ISSN 2616-5775 Vol. 6, Issue 5: 30-38, DOI: 10.25236/AJCIS.2023.060505

Design of Transformer Based English French Translation Model

Xu Bowei*

College of Science, Northeast Forestry University, Harbin, Heilongjiang, 150000, China
**Corresponding author: 910619347@qq.com*

Abstract: *Common Transformer-based translation models use embedding, which can capture the semantic relationships of words in high-dimensional space, and the extracted features are directly transmitted to the attention mechanism of multiple heads. Aiming at the issue of how to better extract the correlation between English and French languages and improve translation quality in English French translation, an English French translation model based on Transform is designed. Combined with the doorstep loop unit, it has the ability to extract semantic information from consecutive sequences, thereby better extracting language information. In terms of improving translation performance, using mask tensors can prevent future information from being used in translation in advance and using greedy algorithms to generate sequences, improve training efficiency and focus translation attention on important information. In terms of optimizing models, using Logsoftmax can alleviate the overflow or underflow problems of Softmax, Replace traditional Transformer's Softmax. and introduce a teacher_ Forcing to avoid errors in the sequence generation process and improve the accuracy of French translation. The experimental results show that under the wmt2014 English French dataset and self-collected dataset, relatively good results have been achieved in effectively extracting association features between two languages and improving translation tasks, with a BLEU of about 0.804.*

Keywords: *embedding; mask tensor; multi-head attachment; vectorization intervention*

1. Introduction

Since the 21st century, with the rapid development of information technology and increasingly frequent exchanges of nationalities [1], machine translation has become increasingly prominent. According to incomplete statistics, the proportion of machine translation in the world every day is higher than the demand for human translation. Translation research, like linguistic research, has gone through a process from convention to description, and then to interpretation. It includes the process of how to translate, what translation actually is, and why to translate [2].

Although machine translation has gradually reached the standard of intelligence after nearly a century of development, and its speed has reached six times the speed of manual translation. This not only saves translation time, but also reduces translation costs to a certain extent [3]. However, there is still a certain gap in the use of high-level thinking, accurate understanding and expression of language, reproduction of style aesthetics, and production of inspirational translations compared to manual translation. The main purpose of machine translation based English French translation is to translate English into French through computers. Traditional statistical machine translation (SMT) has a problem of difficulty in improving system performance [4], while the emergence of neural network which based translation has greatly improved language translation [5-6]. Vaswani [7] and others proposed an attention which based Transformer model in 2017, which has achieved relatively good results in terms of translation speed and quality compared to traditional machine translation. Using parallel corpora to supervise and guide training, however, there are difficulties in obtaining small languages, as well as the need for a large amount of manual translation time costs. However, single corpora are relatively easy to obtain compared to multiple corpora, and the use of single corpora for training has gradually entered people's perspective [8]. According to the translation of source language data into the target language, there is still a problem that the association between the two languages cannot be well handled. In response to this problem, Lample [9] and others are dealing with issues related to the source language and the target language and improving translation accuracy, combining different language data to use BPE (Byte Pair Encoding), and using BPE words together in relevant places to achieve better spatial distribution of word embedding. Ren [[10]et al. In neural machine translation, Wang Tao [11] and others have added predefined bilingual

ISSN 2616-5775 Vol. 6, Issue 5: 30-38, DOI: 10.25236/AJCIS.2023.060505

materials, shared some word vectors, and used additional vectors to enhance the correlation signal, and achieved significant results in translation. Due to the high labor cost of collecting language materials, in the case of possible errors in manual labeling, and the redundancy of decoding information, it is easy to generate over fitting situations. Tao [12] et al. proposed a multi headed attention mechanism to enhance semantic information. Wang [13] et al. proposed a new neural machine translation training method that does not use parallel data pairs, using dictionaries and word embeddings and dual encoders to relate the distribution of two languages in the same semantic space, achieving relatively good results in monolingual machine translation such as English to Russian. Due to the different positions and meanings of the same word in a sentence, it is possible to add location information. It can better capture the relationship between two languages, but may be lacking in capturing the location information of words in the text. In 2022, Google proposed a super strong language model, the Pathways Language Model, with 540 billion parameters. Using sparse algorithms to increase model capacity and other methods, it has achieved higher efficiency in machine translation tasks than in the past.

