Chinese-English Organization Name Translation Based on Correlative Expansion

Feiliang Ren, Muhua Zhu, Huizhen Wang, Jingbo Zhu
Natural Language Processing Lab, Northeastern University, Shenyang, China
{renfeiliang,zhumuhua}@gmail.com
{wanghuizhen,zhujingbo}@mail.neu.edu.cn

Abstract

This paper presents an approach to translating Chinese organization names into English based on correlative expansion. Firstly, some candidate translations are generated by using statistical translation method. And several correlative named entities for the input are retrieved from a correlative named entity list. Secondly, three kinds of expansion methods are used to generate some expanded queries. Finally, these queries are submitted to a search engine, and the refined translation results are mined and re-ranked by using the returned web pages. Experimental results show that this approach outperforms the compared system in overall translation accuracy.

1 Introduction

There are three main types of named entity: location name, person name, and organization name. Organization name translation is a subtask of named entity translation. It is crucial for many NLP tasks, such as cross-language information retrieval, machine translation, question and answering system. For organization name translation, there are two problems among it which are very difficult to handle.

Problem I: There is no uniform rule that can be abided by to select proper translation methods for the inside words of an organization name. For example, a Chinese word "东北", when it is used as a modifier for a university, it is translated to ***Northeastern*** for "***东北***大学/***Northeastern*** University", and is translated to ***Northeast*** for "***东北***林业大学/***Northeast*** Forestry University", and is mapped to Chinese Pinyin ***Dongbei*** for "***东北***财经大学/***Dongbei*** University of Finance and Economics". It is difficult to decide which translation method should be chosen when we translate the inside words of an organization name.

Problem II: There is no uniform rule that can be abided by to select proper translation order and proper treatment of particles Here particles refer to prepositions and articles) for an input organization name. For example, the organization name "中国建设银行/China Construction Bank" and the organization name "中国农业银行/Agricultural Bank of China", they are very similar both in surface forms and in syntax structures, but their translation orders are different, and their treatments of particles are also different.

Generally, there are two strategies usually used for named entity translation in previous research. One is alignment based approach, and the other is generation based approach. Alignment based approach (Chen et al. 2003; Huang et al. 2003; Hassan and Sorensen, 2005; and so on) extracts named entities translation pairs from parallel or comparable corpus by some alignment technologies, and this approach is not suitable to solve the above two problems. Because new organization names are constantly being created, and alignment based method usually fails to cover these new organization names that don't occur in the bilingual corpus.

Traditional generation based approach (Al-Onaizan and Knight, 2002; Jiang et al .2007; Yang et al. 2008; and so on) usually consists of two parts. Firstly, it will generate some candidate translations for the input; then it will re-rank these candidate translations to assign the correct translations high ranks. Cheng and Zong [2008] proposed another generation based approach for organization name translation, which directly translates organization names according to their inherent structures. But their approach still can't solve the above two problems. This is because the amount of organization names is so huge and many of them have their own special translation rules to handle the above two problems. And the inherent structures don't reveal these translation rules. Traditional generation based approach is suitable for organization name translation. But in previous research, the final translation performance depends on the candidate translation gen-

Proceedings of the 2009 Named Entities Workshop, ACL-IJCNLP 2009, pages 143–151,
Suntec, Singapore, 7 August 2009. ©2009 ACL and AFNLP

eration process greatly. If this generation process failed, it is impossible to obtain correct result from the re-ranking process. In response to this, Huang et al. [2005] proposed a novel method that mined key phrase translation form web by using topic-relevant hint words. But in their approach, they removed the candidate translation generation process, which will improve extra difficult during mining phrase. Besides, in their approach, the features considered to obtain topic-relevant words are not so comprehensive, which will affect the quality of returned web pages where the correct translations are expected to be included. There is still much room for the improvement process of the topic-relevant words extraction.

