A Novel Chinese-English ON Translation Method Using Mix-language Web Pages

Feiliang REN
Northeastern University
Shenyang, China
renfeiliang@ise.neu.edu.cn

Jingbo ZHU
Northeastern University
Shenyang, China
zhujingbo@mail.neu.edu.cn

Huizhen WANG
Northeastern University
Shenyang, China
wanghuizhen@ise.neu.edu.cn

Abstract:

In this paper, we propose a novel Chinese-English organization name translation method with the assistance of mix-language web resources. Firstly, all the implicit out-of-vocabulary terms in the input Chinese organization name are recognized by a CRFs model. Then the input Chinese organization name is translated without considering these recognized out-of-vocabulary terms. Secondly, we construct some efficient queries to find the mix-language web pages that contain both the original input organization name and its correct translation. At last, a similarity matching and limited expansion based translation identification approach is proposed to identify the correct translation from the returned web pages. Experimental results show that our method is effective for Chinese organization name translation and can improve performance of Chinese organization name translation significantly.

Keywords:

Chinese organization names translation; named entity translation; web assistant translation method; machine translation

1. Introduction

These Named entity (NE) translation is to translate a NE from one language to the other, which is very important to many Natural Language Processing (NLP) tasks, such as cross-language information retrieval and data-driven machine translation. According to NE types, NE translation usually can be classified into three main sub-tasks: person name (PER) translation, location name (LOC) translation, and organization name (ON) translation. And ON translation is attracting more and more research attention. Many previous works have been done on ON translation. Recently, researchers (Y.Al-Onaizan et al., 2002,Huang et al., 2005, Jiang et al., 2007,Yang et al., 2008, Yang et al., 2009, Ren et al., 2009, and so on) have focused on translating ON with the assistance of web resources. It is believed that there must be some mix-language web pages that contain both the input ON and its correct translation since there are large amount of web pages on the web. Thus ON translation problem is converted to following two problems. One problem is to find the mix-language web pages that contain both the input ON and its correct translation. The other problem is to identify the correct translation from the obtained mix-language web pages.

To solve these two problems effectively, we propose a novel Chinese-English ON translation method using mix-language web page resources, which has following two innovations.

Query construction method considering out-of-vocabulary Solving the first problem usually involve some query construction methods. During query construction, bilingual query is essential to find the mix-language web pages that contain both the input ON and its correct translation. Generally there are two pats in these bilingual queries. One part is the original input ON, which is used as clues for finding the web pages that are related with the input ON. The other part is a translation candidate of the input ON, which is used as clues for finding the mix-language web pages. Previous researchers (Huang et al., 2005, Yang et al., 2009, Ren et al., 2009, and so on) have proposed many novel methods to construct bilingual query. However, to our best knowledge, few of researchers had taken out-of-vocabulary (OOV) problem into consideration during query construction. Different from the OOV problem in traditional machine translation task, there are two kinds of OOV terms in ON translation. One kind is explicit OOV term, and the other kind is implicit OOV term. Explicit OOV terms refer to the words that we can not obtain their translation items from existed bilingual dictionaries or bilingual corpora. Implicit OOV terms refer to the words that their translations in ONs are completely different from their translation items obtained from existed bilingual dictionaries or bilingual corpora. The OOV problem must be well solved. Otherwise, the constructed bilingual queries will mislead search engines to wrong web pages, which is fatal to the ON translation using web resources. In our method, all OOV terms inside an ON will be recognized by a CRFs model firstly. Then, these OOV terms will be ignored during the translation candidate generation process. Finally, some bilingual queries are constructed with

978-1-4244-6899-7/10/$26.00 ©2010 IEEE

some patterns.

