Electronics, Communications and Networks
A.J. Tallón-Ballesteros et al. (Eds.)

doi:10.3233/FAIA231215

Design and Implementation of Natural Language Processing Machine Translation System Based on Seq2Seq Model

Wenqi ZHANG[a], Yiqin BAO[a1] and Wenbin XU[b]
[a] College of Information Engineering of Nanjing XiaoZhuang University, China
[b] Jiangsu United Vocational and Technical College Suzhou Branch, China

Abstract. Aiming at the bottleneck of system performance optimization in Machine translation system, long-distance dependence, difficulty in processing low-frequency words and rare words, a Natural language processing Machine translation system based on Seq2Seq model is designed, which effectively realizes multilingual translation and language communication, realizes accurate translation of multiple languages, and improves the efficiency of language communication. Based on the Natural language processing technology of Seq2Seq model, this paper quickly recognizes and translates the source language text, outputs the target language text, and through training the model, selects the task of French translation into English for experiment. The attention weight shows that the system performs well in the translation task.

Keywords. Machine translation, natural language processing, deep learning, attention weight

1. Introduction

Artificial intelligence is changing from academic-driven to application-driven, and moving from specialized intelligence to general intelligence, which is closer to the level of human intelligence than any other period in history, and has entered a new stage of development, and Natural Language Processing (NLP), as one of today's key application areas of Artificial Intelligence, has made great progress in recent years [1].

With the development of globalization, the communication between different languages has become more and more frequent. As an automatic translation method, machine translation has been widely concerned because of its high efficiency and speed. Under the background of digital and information age, the communication demand of multilingual environment is becoming more and more intense. As an important application of language information processing technology, machine translation plays an increasingly important role in improving the efficiency of language communication and solving the problem of cultural differences. Therefore, in the field of translation, the use of artificial intelligence and neural network machine translation model has become the mainstream of research. With the development of machine learning and natural language processing technology, the use of intelligent machine learning methods to carry out machine translation research of various languages has attracted more attention. However, the language of different countries has a large difference, and it is easy to be affected by the instability of rhyme output in the process of translation, It is necessary to combine

[1] Corresponding author: Yiqin BAO, email: 392335241@qq.com

semantic learning and machine learning methods to establish a semantic parameter parsing model under the data screening mode Type [2] realized machine translation design through data screening and deep learning methods, designed a machine translation system based on Seq2Seq model, and designed the machine translation system through robotics, natural language processing technology, deep learning and context correlation technology.

Based on the powerful natural language processing technology and deep learning technology, by identifying the source language sentences through the decoder and decoder to quickly translate the output target language sentences. The accuracy of machine translation is greatly improved by learning the grammatical requirements and contextual relevance of the translation of the language through big data.

The contributions and innovations of this paper are summarized as follows:

1) Designed the architecture of machine translation system based on Seq2Seq model.

2) Studied the related techniques and implemented the training of the machine translation model.

3) Tested the system for translation.

The rest of the paper is organized as follows. The second section investigates the techniques, the third section designs the architecture of the machine translation system based on Seq2Seq model, the fourth section performs the tests and the fifth section concludes the full paper.

2. Related technology

2.1 Machine Translation

Machine translation refers to the use of machines to translate a natural language in written form or sound form into another natural language in written form or sound form, and it is the earliest problem proposed and studied in the field of natural language processing. So far, machine translation has mainly experienced three eras: rule-based machine translation, statistical machine translation (SMT) and neural machine translation (NMT) [3]. With the development of deep learning technology, the neural machine translation combined with neural network has surpassed other machine translation techniques.

Machine translation is actually the process of using a computer to translate one natural language into another natural language， The principle of machine translation is to use a computer to convert the source language (Source) into the target language (Target)， The process is Source → Analysis → Conversion → Generation → Target, and the machine translation process is shown in Figure 1.

