International Journal of Database Theory and Application
Vol.9, No.1 (2016), pp.143-154
http://dx.doi.org/10.14257/ijdta.2016.9.1.13

Improving Translation of Organization Names Combining Translation Model and Web Mining

Bin Li [1], Yin Zhou[2], Ning Ma[1], Wuqi Liang[1] and Lulu Dong[1]

[1]*Anhui Radio and Television University, No. 398, Tongcheng Road, Hefei City, Anhui Province, China*
[2]*Hubei Engineering University, No.272, Jiaotong Road, Xiaogan City, Hubei Province, China*
szbinlee@126.com, zhouyin05029@foxmail.com, {Maning, Liangwuqi, Donglulu}@ahou.edu.cn

Abstract

Named entity (NE) translation is a fundamental task in machine translation (MT) and cross-language information retrieval (CLIR). Furthermore, Organization name (ON) translation is the most complex among all the NEs. A novel system for translating ONs from Chinese to English, with a translation model and web resources, is proposed. Firstly, we built a translation model with Chunk. Then query expansion was adopted with the translation model and term-subject co-occurrence. Thirdly, we extracted the Chinese Organization names with English sentences using the method of frequency shifting and adjacency information to find English fragments as translation candidates. Finally, we found the best translation by computing the trustworthiness of all candidates. The experimental results showed that the approach returned a better performance than machine translation-based systems.

Keywords*: Organization name translation; translation model; web mining; query expansion.*

1. Introduction

Name entities (NEs) are important parts of any language and typically include: person names, locations, Organization names (ONs), dates and times, monetary amounts, and so on [1]. NE translation is crucial for effective cross-language information retrieval (CLIR) [2-4], and statistical machine translation (SMT) [5-7]. However, the translation of name entities is not successful enough because of the complexity and particularity of a name entity's structure. At the same time, this research is more concerned with the development of machine-based translation techniques.

The translation of Organization names is more complex than translation of other name entities, such as person names, location names, and so on [8]. It is because ONs may contain person names, location names, and indeed other ONs. For example, both the company name, Microsoft, and the location name, Asia, are included in the name Microsoft Research Asia: the translation of the corresponding names of people, places, or ONs are needed before the translation of this ON. Furthermore, the translation of personal names and location names are different in different ONs. For example, the translation of 'Suzhou' is 'Soochow' in the translation of 'Soochow University' compared with 'Suzhou' in the translation of 'Suzhou Institute of Technology'. In some situations, the translation sequences are different in ONs with the same attributes. For example, the translation sequences are different in 'Construction Bank' and 'Bank of China' although they are both banks.

To solve this problem, web-based translation of ONs is necessary. Then we built a

ISSN: 2005-4270 IJDTA

chunk-based translation model and term-subject expansion techniques to expand the query words. Then a method based on the frequency shift and adjacency information was used to extract candidate translation strings. After that, ordering the candidate translating strings according to the translation characteristic and length characteristic produced the Chinese translation of the requisite ONs.

This paper is organised as follows: Section 2 reviews related work, Section 3 covers the system framework, Section 4 discusses ON chunking, Section 5 introduces the query expansion approach, Section 6 covers the approach used to extract translation candidates, Section 7 presents experimental work, and Section 8 provides conclusions and recommends future work.

2. Relevant Research

In the research into name entity translation, research into corresponding translation of ONs is rare [9]. Stalls and Kevin built a model that used phrase-based machine translation system to translate ONs directly [10]: this model contains three modules, including a translation module, a language module, and a reordering module. The translation procedure contains three steps: first, choose the aligning phrases in the English-Chinese translation of ON pairs according to the alignment algorithm, then find all corresponding English phrases or Chinese phrases with regard to the given Chinese ON or English ON. Finally, find the compound mode with the maximum probability.

Chen, et al. studied the compound mode and transfer regulation for Chinese to English name entities [11]. The most important part of this work is to distinguish transliteration from paraphrases of the ONs: it is very difficult to build transfer regulations for ONs because there are many keywords therein. There are no experimental results pertinent to the application of these regulations.

