Research on the Application of the New Standard Model of Translation and the Collocation of the Translation Talents' Quality Combined with the Big Data Technology

HE XIANGJUN1,*, CHEN CHUJIE2

1 Hunan University of Commerce, Changsha, Hunan, China, 410205
2 Guangzhou Development and Research Center of China Construction Bank, Guangzhou, China, 510665

ABSTRACT: After twenty-first Century, with the accelerated process of globalization, the translation between different natural languages has become a major obstacle to the development of human society. At the same time, with the rapid development of computer technology and Internet technology, the application of a new type of translation standard model based on big data technology not only can improve the quality of translation, but also is helpful to the collaborative translation of translators. Through the improvement of the model of traditional machine translation system, a new language model, translation model and decoder were built. In order to further analyze the performance evaluation of the improved system based on the big data technology, the perfect evaluation system was established through the artificial intervention. Finally, through an example, the comparative analysis of the improved machine statistical translation system and the traditional statistical translation system was carried out. It can be seen that the improved translation system and method based on Internet technology can complete the translation process under artificial intervention, the translation quality is better, and the translation efficiency is higher. The improved translation system can be used for the cooperative translation of translation talent, the interactive effect is good, and the efficiency of translation of translators can be improved.

INTRODUCTION

In the human communication, the most direct mean of expression is the natural language, but because of different nationalities, countries and regions, there is a big difference between languages, so there are communication barriers in communicating with each other because of the differences in language, so that it is difficult to convey accurate information to each other. With the development of global economy and the expansion of information communication, it is necessary to take effective methods to carry on the translation between languages, so as to solve the obstacles between different languages. In traditional translation mode, language translation is usually carried out by professional translators, which is time-consuming and laborious.

With the continuous development and maturity of computer technology, machine translation can effectively improve the efficiency of translation [1]. However, this traditional machine translation technology has a big problem, its translation results are easy to make mistakes, and there is not a small gap between this and the direct translation of translators [2]. Since twenty-first Century, with the continuous development of Internet technology, the Internet's big data technology has become more and more influence on people's life. The application of a new type of translation standard model based on big data technology can improve the quality and efficiency of translation, and the combination of it with translation talents can realize the interactive machine translation [3]. Therefore, combined with the big data technology, the new reflection model can be improved, so as to establish the human computer aided translation technology, which not only can provide more satisfied results for customers in the case of manual intervention, but also can improve the
efficiency and quality of translation with the collocation of translators.

2. Research on the theory of translation model

2.1 Machine translation and corpus

Machine translation was put forward in the last century in 50s, the use of machine translation technology deepened people’s understanding of language and related knowledge, but at that time, it just carried on the simple word frequency count, word translation and sequence changes and so on [4]. After entering the twenty-first Century, based on the statistical methods of modern computer and network technology, the application of the corpus methods has been more and more, a large number of language resources and testing platform have been built, machine translation technology has made great progress. Study on machine translation in China began from the last century in 60s, and initially completed the Russian and Chinese translation machine experiment. Today, China’s machine translation technology has made tremendous development, the automatic localization, automatic reading, code conversion and other auxiliary functions have enhanced the user experience. But in many ways there is still a big gap, the accuracy of some of machine translation is not enough, it is difficult to directly use the results of translation [5].

The original statistical translation method is to use the information transmission method to carry on the machine translation, so as to establish the source channel model.

According to Bayes formula:

$$P(S|T) = \frac{P(S) \times P(T|S)}{P(T)}$$  \hspace{1cm} (1)

In the formula, $S$ is the source language sentence, $T$ is the encoded channel output sentence of the source language reduction encoding, $S$ is the maximum possible channel input sentence, the formula is also the basic equation of the statistics machine translation, $P(T|S)$ is a translation model. The framework of statistical machine translation model is shown below:

According to the practice, the value of $m$ is related to the accuracy of the dependency relation between words, the accuracy is higher, the model complexity is relatively high, therefore, it is needed to select a suitable $m$ value (usually the value is in the range of $1 \leq m \leq 7$). If the probability of a given
Research on the Application of the New Standard Model of Translation and the Collocation of the Translation Talents' Quality Combined with the Big Data Technology

**Fig. 3 MT pyramid**

word $w_j$ is related to the first two words, it can be simplified as a three element model:

$$p(w) = p(w_{j-1}) \times p(w_{j-2} | w_{j-1}) \times p(w_{j-1} | w_{j-2})$$ (5)

$$p(w_j | w_{j-1} w_{j-2}) = \frac{\text{count}(w_j w_{j-1} w_{j-2})}{\text{count}(w_{j-1} w_{j-2})}$$ (6)

Cross entropy is generally used to evaluate the quality of the language model, and sometimes the confusion test and others are also used. In the process of the training of the model, because of the lack of corpus, the data is sparse, and then the sharp distribution or zero probability phenomenon occurs, therefore, it is necessary to improve it, so as to improve the low probability, reduce the high probability, and finally realize the approximate uniform of probability distribution [10].

