Hindawi
Mathematical Problems in Engineering
Volume 2022, Article ID 6798505, 10 pages
https://doi.org/10.1155/2022/6798505

Research Article

Deep Learning Classification Model for English Translation Styles Introducing Attention Mechanism

Tian Zhang

College of Foreign Languages, North China University of Water Resources and Electric Power, Zhengzhou, Henan, China

Correspondence should be addressed to Tian Zhang; zhangtian@ncwu.edu.cn

Received 15 January 2022; Revised 14 February 2022; Accepted 26 February 2022; Published 26 March 2022

Academic Editor: Gengxin Sun

Both short-distance association knowledge and long-distance interaction knowledge in the knowledge base contain rich semantics. When learning entity and relation representation in a knowledge base, if we can learn short-distance association knowledge and long-distance interaction knowledge at the same time, we can learn the representation method with rich semantic information and keep the original structure of the knowledge base. Among the knowledge contained in a large number of records, some knowledge reflects individual characteristics and can be called local knowledge; others reflect group characteristics and can be called global knowledge. Using different ways to learn local and global knowledge in the deep learning model will better reflect the difference between the two kinds of knowledge at the model level, and make the model have the ability to understand both individual characteristics and overall characteristics. Through layer-by-layer forward propagation and error back propagation algorithms, the entire network is gradually optimized in an "end-to-end" manner. This "end-to-end" approach leaves some means of introducing prior knowledge flexibly into the model. Although it can reduce the burden on researchers, this "data-driven" approach brings the shortcomings of poor interpretability of learning results and weak generalization ability. Combining the specific prior knowledge implicit in the data with the deep learning algorithm can optimize the algorithm in a targeted manner and avoid blind searching in the solution space, so as to obtain a model with better performance and wider use. To this end, this paper investigates combining prior knowledge with deep learning to design efficient algorithms to address the classification of English translation styles. This paper combines local knowledge with global knowledge and deep learning methods and proposes a memory neural network method combining local knowledge and global knowledge. By recording the local knowledge in the local memory module and simultaneously recording the global knowledge in the global memory module, the method effectively learns the latent information in a large number of records. This paper combines short-distance association knowledge with long-distance interaction knowledge and a distributed representation learning method based on deep learning and proposes a deep learning method combining short-distance association knowledge and long-distance interaction knowledge. On the IWSLT English translation task, experiments show that the method significantly improves translation quality, confirming that grammatical dependencies enhance attention by supplementing dependent grammatical information, resulting in more effective and richer context vectors that more accurately represent contextual situations. Additional experimental analysis showed that the model underwent careful parameter selection and analysis. By mining valuable long-distance interactive knowledge in the knowledge base and using it in the distributed representation learning of the knowledge base, while constraining the short-distance related knowledge and constraining the long-distance interactive knowledge, the learned knowledge can be used to effectively complete the knowledge base distributed representation for discovering new relations.

1. Introduction

Rule-based machine translation is achieved by studying the linguistic information of the source and target languages, mainly based on dictionaries and grammars, etc. to generate translations, and the common difficulty is that it cannot give accurate and sufficient linguistic information to meet the translation needs under different domains. Instance-based machine translation is essentially a translation instance-based machine translation based on the principle of similarity, which searches for matching instances in a corpus consisting of many gold-aligned bilingual sentence pairs and determines the translation with the highest similarity by comparing multiple matching instances [1]. The translation effectiveness of this method is overly dependent on the quality of the bilingual corpus and fails to translate successfully if a suitable matching instance is not searched for. The idea of statistical machine translation is to design a statistical model for generating a language and then use that statistical model for translation. To do this requires statistical analysis of a large parallel corpus using statistical methods, constructing a good statistical translation model, defining the model parameters to be evaluated on the model, and designing algorithms for how the model parameters can be optimized. Statistical machine translation has gone through three stages of development based on words, based on phrases, and based on syntactic information [2].

Computer hardware has been iteratively updated with better and better performance, and both computing power and computing consumption costs have been improved to a great extent. Artificial intelligence has developed from the initial sprout to the present, and the technology at both the academic research level and the practical application level has become more mature, and today, human life has long been closely related to artificial intelligence technology [3]. Machine learning (ML) is a popular research area in AI and a key development discipline in cutting-edge computer technology. Deep learning has also been applied to the study of natural language processing and computer vision and has reached a high level, bringing revolutionary advances in artificial intelligence technology [4]. Under comparison with traditional machine learning, the core idea of deep learning is to imitate the ways of neuronal transmission and learning of the human brain and to read and analyze massive data through computation for automatic feature extraction [5]. In this paper, we investigate the implicit learning capability of the neural machine translation model to stimulate its potential and enhance its performance accordingly in several translation subdomains [6]. Specifically, we first exploit its implicitly acquired multiheaded attention mechanism to achieve improvement on the diversity translation task, then improve the quality of low-resource translation by masking the attention heads within the model, and finally exploit its own potential long-range text modeling capability to achieve a breakthrough on the document translation task [7].

Statistical machine translation views the process of sentence generation as the derivation of phrases or rules, which is essentially a symbolic system in a discrete space [8]. Deep learning turns the traditional discrete-based representation into a continuous-space representation. For example, instead of a discrete representation of words, a distributed representation in real space is used, and the entire sentence can be described as a vector of real numbers [9]. This allows the translation problem to be described in a continuous space, which in turn greatly alleviates problems such as the dimensional catastrophe of the traditional discrete space model. More importantly, the continuous space model can be optimized using methods such as gradient descent, which have good mathematical properties and are easy to implement. However, the effect of machine translation is still far from that of human translation, and neural machine translation has a long way to go. This makes the translation and translation more reductive and fluent. For document translation tasks, extend the end-to-end training method of neural machine translation models, explore its own potential for long-distance text modeling, and build a new document translation paradigm. The current research on neural machine translation still faces many challenges that need to be improved. Therefore, the research on neural machine translation has high academic significance.