Zhang, et al. proposed a model based on the corresponding context of given phrases, to translate ONs [12]. This model was built according to the word-based translation of ONs and consists of a lexical mapping model (LMM) and permutation model (PM). This method divided the procedure of ON translation into two steps. In the first step, it translates the Chinese phrase behind the participle into an English phrase. This procedure translated the Chinese phrase into English phrase through a lexical mapping model. In the second step, adjusting the order of English word sequences is undertaken by using a permutation model. These two procedures are N-gram models which proceed in phrase units. The performance of the method is superior to traditional translation methods based on statistical machine translation.

Chen and Zong investigated a structural formulation of ONs and presented a hierarchical structure-based ON translation model for Chinese-English translation [13]. First, it divides the ONs into three types of chunks. Then it translates the chunks according to synchronous context with free grammars and reorders the sequence of the phrases. In the end, it reorders the sequences between chunks. This method is superior to traditional systems as far as the complexity and performance go.

Fan, et al. proposed a system for translating ONs with the assistance of web resources. They adopted a chunking-based segmentation method to improve the segmentation of Chinese ONs, and then a heuristic query construction method to construct an efficient query to search bilingual Web pages. Finally, they aligned the Chinese ON with English sentences using the asymmetric alignment method to find the best English fragment as the translation equivalent [9].

Lee and Hwang proposed bootstrapping entity translation on weakly comparable corpora. The method is tolerant to asymmetry often found in comparable corpora, by distinguishing different semantics of relations of entity pairs to selectively propagate seed entity translations on weakly comparable corpora [14].

Sellami, et al. mined ON translation from non-parallel corpora using three main steps: firstly, slightly aligning the noisy parallel corpus on a sentence level, secondly, selecting candidate ON translations based on sentence alignment, and thirdly, filtering bad translation pairs based on the context's POS_overlap, type of ON and distance between ON pair translations [15].

Liu analysed several methods of named entity translation, and concluded that web mining can acquire the translation of high-frequency named entities and that it is difficult to translate low-frequency named entities [16].

3. The Framework Structure

The framework for ON translation, based on a translation model and web resources, is shown in Fig. 1. In the first place, training ON translation models according to existing ON translation pairs resources to find an improved strategy based on transverse expansion of the alignment anchor for traditional machine translation alignment results saw us build an internal vocabulary alignment model. The main function of the chunk extraction module is the division of the aligning ONs according to pre-defined forms, which then recorded the relative position information and translation probability information. If the translating ON entities are given, the main procedures of the translating system are the following three steps: first, obtain the corresponding query expansion sets of ON entities, second, search the web resources using the new search words containing their expansion and then obtain the abstract resources of Chinese-English bilingual translation, and finally, extract translation candidates of ON entities from the Chinese-English bilingual resource and order them according to reliability.

