LEVERAGING NATURAL LANGUAGE PROCESSING TOOLS AND TECHNIQUES IN THE ENHANCED EFFECTIVENESS OF MACHINE TRANSLATION TECHNOLOGY

Savar Sharma

PDM University, Bahadurgarh, Haryana

ABSTRACT

The development of machine translation technology closely follows that of information science. It is a breakthrough in natural language processing and a big step forward in the field of artificial intelligence. The purpose of this paper is to study the application of machine translation technology based on natural language processing. This paper introduces the relationship between natural language processing and machine translation technology and expounds on the research importance of machine translation technology in natural language processing and the research status at home and abroad. The application of neural networks in natural language processing machine translation systems is introduced. At the same time, for the existing machine translation system, a neural network machine translation system based on reinforcement learning is proposed and analyses the machine translation of different models through case studies. Finally, the Python software is used to implement the program, and the accuracy of the Transformer-PG-NMT model is obtained.

INTRODUCTION

With the acceleration of the process and the increasing frequency of international social exchanges, under the background of the explosion of Internet big data information, cross-border e-commerce, tourism, foreign trade, cross-border finance, social networking and many other fields, as well as national security, intelligence, military and other departments, all face the problem of rapid processing of multilingual data in business processing [1-2]. Machine translation technology can automatically translate natural language accurately and efficiently, and it plays an important role in promoting political, economic, and cultural exchanges, and has important scientific research value [3-4]. Natural language processing is a branch of computer science and linguistics and has always been one of the hot spots in artificial intelligence research and applications. This technology enables computers to better process human natural language, contributes to better communication between people and computers, and between people, and can also establish machine translation systems, information filtering systems, text recognition systems, information retrieval systems, text processing systems and speech recognition systems, etc. [5-6].

Natural language is a unique tool for humans to communicate ideas and understand the basic laws of the world [7]. The work presented by Vani K aims to explore and compare the performance of syntactic- semantic recognition-based language structures using natural language processing methods. Explore the role of linguistic features (i.e. parts of speech, semantics, and semantics) in identifying text fragments and extracting meaning from the WordNet lexical database using a

combined syntactic-semantic approach. The impact of plagiarism type and severity on content removal was analyzed [8]. Choi H proposed a fine-grained (or 2D) observation system in which each dimension of the environment vector receives a separate observation score. In experiments on the En-De and En-Fi translation tasks, the positive feedback approach improves translation quality in terms of BLEU scores. Furthermore, alignment analysis shows how effective response strategies exploit the internal structure of environmental media [9]. There are many different research fields in natural language processing, and the application of machine translation in natural language has a wide range of applications worth exploring.

This paper studies the idea of utilizing reinforcement learning by first pre-training the model using the cross-entropy loss function, then replacing the real data distribution with the sampled output from the model, removing the dependency on the real data during training, and using the model distribution to minimize the loss function to solve the error accumulation problem of end-to-end neural network machine translation system. The research in this paper can make machine translation, a natural language processing task, have a better effect, which is of great significance for future translation research.

RELATED TECHNOLOGIES OF MACHINE TRANSLATION BASED ON NATURAL LANGUAGE PROCESSING

A.Natural Language Processing

1)Word vector: Converting a natural language understanding problem to a machine learning problem requires converting characters known to humans into mathematical symbols recognized by computers. In NLP, the most commonly used word representation is One-Hot Intuitive Representation. This method represents each word as a one-dimensional vector, of which only one dimension is 1 and the others are 0. This one-hot representation is stored in a compact encoding method, which is very concise, but also suffers from a significant problem, which is the phenomenon of "vocabulary holes". That is to say, when words are based on this representation, the correlation between words cannot be found, and the vector representation between any two words is isolated. Since the similarity of vectors can be measured by traditional methods such as Euclidean distance, the biggest contribution of this method of representing textual representations is to make text vectors have distance units to represent groups or similarities between words [10].

2)Attention mechanism: The attention strategy is to make the model pay more attention to the context-related information of the current translation content during the translation process, and pay less or no attention to irrelevant information, thereby improving the translation effect. The dynamic reasoning process takes the word vector representation corresponding to the word during the conversion process, and assigns different weights to each word in the sentence. The dynamic vector environment representation requires three types of information: the hidden state, the hidden state, and the alignment between the encoder and decoder.