2. Related Work

Residual networks can solve the problems of gradient disappearance and gradient explosion in multilayer networks. Therefore, residual connectivity has become one of the hot research topics in deep networks. A residual network is a network composed of the input and output of a residual connection layer. Both the residual network and the method in this paper increase the information flow path, which is beneficial to the gradient transfer. The difference is that the residual network adds the input of the current layer to the output, while the method in this paper lies in fusing the information of the intermediate layers to the encoder or decoder to supplement the final output, making full use of the intermediate information to improve the modeling ability of the model [10]. The ResNet hierarchical network structure is used to achieve better performance without increasing the computational cost. The pseudoparallel corpus thus constructed is more diverse, so that the pseudoparallel corpus covers a more realistic data distribution, thereby improving the translation performance in resource-rich scenarios. In statistical machine translation, monolingual data is often used to train language models to portray the fluency of translation candidates. In neural machine translation models, both fluency and fidelity are unified in the decoder, and there is no separate language model, and thus no training using monolingual data. However, the monolingual corpus is more readily available compared to the bilingual corpus, and its corpus size is larger than the bilingual corpus. How to utilize the large-scale monolingual corpus to improve the quality of neural machine translation has become a very important research direction [11].

For the first time, shallow and deep fusion methods are proposed to integrate external recurrent neural network-based language models into the encoding-decoding framework. In this approach, the shallow fusion approach linearly combines translation probabilities and language model probabilities, while the deep fusion approach connects recurrent neural network-based language models with decoders to form a new tightly coupled network [12]. Although high-quality and domain-specific translations are crucial in the real world, domain-specific corpora are often scarce or non-existent, leading to poor performance of neural machine translation models in such cases. Therefore, how to exploit parallel corpora outside the domain is important for building domain-specific neural machine translation systems [13]. The first common approach is to train the model with data from the source domain and then fine-tune it with data from the target domain. It has been explored how migration learning of low-resource language pairs can be improved by fine-tuning only a part of the neural network [14]. Domain adaptive methods are evaluated against various experiences, and methods for mixing source and target domain data during fine-tuning are proposed. The effect of using only a small portion of the target domain data during the fine-tuning phase is explored. Another possible approach is based on a multidomain fusion scenario involving the addition of domain indicator markers to each source language sentence [15].

Statistical machine translation also has some areas for improvement. Firstly language models and translation models are more efficient and economical to use, but specific errors are difficult to predict and correct; since translation systems cannot store all native strings and their translations, they are usually translated sentence by sentence for longer passages or documents; language models are usually smoothed, and similar methods are applied to translation models, but again due to differences in sentence length and word order in the language increases the complexity. So statistical machine translation can neither translate in conjunction with contextual semantic scenarios nor handle linguistic rules such as word order, lexicography, and syntax well; also, since statistical machine translation works according to the rule of counting the frequency of phrases in a parallel corpus and selecting the best matching words, the similarity between words is not well represented in statistical machine translation.

3. Deep Learning Combined with Prior Knowledge English Translation Style Classification Analysis

3.1. Optimizing Deep Learning Algorithm Design. Since neural network model training is a "data-driven" approach, the quality of the parallel corpus data largely affects the performance of the network model. At present, the publicly available bilingual data sets are rich in content and large compared with other language pairs, so it is difficult to avoid duplication of utterances and other situations affecting the training quality in the process of text data integration. Therefore, to ensure the quality of the parallel corpus, the data needs to be cleaned before using the parallel data to train the model, and then other preprocessing operations are done. Before performing other processing operations on the original text, the parallel corpus should be cleaned to ensure that the subsequent operations can be performed correctly. The cleaning operations for English text are divided into abbreviation change, separator normalization, etc.; the cleaning operations for English are full and half-corner symbol normalization, space removal, etc.

The translation and alignment are proposed to be learned jointly together in an encoding-decoding framework [16]. We determine the translation with the highest similarity by comparing multiple matching instances. The translation effect of this method is overly dependent on the quality of the bilingual corpus, and if no suitable matching instance is found, the translation cannot be successfully carried out. The machine translation system is based on an attention mechanism. Firstly the input module is responsible for reading the information of the source language words and representing them in a distributed manner, with a feature vector associated with each word position; then the system performs retrieval based on the list of feature vectors; and finally, the tasks are executed according to the content sequence, each time focusing on one or several content depending on the weights. The alignment matrix of the source and target sequences, which shows the importance distribution of each word in the source sequence for the current word to be translated when translating a word, is shown in Figure 1.

As shown in Figure 1, the input is a source sequence X of length n, and tries to output a target sequence of length m_{ty} using a bidirectional RNN as the encoder model with forward hidden states and backward hidden states. The contextual information of the words, computed using equation (1), and the forward and backward representations are spliced as the hidden layer states of the encoder using a simple concatenation, $a^2_{i,j}$ is the alignment matrix of the source and target sequences, and the fraction of the alignment is obtained using the softmax function, as shown in equation (2). The attention mechanism performs a weighted average of the feature vectors h_i^2 with weights to a_t from the context vector c. Artificial intelligence technology has brought revolutionary advances. Compared with traditional machine learning, the core idea of deep learning is to read and analyze massive amounts of data by computing, imitating the way of neuron transmitter transmission and the learning method of the human brain, to achieve the purpose of automatic feature extraction.

$$c_t = \sum_{i=1}^{n} a^3_{i,j} h_i^2, \tag{1}$$

$$a_t = \frac{\exp\left(\mathrm{Score}\left(s_{t-1}, h_t\right)\right)}{\sum_{i=1}^{n} \exp\left(\mathrm{Score}\left(s_{t-1}, h_t\right)\right)}. \tag{2}$$

The multiheaded attention layer does the work of transforming the three basic parameters of the input model, key, value, and query, into a form of data suitable for processing by the scalar multiplicative attention function. At the same time, a cut transformation of the matrix is completed, which uses three linear transformations that are not identical. The small matrix after the cut is processed by the scalar multiplicative attention function and then stitched into the size of the original matrix by the join layer structure. Since random initialization may destabilize the learning process, this problem can be overcome by performing multihead attention in parallel, which allows the results to be connected, where each head (Head) has individually learnable weights. The final output of the entire multiheaded attention layer is obtained from the stitched matrix followed by a linear transformation.

