IJLAI Transactions on Science and Engineering (ISSN: 3006-3418) 2025, vol. 3, no. 1, pp: 59–65

A Novel Chinese-English Neural Machine Translation Model Based on BERT

Linlin Zhang[1]

School of Foreign Languages, Shenyang Normal University
110034 Shenyang, China
Received Mar. 1, 2025; Revised and Accepted Mar. 10, 2025

Abstract. In recent years, neural machine translation has rapidly developed and replaced traditional machine translation, becoming the mainstream paradigm in the field of machine translation. Machine translation can reduce translation costs and improve translation efficiency, bring good news to cultural exchanges and international cooperation, and help national development. However, neural machine translation is highly dependent on large-scale high-quality parallel corpus, and there are problems such as uneven quality and sparse data, so it is imperative to study and explore neural machine translation. The purpose of this paper is to construct pseudo-parallel corpus using data enhancement technology, improve the diversity of Chinese and English materials, and then optimize the translation model to improve the translation effect of the model. Based on BERT pre-training technology, this paper first analyzes the limitations of the traditional Transformer model, and then puts forward two directions for model optimization. On the one hand, in the data preprocessing stage, multi-granularity word segmentation technology is used for word segmentation to help Chinese-English neural machine translation model better understand the text. On the other hand, in the pre-training stage, this paper adopts the strategy of deep integration of BERT dynamic word embedding and original word embedding. On the basis of the original Transformer, a fusion module is added, through which the original word embeddings and BERT dynamic word embeddings are simple linear splicing, and then fed into the encoder. The attention mechanism is used for deep integration and better word vector representation, enabling the Transformer model to take full advantage of the external semantic information introduced by BERT. Finally, the feasibility and effectiveness of the Transformer architecture adopted in this paper are verified by the comparison experiment between RNN and Transformer model. Through the ablation experiments of different word vector representation and different stages using BERT pre-training technology, the effectiveness of BERT dynamic word embedding and deep fusion of word embedding and the rationality of using pre-training technology only in the encoder stage are verified.

Keywords: Transformer, Chinese-English Neural Machine Translation, BERT, Multi-granularity word segmentation technology.

1. Introduction

Machine translation is a popular research direction in the field of NLP. With the continuous development of deep learning, neural machine translation has become the mainstream instead of statistics-based machine translation. In the development of neural machine translation, the most basic and core encoder-decoder architecture has not changed. At the World Machine Translation Evaluation Conference (WMT) [1,2], previous champions of translation tasks in China and the UK have adopted Transformer model as the backbone network. The Transformer model has few parameters and low requirements on computing power. Its parallelized input can greatly improve computing efficiency, and it can break the long-distance limit and has stronger language representation ability. However, this model also has some limitations. Considering the shortcomings of the traditional Transformer model, this chapter improves the model in terms of language granularity and word embedding [3,4].

It is found that the general training process of neural machine translation is as follows: Firstly, it is the pre-processing of corpus, which is reflected by the granularity of selected words. If the word granularity is too large, the data will be sparse. However, although the word granularity is too small, it can alleviate the above problems, but it will lose some semantic information. Therefore, we should pay attention to the pre-processing of corpus. The second is the application of pre-training techniques, since data-driven machine translation was first established, large-scale monolingual data has been used to improve translation results. Therefore, better use of monolingual corpus can improve the quality of translation, better capture the context information and solve the problem of resource scarcity, etc. After completing the two tasks of pre-processing and pre-training, the processed data is used as the input of the model, and the model can begin to train [5,6].

The attention mechanism proposed by Bahdanau [7] could solve the above problems well, and machine translation had reached a new wave. The attention mechanism enables the original sequence model to have the ability

to distinguish, that is, to pay differentiated Attention to the input information. The higher the correlation degree of the input information, the higher the attention will be received. The attention mechanism and RNN-RNN network can be trained simultaneously. Since then, scholars have focused on the optimization of neural machine translation model and the improvement of attention mechanism. Ranathunga et al. [8] improved the encoder structure, adopted bidirectional RNN for encoding, and enhanced the ability of encoder to characterize input information. In order to solve the problem of mismatch between the output results of the training stage and the test stage, Wang et al. [9] made improvements from the perspective of decoder, adopted multi-directional decoding and integrated the output content. Guerreiro et al. [10] proposed a bi-directional decoding mechanism, which could symmetrize the decoding process, to a certain extent, improve the quality of sentences generated by the model.