Inspired by the traditional generation based named entity translation strategy and the approach proposed by Huang et al., we propose an organization name translation approach that mining the correct translations of input organization name from the web. Our aim is to solve the above two problems indirectly by retrieving the web pages that contain the correct translation of the input and mining the correct translation from them. Given an input organization name, firstly, some candidate translations are generated by using statistical translation method. And several correlative named entities for the input are retrieved from a correlative named entity list. Secondly, expanded queries are generated by using three kinds of query expansion methods. Thirdly, these queries are submitted to a search engine, and the final translation results are mined and reranked by using the returned web pages.

The rest of this paper is organized as follows, section 2 presents the extraction process of correlative named entities, section 3 presents a detail description of our translation method for Chinese organization name, and section 4 introduces our parameter evaluation method, and section 5 is the experiments and discussions part, finally conclusions and future work are given in section 6.

2 Extraction of Correlative Named Entities

The key of our approach is to find some web pages that contain the correct translation of the input. With the help of correlative named entities (here if two named entities are correlative, it means that they are usually used to describe the same topic), it is easier to find such web pages. This is because that in the web, one web page usually has one topic. Thus if two named entities are correlative, they are very likely to occur in pair in some web pages.

The correlative named entity list is constructed in advance. During translation, the correlative named entities for the input organization name are retrieved from this list directly. To set up this correlative named entity list, an about 180GB-sized collection of web pages are used. Totally there are about 100M web pages in this collection. Named entities are recognized from every web page by using a NER tool. This NER tool is trained by CRF model[1] with the corpus from SIGHAN-2008[2].

2.1 Features Used

During the extraction of correlative named entities, the following features are considered.

Co-occurrence in a Document The more often two named entities co-occur in a document, the more likely they are correlative. This feature is denoted as $CoD_i(n_1, n_2)$, which means the co-occurrence of named entities n_1 and n_2 in a document D_i. This feature is also the main feature used in Huang et al. [2005].

Co-occurrence in Documents The more often two named entities co-occur in different documents, the more likely they are correlative. This feature is denoted as $CoDs(n_1, n_2)$, which means the number of documents that both n_1 and n_2 occur in.

Distance The closer two named entities is in a document, the more likely they are correlative. This feature is denoted as $DistD_i(n_1, n_2)$, which means the number of words between n_1 and n_2 in a document D_i.

Mutual Information Mutual information is a metric to measure the correlation degree of two words. The higher two named entities' mutual information, the more likely they are correlative. And the mutual information of named entities n_1 and n_2 in a document D_i is computed as following formula.

$$MID_i(n_1, n_2) = p(n_1, n_2) \log \frac{p(n_1, n_2)}{p(n_1) \cdot p(n_2)} \quad (1)$$

Jaccard Similarity Jaccard similarity is also a metric to measure the correlative degree of two words. The higher two named entities' Jaccard

[1] http://www.chasen.org/~taku/software/CRF++/
[2] http://www.china-language.gov.cn/bakeoff08/

similarity, the more likely they are correlative. And Jaccard similarity is computed as following formula.

$$Jaccard(n_1, n_2) = \frac{CoDs(n_1, n_2)}{D(n_1) + D(n_2) - CoDs(n_1, n_2)} \quad (2)$$

where $D(n_i)$ is the number of documents that n_i occurs in, and $CoDs(n_i, n_j)$ is the number of documents that both n_i and n_j occur in.

TF-IDF TF-IDF is a weight computation method usually used in information retrieval. Here for a named entity n_i, TF-IDF is used to measure the importance of its correlative named entities. The TF-IDF value of n_j in a document D_i is computed as following formula.

$$TF - IDF_i(n_j) = tf_{ij} \times \log \frac{N}{D(n_j)} \quad (3)$$

where tf_{ij} is the frequency of n_j in document D_i, N is the number of total documents, and $D(n_j)$ is the number of documents that n_j occurs in.

