Research on the Statistical Machine Translation based on Neural Network Learning

Xueling Wang
School of Foreign Languages, Xinxiang University, Xinxiang 453003, China

Abstract

In order to solve the problem of word alignment in foreign language translation, the principle of neural network learning model, parameters estimation and feature selection algorithm are investigated. Neural network learning model is an effective method to set up statistical language model, which has the strong knowledge representation ability. The training speed of neural network learning method is very slow, thus it consumes a lot of resources. Because the feature space is commonly very large, how to choose typical and less redundant characteristics for model training is very important. Then an improved feature selection algorithm is put forward. Word alignment system based on neural network learning model is established, which includes text pre-processing, model training, name entity recognition and part of speech tagging. At last, experiments are done to test its performance. The results show that the proposed system has high accuracy.

Keywords: Statistical machine translation, Neural network learning, Word alignment, Parameters estimation, Feature selection algorithm.

1. INTRODUCTION

After the reform and opening up, as China's comprehensive national strength continuously improve, China also gradually in all aspects to the internationalization development. Therefore, the national English level also in rapid ascension. However, learning English for Chinese popular has adapted to the traditional education and many don't adapt with difficulty. After continuous reform, level of education in China is rapidly improve, however, due to China's education form for English education effect is not good, cause most of the college students' English study becomes the "dumb English", so it is hard to college students' English into practical English. One important reason is that acquisition of translation knowledge is more difficult. Acquiring translation knowledge by artificial way has limited coverage and high cost. Therefore, mining bilingual knowledge automatically from the massive real bilingual corpus has become a very important way to acquire translation knowledge and one of the key technologies is bilingual alignment technology (Nakov and Ng, 2014). Bilingual corpus is also called parallel corpus, which contains comparative translation information between two languages and is very important translation knowledge source. To obtain the knowledge from the bilingual corpora, bilingual alignment is a key link and accuracy of the knowledge directly depends on the quality and size of the bilingual alignment, so bilingual alignment is very important technology for automatic extraction of bilingual dictionary, machine translation based on statistics and machine translation based on the instance is very important technology (Bloodgood and Callison-Burch, 2014).

Bilingual alignment is also the fundamental technology of cross-language corpus retrieval system. In addition, translation memory storage technology based on bilingual alignment is widely used in the current auxiliary machine translation system. So bilingual alignment has very important value for translation reuse, and is very important for practical application of machine translation. Parallel corpus contains much bilingual lexical information. The task of word alignment is to identify word translation or corresponding relationship from the bilingual texts automatically (Wolk and Marasek, 2015). It sets up alignment relation for bilingual text at vocabulary level, which determines which word of the source language text has corresponding relationship with the word in the target language. Word alignment is usually conducted on sentence aligned bilingual corpus. The alignment of word is more difficult than sentence alignment. Because the word alignment is not only has complicated corresponding relationship such as one-to-one, one-to-many, many-to-one and many-to-many. Certain words could not find any corresponding relation, and there is
no restriction relation on the order. There are a lot of cross alignment phenomenon, and even one word may is related to multiple discrete words. Therefore, the word alignment is a challenging research topic in the field of natural language processing (Bertoldi et al., 2014).

Based on neural network learning approach, word alignment was investigated (Allauzen et al., 2014). Firstly, it automatically identified the English naming entities. In order to avoid the error propagation, it directly extracted candidate Chinese naming entities on the Chinese text without Chinese word segmentation and labelling. Then the neural network learning model was used to calculate alignment. Boost strapping method was used to train the neural network learning model. The current mainstream method can be divided into two categories, heuristic method based on hypothesis testing and the method based on statistical translation model. Heuristic method is simple. Through the analysis of the word correlation in the source language and target language, word alignment is obtained, which is often used in the actual work. Heuristic method is easy to implement and understand. But its shortcoming is also very obvious. Form of specific correlation function is selected, and some functions include parameters which are adjusted by experience, thus the quality of this alignment method is poor. Practice shows that the heuristic alignment model is worse than statistical translation model. Unsupervised word alignment with arbitrary features was investigated by Chris Dyer (Durrani et al., 2015). For statistical machine translation, discriminative weighted alignment matrices were put forward by Tomeh N (Kenny et al, 2014). A discriminative matching algorithm used for word alignment was also proposed by Taskar B (Zhao et al., 2014).