Erdogmus et al. [11] proposed the minimum training criterion to solve the mismatch between training and testing in the codec architecture, thereby improving the translation effect. The modeling method using past and future information proposed by Ney et al. [12] solves the problems of over-translation and missing translation in actual translation to a certain extent. In 2018, the Facebook team [13] combined the CNN network structure with the sequence-to-sequence model, and proposed a encoder-decoder framework based on CNN as the backbone network, so that the encoder and decoder could run simultaneously. The model performance was similar to that of the RNN as the backbone network, but the running speed was faster. In the same year, Google team [14] proposed a Transformer based NMT model, which did not use RNN and CNN networks, and its network architecture was composed of multi-head attention mechanism and feed-forward neural network. The multi-head attention module innovated the attention mechanism, the core of which was the self-attention mechanism. The advantage of this module was that it had a strong feature extraction ability.

1.1. Segmentation of Chinese-English Materials Based on Word Segmentation Algorithm

It is necessary to preprocess Chinese corpus by word segmentation in machine translation because: there are no Spaces or other delimiters in Chinese to define word boundaries, and the semantic information of each word can be captured by word segmentation. Chinese is rich in vocabulary and has a huge vocabulary size, which can be effectively reduced by word segmentation. Translating the phrase as a whole helps to preserve the full meaning of the phrase. In conclusion, the preprocessing of Chinese word segmentation is an indispensable step in machine translation, which can help machine translation models better understand the text and improve the quality and accuracy of translation. In this paper, stuttering word segmentation and BPE (Byte Pair Encoding) [15] are adopted, which combines the traditional dictionary-based word segmentation and data-driven sub-word segmentation technology, and can effectively deal with the word segmentation problem in Chinese corpus.

Stuttering word segmentation is a word segmentation tool based on dictionaries and rules, which has wide application and good performance. It can divide Chinese text into a series of words and identify common words, proper nouns, idioms and so on. Stuttering word segmentation takes into account the context information of the words, and it can better deal with the common words in Chinese corpus by means of statistics and rule matching. BPE is an unsupervised subword segmentation algorithm that builds richer language representations by continuously merging words in a vocabulary into larger subword units. The BPE algorithm first divides each word into a sequence of character levels [16], and then combines the character pairs with the highest frequency according to the frequency of occurrence to form a larger subword unit. BPE is an unsupervised subword segmentation algorithm that builds richer language representations by continuously merging words in a vocabulary into larger subword units. The BPE algorithm first divides each word into a sequence of character levels, and then combines the character pairs with the highest frequency according to the frequency of occurrence to form a larger subword unit. This process is iterated until a preset vocabulary size or stop guideline is reached. The specific process of BPE word segmentation algorithm is shown in Figure 1.

When Chinese corpus is segmented, the text is segmented into word sequences by stuttering word segmentation, and then BPE algorithm is applied to further segmented word sequences. The advantage of doing so is that on the basis of retaining the traditional word segmentation, the language representation can be further refined by BPE algorithm, especially for some rare words, proper nouns, etc., which can be better segmented and expressed. Word segmentation preprocessing of English corpus is also necessary in natural language processing tasks. Here are a few important reasons:

(1) Word understanding: English words are usually separated by Spaces or punctuation marks, but sometimes a word may be composed of multiple phrases, such as compounds, abbreviations, etc. Through word segmentation preprocessing, these compound words are divided into independent word units, which is helpful to understand the semantics and meaning of each word more accurately.

(2) Part of speech change: English words often have part of speech changes, such as tense, plural forms, etc. By word segmentation, the morphologies of these words can be taken into account, and their roots can be used as the basic unit to improve the recognition and matching ability of the model.