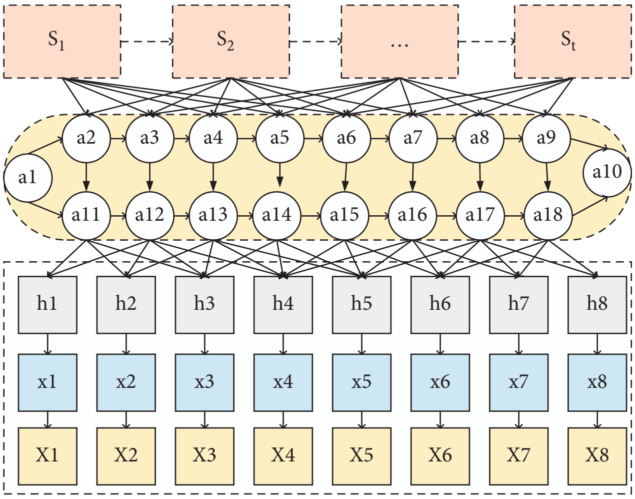

Figure 1: Improved deep learning algorithm architecture.

$$\begin{gathered}\text{Attention}\,(\mathrm{Q},\mathrm{K},\mathrm{V}) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)\cdot V^2,\\ \text{head}_{\mathrm{i}} = \text{Attention}\left(QW_i^Q, KW_i^K, VW_i^V\right).\end{gathered} \tag{3}$$

Choosing a suitable model among the many word vector models that replace the random initial values in end-to-end neural network-based machine translation systems as inputs to the input layer of the neural network is crucial to enable the neural network translation model to converge to a better optimal solution. In addition to the model itself, the accuracy of the word vector is very dependent on the choice of the corpus, and the corpus of different sizes and language families can greatly affect the performance of the word vector, and the setting of the model parameters, and the number of iterations when training the word vector model can affect the effectiveness of the word vector model. Therefore, how to select a word vector model in a machine translation task is the first problem that should be solved. The existing word vector models have their advantages and disadvantages, but none of them consider the influence of the location information of the source language words on the text feature representation.

As introduced in the model structure, the co-occurrence matrix is denoted by X and each element of the matrix that is specifically different is represented by i, X_i, j, and X_j means during the operation, the number of times that word i and word j appear together in a window in the whole corpus, and the size of the window can be set.

$$F\left(w_i, w_j, w_k\right) = \frac{P_{i,k}^2}{P_{j,k}^2}, \tag{4}$$

$X_i = \sum_k x_{i,k}$ denotes the number of occurrences of all words in the context of the word i; $P_{i,j} = X_{i,j}/x_{i,k}$ denotes the probability of occurrence of the word j in the context of the word i. The word vector function F is used to express its co-occurrence probability ratio, w is a D-dimensional word vector, and the probability ratio on the right-hand side of the equation is obtained from the content of the corpus and calculated by equation (5).

Due to the large corpus, the probability ratio on the right-hand side of the equation is difficult to obtain, so deformation of the formula, F is to encode into the $P_{i,k}^2/P_{j,k}^2$ vector space, that is, to represent $P_{i,k}^2$ and $P_{j,k}^2$ distance in the vector space, usually using the difference between the two, as shown in equation (6).

$$F\left(w_i, w_j, w_k\right) = \frac{P_{i,k}^2}{P_{j,k}^2}. \tag{5}$$

The left side of the equation is a vector while the right side of the equation is a scalar, thus continuing to transform the left side of the equation [17]. In this paper, we choose to build on the existing word vector model to further investigate and validate the performance of the existing word vector model through a machine translation task. Since the CBOW model can obtain better grammatical information by predicting the target word through context learning, it has higher accuracy in grammar tests and uses a distributed representation of word vectors for each word in both the input and output layers, merges the projections of the input layer into the projection layer, and uses the mean value to represent individual word vectors with low model complexity.

It is foreseeable that if the masking covers more attention heads, it will make the translation quality gradually decrease. Furthermore, all the heads are gradually blocked according to the order of importance, and the curves are plotted in the order of large to small and small to large, respectively, and the results in Figure 2 can be obtained. It can be concluded that the curve of blocking unimportant attention heads at the beginning decreases slowly, while the curve of blocking from important attention heads decreases very fast.

Another idea is to shield the most important head from the training process, using the analogy of a football player who trains the left foot more if the dominant foot is the right foot, or trains the right foot more if the left foot becomes the dominant foot [14]. The core idea of the strategy is to inhibit the use of the most important head, allowing the rest of the heads to be trained more. Specifically, the implementation starts with a feedforward calculation and backpropagation by the original network, which uses equation (6) to calculate the importance of all heads, but does not perform parameter updates. After filtering out the most important parts of the current attention heads by sorting, they are masked out and the remaining network is used to reperform the loss function computation and parameter update.

$$I_h = E_{x\sim X}\left|\frac{\partial \hbar\,(x)}{\partial \xi_h}\right|, \tag{6}$$

where X is the data distribution and $\hbar\,(x)$ is the loss function over the sample x. Substituting equation (6) into equation (7) yields the final expression for the I_h.

$$I_h = E_{x\sim X}\left|Att\,(x)\frac{\partial \hbar\,(x)}{\partial \xi_h}\right|. \tag{7}$$