2.2 Feature Combination

During the process of feature combination, every feature is normalized, and the final correlative degree of two named entities is the linear combination of these normalized features, and it is computed as following formula.

$$C(n_i, n_j) = \lambda_1 \frac{\sum_k CoD_k(n_i, n_j)}{\sum_j \sum_k CoD_k(n_i, n_j)} + \lambda_2 \frac{CoDs(n_i, n_j)}{\sum_j CoDs(n_i, n_j)} + \lambda_3 \frac{\sum_k 1/DistD_k(n_i, n_j)}{\sum_j \sum_k 1/DistD_k(n_i, n_j)} + \lambda_4 \frac{\sum_k MID_k(n_i, n_j)}{\sum_j \sum_k MID_k(n_i, n_j)} + \lambda_5 \frac{Jaccard(n_i, n_j)}{\sum_j Jaccard(n_i, n_j)} + \lambda_6 \frac{\sum_k TF - IDF_k(n_j)}{\sum_k \sum_j TF - IDF_k(n_j)} \quad (4)$$

Finally, for every organization name n_i, its top-K correlative named entities are selected to construct the correlative named entity list.

During translation, the correlative words for the input can be retrieved from this correlative list directly. If the input is not included in this list, the same method as in Huang et al. [2005] is used to obtain the needed correlative words.

3 Organization Name Translation Based on Correlative Expansion

3.1 Statistical Translation Module

The first step of our approach is to generate some candidate translations for every input organization name. As shown in table 1, these candidate translations are used as query stems during query expansion. We use Moses[3], a state of the art public machine translation tool, to generate such candidate translations. Here Moses is trained with the bilingual corpus that is from the 4th China Workshop on Machine Translation[4]. Total there are 868,947 bilingual Chinese-English sentence pairs on news domain in this bilingual corpus. Moses receives an organization name as input, and outputs the N-best results as the candidate translations of the input organization name. Total there are six features used in Moses: phrase translation probability, inverse phrase translation probability, lexical translation probability, inverse lexical translation probability, language model, and sentence length penalty. All the needed parameters are trained with MERT method (Och, 2003) by using a held-out development set.

3.2 Query Expansions

Because the amount of available web pages is so huge, the query submitted to search engine must be well designed. Otherwise, the search engine will return large amount of un-related web pages. This will enlarge the difficulty of mining phase. Here three kinds of expansion methods are proposed to generate some queries by combining the clues given by statistical translation method and the clues given by correlative named entities of the input. And these correlative named entities are retrieved from the correlative named entities list before the query expansions process. These three kinds of expansions are explained as follows.

3.2.1 Monolingual Expansion

Given an input organization name n_i, suppose s_i is one of its candidate translations, and n_j is one of its correlative named entities. If n_j can be reliably translated[5], we expand s_i with this reli-

[3] http://www.statmt.org/moses/
[4] http://www.nlpr.ia.ac.cn/cwmt-2008
[5] A word can be reliably translated means either it has a unique dictionary translation or it is a Chinese

able translation $t(n_j)$ to form a query "$s_i + t(n_j)$". This kind of expansion is called as monolingual expansion.

For two named entities, if they are correlative, their translations are likely correlative too. So their translations are also likely to occur in pair in some web pages. Suppose a query generated by this expansion is "$s_i + t(n_j)$", if the candidate translation s_i is the correct translation of the input, there must be some returned web pages that contain s_i completely. Otherwise, it is still possible to obtain some returned web pages that contain the correct translation. This is because that the search engine will return both the web pages that include the query completely and the web pages that include the query partly. And for a translation candidate s_i and the correct translation s_i', they are very likely to have some common words, so some of their returned web pages may overlap each other. Thus it can be expected that when we submit "$s_i + t(n_j)$" to search engine, it will return some web pages that include "$s_i' + t(n_j)$" or include s_i'. This is very helpful for the mining phase.

3.2.2 Bilingual Expansion

Given an input organization name n_i, suppose s_i is one of its candidate translations, we expand s_i with n_i to form a query "$s_i + n_i$". This kind of expansion is called as bilingual expansion.