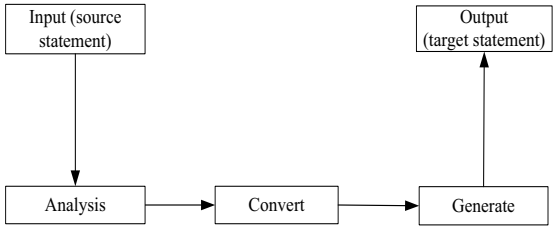

Figure 1. Machine translation process

2.2 Natural Language Processing

Natural Language Processing abbreviated as NLP, may be regarded as "natural language" and "processing" two parts, natural language is different from computer language, is the human development process as a tool for communication of language and text, spoken and written language is a natural language, such as people use the Chinese language, Arabic, Korean, and English, etc. [4-5].Natural Language Processing is an interdisciplinary discipline that integrates computer science, artificial intelligence and linguistics to study how computers can learn to process human language and understand it through techniques such as machine learning, as shown in Figure 2:

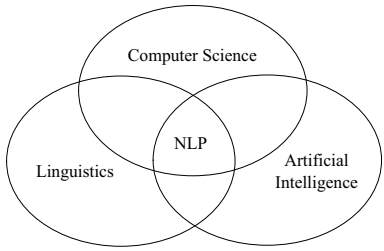

Figure 2. Natural Language Processing

The process of computer processing of natural language can be described in four ways: normalization, algorithmization, systematization and application.

1)Normalization: the natural language problem to be studied is built into a formal model and normalized into a mathematical form for rigorous and consistent representation. This process mainly involves the description and analysis of syntax, semantics and pragmatics.

2)Algorithmization: transforming mathematical models into executable algorithmic steps so that computers can understand and process natural language. This process involves the design and implementation of processing steps such as how to perform disambiguation, syntactic analysis, and semantic understanding of natural language.

3) Systematization: Based on the algorithm, a natural language processing system is built, i.e., the algorithm is transformed into a practically usable software system. This process includes system architecture design, algorithm implementation, performance optimization, etc. The purpose is to build an efficient and reliable natural language processing system.

4)Applicationization: The natural language processing system is evaluated and improved so that it can really meet the practical needs. In the applicationization phase, various tests, evaluations and optimizations are conducted to ensure the applicability and reliability of the system in different domains and tasks.

These four aspects describe the entire process of computer processing of natural language, from problem modeling to algorithm design to system implementation and application evaluation, leading to the understanding and processing of natural language.

3. Implementation of Machine Translation Based on Seq2Seq Model

3.1 Seq2Seq model

Seq2Seq (Sequence to Sequence) is a method that can generate another sequence based on a given sequence through a specific method. It was proposed in 2014. Its input is a

sequence, and its output is also a sequence. The Seq2Seq framework [6] is an important RNN model, also known as the Encoder-Decoder model, which can be understood as an N×M model. The model consists of two parts: The Encoder is used to encode the information of a sequence, encoding the information of a sequence of arbitrary length into a vector c. The Decoder is the decoder, and the Decoder is the decoder. The Decoder is the decoder, which can decode the information and output it as a sequence after getting the context information vector c. The structure of the Seq2Seq model is shown in Figure 3.

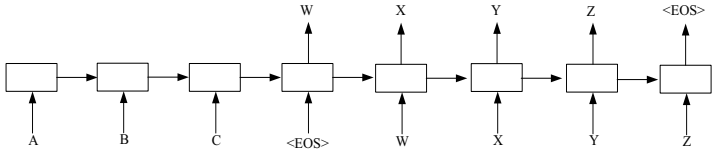

Figure 3. Seq2Seq model structure

3.2 Seq2Seq *architecture*

Most of the Seq2seq models in NLP use the encoder decoder framework, which encodes the input Sequence with an Encoder and outputs the Sequence with a Decoder. The framework of the Seq2seq model is to obtain a hidden state from a sequence through a decoder, and then use the decoder to obtain the final desired sequence through this hidden state. The Seq2Seq model mainly includes the input layer, encoder, decoder module, and output layer. The input layer represents the source language in vector form, the encoder converts the source language sentence into a series of hidden layer representations, the decoder generates the target language based on the hidden layer representation, and the output layer represents the target language as a word vector. As shown in Figure 4, a text translation task from Chinese to English is presented, which well demonstrates the construction of the seq2seq model.

