A Practical Chinese-English ON Translation Method Based on ON's Distribution Characteristics on the Web

Feiliang REN
Northeastern University, Shenyang, 110819, P.R.China

renfeiliang@ise.neu.edu.cn

ABSTRACT

In this paper, we present a demo that translate Chinese-English organization name based on the input organization name's distribution characteristics on the web. Specifically, we first experimentally validate two assumptions that are often used in organization name translation using web resources. From experimental results, we find out several distribution characteristics of Chinese organization name on the web. Then, we propose a web mining method based on these distribution characteristics. Experimental results show that our method is effective. It can improve the inclusion rate of correct translations for those Chinese organization names whose correct translations often occur on the web, and it can also improve the BLEU score and accuracy for those Chinese organization names whose correct translations rarely occur on the web.

KEYWORDS : organization name translation, distribution characteristics, web resource, machine translation

Proceedings of COLING 2012: Demonstration Papers, pages 377–384,
COLING 2012, Mumbai, December 2012.

1 Introduction

Named entity (NE) translation is very important to many Natural Language Processing (NLP) tasks, such as cross-language information retrieval and data-driven machine translation. Generally NE translation has three main sub-tasks which are person name (PER) translation, location name (LOC) translation, and organization name (ON) translation. And ON translation is attracting more and more research attention.

There are a large amount of resources on the web, thus researchers usually assume that for every NE to be translated, its correct translation exists somewhere on the web. Based on this assumption, recent 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. And the performance of ON translation using web resources depends greatly on the solution of the following problem: how can we find the web pages that contain the correct translations effectively? Solving this problem usually involves some query construction methods. Some researchers (Huang et al., 2005, Yang et al., 2009, Ren et al., 2009, and so on) prefer to constructing bilingual queries to find the web pages that contain the correct translations. They further propose another assumption that both the input NE and its correct translation exist somewhere in some mix-language web pages. Based on this assumption, they think bilingual queries are the most useful clues that can be used to find these web pages.

These two assumptions are essential for those ON translation methods using web resources. Their validity will determine the validity of those ON translation methods using web resources. So we think it is very necessary to validate these two assumptions experimentally. However, to our best knowledge, there are no related works on the research of validating these two assumptions.

In this paper, we focus on the following two issues. The first issue is to experimentally validate these two assumptions. The second issue is to propose an effective web mining method for Chinese ON translation.

2 Our Basic Idea

In this section, we carried out some experiments to validate these two assumptions. We use some bilingual ON translation pairs as test data to validate whether these ON translation pairs and their monolingual ON parts can be easily found on the web. The test data are extracted from LDC2005T34. In order to analyze the validity of these two assumptions thoroughly, we divide these ON translation pairs into different groups according to the keyword types of their Chinese parts. And 20 groups that have the maximal amount of ON translation pairs are selected as final test data. In our experiments, following three kinds of queries are constructed for every ON translation pair in the test data:

Q1: only the Chinese ON.

Q2: only the English ON.

Q3: the ON translation pair.

We obtain at most 10 web pages from Bing (http://www.bing.com/) for every query, and compare the inclusion rates of different test groups respectively. For Q1 and Q2, the inclusion rate is defined as the percentage of ON translation pairs whose queries are completely contained

in the returned web pages. And for Q3, the inclusion rate is defined as the percentage of ON translation pairs whose Chinese parts and English parts are both contained in the returned web pages. And the experimental results are shown in table 1.

Keyword types	Num	inclusion rate(%)			
		Q1	Q2	Q3	Q3'
committee	3444	77.09	72.97	13.30	20.76
company	2315	65.49	61.17	7.17	12.53
factory	971	67.35	47.79	11.23	13.80
college	572	95.80	86.19	29.20	34.97
institute	531	85.31	71.37	18.27	26.55
center	467	75.16	51.61	8.99	12.21
bureau	409	83.86	81.42	16.38	24.21
agency	349	75.93	73.07	17.48	21.49
university	294	97.28	94.22	47.28	62.93
ministry	258	82.56	83.33	16.67	26.74
bank	180	81.67	88.33	28.89	33.89
party	137	85.40	89.05	18.25	26.28
organization	131	67.94	83.97	12.98	21.37
restaurant	126	81.75	92.86	30.95	48.41
union	125	72.00	87.20	16.00	22.40
group	123	51.22	70.73	4.07	6.50
school	98	82.65	45.92	6.12	8.16
area	76	67.11	85.53	23.68	27.63
team	73	83.56	89.04	12.33	16.44
hospital	56	92.86	71.43	28.57	39.29
Total	10735	75.81	69.91	14.49	20.75

TABLE 1. INCLUSION RATE COMPARISONS

From table 1 we can draw following conclusions.