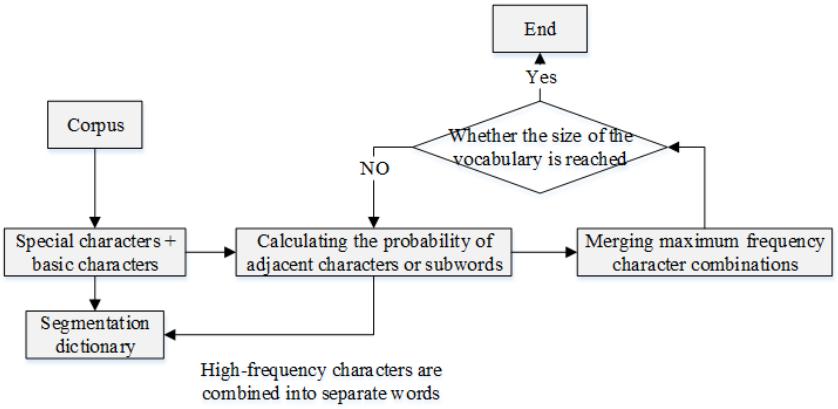

Fig. 1. BPE word segmentation algorithm flow

(3) Grammatical structure and modeling efficiency: Word segmentation can better capture the order and relationship of words in phrases and sentences, which helps the model to accurately understand the sentence grammar and contextual meaning; It can also divide the English text into discrete word units, reduce the size of the vocabulary, and improve the efficiency and accuracy of model modeling.

The BPE word segmentation algorithm is applied to the prepared English corpus, and the vocabulary is constructed by merging the most common character sequences step by step to realize English word segmentation. Figure 2 shows the specific process of obtaining a bilingual dictionary using the word segmentation algorithm in this paper.

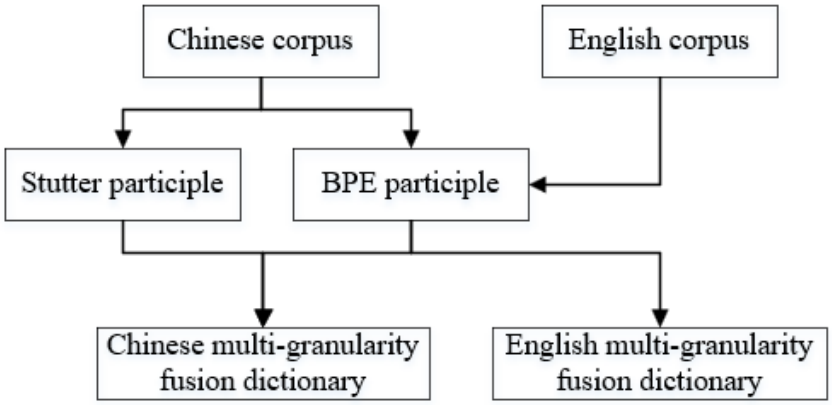

Fig. 2. Word segmentation process

1.2. BERT Pre-training Model

This paper describes Onehot and word2vec text characterization methods used by representative neural machine translation systems. Among them, word2vec is the most commonly used, which is a fixed representation based on word vectors, and can not represent the polysemy of words in practical situations. In order to solve the problem of word polysemy, scholars began to study dynamic representation methods based on word vectors. ELMo, GPT and BERT1 are proposed. Since these three methods are all large-scale pre-training models, the models can generate dynamic word vectors according to specific downstream tasks when applied. The characteristics of these three methods are shown in Table 1.