Similarity matching and limited expansion based translation identification method Traditional translation identification methods usually recognize all the target language NEs in the returned mix-language web pages firstly. Then the recognized NEs will be aligned to the input ON with some probabilities. The NE with the highest alignment probability will be selected as the final translation result. However, the NE recognition process is not so perfect and it often bring in some recognition errors. Besides, NE nesting problem may also bring in some boundary recognition errors. These errors are fatal for identifying the correct translation. Moreover, traditional NE alignment methods usually don't use the obtained translation candidate at all. They usually align the input ON and the recognized NEs from scratch. In our option, since many factors have been considered carefully during translation candidate generation process, those generated translation candidates are sure to have some similarity or relationship with the correct translation. So it is not suitable to discard the generated translation candidates during alignment phase. To solve these problems, our translation identification method doesn't need recognize NE on the target language side, it identifies the correct translation directly from the target language sentences with a similarity matching and limited expansion method. Our method can avoid the influence of improper NE recognition results and can make full use of generated translation candidate.

In order to illustrate our approach clearly, we give an example of translating the Chinese ON "青岛 海尔 电冰箱 股份 有限 公司" into its English translation "Qingdao Haier Refrigerator Co., Ltd.".

Step 1: Firstly we recognize the OOV term "海尔 /Haier" with a CRFs model. Then the rest terms will be translated by a statistical translation method. As a result, we will obtain such a translation candidate: "Qingdao Refrigerator Co., Ltd.".

Step 2: We will construct several bilingual queries like "青岛 海尔 电冰箱 股份 有限 公司 + Qingdao Refrigerator Co., Ltd" and so on. Then these queries are used to find the mix-language web pages that contain both the input ON and its correct translation.

Step 3: We extract the correct translation "Qingdao Haier Refrigerator Co., Ltd" from the returned mix-language web pages with a similarity matching and limited expansion based translation identification method..

The rest of this paper is organized as follows. Section 2 presents our translation candidate generation strategy. In section 3, we introduce our query construction method. Our translation identification method is introduced in section 4. In section 5, we present our experiments and discussions. Finally conclusions and future work are given in section 6.

2. Translation Candidate Generation

The first step of our approach is to generate several translation candidates for the input ON. But if there were several OOV items in the input ON, it would absolutely introduce some translation errors if we translate theses OOV terms directly. These translation errors will mislead search engines to wrong web pages, which is very harmful for ON translation.

In our approach, we ignore these OOV terms during translation candidate generation. We think ignoring them would be better than translating them toughly. We think that without the noises caused by OOV translation errors, it would be easier to find the mix-language web pages that contain the input ON and its correct translation because of the fuzzy matching scheme in search engines. Accordingly, there are two subtasks in our translation candidate generation phase. One is OOV recognition task which is expected to recognize all the OOV terms in the input ON. The other is to translate the rest words of the input ON where OOV terms have been removed.

2.1. Conditional Random Fields Model for OOV Recognition

There are two kinds of OOV terms in ON translation: explicit OOV term and implicit OOV term. Explicit OOV terms can be easily recognized during translation. So here we focus on recognizing implicit OOV terms. We view the implicit OOV recognition problem as a sequence labeling problem. Since the Conditional Random Fields model has achieved great success for sequence labeling tasks, it is employed here. We use five kinds of features which are proved to be efficient for implicit OOV term recognition through experiments. These features are listed as followings.

Position Feature This kind of feature is used to denote the position of current word in the input ON. We use B/I/E tags, which B denotes that current word is the first words of the input ON, I denotes that current word is in the middle of the input ON, E denotes that current word is the last word of the input ON.

Length Feature This kind of feature is used to denote whether current word is a single character word.

Location Feature This kind of feature is used to denote the location of current word in the input ON.

POS Feature This kind of feature is used to denote the part-of-speech tag of current word.

Context Feature This kind of feature is to denote the context information from previous word and success word of current word. This context information includes length features of previous word and success word, POS features of previous word and success word, and position features of previous word and success word.

We download 20729 Chinese-English bilingual ON pairs from web2. And the implicit OOV terms in this corpus are labeled with following rule. For a word inside a Chinese ON, if we can find its some translation items either from a bilingual dictionary or from the SMT training corpus which will be introduced in the next section, but these translation items don't appear in the English translation part of the Chinese ON, the Chinese word is labeled as an implicit OOV item.

