from LDC corpus[1]. Table 1 shows the detailed statistics of our training data. To test different approaches on Chinese-to-English translation task, we use NIST 2003(MT03) dataset as the validation set, and NIST 2004(MT04), NIST 2005(MT05), NIST 2006(MT06) datasets as our test sets. For English-to-Chinese translation task, we also use NIST 2003(MT03) dataset as the validation set, and NIST 2008(MT08) will be used as test set.

Table 1. The characteristics of our training dataset on the LDC corpus.

Corpora		Chinese	English
LDC corpora	#Sent.	2.09M	
	#Word	43.14M	47.73M
	Vocab	0.39M	0.23M

4.2 Training Details

We build the described models modified from the Zoph_RNN[2] toolkit which is written in C++/CUDA and provides efficient training across multiple GPUs. Our training procedure and hyper parameter choices are similar to those used

[1] The corpora include LDC2000T50, LDC2002T01, LDC2002E18, LDC2003E07, LDC2003E14, LDC2003T17 and LDC2004T07.

[2] https://github.com/isi-nlp/Zoph_RNN

by Luong et al. [14]. In the NMT architecture as illustrated in Figure 1, the encoder has three stacked LSTM layers including a bidirectional layer, followed by a global attention layer, and the decoder contains two stacked LSTM layers followed by the softmax layer.

The word embedding dimension and the size of hidden layers are all set to 1000. We limit the maximum length in training corpus to 120. Parameter optimization is performed using both stochastic gradient descent(SGD) method and Adam method [10]. For the first three epoches, We train using the Adam optimizer and a fixed learning rate of 0.001 without decay. For the remaining six epoches, we train using SGD, and we set learning rate to 0.1 at the beginning and halve the threshold while the perplexity go up on the development set. We set minibatch size to 128. Dropout was also applied on each layer to avoid overfitting, and the dropout rate is set to 0.2. At test time, we employ beam search with beam size b = 12.

4.3 Data Segmentation

For Chinese word segmentation, we use our in-house segmentation tools. For English corpus, the training data is tokenized with the Moses tokenizer. We carry out Chinese-to-English translation experiment on 30k vocabulary and 15k vocabulary for both sides respectively, and we also conduct English-to-Chinese translation experiment on 30k vocabulary size. The word level translation granularity is set to our baseline method.

For character level, we only segment the Chinese sentences into characters and the English sentences remain the same. For hybrid word-characters level, we segment training corpus for both sides. We rank the word frequency from greatest to least in training corpus, and in order to prevent the pollution from the very rare word, we have to set a segmentation point relatively higher. For 30k vocabulary, the word frequency below 64 is segmented into characters on Chinese side, and the segmentation point is set to 22 on English side. For 15k vocabulary, we set the segmentation point to 350 and 96 on Chinese side and English side respectively. For 60k vocabulary, the frequency of Chinese words below 14 and that of English words below 6 are split into characters.

For subword level, two different approaches are used. In BPE method[3], the number of merge operations is set to 30000 on 30k vocabulary size, 15000 on 15k vocabulary size and 60000 on 60k vocabulary size. For Chinese sentences, we segment the training corpus using our in-house segmentation tools first, and then we can apply the BPE method same as English sentences. Considering the essence of WPM method[4], we do not have to segment words for Chinese and tokenize sentences for English. That is to say, we can train the WPM without pre-processing step. Hence, for WPM method, we conduct our experiments both on the sentences trained on the raw corpus and the sentences trained on the segmented corpus.

[3] https://github.com/rsennrich/subword-nmt
[4] https://github.com/google/sentencepiece

4.4 Results on Chinese-to-English Translation

30k Vocabulary Size We list the BLEU scores of different translation granularities on 30k vocabulary in Table 2.

Table 2. Translation results (BLEU score) of 30k vocabulary for Chinese-to-English translation.

