Hindawi
Mobile Information Systems
Volume 2021, Article ID 5744248, 7 pages
https://doi.org/10.1155/2021/5744248

Research Article

Research on Uyghur-Chinese Neural Machine Translation Based on the Transformer at Multistrategy Segmentation Granularity

Zhiwang Xu,[1,2] Huibin Qin ,[1] and Yongzhu Hua[1]

[1]*Institute of Electron Device and Application, Hangzhou Dianzi University, Hangzhou 310018, Zhejiang, China*
[2]*Shaoxing University Yuanpei College, Shaoxing 312000, Zhejiang, China*

Correspondence should be addressed to Huibin Qin; xzw1985@hdu.edu.cn

Received 12 May 2021; Revised 5 June 2021; Accepted 20 June 2021; Published 28 June 2021

Academic Editor: Sang-Bing Tsai

In recent years, machine translation based on neural networks has become the mainstream method in the field of machine translation, but there are still challenges of insufficient parallel corpus and sparse data in the field of low resource translation. Existing machine translation models are usually trained on word-granularity segmentation datasets. However, different segmentation granularities contain different grammatical and semantic features and information. Only considering word granularity will restrict the efficient training of neural machine translation systems. Aiming at the problem of data sparseness caused by the lack of Uyghur-Chinese parallel corpus and complex Uyghur morphology, this paper proposes a multistrategy segmentation granular training method for syllables, marked syllable, words, and syllable word fusion and targets traditional recurrent neural networks and convolutional neural networks; the disadvantage of the network is to build a Transformer Uyghur-Chinese Neural Machine Translation model based entirely on the multihead self-attention mechanism. In CCMT2019, dimension results on Uyghur-Chinese bilingual datasets show that the effect of multiple translation granularity training method is significantly better than the rest of granularity segmentation translation systems, while the Transformer model can obtain higher BLEU value than Uyghur-Chinese translation model based on Self-Attention-RNN.

1. Introduction

Machine translation is an important branch of artificial intelligence and natural language processing means automatically with a natural language sequence $X = \{x_1, x_2, \ldots, x_n\}$ turning into another sequence $Y = \{y_1, y_2, \ldots, y_m\}$ having the same natural language semantics process. Machine translation can be divided into rule-based machine translation, instance-based machine translation, statistics-based machine translation, and neural network-based machine translation [1].

Both statistical machine translation and neural network machine translation rely on large-scale bilingual parallel corpus. The Transformer [2] model used in this paper has a good translation effect in resource-rich languages, but in the series of small language translation tasks such as Uyghur, there is a problem of insufficient parallel corpus, which is difficult to meet the training needs of the Transformer model. At present, due to the lack of Uyghur-Chinese parallel corpus, lack of resources, and the low quality of some of the existing data, there is a serious resource asymmetry and imbalance between Uyghur and Chinese. Secondly, Uyghur is a typical sticky language with complex shapes, words are composed of stems and affixes, and the same stem and different affixes constitute different new words. Therefore, the recognition and translation for Uyghur language have a data sparseness and OOV (out of vocabulary) problem during language training.

So, this paper aims at the task of Uyghur-Chinese translation under the scare resources situation and compares four strategies of syllable segmentation, marked syllable segmentation, word segmentation, and syllable word fusion segmentation through experiments of the translation quality of the next model.

2. Transformer Model

The Transformer model depends on the attention mechanism and also uses encoder-decoder architecture, but its structure is more complicated than attention. The encoding end is composed of 6 encoders stacked together, and the decoding end is the same. Each encoder contains two layers: a self-attention layer and a feed-forward neural network. Self-attention can help the current node to focus on the current word, so as to obtain the semantics of the context. Each decoder also comprises two network encoders mentioned, the two layers are in an intermediate layer and a layer of attention, and the current node helps to obtain the current contents key concern. The structure of Transformer corresponds exactly to the input and output of the translation model, and the effect of different shard granularity on translation performance can be better observed under the same model, as shown in Figure 1.