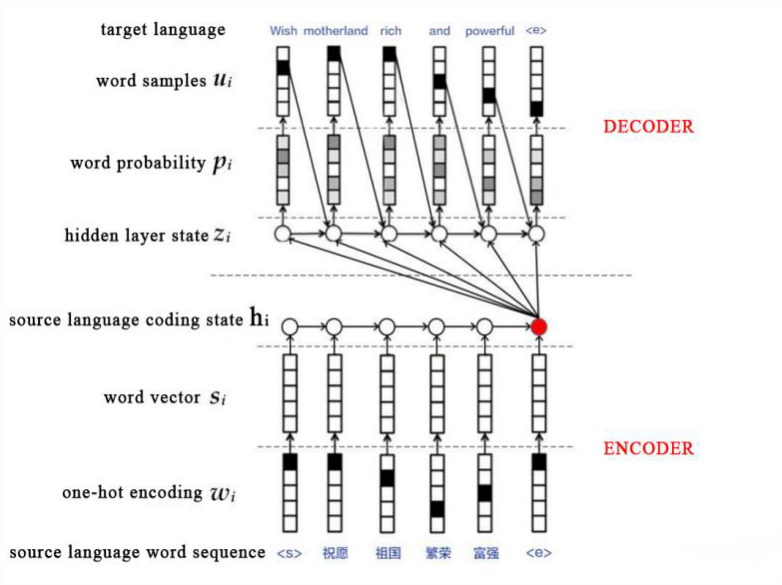

Figure 4. Seq2seq architecture diagram

3.3 Training Models

he dataset we are using is the Multi30k dataset, which is a dataset with approximately 30000 parallel English, German, and French sentences, each containing~12 words per sentence. We need the input tensor (the word index in the input sentence) and the target tensor (the target sentence's word index) for training. In creating these tensors, we add EOS tags to both sequences. Training is done by running the input sentences through the encoder and keeping track of each output and the latest hidden state. The decoder is then given the <SOS> flag as its first input and the last hidden state of the encoder as its first hidden state.

The entire training process is:

1) Start timing;
2) Initializing the optimizer and criterion;
3) Creating the translation pair dataset for training;
4) Creating the loss array for drawing.

While the model is being trained, the trainer is called several times and the progress is displayed, as shown in Figure 5.

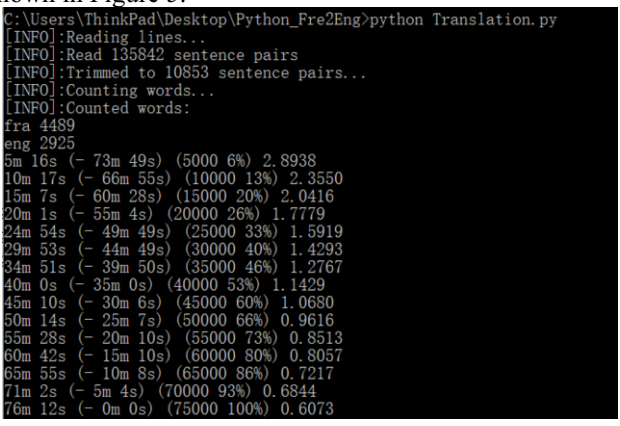

Figure 5. Training progress

After the training progress was completed, the average loss was outputted, and as shown in Figure 6, the average loss was reduced to about 0.5.

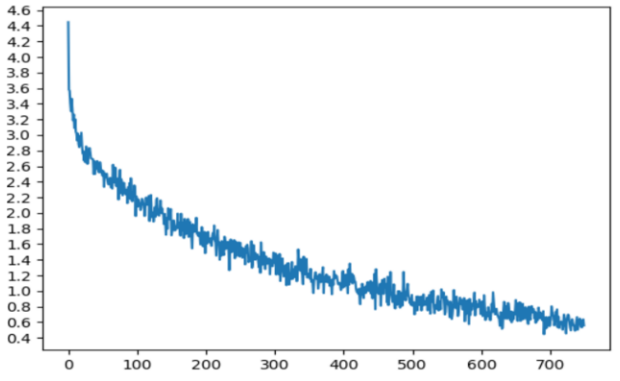

Figure 6. Average Loss

4. Model Testing

By training the model, testing, this paper uses Pycharm platform as the program deployment platform, the test in this paper is to translate French into English, select a few sentences in the test set to test, the translation results are shown in Figure 7.

```
> tu es completement ignorant .
= you re totally ignorant .
< you re totally ignorant . <EOS>

> je suis obstinee .
= i m stubborn .
< i m stubborn . <EOS>

> nous sommes pieges !
= we re trapped !
< we re trapped ! <EOS>

> tu n es pas la bienvenue ici .
= you re not welcome here .
< you are not welcome . <EOS>

> j essaie de me rappeler .
= i m trying to remember .
< i m trying to sleep . <EOS>

> vous etes gare en double file .
= you re double parked .
< you re double parked . <EOS>

input = elle a cinq ans de moins que moi .
output = she s been years younger than me . <EOS>
input = elle est trop petit .
output = she s too loud . <EOS>
input = je ne crains pas de mourir .
output = i m not scared to die . <EOS>
input = c est un jeune directeur plein de talent .
output = he s a talented young director . <EOS>
```