Word alignment technology based on statistical translation model has perfect theory and the determination of parameters has scientific statistical basis. It has good alignment performance. The defect is that it needs to compute a huge number of model parameters (Kenny and Dorothy, 2014; Bojar et al., 2014). The computational complexity is high. Another problem is that dual language is commonly used for parameter estimation (Pecina et al., 2014; Labaka et al., 2014). Besides, neural network was also used in translation model (Banerjee et al., 2015; Lu et al., 2014; Clark et al., 2014). At present, word alignment technology is not mature. Accuracy and recall rate remains to be further improved (Costajussà, 2015). Accuracy rate and recall rate has large volatility when corpus and its size is different (Farzi and Faili, 2015). Most of the data are acquired from the corpus of western languages. In the next section, principle of neural network learning and feature selection is investigated. In section 3, word alignment model based on neural network learning is put forward. In section 4, experiments are done to test the performance of proposed word alignment model. At last, we give some remarks.

2. MATERIALS AND METHODS

Supposing the output of a random process is y. When y is generated, this random process may be influenced by some relevant context information, yeY, xeX. X and Y are finite sets. Our task is to construct a statistical model to express this random process accurately. When context x is given, the condition probability p(y/x) of output y should be estimated. A collection of all condition probability distribution is P. Some training samples are selected. That is (x1, y1),(x2, y2),...(xN, yN). The empirical probability distribution is

\[ p(x, y) = \frac{1}{N} \cdot \text{the number of occurrences of (x, y)} \]  \hspace{1cm} (1)

Firstly, the characteristic function is introduced. For example, the probability of the word make being translated into meaning of write is small. If the word make is followed with treaty, the probability of the word make being translated into meaning of write is big. In order to express this event, following characteristic function can be used.

\[ f(x, y) = \begin{cases} 1, & \text{if y="write" and "tready" follows} \\ 0, & \text{otherwise} \end{cases} \]  \hspace{1cm} (2)

It is a binary function, event is the \( Y \times X \) space is mapped into (0, 1) space. For any characteristic function \( f \), experience expectation on training sample is
\[ E_p f_i = \sum_{x,y} \tilde{p}(x, y) f_i(x, y) \] (3)

The expectation of characteristic function \( f_i \) is

\[ E_p f_i = \sum_{x,y} \tilde{p}(x) p(y / x) f_i(x, y) \] (4)

\( \tilde{p}(x) \) represents experience marginal distribution of \( x \) in the training sample. The expectation value calculated by model should be consistent with experience expectation value.

\[ C = \{ p / E_p f_i = E_{p^*} f_i, i \in [1, 2, ..., K] \} \] (5)

\( C \) represents a series of probability distribution. The core idea of neural network learning is to choose the model with largest entropy in these models. In all probability distribution, \( p^* \) is selected, which meets the following equation.

\[ H(p) = -\sum_{x,y} \tilde{p}(x) p(y / x) \log p(y / x) \] (6)

\[ p^* = \arg \max_{p \in C} H(p) \] (7)

\( H(p) \) represents condition entropy used to represent evenness of condition probability \( p(y/x) \). This is an optimization problem, the introduction of Lagrange operator makes us get the form of solution.

\[ p^*(y / x) = \frac{1}{Z(x)} \exp \left( \sum_{i=1}^{K} \lambda_i f_i(x, y) \right) \] (8)

\[ Z(x) = \sum_{x} \exp \left( \sum_{i=1}^{K} \lambda_i f_i(x, y) \right) \] (9)

This task is converted to solve optimal solution of \( \lambda_i, i = 1, 2, ..., K \). It can be solved by generalized iterative scaling algorithm. For any \((x, y) \in X \times Y\), the sum of characteristic function is

\[ \sum_{i=1}^{K} f_i(x, y) = C, \lambda_i^{(0)} = 1 \] (10)

\[ \lambda_i^{(n+1)} = \lambda_i^{(n)} \left[ \frac{E_p f_i}{E_{p^*} f_i} \right]^{\frac{1}{c}} \] (11)

\[ E_{p^*} f_i = \sum_{x,y} \tilde{p}(x) p^{(n)} (x / y) f_i(x, y) \] (12)
\[ p^{(n)}(x \mid y) = \frac{1}{Z(x)} \exp \left( \sum_{i=1}^{n} \lambda^{(n)}_i f_i(x, y) \right) \]  \hspace{1cm} (13)  

Log-likelihood value of \( \tilde{p} \) is

\[ L(p) = \sum_{x,y} \tilde{p}(x, y) \log p(y \mid x) \]  \hspace{1cm} (14)  

\[ L(p^{(n+1)}) \geq L(p^{(n)}), \lim_{n \to \infty} p^{(n)} = p^* \]  \hspace{1cm} (15)  

