A Method to Improve Alignment Performance of Uygur and Chinese Words through Corpus Filtering

Qinqin Li\textsuperscript{a,b} and Xiaodong Yan\textsuperscript{a,b,1}

\textsuperscript{a} (College of Information Engineering, Minzu University of China, Beijing 100081)
\textsuperscript{b} (National Center for Language Resources Monitoring and Research Center for Ethnic Minorities, Beijing 100081)
Email:1595969552@qq.com

Abstract. Word alignment is the most important part of the statistical machine translation system. The translation model and the ordinal model in statistical machine translation are constructed on the basis of word alignment results, and the errors in the word alignment stage will continue to these models. In the model, even larger mistakes may be caused in these models due to word alignment errors. The research of word alignment technology provides basic construction for corpus construction, speech recognition, bilingual dictionary compilation and information retrieval in the field of natural language processing. However, the research on word alignment technology in Chinese-Uighur is relatively late. We mainly study the alignment of Chinese and Uighur words based on the sentence level and apply the bilingual corpus filtering method based on the degree of alignment perplexity to the alignment of Chinese-Uighur words. The experimental results show that the method is feasible and has achieved the expected results in the primary stage, and the method provides a certain basis for the follow-up research of word alignment techniques.

1. Introduction

With the rapid increase of the amount of information in the world, the increasing frequency of international exchanges, and the rapid spread and development of computer network technology, language-related communication barriers become more and more obvious and serious. As a result, the potential demand for machine translation is also increasing. As the name suggests, machine translation refers to the process by which computers learn how to accurately and completely transform a natural language into another natural language. It involves many disciplines such as linguistics, mathematics, and computer science, and is a typical multi-edge intersection discipline. Machine translation needs to combine these different disciplines and conduct comprehensive research [1].

A key aspect of acquiring knowledge from a bilingual corpus is bilingual alignment. Bilingual alignment has different levels of alignment: text alignment, paragraph alignment, sentence alignment, and word alignment. The word alignment plays an important role in fully exploiting the corpus and provides the basis for natural language processing such as speech recognition, bilingual dictionary compiling and information retrieval. Therefore, bilingual word alignment research is very valuable for machine translation, and it is also very important in the practical application of machine translation as a whole [2].

In 1990, IBM's P. F. Brown and others published the paper "Statistical Machine Translation Methodology" in the journal "Computational Linguistics", and in 1993 published the paper...
"Mathematics of Statistical Machine Translation: Parameter Estimation". Introduced in statistical machine translation. After that, it attracted the attention of researchers, and there have been some improvements based on the source channel model: HMM-based word alignment model, structure-based word alignment model and so on.

In 1991, Gale analyzed the statistical word alignment method of Brown et al. on the one hand; it needed to consume a large amount of system memory while training parameters. On the other hand, he was hesitant to judge whether parameter training was robust. Therefore, Gale et al. proposed to calculate a $\Phi^2$ quantity similar to the $X^2$ statistic by using a simultaneous list of the frequency of word pairs in a bilingual corpus, and determine the strength of the correlation between word pairs based on the $\Phi^2$ value [3].

In 1993, according to the characteristics of the homologous words, Church aligned the words on the noisy text and proposed a character-based alignment method char-align, which mainly works in the cross-translation texts with congnate words. The strings of the homologous words are composed of many similar characters, and the matching words are obtained by searching the aligned paths of the translations on the characters [4].

In 1997, Ker analyzed the methods of using statistical methods from Brown and Gale et al. to obtain bilingual word pairs from bilingual corpus and found that these methods can only obtain limited vocabulary coverage even if they are trained in large bilingual corpus rate. Based on the above shortcomings, Ker proposed a class-based word alignment method that uses two synonym dictionaries for word classification in two languages to use a sentence-sized corpus on a small scale (25,000 sentence pairs). The method of greedy learning automatically extracts the corresponding rules of the synonym classes (the synonym classes of the source language and the synonym classes of the target language), and then uses the information of the corresponding rules of the synonym classes of statistics in the corpus for word alignment, and in order to obtain higher With the coverage, they also introduced additional knowledge resources - bilingual dictionaries [5].