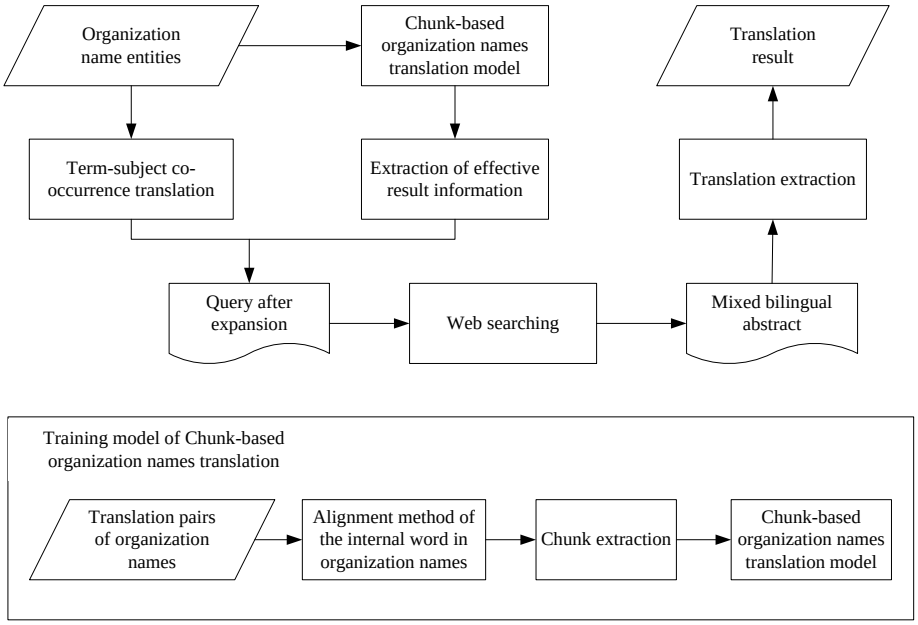

Figure 1. The Framework for ON Translation based on the Translation Model and Web Resources

4. ON Translation by Chunk-based Model

Research into the existing ON dictionary allowed the team to conclude that the organising form of the ONs definite. It usually contains area or range modifiers, an ordinal modifier, a general modifier, modifiers on behalf of area or function and keywords, *etc*. Furthermore, the repetition rate of range modifier and keywords in the

corpus is higher than the other modifiers.

In most situations, it is not necessary to adjust the word orders in the translation of company names. It is more important that the translation of transliteration entities in entrepreneur names be done. The solutions of problems were based on the aforementioned transliteration methods. It usually adjusts the word order in the process of translating administrative ONs. It was further concluded that a chunk is the minimum unit in word aligning administrative ON translations, and the frequency of reordering between chunks was higher than that in chunks.

4.1. The ON Internal Alignment Method based on Transverse Expansion of the Aligning Anchor

According to existing research into the structure and translation attributes of current ON, the internal words within Chinese ONs are notional words. They are translated into one or more English words. The words in English ONs are also notional names except words such as: 'of', 'with', 'the', 'and', 'for', *etc*. Furthermore, the word alignment is shown to be a lumpish alignment structure within a given ON. To build an internal alignment method of ONs based on the transverse expansion of the aligning anchor, we found the probability of an internal aligning of word strings arising and then selected the global optimum aligning method.

First, we processed the word to a bit operation of Chinese to English translation of ONs with GIZA++ alignment tools which are usually used in machine translation. Both Chinese to English, and English to Chinese, translations were considered. At most, one corresponding English word was permissible with each Chinese word in the process of English to Chinese alignment with GIZA++ tools. Similarly, each English word corresponded to only one Chinese word in the alignment of Chinese to English translation. The aligning anchor is a pair of Chinese and English words that aligned exactly in both directions. To optimise the word alignment result on the basis of the first step used in this experiment, a new method was proposed.

The first step was the same as the method above: it then found its aligning anchor based on the intersection set of the aligning results in both directions. Secondly, the procedure of extracting candidate strings involved expanding in transverse directions based on every anchor obtained in the last step until the next aligning anchor, and then adding the expansion words to the current anchor as the candidate strings. Thirdly, we calculated the translation reliability of bilingual word strings, and finally, for each naming entity translation pair, obtained the optimum alignment results with a greedy algorithm. A detailed explanation of these algorithms is given next.

4.1.1. Calculation Method for Translation Reliability: We used a method similar to the tf-idf method to mark the translating segment. For example, the translation reliability of a given Chinese string 'o' and English string 'e' can be calculated thus:

$$score(\{o,e\}) = f(\{o,e\}) \times idf(o) \times \log_2(|o|+1) \quad (1)$$

Where $f(\{o,e\}) = \frac{tf(\{o,e\})}{\max_i(tf(\{o_i,e\}))}$ represents the co-occurrence frequency of 'o' and 'e', $idf(o) = \log_2 \frac{N}{df(o)}$ represents the number of 'e' which is the complementary translation of

'o'. $\log_2(|o|+1)$ is the punishment parameter of the length of the Chinese equivalent, Chinese segment 'o' is the translation of English segment 'e', and *N* represents the number of categories of all English entities.

4.1.2. The Acquisition Algorithm of Optimum Alignment: On the basis of calculating probabilities of every pair of candidate Chinese string 'c' and English string 'e', we used a greedy strategy to find the optimum alignment results. The whole procedure consisted of six steps.