Finally, all the labeled Chinese ONs are selected as training data to train the CRFs model.

2.2 Translation Candidate Generation

The non-OOV terms in the input ON should be translated after OOV recognition. Here we use NEUTrans [Xiao et al., 2009] system to generate translation candidates for the input ON that implicit OOV terms have been removed. NEUTrans system is one of the states of art Chinese-to-English translation systems. It has participated in the news domain Chinese-to-English single-system translation task in the 5th China workshop of Machine Translation (CWMT2009) and has achieved very promising translation result (Its BLEU score is the second highest among all participated systems, and its BLEU score is also far higher than those systems that are based on Moses). There are six features used in NEUTrans: phrase translation probability, inverse phrase translation probability, lexical translation probability, inverse lexical translation probability, 5-gram language model, and sentence length penalty. All the needed parameters are trained with MERT method [Och, 2003]. The training corpus for NEUTrans consists about 370K bilingual sentences that are extracted from the corpora of LDC2005T10, LDC2003E07, LDC2003E14 and LDC2005T06. The language model is trained on the Xinhua portion of English Gigaword corpus. The development sets are NIST Chinese-to-English evaluation sets of MT03 (99 documents, 919 sentences).

3. Query Construction Strategy

Query construction is very important for the ON translation using web resources. The constructed queries must provide enough clues to find the needed web pages,. [Ren et al., 2009] proposed a correlative expansion based query construction method. In their method, two kinds of information are used. One is the information obtained from the original input ON and its translation candidate. And the other is the information obtained from some correlative words of the input ON. These two kinds of information will provide more comprehensive clue for search engines, thus it is easier to find the needed web pages for the ON translation using web pages. In this paper, we also use some correlative words information during query construction. And the extraction method of these correlative words is the same as [Ren et al., 2009].

Two factors are considered in our query method. The first factor is that the returned web pages must be highly related with the input ON. The other factor is that there must be some target language text in the returned web pages where the correct translation may be contained. Based on these factors, following patterns are used to construct the needed bilingual queries.

Query Pattern 1: ON+TransCand(ON)

Query Pattern 2: ON+TransCand(CWi)

Query Pattern 3: CWi+TransCand(CWi)

In above patterns, *ON* means the original input *ON*, *TransCand(ON)* is the translation candidate of *ON* that all the implicit OOV terms have been ignored, CWi is the *ith* correlative words of *ON*, and *TransCand(CWi)* is the translation candidate of *CWi*. *TransCand(CWi)* is generated with the same method for generating *TransCand(ON)*.

4. Translation Candidate Generation

4.1. Limitation of Traditional Method

Traditional translation identification methods usually recognize all the target language NEs in the returned mix-language web pages firstly. Then the recognized NEs will be aligned to the input ON with some probabilities. However, the NE recognition process is not so perfect and it often brings in some recognition errors. Besides, NE nesting problem may also bring in some boundary recognition errors. For example, the input Chinese ON is "厦门大学", its correct translation "Xiamen University" is contained in the returned mix-language web pages. But the NE recognition process outputs "School of Mathematical Sciences Xiamen University", where the correct translation is nested. These NE recognition errors are fatal for identifying the correct translation.

Moreover, during the NE alignment phase, traditional methods usually compute the alignment probability between input ON and target NE literally. But ONs cannot always be translated literally. There may be some insertion or deletion operations in translation [Chen et al., 2008]. For example, for a given the input ON "扶贫 开发 领导 小组", in its correct translation "Leading Group in Charge of Aid-the-Poor Projects", "开发" is removed and "in Charge of "is inserted. Besides, OOV problem can also increase the difficulty of this alignment process. On the other hand, since many factors have been considered carefully during translation candidate generation process, those generated translation candidates are sure to have some similarity or

[2] http://www.fane.cn/dic/No3/300~358.htm

relationship with the correct translations. So it is not suitable to discard the translation candidate information during alignment phase.

