A Chinese-English Organization Name Translation System Using Heuristic Web Mining and Asymmetric Alignment

Fan Yang, Jun Zhao, Kang Liu
National Laboratory of Pattern Recognition
Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China
{fyang,jzhao,kliu}@nlpr.ia.ac.cn

Abstract

In this paper, we propose a novel system for translating organization names from Chinese to English with the assistance of web resources. Firstly, we adopt a chunking-based segmentation method to improve the segmentation of Chinese organization names which is plagued by the OOV problem. Then a heuristic query construction method is employed to construct an efficient query which can be used to search the bilingual Web pages containing translation equivalents. Finally, we align the Chinese organization name with English sentences using the asymmetric alignment method to find the best English fragment as the translation equivalent. The experimental results show that the proposed method outperforms the baseline statistical machine translation system by 30.42%.

1 Introduction

The task of Named Entity (NE) translation is to translate a named entity from the source language to the target language, which plays an important role in machine translation and cross-language information retrieval (CLIR). The organization name (ON) translation is the most difficult subtask in NE translation. The structure of ON is complex and usually nested, including person name, location name and sub-ON etc. For example, the organization name "北京诺基亚通信有限公司 (Beijing Nokia Communication Ltd.)" contains a company name (诺基亚/Nokia) and a location name (北京/Beijing). Therefore, the translation of organization names should combine transliteration and translation together.

Many previous researchers have tried to solve ON translation problem by building a statistical model or with the assistance of web resources. The performance of ON translation using web knowledge is determined by the solution of the following two problems:

- *The efficiency of web page searching*: how can we find the web pages which contain the translation equivalent when the amount of the returned web pages is limited?
- *The reliability of the extraction method*: how reliably can we extract the translation equivalent from the web pages that we obtained in the searching phase?

For solving these two problems, we propose a Chinese-English organization name translation system using heuristic web mining and asymmetric alignment, which has three innovations.

1) *Chunking-based segmentation*: A Chinese ON is a character sequences, we need to segment it before translation. But the OOV words always make the ON segmentation much more difficult. We adopt a new two-phase method here. First, the Chinese ON is chunked and each chunk is classified into four types. Then, different types of chunks are segmented separately using different strategies. Through chunking the Chinese ON first, the OOVs can be partitioned into one chunk which will not be segmented in the next phase. In this way, the performance of segmentation is improved.

2) *Heuristic Query construction*: We need to obtain the bilingual web pages that contain both the input Chinese ON and its translation equivalent. But in most cases, if we just send the Chinese ON to the search engine, we will always get the Chinese monolingual web pages which don't contain any English word sequences, let alone the English translation equivalent. So we propose a heuristic query construction method to generate an efficient bilingual query. Some words in the Chinese ON are selected and their translations are added into the query. These English words will act as clues for searching

Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP, pages 387–395,
Suntec, Singapore, 2-7 August 2009.

bilingual web pages. The selection of the Chinese words to be translated will take into consideration both the translation confidence of the words and the information contents that they contain for the whole ON.

3) *Asymmetric alignment*: When we extract the translation equivalent from the web pages, the traditional method should recognize the named entities in the target language sentence first, and then the extracted NEs will be aligned with the source ON. However, the named entity recognition (NER) will always introduce some mistakes. In order to avoid NER mistakes, we propose an asymmetric alignment method which align the Chinese ON with an English sentence directly and then extract the English fragment with the largest alignment score as the equivalent. The asymmetric alignment method can avoid the influence of improper results of NER and generate an explicit matching between the source and the target phrases which can guarantee the precision of alignment.

In order to illustrate the above ideas clearly, we give an example of translating the Chinese ON "中国华融资产管理公司 (China Huarong Asset Management Corporation)".

*Step*1: We first chunk the ON, where "LC", "NC", "MC" and "KC" are the four types of chunks defined in Section 4.2.

中国(China)/LC 华融(Huarong)/NC 资产管理(asset management)/MC 公司(corporation)/KC

*Step*2: We segment the ON based on the chunking results.