Unlike conventional mainstream-based machine translation of Seq2Seq based on RNN model framework, the Transformer framework replaced the RNN with an attention mechanism to build the entire model framework; the Transformer framework is still an encoder-decoder structure. The encoder on the left of Figure 1 is composed of a multihead attention network and a simple fully connected feed-forward neural network. A residual connection is added between the two networks, and for the layer standardization operation, the decoder on the right is composed of two multihead attention networks and a fully connected feed-forward network. It also uses residual connection and layer standardization operations [3]. The Encoder is made up of $N = 6$ identical layers. The layers are the cells on the left of the image above, and there is a "Nx" on the far left; here is x6. Each layer is composed of two sublayers, namely multihead self-attention mechanism and fully connected feed-forward network. Each sublayer has a residual connection and a normalization. The decoder has the same structure as Encoder, but adds a sublayer of attention. During the training, all the decodes are produced at one time and the ground truth in the previous step is used for prediction. In forecasting, because there is no ground truth, we need to make predictions one by one.

Multihead attention is shown in Figure 2. The case will be appreciated that as parameter is not shared, the multiple dot product performs scaling attention by h different linear transformation Q, K, V projection. Then, the different attention results are stitched together and finally output through a linear map. This has the advantage of allowing the model to learn relevant information in different representation subspaces [4].

Through multihead attention, the model can obtain position information in different subspaces [5]. The calculation formula for multihead attention can be formulated as

$$\begin{aligned} \text{multiHead}(Q, K, V) &= \text{concat}\left(\text{head}_1, \text{head}_2, \ldots, \text{head}_h\right)W^O, \\ \text{head}_i &= \text{attention}\left(\text{QW}_i^Q, \text{KW}_i^K, \text{VW}_i^V\right), \end{aligned} \tag{1}$$

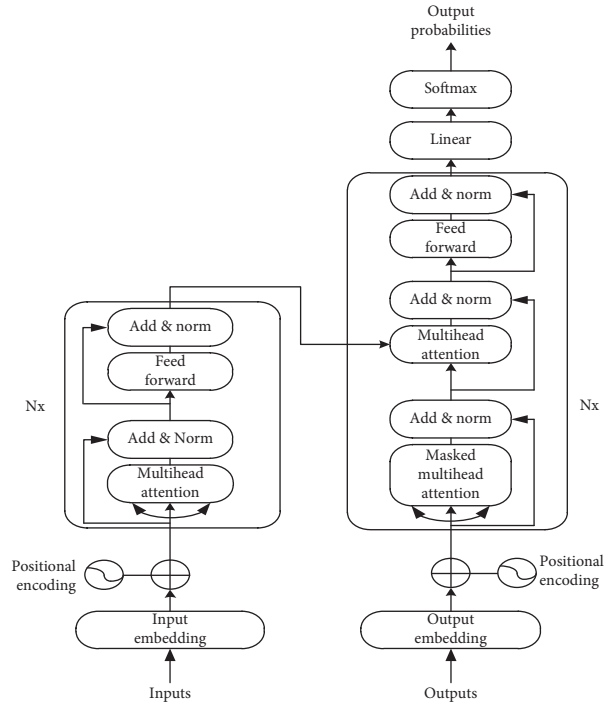

Figure 1: Transformer frame diagram.