4.2. Similarity Matching and Limited Expansion Based Translation Identification Method

To overcome these shortcomings of traditional translation identification methods, we propose a similarity matching and limited expansion based translation identification algorithm. It doesn't need recognize NE on the target language side, and it identifies the correct translation directly from the target language sentences by similarity matching and limited expansion method.

In our method, firstly we extract all the continuous target texts from the returned mix-language web page. Then we anchor a possible correct translation boundary in one of a continuous target text by a similarity matching scheme. Finally this boundary is adjusted with a limited expansion strategy. An example of our translation identification process is shown in Fig1.

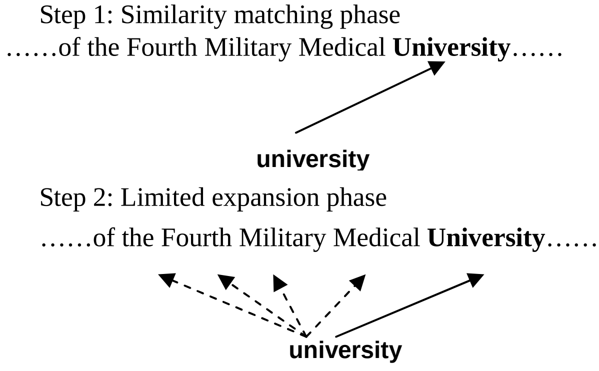
Step 1: Similarity matching phase
……of the Fourth Military Medical **University**……
university
Step 2: Limited expansion phase
……of the Fourth Military Medical **University**……
university

Fig1. An example of our identification method

In Fig1, the input Chinese ON is "第四 军医 大学", during OOV recognition process, "第四" is recognized as an implicit OOV term. And during translation candidate generation, "军医" is recognized as an explicit OOV term. Both of them will be ignored during translation candidate generation process. And the final generated translation candidate is "university".

During similarity matching phase, our method will try to find a continuous target text that has the maximal similarity with the original translation candidate. This continuous text determines the initial boundary for the possible correct translation. This process is designed based on the idea that the generated candidate is sure to have some similarity or relationship with the correct translation. Besides, this process can also solve the reordering problem in Chinese ON translation.

During the limited expansion phase, our method will expand the initial boundary with some expansion rules. This process is designed to solve the OOV translation problem and the translation item selection problem.

Our concrete method is shown in figure 2.

Input: A continuous English text $ES_{1,l}$, input Chinese ON *CO* and its translation candidate $TO_{1,m}$
Output: An English correct translation candidate $EO_{k,k+n}$

Algorithm:
1. Find the minimal continuous text $ES_{a,b}$ that covers the most words in $TO_{1,m}$ and subjects to $1 \le a, b \le l$.
2. For every possible continuous text fragment $ES_{k,j}$ that subjects to $a \le k, j \le b$, we compute the similarity between it and $TO_{1,m}$.
3. Select the top-k similarity results. And for each of these results $ES_{k,j}$, we expand it with some limited expansion rules. Repeat this procedure until $ES_{k,j}$ can not be expanded any more. And we denote the expanded $ES_{k,j}$ as $ES_{k,j}$'.
4. For each of the expanded result $ES_{k,j}$', we take "*CO* + $ES_{k,j}$'"as query and submit this query to search engine.
5. We rank these expanded results by the returned web count value obtained in step 4. The text fragment that has the highest web count value will be returned.

Fig 2. Our Translation Identification Algorithm

In the step 2 of our algorithm, the similarity is computed by following formula 1.

$$Sim(A,B) = \frac{\delta * SameWord(A,B)}{Len(A) + Len(B)} \quad (1)$$

In formula 1, δ is a factor which is used to adjust the confidence of similarity. A larger δ value will be assigned if input ON is completely contained in the mix-language web page where ES1,l is extracted. This similarity matching scheme can also solve the reordering problem in ON translation.