中国(china) 华融(Huarong) 资产(asset) 管理(management) 公司(corporation)

If we do not chunk the ON first, the OOV word "华融(Huarong)" may be segmented as "华融". This result will certainly lead to translation errors.

Step 3: Query construction:

We select the words "资产" and "管理" to translate and a bilingual query is constructed as: "中国华融资产管理公司" + asset + management

If we don't add some English words into the query, we may not obtain the web pages which contain the English phrase "*China Huarong Asset Management Corporation*". In that case, we can not extract the translation equivalent.

Step 4: Asymmetric Alignment: We extract a sentence "…*President of China Huarong Asset Management Corporation*…" from the returned snippets. Then the best fragment of the sentence "*China Huarong Asset Management Corporation*" will be extracted as the translation equivalent. We don't need to implement English NER process which may make mistakes.

The remainder of the paper is structured as follows. Section 2 reviews the related works. In Section 3, we present the framework of our system. We discuss the details of the ON chunking in Section 4. In Section 5, we introduce the approach of heuristic query construction. In section 6, we will analyze the asymmetric alignment method. The experiments are reported in Section 7. The last section gives the conclusion and future work.

2 Related Work

In the past few years, researchers have proposed many approaches for organization translation. There are three main types of methods. The first type of methods translates ONs by building a statistical translation model. The model can be built on the granularity of word [Stalls et al., 1998], phrase [Min Zhang et al., 2005] or structure [Yufeng Chen et al., 2007]. The second type of methods finds the translation equivalent based on the results of alignment from the source ON to the target ON [Huang et al., 2003; Feng et al., 2004; Lee et al., 2006]. The ONs are extracted from two corpora. The corpora can be parallel corpora [Moore et al., 2003] or content-aligned corpora [Kumano et al., 2004]. The third type of methods introduces the web resources into ON translation. [Al-Onaizan et al., 2002] uses the web knowledge to assist NE translation and [Huang et al., 2004; Zhang et al., 2005; Chen et al., 2006] extracts the translation equivalents from web pages directly.

The above three types of methods have their advantages and shortcomings. The statistical translation model can give an output for any input. But the performance is not good enough on complex ONs. The method of extracting translation equivalents from bilingual corpora can obtain high-quality translation equivalents. But the quantity of the results depends heavily on the amount and coverage of the corpora. So this kind of method is fit for building a reliable ON dictionary. In the third type of method, with the assistance of web pages, the task of ON translation can be viewed as a two-stage process. Firstly, the web pages that may contain the target translation are found through a search engine. Then the translation equivalent will be extracted from the web pages based on the alignment score with the original ON. This method will not

depend on the quantity and quality of the corpora and can be used for translating complex ONs.

3 The Framework of Our System

The Framework of our ON translation system shown in Figure 1 has four modules.

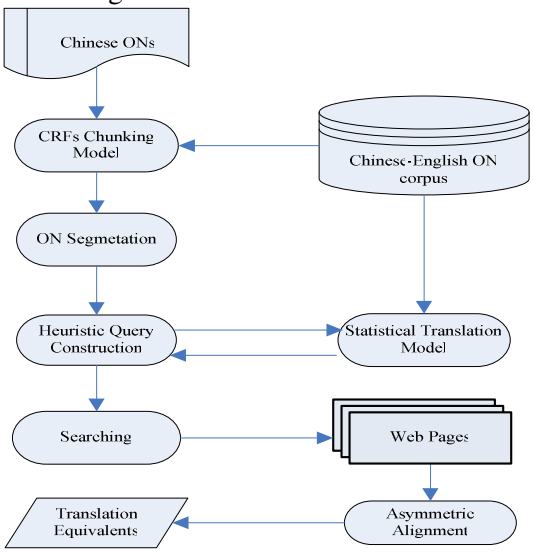


Figure 1. System framework

1) *Chunking-based ON Segmentation Module*: The input of this module is a Chinese ON. The Chunking model will partition the ON into chunks, and label each chunk using one of four classes. Then, different segmentation strategies will be executed for different types of chunks.