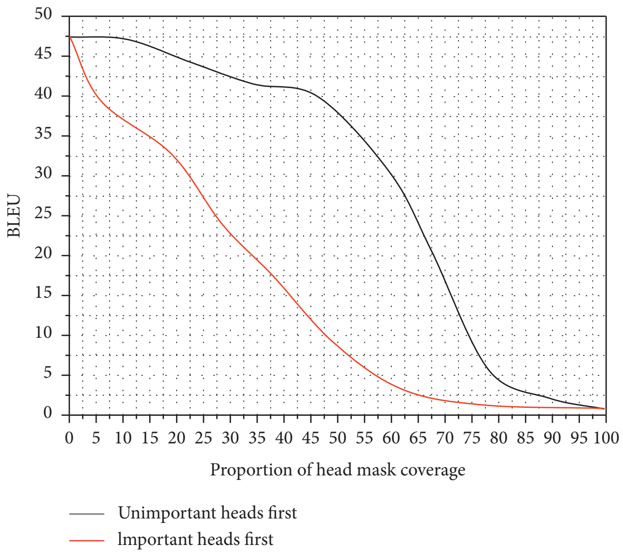

Figure 2: Gradual blocking of all attention heads in order of importance (ascending or descending).

The proposed method further enhances the corpus based on the reverse translation by using word substitution methods to enhance the training data containing low-frequency words through language model substitution for low-frequency words, which constructs a more diverse pseudoparallel corpus and makes the pseudoparallel corpus covers a more realistic data distribution, thus improving the translation performance in resource-rich scenarios. In low-resource scenarios, the word substitution method additionally adds a linguistic error correction module to eliminate errors such as syntactic semantics generated by substitution.

To assess the goodness of translation results, different translation evaluation metrics exist, which can be generally classified into two categories: manual evaluation and automatic evaluation. The manual assessment method can compare the translation result and the reference translation well to determine the degree of restoration of text content and sentence meaning and thus measure the fidelity and fluency of the translation result. However, the cost of the manual evaluation is very high, requiring not only the quality of professionals, but also a long time to evaluate, and the subjective differences of different translators on the same translation result will lead to different evaluated results. More importantly, continuous space models can be optimized with methods such as gradient descent, which have good mathematical properties and are easy to implement. Due to the abovementioned limitations, manual evaluation is not suitable as a common evaluation index for neural machine translation.

3.2. Experimental Design of a Priori Knowledge English Translation Style Classification. Furthermore, in our experiments, the standard deviation is set to $h/2$, and h is empirically set to the depth of the dependency tree, which is the hierarchical order of the grammar tree. We can obtain the grammar branch distances between arbitrary words from the hierarchical structure of the dependency tree. To eliminate unwanted interfering words to some extent without losing a moderate focus on words on different branches, we chose to set the maximum grammar branch distance to the depth of the dependency tree.

$$e_{TS}^{sbd} = e_{ts} \exp\left(\frac{\left(B\left[p_t\right]\left[s\right]^2\right)}{s\sigma^2}\right). \tag{8}$$

For Chinese-English language pairs, we selected specific training data from LDC Corpora. The entire training corpus consists of 2.6 million sentence pairs containing 65.1 million Chinese characters and 67.1 million English words, respectively. In addition, 8 million Chinese sentences and 8 million English sentences were also randomly selected from the Xinhua section of Gigaword Corpus as the monolingual dataset. During the model training, any sentence with more than 60 words in the training data, bilingual data, and pseudobilingual data will be removed. For translations in this direction from Chinese to English, the NIST2006 dataset is used as the validation set, and the NIST2003, NIST2005, NIST2008, and NIST2012 datasets are used as the test sets. In all validation and test sets, there are four English reference translations for each Chinese sentence [18]. And for translations in this direction from English to Chinese, we use the reverse NIST dataset: the first English sentence of the four English reference translations is treated as the source language sentence, and the Chinese sentence is treated as a single reference translation. For the dictionary of the translation model, the entire lexical dictionary was restricted to contain only the most frequent 50,000 words occurring at the source and target language ends, and the other words were converted to <UNK> symbols.

During the whole training process, if the objective function is close to the ideal semisupervised objective function, the potential gain to the translation model is smaller (Figure 3). Also, due to the many uncertainties in the training process, sometimes the translation performance of the model will occasionally drop a bit. Before performing other processing operations on the original text, the parallel corpus should be cleaned to ensure that the subsequent operations can be performed correctly. As mentioned in the previous analysis of the semisupervised objective function, as the translation model from the target language to the source language gets closer to the ideal translation probability, the objective function used in the joint training method will more closely approximate the ideal semisupervised objective function. Throughout the training process, the closer the objective function approximates the ideal semisupervised objective function, the smaller the potential gain to the translation model will be. In addition, since there is a lot of uncertainty in the training process, sometimes the translation performance of the model will occasionally degrade a bit.

This same problem severely affects most neural machine translation models. Neural machine translation models are often trained using only normal bilingual sentences, and the models self-recursively produce translated utterances word

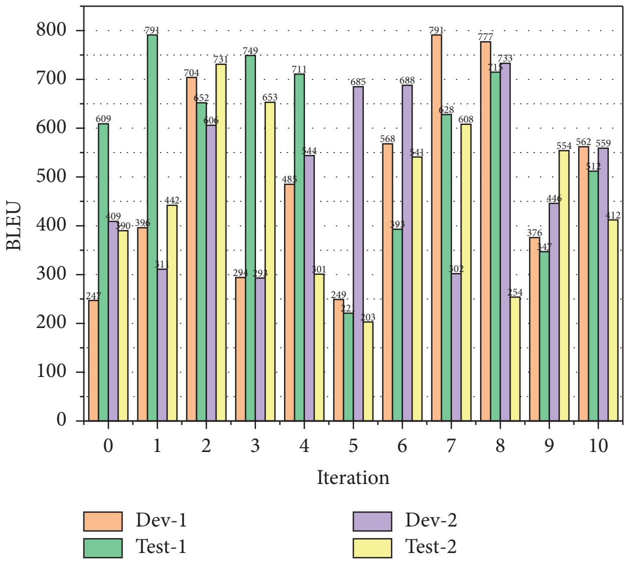

Figure 3: Translation performance variation on the validation and test sets of the translations.