In the typical statistical machine models, the most famous is the Machine Translation Pyramid. According to the different levels of knowledge, the Machine Translation Pyramid combines the knowledge of the use of state, and it is divided into the intermediate language, the semantic level, the syntactic level, the phrase level and the word level from top to bottom [11]. According to Pyramid, it can be seen that the statistical machine translation models mainly include the translation model based on the phrase, the word and the syntax, according to the target language syntax and the different processing depth of source language, it also can be divided into the tree to string, the string to tree, the tree to tree of the three models [12].

Generally speaking, the translation model based on word and the translation model based on syntax translate a single word in the source language sentence into the target language, but because of the difference of the generating process, there are some differences in the adjustment of the word order of the target language [13]. If the translation is regarded as the problem of the noise channel, that is, the target language is affected by the noise channel, and the other end of the target language is the source language.

**Fig. 4 Noise channel model**

The existing translation model $p(S|T)$ only considers the translation between words. Assuming that the source language is $S = s_1 \cdots s_m$, the target language is $T = t_1 \cdots t_n$, so there are $(l+1)^m$ ways of alignment between the two, then the joint likelihood probability of $T$ and word alignment $A$ can be obtained:

$$p(S|T) = \frac{\mathcal{E}}{(l+1)^m} \prod_{i=0}^{m} p(s_i | t_i)$$ (7)

Based on the change of the position of the word translation, the probability of distortion is introduced, according to the situation of the multiple translation of a word, the reproduction probability is introduced, then there are:

$$p(S|T) = \mathcal{E} \prod_{i=0}^{m} p(s_j | t_i) p(a_j | j,m,l)$$ (8)

$$p(S|T) \approx \prod_{i=1}^{l} \phi(n_{t_i}) \prod_{j=1}^{m} p(s_j | t_{a_j}) \prod_{j=1}^{a_0} d(a_j | j,m,l)$$ (9)

The translation model based on the word regards the translation of the word as the basic unit; it lacks the consideration of context information, so it is often confused with the meaning of words. Due to the large vocabulary in the corpus, the control translation will produce a lot of translation lists, therefore, it is needed to train it to obtain the corresponding probability table, this makes the subsequent decoding process become more complex and difficult, the memory space consumption is large, so that the efficiency of the translation process is low, and the quality of the translation results is not high [14]. The translation model based on the phrase has better effect because of its successive sub sequences, which needs to experience the word segmentation, phrase division and training of different stages, but eventually it will still have certain mistakes, therefore, it is needed to reduce the noise and eliminate the noise [15].
After the completion of the machine translation system, the quality evaluation needs to be carried out, including the fluency, composition of the translation, and the amount of information of these three aspects. There are two kinds of commonly used evaluation standards, one is to carry out the machine automatic calculation of the output translated text through the establishment of the data model, so as to calculate the final score; the other is to carry out the manual evaluation for the output translated text by the translation related professionals [16]. Two different evaluation methods have different characteristics, the evaluation cost of using the machine is relatively low, the speed is high, but the accuracy is not high. Artificial evaluation methods need more time, and the cost is high, but the accuracy of evaluation results is high [17].

3. Application of a new model of translation model combined with big data technology

3.1 Design of system structure

In order to improve the accuracy of machine translation, it is needed to use the Internet technology to carry out the design of machine translation technology for human-computer mutual assistance, mainly including the language module, translation module and decoder module [18]. In addition, it is needed to establish the evaluation system of machine translation, so as to carry on the accuracy analysis of the result of the translation, and the improvement of the design system. Translation probability table has three parts; the first is to generate the phrase table, which is extracted from the phrases in the bilingual parallel corpus. The second is to regard the continuous random sequence as phrase, and regard it as a translation unit, but if the frequency of the sub phrase is the same as the longer phrase (including the phrase), then the sub phrase is used as the useful information, and the statistical analysis is used to score the rationality of phrases, so as to reduce the search space of the phrase table [19]. Finally, the results of the target text are evaluated by the manual evaluation, and a weight is obtained by the mathematical statistics, which can be used as the characteristic parameter of the translation probability to improve the quality of the target text.
upper case letters are converted to the lowercase letters, the Chinese corpus processing is the conversion of the full width and half angle of word segmentation. In addition, the punctuation mark between the words is separated by the use of the space. For the long sentences with many words, in the initial pretreatment, some words need to be removed, and the number of words is limited in a fixed range.