We can know that BERT model has the following advantages over ELMo and GPT in dynamic word embedding:

(1) Bidirectional context modeling: BERT is a bidirectional model, which can utilize the context information at the same time, not only depending on the left or right context. This allows BERT to better understand the meaning and relationships of words in context, leading to more accurate embedding representations;

Table 1. Pre-training model features

Pre-trained model	Model structure	Model layer	Language model
ELMO	LSTM	One layer static vector+two layers LSTM	Bidirectional language model
GPT	Transformer	Multilayer Decoder	One-way language model
BERT	Transformer	Multilayer Encoder	Bidirectional language model

(2) Deep bidirectional Transformer architecture: BERT adopts multi-layer bidirectional Transformer architecture, which enables the model to carry out more in-depth semantic modeling and feature extraction. This deep architecture enables BERT to capture more complex language structures and semantic relationships;

(3) Pre-training - fine-tuning framework: In the pre-training phase, a large number of unlabeled text data is trained and a general language representation is learned from it. In the fine-tuning stage, BERT can adapt to the feature extraction and representation learning of specific tasks through supervised fine-tuning on specific tasks. This framework allows BERT to better adapt to different downstream tasks and perform well with small amounts of labeled data;

(4) Global and local context attention: BERT introduces self-attention mechanism, which can pay attention to global and local context information at the same time. This mechanism enables BERT to handle long-distance dependencies and capture global semantic relationships, thus providing more accurate dynamic word embedding.

(5) Extensive pre-training tasks: BERT uses a variety of tasks in the pre-training phase, including masking language modeling and next sentence prediction. These tasks enable BERT to learn a wealth of language representation and semantic knowledge, providing a more comprehensive and diversified dynamic word embeddings.

To sum up, BERT has better performance and performance in dynamic word embedding than ELMO and GPT, and can accurately and effectively capture context information, semantic relations and global dependencies, so as to provide more accurate and rich word embedding representation. Therefore, the BERT model is used as an additional coding network to improve the encoding capability of the encoder, and the data of the translation model is enhanced at the word vector level.

1.3. Model Overall Architecture

In this paper, before training the Chinese-English neural machine translation model, BERT pre-training model is used to improve the model translation performance and translation speed. Firstly, source language sequences are input into BERT pre-training model and original word embedding module respectively, and two kinds of representations are obtained: dynamic word vector and original word embedding. Then the two representations are transformed linearly, and simple representation fusion is carried out. The new representation vector obtained is taken as the input of the encoder, and deep fusion is carried out through the encoder. Finally, neural machine translation is carried out. The model consists of three parts, namely, word embedding module, fusion module and encoder-decoder module, which are explained in this section.

In BERT and word embedding representation fusion module, $E_{bert-out}$ and word embedding represent $E_{embedding}$ to calculate attention, the output of the word embedding part $E_{embedding}$ as Q, $E_{bert-out}$ as K, to calculate the weight. $E_{bert-out}$ as V is multiplied with the calculated weights to establish the relationship between the original word embeddings and BERT dynamic word embeddings. The specific process of representation fusion is as follows:

$$Q = E_{embedding}. \tag{1}$$

$$V = K = E_{bert-out}. \tag{2}$$

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V. \tag{3}$$

$$E'_{bert-out} = Attention(Q, K, V). \tag{4}$$

$$Q = V = K = E_{embedding}. \tag{5}$$

$$E'_{embedding} = Attention(Q, K, V). \tag{6}$$

The new text sequence hidden state $E_{bert-embedding}$ is obtained by combining the new representation $E'_{bert-out}$ with the enhanced representation $E'_{embedding}$. The specific process of representation embedding is as follows:

$$E_{contact} = Contact('_{bert-out}, E'_{embedding}). \tag{7}$$

$$E_{bert-embedding} = Linear(E_{contact}). \tag{8}$$

2. Experiment and Result Analysis

This section describes and explains the experimental preparation and experimental results of the BERT-based Chinese-English neural machine translation model Transformer. Due to the high cost of pre-training, BERT-base-Chinese, a model developed by Tsinghua University and trained on Chinese data sets, is adopted in this paper. This model has been trained on a large number of data sets, and has rich semantic information, which can be used to represent Chinese sequences.