In the step 3 of our algorithm, the limited expansion rules are listed as followings.

Left Expansion Rule If the word w_{k-1} is subjected to following conditions, add wk-1 into $ES_{k,j}$, and update k to *k-1*.

(1) The first letter of w_{k-1} is capitalized.

(2) w_{k-1} must have at least one dictionary translation item that appears in *CO* or w_{k-1} must not have any dictionary translation items.

Right Expansion Rule If the word w_{j+1} is subjected to following conditions, add w_{j+1} into $ES_{k,j}$, and update *j* to *j+1*.

(1) The first letter of w_{j+1} is capitalized.

(2) w_{j+1} must have at least one dictionary translation item that appears in *CO* or w_{j+1} must not have any dictionary translation items.

Function Word Expansion Rule If the word w_{k-1} (or w_{j+1}) is a function word (such as articles or prepositions), add w_{k-1} (or w_{j+1}) into $ES_{k,j}$, and update *k* to *k-1*(or update *j* to *j+1*).

In above three expansion rules, the condition 2 in left expansion rule and right expansion rule is required to solve the out-of-vocabulary (OOV) translation problem and the translation item selection problem in ON translation.

After the function word expansion, some errors maybe introduced. So a post processing rule is necessary.

Post Processing Rule If the first word wk of $ES_{k,j}$ is a function word except "the", remove wk from $ES_{k,j}$; if the last word w_j of $ES_{k,j}$ is a function word, remove w_j from $ES_{k,j}$. Repeat this procedure until $ES_{k,j}$ can not be updated any more.

5. Experiments and Discussions

We use a Chinese to English ON translation task to evaluate our approach. The experiments consist of four parts. Firstly, we evaluate the efficiency of our OOV processing strategy upon ON translation. Secondly, we evaluate the efficiency of our query construction strategy. Thirdly, we investigate the effect of our translation identification algorithm. And finally, we evaluate how much our approach can improve recall and precision for ON translation. And the search engine used here is Bing[3].

5.1. Test Set

We download about 2,000 Chinese web pages from Sina[4]. Then all NEs are recognized from these web pages. Among these recognized NEs, 356 ONs are selected as test set. These test ONs are then translated by two bilingual speakers and the translation results are used as reference set. Some statistics of the test set are listed in table 1.

ON types	Num	Average Length	Percentage (%)
Research institute	137	6 characters	38.48
Government agency	98	8 characters	27.53
Company	39	7 characters	10.96
Others	82	7 characters	23.03
Total	356	7 characters	100

Table 1. Statistics of Test Set

5.2. The Efficiency of Our OOV Processing Strategy

To evaluate the efficiency of our OOV processing strategy, we compare the inclusion rate with two translation candidate generation strategies that the same query construction strategy is used. The inclusion rate is defined as the percentage of ONs whose correct translations are contained in the returned web pages [Huang et al., 2005]. The first strategy (denoted as baseline in table 2) uses NEUTrans to output a translation candidate for the input ON directly; and the second strategy is our approach, which OOV terms will be recognized and ignored during translation candidate generation. Query pattern 1 is used to construct query. The experimental results are shown in Table 2.

	Baseline	Our method
Inclusion Rate	0.61	0.79

Table 2. Inclusion rate with different translation candidate generation strategy

From experimental results in table 2 we can see that the inclusion rate of our method improves about 30% compared with baseline system. Many translation errors which would mislead search engines have been avoided when OOV terms are recognized and ignored during translation candidate generation. For example, the input Chinese ON is "东软 集团/Neusoft Group", and its NEUTrans output translation candidate is "east soft group". When we use this candidate to construct queries, its correct translation won't appear in returned web pages. On the contrary, in our system, the word "东软" is recognized as an OOV term and is ignored during translation candidate generation. Thus the corresponding constructed query is "东软 集团 group". With this query, search engine returns the web pages that contain the correct translation.