In 2000, Huang et al. were inspired by Ker's use of linguistic knowledge in word alignment methods to study the application of word similarity measurement methods based on linguistic knowledge in word alignment, such as the comparison of word parts of speech, similarity of parts of speech The internal structure of vocabulary, syntactic comparison, etc.[9]

However, the Uyghur language started relatively late in the field of natural language processing. Up to now, the research on the alignment technique of Uyghur and Chinese words, Xinjiang University Liu Jianming used IBM model to develop the research on the alignment of Uyghur and Chinese words [8]. Based on GIZA++, the University of Xinjiang University's Ma Hee-Hui proposed a selective separation-rejection scheme to contribute to improving the alignment of Uyghur and Chinese words [9]. Li Ping of Xinjiang Normal University proposed the alignment study of Chinese and Uyghur words based on HMM and dictionaries [10]; Tan Xun of Xinjiang University proposed the clustering of words in Uighur sentences and the study of the alignment of Chinese and Uyghur words [11]. Nowadays, research on Uyghur natural language processing is progressing steadily and the word alignment technology of Uyugr and Han seems particularly important. Therefore, the research on the Alignment Technique of Chinese and Uyghurs has certain exploration space and research value in the field of natural language processing.

In this paper, according to the characteristics of IBM word alignment model 3, a bilingual corpus filtering method based on the degree of alignment perplexity is used to filter bilingual expectation and reduce the error sentences in bilingual corpus, thereby improving the performance of bilingual word alignment.

2. Word Alignment Based on IBM Model

The study of the word alignment method should first mention the research of word alignment based on the noise source channel model used by Brown's most influential IBM. [9] The basic principle of the statistical machine translation method based on the noise source channel model can be described as a decoding process. The translation system is regarded as a source channel. For a target language string (sentence) V, find the largest possible source language. String (sentence) C. According to Bayesian formula:
\[ P(C|V) = \frac{P(V|C)P(C)}{P(V)} \]  
(1)

Since the right side of equation (1) has no relation between \( P(V) \) and \( C \), finding the maximum value of the above equation is equivalent to finding the maximum value of the right side of the equation:

\[ C = \operatorname{arg\,max} P(C)*P(V|C) \]  
(2)

Brown called formula (2) as the basic equation for statistical machine translation, where \( P(C) \) is a language model that reflects the fluency of sentences; \( P(V|C) \) is a translation model that reflects the fidelity of sentences. Word alignment is the "implicit function" variable introduced by the translation model to implement the translation process. Among them IBM Model1-5 can be regarded as the representative of translation model.

### 2.1. IBM Model 1

Suppose \( C \) is the source language sentence (Uyghur sentences), \( V \) is the sentence of the \( C \) target language (Chinese sentence), if \( c \) is the word in \( C \), \( C = (c_1, c_2, \ldots, c_m) \), say \( v \) is \( c \) translate Words in \( V \), \( V=(v_1, v_2, \ldots, v_n) \). The most important parameter of the IBM model is the word translation probability (Corpus Translation Probabilities) derived from the Bayesian formula to obtain the Uyghur word translation probability \( P(C|V) \):

\[ P(C|V) = \frac{P(V|C)P(C)}{P(V)} \]  
(2.1)

To get a \( C \) such that the maximum \( P(C|V) \) value can be expressed using the \( \operatorname{arg\,max}c \) function (decoding algorithm), denoted as \( C^* \):

\[ C^* = \operatorname{arg\,max} P(C|V) \]
\[ = \operatorname{arg\,max} \frac{P(V|C)P(C)}{P(V)} \]  
(2.2)

\( P(V) \) is independent of \( C \), so \( P(V) \) can be omitted during maximization. In this way, you get:

\[ C^* = \arg \max cP(C|V) \]
\[ = \arg \max cP(V|C)P(C) \]  
(2.3)

Use knowleEquation (2.4) is called Fundamental Equation of SMT, \( P(V|C) \) is called Translation Model, and \( P(C) \) is Language Model.