by word during decoding. In this way, translation errors that occur early on can mislead the model for subsequent translations, resulting in the generation of translated sentences that often have the correct prefix but the wrong suffix. And this problem becomes more severe when the length of the sentences to be translated increases. Therefore, even if the translation model is trained on a sufficiently rich parallel corpus, the model often fails to take full advantage of the parallel corpus and produce high-quality translation candidates.

In practice, irrespective of whether left-to-right or right-to-left orientation is used for decoding, translation models based on self-recursive structures suffer from exposure bias problems, which produce undesirable translated sentences. Models based on left-to-right decoding produce translation candidates with good prefixes but incorrect suffixes, while models based on right-to-left decoding produce translation candidates with good suffixes but incorrect prefixes [19]. However, we find that the model can avoid the translation errors generated by the left-to-right decoding approach if a right-to-left decoding translation order is used. However, we found that if the translation order of right-to-left decoding is used, the model can avoid the translation errors generated by the left-to-right decoding method. At the same time, the problems caused by the right-to-left decoding method can also be avoided by the model using the left-to-right decoding method. Also, the problems posed by the right-to-left decoding approach can be exactly avoided by the model using the left-to-right decoding approach. This strongly suggests that there is a degree of complementarity between the translations of these two-way decoding approaches, and if more efficient methods can be devised to take full advantage of this property, the exposure bias problem can hopefully be effectively mitigated, as shown in Table 1.

The original study only described that the dependency syntactic closeness of two words in a dependency tree diminishes as their syntactic distance becomes larger, without considering the dependency closeness between different parent-child node pairs, and between different child node pairs to the same parent node, and certainly without considering the case of grandparent node pairs. Often, the default is that the degree of dependency intimacy between different parent-child or grandchild node pairs is equal.

With the same development environment and parameter settings, the mainstream statistical machine translation model, Maverick statistical machine translation model, is selected as the baseline system to compare with the neural network model as a comparison; the recurrent neural network model, long and short-term memory network model, and bidirectional recurrent neural network model are constructed for training according to the above parameters, and the attention mechanism algorithm is added to them, together with the Transformer model for comparison. This experiment illustrates the effectiveness of machine translation models based on deep learning neural networks, and the performance of several networks in machine translation tasks is derived based on the comparison experiments. The results obtained from the validation set and the two test sets show that firstly, several neural network-based translation models have higher BLEU scores than the baseline system, with Transformer having the highest BLEU-4 score; among the models that do not incorporate the attention mechanism, the long short-term memory network translation model has the best performance, followed by the bidirectional recurrent neural network model, and finally, the unidirectional recurrent neural network model. From the control group of the four networks incorporating the attention mechanism, the performance of each translation model after incorporating the attention mechanism showed a considerable improvement, and the inclusion of the attention mechanism significantly reduced the gap with the baseline.

4. Analysis of Results

4.1. Improved Deep Learning Algorithm Results. The annotation of the text sequences is done using the toolkit NLTK from the Python standard library and collated into lexical sequences that correspond to the source language corpus. In the model training phase, joint training is performed on both tasks using alternating training. For the lexical annotation task, the lexical categories are constructed separately as category tagging tables for training, and the word lists are generated separately for both languages by the BPE approach. Considering that machine translation is the main task and the number of lexical labels is small relative to the target language vocabulary, the lexical annotation task is trained first and then the machine translation task is trained. All parameters are initialized using a Gaussian random distribution and decoding is performed using bundle search. For effective comparison, the Transformer base model was chosen as the baseline system. The translation quality evaluation metric is the BLEU-4 score.

To effectively compare the models in this experiment and to explore the advantages and disadvantages of the global sharing approach, we simultaneously trained a multitask model with shared local parameters. This model is the same

Table 1: Training of translation models decoded from left-to-right.

Input	$D = \{(x^n, y^n)\}_{n=1}^{N}$
Procedure	$P(y\|x;\theta)$
While	$P(y\|x_1;\theta)$
Do	Training process
If	Sample translated sentence pairs from bilingual corpus D
Else	Then use them to construct pseudosentence pairs
Do	Use them to construct pseudosentence pairs with weights
Output	End procedure

multitask learning model used to solve the target language lexical annotation problem and the machine translation problem. Unlike the global weight-sharing approach in this paper, as described in the previous section, the model only shares the encoder network layer and builds two decoders separately, and all network layer weights in the encoder are trained independently of each other, i.e., the local sharing approach. The base Transformer model is selected as the baseline system for this experiment, as shown in Figure 4.

Analysis of the experimental results based on the content of Figure 4 shows that the model in this experiment performs better than the baseline system on both test sets, with the global weight-sharing model improving the BLEU-4 score by 1.98 points on average, while the global sharing model improves the BLEU-4 score by 1.19 points on average compared to the local sharing model that only shares the encoder layer network weights. Thus, it can be concluded that sharing weights with the joint training task in the translation task is effective, while sharing weights on the decoder side as well can further improve the performance of the machine translation task.

Each time, focuses on one or more contents according to different weights. The alignment matrix of the source sequence and the target sequence shows the importance distribution of each word in the source sequence to the current word to be translated when translating a word.