3)Machine translation: Machine translation is a secondary function in natural language processing, mainly converting sentences in one language into sentences in another language, and the converted sentences have the same meaning as the original sentences. A more formal description is: Suppose

there is a sentence X={x1,x2,x3,.,.xn} described in a language, where x is a specific character in the language writing table, and the goal of machine translation is to find the sentence in another. Express Y={y1,y2,Y3,...,Ym} in a higher-order language P(Y | X). Machine translation using deep networks has become more and more popular.

B.Convolutional Neural Network Coding

CNN is a multi-layer neural network. A typical convolutional network consists of convolutional layers and aggregation layers. The convolutional layer alternately cooperates to create a multivariate group, the convolutional layer features layer by layer, the pooling layer reduces the output feature vector through the convolutional layer and finally completes the classification through the fully connected layer. The convolution layer can be considered to be inspired by the concept of local receptive field, and extract the underlying features of the object through the convolution kernel of shared parameters. The pooling layer mainly reduces the order of magnitude of the network parameters. CNN encoding generally acts on a two-dimensional word vector matrix, each row represents a word vector of a word, and a word vector dimension is taken as a whole, and the filter of CNN usually covers the upper and lower rows.

First, the text is represented as a vector matrix, that is, the text matrix is convolved with convolution kernels (filters) of different lengths. The size of the filters is equal to the length of the word spoiler, and then using the largest vector extracted from each filter, one number per filter, a vector representing the sentence is obtained by combining these filters.

C. Translator

The transformer completely abandons CNN and only uses an attention mechanism to build the entire model. Transformer can directly obtain the global context information and has a better effect. Transformer is a sequence-to-sequence model, which consists of an encoder that deeply represents the input sentence and a decoder that generates the target sentence.

Before the encoder and decoder process the sentence, to enable the characters to carry more information that is relevant to themselves and helpful for translation, it is necessary to use embedding techniques to embed the characters. There are two pieces of information related to character representation that are very important: semantic information and positional information. To preserve the semantic information of characters, word embedding technology is used to map each character into a high-dimensional space. The embedded characters are represented in the form of d-dimensional vectors, where the d-dimensional embedding space dimension. In this mapped space, characters with the same meaning have a short distance, while characters with different meanings have a farther distance.

D. Neural Network Machine Translation System Based on Reinforcement Learning

The policy gradient algorithm is applied to the system, and the policy function of the PG-NMT model is parameterised.

Then the parameter θ is learned by the method of neuralnetwork, and the objective function LT.

In order to avoid the high deviation caused by a single sampling, the PG-NMT system adopts the method of batch sampling to design the algorithm and uses the average reward of multiple samples as the baseline reward of the model. Matching is easy to cause a large deviation in the calculation of the whole sentence reward, which leads to the problem of poor model convergence.

EXPERIMENTAL DATA AND PARAMETER SETTINGS

A.Data

All models in this experiment are implemented using a torch. The parallel corpus information used is shown in Table 1.

Table I. Information on the Parallel Corpus Used

Translation task	Training set	validation set	test set
Chinese-English	156k	5000	1000
Moral Education - French	1.2M	10000	1000
Japanese-Korean	10.1M	35420	1000

B.Parameter Setting

1)CNN model parameter settings: Both the encoder and the decoder are initialized with a 15-layer convolutional network. The first 6 layers use a 512-dimensional convolution kernel with a width of 3, the middle 4 layers use a 768-dimensional convolution kernel with a width of 3, and the next 3 layers are 1024-dimensional with a width of 3. Convolution kernels, the last layer uses a 2048-dimensional convolution kernel with a width of 1, and the last layer uses a 4096-dimensional convolution kernel with a width of 1. Before the fully connected layer, the output generated by the decoder, all word vectors are 768 dimensions.

2)Transformer model parameter settings: The reinforcement learning algorithm is applied to the system.

On the Chinese-English machine translation task, both the encoder and the decoder use 6 layers of attention, and each layer uses 4 attention heads. , the word vector dimension is512dimensions; for the moral education-French machine translation task, both the encoder and the decoder use 6 layers of attention, each layer uses 8 attention heads, and the word vector dimension is 512 dimensions; for Japanese-Korean machine translation For the translation task, both the encoder and the decoder use 6 layers of attention, each layer uses 16 attention heads, and the word vector dimension is 1024 dimensions.