Figure 7. Test translation results

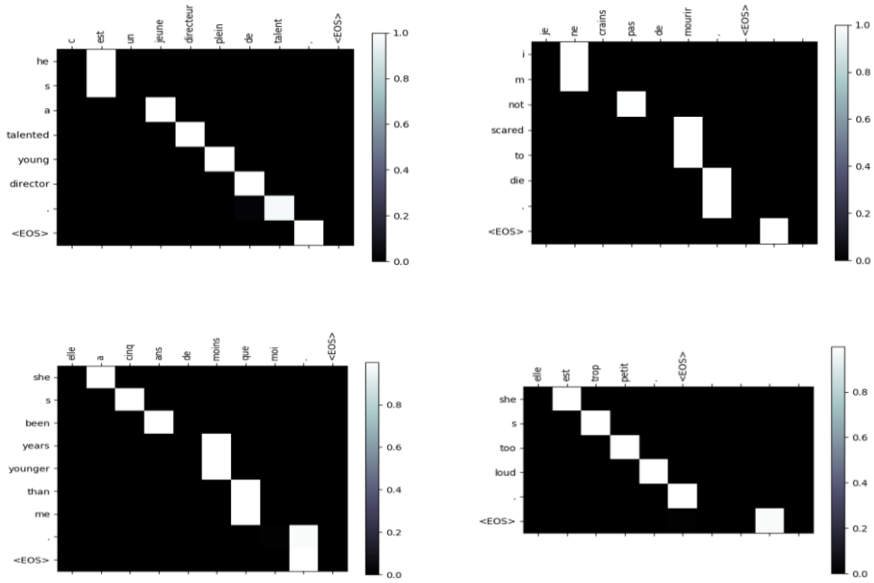

Figure 8. Attention weighting results

Attention mechanism is a very effective and widely used technique applied in several task areas such as natural language processing, computer vision, and so on. Based on the calculated attentional weights, the translation results can be evaluated. As Figure 8 shows the translation results of the system for a given input sentence and shows the attention weights, the results show that the translation results of this system are relatively good.

5. Conclusions

This paper presents the design and implementation of a natural language processing machine translation system based on Seq2Seq model. The system framework includes four main modules, namely, input layer, encoder, decoder and output layer, which are respectively responsible for transforming source language sentences into hidden layer representations and generating target language sentences. The system implementation aspect includes specific model selection, parameter setting and training strategy. By training the model and testing it, the task of translating from French to English was selected for experimentation, which shows that the system performs well on the translation task. This paper can also provide valuable references and insights for further research and application of the natural language processing machine translation system based on Seq2Seq model.

References

[1] Li Fengsu. A comparative study on the quality of human-computer English-Chinese translation in the era of artificial intelligence[J]. Foreign language community,2022(04):72-79.
[2] Yuan Min. Design of Korean machine translation system based on improved neural network[J]. Automation and Instrumentation,2023(01):212-215+220.DOI:10.14016/j.cnki.1001-9227.2023.01.212.
[3] Li-Yuan Zhang. Machine translation based on bidirectional decoding consistency at the target end[D]. Harbin Institute of Technology,2019.DOI:10.27061/d.cnki.ghgdu.2019.001562.
[4] Peipei Shi. Research on Chinese text classification based on hybrid neural network model[D]. Southwest University of Finance and Economics, 2020.DOI:10.27412/d.cnki.gxncu.2020.002689.
[5] WU Xiao-kun, ZHAO Tian-fang. Application of Natural Language Processing in Social Communication:A Review and Future Perspectives[J]. Computer Science, 2020, 47(6): 184-193.
[6] Xi Yawen. Research and analysis on key technology of LSTM text classification based on deep migration[D]. Southwest Jiaotong University, 2019.DOI:10.27414/d.cnki.gxnju.2019.001615.