Computational linguists Zhiwei Feng in the present situation and problems of the machine translation [5] a pyramid model of the development of the mentioned machine translation, the pyramid describes the developing course of statistical machine translation, at the bottom of the pyramid model is the most simple, easy to implement, the pyramid is upward development, need of the more complicated model. The ultimate goal of statistical machine translation is to realize the automatic translation based on intermediate language. The current statistical machine translation model and a lot of research focused on syntax and phrase between models, and on the study of semantics and intermediate language is still in the exploration phase. Statistical machine translation in the source language word pyramid as shown below, as shown in figure 1:

![Figure 1. The pyramid of statistical machine translation](image)

Statistical machine translation model, roughly experienced models based on word, phrase based model, based on three stages of syntactic model. Translation model based on word in the context of information too little, the translation effect is not good, but it is the foundation of the other two models; Based on model scalability of the phrase is very good, as long as there is sufficient corpus, it can get very good translation model, but this kind of model is not usually a good deal with the order; Based on syntactic model is usually the result of the translation of sequence has good guidance, but because of its reliance on syntax analyser, scalability than a model based on the phrase. Researchers hope translation model can integrated the advantages of the above two kinds of models, therefore proposed the BTG grammar.

Condition neural network learning method is a kind of supervised machine learning methods, so each training sample is composed of one instance \( x \) and its target concept class \( y \). For the aligned Chinese word string \( c_1^I = c_1, K, c_j, K, c_j \) and objective language(English) \( f_1^I = j_1, K, j_i, K, j_f \).
We must give a specified aligned target word to each source word, and only one target word is specified. Each source can be aligned with any word in the target language, rather than just those words in the target list. \( p(y/x) \) represents the probability of outputting \( y \), when context \( x \) is given.

\[
p(y/x) = \frac{1}{Z(x)} \exp\left(\sum_{i=1}^{K} \lambda_i f_i(x, y)\right) \quad (17)
\]

Let \( f_i(x, y) \) represents feature function, \( K \) represents the number of feature, \( Z(x) \) represents standardized factor, which makes \( \sum_{y} p(y/x) = 1 \).

\( \lambda_i \) is weight parameter of feature function \( f_i(x, y) \). The optimal parameter \( \lambda_i \) can be worked out using generalized iterative scaling algorithm. The word in objective string is output, which is \( j_1, I \).

The problem of feature selection is as follows. In the candidate feature set \( F \), only a small part of subset can be used in the model. \( C(S) \) represented model determined by effective feature set \( S \). Only some part of \( C(S) \) meets the equation \( p(f) = p(f) \). The model set space is \( C(S \cup f) \).

\[
C(S \cup f) = \{ p \in P \mid p(f) = p(f), f \in S \cup \overline{f} \}. \quad (19)
\]

The optimization model is

\[
P_{S \cup f} = \arg \max_{p \in C(S \cup f)} H(p) \quad (20)
\]

Incremental feature selection algorithm has bad performance when it is used in feature selection. So an improved feature selection algorithm based on incremental feature selection is put forward. \((X, Y)\) is an event space of language text, \( y \in Y \) represents an event of current word and \( x \in X \) represents context information feature. In the sample set, candidate feature set \( F \) of each \( y \) is calculated. \( F = \{F_1, F_2, K, F_N\} \) represents feature candidate set of the whole sample space. The appearance times of \( x \) in \( y \) is \( cont(x) \). If \( cont(x) \) is bigger than a determinative value \( c_n \), it is added into effective feature set \( F \). If \( cont(x) \) is smaller than a determinative value \( c_n \), the approximate gain calculation is carried out. In each feature selection step, keep all the candidate features ordered from big to small according to gain. Every time the character ranked first is checked. If the gain of this characteristic belongs to the current step, this feature is the one with the biggest gain. This characteristic is selected and the current feature selection step is over. It steps into the next feature selection. Otherwise, calculate the gain of current step to replace the original value. Then the gain is inserted it into the characteristics column and the feature in the first order is checked. The effective characteristics set \( F \) are obtained.

\[
p^g_{S,f} = \frac{1}{Z(x)} p_S(y, x) e^{\alpha f(x,y)} \quad (21)
\]
\[ G_{S,f}(\alpha) = L(p_{S,f}^\alpha) - L(p_S^\alpha) \]
\[ = -\sum_{x} \overline{p}(x) \log Z_{\alpha}(x) + \alpha \overline{p}(f) \]  

(22)

According to derivative of \( G_{S,f}(\alpha) \), an equation is obtained. Newton method is used to solve this equation and \( \alpha \) is figured out. \( \alpha \) is substituted into equation \( G_{S,f}(\alpha) \) to get gain of each feature. Then feature with the biggest feature gain is obtained. The process of feature selection is as follows.