APPLICATION ANALYSIS OF MACHINE TRANSLATION TECHNOLOGY

A.Performance Comparison of Reinforcement Learning Algorithms

The concepts of reinforcement learning and deep learning algorithms are integrated into the architecture of machine learning algorithms, and a two-step machine learning system based on

reinforcement learning algorithms is constructed based on CNN. The results are shown in Figure 1. It can be seen that the machine transformation system using reinforcement learning can achieve better performance than the basic model system of CNN and Transformer; the performance of the machine learning transformation system based on the Transformer model is better than that of CNN.

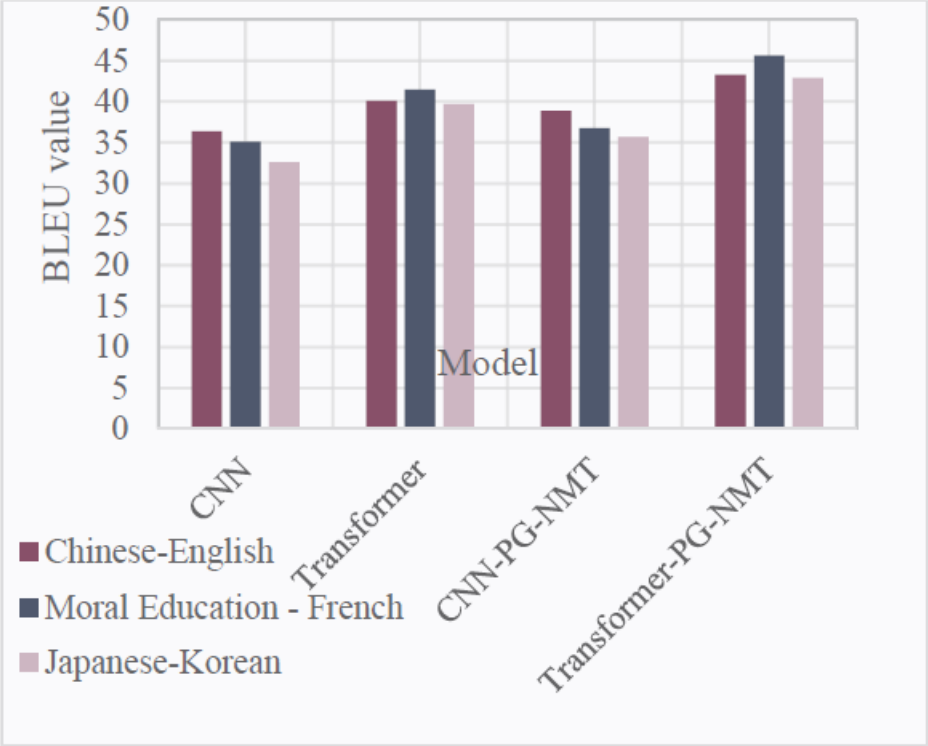

Fig. 1. Performance of different reinforcement learning methods in machine translation models.

For the Chinese and English datasets, the BLEU value of the CNN-PG-NMT model is 2.5 higher than that of the machine-defined CNN, and the BLEU value of the Transformer-PG- NMT model is 3.2 higher than that of the machine-defined CNN explained CNN. The Transformer-PG-NMT model has the highest performance with a BLEU value of 43.3; for the Social Education Dataset-French, the CNN-PG-NMT model improves the BLEU value of the CNN model for machine interpretation by 1.7, while the Transformer-PG- NMT model Outperforms models in machine interpretation. The BLEU value of the CNN-PG-NMT model is improved by 4.1, while the Transformer-PG-NMT model has the highest performance with a BLEU value of 45.6; for the Japanese and Korean datasets, the CNN-PG- NMT model has higher performance than the CNN machine interpretation method. BLEU value.

B. Deviation Analysis of the System

As shown in Figure 2, in the first part of training, the machine learning method based on reinforcement learning CNN will have good value when calculating the reward value.

Transformer-based reinforcement learning machine learning also has negative returns in early training sessions. Compared with the CNN model that describes the RL machine, the value of the loss model is based on the Transformer model.

