English Machine Translation Model Based on Artificial Intelligence

Xueting Li¹*, Xing Hao²

¹Modern College of Northwest University, Xi’an, Shaanxi 710130, China
²Northwestern Polytechnical University, Xi’an, Shaanxi 710072, China

*Corresponding author: 651030138@qq.com

Abstract. Machine translation is a language processing method that is widely used at present. However, because the machine translation mechanism extremely relies on rigid grammatical rules, the system often has large understanding deviations in the process of processing the original text, and it is often difficult to accurately grasp the semantic features. In this context, the author proposed an optimization method for English machine translation which is based on artificial intelligence on the premise of weakening grammatical rules. The above method first analyzes the grammar rules of English, and then obtains the semantic Gaussian marginalization rectangular window function. These functions project a large amount of information entropy data of English semantics. In order to obtain the nonlinear spectral features of adaptive matching semantics, the semantic correlation factors of words are added to the text information entropy and information gain. Simulation results show that the proposed method achieves good machine translation effect and has good development space.

Keywords: Machine Translation, Weakening Rules, Artificial Intelligence, Adaptive Matching.

1. Introduction

China's comprehensive national power and international competitiveness have been greatly improved in the past few decades, and with it comes the deepening of trade and cultural exchanges with other countries and nations. Since the 1920s, English has been used more and more widely as the world's official language, and it is also an effective tool for the communication between China and other countries in the modern society [1-2]. In order to achieve cross-cultural communication in a timely manner, various English translation machines came into being, and the history of English machine translation can be traced back to the eighties of last century. Over last decades, great changes have taken place in English machine translation technology. There are so many translation models available that they can satisfy the demand of most people. However, there are some issues that cannot be ignored. At present, most English machine translation mechanisms are characterized by word sense disambiguation and semantic role tagging. Based on the semantic network, the translation model designed in the study not only has independent semantic expression capability, but is also equipped with the capacity to describe the relation between words, providing accurate translation services for users in various fields.
It was in the 1940s that the machine translation first appeared [3]. It is a new topic with the birth of computer technology. Its goal is to translate one natural language into another on the platform of computer. However, for a variety of natural languages, most of the current machine translation models put too much emphasis on the manual summary of language experts, resulting in insufficient processing of language details and even errors. Therefore, how to effectively optimize the results of English machine translation under the premise of weakening the grammatical rules has become a very important topic [4-6]. The corresponding optimization method can adaptively match the non-linear spectral features of the semantics, and realize the characteristic extraction and recognition of semantic information fusion [7]. It will be an effective way to deal with the problems which were mentioned above. On the premise of weakening the role of grammatical rules, optimizing the existing English translation mechanisms has great practical significance. This subject has received increasing attention and has made great progress.

2. Design and improvement of English machine translation model

2.1. Design of English machine translation model based on semantic network

Linguists strictly analyze the translation results and update the knowledge base to improve the use of English translators. The function of the computer scientist is to input fixed values to the machine. After the system is designed successfully, the flexibility of the translation results is low and cannot be adjusted. Therefore, semantic network is integrated into the research paradigm of rule-based English machine translation to shape the model of semantic Network-based English machine translation, as shown in Figure 1.

Figure 1 English machine translation model based on semantic network

Figure 1 introduces the semantic network part. The translation result of statistical machine translation model based on semantic network is more accurate, which makes the learning effect of language learners more obvious and enables them to learn English efficiently in limited time. The knowledge base has become a corpus containing a lot of information, and the design of algorithms by computer scientists is more comprehensive and systematic.

2.2. English machine translation method based on phrase synthesis semantic statistics

2.2.1. Translation similarity model. The similarity degree of two different multidimensional vectors \( \hat{u} \) and \( \hat{v} \) in the same semantic space can be described by semantic similarity. The greater the similarity of vectors \( \hat{u} \) and \( \hat{v} \), the greater the semantic similarity of vectors \( \hat{u} \) and \( \hat{v} \). “Similarity” is increasingly widely used. For example, natural language analysis covers the concept of semantic similarity, while the specific performance of semantic similarity is applied into the translation similarity model in machine translation. In verification analysis, the translation semantic similarity of two vectors \( \hat{u} \) and \( \hat{v} \) can be obtained in various ways. This paper gives an example of cosine similarity, which is the most commonly used method.
2.2.2. Cosine similarity. Cosine similarity is also called cosine distance. In multidimensional space, cosine of the angle between two vectors is used as the standard to measure the difference between the two vectors. When the cosine value increases, the angle between the two semantic vectors will decrease, and the semantics of the two words will be closer; when the cosine value decreases, the angle between the two semantic vectors will increase, and the semantics of the two words will not be closer. The English translation similarity of bilingual vectors \( \vec{u} = [a_1, a_2, ..., a_n] \) and \( \vec{v} = [b_1, b_2, ..., b_n] \) is as follows [8-10]:

\[
Sim(\vec{u}, \vec{v}) = \frac{\vec{u} \cdot \vec{v}}{\|\vec{u}\| \times \|\vec{v}\|} = \frac{\left( \sum_{i=1}^{n} a_i \times b_i \right)}{\sqrt{\sum_{i=1}^{n} a_i^2 \times \sum_{i=1}^{n} b_i^2}}
\]