Bilingual expansion is very useful to verify whether a candidate translation is the correct translation. To give readers more information or they are not sure about the translation of original named entity, the Chinese authors usually include both the original form of a named entity and its translation in the mix-language web pages [Fei Huang et al, 2005]. So the correct translation pair is likely to obtain more supports from the returned web pages than those incorrect translation pairs. Thus bilingual expansion is very useful for the re-ranking phase.

Besides, for an input organization name, if one of its incorrect candidate translations s_i is very similar to the correct translation s_i' in surface form, the correct translation is also likely to be contained in the returned web pages by using this kind of queries. The reason for this is the search mechanism of search engine, which has been explained above in monolingual expansion.

3.2.3 Mix-language Expansion

Given an input organization name n_i, suppose s_i is one of its candidate translations, and n_j is one of its correlative named entities. We expand s_i with n_j to form a query "$s_i + n_j$". This kind of expansion is called as mix-language expansion.

Mix-language expansion is a necessary complement to the other two expansions. Besides, this mix-language expansion is more prone to obtain some mix-language web pages that may contain both the original input organization name and its correct translation.

3.3 Mining

When the expanded queries are submitted to search engine, the correct translation of the input organization name may be contained in the returned web pages. Because the translation of an organization name must be also an organization name, we mine the correct translation of the input among the English organization names. Here we use the Stanford named entity recognition toolkits[6] to recognize all the English organization names in the returned web pages. Then align these recognized organization names to the input by considering the following features.

Mutual Translation Probability The translation probability measures the semantic equivalence between a source organization name and its target candidate translation. And mutual translation probability measures this semantic equivalence in two directions. For simplicity, here we use IBM model-1(Brown et al. 1993), which computes two organization names' translation probability using the following formula.

$$p(f \mid e) = \frac{1}{L^J} \prod_{j=1}^{J} \sum_{l=1}^{L} p(f_j \mid e_l) \qquad (6)$$

where $p(f_j \mid e_l)$ is the lexical translation probability. Suppose the input organization name is n_i, s_i is one of the recognized English organi-

person name and can be translated by Pinyin mapping.

[6] http://nlp.stanford.edu/software/CRF-NER.shtml

zation names, the mutual translation probability of n_i and s_i is computed as:

$$mp(n_i, s_i) = \lambda p(n_i \mid s_i) + (1-\lambda) p(s_i \mid n_i) \quad (7)$$

Golden Translation Ratio For two organization names, their golden translation ratio is defined as the percentage of words in one organization name whose reliable transactions can be found in another organization name. This feature is used to measure the probability of one named entity is the translation of the other. It is computed as following formula.

$$GR(n_i, s_j) = \lambda \frac{G(n_i, s_j)}{|n_i|} + (1-\lambda) \frac{G(s_j, n_i)}{|s_j|} \quad (8)$$

where $G(n_i, s_j)$ is the number of golden translated words from n_i to s_j, and $G(s_j, n_i)$ is the number of golden translated words from s_j to n_i.

Co-occurrence In Web Pages For an input organization name n_i and a recognized candidate translation s_j, the more often they co-occur in different web pages, the more likely they are translations of each other. This feature is denoted as $CoS(n_i, s_j)$, which means the number of web pages that both n_1 and s_j occur in.

Input Matching Ratio This feature is defined as the percentage of the words in the input that can be found in a returned web page. For those mix-language web pages, this feature is used to measure the probability of the correct translation occurring in a returned web page. It is computed as the following formula.

$$IMR(n_i, s_k) = \frac{|n_i \cap s_k|}{|n_i|} \quad (9)$$

where s_k is the $k-th$ returned web page.