Segmentation (30k)	MT03(dev)	MT04	MT05	MT06	Ave
Word level	41.48	43.67	41.37	41.92	42.11
Character level	42.72	44.12	41.29	41.83	42.49
Hybrid word-characters level	43.24	45.18	**42.96**	42.89	43.57
BPE method	43.78	**45.47**	42.37	**43.37**	**43.75**
WPM method (raw)	41.96	43.38	40.84	40.98	41.79
WPM method	**44.12**	44.96	42.34	42.18	43.40

Row 1 is translation result of the state-of-the-art NMT system with word level. For the character level granularity (Row 2), the translation quality is higher than the word level by only 0.38 BLEU points. The last three lines in Table 2 are subword level translation granularity, which contains BPE method and WPM method. BPE method (Row 4) achieves the best translation performance, which gets an improvement of 1.64 BLEU points over the word level. As for the WPM method (Row 6), the gap between this method and BPE method is narrow. Moreover, hybrid word-character level model (Row 3) outperforms the word level by 1.46 BLEU points, and translation quality of this method is very close to the BPE method. Experiments show that hybrid word-character level granularity and BPE method of subword level granularity are our choices for translation granularity on Chinese-to-English translation task.

Comparison in Sentences of Different Lengths We execute different translation granularities on the training corpus. To make a comparison, We randomly choose 10000 sentences. Table 3 show the average sentence length of different methods on all granularities.

Table 3. Sentence length of different translation granularities.

Language	word	character	Hybrid	BPE	WPM	WPM(raw)
Source(Chinese)	20.60	33.84	22.07	21.56	22.13	18.17
Target(English)	22.85	22.85	25.00	23.52	24.43	23.85

A well-known flaw of NMT model is the inability to properly translate long sentences. However, most of translation granularities will go below the word level. Therefore, as shown in Table 3, we can get longer sentences than the word

level. We wonder what the translation performance of different lengths are on all translation granularities. We follow Bahdanau et al. [1] to group sentences of similar lengths together and compute a BLEU score per group, as demonstrated in Figure 3.

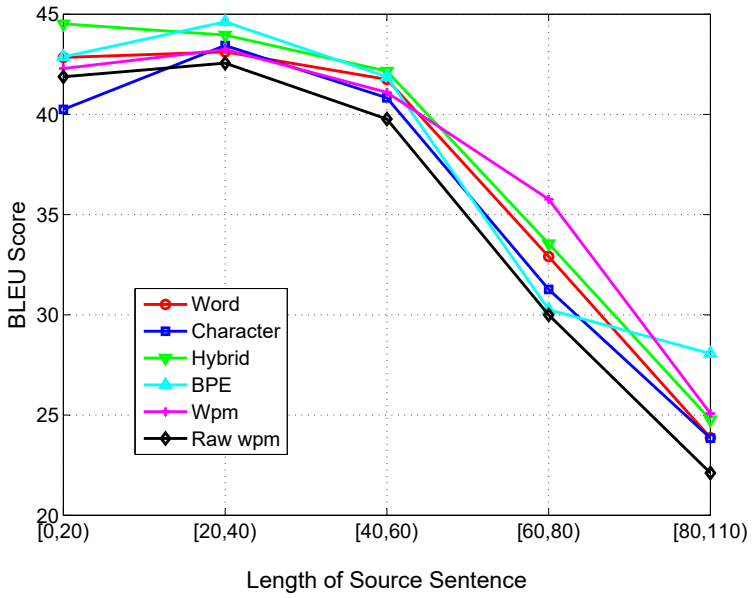

Fig. 3. Length Analysis - translation qualities(BLEU score) of different lengths.

In order to make the comparison fair, length refers to the number of tokens split in word level. As above mentioned, hybrid word-character level model is one of suitable granularity choices for Chinese-to-English translation. We can find when the length of sentences is below 20, the translation result of this model outperforms the other models to a great extent. But with the length going up, the advantage over other models is diminishing. The character level granularity performs bad for the sentences whose length are below 20. We think the reason may be that when the sentences are short, the representation of sentence in character level cannot express the sentence meaning well. As for BPE method, we find a strange phenomenon. When the number of words in source sentence is from 60 to 80, the translation performance of BPE method is not so good. However, this method can achieve almost 3 BLEU points higher than next-best approach when the source sentence is longer than 80 words. As shown in Figure 3, we can see WPM method does not perform well lower than 60 words in source language. But when the length of sentences is between 60 and 80, this method even outperforms the BPE method by up to 5.51 BLEU points. In this

experiment, we conclude that subword model is more effective than other models in handling long sentences.

15k Vocabulary Size We concern what the translation results of different translation granularities are on smaller vocabulary size. We also carry out the experiment on Chinese-to-English task of 15k vocabulary size.