(1)

2.2.3. Weighted vector addition. Through formula (2), we can get the composite semantic vector \( \vec{p} \) of two word semantic vectors \( \vec{u} \) and \( \vec{v} \) in the same corpus:

\[
\vec{p} = \vec{u} + \vec{v} = \left[ (a_1 + b_1), (a_2 + b_2), ..., (a_i + b_i) \right]
\]

(2)

Set the semantic vector of the word "machine" as the 5-Dimensional vector \( \vec{u} = [2,6,8,7,1] \), and the semantic vector of the word "translation" as the 5-Dimensional vector \( \vec{v} = [1,3,4,5,6] \). Then according to formula (2), the semantic vector of the synthetic phrase "machine translation" is \( \vec{p} = [3,9,12,7] \).

The above error can be effectively solved by adding weighted vectors to obtain the synthetic semantic vector of the synthetic phrase:

\[
\vec{p} = \alpha \vec{u} + \beta \vec{v}
\]

(3)

Similarly, the steps of synthesizing semantic "machine translation" are analyzed. After corpus training, the weight of "machine" in the "machine translation" phrase is \( \alpha = 0.6 \), and the weight of translation is \( \beta = 0.4 \). According to formula (3), the semantic vector of "machine translation" is \( \vec{p} = [2.6,4.8,6.4,6.2,3.0] \).

In this case, the phrase "translation machine" is regarded as a new phrase, and new weight training is implemented for "translation" and "machine". The weights are 0.3 and 0.7 respectively. The semantic vector of "translation machine" is \( \vec{q} = [1.7,5.1,6.8,6.4,2.5] \). By comparing the results of the two semantic vectors, it is found that the differences between the two phrases can be easily distinguished this time.

The synthetic semantic vector of multi word phrases can be obtained by formula (4):

\[
\vec{p} = \sum_{i=1}^{n} \lambda_i \vec{w}_i
\]

(4)

Where: the word semantic vector of each combination unit of a multi word phrase is represented by \( \vec{w}_i \), and the weight of each combination unit is represented by \( \lambda_i \).

2.3. Improvement of English machine translation model based on semantic network

This paper discusses the current level of machine translation, and then looks forward to its future development trend.

Future research should introduce deeper language and knowledge reserves, more accurate and advanced technology. The specific processing mechanism is to introduce the participation of linguists in the process of machine translation, so as to continuously add new relevant information and update the data of the corpus. From the perspective of scientificity and effectiveness, this method can effectively improve the traditional processing models of machine translation and realize the effective adjustment of the work flow of the research paradigm, thereby constructing the New Semantic Web-
based English machine translation model, as shown in Figure 2.

![Figure 2](image)

**Figure 2** Improved model of English machine translation based on semantic network

### 3. Experimental analysis

#### 3.1. Experimental data

In order to verify whether the English machine translation model put forward in this research can get accurate translation results, a partial subset of the LDC corpus was adopted as the experimental data for this study, including 4 million pairs of parallel sentences, 98.9 million Chinese words and 112.6 million English words. NIST 05 was adopted as the experimental development set, which included 1082 Chinese sentences, each with 4 translation results. In other words, there are a total of 4328 English sentences in this experimental development set. This experimental test sets are NIST 06 and NIST 08 respectively. The test set NIST 06 contains 1 664 sentences and 4 English sentences, that is, 6 656 English sentences. Test set NIST 06 contains 1357 Chinese sentences and 4 translation results for each one, that is, there are totally 5428 English sentences.

#### 3.2. Experimental setup

The decoder of the English machine translation model designed in this research is the C++ implementation version of the layered phrase decoder, and the specific processing steps are as follows. First of all, the GIZA++ tool is used to realize the alignment of word information in Chinese-English and Chinese-English directions, and the heuristic effect of grow-diag-final-and is used to realize the state of many-to-many word alignment. Generally speaking, the system will present better translation performance when the number of cross-connections of word alignment is small in translation results. The quaternary English language model is obtained in Xinhua part of Gigaword by using SRILM tool. Because of the instability of MERT, it is necessary to repeat the experiment three times in the way proposed by Clark et al., and take the final average value as the experimental result.