- Extract all { c,e } pairs contained in this entity when naming some special entity.
- Order the pairs, by score, in descending order and then preserve them in the scoreArray set.
- Delete the first element{ cc,ee }from scoreArray set, and refresh this name entity based on { cc,ee } 's corresponding bit.
- Delete all { cc,* } and { *,ee } elements in scoreArray.
- Repeat the second step until scoreArray is an empty set.
- Otain the optimum corresponding bit of each name entity pair.

4.2. The Chunk-based Translation of ONs

We built a translation model that used the chunk as its translation unit. The most important part was the extraction of candidate chunks, calculation of probabilities and the translation decoding algorithm based on the context-free method.

In this experiment, the method of translating ONs is based on synchronous context-free grammars. The ONs consist of keywords, area or range modifiers, and the other modifiers. In the first place, we divided the pairs of each aligning ON entity into three parts, and then deduced the position information by keeping the former two parts in the whole name entity. We got a series of deduction regulations and corresponding reliabilities. The translation procedure contained two steps: the first was called division of chunks, which meant that we divided the given ONs into three parts. The second part was the translation of each entity by deduction. The translation order was area or range modifier part, keyword part, and finally the other modifier. If there was no training corpus, we mixed the traditional machine translation method with a transliteration method.

For example, the 'National Committee of Institute Safety' is extracted as three rules after the training process. Rule one, 'National', rule two, 'Committee of', rule three, 'Industry Safety' and a translation probability of reliability deduced.

The translation procedure of 'National Committee of Institute Safety' is described as follows: divide this name entity into area or range modifiers (National), other modifiers (Safety), keywords (Committee of): the whole process consists of three steps:

- Use rule one, 'National Committee of Institute Safety, National';
- Use rule two, 'National Committee of Institute Safety, National Committee of';
- Use rule three, 'National Committee of Institute Safety, National Committee of Institute Safety'.

5. Construction of the Query Expansion Method

The construction of a query expansion considers both external characteristics and internal characteristics. We made the effective information of extraction translation results into the internal characteristic of the vocabulary. At the same time, taking the combined co-occurrence words as the external characteristic, this method can obtain an effective bilingual abstract resource because it considered both the internal characteristics of an ON

entity and the co-occurrence information of emerging web pages. Furthermore, errors were unavoidable when identifying ON entities because of the shortage of abstract words. In this experiment, we extracted the translations directly from bilingual abstracts. In the extraction process, the translation information, length information, and transliteration information of candidate strings were all considered, then the candidate translation string with the highest total scores was output. The weighted probability algorithm, combined the co-occurrence thesaurus translation to construct query expansion, and was adopted for extraction of effective information from the translation results.

The selection of query expansion exerts a significant influence on the quality and quantity of bilingual resources. Through analysing the searched abstract results after expansion, we found that its quality had improved compared to the returned results searching with resources only. It almost contained the correct translation of ON entities.

According to current research into query expansion, we adopted both the translation results and co-occurrence thesaurus translation to construct query expansion in ON entity translations.

5.1. Method based on Thesaurus Translation

To construct the query expansion method based on thesaurus translation, we need to submit the source search word to a search engine to obtain the abstract information of the source language in the first place. In the second step, we extract the co-occurrence subject vocabulary of source searches from the obtained abstract information. In this step, the TF-IDF method is necessary. After obtaining the subject vocabulary, we will search for a translation of these subject vocabularies from a bilingual dictionary as the final expansion sets of this method [17, 18].

5.2. Searching Construction: ON Translation Results

In the construction process, we evaluated certain statistics about the smallest N translation units with the greatest weighted probability in the top-N translation results. This set formed the query expansion set. The calculation of weighted frequency probability was as follows:

$$weight(c) = N \times \sum_{i=1}^{N} \delta(i) \frac{p(T_i \mid \alpha)}{\sum_{i=1}^{N} p(T_i \mid \alpha)} \tag{2}$$

Where N is the number of transliteration result, T_i is the i*th* translation result of α, $p(T_i \mid \alpha)$ is the reliability of the i*th* translation result, c represents a certain Chinese word or expression, and:

$$\delta(i) = \begin{cases} 1 & if\ c\ exists\ in\ T_i \\ 0 & else \end{cases}.$$