To calculate the translation model \( P(V|C) \), we need to introduce the alignment function \( a(i,j) \):

\[ P(V|C) = \sum_{a} P(V, a | C) \]
\[ = \frac{\varepsilon}{(m+1)^{n}} \prod_{j=1}^{n} \sum_{i=1}^{m} t(v_j | C_{a_i}) \]  
(2.4)

Among them, the parameter is a constant, and \( t(v|c) \) is the translation probability distribution, which gives the probability that each Chinese word is translated into any one Uyghur word, and this probability satisfies the normalization condition.

The calculation of language model \( P(C) \) is based on n-grams: the most widely used and most effective language model.

For \( n=3 \) (Tri-gram):

\[ P(C) = P(C_1C_2\cdots C_n) \]
\[ = P(C_1) \times P(C_2|C_1) \times \prod_{i=3}^{n} P(C_i|C_{i-2}C_{i-1}) \]  
(2.5)
2.2. IBM Model 2
IBM model 1 aligns each Uyghur word \( V=(v_1,v_2,\ldots,v_n) \) in Chinese and Uyghur pairs with the given Chinese word \( C=(c_1,c_2,\ldots,c_n) \). The probability is considered as equal, and the position of the word in the sentence is ignored. IBM model 2 takes into account the different positions of the sentences and the effect of different sentence pairs on the length, which may lead to different probability of alignment between any two words C and V. Therefore, IBM model 2 needs to introduce the concept of alignment probability. Marked as \( P(a_{i,j,n,m}) \), the normalized conditions are: \( \sum_{i=0}^{n} P(a_{i,j,n,m}) = 1 \). C, and \( V_j \) translate each other (or partially translate), \( a_i = i \).

2.3. IBM Model 3
In IBM model 1-2, only the one-to-one alignment of words in Chinese sentences and dimensional sentences are considered. However, in actual translation, words in Chinese sentences and words in dimensional sentences are in many cases. There is not only a one-to-one alignment of words. As a result, IBM Model 3 introduces an output rate based on the first two models. The goal is to solve one-to-many word alignment, which is a one-to-many (not necessarily continuous) relationship between Uyghur words and Chinese words. The quantitative relationship.

Defects of IBM Model 3:
- IBM Model 3 did not consider the "many-to-one" situation and caused the model to be incomplete.
- If a model has this characteristic: it cannot be able to focus all the probabilities on the events of interest.

It can be said that this model is not sufficient or is flawed.

3. Training Corpus Filtering Method Based on Confusion Degree
The corpus filtering method can remove the double-sentences with serious errors in the word alignment phase, thus improving word alignment performance. Experiments show that the corpus-based corpus filtering method has achieved good results in the research of word alignment in Chinese and English.[12]

3.1. Monolingual Statement Confusion
In statistical machine translation, an n-gram language model is usually introduced to determine whether the automatically generated translation result is a reasonable target language sentence. Assuming a Chinese sentence \( e = e_1 e_2 \cdots e_K \) composed of \( K \) words, the score of the English sentence \( e \) is calculated on the N-gram language model LM as formula (3.1):

\[
P(e) = \prod_{i=1}^{K} P_{LM}(e_i | e_{i-N+1} \cdots e_{i-1}) \quad (3.1)
\]

For sentences of the same length, it is clear that the greater the probability \( P(e) \), the more "reasonable" the sentence. However, since the n-gram probability in the n-gram language model is less than 1, the longer the sentence, the smaller the probability. Perplexity is proposed to evaluate whether a sentence of any length is reasonable. The starting point of this concept is very simple, that is to calculate the geometric mean of the probability of all the N pieces in the sentence, and then take the reciprocal number, the specific formula is as follows:

\[
PP(e) = \left( \prod_{i=1}^{K} P_{LM}(e_i | e_{i-N+1} \cdots e_{i-1}) \right)^{-\frac{1}{K}} \quad (3.2)
\]