In response to the above relevant information, this scheme improves the English French translation model based on the conventional translation model of Transform. It uses a doorstep loop unit to extract semantic association information between captured sequences, enhance language semantic information, and combine mask tensors to avoid early use of future information during translation. The multi header attention mechanism allows for richer expression of word meanings, focuses on more useful information, and combines greedy algorithms. Generate language sequences, improve the ability to store too much information in the past, and combine offsets to handle different text lengths. Improve the previous single length tensor problem, improve translation accuracy, and ultimately achieve translation.

2. General ideas for English French translation

An improved English French translation based on Transform is proposed, with the main structure shown in Figure 1.

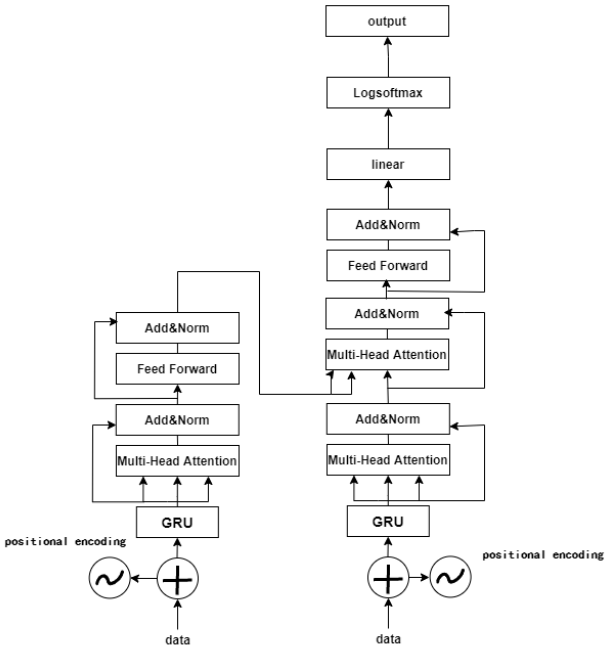

Figure 1: Structure.

The work content of the model is character representation. After a series of pre- processing such as de stressing based on English and French corpora, word segmentation is performed, and corresponding English and French dictionaries are constructed based on word frequency. The text corresponding to English and French is mapped to corresponding numerical tensors based on the dictionary, and word embedding is used to expand the vocabulary dimension, enhance the semantic association of words in high-dimensional space, and capture information between words. Combined with a position encoder, it captures the position information of words at different positions in a sentence. Enhance the semantic information of sentences. Extracting comprehensive language information, representing the obtained English corpus text tensor with characters, and passing it into the GRU module. Gated cyclic neural

ISSN 2616-5775 Vol. 6, Issue 5: 30-38, DOI: 10.25236/AJCIS.2023.060505

networks capture and extract the semantic relationships between long sequence texts. At the same time, it can alleviate the gradient disappearance and gradient decline problems caused by extracting associated semantic information in traditional RNNs, and compared to the long and short memory cyclic neural networks LSTM, which can also capture the associated semantic information of long sequence texts, Its computational complexity is relatively small, obtaining comprehensive semantic information from the language corpus. Using the obtained comprehensive semantic information features as input to the multi header attention mechanism combined with mask tensors, a greedy algorithm is used to construct and generate language sequences and achieve translation.