3.2 Improvement and optimization of the new model of translation standard

In general, the more the knowledge and rules of the language in the computer are, the higher the quality of the translation is. But the data and information contained in the natural language of human beings are very large, the data can be input as much as possible, but not all the data can be input in the computer, and even with the development of human society, the new language knowledge and language phenomenon may appear [20]. Therefore, the big data technology and manual intervention can be used to improve the credibility of the machine translation, and the quality of the translation of the text can be improved through the context of the sentence. Compared with the traditional machine translation method, the improvement of the translation model is mainly aimed at the artificial intervention to improve the quality of translation. Through the Internet, the parallel corpus is obtained, and the training data preprocessing is carried out. The weight of the translated text is calculated by using the artificial scoring results, and then the weight can be modified with the co-occurrence probability, so as to obtain the translation probability table. In order to avoid the unreasonable results of the score caused by human factors, there are two measures: One is to improve the quality of translation talents, so as to make the translation results more professional; the other is to use the normal distribution to carry on the discrete score processing, and divide the manual artificial scoring results, and then the weight can be calculated, and the probability distribution function is calculated as:

\[ F(t_i) = \int_{-\infty}^{t_i} f(t) dt \]  

(13)

The sentence weight is calculated as:

\[ w_i = \sum_{i=1}^{n} x_i P_i \]  

(14)

The word alignment model is constructed, which provides the basis for the establishment of translation model. First of all, the correspondence between the target language and the source language is established, and the translation probability is calculated, the word alignment probability is calculated as:

\[ p(d|e, f, w) = \frac{c(e, f)}{c(f)} \times w \]  

(15)

In the formula (15), \( e \) represents the source language, \( f \) represents the target language, \( c(e, f) \) is the number of appearance times of the two, \( a \) is a word alignment of \( e, f \). In the decoding process, the product of the weights and the co-occurrence probability can get the word alignment probability, and the value is used as a part of the translation probability. When the manual scoring data statistics is carried out, the size of the weight reflects the quality of the translated text, the variance directly reflects the fluctuation of the score.

Statistical machine translation system is mainly based on translation model and language model, in which, the language model directly reflects the fluency of the target language, which affects the choice of translation text. According to the causal hypothesis of the Markov chain, N-gram statistical language model is used, then:

\[ p(s) = \prod_{i=1}^{l} p(w_i|w_{i-1}) \approx \prod_{i=1}^{l} p(w_i|w_{i-1}w_{i-2}) \]  

(16)

The processing model of statistical natural language is derived from the corpus, and its parameters are estimated by using the maximum likelihood estimation method:

<table>
<thead>
<tr>
<th>T</th>
<th>t1</th>
<th>t2</th>
<th>…</th>
<th>ti</th>
<th>…</th>
<th>tn</th>
</tr>
</thead>
<tbody>
<tr>
<td>P</td>
<td>p1</td>
<td>p2</td>
<td>…</td>
<td>pi</td>
<td>…</td>
<td>pn</td>
</tr>
</tbody>
</table>
3.3 The application of the new model of translation standard and the matching of the talents' quality

The method of evaluating language model is to use the existing model to calculate the probability of test data, including the cross entropy, the perplexity and other derived tests. The perplexity \( PP(S) \) of the word is calculated as:

\[
PP_S = 2^{\frac{1}{W_s} \sum_{t=1}^{W_s} \log_2 p(w_t|w_{t-1})} = 2^{\frac{1}{W_s} \sum_{t=1}^{W_s} \log_2 p(w_t|w_{t-1})}
\]

Different smoothing methods are used to make the perplexity of the language model of test set small, and its basic constraint is:

\[
\sum_{w_i} P_{mle}(w_i|w_1, w_2, \cdots, w_{i-1}) = 1
\]

As one of the key modules of machine translation, the decoding model aims to find the most appropriate translation based on the information of the source language. In the machine learning of the bilingual corpus, the phrase translation and the target language model are obtained, and other models are added to the current model according to the current log reference model thinking. At the beginning of the decoding, the source language sentences are segmented to obtain a number of phrases, and the probability of the model is calculated, then the translation of the source language sentence phrase is expanded, the final expansion state can complete the target language sentence. There are three main information of translation state, namely, the source language, the target language information and the translation probability, the expansion of translation state includes the evaluation of the state of the current translation and the non-translation part. In which, the score of the non-translation part is the evaluation of the current status to the target state, and the status of the current phrase represents the weighted sum of the initial to the current state of the model. The specific flow chart of decoding is shown below:
4. Application analysis