2) *Statistical Organization Translation Module*: The input of the module is a word set in which the words are selected from the Chinese ON. The module will output the translation of these words.

3) *Web Retrieval Module*: When input a Chinese ON, this module generates a query which contains both the ON and some words' translation output from the translation module. Then we can obtain the snippets that may contain the translation of the ON from the search engine. The English sentences will be extracted from these snippets.

4) *NE Alignment Module*: In this module, the asymmetric alignment method is employed to align the Chinese ON with these English sentences obtained in Web retrieval module. The best part of the English sentences will be extracted as the translation equivalent.

4 The Chunking-based Segmentation for Chinese ONs

In this section, we will illustrate a chunking-based Chinese ON segmentation method, which can efficiently deal with the ONs containing OOVs.

4.1 The Problems in ON Segmentation

The performance of the statistical ON translation model is dependent on the precision of the Chinese ON segmentation to some extent. When Chinese words are aligned with English words, the mistakes made in Chinese segmentation may result in wrong alignment results. We also need correct segmentation results when decoding. But Chinese ONs usually contain some OOVs that are hard to segment, especially the ONs containing names of people or brand names. To solve this problem, we try to chunk Chinese ONs firstly and the OOVs will be partitioned into one chunk. Then the segmentation will be executed for every chunk except the chunks containing OOVs.

4.2 Four Types of Chunks

We define the following four types of chunks for Chinese ONs:

- *Location Chunk (LC)*: LC contains the location information of an ON.
- *Name Chunk (NC)*: NC contains the name or brand information of an ON. In most cases, Name chunks should be transliterated.
- *Modification Chunk (MC)*: MC contains the modification information of an ON.
- *Key word Chunk (KC)*: KC contains the type information of an ON.

The following is an example of an ON containing these four types of chunks.

北京(Beijing)/LC 百富勤(Peregrine)/NC 投资咨询(investment consulting)/MC 有限公司(co.)/KC

In the above example, the OOV "百富勤(Peregrine)" is partitioned into name chunk. Then the name chunk will not be segmented.

4.3 The CRFs Model for Chunking

Considered as a discriminative probabilistic model for sequence joint labeling and with the advantage of flexible feature fusion ability, Conditional Random Fields (CRFs) [J.Lafferty et al., 2001] is believed to be one of the best probabilistic models for sequence labeling tasks. So the CRFs model is employed for chunking.

We select 6 types of features which are proved to be efficient for chunking through experiments. The templates of features are shown in Table 1,

Description	Features
current/previous/success character	C_0、C_{-1}、C_1
whether the characters is a word	$W(C_{-2}C_{-1}C_0)$、$W(C_0C_1C_2)$、$W(C_{-1}C_0C_1)$
whether the characters is a location name	$L(C_{-2}C_{-1}C_0)$、$L(C_0C_1C_2)$、$L(C_{-1}C_0C_1)$
whether the characters is an ON suffix	$SK(C_{-2}C_{-1}C_0)$、$SK(C_0C_1C_2)$、$SK(C_{-1}C_0C_1)$
whether the characters is a location suffix	$SL(C_{-2}C_{-1}C_0)$、$SL(C_0C_1C_2)$、$SL(C_{-1}C_0C_1)$
relative position in the sentence	$POS(C_0)$

Table 1. Features used in CRFs model

where C_i denotes a Chinese character, i denotes the position relative to the current character. We also use bigram and unigram features but only show trigram templates in Table 1.

5 Heuristic Query Construction

In order to use the web information to assist Chinese-English ON translation, we must firstly retrieve the bilingual web pages effectively. So we should develop a method to construct efficient queries which are used to obtain web pages through the search engine.

5.1 The Limitation of Monolingual Query

We expect to find the web pages where the Chinese ON and its translation equivalent co-occur. If we just use a Chinese ON as the query, we will always obtain the monolingual web pages only containing the Chinese ON. In order to solve the problem, some words in the Chinese ON can be translated into English, and the English words will be added into the query as the clues to search the bilingual web pages.