3. An Improved English French Translation Model

3.1. Character representation

Before inputting English corpus into a computer for translation, it is necessary to process English corpus data into a language that can be recognized by the computer, which is also the primary task of natural language processing. Word vector refers to the representation of a word as a sparse vector, with discrete representation and continuous representation. Discrete is commonly used to represent a word vector as a one-hot encoding. This method is simple in calculation. Using the location of the status register, 1 indicates validity, and the other locations are 0, which has uniqueness. However, it also separates the association between words. When dealing with long sequence text problems, it consumes large memory, and the data presents sparsity. Successively, continuous word bag (CBOW), continuous skip gram (skip gram), Braun clustering, singular value decomposition, and other methods have emerged. This article uses word embedding to capture the distribution information of words in high-dimensional space, and words with similar meanings have similar forms in representation. It calculates the cosine value of the angle between different word vectors to obtain the similarity between words, revealing the semantic relationship between words in text, and uses point mutual information to describe the proximity of the target words to the context words. Thereby not clinging to the semantic association between the vocabulary and the context [14]. Embedding maps one dot in Keras to a vector, which is a representation of a language model. Embedding's work converts x into y in the formal expression (1) as follows:

$$x = [x_1, x_2, x_3, \dots x_{n-1}, x_n]$$
$$y = [[y_{11}, y_{12}, y_{13}, \dots y_{1m}], [y_{21}, y_{22}, y_{23}, \dots y_{2m}], \dots [y_{n1}, y_{n2}, y_{n3}, \dots y_{nm}]] \quad (1)$$

N is the number of words, and m is the embedding dimension of the specified word.

Due to the fact that the same word may generate different semantic information in different locations, while word embedding does not capture the location information of the word. This article adds Position Embedding (PE) [15] to capture the location information of the word, enhance the semantic information of the corpus text, and further deepen the understanding of the corpus text information. The calculation formula is as follows (2):

$$u(i, 2n) = \sin(\frac{i}{10000^{\frac{2n}{d}}})$$
$$u(i, 2n+1) = \cos(\frac{i}{10000^{\frac{2n}{d}}}) \quad (2)$$

i is the position of the vocabulary in the corpus text, d is the input dimension, n is the dimension of the word vector; $in \in [0, \frac{d}{2}]$; $2n$ represents even numbered positions, using sine encoding, $2n+1$ represents odd numbered positions, and uses cosine encoding. In this text, the vocabulary has 256 dimensions, and the six-dimensional images of the word vector are taken. It can be seen that at the same latitude, the coding values of different positions of the vocabulary are different, and the different dimensions of the same vocabulary are different. When there is a positional offset between k words, $u(i+k)$ is a linear function of $u(i)$, which can be used to learn the relative positional relationship [16]. Here, we show the location information of the first 50 lengths in the case of dimensions 7, 8, 9, and 10, and discard it using a dropout0.1 ratio. It can be seen that the information at different locations of the added vocabulary may have different semantic information due to the location of the vocabulary, as shown in Figure 2 below:

ISSN 2616-5775 Vol. 6, Issue 5: 30-38, DOI: 10.25236/AJCIS.2023.060505

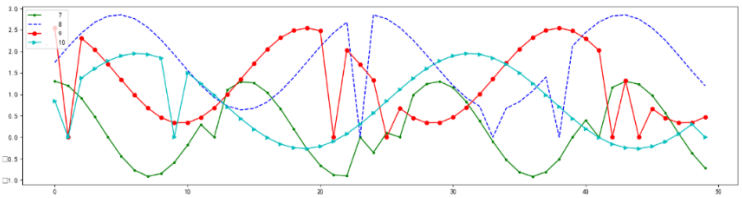

Figure 2: Different locations of the added vocabulary.

3.2. Extracting comprehensive language information

The English French translation encoder is shown like figure 3:

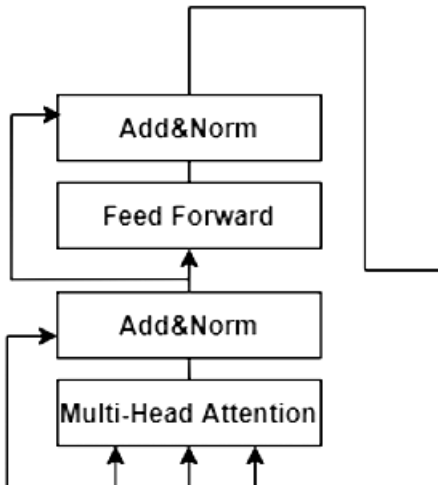

Figure 3: Encoder.