In order to verify the effectiveness of deep fusion between BERT dynamic word embedding and original word embedding, the following four neural machine translation methods are tested in this section. (1) RNN+Attention; (2) BERT-RNN: Neural machine translation method integrating BERT model into RNN model. In BERT model, the word vector dimension is 768, the optimizer is Adam, $\beta 1 = 0.9$, $\beta 2 = 0.999$, and the learning rate is 0.1; (3) Transformer; (4) BERT-TF: In this experiment, Transformer model uses the neural machine translation method with deep integration of BERT dynamic word embedding and original word embedding, where the parameters of BERT model are the same as those of model (2).

Table 2. Comparative experimental results

Model	BLEU value
RNN+attention	24.72
BERT+RNN	25.23
Transformer	25.89
Proposed	26.68

After analyzing the experimental results in Table 2, it can be seen that after deep integration of BERT dynamic word embeddings and original word embeddings, RNN and Transformer models increase BLEU values by 0.51 and 0.78, respectively. This shows that this operation is effective for improving the translation performance of the model. By obtaining dynamic word embedding through BERT pre-training models, richer semantic information can be obtained, thus improving the translation ability of the model. Moreover, for different backbone networks, the improved Transformer model has better translation effect than the improved RNN+Attention model. Therefore, it is feasible and effective to introduce BERT pre-training model with Transformer as the backbone network in this chapter.

2.1. Ablation Experiments

The word Embedding method of the original Transformer model adopts random initialization, using the embedding layer to randomly initialize a size $N \times d_{model}$. Where N is the size of the vocabulary. d_{model} is the model dimension, and the Trainable parameter is True. After the training, each word in the thesaurus corresponds to a unique vector representation, so the obtained word vectors cannot obtain context information and solve the polysemy problem of a word. Therefore, the performance of the Chinese-English translation model has certain space for improvement.

In order to verify the effectiveness of the deep integration of BERT dynamic word embedding and original word embedding in this paper, the following four word embedding modules are used to conduct a comparative experiment to evaluate the influence of dynamic word vectors and original static word vectors on the performance of Chinese-English neural machine translation.

(1) Embedding: Only the original word embedding is used to represent the source language;

(2) BERT: Only BERT dynamic word embedding is used to characterize the source language, where the optimizer is Adam.

(3) Embedding-bert-concat: embedding the source language Embedding and BERT dynamic words;

(4) Proposed: the Attention mechanism is used to deeply integrate the source language Embedding of the embedding of the source language Embedding and BERT dynamic word embedding.

In order to ensure the effectiveness of the experiments, the same data set is used in the above experiments, and the experimental results are shown in Table 3.

Table 3. Word embedding module fusion experiments

Word embedding module	BLEU value
Embedding	25.89
BERT	26.59
Embedding-BERT-concat	24.38
Proposed	27.68

As can be seen from Table 3, different word embeddings have different translation effects on the model. The effect of only using BERT dynamic word embedding is slightly better than that of the original word embedding. However, it only improves less than 1 BLEU value, which is far higher than the BLEU value that integrates BERT dynamic word embedding and original word embedding with attention mechanism. Meanwhile, in the experiment, the BLEU value obtained by linear concatenation of BERT dynamic word Embedding and embedding is lower than that of single word embedding. This shows that the deep integration of BERT dynamic word embeddings with original word embeddings adopted in this section can obtain better word vector representation, and can make full use of the external semantic information provided by BERT model, so as to improve the translation ability of the model.

2.2. Decoder Using BERT Dynamic Word Embedding Experiments

In order to investigate whether the pre-trained model is used in the decoding stage, the following experiments are set up in this section: (1) BERT-DEC: Only BERT dynamic word embedding is used at the decoder end; (2) BERT-ENC: BERT dynamic word embedding is used only at the encoder end; (3) BERT-ENC-DEC: BERT dynamic word embedding is used in both encoder and decoder end; (4) Transformer: The BERT pre-trained model is not used, and only the original word is embedded for the encoder input

Table 4. Experimental results of BERT dynamic word embedding at different stages

Chinese-English neural machine translation model	BLEU value
BERT-DEC	24.56
BERT-ENC	27.68
BERT-ENC-DEC	26.04
Transformer	25.89

It can be seen from Table 4 that using BERT dynamic word embedding only at the decoding end has the worst translation effect. The effect of BERT dynamic word embedding on codec side is worse than that on encoding side only. The results show that dynamic fusion at the decoder side cannot improve the translation performance of the model, which proves the feasibility and effectiveness of dynamic fusion only at the coding side in this paper.