5.2 The Strategy of Query Construction

We use the metric of precision here to evaluate the possibility in which the translation equivalent is contained in the snippets returned by the search engine. That means, on the condition that we obtain a fixed number of snippets, the more the snippets which contain the translation equivalent are obtained, the higher the precision is. There are two factors to be considered. The first is how efficient the added English words can improve the precision. The second is how to avoid adding wrong translations which may bring down the precision. The first factor means that we should select the most informative words in the Chinese ON. The second factor means that we should consider the confidence of the SMT model at the same time. For example:

天津/LC 本田/NC 摩托车/MC 有限公司/KC
(Tianjin Honda motor co. ltd.)

There are three strategies of constructing queries as follows:

Q1.“天津本田摩托车有限公司” Honda

Q2.“天津本田摩托车有限公司” Ltd.

Q3.“天津本田摩托车有限公司” Motor Tianjin

In the first strategy, we translate the word “本田(Honda)” which is the most informative word in the ON. But its translation confidence is very low, which means that the statistical model gives wrong results usually. The mistakes in translation will mislead the search engine. In the second strategy, we translate the word which has the largest translation confidence. Unfortunately the word is so common that it can’t give any help in filtering out useless web pages. In the third strategy, the words which have sufficient translation confidence and information content are selected.

5.3 Heuristically Selecting the Words to be Translated

The mutual information is used to evaluate the importance of the words in a Chinese ON. We calculate the mutual information on the granularity of words in formula 1 and chunks in formula 2. The integration of the two kinds of mutual information is in formula 3.

$$MIW(x,Y)=\sum_{y\in Y}\log\frac{p(x,y)}{p(x)p(y)} \quad (1)$$

$$MIC(c,Y)=\sum_{y\in Y}\log\frac{p(y,c)}{p(y)p(c)} \quad (2)$$

$$IC(x,Y)=\alpha MIW(x,Y)+(1-\alpha)MIC(c_x,Y) \quad (3)$$

Here, $MIW(x,Y)$ denotes the mutual information of word x with ON Y. That is the summation of the mutual information of x with every word in Y. $MIC(c,Y)$ is similar. c_x denotes the label of the chunk containing x.

We should also consider the risk of obtaining wrong translation results. We can see that the name chunk usually has the largest mutual information. However, the name chunk always needs to be transliterated, and transliteration is often more difficult than translation by lexicon. So we set a threshold T_c for translation confidence. We only select the words whose translation confidences are higher than T_c, with their mutual information from high to low.

6 Asymmetric Alignment Method for Equivalent Extraction

After we have obtained the web pages with the assistant of search engine, we extract the equivalent candidates from the bilingual web pages. So we first extract the pure English sentences and then an asymmetric alignment method is executed to find the best fragment of the English sentences as the equivalent candidate.

6.1 Traditional Alignment Method

To find the translation candidates, the traditional method has three main steps.

1) The NEs in the source and the target language sentences are extracted separately. The NE collections are S_{ne} and T_{ne}.

2) For each NE in S_{ne}, calculate the alignment probability with every NE in T_{ne}.

3) For each NE in S_{ne}, the NE in T_{ne} which has the highest alignment probability will be selected as its translation equivalent.

This method has two main shortcomings:

1) Traditional alignment method needs the NER process in both sides, but the NER process may often bring in some mistakes.

2) Traditional alignment method evaluates the alignment probability coarsely. In other words, we don't know exactly which target word(s) should be aligned to for the source word. A coarse alignment method may have negative effect on translation equivalent extraction.

6.2 The Asymmetric Alignment Method

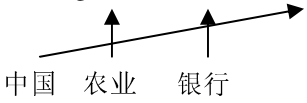
To solve the above two problems, we propose an asymmetric alignment method. The alignment method is so called "asymmetric" for that it aligns a phrase with a sentence, in other words, the alignment is conducted between two objects with different granularities. The NER process is not necessary for that we align the Chinese ON with English sentences directly.