According to the character representation module, the processed feature representation of the obtained language corpus can be further captured by inputting it to enhance text understanding. The scheme uses a Gate Recurrent Unit (GRU) [17] to capture the semantics of the language corpus and extract relevant information between the corpus texts. The Gate Recurrent Unit is a variant of traditional RNN, It can capture the semantic association between long sequences and extract text information, extract some specific pre or post features in language syntax, alleviate the gradient disappearance and gradient explosion problems existing in RNN, and has a simpler structure than Long Short Term Memory (LSTM), with less computational complexity than LSTM. It is composed of reset gates and update gates. The structure of LSTM like figure 4 is shown below:

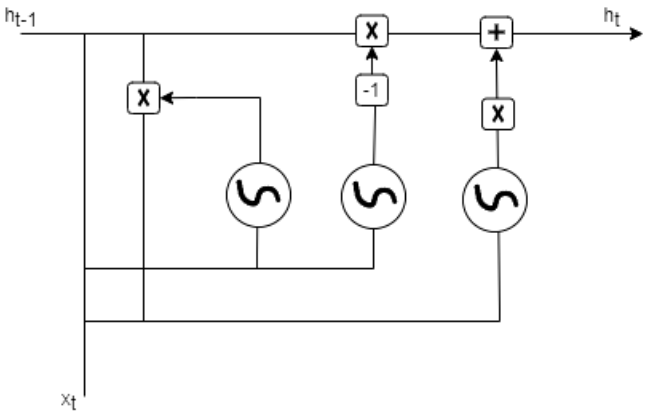

Figure 4: The structure of LSTM.

ISSN 2616-5775 Vol. 6, Issue 5: 30-38, DOI: 10.25236/AJCIS.2023.060505

The expression for how much information transmitted from the previous time step of reset gate control can be utilized is as follows (3):

$$r(t) = \sigma(W_r * [h_{t-1}, x_t])$$
$$\widetilde{h_t} = \tanh(W * [r_t * h_{t-1}, x_t)) \quad (3)$$

The update gate control retains the results of the previous time step and the proportion of the current time step. The expression is as follows (4):

$$z(t) = \sigma(W_z * [h_{t-1}, x_t])$$
$$h_t = (1 - z) * h_{t-1} + z_t * \widetilde{h_t} \quad (4)$$

Reset gate $r(t), t$ is time,W_r is the weight of the reset door, x_t Input text tensor,h_t is the hidden layer output for time t,update gates is $z(t), W_z$ is the weight of the update door. After obtaining the comprehensive language information features, they are input as part of the multi header attention mechanism and combined with mask tensors.

From the source side information passing through the door loop unit, it is assumed that, as shown in the figure, the mask Q_{x2} can be used to query$x1, x2$.

Multi headed attention mechanism is an integrated representation of multiple independent attention mechanisms. The input Query (Q), Key (K), and Value (V) are changed through a linear layer to obtain the feature representation of feature vectors in different subspaces in the same attention mechanism. Compared to single headed attention mechanism, multi headed attention mechanism [18] can alleviate the need for the same vocabulary to type formulas in different features here. The use of attention mechanisms can lead to errors in dimensionality, obtaining richer lexical features, and extracting more useful information from limited computational resources. Attention mechanism: Bahdanau et al. first applied attention mechanism to the field of natural language processing in "Neural Machine Translation by Jointly Learning to Align and Translate" for machine translation tasks. The earliest proposed attention mechanism was in the field of visual images. Google Mind adopted attention mechanism in "Recurrent Models of Visual Attention" for image classification tasks. The attention mechanism can focus attention on more important information, filter out useless information, improve operation speed, and enhance the expression ability of the model. Its attention calculation formula is as follows (5).

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (5)$$

Q is the query vector, K is the keyword vector, and V value vector.