As shown in formula (2.2), if the probability of the N-gram fragments in \( e \) is relatively high, then \( e \) is a more reasonable sentence; and the calculation formula of the confusion degree is to find the reciprocal of these probabilities, so the more reasonable the sentence, the smaller the confusion. In general, the probability of N-pieces in the n-gram language model is very small. In order to improve the accuracy
of calculation, the logarithm of the degree of confusion is mostly used in practical applications. The formula is as follows:

$$\log PP(e) = -\frac{1}{K} \sum_{i=1}^{K} \log \left( P_{LM}(e_{i-N+1}, \ldots, e_{i-1}) \right)$$

(3.3)

From equations (3.2) and (3.3), it can be seen that the size of the confusion degree has nothing to do with the length of the sentence, since it only takes the average of the probabilities of all the N-gram fragments in e.

3.2. Alignment Perplexity of Double Statement Pairs

In statistical machine translation, word alignment information is the most important and basic work in clause-aligned corpora. All unsupervised word alignment algorithms break the sentence alignment probabilities into word alignment probabilities and some product of the word location related probabilities. The current generic IBM word alignment model 1-5 is a kind of one-way word alignment model. Such models make such an assumption: In the alignment process from the target language $f = f_1 \cdots f_m$ to the source language $e = e_1 \cdots e_n$, each target language word $f$ can only be mapped once to the source language, there are no more than two source language words aligned to the same target language word. Therefore, when aligning a one-way word from the target language to the source language, a word alignment of $m$ is generated. When the one-way word alignment from the source language to the target language is relatively performed, a total of $n$ word alignments are generated.

When the word alignment is generated, the word alignment probability table on the entire corpus will be generated together. For a monolingual sentence, the n-gram language model can be used to estimate its degree of confusion. Similarly, for a double-sentence pair containing word alignment information, the word alignment probability can be used to estimate the degree of confusion. The specific formula is:

$$\log PP(f \| e) = -\frac{1}{m} \sum_{i=1}^{m} \log \left( P(e_i | f_i) \right)$$

(3.4)

$$\log PP(e \| f) = -\frac{1}{n} \sum_{i=1}^{n} \log \left( P(f_i | e_i) \right)$$

(3.5)

The reason why equations (2.4) and (2.5) use reverse word alignment is because the IBM word alignment model is essentially a noise channel model. When estimating the alignment probability $P(f|e)$ of the target language sentence $f$ to the source language sentence $e$, a word alignment probability table ($P(f|e)$) is generated in the source language word $e$ as a condition, and in this direction the word alignment on the same source language word $e$ may correspond to multiple target language words $f$, and may not correspond to any word. When calculating the perplexity $PP(f|e)$, each word in $f$ is assigned an alignment with probability $P(e|f)$. The geometric mean of these word alignment probabilities is the target language sentence $f$ Alignment to the degree of confusion on the source language sentence $e$. Like the perplexity of a monolingual sentence, the perplexity of a double-sentence pair is also not related to the length of the sentence. If the training corpus has completed word alignment in both directions and has obtained a conditional probability table with word alignment in both directions, the alignment perplexity for the double-sentence pair $(f, e)$ is $(\log PP(f|e) + \log PP(e|f))/2$. The greater the value of this alignment perplexity, the more “unlike” the alignment sentence pair is. By setting the alignment perplexity threshold, the double sentences in the training corpus with the alignment perplexity greater than this threshold can be completely filtered out.

4. Experiment and Result Analysis

This thesis is divided into two parts. The first part uses IBM word-alignment model 3 to achieve the alignment of Chinese and Uyghur words; the second part uses the double-sentence pair's word-alignment perplexity to filter the training corpus. After filtering, the traditional IBM word alignment model is used to achieve the alignment of Uyghur and Chinese words. experiment.
Then according to the word alignment results obtained by the two methods in the training corpus, the alignment accuracy, recall rate, and F value of these two word alignment results are analyzed.