[Wai Lam et al., 2007] proposed a method which uses the KM algorithm to find the optimal explicit matching between a Chinese ON and a given English ON. KM algorithm [Kuhn, 1955] is a traditional graphic algorithm for finding the maximum matching in bipartite weighted graph. In this paper, the KM algorithm is extended to be an asymmetric alignment method. So we can obtain an explicit matching between a Chinese ON and a fragment of English sentence.

A Chinese NE $CO=\{CW_1, CW_2, \ldots, CW_n\}$ is a sequence of Chinese words CW_i and the English sentence $ES=\{EW_1, EW_2, \ldots, EW_m\}$ is a sequence of English words EW_i. Our goal is to find a fragment $EW_{i,i+n}=\{EW_i, \ldots, EW_{i+n}\}$ in ES, which has the highest alignment score with CO. Through executing the extended KM algorithm, we can obtain an explicit matching L. For any CW_i, we can get its corresponding English word EW_j, written as $L(CW_i)=EW_j$ and vice versa. We find the optimal matching L between two phrases, and calculate the alignment score based on L. An example of the asymmetric alignment will be given in Fig2.

Step 1:

… [(The) Agriculture Bank (of) China] (is) (the) four



Step 2:

(The) Agriculture [Bank (of) China] (is) (the) four]…

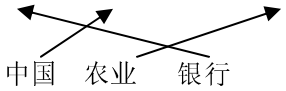


Fig2. An example of asymmetric alignment

In Fig2, the Chinese ON "中国农业银行" is aligned to an English sentence "… the Agriculture Bank of China is the four…". The stop words in parentheses are deleted for they have no meaning in Chinese. In *step* 1, the English fragment contained in the square brackets is aligned with the Chinese ON. We can obtain an explicit matching L_1, shown by arrows, and an alignment score. In *step* 2, the square brackets move right by one word, we can obtain a new matching L_2 and its corresponding alignment score, and so on. When we have calculated every consequent fragment in English sentence, we can find the best fragment "the Agriculture Bank of China" according to the alignment score as the translation equivalent.

The algorithm is shown in Fig3. Where, m is the number of words in an English sentence and n is the number of words in a Chinese ON. KM algorithm will generate an equivalent sub-graph by setting a value to each vertex. The edge whose weight is equal to the summation of the values of its two vertexes will be added into the sub-graph. Then the Hungary algorithm will be executed in the equivalent sub-graph to find the optimal matching. We find the optimal matching between $CW_{1,n}$ and $EW_{1,n}$ first. Then we move the window right and find the optimal matching between $CW_{1,n}$ and $EW_{2,n+1}$. The process will continue until the window arrives at the right most of the

English sentence. When the window moves right, we only need to find a new matching for the new added English vertex EW_{end} and the Chinese vertex C_{drop} which has been matched with EW_{start} in the last step. In the Hungary algorithm, the matching is added through finding an augmenting path. So we only need to find one augmenting path each time. The time complexity of finding an augmenting path is $O(n^3)$. So the whole complexity of asymmetric alignment is $O(m*n^3)$.

Algorithm: Asymmetric Alignment Algorithm

Input: A segmented Chinese ON *CO* and an English sentence ES.
Output: an English fragment $EW_{k,k+n}$

1. Let *start*=1, *end*=*n*, L_0=*null*
2. Using KM algorithm to find the optimal matching between two phrases $CW_{1,n}$ and $EW_{start,end}$ based on the previous matching $L_{start-1}$. We obtain a matching L_{start} and calculate the alignment score S_{start} based on L_{start}.
3. $CW_{drop} = L(EW_{start})$ $L(CW_{drop})$=*null*.
4. If (*end*==*m*) go to 7, else *start*=*start*+1, *end*=*end*+1.
5. Calculate the feasible vertex labeling for the vertexes CW_{drop} and EW_{end}
6. Go to 2.
7. The fragment $EW_{k,k+n-1}$ which has the highest alignment score will be returned.