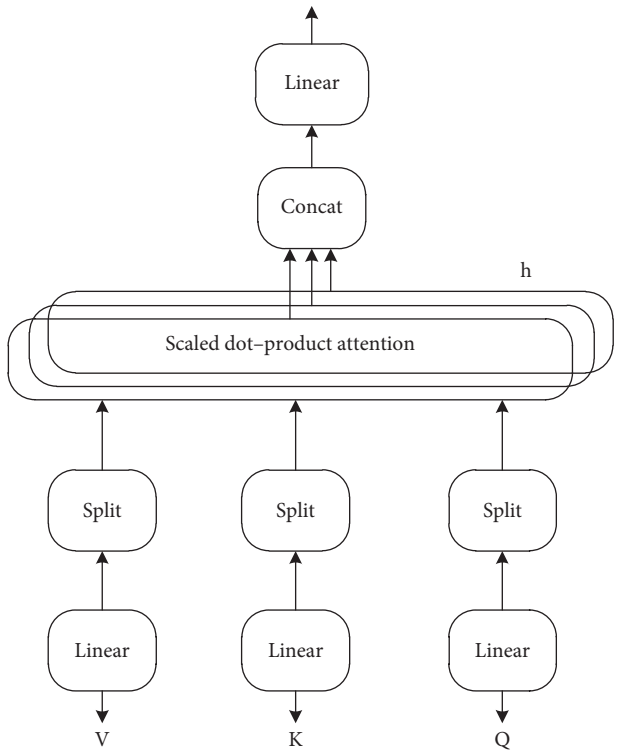

Figure 2: Structure diagram of multihead attention.

where $W^O \in R^{d_{\text{model}} \times d_k}$, $W_i^Q \in R^{d_{\text{model}} \times d_k}$, $W_i^K \in R^{d_{\text{model}} \times d_k}$, and $W_i^V \in R^{d_{\text{model}} \times d_k}$ are linearly mapped parameter matrices.

First, multihead attention is used at the encoder side to connect, K and V are the output of the encoder layer, and Q is the input of the multihead attention in the decoder. The

encoder and decoder attention to translate and align is used, and then both the encoder and the decoder use multihead self-attention to learn the representation of the text [6, 7].

When calculating attention, it is mainly divided into three steps [8]: first, the query and key are used to calculate the similarity weight; second, the Softmax function is used to normalize; third, the weight and the corresponding key value are used to value weighted summation. The calculation formula can be formulated as

$$\text{attention}(Q, K, V) = \text{softmax}(\text{sim}(Q, K))V. \quad (2)$$

The scalar dot product attention structure [9] is shown in Figure 3. Scaling the dot product attention is to use the dot product to calculate the similarity, and then divide $\sqrt{d_k}$ by the adjustment to prevent the inner product from being too large [10]. The calculation formula can be formulated as

$$\text{attention}(Q, K, V) = \text{softmax}\left(\frac{\text{QK}^T}{\sqrt{d_k}}\right)V. \quad (3)$$

Since the position of each word is related to translation, it is necessary to encode the position of each word [11]. In the RNN structure, the position information is automatically recorded in the hidden layer of the RNN cycle. In the Transformer model, because no cycle or convolution is used, in order to use the sequence information of the sequence, the relative and absolute position information needs to be inputted into the model. As the position code is introduced, the position code is applied to the input terminal; that is, the position information input at every moment in the input sequence is encoded [12, 13]. The calculation formula can be formulated as

$$\begin{aligned} \text{PE}_{2i}(p) &= \sin\left(\frac{p}{10000^{2i/d_{\text{model}}}}\right), \\ \text{PE}_{2i+1}(p) &= \cos\left(\frac{p}{10000^{2i/d_{\text{model}}}}\right). \end{aligned} \quad (4)$$

In the formula, p represents the position and d_{model} represents the dimension of the word vector. In the formula, sine or cosine functions are used according to different positions, and the vector dimension of words is controlled by d_{model} in the formula.

In Figure 1, each encoder and decoder module finally contain a fully connected feed-forward neural network, which is applied independently and identically to each location. The feed-forward network consists of performing two linear transformations on the input. The calculation formula can be formulated as

$$\text{FFN}(x) = \max(0, xW_1 + b_1)W_2 + b_2. \quad (5)$$

In the formula, x represents the input, W_1 represents the parameter matrix of the first linear transformation, b_1 represents the offset vector of the first linear transformation, W_2 represents the parameter matrix of the second linear transformation, and b_2 represents the offset vector of the second linear transformation [14].