4.1 Running test

According to the new translation standard model, the data was collected and used, and a statistics machine translation system based on network was built, then the system test and data analysis were carried out. The preparatory work of the bilingual parallel corpus was completed in the tested system, and the language model, translation model, decoder and automatic evaluation system were constructed. The system used Moses as the open source statistical machine translation tool, the whole system used C++ language, and the core framework was the SMT framework based on the phrase. In addition, the function of the artificial score was added to the system, and the translation professionals carried out the evaluation of the translation. Finally, the improved translation system and the traditional translation system were compared and analyzed, so as to analyze the advantages and disadvantages of the system.

Evaluation method used BLEU. Assuming that the number of the candidate target and the reference target was n, at the same time, the faithfulness and fluency of the manual evaluation were considered, the score formula was:

\[
\log \text{BLEU} = \min \left(1 - \frac{c}{r}, 0\right) + \sum_{n=1}^{N} w_n \log p_n
\]

Among them, \(c\) was the length of the candidate file, \(r\) was the length of the effective reference, \(N\) was the maximum number of \(n\) meta syntax.

The training corpus included 20 thousand Chinese-English sentences, which could comprehensively reflect the quality of the system's translation results, and could be used to train statistical translation model parameters. The characteristics of data set are shown as follows:

<table>
<thead>
<tr>
<th>Data set</th>
<th>Chinese</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Number of sentences</td>
<td>Word number</td>
</tr>
<tr>
<td>Training</td>
<td>2×104</td>
<td>180K</td>
</tr>
<tr>
<td>Test set</td>
<td>500</td>
<td>3694</td>
</tr>
<tr>
<td></td>
<td>8000</td>
<td>68K</td>
</tr>
</tbody>
</table>

Moses was used to carry on the training; the process included the data preprocessing, word alignment, word order and phrase extraction and so on. Bi directional alignment results were obtained after two-way training for bilingual corpora, the results included the sentence length information and the word target sentence, the source sentences in a sentence and the position relationship between the words and expressions. The center point was the intersection of the two directions, and then the adjacent eight points were examined, if it belonged to the union, it should be extended as the alignment point and added to the sequence alignment.

In the analysis of word order, the comparative analysis of the word order of English and Chinese of the two languages was carried out, in this regard; the translation process was divided into two parts: the revision of word order and the word segmentation and transformation. The specific translation model is shown below:

![Fig. 11 Language translation model](image)
significantly improved. The number of personals for the evaluation was more, the quality of the target was objectively reflected, the conclusion was more objective.

<table>
<thead>
<tr>
<th>Translation system</th>
<th>Number of phrases (10000)</th>
<th>BLEU-4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Traditional systems</td>
<td>140</td>
<td>0.244</td>
</tr>
<tr>
<td>Improved system</td>
<td>114</td>
<td>0.315</td>
</tr>
</tbody>
</table>

The experimental results also show that the benchmark system based on the Moses can effectively realize the mutual assistance, and the statistical translation system operates normally, and its feasibility is positive. And the application of Internet technology in translation can effectively improve the quality of translation and the efficiency of the system, compared to the traditional translation system, the translation effect is better.

5. Conclusions

Machine translation system makes full use of computer technology and Internet technology, can realize the translation among different languages; it is helpful for people to communicate with each other. However, because there is a large deviation in the sequence of the translation results of the translation model, the traditional system is difficult to be used directly. Therefore, through the improvement of statistical translation model, the efficiency of translation and the quality of translation can be improved. The improved translation system includes the preprocessing of bilingual corpus, word segmentation, phrase extraction and artificial evaluation. Through the comparison of the results of the improved translation system and the traditional translation system, it can be seen that: The improved machine translation system can operate normally, and the reliability is high; Compared with the traditional translation system, the translation quality is higher, and the translation efficiency is significantly improved; The evaluation of translation results carried out by manual work will be affected by some subjective factors, but with the increase of the number of personals, the accuracy of the evaluation of the quality of the target text is higher. Therefore, this model can be applied to the improvement and perfection of the statistical Machine Translation system, and can be applied to the specific translation work.

7. REFERENCES

9. Helz G R, Bura -