Fig3. The asymmetric alignment algorithm

6.3 Obtain the Translation Equivalent

For each English sentence, we can obtain a fragment $ES_{i,i+n}$ which has the highest alignment score. We will also take into consideration the frequency information of the fragment and its distance away from the Chinese ON. We use formula (4) to obtain a final score for each translation candidate ET_i and select the largest one as translation result.

$$S(ET_i) = \alpha SA_i + \beta \log(C_i + 1) + \gamma \log(1/D_i + 1) \quad (4)$$

Where C_i denotes the frequency of ET_i, and D_i denotes the nearest distance between ET_i and the Chinese ON.

7 Experiments

We carried out experiments to investigate the performance improvement of ON translation under the assistance of web knowledge.

7.1 Experimental Data

Our experiment data are extracted from LDC2005T34. There are two corpora, ldc_propernames_org_ce_v1.beta (Indus_corpus for short) and ldc_propernames_industry_ce_v1.beta (Org_corpus for short). Some pre-process will be executed to filter out some noisy translation pairs. For example, the translation pairs involving other languages such as Japanese and Korean will be filtered out. There are 65,835 translation pairs that we used as the training corpus and the chunk labels are added manually.

We randomly select 250 translation pairs from the Org_corpus and 253 translation pairs from the Indus_corpus. Altogether, there are 503 translation pairs as the testing set.

7.2 The Effect of Chunking-based Segmentation upon ON Translation

In order to evaluate the influence of segmentation results upon the statistical ON translation system, we compare the results of two translation models. One model uses chunking-based segmentation results as input, while the other uses traditional segmentation results.

To train the CRFs-chunking model, we randomly selected 59,200 pairs of equivalent translations from Indus_corpus and org_corpus. We tested the performance on the set which contains 6,635 Chinese ONs and the results are shown as Table-2.

For constructing a statistical ON translation model, we use *GIZA++*[1] to align the Chinese NEs and the English NEs in the training set. Then the phrase-based machine translation system *MOSES*[2] is adopted to translate the 503 Chinese NEs in testing set into English.

	Precision	Recall	F-measure
LC	0.8083	0.7973	0.8028
NC	0.8962	0.8747	0.8853
MC	0.9104	0.9073	0.9088
KC	0.9844	0.9821	0.9833
All	0.9437	0.9372	0.9404

Table 2. The test results of CRFs-chunking model

We have two metrics to evaluate the translation results. The first metric *L*1 is used to evaluate whether the translation result is exactly the same as the answer. The second metric *L*2 is used to evaluate whether the translation result contains almost the same words as the answer,

[1] http://www.fjoch.com/GIZA++.html
[2] http://www.statmt.org/moses/

without considering the order of words. The results are shown in Table-3:

	chunking-based segmentation	traditional segmentation
*L*1	21.47%	18.29%
*L*2	40.76%	36.78%

Table 3. Comparison of segmentation influence

From the above experimental data, we can see that the chunking-based segmentation improves *L*1 precision from 18.29% to 21.47% and *L*2 precision from 36.78% to 40.76% in comparison with the traditional segmentation method. Because the segmentation results will be used in alignment, the errors will affect the computation of alignment probability. The chunking based segmentation can generate better segmentation results; therefore better alignment probabilities can be obtained.

7.3 The Efficiency of Query Construction

The heuristic query construction method aims to improve the efficiency of Web searching. The performance of searching for translation equivalents mostly depends on how to construct the query. To test its validity, we design four kinds of queries and evaluate their ability using the metric of average precision in formula 5 and macro average precision (MAP) in formula 6,

$$Average\,\mathrm{Precision} = \frac{1}{N}\sum_{i=1}^{N}\frac{H_i}{S_i} \qquad (5)$$

where H_i is the count of snippets that contain at least one equivalent for the *i*th query. And S_i is the total number of snippets we got for the *i*th query,

$$MAP = \frac{1}{N}\sum_{j=1}^{N}\frac{1}{H_j}\sum_{i=1}^{H_j}\frac{i}{R(i)} \qquad (6)$$

where *R*(i) is the order of snippet where the *i*th equivalent occurs. We construct four kinds of queries for the 503 Chinese ONs in testing set as follows:

*Q*1: only the Chinese ON.