The first one is that the correct translations for most of Chinese ONs do exist on the web, but more effective clues are needed to find them. Besides, experimental results also tell us that only bilingual query is not enough to find the web pages that contain the correct translations. This conclusion can be further confirmed by another experiment whose results are denoted as Q3' in table 1. The query construction method for Q3' is the same as the method for Q3, but the definition of inclusion rate between them is different. And the inclusion rate for Q3' is defined as the percentage of ON translation pairs whose English parts are contained in the returned web pages. In fact, the experimental results of Q3' are the upper bound of the correct translation inclusion rate that can be obtained by using bilingual queries. But they are still far lower than the results of Q2, which is the true inclusion rate of the correct translations.

The second conclusion is that the inclusion rates for different types of Chinese ON are different greatly. To further investigate this conclusion, we pick out some ON translation pairs that have

the highest inclusion rate for Q2 and Q3'. We find out that most of the ONs in these ON translation pairs are multinational companies, government agencies, research institutes, university names and so on. The ONs in these ON translation pairs often occur on the web and the available web resources for them are huge. For such kind of an ON translation pair, it is easier to find both some monolingual web pages that contain one of its monolingual ONs and some mix-language web pages that contain the source ON part and the target ON part. On the other hand, we also pick out some ON translation pairs that have the lowest inclusion rate for Q2 and Q3'. We find out that the ONs in these ON translation pairs rarely occur on the web directly, and most of the ONs in these ON translation pairs have following two characteristics. Firstly they usually have a lot of modifiers, and the lengths of them are long too. Secondly, there are usually some nested sub-ONs in them. Based on these characteristics, we think if the ON translation pairs that rarely occur on the web were segmented into several small translation pairs (such as chunk translation pairs), the inclusion rate of these small translation pairs would improve. And further experiments confirm our idea. These experimental results are shown in table 2. In table 2, the test data are extracted from the ON translation pairs whose English parts cannot be found on the web. Three types of chunks in the Chinese ON parts are defined as [Chen and Zong, 2008] did, and these chunks are Regionally Restrictive Chunk (RC), Keyword Chunk (KC), and Middle Specification Chunk (MC). The test ON translation pairs are chunked and aligned manually and every chunk translation pair is viewed as a new ON translation pair. From table 2 we find that both the monolingual chunk parts and the chunk translation pairs are easier to be found on the web. We take our above idea as the third conclusion. Moreover, from the inclusion rates for Q1 and Q2 in table 2 we can see that smaller text units are easier to be found on the web, so this conclusion is also suitable to those ON translation pairs that often occur on the web.

Keyword types	Chunk num	Inclusion rate (%)			
		Q1	Q2	Q3	Q3'
committee	2514	100	100	38.46	55.33
company	2248	100	100	26.25	34.48
factory	1065	100	100	41.22	46.2
college	182	100	100	59.34	75.82
institute	441	100	100	63.72	79.59
center	588	100	100	59.35	65.99
bureau	175	100	100	70.86	86.86
agency	263	100	100	65.02	79.85
university	37	100	100	75.68	94.59
ministry	108	100	100	83.33	92.59
bank	44	100	100	61.36	65.91
party	30	100	100	60	66.67
organization	27	100	100	88.89	88.89
restaurant	42	100	100	80.95	88.1
union	44	100	100	77.27	77.27
group	94	100	100	46.81	55.32
school	154	100	100	77.27	81.17
area	25	100	100	64	72

team	23	100	100	60.87	60.87
hospital	40	100	100	57.5	62.5
Total	8144	100	100	42.98	54.15

TABLE 2. INCLUSION RATE OF CHUNK TRANSLATION PAIRS

3 Our Web Mining Method for Chinese ON Translation

3.1 Motivations of Our Method

Based on above analysis, we design a web mining method for Chinese-English ON translation as shown in Fig 1.