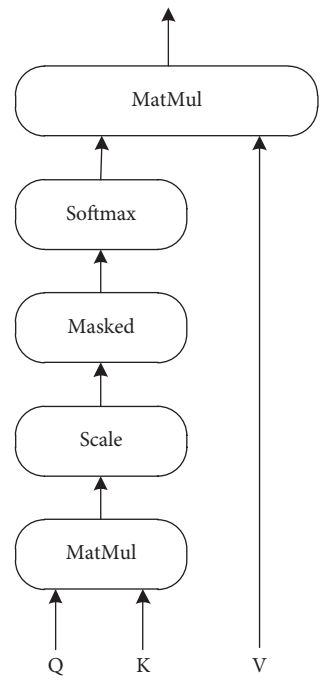

Figure 3: Scaling dot product attention structure diagram.

3. Multigranularity Segmentation

On the problem of language translation where parallel corpus is scarce, if the sparsity problem is not solved, it will seriously affect the application of neural machine translation between languages. Segmenting the corpus can reduce low-frequency words, increase the generalization ability of the model, and improve the effect of machine translation [15]. However, a large segmentation unit can save relatively complete local features, but it will aggravate the problem of data sparseness. Small segmentation granularity units can alleviate the problem of sparse data, but some local features will be lost in comparison. Therefore, this paper attempts to segment the Uyghur-Chinese bilingual corpus with different granularities to alleviate the problem of data sparseness. The granularity of syllables, marked syllables, words, and syllable word fusion is chosen for the experiment because this granularity can describe the characteristics of language very well, and they are very representative. From the experiment of this granularity, language features can be obtained in a more comprehensive way.

3.1. Syllable Strategy. There are 32 letters in Uyghur language, including 24 consonants and 8 vowels. At the same time, each letter has a different form, a total of about 130 kinds. In Uyghur, sentences are composed of one or more words, separated by spaces, and each word is composed of one or more syllables. Syllables are the smallest speech structure and the smallest speech segment that human hearing can feel naturally. Uyghur middle syllables consist of a vowel or a vowel plus multiple consonants, and each syllable contains certain semantic information. This feature is just like the composition of Chinese pinyin. Pinyin

consists of finals and initials, although there is no syllable in Chinese, but we can think of a Chinese pinyin as a syllable unit.

Uyghur syllable segmentation has certain rules: the syllable structure is (attack) + lead tone + (radio), wherein the syllables must have a neck and must be a vowel sound, and the sound from the radio may or may not be in the attack and radio. C represents consonants, V represents vowels, and there are 12 types of word syllables, which can be expressed as

$$\begin{aligned} &\text{V, VC, CV, CVC, VCC, CVCC, CCV, CCVC,} \\ &\text{CCVCC, CVV, CVVC, CCCV.} \end{aligned} \tag{6}$$

Among them, the first six are common Uyghur word syllable types, and the last six are foreign word syllable types. Generally speaking, Uyghur has the highest frequency of CV and CVC syllables.

3.2. Marked Syllable Strategy. We divide the Uyghur data into syllables with certain semantic information, and the Chinese data into single characters, which can reduce the number of translation units and increase the frequency of occurrence. The increase in the frequency of each translation unit increases the learning ability of the network model. The reduction in the number of translation units can reduce the size of the vocabulary, reduce the complex calculation of the model, and shorten the training time of the model. At the same time, it can effectively solve the problem of OOV and alleviate the data sparseness problem of Uyghur neural machine translation, thus improving the quality of translation.

3.3. Words Strategy. At present, a large number of machine translation systems are trained in terms of word level units. Syllable-level machine translation systems may encounter problems such as missing semantics or scattered data information. There is no clear separator between words in Uyghur, which makes machine learning alignment and translation more difficult. In order to obtain good characteristic information on the level of words, using word segmentation tools to segment data. Finally, manual correction is carried out.