4.1. Data Preparation
The training corpus used in this experiment is 1000 sentences of Vernain and Han sentences. These entire corpuses have undergone preliminary preprocessing, and the Chinese part has been segmented with the jieba segmentation.

The two pairs of VS-HAN sentences are respectively stored in two files as aligned corpora. One line in the file is used as a line of independent sentences, and the line number of the sentence and corresponding Chinese sentence line number are corresponding sentences.

4.2. Experimental Results And Analysis
In order to test the alignment effects of different word alignment models and the effect of different word alignment results on machine translation, word alignments of the three-word IBM model alignment and the training corpus-based filtering method based on the perplexity were performed on the training corpus. These alignment results are obtained by one-way alignment of GIZA++.

| Word alignment                          | P/% | R/% | F/% |
|----------------------------------------|-----|-----|-----|
| IBM Model 3                            | 30  | 28.7| 29.3|
| Confusion filtering method + IBM Model 3| 31.6| 29.3| 30.41|

The experimental results given in Table 1 indicate that misaligned sentence pairs in the training corpus affect the results of bilingual word alignment. In the experiment of filtering training corpora with perplexity, the accuracy rate, recall rate, and F value are better than those obtained by the IBM Model model. This shows that the ambiguity-based filtering method used in this paper helps to improve the performance of Alignment of Chinese and Uighur words.

5. Conclusion
This article describes the use of training-based corpus filtering methods based on perplexity. The statistical translation model in machine translation systems is based on sentence alignment. In practice, most sentence-aligned corpora are from the textual level. Aligned corpora are automatically extracted, so there is often a large amount of alignment errors in the training corpora. The perplexity-based filtering method introduced in this paper can effectively filter out the alignment sentence pairs that affect the translation quality error, improve the word alignment performance, and improve the automatic translation quality.

6. Acknowledgment
This work is supported by National Nature Science Foundation (61331013).

7. References
[1] Manning, C.D., et al., translated by Yuan Chunfa et al. Statistical Natural Language Processing Foundation [M]. Beijing: Publishing House of Electronics Industry, 2005.1.
[2] Lü Xueqiang, Wu Honglin, Yao Tianshun. Alignment of English and Chinese Words without Bilingual Dictionaries [J]. August 2004, 27 (8): 7301-7401.
[3] Gale, W. and Church, K. Identifying Word Correspondences in Parallel Texts[A].Proceedings of the 4th DARPA Speech and Natural Language Workshop[C], Pacific Grove, CA, 1991.152-157.
[4] Church K.W. Char align: A program for aligning parallel texts at the character level[A].Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics [C]. Columbus, Ohio, 1993.1-8.
[5] Huang, 2000] Jin-Xia Huang, and Key-Sun Choi. Chinese-Korean word alignment based on linguistic comparison[C]. In: Annual Meeting of the Association for Computational Linguistics, 2000, 392-399.

[6] Deng Dan. Research on Chinese-English word alignment algorithms based on bilingual dictionaries [J]. Computer Engineering. August 2005.

[7] Lu Yajuan. Alignment of Words in Bilingual Corpora Combining Statistics and Dictionary Methods [J]. The 6th Joint Academic Conference on Computational Linguistics. 2001.

[8] Liu Jianming, Tuergen and Yibu Layin. Research on Word Alignment in Chinese and Uyghur based on Statistical Machine Translation. Xinjiang University, 2010, 5

[9] Maiheynhoti, Mairehaba Aili, Milwand Snow and Laihei. The effect of different suffixes on Uighur word alignment in Uighur. Xinjiang University, 2015, 8.

[10] Li Ping, Yang Yong, Ren Ge, Saimiti Ai Li. Alignment of Chinese and Uyghur words based on HMM and dictionaries. Xinjiang Normal University, 2017.

[11] Tan Xun, Tuergen Ibrahim, Uyghur Sentence Clustering and Uyghur Word Alignment Research, Xinjiang University, 2012, 5.

[12] Liang Huashen, Zhao Tiejun. Filtration of bilingual corpora in statistical machine translation and improvement of word alignment. Harbin Institute of Technology, 2013, 8.