Table 4. Translation results (BLEU score) of 15k vocabulary for Chinese-to-English translation.

Segmentation (15k)	MT03(dev)	MT04	MT05	MT06	Ave
Word level	39.03	42.42	38.84	39.58	39.97
Character level	42.60	43.60	40.85	41.29	42.09
Hybrid word-characters level	43.58	44.25	42.29	42.37	43.12
BPE method	**44.17**	44.89	42.79	**42.72**	43.64
WPM method (raw)	43.31	43.62	41.63	41.23	42.46
WPM method	44.03	**45.15**	**43.05**	42.63	**43.72**

Compared to 30k vocabulary size, the translation performance of word level (Row 1) on 15k vocabulary size is reduced by 2.14 BLEU points. However, character level (Row 2) and hybrid word-character level (Row 3) achieve 42.09 and 43.12 BLEU points respectively, which is on par with quality of translation on 30k vocabulary. Both these two models exceed word level to a great extent. We infer the reason is that both character level and hybrid word-character level can represent source side and target side sentences better than the word level even if the vocabulary size is small. For subword model, translation performance of these methods remain almost the same as 30k vocabulary, which is beyond our imagination. We can find in Table 4, WPM method (Row 6) outperforms other models, and to our surprise, translation results of both WPM method and WPM methods with raw corpus (Row 5) obtain a higher BLEU points than 30k vocabulary size. We analyze the reason of this phenomenon is that the subword model is not constrained by the vocabulary size. Although the WPM method achieves the best results for the 15k vocabulary size, this method also belongs to subword level translation granularity. We can conclude that subword translation granularity is more suitable for Chinese-to-English translation task.

60k Vocabulary Size In order to make a comparison of these translation granularities on larger vocabulary size, we perform the our experiment of 60k vocabulary size on Chinese-to-English translation task.

We can find in Table 5, the word and character level (Row 1 and Row 2) on 60k vocabulary size are increased by 1.15 and 1.11 BLEU points respectively compared to 30 vocabulary size. However, to our surprise, all the translation results of subword level granularities on 60k vocabulary are below to the 30k vocabulary size. With the increase of vocabulary size, we add more fine-grained

Table 5. Translation results (BLEU score) of 60k vocabulary for Chinese-to-English translation.

Segmentation (60k)	MT03(dev)	MT04	MT05	MT06	Ave
Word level	42.92	44.42	41.99	42.48	42.95
Character level	43.01	44.38	41.35	42.44	42.80
Hybrid word-characters level	**43.84**	45.11	**43.24**	**43.68**	**43.97**
BPE method	43.17	44.75	42.85	43.23	43.50
WPM method (raw)	40.85	42.64	38.87	40.48	40.71
WPM method	43.75	**44.88**	41.49	41.55	42.92

subword segmentation units into vocabulary. We infer that large amount of subword units do not have beneficial effect on the translation results. As for hybrid word-character level, this method achieves 43.97 BLEU points, which is highest among all the translation granularities on 60k vocabulary size. Compared with Table 2, hybrid word-character level outperforms the best translation result on 30k vocabulary size (BPE method) by 0.22 BLEU points.

Different Granularities on Both Sides We also conduct experiments that we use different translation granularities on source and target side. In order to carry out the experiments easily, we only compare several granularities pairs.

Table 6. Translation results (BLEU score) of 30k vocabulary for different granularities on Chinese-to-English translation.

Segmentation (30k)	MT03(dev)	MT04	MT05	MT06	Ave
Word_BPE	41.11	43.63	40.57	41.87	41.80
Word_Hybrid	41.36	43.28	40.83	41.29	41.69
Hybrid_Word	43.53	44.77	42.69	42.56	43.39
Hybrid_BPE	**44.46**	**45.21**	**43.79**	**43.56**	**44.26**
BPE_Word	43.23	45.00	42.13	42.75	43.28
BPE_Hybrid	44.28	44.89	43.29	42.45	43.73

In Table 6, we can find that when the source translation granularity is word level (Row 2 and Row 3), the translation performances are relative poor, even worse than the word level of both sides in Table 2. As for BPE method on source side, the hybrid word-character on target side obtains 43.73 BLEU points (Row 6), which is close to the best translation result in Table 2. Hybrid_BPE method achieves up to 44.26 BLEU points (Row 4), which is even higher than BPE method by up to 0.51 BLEU points. This method can acquire best translation result for Chinese-to-English translation task.