3.4. Syllable Word Fusion Strategy. With rare language resources in training process, a large vocabulary causes low-frequency words to be represented as subword units during training, and the model also needs to learn these high-dimensional representation capabilities. For this reason, because of word segmentation, this paper divides the Uyghur text according to the Uyghur compact lattice recognition method and merges rules, statistics, and reduction. The specific steps are as follows: first, on the basis of word segmentation, first extract the words carrying pseudo-condensed syllables and use the rules to identify the syllables carried in them; if the rules cannot be recognized, then use the reduction method to recognize them, if the reduction method cannot recognize the rules, then use the maximum entropy model recognition to finally achieve the effect of syllable fusion. One feature of this fusion method is that the syllable and word-based fusion model can be controlled by changing the size of the vocabulary.

4. Test and Result Analysis

In this paper the experimental data selected machine translation evaluation (CCMT 2019) in Uyghur-Chinese parallel corpus, in which the training set has 170,000 Uyghur-Chinese parallel sentence pairs and the validation set has 1000 parallel sentence pairs. In order to analyze and compare the effect of the segmentation models based on different granularity on translation effects, four comparative experiments of syllabic granularity, marked syllable granularity, word granularity, and syllable word fusion granularity were conducted, respectively. The translation model uses Transformer model and Self-Attention-RNN model for comparison. The specific data set information is shown in Table 1.

Table 2 shows the translation of the test results of the model at different particle sizes; the scoring tool used is the value of Bilingual Evaluation understudy (BLEU).

At the same time, the experiment tested the number of training cycles required by the model to achieve a stable translation effect under different segmentation strategies. Figures 4 and 5 show how the BLEU value under the four segmentation strategies of the Self-Attention-RNN model changes as the training period increases, Figures 6 and 7 show how the BLEU value under the four segmentation strategies of the Transformer model changes as the training period increases.

Analyzing the experimental results, the following conclusions can be obtained.

(1) The Uyghur-Chinese translation model based on Transformer is indeed better than the translation model based on Self-Attention-RNN. Under the same parameters, the BLEU value can be increased by about 1.

(2) When training with syllables as the segmentation granularity, the model's translation effect is the worst. Here, due to the loss of its semantic information when syllables are segmented, the information expressed in the original text cannot be retained, and a lot of noise is introduced. Likewise, in accordance with Chinese characters, segmentation will lose semantic information carried by the original, not the corresponding dimension language in the corresponding translation, resulting in bilingual parallel corpus being ineffective, so having poor translation effects.

(3) The effect of using words as the granularity of segmentation is obviously better than that of using syllables as the granularity of granularity, because the semantic information of the original text can be relatively stable and retained at the word level. In this way, corresponding words can be found in parallel sentences better, so as to help the system obtain more semantic features and better translation effect.

Table 1: Corpus data at different granularities.

Different granularity	Training set		Test set	
	Number of sentences	Average sentence length	Number of sentences	Average sentence length
Syllable	147434	81.58	1000	46.21
Marked syllable	147434	81.65	1000	46.36
Words	147434	82.64	1000	46.97
Syllable word fusion	147434	99.69	1000	57.35

Table 2: Test results.

Task	Model	Different granularity	BLEU %
Uyghur-Chinese	Self-attention-RNN	Syllable	12.68
		Marked syllable	47.98
		Words	45.61
		Syllable word fusion	51.74
	Transformer	Syllable	14.7
		Marked syllable	48.12
		Words	46.09
		Syllable word fusion	52.21
Chinese-Uyghur	Self-attention-RNN	Syllable	8.68
		Marked syllable	61.96
		Words	60.45
		Syllable word fusion	63.41
	Transformer	Syllable	11.19
		Marked syllable	63.01
		Words	62.81
		Syllable word fusion	63.88