The input is feature candidate set \( F_c = \{ x_1, x_2, K, x_k \} \) and feature appearance frequency threshold \( cn \). The output is effective feature set \( F \).

Step1. \( F \) is initialized to be zero.

Step2. Calculate gain of the entire candidate feature and arrange the gain from small too big.

Step3. For each \( y_j \in Y, j=1,2,K,N \). Calculate each candidate feature \( x_i \in F_c, i=1,2,K,n \). If \( \text{cont}(x_i) \geq cn \), equation \( F = F \cup (x_i, y_j) \) and \( F_c = F_c - x_i \) is established. Otherwise, turn to step 4.

Step4. The approximate gain calculation is carried out. If this gain is the biggest at present, \( F = F \cup (x_i, y_j) \) and \( F_c = F_c - x_i \). Otherwise, this gain value is inserted into gain queue, gain of candidate feature is ordered according to the order from big too small.

Step5. If the algorithm meets the termination condition, the algorithm stops. Otherwise, turn to step 2.

Similarly, Ochs presents an alignment template model, the model is divided into levels phrase alignment and word alignment, the difference is Ochs will align phrase alignment of generalization based on parts of speech of the template, and using the linear logarithmic model framework as a whole. Koehn in drab phrase model on the basis of considering the adjustable factors put forward a phrase translation model based on word alignment. And according to the model implements the translation system based on phrases Moses Pharaoh and upgrade version.

Moses had become a contrast experiment was carried out in the current statistical machine translation of one of the most popular baseline system. Marco et al., the joint probability instead of conditional probability, the model was proposed based on joint probability of the phrase. This model is similar to the model of IBM: the different is that it USES a combined strategy to combine the phrase. Statistical translation model based on syntactic research have also made great progress in recent years, Wooldridge presented the first bilingual syntactic parsing grammar BTG and ITG, and the thoughts of the two methods from the context-free grammar, context-free grammar used in single language syntax analysis, and BTG and ITG grammar will this idea to promote to the bilingual.

BTG and ITG grammar for the first time will be the process of statistical machine translation in the form of syntactic analysis, is just the end of the two grammatical operators are words. DeyiXiong made the BTG grammar model extended to phrase and put forward a kind of syntactic level limit, reward continuous phrases translation. Wei Jiang inspired by BTG grammar puts forward a model of hierarchical phrase, phrases hierarchy model used is a bilingual
context-free grammar probability of synchronization. On the rules of a longer phrase, using a placeholder X instead of shorter translate phrases on both ends of the rules, to generate the translation templates. By a bonding rules code translation as a result, to ensure that the entire translation process can generate a complete analysis of bilingual syntactic tree.

Yamata tree-string model is proposed, the model is the earliest transcription tree model, using the syntax tree of the source language and target language sentence for training. His translation process can be divided into three steps: the sequence, inserted, translation. First use of the source language syntax tree in the sequence and then on the syntax tree node insert some may be ignored the utterances of target languages, finally use the target language words to replace the source language words syntax tree. Yang Liu in addition kind of tree - string model is put forward, in the process of model training using the source language syntax tree, and through to remove certain syntactic nodes to translation model generalization. By iterating through the source language in the process of decoding, the syntactic nodes using the rules of syntax tree cover in the translation table to get result of translation.

Galley put forward a new string-tree model, based on his model can describe the layers of tree structure transformation rules. Galley method of extraction and the source language phrase, the target language structure tree and the alignment between them to maintain consistent minimum rules, they cannot be broken up into other rules. Galley further put forward the method of complex rules extraction by minimum rules and found that the compound rules to promote the performance of the system has a lot of help.

To some extent, can put the minimum rules for the word analogy, analogy and composite rules for phrases, based on the phrase of SMT, which are better than those based on words, because the phrase bound more context information, such as local word choice and phrase, the word order of compound rules relative to the minimum rules, also contains more context information, which is a major cause of composite rules of translating system performance.

Macro presents a string-tree model, an application of actual transcription and tree model, training to use the source language and target language syntax tree extract rules, when decoding by producing in the target language syntax tree rules cohesion was adjusted order. Graehl gave a detailed description to the tree transcription model framework, and put forward a set of general training series-tree model and the method of tree - a tree model. M waves proposed a more complex forest model. The basic idea is to source language sentence syntactic analysis of n - best results for compression forest, on the basis of bilingual word alignment from the forest of the source language syntax and sentence on the extraction of the target language translation rules. Decoding when first produced in the source language sentence n-best syntactic analysis, and then iterate through all the syntactic nodes, using the rule table to cover other way to the tree, the last in the forest in search of the highest scoring compression results. Forest model on syntactic nodes only when extracting rules extraction, Hui Zhang extends the model that the rule extraction stage allows to extract multiple syntactic tree node sequence (tree-sequence).