4.5 Results on English-to-Chinese Translation

30k Vocabulary Size We evaluate different translation granularities on the English-to-Chinese translation tasks, whose results are presented in Table 7.

Table 7. Translation results (BLEU score) for English-to-Chinese translation.

Segmentation (30k)	MT03(dev)	MT04	MT05	MT06	MT08	Ave
Word level	17.44	21.67	18.53	19.27	22.80	19.94
Character level	18.18	20.11	17.36	18.80	23.75	19.64
Hybrid word-characters level	19.81	23.28	20.99	**21.59**	**26.13**	**22.36**
BPE method	19.43	23.23	19.77	20.24	24.30	21.39
WPM method (raw)	18.66	21.19	18.34	18.43	19.06	19.14
WPM method	**20.78**	**24.05**	**21.07**	21.54	23.27	22.14

We find that hybrid word-character level (Row 3) granularity obtains significant accuracy improvements over word level and this granularity is also superior to other granularities on large-scale English-to-Chinese translation. BPE method (Row 4) in this task does not perform well as Chinese-to-English task, the translation quality of it is lower than hybrid word-character model by up to 0.97 BLEU points. However, another subword level translation granularity WPM method (Row 6) achieves 22.14 BLEU points, which is near the hybrid word-character level. Although the vocabulary of character level on Chinese side is only 7.2k, it can also obtain 19.64 BLEU points (Row 2), which is on par with translation performance of word level.

Different Granularities on Both Sides As Chinese-to-English translation task, we carry out experiments on English-to-Chinese translation for different granularities. According to Table 6, Hybrid_BPE and BPE_Hybrid methods acquire relative higher translation quality than other methods. Therefore, in this section we only use these two methods to test which is most suitable for English-to-Chinese translation task.

Table 8. Translation results (BLEU score) of 30k vocabulary for different granularities on English-to-Chinese translation.

Segmentation (30k)	MT03(dev)	MT04	MT05	MT06	MT08	Ave
Hybrid_BPE	20.31	22.16	**20.65**	19.87	25.65	21.73
BPE_Hybrid	**20.36**	**23.35**	20.03	**20.52**	**26.35**	**22.12**

Table 8 shows that translation performances of both two methods are below to the Hybrid word-character granularity in Table 7. BPE_Hybrid method (Row 2) achieves 22.12 BLEU points, which is higher than Hybrid_BPE method by 0.39 BLEU points and is near the translation quality of WPM method in Table 7.

5 Related Work

The recently proposed neural machine translation has drawn more and more attention. Most of existing work in neural machine translation focus on handling rare words [12, 20, 15], integrating SMT strategies [6, 28, 24, 21], designing the better framework [23, 14, 16] and addressing the low resource scenario [2, 27, 19].

As for strategies for dealing with rare and unknown words, a number of authors have endeavored to explore methods for addressing them. Luong et al. [14] and Li et al. [12] propose simple alignment-based technique that can replace out-of-vocabulary words with similar words. Jean et al. [7] use a large vocabulary with a method based on importance sampling.

In addition, another direction to achieve rare words problem in NMT is changing the granularity of segmentation. Chung et al. [4] focus on handling translation at the level of characters without any word segmentation only on target side. Luong et al. [13] propose a novel hybrid architecture that combines the strength of both word and character-based models. Sennrich et al. [20] use BPE method to encode rare and unknown words as sequences of subword units. Wu et al. [25] use both WPM method and hybrid word-character model in their online translation system. However, there is no study that shows which translation granularity is suitable for translation tasks involving Chinese language. Our goal in this work is to make an empirical comparison of different translation granularities for bidirectional Chinese-English translation tasks.

6 Conclusion

In this work, we provide an extensive comparison for translation granularities in Chinese-English NMT, such as word, character, subword and hybrid word-character. We have also discussed the advantages and disadvantages of various translation granularities in detail. The experiments demonstrate that the subword model best fits Chinese-to-English translation even with smaller vocabulary size, while the hybrid word-character approach obtains the highest performance on English-to-Chinese translation. In addition, experiments on different granularities show that Hybrid_BPE method can acquire best result for Chinese-to-English translation task.