6. Extraction of Translation Results

In this effective query expansion method, we need to obtain the bilingual webpage that consists of ON entity translations, but unavoidable errors will be introduced in the process of identifying these ON entities. Therefore, it was impossible to identify the ON entities of the bilingual web pages. The solution involved extracting the ON translation structure. In the first step, we extracted the candidate translating string combining the frequency

shift with adjacency information, and then calculated the translation similarity, co-occurrence information, length information, and transliteration information between candidate translation strings and entities to be translated respectively. Finally, we considered the scores of several characteristics and output the translation sequence according to the sequence of combined scores.

6.1. Extraction of Candidate Translation String

In this experiment, adopting frequency shift and adjacency information to extract candidate translating string [19], the calculation was as follows:

$$R(s) = \frac{left\_n(s) \times freq(s) \times right\_n(s)}{1 + \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2}} \tag{3}$$

Where s is an expression composed of several words, freq(s) represents the frequency of occurrence of s, x_i represents the frequency of an arbitrary word in s, $\overline{x}$ represents the average frequency of all words in s, left_n represents the total number of different words adjoining the left-hand side of s, right_n represents the total number of different words adjoining the right-hand side of s. We judged whether the left and right adjoining words of a candidate string constructed a translation string of an ON by calculating R(s). If the answer is yes, we constructed a new candidate string and then added it to the candidate string sets.

6.2. Sequencing of Candidate Translation Strings

For the candidate translation string remaining after extraction, we calculated its reliability, and then choose the best one as the translation, as follows:

$$score(s) = \lambda \times w\_translation(s) + (1 - \lambda) \times w\_length(s) \tag{4}$$

Where s represents an English candidate translation string, w_translation(s) represents the translation characteristic value and w_length(s) represents the length characteristic value of s, and is the weight of the translation characteristic value. The calculation of w_translation(s) and w_length(s) was as follows:

$$w\_translation(s) = \frac{length(s')}{length(s)} \sum_{q \in Q} \delta(q,s) \times score(q) \tag{5}$$

$$\delta(q,s) = \begin{cases} \dfrac{length(q)}{length(s)}, & if\ s\ exists\ in\ q \\ 0, & else \end{cases} \tag{6}$$

Where Q is the set of query expansion strings, length(s) represents the length of a candidate translation string, it is also the number of English words; $length(s')$ represents the number of English words corresponding to the Chinese ON in each candidate translation string; and length(q) represents the length of q (also the number of words.

The similarity of Chinese ONs and English translations is seen in the length characteristic. The greater the comparability of ONs in two languages, the greater the

length value (*i.e*. the greater the probability of mutual translation of the ON in two languages). The length characteristic is given by:

$$w\_length(s) = \frac{\min(|s|,|c|)}{\max(|s|,|c|)} \quad (7)$$

Where |s| represents the number of words in a candidate translation string, and |c| represents the number of words in the Chinese ON.

7. Results and Analysis of Experimental Data

In this experiment, the corpus is LDC2005T34. It contains two sub-sets: the entrepreneur ON corpus (54,747 pairs) and the administration ON corpus (30,800 pairs). We choose 68,438 pairs from them as the training data; the remaining 17,109 pairs formed the test data set.

7.1. Performance Test: Alignment Methods

To examine the effect on translation performance arising from the alignment and anchor left-right expansion method, we adopted a heuristic expansion alignment method based on GIZA++. We made the alignment training set, obtained through the above two methods respectively, as inputs, and then we evaluated the key statistics of the translation model system based on various expressions. We then adopted two translation measure methods statistically assess each result respectively. MODEL_1 required the words and the sequence of the translation results was the same as the correct answer, and MODEL_2 only required the words of the translation results to be the same as the correct answer. The statistical results are shown in Table 1.

From Table 1, the accuracy rate of MODEL_2 was almost twice that of MODEL_1. At the same time, the alignment anchor method has improved the result by 3.4% and 6.3% compared with the GAZA++ method in MODEL_1 and MODEL_2 respectively.