Input: a Chinese ON O_c to be translated
Output: a Query set *QS* and a recommended translation result

Algorithm:

1. Segment O_c into RC_c, MC_c, and KC_c,.
2. Generate translation candidates for O_c and the segmented chunks, denote these translation candidates as O_e, RC_e, MC_e, and KC_e. Add "O_c + O_e", "O_c + RC_e", "O_c + KC_e", and "O_c + MC_e" into *QS*. Take RC_e, MC_e, and KC_e as queries respectively and goto step 3.
3. Submit input query to search engine and revise this query according to some rules. Repeat this procedure until the input query cannot be revised any more.
4. Take "RC_e + MC_e", "KC_e + RC_e" and "KC_e + MC_e" as queries respectively and goto step 3.
5. Take "RC_e + MC_e + KC_e", "KC_e + RC_e + MC_e" and "KC_e + MC_e + RC_e" as queries respectively and goto step 3.
6. If there is a query that has been revised in step 5, take it as the recommended translation result. Otherwise, "RC_e MC_e KC_e" is selected as the recommended translation result. Add this recommended result into *QS*.
7. Return *QS* and the recommended translation result.

FIGURE 1. OUR WEB MINING METHOD

In Fig 1, we use NEUTrans [Xiao et al., 2009] system to generate translation candidates for the input. The training corpus for NEUTrans consists of about 370K bilingual sentences that are extracted from the corpora of LDC2005T10, LDC2003E07, LDC2003E14 and LDC2005T06.

In the third step of Fig 1, for a given query q (its original Chinese source text is denoted as q_c) and the returned web pages, one of following rules is used to revise q, and we denote the revised result as q'.

Rule 1: If q is completely contained in the web pages, take q as q'.

Rule 2: If q cannot be completely contained in the web pages, and if we can find such a continuous English text s in a web page that is subjected to following three conditions, take s as q'.

(1) Submit s to search engine and it can be completely contained in the returned web pages.

(2) s has the largest similarity with q. And the similarity is computed with following formula 1.

$$Sim(q,s) = \frac{SameWord(q,s)}{Len(q) + Len(s)} \quad (1)$$

This condition is required to solve the reordering problem in ON translation.

(3) If there is a word w_i in s that does not appear in q, we require that w_i must have at least one dictionary translation item that appears in q_c or w_i must not have any dictionary translation items. This condition is required to solve the out-of-vocabulary (OOV) translation problem and the translation item selection problem in ON translation.

Rule 3: If q cannot be revised by rule 1 and rule 2, take q as q' directly.

4 Experiments

In this section, we evaluate the obtained recommended translation results with the metrics of BLEU score and accuracy. In this experiment, test data consists of 500 Chinese ONs that are randomly selected from those Chinese ONs whose correct translations don't exist on the web. The entire test ONs are chunked manually. Experimental results are shown in table 3.

Num	Average Length	BLEU Score		Top1 Accuracy	
		NEUTrans	Our	NEUTrans	Our
500	3.7 words	0.1811	0.2326	11.6%	19.4%

TABLE 3. RESULTS OF THE SECOND EXPERIMENT

From these results we can see that our method can improve the efficiency of web mining greatly for Chinese ON translation. Compared with the baseline systems, our method obtains higher inclusion rate and higher translation performance.

Conclusions

The main contribution of this paper is that we validate the two assumptions that are often used in ON translation using web resources. Another contribution of this paper is that we find out some distribution characteristics of Chinese ON on the web. These distribution characteristics are very useful for designing appropriate web mining method for Chinese ON translation using web resources. Besides, we propose a novel web mining method based on these distribution characteristic for Chinese ON translation. And experimental results show that our method is effective.

Acknowledgements

This paper is supported by the National Natural Science Foundation of China (Grand No. is 61003159, 61100089, 61073140, and 61272376).

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] 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.

[9] 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.

[10] Hsin-Hsi Chen and Yi-Lin Chu. 2004. Pattern discovery in named organization corpus. Proceedings of 4th International Conference on Language, Resource and Evaluation. pp301-303.

[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.