5.3. The Efficiency of Three Query Construction Patterns

To evaluate the efficiency of our three query construction patterns respectively, we use them separately one by one, and compare their inclusion rates. Experimental results are shown in Table 3.

			#of snippets used		
			1	5	10
# of correlative NEs used	Pattern1	0	0.351	0.442	0.714
	Pattern2	1	0.112	0.119	0.123
		5	0.152	0.153	0.179
		10	0.185	0.199	0.214
	Pattern3	1	0.048	0.078	0.084
		5	0.061	0.087	0.087
		10	0.092	0.114	0.121

Table 3. Inclusion rate comparisons with different query construction patterns

From Table 3 we can see that query pattern 1 play a major role on improving the inclusion rate. Compared with query pattern 2 and query pattern 3, query pattern achieves higher inclusion rate. We think this is because that in query pattern 1, both the source language part and the target language part are highly relative with our search goal. And we think this maybe also the reason that under same conditions, query pattern 2 achieves higher

[3] http://www.bing.com/

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

inclusion than query pattern 3. Besides, we think that query pattern 2 and pattern 3 are necessary supplements for query pattern 1. For search engines, the more clues are provides, the more possible to find the needed mix-language web pages. And query pattern 2 and query pattern 3 provide clues for search engines from another views. Thus, we think by using these different clues, search engines will be easier to find the needed results. To confirm our this idea, we carry out another experiment that use all these query patterns, and we evaluate the inclusion rate again. The experimental results are shown in table 4.

From table 4 we can see that inclusion rate improves significantly when correlative NEs are used, especially when the number of correlative NEs increases. This experiment also indicates that to further improve the inclusion rate, more comprehensive clues should be provided to search engines.

		#of snippets used		
		1	5	10
# of correlative NEs used	1	0.354	0.453	0.716
	5	0.354	0.499	0.757
	10	0.363	0.502	0.843

Table 4. Inclusion rate all query patterns used

5.4. The Efficiency of Our Translation Identification Algorithm

To evaluate the efficiency of our translation identification algorithm, we compare the identification accuracy between our algorithm and the identification approach used by [Ren et al., 2009]. Their approach firstly performs the NE recognition process, then compute the alignment probability between the input ON and the recognized NEs. In our method, we use 10 correlative NEs to construct queries, and we also all three kinds of query patterns. Experimental results are shown in table 5.

	Top-K outputs		
	1	5	10
Ren's approach	0.561	0.753	0.797
Our algorithm	0.566	0.789	0.825

Table 5. Accuracy comparison

From table 5 we can see that our algorithm outperforms Ren's approach. Many NE recognition errors have been avoided successfully by our approach. Our method can easily identify the correct translation for such examples like "扶贫 开发 领导 小组" which will tend to be wrongly identified by traditional methods. Besides, because neither the NE recognition process nor the NE alignment process is used in our method, our identification speed is far faster than traditional identification approaches.

5.5. Our Approach vs. Other Approaches

In this section, we compare our approach with other two methods: NEUTrans and the approach proposed by Huang et al. [2005]. We compare the accuracy of Top-K results. In our approach and Huang et al.'s approach, we use 10 correlative words query construction and use 10 returned web pages for mining the correct translation result. The experimental results are shown in Table 6.

	NEUTrans Results	Huang's Results	Our Results
Top 1	0.352	0.443	0.566
Top 5	0.367	0.656	0.789
Top 10	0.367	0.743	0.825

Table 6. Our approach vs. other approaches

From table 6 we can see that although NEUTrans is one of the state-of-the-art Chinese-English translation systems, it can hardly translate ONs well. In fact, there are usually three problems in ON translation that are hard to solve for NEUTrans. Besides the two problems that have been analyzed in [Ren et al., 2009], the OOV problem is another obstacle for NEUTrans. For example, when translating the ON "国际 海啸 信息 中心 /International Tsunami Information Centre", because the word "海啸" is an OOV word, NEUTrans fails to give correct translation. But for those approaches that use web information, all of these problems are less serious. This is the reason that NEUTrans 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 6. We think this is because of the following three reasons. The first reason is that in our approach, we design a novel translation candidate generation method, which can provide more 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. In most of the time (except for the case that the input is not included in the correlative word list), our approach is easier to obtain better correlative words for the input. The third reason is that our translation identification algorithm is better.