Table 1. The Statistical Translation Results with Different Alignment Methods

	Alignment anchor left-right expansion method	Heuristic expansion method based on GIZA++
MODEL_1	23.2% (3969 pairs)	19.8% (3385 pairs)
MODEL_2	43.5% (7443 pairs)	37.2% (6365 pairs)

7.2. Comparison of Query Expansion Methods

Here, we defined the search efficiency by calculating the number of correct translation result segments in a certain number of webpage segments. The more segments with correct translation results, the higher the search efficiency. There was a significant difference in search efficiency with different query expansion construction methods [20]. Usually, the average inclusion rate is used for evaluating the search efficiency, as follows:

$$Inclusion\_rate = \frac{1}{N}\sum_{i=1}^{N}\frac{D_i}{S_i} \quad (8)$$

Where Si represents the number of abstracts of the obtained webpage in ith search, Di represents the number of webpage segments for which a correct translation exists at least once. In this experiment, we chose 1,000 ONs to form the test set, and then used four search construction methods to obtain different keywords, and finally we submitted the result to a search engine to obtain the former 100 returned webpage segments.

Method-1: only use Chinese ONs as search keywords

Method-2: use both Chinese ONs and the translation of co-occurrence thesaurus expansions (Chinese-English general dictionary) as a query expansion.

Method-3: use both the Chinese ONs and the translation of the co-occurrence thesaurus expansion (the dictionary constructed based on optimum alignment result in this article) as a query expansion.

Method-4: merge the chunk-based translation model with the basis of the third method. The average inclusion rate with different expansion methods is shown in Table 2.

Table 2. Comparison Average Inclusion Rates from Different Expansion Methods

	Method-1	Method-2	Method-3	Method-4
Average inclusion rate	4.3%	13.4%	27.3%	**30.6%**

We examined data from all test sets with the above four search construction methods. At the same time, we chose translation result obtained by statistical machine translation techniques based on expressions as a comparison. In this experiment, we chose the TOP1, TOP5, TOP10, and TOP15 candidate translation strings to analyse translation accuracy. We found that the translation accuracy was greatest when the weight of the translation characteristic value was 0.96. The results of the experiments are shown in Figure 2 which shows that the translation accuracy was improved as the choice of candidate translation string improved for each method. For example, the translation accuracy of TOP15 increased compared with TOP1, TOP5 and TOP10. The translation accuracy of the forth method was the highest: this is explained in four ways.

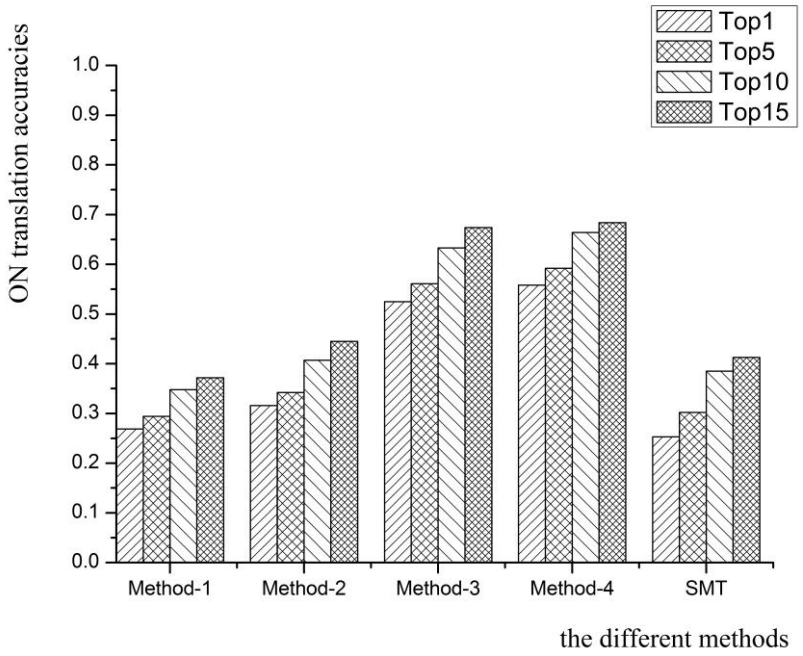

Figure 2. ON Translation Accuracies

It increased the reliability of obtaining bilingual webpage candidates because of its expanded search.