Multitask learning corresponds to increasing the number of samples used to train the model. Since all samples are not ideally distributed, there is some noise and the risk of overfitting exists when training separate models. This noise can be balanced by multitask learning under different samples, and sharing parameters weakens the network capability to some extent, reducing the possibility of task overfitting and improving the generalization ability of the model. In addition, multitask learning can further deepen the model's focus on the common features of multiple related tasks. And the global weight-sharing can enable the model to be applied to the main task by focusing on more implicit information and mining the hidden patterns under different task levels.

In the English-French dataset machine translation task, the PW-CBOW word vector model achieves the best result with a BLEU value of 43.97 in the end-to-end neural network machine translation system; none of the translation performance based on several other word vector models exceeds that of the randomly generated word vector model of the baseline system. Figure 5 shows the performance gain rates for each word vector model in the WMT 14 English-French machine translation task, where the machine translation performance using the PW-CBOW model has reached the best results for the same conditions.

To further verify the performance of the PW-CBOW word vector model in this chapter, we divided three different sizes of data from the English-French dataset, 3 million pairs, 10 million pairs, and 33 million pairs. To verify the effect of the word vector model on the performance of the end-to-end neural network machine translation system, according to the experimental results in Figure 5, we can see that the end-to-end PW-CBOW word vector model-based neural network machine translation system achieves a BLEU value of 43.97 for the WMT 14 English-French dataset (3300w) machine translation task, a BLEU value of 34.35 for the WMT 14 English-French dataset (1000w) machine translation task, and a BLEU value of 21.35 for the WMT 14 English-French dataset (300w) machine translation task. These noises can be balanced through multitask learning under different samples, and the shared parameters weaken the network ability to a certain extent, reduce the possibility of task overfitting, and improve the generalization ability of the model. In addition, multitask learning can further deepen the model's attention to common features of multiple related tasks. In the WMT 14 English-French dataset (300w) machine translation task, the BLEU value reaches 21.12, which is the best result.

It just adds extra document information. This approach requires not only that the document's data be parallel, but also that the sentences within it be parallel. Any data with an unequal number of sentences or slight shifts in order cannot be used for training. However, this is rather demanding for many scenarios, such as bilingual fiction data, which often only satisfies chapter-level parallelism and does not guarantee sentence-level parallelism. This constraint limits the size of data that can be used for training and thus limits the space available for model enhancement.

4.2. Experimental Results. To investigate the relationship between translation quality and length, the test documents were randomly sliced into sequences of different lengths and evaluated using the trained models. The results are shown in Figure 6, where the model trained on the PDC-Sent corpus only suffers from severe quality degradation when translating long texts, while the model trained on PDC-Doc only suffers from poor quality when translating short sentences. The model trained on the mixed corpus PDC-Mix achieves good translation quality in all scenarios. This results in an all-in-one model that has excellent translation performance regardless of the length of the sequence. The "document-to-document" translation model breaks the length limitation of translation sequences.

We choose global-attention as the experimental baseline system, and the selection of parameter settings is made by the average BLEU scores of the validation and test sets in the English-German translation task. Figure 7 shows the translation results of the trained models of the baseline

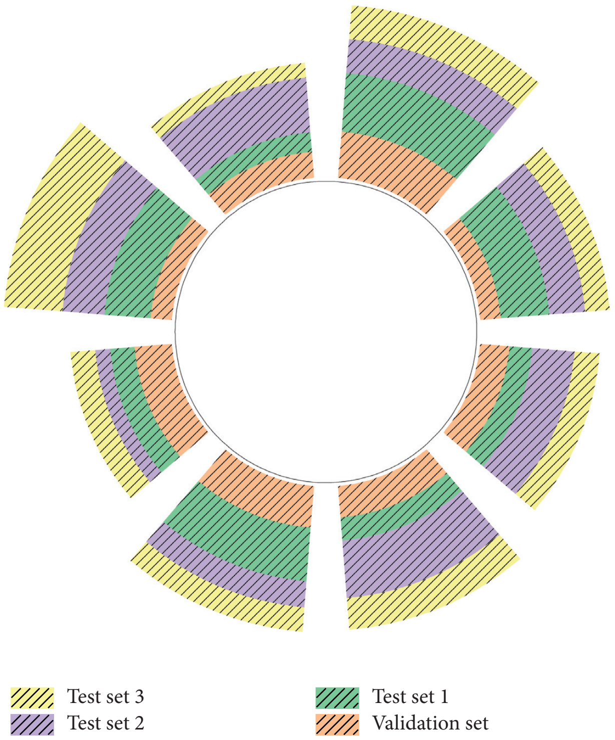

Figure 4: Translation model BLEU-4 scores.

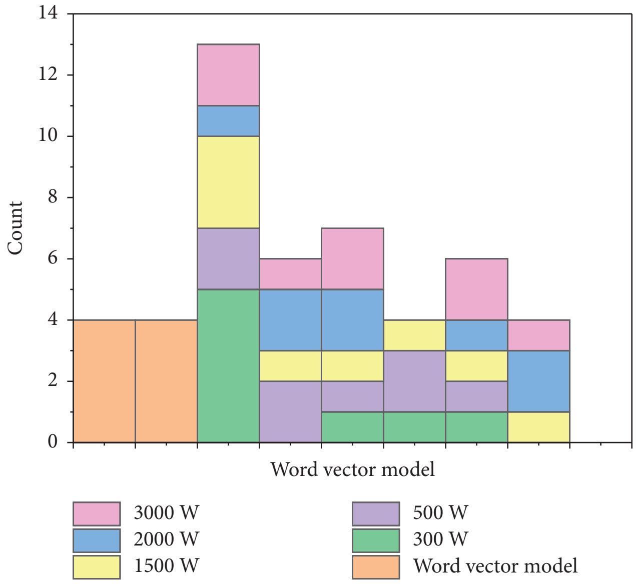

Figure 5: Performance of English-French machine translation for different corpus sizes.