3. RESULTS AND DISCUSSION

The text is preprocessed. Special markers are used to divide the input text into short Chinese character string. These markers include punctuation marks, numbers, letters and other non-Chinese characters. Then feature is selected and generalized iterative scaling algorithm is used to calculate optimal parameters. Here, name entity recognition is joined. Proper nouns and meaningful phrases are classified. The organization name, person name, place name, proper nouns, time expression and quantity expression are recognized. In addition to rely on the word itself, the dependence on the part of speech is increased. The part of speech of current word and part of speech of before and after the current word are taken as features integrated into the neural network learning model. In the same way, we can put the part of speech information related to target word as characteristics. Part of speech information is automatically labeled by Chinese and English part of speech tagger respectively. After analysis, we can see that the part of speech of the target language has great influence on improvement of the test precision. If more context information is used, a higher precision may be obtained. On the contrary, part of speech of the source language has little effect and it even has a negative effect sometimes. That is to say, context information of the target word is more important than context information of source word. The best accuracy of word alignment based on neural network learning is 89.6%. It can be seen that the proposed word alignment algorithm has good performance.
The Bayes error and model error is automatic translation problems inherent errors, it is difficult to improve. This article obtains the mistakes from the training; analysis of statistical machine translation model construction process, each link may be introduced by the error, and discusses some effective ways to reduce these errors. In the phrase based statistical machine translation model training process as shown in figure 2.

Model training process, first of all, starting from the bilingual sentence alignment corpora, after word alignment, phrase translation extraction rules, several steps such as short tone sequence rules extraction in bilingual translation model, at the same time in the target language training in monolingual corpora language model, at last, through training methods to adjust model parameters in the model the weight of each feature.

Experiment by alignment of this section is to filter the confusion degree of training corpus, get confused degree of five-sixteen respectively to filter the training corpus; the filtering result is shown in figure 3. Notes: Figure x axis as the corpus in the filter when the confusion degree threshold, the y axis for filtering after training corpus the rest of the sentence on the quantity. It can be seen when confused degree threshold value is less than 12 training corpora began to decrease. In the actual training corpus, some words of translation itself is very good, but because it is free translation or a part of Chinese word segmentation is done is not good enough, affected the word alignment. In other words, the aligned due from chapter and extracted automatically, only half sentence is aligned to each other. Although these words to the whole word alignment do not good enough, but for some individual words, still can find a good alignment, so the words of the confusion degree is often greater than 12.

To test corpus filtering method based on degree of confusion on the result of automatic translation, with filtering of the proceeds of the 12 training corpus structure phrase model and sequence model respectively. Experiment on 12 phrase model and the sequence model of NIST 2005, 2006, 2008 Chinese-English translation corpora were tested, the result is shown in figure 4.

![Figure 2. The training process of phrase-based SMT models](image)

![Figure 3. Filtering the training corpus by alignment perplexity](image)
Experiment proves that training corpus alignment errors in the words of is it going to affect the results of the automatic translation. In confusion degree of training corpus filtration experiment, dramatically reduce the training corpus is started in the confusion degree threshold value is less than 12, in get BLEU - 4 highest score on the test set is a range of degree of confusion for 7-8. This suggests that the alignment degree of confusion between 8 and 12 aligned about 400000 words of no help to the improvement of the performance of the translation system, not only will have the opposite effect. And although alignment confusion degree under 7 to is a good translation of a sentence to, but the clauses of proportion in the whole training corpus is too small, not enough to higher effect of automatic translation. Experiments prove that degree of confusion between 7 and 8 alignment sentence is a critical point, for just to join in the training corpus alignment confusion degree under 7 other pair, large may improve the performance of the translation system; On the contrary, to join in the training corpus alignment confusion degree in eight or more other pair, most likely will cause adverse effect to the translation of system performance.

4. CONCLUSION

In order to solve the problem of word alignment in foreign language translation, the basic principle of neural network learning model is investigated, including the parameter estimation and feature selection algorithms. Based on the analysis and comparison of feature selection algorithm, we put forward the improved method. Word alignment system based on neural network learning model is put forward. The system includes text pre-processing, model training, name entity recognition, part-of-speech tagging and other function modules. Finally the performance of the system is verified by experiment. It has achieved high word alignment accuracy.

REFERENCES