The multi-headed attention mechanism uses a soft attention mechanism, and its specific structure is shown in Figure 5:

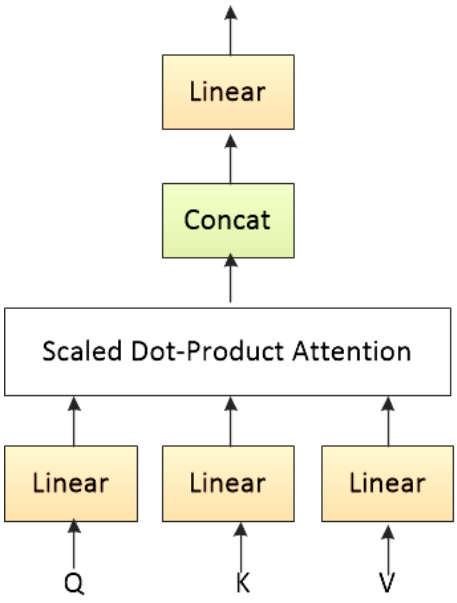

Figure 5: Multi-headed attention.

ISSN 2616-5775 Vol. 6, Issue 5: 30-38, DOI: 10.25236/AJCIS.2023.060505

The calculation rule using formula (5) is combined with the following calculation formula (6):

$$MultiHead(Q,K,V) = Concat(head_1,\dots,head_h)W^O$$
$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \quad (6)$$

W^Q is the weight of Query,$W^K$is the weight of Key,W^V is the weight of Value.

3.3. English French Translation

In generating language sequences, the multi header attention mechanism combines a mask [19], which is a tensor composed of 0 and 1. The scheme word embedding dimension is 256. Here, a 16-dimensional mask representation is shown, as shown in Figure 6 below.

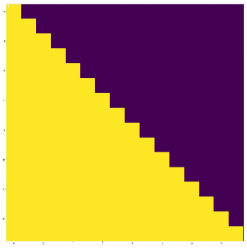

Figure 6: 16-dimensional mask.

As can be seen from the figure, for example, if the position is 1, only the information before 1 can be seen, but the information after 1 cannot be seen. This serves to prevent the use of future information when extracting information, focus on important information, and improve the accuracy of extracting information. The decoder uses a greedy algorithm to generate language sequences and achieve translation. The greedy algorithm is defined as not considering the overall optimal solution, but considering the local optimal solution when solving a problem.

In the classification stage, LogSoftmax is used to replace traditional Softmax to alleviate the overflow or underflow problems of Softmax. The formula is as follows (7):

$$LogSoftmax(x_i) = \log(\frac{\exp(x_i)}{\sum_j \exp(x_j)}) \quad (7)$$

4. Experimental results and analysis

4.1. Data Set

The English and French dataset selected in this article is WMT2014 as the training set, which contains 36M sentences and splits the tags into a vocabulary of 32000 words. In order to improve the robustness of the model, 1000 sentences of the corpus were collected and aligned using last align. In the English French translation task, the corpus script was randomly extracted, and the training set and test set were divided. The training set and test set were 9:1.

4.2. Experimental Environment and Experimental Related Operations

The experimental environment for this article is a Windows 10 operating system, with an Intel (R) Core (TM) i5-8265U (1.80 GHz) CPU, 8GB of running memory, and programming based on Python 3.9. The specific information is shown in Table 1 below:

Table 1: Experimental Environment.

Experimental environment	Environment configuration
operating system	Windows10
CPU	Intel(R) Core (TM) i5-8265U (1.80 GHz)
Running memory	8GB
programming language	python3.9

ISSN 2616-5775 Vol. 6, Issue 5: 30-38, DOI: 10.25236/AJCIS.2023.060505

4.2.1. English French translation data processing

First, collect English corpus and corresponding French predictions, remove duplicates, find out whether there are writing errors, remove accents, remove unconventional punctuation marks, and convert them to ASCII (American Information Exchange Code) data format, and segment words. Construct corresponding English and French dictionaries based on word frequency, and convert the corpus text into numerical sequences as input to the character representation module.