Correlative Named Entities Matching Ratio This feature is defined as the percentage of the words in a correlative named entity that can be found in a returned web page. This feature is also used to measure the probability of the correct translation occurring in a returned web page. It is computed as the following formula.

$$CW\_MR(c_i, s_k) = \frac{|c_i \cap s_k|}{|c_i|} \quad (10)$$

The final confidence score of n_i and t_j to be a translation pair is measured by following formula. As in formula 4, here every factor will be is normalized during computation.

$$C(n_i, t_j) = \lambda_1 mp(n_i, t_j) + \lambda_2 GR(n_i, t_j) + \lambda_3 \frac{CoSs(n_i, n_j)}{\sum_j CoS(n_i, n_j)} + \frac{\lambda_4}{K} \sum_k IMR(n_i, s_k) + \frac{\lambda_5}{K \times I} \sum_i \sum_k CW\_MR(c_i, s_k) \quad (11)$$

where K is the number of returned web pages, I is the number of correlative named entities for the input organization name.

For every input organization name, we remain a fixed number of mined candidate translations with the highest confidence scores. And add them to the original candidate translation set to form a revised candidate translation set.

3.4 Re-ranking

The aim of mining is to improve recall. And in the re-ranking phase, we hope to improve precision by assigning the correct translation a higher rank. The features considered here for the re-ranking phase are listed as follows.

Confidence Score The confidence score of n_i and t_j is not only useful for the mining phase, but also is useful for the re-ranking phase. The higher this score, the higher rank this candidate translation should be assigned.

Inclusion Ratio For Bilingual Query This feature is defined as the percentage of the returned web pages that the bilingual query is completely matched. It is computed as the following formula.

$$EHR\_BQ(q_i) = \frac{h(q_i)}{H(q_i)} \quad (12)$$

where $h(q_i)$ is the number of web pages that match the query q_i completely, and $H(q_i)$ is the total number of returned web pages for query q_i.

Candidate Inclusion Ratio for Monolingual Query and Mix-language Query This feature is defined as the percentage of the returned web pages that the candidate translation is completed matched. This feature for monolingual query is computed as formula 13, and this feature for mix-language query is computed as formula 14.

$$ECHR\_MlQ(s_i) = h(s_i) \Big/ H(q_i) \quad (13)$$

$$ECHR\_MixQ(s_i) = h(s_i) \Big/ H(q_i) \quad (14)$$

where $h(s_i)$ is the number of web pages that match the candidate translation s_i completely, and

$H(q_i)$ is the total number of returned web pages for query q_i.

Finally, the above features are combined with following formula.

$$R(n_i,t_j)=\lambda_1 C(n_i,t_j)+\frac{\lambda_2}{N}\sum_i EHR\_BQ(q_i)+\frac{\lambda_3}{M}\sum_i ECHR\_MlQ(s_i)+\frac{\lambda_4}{L}\sum_i ECHR\_MixQ(s_i) \quad (15)$$

where N is the number of candidate translations, M and L are the number of monolingual queries and mix-language queries respectively.

At last the revised candidate translation set is re-ranked according to this formula, and the top-K results are outputted as the input's translation results.

4 Parameters Evaluations

In above formula (4), formula (11) and formula (15), the parameters λ_i are interpolation feature weights, which reflect the importance of different features. We use some held-our organization name pairs as development set to train these parameters. For those parameters in formula (4), we used those considered features solely one by one, and evaluated their importance according to their corresponding inclusion ratio of correct translations when using mix-language expansion and the final weights are assigned according to the following formula.

$$\lambda_i=\frac{InclusionRate_i}{\sum_i InclusionRate_i} \quad (16)$$

Where $InclusionRate_i$ is the inclusion rate when considered feature f_i only. The inclusion rate is defined as the percentage of correct translations that are contained in the returned web pages as Huang et al.[2005] did.

To obtain the parameters in formula (11), we used those considered features solely one by one, and computed their corresponding precision on development set respectively, and final weights are assigned according to following formula.

$$\lambda_i=P_i \Big/ \sum_i P_i \quad (17)$$

Where P_i is the precision when considered feature f_i only. And for the parameters in formula (15), their assignment method is the same with the method used for formula (11).

5 Experiments and Discussions

We use a Chinese to English organization name translation task to evaluate our approach. The experiments consist of four parts. Firstly, we evaluate the contribution of the correlative named entities for obtaining the web pages that contain the correct translation of the input. Secondly, we evaluate the contribution of different query expansion methods. Thirdly, we investigate to which extents our approach can solve the two problems mentioned in section 1. Finally, we evaluate how much our approach can improve the overall recall and precision. Note that for simplicity, we use 10-best outputs from Moses as the original candidate translations for every input. And the search engine used here is Live[7].