The different expansion dictionary used an optimum alignment result. Its coverage and scale were larger than that of a general dictionary. At the same time, it included some special nouns pairs. Furthermore, the ambiguity of a word was smaller than in the general dictionary because the words are from the training corpus.

We merged the co-occurrence term-subject expansion method with the chunk-based method and built a higher pertinence query expansion of the test corpus to improve the accuracy of bilingual abstraction of webpage search data.

Compared with expression-based statistical machine translation techniques, the web-based mining method was insensitive to word order when searching web pages. It meant that this search satisfied the requirements of the candidate translation string if the word translation of a given query expansion was included in the searched web pages. There was no rigorous requirement imposed upon the position of a translation within the word expressions. Furthermore, it avoided the translation errors due to the usage of synonyms in the expression-based machine translation results.

8. Conclusions and Future Work

The translation of name entities is important for machine translation and cross-language information searching. At the same time, ON translation is more difficult compared with other name entities. In this study, we introduced a web-based Chinese to English ON translation method which was a multi-query expansion and merging method. In the first place, we adopted an internal word alignment method, based on the anchored left-right expansion of ONs, to extract a bilingual dictionary and construct query expansions based on co-occurrence term-subject translations. At the same time, we constructed query expansions by chunk-based ON translation, and then combined the two query expansion methods. From the result, it was found that the inclusion rate of existing translation strings in the translation result had been improved largely through the use of query expansion techniques. Thereafter, we extracted the candidate translating string in

bilingual web pages through frequency shift and adjacency information. It indicated that the TOP1 translation accuracy of this method had been improved by 30.5% over an expression-based statistical machine translation technique.

There were also some inefficiencies in this study. Some improvements should be implemented with regard to the following.

- Try to merge other query expansion methods the better to obtain bilingual abstracts of web pages to enhance the quality of translation string abstracts.
- Try to use linguistical characteristics (such as lexical category, grammar, and so on) as candidate unit models to improve the quality of candidate translation strings.
- Adopt more characteristics as translating selection models besides translation characteristic and length characteristic, and obtain the optimum translation by comparing the composite results of several characteristics.

Acknowledgment

This work was supported by the Key Programme of the Foundation for Young Talents in the Colleges of Anhui Province under Grant 2013SQRL097ZD; the Key Programme of the Natural Science Foundation of Anhui Educational Committee under Grant KJ2014A081; and the Foundation for Young Talents in Anhui Radio & Television University under Grant qn11-19.