4.2.2. Training of English French Translation Model

The encoder and decoder each use a random gradient descent method to optimize training, and use a back propagation algorithm to update parameters with cross entropy loss. The cross entropy loss function is as follows (8):

$$H(p,q) = -\sum_{i=1}^{n} p(x_i)\log(q(x_i)) \tag{8}$$

$p(x)$ Is the true French probability distribution,$q(x)$ is the translated French probability distribution.

The number of iteration steps is set to 75000 and the word embedding is set to 256. In order to prevent model overfitting and improve model prediction accuracy, a dropout is used and set to 0.1. As the output of the previous time step of the GRU is used as a part of the input of the next time step, in order to prevent errors that may exist in the previous time step, resulting in cumulative increases in errors and affecting model performance, a teacher is introduced here_ Forcing, which is set to 0.5, can avoid errors in the sequence generation process, reduce errors, correct model predictions, and accelerate model convergence speed, allowing the model to converge faster and more smoothly. The mask effect is shown in Figure 7:

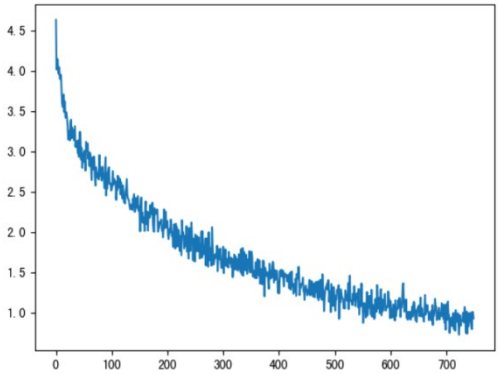

Figure 7: Training loss.

As can be seen from Figure 1, the model can converge and fit the data well, ranging from 0.8 to 0.4 after 70000 iterations.

4.3. Experimental results and evaluation indicators

Using five examples, the prediction results of the model are as follows table 2:

Table 2: The prediction results of the model.

input	output	Correct translation
you are a trouble maker	Vous êtes un fauteur de troubles	Vous êtes un fauteur de troubles
we are short of money	Nous manquons d'argent	Nous manquons d'argent
we are doing this for money	Nous le faisons pour l'argent	Nous le faisons pour l'argent
I am getting used to eat alone	Je me suis habitué à manger seul seul	Je me suis habitué à manger seul
I am ordering you to leave immediately	Je vous ordonne de partir tout de suite.	Je vous ordonne de partir immédiatement

ISSN 2616-5775 Vol. 6, Issue 5: 30-38, DOI: 10.25236/AJCIS.2023.060505

As can be seen from the above example, there are errors in the fourth and fifth sentences. Those with English sentence lengths of less than 6 can be translated almost in line with the translation, while those with more than 6 have repetitive words or synonyms output, which is different from the original translation. According to BLEU, about 0.804 is not considered from the perspective of translation length, and the impact of language length is ignored. In actual scene translation, it is difficult to maintain complete consistency with the translation, Language terms also have a one to many possibility.

5. Conclusion

This research combines the current popular Transform architecture and the idea of seq2seq to design an English-French translation model. On the issue of how to better extract the correlation between English and French languages and effectively translate, it designs an English-French translation based on the Transform architecture, which has achieved certain results. The current problem is that GRU cannot process data in parallel, which to some extent affects the training speed, Secondly, for word segmentation operations in English and French, using the most basic blank word segmentation may have incorrect segmentation results, polysemy, and other issues that affect information extraction, as well as language features such as the difference between terminology and text. Next, in terms of translation, we need to find ways to achieve generalization and common use. Only English French translation can be used to a certain extent, limiting other applications. Next, we will focus on how to improve data parallel processing. We will refer to the parallel layer technology proposed by Google's Pathways Language Model, while improving the generalization and robustness of translation, making it more valuable in applications. Further research will be conducted to meet market requirements.

Acknowledgements

Fund project: Achievements of the 2020 level counselor special project of Hunan University of Humanities and Technology, "Research on collaborative education between college counselors and professional teachers", project number (2020FDY02).