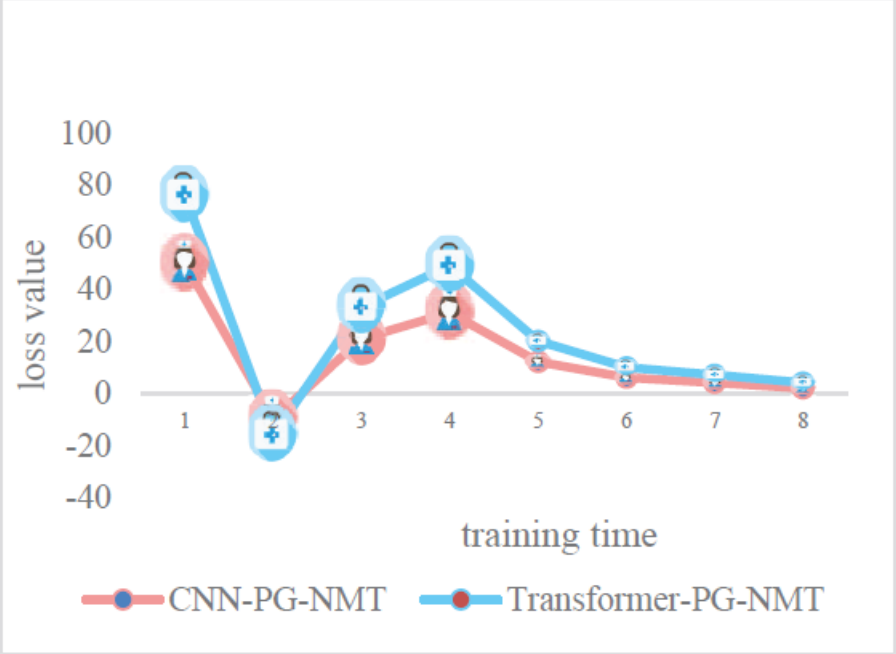

Fig. 2. Analysis of loss value of CNN-based reinforcement learning model.

CONCLUSIONS

As one of the important fields of artificial intelligence research, natural language processing is the key to the computer's cognitive learning of real knowledge. Words are the basic unit of natural language, and the research on word representation theory is of great significance for the improvement of natural language processing technology. This paper studies and applies machine translation based on the existing theoretical research results of natural language processing technology. This paper conducts an in-depth study on the translation system incorporating reinforcement learning theory, and on this basis studies how to make the model better maintain the advantages of the interpretability and flexibility of the translation system, while introducing the powerful learning ability of neural networks and generalization ability, which are then applied to text translation tasks with good results.

REFERENCES

[1]Balsmeieri B, Assaf M, Chesebro T, et al. Machine learning andnatural language processing on the patent corpus: data, tools, and newmeasures[J]. Journal of Economics & Management Strategy, 2018,27(3):535-553.

[2]Selby L V, Narain W R, Russo A, et al. Autonomous detection,grading, and reporting of postoperative complications using naturallanguage processing[J]. Surgery, 2018, 164(6):1300-1305.

[3]Mcgregor K A, Whicker M E. Natural Language Processing Approaches to Understand HPV Vaccination Sentiment[J]. Journal of Adolescent Health, 2018, 62(2):S27-S28.

[4]Mizera-Pietraszko J. Natural language processing for social media (2nd ed.)[J]. Computing reviews, 2020, 61(1):23-24.

[5]Zhekova M, Totkov G. Model of process and model of natural language processing system[J]. IOP Conference Series Materials Science and Engineering, 2020, 878(12028):1-12.

[6] Morgan D G, Chorneyko K, Swain D, et al. 279 – Validation of a Natural Language Processing Algorithm to Identify Colonic Adenomas Across a Health System[J]. Gastroenterology, 2019, 156(6):S-56.

[7]Vgp A, Smnf B . Do no harm: Natural language processing of social media supports safety of aseptic allergen immunotherapy procedures[J]. Journal of Allergy and Clinical Immunology, 2019, 144(1):38-40.

[8]Vani K, Gupta D. Unmasking text plagiarism using syntactic-semantic based natural language processing techniques: Comparisons, analysis and challenges[J]. Information Processing & Management, 2018, 54(3):408-432.

[9]Choi H, Cho K, Bengio Y. Fine-Grained Attention Mechanism for Neural Machine Translation[J]. NEUROCOMPUTING, 2018, 284(APR.5):171-176.

[10]Maruf S, Saleh F, Haffari G. A Survey on Document-level Neural Machine Translation: Methods and Evaluation[J]. ACM Computing Surveys, 2021, 54(2):1-36.