*Q*2: the Chinese ON and the results of the statistical translation model.

*Q*3: the Chinese ON and some parts' translation selected by the heuristic query construction method.

*Q*4: the Chinese ON and its correct English translation equivalent.

We obtain at most 100 snippets from *Google* for every query. Sometimes there are not enough snippets as we expect. We set α in formula 4 at 0.7，and the threshold of translation confidence at 0.05. The results are shown as Table 4.

	Average precision	MAP
*Q*1	0.031	0.0527
*Q*2	0.187	0.2061
*Q*3	0.265	0.3129
*Q*4	1.000	1.0000

Table 4. Comparison of four types query

Here we can see that, the result of *Q*4 is the upper bound of the performance, and the *Q*1 is the lower bound of the performance. We concentrate on the comparison between *Q*2 and *Q*3. *Q*2 contains the translations of every word in a Chinese ON, while *Q*3 just contains the translations of the words we select using the heuristic method. *Q*2 may give more information to search engine about which web pages we expect to obtain, but it also brings in translation mistakes that may mislead the search engine. The results show that *Q*3 is better than *Q*2, which proves that a careful clue selection is needed.

7.4 The Effect of Asymmetric Alignment Algorithm

The asymmetric alignment method can avoid the mistakes made in the NER process and give an explicit alignment matching. We will compare the asymmetric alignment algorithm with the traditional alignment method on performance. We adopt two methods to align the Chinese NE with the English sentences. The first method has two phases, the English ONs are extracted from English sentences firstly, and then the English ONs are aligned with the Chinese ON. Lastly, the English ON with the highest alignment score will be selected as the translation equivalent. We use the software *Lingpipe*[3] to recognize NEs in the English sentences. The alignment probability can be calculated as formula 7:

$$Score(C,E) = \sum_i \sum_j p(e_i \mid c_j) \qquad (7)$$

The second method is our asymmetric alignment algorithm. Our method is different from the one in [Wai Lam et al., 2007] which segmented a Chinese ON using an English ON as suggestion. We segment the Chinese ON using the chunking-based segmentation method. The English sentences extracted from snippets will be preprocessed. Some stop words will be deleted, such as "*the*", "*of*", "*on*" etc. To execute the extended KM algorithm for finding the best alignment matching, we must assure that the vertex number in each side of the bipartite is the

[3] http://www.alias-i.com/lingpipe/

same. So we will execute a phrase combination process before alignment, which combines some frequently occurring consequent English words into single vertex, such as "*limited company*" etc. The combination is based on the phrase pair table which is generated from phrase-based SMT system. The results are shown in Table 5:

	Asymmetric Alignment	Traditional method	Statistical model
Top1	**48.71%**	36.18%	**18.29%**
Top5	53.68%	46.12%	**--**

Table 5. Comparison the precision of alignment method

From the results (column 1 and column 2) we can see that, the Asymmetric alignment method outperforms the traditional alignment method. Our method can overcome the mistakes introduced in the NER process. On the other hand, in our asymmetric alignment method, there are two main reasons which may result in mistakes, one is that the correct equivalent doesn't occur in the snippet; the other is that some English ONs can't be aligned to the Chinese ON word by word.

7.5 Comparison between Statistical ON Translation Model and Our Method

Compared with the statistical ON translation model, we can see that the performance is improved from 18.29% to 48.71% (the bold data shown in column 1 and column 3 of Table 5) by using our Chinese-English ON translation system. Transforming the translation problem into the problem of searching for the correct translation equivalent in web pages has three advantages. First, word order determination is difficult in statistical machine translation (SMT), while search engines are insensitive to this problem. Second, SMT often loses some function word such as "*the*", "*a*", "*of*", etc, while our method can avoid this problem because such words are stop words in search engines. Third, SMT often makes mistakes in the selection of synonyms. This problem can be solved by the fuzzy matching of search engines. In summary, web assistant method makes Chinese ON translation easier than traditional SMT method.