Acknowledgments

The research work has been funded by the Natural Science Foundation of China under Grant No. 61333018 and No. 61402478, and it is also supported by the Strategic Priority Research Program of the CAS under Grant No. XDB02070007.

References

1. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. In Proceedings of ICLR 2015 (2015)

2. Cheng, Y., Liu, Y., Yang, Q., Sun, M., Xu, W.: Joint training for pivot-based neural machine translation. arXiv preprint arXiv:1611.04928v2 (2017)
3. Chiang, D.: A hierarchical phrase-based model for statistical machine translation. In Proceedings of ACL 2005 (2005)
4. Chung, J., Cho, K., Bengio, Y.: A character-level decoder without explicit segmentation for neural machine translation (2016)
5. Gage, P.: A new algorithm for data compression. R & D Publications, Inc. (1994)
6. He, W., He, Z., Wu, H., Wang, H.: Improved neural machine translation with SMT features. In Proceedings of AAAI 2016 (2016)
7. Jean, S., Cho, K., Memisevic, R., Bengio, Y.: On using very large target vocabulary for neural machine translation. Computer Science (2014)
8. Junczys-Dowmunt, M., Dwojak, T., Hoang, H.: Is neural machine translation ready for deployment? A case study on 30 translation directions. In Proceedings of IWSLT 2016 (2016)
9. Kalchbrenner, N., Blunsom, P.: Recurrent continuous translation models. In Proceedings of EMNLP 2013 (2013)
10. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. Computer Science (2014)
11. Koehn, P., Och, F.J., Marcu, D.: Statistical phrase-based translation. In Proceedings of ACL-NAACL 2013 (2003)
12. Li, X., Zhang, J., Zong, C.: Towards zero unknown word in neural machine translation. In Proceedings of IJCAI 2016 (2016)
13. Luong, M.T., Manning, C.D.: Achieving open vocabulary neural machine translation with hybrid word-character models (2016)
14. Luong, M.T., Pham, H., Manning, C.D.: Effective approaches to attention-based neural machine translation. In Proceedings of EMNLP 2015 (2015)
15. Luong, M.T., Sutskever, I., Le, Q.V., Vinyals, O., Zaremba, W.: Addressing the rare word problem in neural machine translation. In Proceedings of ACL 2015 (2015)
16. Meng, F., Lu, Z., Li, H., Liu, Q.: Interactive attention for neural machine translation. In Proceedings of COLING 2016 (2016)
17. Papineni, K., Roukos, S., Ward, T., Zhu, W.: Bleu: a methof for automatic evaluation of machine translation. In Proceedings of ACL 2002 (2002)
18. Schuster, M., Nakajima, K.: Japanese and korean voice search 22(10), 5149–5152 (2012)
19. Sennrich, R., Haddow, B., Birch, A.: Improving neural machine translation models with monolingual data. In Proceedings of ACL 2016 (2016)
20. Sennrich, R., Haddow, B., Birch, A.: Neural machine translation of rare words with subword units. In Proceedings of ACL 2016 (2016)
21. Shen, S., Cheng, Y., He, Z., He, W., Wu, H., Sun, M., Liu, Y.: Minimum risk training for neural machine translation. In Proceedings of ACL 2016 (2016)
22. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. In Proceedings of NIPS 2014 (2014)
23. Tu, Z., Lu, Z., Liu, Y., Liu, X., Li, H.: Modeling coverage for neural machine translation. In Proceedings of ACL 2016 (2016)
24. Wang, X., Lu, Z., Tu, Z., Li, H., Xiong, D., Zhang, M.: Neural machine translation advised by statistical machine translation. In Proceedings of AAAI 2017 (2017)
25. Wu, Y., Schuster, M., Chen, Z., Le, Q.V., Mohammad Norouzi, e.a.: Googles neural machine translation system: bridging the gap between human and machine translation. arXiv preprint arXix:1609.08144 (2016)

26. Zhai, F., Zhang, J., Zhou, Y., Zong, C., et al.: Tree-based translation without using parse trees. In Proceedings of COLING 2012 (2012)
27. Zhang, J., Zong, C.: Bridging neural machine translation and bilingual dictionaries. arXiv preprint arXiv:1610.07272 (2016)
28. Zhou, L., Hu, W., Zhang, J., Zong, C.: Neural system combination for machine translation. arXiv preprint arXiv:1704.06393 (2017)