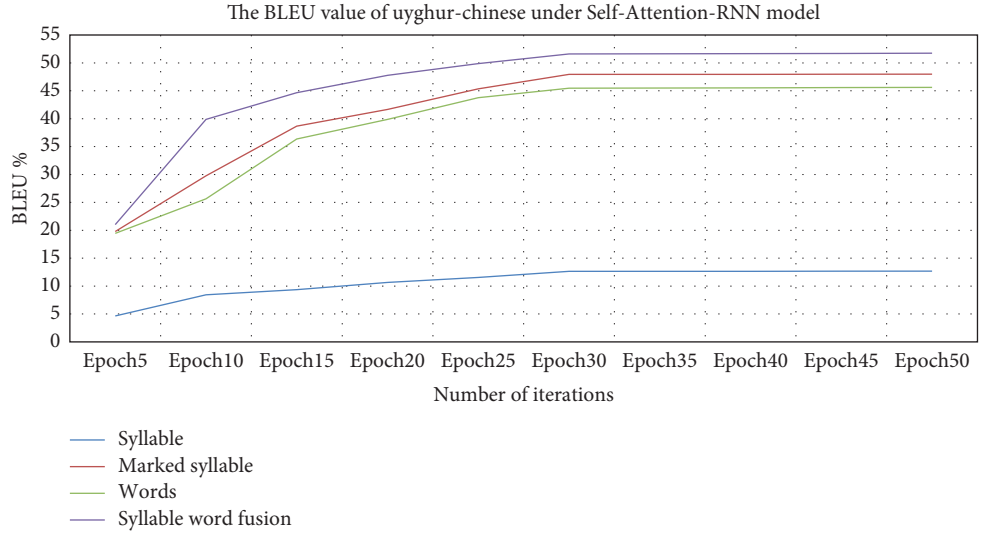

Figure 4: The relationship between training period of different segmentation strategies and BLEU under Self-Attention-RNN model (Uyghur-Chinese).

(4) Using marked syllables as the segmentation granularity is better than using words as the segmentation granularity. The Uyghur data are divided into syllables with certain semantic information. The Chinese data are divided into single characters, which can reduce the number of translation units. The frequency of occurrence increases. The increase in the frequency of each translation unit increases the learning ability of the network model. The reduction in the number of translation units can reduce the size of the vocabulary, reduce the complex calculation of the model, and shorten the training time of the model. At the same time, it can effectively solve the problem of OOV and alleviate the data sparseness problem of Uyghur neural machine translation, thus improving the quality of translation.

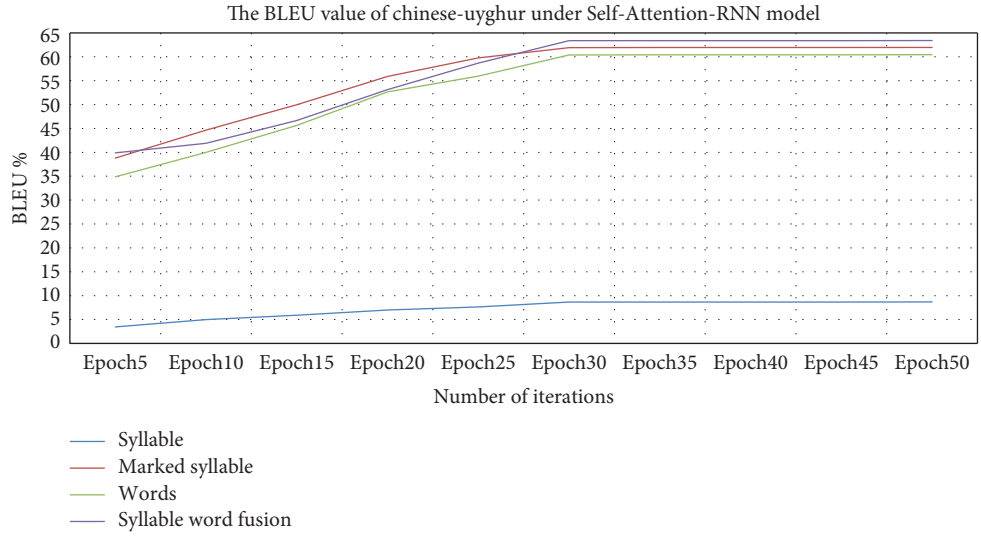

Figure 5: The relationship between training period of different segmentation strategies and BLEU under Self-Attention-RNN model (Chinese-Uyghur).