References

[1] Ren Wen. The Challenge and Orientation of Machine Translation Ethics [J]. Shanghai Translation, Issue, 2019(5):46-52+95.
[2] Zheng Man, Hu Xianyao. The Theory, Development, and Prospect of Constructive Translation Theory [J]. Journal of Xi'an International Studies University, 2022, 30(3):5.
[3] Fan Wuqiu, Wang Yu. Spiritual Interaction between Translator and Text: A Bottleneck to Be Breakthrough in Machine Translation [J]. Foreign Language Teaching Theory and Practice, 2022(03): 128-137.
[4] Liu Qun. Survey on statistical machine translation [J]. Journal of Chinese Information Processing, 2003, 17(4):1-12.
[5] Gao Minghu, Yu Zhiqiang. Asummary review of neural machine translation [J]. Journal of Yunnan University of Nationalities (Natural Sciences Edition), 2019, 28(1):72-76.
[6] Kalchbrenner N, Blunsom P. Recurrent continuous translation models [C] //Proc of 2013 Conference on Empirical Methods in Natural Language Processing, 2013:1700-1709.
[7] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need [C]// Proc of the 31st International Conference on Neural Information Processing Systems, 2017:5998-6008.
[8] Artetxe M, Labaka G, Agirre E, et al.Unspervised neural machine translation [J]. arXiv: 1710. 11041v2, 2018.
[9] Lamole G, Ott M, Conneau A, et al. Phrase-based & neural unsupervised machine translation [C] // Proc of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018:5039-5049.
[10] Ren S, Wu Y, Liu S, et al. A Retrieve-and-Rewrite initialization method for unsupervised machine translation[C]//Proc of the 58th Annual Meeting of Association for Computational Linguistics, 2020:3498-3504.
[11] Wang Tao, Xiong Deyi. Enhancing Neural Machine Translation with Predefined Bilingual Pairs [J]. Chinese Journal of Information Technology, 2022, 36 (06): 36-43.
[12] Tao C., Gao S., Shang M., Wu W, Zhao D, Yan R, Get the point of my utterance! Learning towards effective responses with multi-head attrntion mechanism [C/OL]. [2022-06-12]. http: // doi. org/ 10. 24963/ ijcai.2018/614

ISSN 2616-5775 Vol. 6, Issue 5: 30-38, DOI: 10.25236/AJCIS.2023.060505

[13] Wang Xu, Jia Hao, Ji Baijun, Duan Xiangyu. Neural Machine Translation Based on Dictionary Model Fusion [J]. Computer Engineering and Science, 2022, 44 (08): 1481-1487
[14] Feng Zhiwei. Three Methods of Generating Word Vector [J]. Foreign Language Audiovisual Teaching, 2021 (01): 18-26+3.
[15] Gengchen Mai, Krzysztof Janowicz, Yingjie Hu, Song Gao, Bo Yan, Rui Zhu, Ling Cai & Ni Lao. A review of location encoding for GeoAI: methods and applications [J]. International Journal of Geographical Information Science, 2022, 36(4):639-673.
[16] Jiang Menghan, Li Shaomei, Zheng Honghao, Zhang Jianpeng. A Rumor Detection Model Based on Improved Location Coding [J]. Computer Science, 2022, 49 (08): 330-335.
[17] Huang Hao, Zhou Lihua, Huang Yaqun, Jiang Yiting (2022). Early detection of false information based on hybrid depth model Journal of Shandong University (Engineering Edition), 2022 (52 (04)), 89 – 98+109.
[18] Kang Xiaomian, Zong Chengqing. Neural Machine Translation Based on Textual Structure Multitask Learning [J]. Journal of Software, 2022, 33(10):3806-3818.
[19] Liu Hao, Hong Yu, Zhu Qiaoming. Unsupervised Domain Adaptive Machine Reading Comprehension Method [J]. Journal of Computer Science, 2022, 45(10):2133-2150.