5.1 Test Set

The test set consists of 247 Chinese organization names recognized from 2,000 web pages that are downloaded from Sina[8]. These test organization names are translated by a bilingual speaker given the text they appear in. And these translations are verified from their official government web pages respectively. During translation, we don't use any contextual information.

5.2 Contribution of Correlative Named Entities

The contribution of correlative named entities is evaluated by inclusion rate, and we compare the inclusion rate with different amount of correlative named entities and different amount of returned web pages. The experimental results are shown in Table 1 (here we use all these three kinds of expanding strategies).

		# of correlative named entities used		
		1	5	10
#of web pages used	1	0.17	0.39	0.47
	5	0.29	0.63	0.78
	10	0.32	0.76	0.82

Table 1. Comparisons of inclusion rate

From these results we can find that our approach obtains an inclusion rate of 82% when we use 10 correlative named entities and 10 returned web pages. We notice that there are some Chinese organization names whose correct English translations have multiple standards. For example, the organization name "国防部"is translated

[7] http://www.live.com/

[8] http://news.sina.com.cn/

into “Department of Defense” when it refers to a department in US, but is translated into “Ministry of Defence” when it refers to a department in UK or in Singapore. This problem affects the actual inclusion rate of our approach. Another factor that affects the inclusion rate is the search engine used. There is a small difference in the inclusion rate when different search engines are used. For example, the Chinese organization name “中信银行/China CITIC Bank”, because the word “中信” is an out-of-vocabulary word, the best output from Moses is “of the bank”. With such candidate translation, none of our three expansion methods works. But when we used Google as search engine instead, we mined the correct translation.

From these results we can conclude that by using correlative named entities, the returned web pages are more likely to contain the correct translations of the input organization names.

5.3 Contribution of Three Query Expansion Methods

In this section, we evaluate the contribution of these three query expansion methods respectively. To do this, we use them one by one during translation, and compare their inclusion rates respectively. Experimental results are shown in Table 2.

			#of web pages used		
			1	5	10
# of correlative named entities used	Monolingual Expansion Only	1	0.002	0.002	0.004
		5	0.017	0.019	0.019
		10	0.021	0.037	0.051
	Bilingual Expansion Only	1	0.112	0.159	0.174
		5	0.267	0.327	0.472
		10	0.285	0.414	0.669
	Mix-language Expansion Only	1	0.098	0.138	0.161
		5	0.231	0.307	0.386
		10	0.249	0.398	0.652

Table 2. Inclusion rate of different kinds of query expansion methods

From Table 2 we can see that bilingual expansion and mix-language expansion play greater roles than monolingual expansion in obtaining the web pages that contain the correct translations of the inputs. This is because the condition of generating monolingual queries is too strict, which requires a reliable translation for the correlative named entity. In most cases, this condition cannot be satisfied. So for many input organization names, we cannot generate any monolingual queries for them at all. This is the reason why monolingual expansion obtains so poorer an inclusion rate compared with the other two expansions. To evaluate the true contribution of monolingual expansion method, we carry out another experiment. We select 10 organization names randomly from the test set, and translate all of their correlative named entities into English by a bilingual speaker. Then we evaluate the inclusion rate again on this new test set. The experimental results are shown in Table 3.

		# of correlative named entities used		
		1	5	10
#of web pages used	1	0.2	0.3	0.6
	5	0.4	0.7	0.9
	10	0.4	0.8	0.9

Table 3. Inclusion rate for monolingual expansion on new test set

From Table 3 we can conclude that, if most of the correlative named entities can be reliably translated, the queries generated by this monolingual expansion will play greater role in obtaining the web pages that contain the correct translations of the inputs.

From those results in Table 2 we can conclude that, these three kinds of expansions complement each other. Using them together can obtain higher inclusion rate than using anyone of them only.