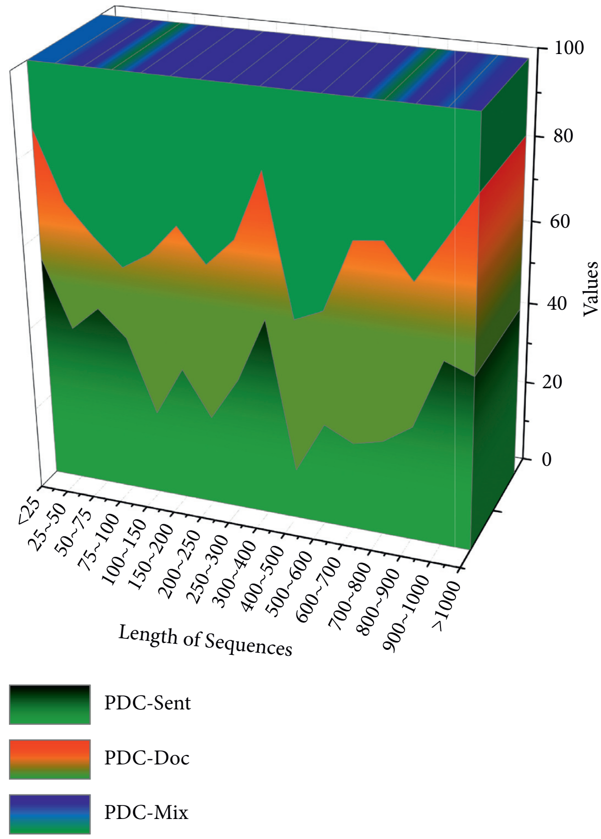

Figure 6: Relationship between BLEU and sequence length in document-to-document translation.

system with different batch sizes, expecting to find the optimal setting of the batch size in this way. According to the experimental comparison results given, it is easy to find that the translation effect of the model using a batch size of 40 fails to reach the best, but the difference with the best BLEU score of batch size 64 is only 0.21, and as the batch size increases from 40 to 64 and 80, the BLEU score is a general trend that increases first and then decreases. This also verifies that within a certain range, the larger the batch size is the more accurate the direction of decline in training loss, while increasing to a certain degree, the direction of decline is not changing. Here, with our experimental setup with a batch size of 40, we completed other comparative experiments with different model systems, and the experimental effect difference between batch sizes of 40 and 64 is not obvious. To reduce the number of repeated experiments and save time spent on the experiments, we chose a batch size of 40 on this dataset to achieve a better convergence effect.

Also, we explored the effect of beam search on decoding effectiveness for different beamwidths, and the results are shown in Figure 7. It is not difficult to find that the optimal beamwidth for this dataset is 5. This is different from the common conclusion that increasing the beamwidth provides better translations because there is a saturation point in the process of increasing the beamwidth beyond which increasing the beamwidth no longer works to improve translations. Second, candidates with high scores may tend to have similar compositions, so even a search range of width 5 is sufficient, and continuing to increase it simply introduces unnecessary candidate words.

The former, by dynamically tuning to subtly achieve a reasonable distinction between the dependency intimacy of different word pairs. Afterward, grammar-aware attention is guided to generate source-side dependency contexts based on the two novel types of distances to improve the

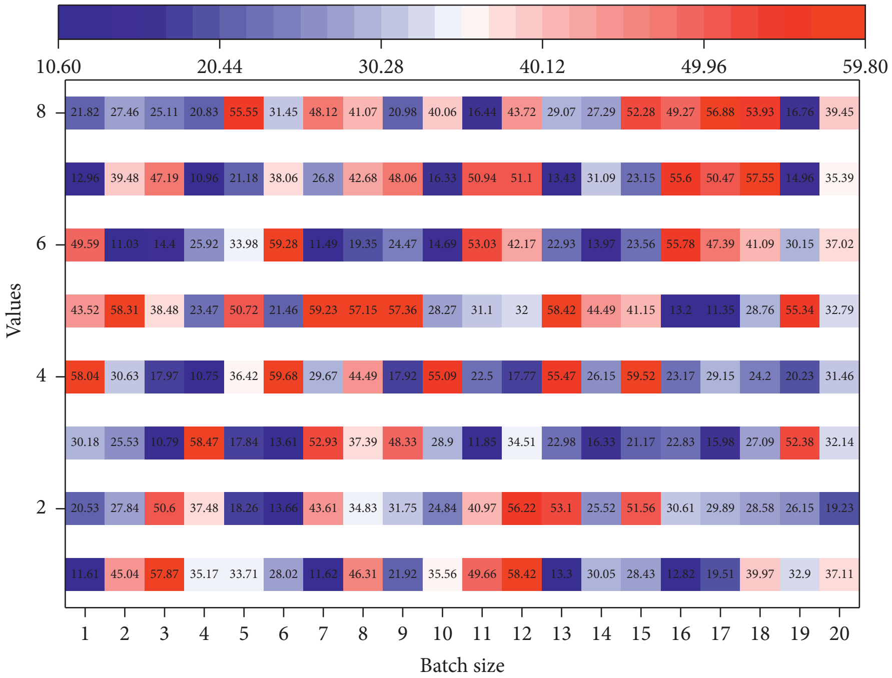

Figure 7: Translation effect of the baseline system model with different beamwidths.

performance of predicting target words. Experiments were conducted on an English-to-German translation task, and the experimental results showed that grammatical dependencies serve as additional information to obtain more linguistic information from the source, helping to enhance attention to compute a more efficient, context-rich vector for predicting the target word, significantly improving translation quality.