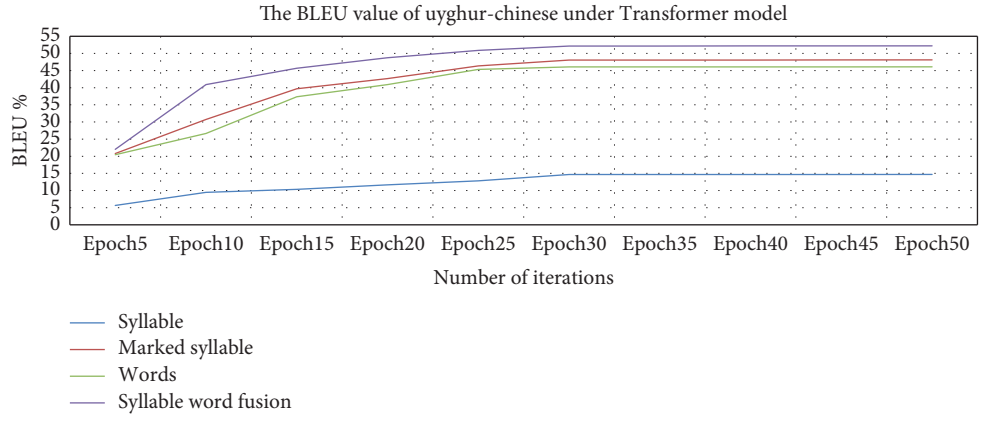

Figure 6: The relationship between training period of different segmentation strategies and BLEU under Transformer model (Uyghur-Chinese).

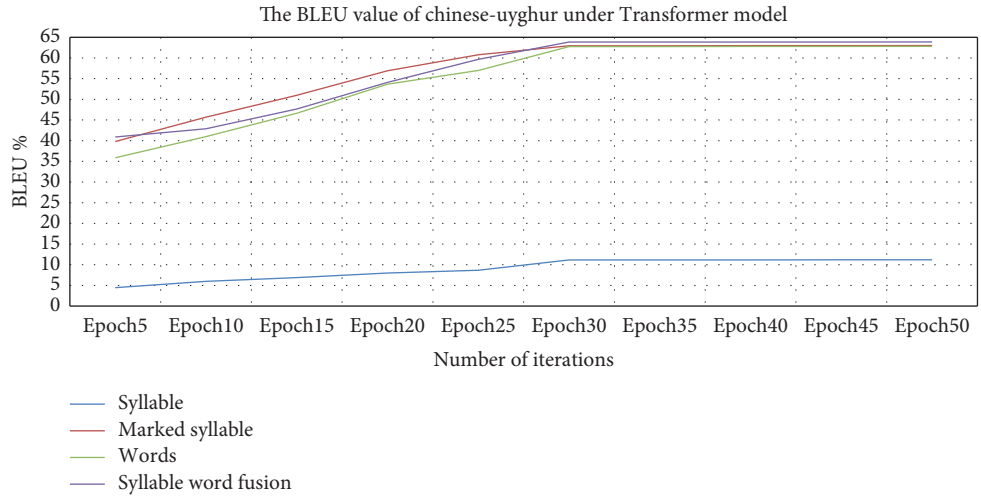

Figure 7: The relationship between training period of different segmentation strategies and BLEU under Transformer model (Chinese-Uyghur).

(5) In syllables, word segmentation size as fusion combines the advantages of syllables and words divided, as compared to the other three segmentations of BLEU values which were improved significantly.