8 Conclusion

In this paper, we present a new approach which translates the Chinese ON into English with the assistance of web resources. We first adopt the chunking-based segmentation method to improve the ON segmentation. Then a heuristic query construction method is employed to construct a query which can search translation equivalent more efficiently. At last, the asymmetric alignment method aligns the Chinese ON with English sentences directly. The performance of ON translation is improved from 18.29% to 48.71%. It proves that our system can work well on the Chinese-English ON translation task. In the future, we will try to apply this method in mining the NE translation equivalents from monolingual web pages. In addition, the asymmetric alignment algorithm also has some space to be improved.

Acknowledgement

The work is supported by the National High Technology Development 863 Program of China under Grants no. 2006AA01Z144, and the National Natural Science Foundation of China under Grants no. 60673042 and 60875041.

References

Yaser Al-Onaizan and Kevin Knight. 2002. Translating named entities using monolingual and bilingual resources. In Proc of ACL-2002.

Yufeng Chen, Chenqing Zong. 2007. A Structure-Based Model for Chinese Organization Name Translation. In Proc. of ACM Transactions on Asian Language Information Processing (TALIP)

Donghui Feng, Yajuan Lv, Ming Zhou. 2004. A new approach for English-Chinese named entity alignment. In Proc. of EMNLP 2004.

Fei Huang, Stephan Vogal. 2002. Improved named entity translation and bilingual named entity extraction. In Proc. of the 4th IEEE International Conference on Multimodal Interface.

Fei Huang, Stephan Vogal, Alex Waibel. 2003. Automatic extraction of named entity translingual equivalence based on multi-feature cost minimization. In Proc. of the 2003 Annual Conference of the ACL, Workshop on Multilingual and Mixed-language Named Entity Recognition

Masaaki Nagata, Teruka Saito, and Kenji Suzuki. 2001. Using the Web as a Bilingual Dictionary. In Proc. of ACL 2001 Workshop on Data-driven Methods in Machine Translation.

David Chiang. 2005. A hierarchical phrase-based model for statistical machine translation. In Proc. of ACL 2005.

Conrad Chen, Hsin-His Chen. 2006. A High-Accurate Chinese-English NE Backward Translation System Combining Both Lexical Information and Web Statistics. In Proc. of ACL 2006.

Wai Lam, Shing-Kit Chan. 2007. Named Entity Translation Matching and Learning: With Application for Mining Unseen Translations. In Proc. of ACM Transactions on Information Systems.

Chun-Jen Lee, Jason S. Chang, Jyh-Shing R. Jang. 2006. Alignment of bilingual named entities in parallel corpora using statistical models and multiple knowledge sources. In Proc. of ACM Transactions on Asian Language Information Processing (TALIP).

Kuhn, H. 1955. The Hungarian method for the assignment problem. Naval Rese. Logist. Quart 2,83-97.

Min Zhang., Haizhou Li, Su Jian, Hendra Setiawan. 2005. A phrase-based context-dependent joint probability model for named entity translation. In Proc. of the 2nd International Joint Conference on Natural Language Processing(IJCNLP)

Ying Zhang, Fei Huang, Stephan Vogel. 2005. Mining translations of OOV terms from the web through cross-lingual query expansion. In Proc. of the 28th ACM SIGIR.

Bonnie Glover Stalls and Kevin Knight. 1998. Translating names and technical terms in Arabic text. In Proc. of the COLING/ACL Workshop on Computational Approaches to Semitic Language.

J. Lafferty, A. McCallum, and F. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In Proc. ICML-2001.

Tadashi Kumano, Hideki Kashioka, Hideki Tanaka and Takahiro Fukusima. 2004. Acquiring bilingual named entity translations from content-aligned corpora. In Proc. IJCNLP-04.

Robert C. Moore. 2003. Learning translation of named-entity phrases from parallel corpora. In Proc. of 10th conference of the European chapter of ACL.