#### 3.3. Experimental result

1) To ascertain the accuracy of the model designed in this paper in English translation, the hierarchical phrase translation model, the word distribution semantic information model and this model are tested experimentally. The translation results of different data sets are shown in Table 1. The evaluation index of this experiment is BLEU value, which can be obtained from analysis table 1. Based on the test sets NIST06 and NIST08, the translation results obtained by this model are increased by 0.35 and 0.23 respectively compared with the translation results of hierarchical phrase translation model. Similarly, based on the test sets NIST 06 and NIST 08, the translation results of this model are 0.12 and 0.03 higher than those of the hierarchical phrase translation model with word distribution semantic information, respectively. It can be seen that the translation results provided by the model present better accuracy and scientificity. The translation results obtained by significance test conform to the condition of $\rho < 0.05$, which shows that the performance of the translation results has improved significantly.
2) In order to verify that the performance of the model designed in this paper is higher than that of other models, this paper explains "Lanzhou Price Bureau limits the price of beef noodles" in NIST 08 test set. Only because the sentence "too much increase" is translated into English by using the hierarchical phrase model, the hierarchical phrase translation model with word distribution semantic information and the model proposed in this research. The experimental results can be seen in Table 2 below.

Table 2 Comparison table of translation examples

| Source language sentences | Lanzhou Price Bureau on beef noodle price limit to make an explanation: only because the increase is too large |
|---------------------------|------------------------------------------------------------------------------------------------------------|
| Reference translation     | Lanzhou price bureau gives explanation of price control on beef noodles: it is only because the raises have been too large |
| Hierarchical phrase       | Lanzhou explained beef noodles reduce: only because of the excessive price |
| translation model         |                                                                                                           |
| Adding words to           | Lanzhou explained that beef noodles reduce: only because of the excessive price increase |
| distribute semantic       |                                                                                                           |
| information               |                                                                                                           |
| Model in this research    | Lanzhou gives explanation of beef noodles reduce: only because of the excessive price raises |

According to the analysis of Table 2, in the specific process of translation, the term “Price Bureau” is not translated by the three models, and then the term "explain" is analyzed. The result of the first two models is explained. The translation results provided by this model are the most consistent with the reference translation, indicating that the translation results of this model are more accurate than those of other models.

3) As mentioned in the experimental setup, the number of cross-connections for word alignment in the translation result is inversely proportional to the translation performance of the system. Three models were used to analyze the cross connection number of English translation results.

It can be seen from the analysis of Table 3 that the number of cross connections in the translation results of the hierarchical phrase model is 29.2, and that of the hierarchical phrase model with word distribution semantic information is 4.7 less than that of the former one, indicating that the performance of the translation results is improved. In this model, the number of cross links in the translation result is 17, which is significantly less than that of the other two models, and it shows that it has better translation performance.

Table 3 Average of the number of aligned cross connections per sentence pair from Chinese to English

| Experimental model                  | Number of cross connections |
|-------------------------------------|-----------------------------|
| Hierarchical phrase translation model | 29.2                        |
| Adding words to distribute semantic information | 23.9                        |
| Model in this research              | 17                          |
4. Conclusion
In general, the performance of the English machine translation model put forward in this research is very good. In addition to the ability of independent semantic expression, it can also accurately express the relationship between words, thereby giving translation accurately in accordance with the meaning of the original text. This model has great practical significance. On the one hand, it can act as a medium that allows users to quickly obtain English translation results for reference, helping them save time. On the other hand, it can promote cross-cultural exchanges and trade activities. In the future, big data and multiple information will become the development direction of English machine translation.

References
[1] Hu Wenzhong, Sun Youzhong. Highlighting Discipline Characteristics and Strengthening Humanistic Education -- on Current English Teaching Reform. Foreign Language Teaching and Research, 2006, 38 (005): 243-247.
[2] Liu Lude. The Enlightenment of Problem-based Learning on Teaching Reform. Education Research, 2002 (2): 73-77
[3] Zheng Xinmin, Jiang Qunying. a Study on "teacher Belief" in College English Teaching Reform. Foreign Language Circles, 2005 (6): 16-22
[4] Zhou Yuanqing. The Construction of Excellent Course Materials is an Important Measure of Teaching Reform and Innovation. China Higher Education Research, 2003 (1): 12-12
[5] Li Zhiyi, Zhu Hong, Liu Zhijun. Guiding the Teaching Reform of Higher Engineering Education with the Concept of Achievement Oriented Education. Higher Engineering Education Research, 2014, 000 (002): 29-34
[6] Ye Lan. Let the Classroom Radiate Vitality -- on the Deepening of Teaching Reform in Primary and Secondary Schools. Education Research, 1997 (09): 3-8
[7] Liang Dingfang. My View on Foreign Language Teaching Reform. Foreign Language Teaching Theory and Practice, 2001 (1): 8-11
[8] Huang Jianbin, Shao Yongzhen. the Way out of College English Teaching Reform. Foreign Language Circles, 1998 (04): 20-22
[9] Liu Lude. the Enlightenment of Problem-based Learning on Teaching Reform. Education Research, 2002 (2): 73-77
[10] Li Guojie, Cheng Xueqi. Big data research: a major strategic field of future science and technology and economic and social development -- Research Status and scientific thinking of big data. Journal of Chinese Academy of Sciences, 2012, 27 (6): 647-657