References

[1] I. Ahmed and S. R. Named Entity Recognition by Using Maximum Entropy. International Journal of Database Theory and Application, vol. 7, no. 1, **(2014)**, pp. 1-10.
[2] R. Steinberger, B. Pouliquen, J. Hagman. Cross-lingual document similarity calculation using the multilingual thesaurus EUROVOC. Cicling, **(2002)**, pp. 415-424.
[3] R. Udupa, K. Saravanan, A. Bakalov, A. Bhole. "They Are Out There, If You Know Where to Look": Mining Transliterations of OOV Query Terms for Cross-Language Information Retrieval. Advances in Information Retrieval, Springer Berlin Heidelberg, **(2009)**, pp. 437-448.
[4] M. Potthast, B. Stein, M. Anderka. A Wikipedia-Based Multilingual Retrieval Model. Advances in Information Retrieval, Springer Berlin Heidelberg, **(2008)**, pp. 522-530.
[5] P. F. Brown, S. A. D. Pietra, V. J. D. Pietra, R. L. Mercer. The Mathematics of Statistical Machine Translation: Parameter Estimation. Computational Linguistics **(1993)**, pp. 263-311.
[6] D. Chiang. Hierarchical Phrase-Based Translation[J]. Computational Linguistics, vol. 33, no. 2, **(2007)**, pp. 201-228.
[7] H. Hassan, J. Sorensen. An Integrated Approach for Arabic-English Named Entity Translation. Proceedings of the Acl Workshop on Computational Approaches to Semitic Languages, **(2005)**, pp. 87-93.
[8] B. Li, Y. Zhou, N. Ma, L. L. Dong, W. Q. Liang. Chinese-English Translation of Organization Names Based on a Translation Model and Web Mining. Proceedings of the 3rd International Conference on Computer and Computing Science, Hanoi, Vietnam, **(2015)** October 22-24.
[9] Y. Fan, J. Zhao, K. Liu. A Chinese-English Organization Name Translation System Using Heuristic Web Mining and Asymmetric Alignment. Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, Association for Computational Linguistics, **(2009)**, pp. 387-395.
[10] B. G. Stalls, K. Kevin. Translating names and technical terms in Arabic text. Proceedings of the Workshop on Computational Approaches to Semitic Languages. Association for Computational Linguistics, **(1998)**, pp. 34-42.
[11] H. H. Chen, C. Yang, Y. Lin. Learning formulation and transformation rules for multilingual named entities[C]// Proceedings of the ACL 2003 Workshop on MMLNER **(2003)**, pp.1-8.
[12] M. Zhang, H. Li, J. Su, H. Setiawan. A Phrase-Based Context-Dependent Joint Probability Model for Named Entity Translation. Natural Language Processing – IJCNLP 2005, Springer Berlin Heidelberg, **(2005)**, pp. 600-611.
[13] Y. Chen, C. Zong. A Structure-Based Model for Chinese Organization Name Translation. ACM Transactions on Asian Language Information Processing, vol. 7, no. 1, **(2008)**, pp. 1-30.
[14] T. Lee, S. W. Hwang. Bootstrapping Entity Translation on Weakly Comparable Corpora. Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, **(2013)**, pp. 631–640.
[15] R. Sellami, L. B. Hadrich, F. Sadat. Mining Named Entity Translation from Non Parallel Corpora[J]. Proceedings of the Twenty-Seventh International Florida Artificial Intelligence Research Society Conference, **(2014)**, pp. 219-224.
[16] Y. Liu. The Technical Analyses of Named Entity Translation. 2015 International Symposium on

Computers & Informatics. Atlantis Press, **(2015)**, pp. 2028-2037.
[17] Y. D. Ge, Y. Hong, J. M. Yao, Q. M. Zhu. Improving Web-Based OOV Translation Mining for Query Translation. Information Retrieval Technology. Springer Berlin Heidelberg, **(2010)**, pp. 576-587.
[18] J. H. Wang, J. W. Teng, P. J. Cheng, et al. Translating Unknown Cross-Lingual Queries in Digital Libraries Using a Web-based Approach. Digital Libraries, Joint Conference on ACM, **(2004)**, pp. 108-116.
[19] C. Lu, Y. Xu, and S. Geva. Web-based query translation for english-chinese CLIR. Computational Linguistics and Chinese Language Processing, **(2008),** pp. 61-90.
[20] P. J. Cheng, J. W. Teng, R. C. Chen, J. H. Wang, W. H. Lu, L. F. Chien. Translating unknown queries with web corpora for cross-language information retrieval. Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval ACM, **(2004)**, pp. 146-153.

Authors

Bin Li, he is a lecturer at Anhui Radio and Television University. He has lead or participated in several research projects, including the National Natural Science Foundation of China, the Key Programme of the Foundation for Young Talents in the Colleges of Anhui Province, the Key Programme of the Natural Science Foundation of Anhui Educational Committee and etc. His research interests include data mining, machine translation, recommended system and so on.

Yin Zhou, she is a lecturer at College of Economic and Management of Hubei Engineering University. Her research interests include data mining, sentiment analysis and so on.

Ning Ma, he is a lecturer at Anhui Radio and Television University. He has authored many papers in journals, conferences and so on. His research interests include data mining, neural network.

Wuqi Liang, he is an Associate Professor at Anhui Radio and Television University. He has authored many papers in journals and lead or participated in several research projects. His research interests include data mining, text classification and so on.

Lulu Dong, she is a teaching assistant at Anhui Radio and Television University. Her research interests include data mining, text classification.