3. Conclusion

The main content of this paper is to improve the traditional Transformer model and propose a new Chinese-English neural machine translation model. The innovation of the model is mainly embodied in the choice of word segmentation granularity and the dynamic word embedding based on BERT. Finally, the effectiveness of Transformer backbone network is verified by comparison experiments of different model architectures, and the effectiveness of BERT dynamic word embedding module is verified by two sets of ablation experiments. applied to other encryption schemes.

4. Conflict of Interest

The authors declare that there are no conflict of interests, we do not have any possible conflicts of interest.

Acknowledgments. This work was supported by the Project of the Research Center for the Theory System of Socialism with Chinese Characteristics in Shenyang Normal University, 2023. Project number: ZTSYB2023012. Project name: A Study of Xi Jinping Thought on Socialism with Chinese Characteristics in the Course of Foreign Affairs Translation.

References

1. Rivera-Trigueros I. Machine translation systems and quality assessment: a systematic review[J]. Language Resources and Evaluation, 2022, 56(2): 593-619.
2. Klimova B, Pikhart M, Benites A D, et al. Neural machine translation in foreign language teaching and learning: a systematic review[J]. Education and Information Technologies, 2023, 28(1): 663-682.
3. Peng K, Ding L, Zhong Q, et al. Towards making the most of chatgpt for machine translation[J]. arXiv preprint arXiv:2303.13780, 2023.
4. Faradiba C F, Aini N. The Use of Machine Translation for Legal Documents of University Students in English Department Class[C]//Proceeding International Conference on Religion, Science and Education. 2024, 3: 301-307.
5. Zhang B, Haddow B, Birch A. Prompting large language model for machine translation: A case study[C]//International Conference on Machine Learning. PMLR, 2023: 41092-41110.
6. Costa-Juss M R, Cross J, elebi O, et al. No language left behind: Scaling human-centered machine translation[J]. arXiv preprint arXiv:2207.04672, 2022.
7. Soydaner D. Attention mechanism in neural networks: where it comes and where it goes[J]. Neural Computing and Applications, 2022, 34(16): 13371-13385.
8. Ranathunga S, Lee E S A, Prifti Skenduli M, et al. Neural machine translation for low-resource languages: A survey[J]. ACM Computing Surveys, 2023, 55(11): 1-37.
9. Wang L, Lyu C, Ji T, et al. Document-level machine translation with large language models[J]. arXiv preprint arXiv:2304.02210, 2023.
10. Guerreiro N M, Rei R, Stigt D, et al. xcomet: Transparent machine translation evaluation through fine-grained error detection[J]. Transactions of the Association for Computational Linguistics, 2024, 12: 979-995.
11. Erdogmus D, Principe J C. Convergence properties and data efficiency of the minimum error entropy criterion in adaline training[J]. IEEE Transactions on Signal Processing, 2003, 51(7): 1966-1978.
12. Ney H. On the probabilistic interpretation of neural network classifiers and discriminative training criteria[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1995, 17(2): 107-119.
13. Li C, Zhang Z, Lee W S, et al. Convolutional sequence to sequence model for human dynamics[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 5226-5234.
14. Gupta R, Besacier L, Dymetman M, et al. Character-based nmt with transformer[J]. arXiv preprint arXiv:1911.04997, 2019.
15. Bostrom K, Durrett G. Byte pair encoding is suboptimal for language model pretraining[J]. arXiv preprint arXiv:2004.03720, 2020.
16. Ghosh S, Caliskan A. Chatgpt perpetuates gender bias in machine translation and ignores non-gendered pronouns: Findings across bengali and five other low-resource languages[C]//Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society. 2023: 901-912.

Biography

Linlin Zhang is with the School of Foreign Languages, Shenyang Normal University. Research direction is English education and application and AI.