From table 6 we also notice that the final recall of our approach is lower than the inclusion rate show in table 4. This means that our approach doesn't mine all the correct translations that are contained in the returned web pages. According to our analysis, primarily the following error types that are listed in table 7 attribute these wrong results.

From Table 7 we can see that, boundary error is one of the main types of identification errors in our approach. This is because that many words' first letters are capitalized in a web page, thus, our approach tends to merge them in a translation candidate during the translation identification, which will introduce a boundary error. We think this problem should be further investigated.

Another main type of identification error is caused by the OOV recognition. As shown in Table 7, for the input ON “西安 公路 交通 大学”, if the word “公路” is recognized as an OOV term, our approach is more likely to mine the “Xi'an JiaoTong University” as final translation result.

Another identification error is caused because that some of the input ONs are not clearly expressed. For example, an input ON “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.

Input	Output	Reference	Error type	Percentage (%)
太原理工大学	Taiyuan University of Technology , Taiyuan	Taiyuan University of Technology	Boundary error	31
西安公路交通大学	Xi'an JiaoTong University	Xi'an Highway University	OOV recognition error	14
伯克利分校	Berkeley	University of California, Berkeley	Ambiguity input	9

Table 7. Identification Error Examples

6. Conclusions and Future Work

In this paper, we present a new ON translation approach with the assistance of web resources. And there are two contributions in our work. The first contribution is that we take OOV problem into consideration during query construction, which is very useful for improving the inclusion rate for ON translation. The second contribution is that we propose a novel translation identification algorithm which doesn’t need the NER process and can make full use of the generated translation candidates at the same time. Specially, during translation candidate generation, our method firstly recognizes all the implicit OOV terms in the input ON; then our method generates translation candidate for the rest part of the input ON with a state-of-the-art SMT system. With the generated translation candidate, three query patterns are used to construct query. Finally, a similarity matching and limited expansion based translation identification strategy is proposed to extract the correct translation from the returned web pages. Experimental results show that for most of the input ONs, their correct translations are contained in the returned web pages. And thus recall and precision are also improved correspondingly.

In the future, we will try to solve the identification errors listed in table 7. And we will also further investigate more efficient query construction strategy so that further improve the inclusion rate.

Acknowledgements

This paper is supported by the Open Project Program of the National Laboratory of Pattern Recognition (NLPR), and also is supported by the “the Fundamental Research Funds for the Central Universities” and National Science Foundation of China (60873091).

References

[1] 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.

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

[3] Fan Yang, Jun Zhao, Kang Liu. A Chinese-English Organization Name Translation System Using Heuristic Web Mining and Asymmetric Alignment. Proceedings of the 47th Annual Meeting of ACL and the 4th IJCNLP of the AFNLP. 2009. pp387-395

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

[5] 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.

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

[7] Feiliang Ren, Muhua Zhu, Huizhen Wang, Jingbo Zho, Chinese-English Organization Name Translation Based on Correlative Expansion. Proceedings of the 2009 Named Entities Workshop, ACL-IJCNLP 2009. pp143-151

[8] Feiliang Ren, Jingbo Zhu, Huizhen Wang. Translate Chinese Organization Names Using Examples and Web. Proceedings of 2009 IEEE International Conference on Natural Language Processing and Knowledge Engineering. 2009. pp83-89.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

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

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

[15] 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.

[16] Tong Xiao, Rushan Chen, Tianning Li, Muhua Zhu, Jingbo Zhu, Huizhen Wang and Feiliang Ren. 2009. NEUTrans: a Phrase-Based SMT System for CWMT2009. Proceedings of 5th China Workshop on Machine Translation.

[17] 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.

[18] 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.