5. Conclusion

To address the problem of inadequate utilization of information in existing NMT systems with multilayer networks, an optimization method using multilayer information fusion networks is proposed. Since random initialization can destabilize the learning process, this problem can be overcome by performing multihead attention in parallel by concatenating the results, where each head has a separate learnable weight. But the difference between the best BLEU score with a batch size of 64 is only 0.21, and as the batch size increases from 40 to 64, 80, the BLEU score shows an overall trend of first increasing and then decreasing. This also verifies that within a certain range, the larger the batch size, the more accurate the decreasing direction of the training loss is, and when it increases to a certain extent, the decreasing direction does not change. And three fusion methods are proposed, which are arithmetic mean fusion, linear transformation fusion, and gate mechanism fusion. The experimental results show that the multilayer information fusion method helps to improve the quality of machine translation, of which the arithmetic average fusion method has the least impact on the model training speed. In this way, statistical machine translation can construct more reliable phrase translation tables and simultaneously eliminate the uncommon mistranslation patterns generated by neural machine translation models during unsupervised training. In addition, an expectation-maximization algorithm is used to unify the updating of statistical machine translation and neural machine translation models, where all models can be trained jointly and benefit progressively. With this framework, the negative impact caused by errors during unsupervised training can be mitigated by statistical machine translation, while neural machine translation can compensate for the inherent lack of translation fluency in statistical machine translation. The results of large-scale machine translation experiments verify that this method can effectively mitigate the problem of pseudodata noise, thus achieving state-of-the-art unsupervised neural machine translation performance. Furthermore, this unsupervised training method is extended to the language style migration task, and corresponding experiments further confirm the usefulness of the method.

Data Availability

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

Conflicts of Interest

The authors declare that there are no conflicts of interest with this study.

References

[1] M. Yang, S. Liu, K. Chen, H. Zhang, E. Zhao, and T. Zhao, “A hierarchical clustering approach to fuzzy semantic representation of rare words in neural machine translation,” *IEEE*

Transactions on Fuzzy Systems, vol. 28, no. 5, pp. 992–1002, 2020.

[2] N. Briggs, "Neural machine translation tools in the language learning classroom: students' use, perceptions, and analyses," *Jalt call journal*, vol. 14, no. 1, pp. 2–24, 2018.

[3] E. Korot, Z. Guan, D. Ferraz et al., "Code-free deep learning for multi-modality medical image classification," *Nature Machine Intelligence*, vol. 3, no. 4, pp. 288–298, 2021.

[4] F. Teng, Z. Ma, J. Chen, M. Xiao, and L. Huang, "Automatic medical code assignment via deep learning approach for intelligent healthcare," *IEEE journal of biomedical and health informatics*, vol. 24, no. 9, pp. 2506–2515, 2020.

[5] Y. Muravev, "Teaching legal English translation by the case method in Russian-English language pair," *Humanities & Social Sciences Reviews*, vol. 8, no. 4, pp. 961–971, 2020.

[6] S. Zulfiqar, M. F. Wahab, M. I. Sarwar, and I. Lieberwirth, "Is machine translation a reliable tool for reading German scientific databases and research articles?" *Journal of Chemical Information and Modeling*, vol. 58, no. 11, pp. 2214–2223, 2018.

[7] A. H. S. Hamdany, R. R. Omar-Nima, and L. H. Albak, "Translating cuneiform symbols using artificial neural network," *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, vol. 19, no. 2, pp. 438–443, 2021.

[8] L. Yu, L. Sartran, W. Stokowiec et al., "Better document-level machine translation with Bayes' rule," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 346–360, 2020.

[9] R. Satapathy, E. Cambria, A. Nanetti, and A. Hussain, "A review of shorthand systems: from brachygraphy to microtext and beyond," *Cognitive Computation*, vol. 12, no. 4, pp. 778–792, 2020.

[10] Y. Jia, M. Carl, and X. Wang, "How does the post-editing of neural machine translation compare with from-scratch translation? A product and process study," *The Journal of Specialised Translation*, vol. 31, pp. 60–86, 2019.

[11] J. Bobadilla, Á. González-Prieto, F. Ortega, and R. Lara-Cabrera, "Deep learning feature selection to unhide demographic recommender systems factors," *Neural Computing & Applications*, vol. 33, no. 12, pp. 7291–7308, 2021.

[12] S. Altaf, S. Iqbal, and M. W. Soomro, "Efficient natural language classification algorithm for detecting duplicate unsupervised features," *Informatics and Automation*, vol. 20, no. 3, pp. 623–653, 2021.

[13] J. G. Makin, D. A. Moses, and E. F. Chang, "Machine translation of cortical activity to text with an encoder-decoder framework," *Nature Neuroscience*, vol. 23, no. 4, pp. 575–582, 2020.

[14] B. Zhang, D. Xiong, J. Su, and J. Luo, "Future-aware knowledge distillation for neural machine translation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 12, pp. 2278–2287, 2019.

[15] J. Kang, R. Fernandez-Beltran, P. Duan, S. Liu, and A. J. Plaza, "Deep unsupervised embedding for remotely sensed images based on spatially augmented momentum contrast," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 3, pp. 2598–2610, 2020.

[16] L. A. Bolaños, D. Xiao, N. L. Ford et al., "A three-dimensional virtual mouse generates synthetic training data for behavioral analysis," *Nature Methods*, vol. 18, no. 4, pp. 378–381, 2021.

[17] U. Sulubacak, O. Caglayan, S. A. Grönroos et al., "Multimodal machine translation through visuals and speech," *Machine Translation*, vol. 34, no. 2, pp. 97–147, 2020.

[18] S. M. Sarsam, H. Al-Samarraie, A. I. Alzahrani, and B. Wright, "Sarcasm detection using machine learning algorithms in Twitter: a systematic review," *International Journal of Market Research*, vol. 62, no. 5, pp. 578–598, 2020.

[19] C. Ducar and D. H. Schocket, "Machine translation and the L2 classroom: pedagogical solutions for making peace with Google translate," *Foreign Language Annals*, vol. 51, no. 4, pp. 779–795, 2018.