5.4 Efficiency on Solving Problem I and Problem II

In this section, we investigate to which extents our approach can solve the two problems mentioned in section 1.We compare the wrong translation numbers caused by these two problems (another main kind of translation error is caused by the translation of out-of-vocabulary words) between Moses and our approach. The experimental results are shown in Table 4.

	Moses Results	Our method
Problem I	44	3
Problem II	30	0

Table 4. Comparison of error numbers

From Table 4 we can see that our approach is very effective on solving these two problems. Almost all of the errors caused by these two problems are corrected by our approach. Only three wrong translations are not corrected. This is because that there are some Chinese organization names whose correct English translations have multiple standards, such as the correct translation of organization name “国防部”depends on its nationality, which has been explained in section 5.2.

5.5 Our Approach vs. Other Approaches

In this section, we compare our approach with other two methods: Moses and the approach proposed by Huang et al. [2005]. We compare their accuracy of Top-K results. For both our approach and Huang et al.'s approach, we use 10 correlative words for each input organization name and use 10 returned web pages for mining the correct translation result. The experimental results are shown in Table 5.

	Moses Results	Huang's Results	Our Results
Top 1	0.09	0.44	0.53
Top 5	0.18	0.61	0.73
Top 10	0.31	0.68	0.79

Table 5. Moses results vs. our results

Moses is a state-of-the-art translation method, but it can hardly handle the organization name translation well. In addition to the errors caused by the above two problems mentioned in section 1, the out-of-vocabulary problem is another obstacle for Moses. For example, when translating the organization name "国际海啸信息中心/International Tsunami Information Centre", because the word "海啸" is an out-of-vocabulary word, Moses fails to give correct translation. But for those approaches that have a web mining process during translation, both the out-of-vocabulary problem and the two problems mentioned in section 1 are less serious. This is the reason that Moses obtains the lowest performance compared with the other two approaches. Our approach is also superior to Huang's method, as shown in the above table. We think this is because of the following three reasons. The first reason is that in our approach, we use a translation candidate generation process. Although these candidates are usually not so good, they can still provide some very useful clue information for the web retrieval process. The second reason is that the features considered for correlative words extraction in our approach are more comprehensive. Most of the time (except for the case that the input is not included in the correlative word list) our approach is more prone to obtain better correlative words for the input. The third reason is that our approach use more query expansion strategies than Huang's approach. These expansion strategies may complement each other and improve the probability of obtaining the web pages that contain the correct translations For example, both Moses and Huang's approach failed to translate the organization name "国际海啸信息中心". But in our approach, with the candidate translation "International Information Centre" that is generated by Moses, our approach still can obtain the web page that contains the correct translation when using bilingual expansion. Thus the correct translation "International Tsunami Information Centre" is mined out during the sequent mining process.

From table 5 we also notice that the final recall of our approach is a little lower than the inclusion rate as show in table 1. This means that our approach doesn't mine all the correct translations that are contained in the returned web pages. One of the reasons is that some of the input organization names are not clearly expressed. For example, an input organization name "伯克利分校", although its correct translation "University of California, Berkeley" is contained in the returned web pages, this correct translation cannot be mined out by our approach. But if it is expressed as "加利福尼亚大学伯克利分校", its correct translation can be mined from the returned web pages easily. Besides, the recognition errors of NER toolkits will also reduce the final recall of our approach.

6 Conclusions and Future Work

In this paper, we present a new organization name translation approach. It uses some correlative named entities of the input and some query expansion strategies to help the search engine to retrieve those web pages that contain the correct translation of the input. Experimental results show that for most of the inputs, their correct translations are contained in the returned web pages. By mining these correct translations and re-ranking them, the two problems mentioned in section 1 are solved effectively. And recall and precision are also improved correspondingly.

In the future, we will try to improve the extraction perform of correlative named entities. We will also try to apply this approach to the person name translation and location name translation.