5. Conclusion

In this paper, we study the training method of multisegment granularity for the Uyghur-Chinese translation with scarce resources. Through syllables, words, and syllable word fusion, it can effectively solve the problem of translation of prepositions and conjunctions from Chinese language but not in Uyghur and avoid translation difficulties at the vocabulary and syntactic level. Meanwhile, the low-frequency words cut into relatively high frequency subword fragment, and the data sparseness problem to alleviate the effect of the translation model is significantly improved from different corpus with different models and segmentation of different corpora under the same model. Through the above experiments, it can be concluded that different granularity cutting has a greater impact on the effect of machine translation, and subsequent attempts can be made to introduce more granularity segmentation and multigranularity fusion. In the next work, we hope to combine different granularity segmentation and improve the encoder to encode at the syllable and word level at the same time, in order to effectively obtain more feature attributes, avoid the cumbersome process in the middle, and apply the method proposed in this article to other translation tasks of sticky words.

Data Availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Conflicts of Interest

The authors declare no conflicts of interest with respect to the research, authorship, and/or publication of this article.

References

[1] X. Xu, D. Cao, Y. Zhou, and J. Gao, "Application of neural network algorithm in fault diagnosis of mechanical intelligence," *Mechanical Systems and Signal Processing*, vol. 141, Article ID 106625, 2020.

[2] A. Vaswani, N. Shazzer, N. Parmar et al., "Attention is all you need," pp. 5998–6008, 2017, https://arxiv.org/abs/1706.03762.

[3] J. l Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," 2016, https://arxiv.org/abs/1607.06450.

[4] K. Armeni, R. M. Willems, and S. L. Frank, "Probabilistic language models in cognitive neuroscience: promises and pitfalls," *Neuroscience & Biobehavioral Reviews*, vol. 83, pp. 579–588, 2017.

[5] S. Tong, P. N. Garner, and H. Bourlard, "Cross-lingual adaptation of a CTC-based multilingual acoustic model," *Speech Communication*, vol. 104, pp. 39–46, 2018.

[6] D. Bahdanau, J. Chorowski, D. Serdyuk et al., "End-to-end attention-based large vocabulary speech recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4945–4949, Shanghai, China, March 2016.

[7] C. Li, S. Zhang, P. Liu, F. Sun, J. M. Cioffi, and L. Yang, "Overhearing protocol design exploiting intercell interference in cooperative green networks," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 1, pp. 441–446, 2016.

[8] D. Yu and J. Li, "Recent progresses in deep learning based acoustic models," *IEEE/CAA Journal of Automatica Sinica*, vol. 4, no. 3, pp. 396–409, 2017.

[9] A. Vaswani, N. Shazeer, N. Parmar et al., "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 4, no. 3, pp. 5998–6008, 2017.

[10] L. Dong, S. Xu, and B. Xu, "Speech-transformer: a no-recurrence sequence-to-sequence model for speech recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5884–5888, Alberta, Canada, April 2018.

[11] R. Sennrich, B. Haddow, and A. Birch, "Neural machine translation of rare words with sub word units," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pp. 1715–1725, Berlin, Germany, August 2016.

[12] Y. Miao, M. Gowayyed, and F. Metze, "EESEN: end-to-end speech recognition using deep RNN models and WFST-based decoding," in *Proceedings of IEEE Workshop on Automatic Speech Recognition and Under-standing (ASRU)*, pp. 167–174, Scottsdale, AZ, USA, December 2015.

[13] N. Chen, B. Rong, X. Zhang, and M. Kadoch, "Scalable and flexible massive MIMO precoding for 5G H-CRAN," *IEEE Wireless Communications*, vol. 24, no. 1, pp. 46–52, 2017.

[14] Y. T. Chen, C. H. Chen, S. Wu, and C. C. Lo, "A two-step approach for classifying music genre on the strength of AHP weighted musical features," *Mathematics*, vol. 7, no. 1, p. 19, 2019.

[15] M. Ali, L. Tang Jung, A.-H. Abdel-Aty, M. Y. Abubakar, M. Elhoseny, and I. Ali, "Semantic-k-NN algorithm: an enhanced version of traditional k-NN algorithm," *Expert Systems with Applications*, vol. 151, Article ID 113374, 2020.