Advanced Science and Technology Letters
Vol.111 (COMCOMS 2015), pp.211-216
http://dx.doi.org/10.14257/astl.2015.111.41

Chinese-English Translation of Organization Names Based on a Translation Model and Web Mining

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

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

Abstract. Organization name (ON) translation is the most complex among all the Named entities (NEs). A novel system for translating ONs from Chinese to English, with a translation model and web resources is proposed. Firstly, a translation model was built with Chunk and query expansion was adopted with the translation model and term-subject co-occurrence. Secondly, 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 get the best translation by computing the trustworthiness with candidates. The experimental results show that the approach reaches a better performance than machine translation-based system.

Keywords: Organization name translation; translation model; web mining.

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. NE translation is crucial for effective cross-language information retrieval (CLIR)[1], and statistical machine translation (SMT)[2]. 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. It is because ONs may contain person names, location names, and indeed other ONs.

In the research into name entity translation, research into corresponding translation of ONs is rare[3]. Stalls and Kevin built a model that used phrase-based machine translation system to translate ONs directly[4]. Chen, *et al.* studied the compound mode and transfer regulation for Chinese to English name entities[5]. 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. Zhang, *et al.*

ISSN: 2287-1233 ASTL

proposed a model based on the corresponding context of given phrases, to translate ONs[6]. This model was built according to the word-based translation of ONs and consists of a lexical mapping model (LMM) and permutation model (PM). Chen and Zong investigated a structural formulation of ONs and presented a hierarchical structure-based ON translation model for Chinese-English translation[7]. Lee and Hwang proposed bootstrapping entity translation on weakly comparable corpora[8].

2 ON Translation by Chunk-based Model

We have a research on the existing organization names dictionary and we conclude that the organizing form of organization names is 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. 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.

2.1 The ON Internal Alignment Method Based on Transverse Expansion of The Aligning Anchor

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. And 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.

2.2 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.

3 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. We adopted both the translation results and co-occurrence thesaurus translation to construct query expansion in ON entity translations.

3.1 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. Then, we will search for a translation of these subject vocabularies from a bilingual dictionary as the final expansion sets of this method.

3.2 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:

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

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

4 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.

5 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. In this experiment, we used four search construction methods to obtain different keywords.

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.

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, and TOP10 candidate translation strings to analyse translation accuracy. The results of the experiments are shown in Fig. 1 which shows that the translation accuracy was improved as the choice of candidate translation string improved for each method.

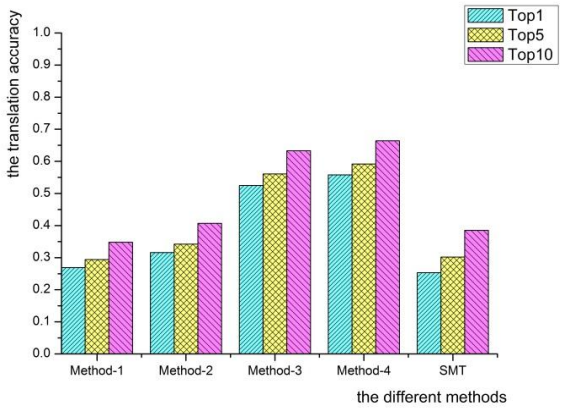

Fig. 1. ON translation accuracies

For example, the translation accuracy of TOP10 increased compared with TOP1 and TOP5. The translation accuracy of the forth method was the highest: this is explained in

four ways. 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.

6 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. 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. 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.

Acknowledgement. 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.

References

1. 1 R. Steinberger, B. Pouliquen, J. Hagman. Cross-lingual document similarity calculation using the multilingual thesaurus EUROVOC. Cicling, (2002), pp. 415-424.
2. D. Chiang. Hierarchical Phrase-Based Translation [J]. Computational Linguistics, (2007, 33(2)), pp. 201-228.
3. Y. Fan, J. Zhao, K. Liu. A Chinese-English Organization Name Translation System Using Heuristic Web Mining and Asymmetric Alignment [J]. ACL Proceedings of the Joint Conference of Annual Meeting of the ACL & Intern, (2009), pp. 387-395.
4. 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.
5. 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.
6. M. Zhang, H. Li, J. Su, et al. A Phrase-Based Context-Dependent Joint Probability Model for Named Entity Translation. Natural Language Processing – IJCNLP 2005, (Springer Berlin Heidelberg, 2005), pp. 600-611.
7. Y. Chen, C. Zong. A Structure-Based Model for Chinese Organization Name Translation. ACM Transactions on Asian Language Information Processing, (2008, 7(1)), pp. 1-30.
8. 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.