Acknowledgments

This work was supported by the open fund of National Laboratory of Pattern Recognition, Institute of Automation Chinese Academy of Science, P.R.C, and was also supported in part by National Science Foundation of China (60873091), Natural Science Foundation of Liaoning Province (20072032) and Shenyang Science and Technology Plan (1081235-1-00).

References

Chen Hsin-Hsi, Changhua Yang, and Ying Lin. 2003. Learning formulation and transformation rules for multilingual named entities. Proceedings of the ACL 2003 Workshop on Multilingual and Mixed-language Named Entity Recognition. pp1-8.

Dekang Lin, Shaojun Zhao, Durme Benjamin Van Drume, Marius Pasca. Mining Parenthetical Translations from the Web by Word Alignment, ACL08. pp994-1002.

Fan Yang, Jun Zhao, Bo Zou, Kang Liu, Feifan Liu. 2008. Chinese-English Backward Transliteration Assisted with Mining Monolingual Web Pages. ACL2008. pp541-549.

Fei Huang, Stephan Vogel and Alex Waibel. 2003. Automatic Extraction of Named Entity Translingual Equivalence Based on Multi-feature Cost Minimization. Proceedings of the 2003 Annual Conference of the Association for Computational Linguistics, Workshop on Multilingual and Mixed-language Named Entity Recognition.

Fei Huang, Stephan vogel and Alex Waibel. 2004. Improving Named Entity Translation Combining Phonetic and Semantic Similarities. Proceedings of the HLT/NAACL. pp281-288.

Fei Huang, Ying Zhang, Stephan Vogel. 2005. Mining Key Phrase Translations from Web Corpora. HLT-EMNLP2005, pp483-490.

Feng, Donghui, Yajuan LV, and Ming Zhou. 2004. A new approach for English-Chinese named entity alignment. Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2004), pp372-379.

Franz Josef Och. 2003. Minimum Error Rate Training in Statistical Machine Translation. ACL2003. pp160-167.

Jin-Shea Kuo, Haizhou Li, Ying-Kuei Yang. Learning Transliteration Lexicon from the Web. COLING/ACL2006. pp1129-1136.

Hany Hassan and Jeffrey Sorensen. 2005. An Integrated Approach for Arabic-English Named Entity Translation. Proceedings of ACL Workshop on Computational Approaches to Semitic Languages. pp87-93.

Lee, Chun-Jen and Jason S.Chang and Jyh-Shing Roger Jang. 2004a. Bilingual named-entity pairs extraction from parallel corpora. Proceedings of IJCNLP-04 Workshop on Named Entity Recognition for Natural Language Processing Application. pp9-16.

Lee, Chun-Jen, Jason S.Chang and Thomas C. Chuang. 2004b. Alignment of bilingual named entities in parallel corpora using statistical model. Lecture Notes in Artificial Intelligence. 3265:144-153.

Lee, Chun-Jen, Jason S.Chang, and Jyh-Shing Roger Jang. 2005. Extraction of transliteration pairs from parallel corpora using a sta Acquisition of English-Chinese transliterated word pairs from parallel-aligned text using a statistical transliteration model. Information Sciences.

Long Jiang, Ming Zhou, Lee-Feng Chien, Cheng Niu. [2007]. Named Entity Translation with Web Mining and Transliteration. IJCAI-2007.

Moore, Robert C. 2003. Learning translations of named-entity phrases form parallel corpora. ACL-2003. pp259-266.

Peter F. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. 1993. The Mathematics of Statistical Machine Translation: Parameter Estimation. Computational Linguistics, 19(2):263-311.

Y. Al-Onaizan and K. Knight. 2002. Translating named entities using monolingual and bilingual resources. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pp400-408.

Ying Zhang and Phil Vines Using the Web for Automated Translation Extraction in Cross-Language Information Retrieval. SIGIR2004,pp162-169.

Yufeng Chen, Chengqing Zong. A Structure-based Model for Chinese Organization Name Translation. ACM Transactions on Asian Language Information Processing, 2008, 7(1), pp1-30.