Study on Word Alignment for Re-ordering of Web-mined OOV Translation Candidates

Shuang LI Meng SUN Yang YANG Jian-Min YAO

Provincial Key Laboratory of Computer Information Processing Technology
Soochow University, Suzhou, China, 215006
lisa.s.li@hotmail.com, jyao@suda.edu.cn

Abstract

Web information retrieval technology has been widespread concerned by researchers. Web-based search of the OOV Translation Mining has also become hot spots. In this paper, the re-ordering of OOV translation candidates is studied, which is the result of web mining. Automatic word alignment technology is used to calculate the weighted points for each candidate then sort the results, with the closely right candidate to be top ranking. The approach can make good performance in such aspects like equal frequency or low frequency circumstances. We take some OOV phrases in different fields as test corpora for web mining, and evaluate out the method on it, and the result is encouraging.

Keywords: Web-based Data Mining; Word Alignment; OOV Translation; Natural Language Processing

1. Motivation of the Study

OOV (out-of-vocabulary)refers to the words and phrases that are not in the dictionary, e.g. named-entities like person, organization and location names, book/movie titles, technical terminology (medical, scientific, technological, military, etc.), and newly-coined words because of social development.

Although natural language understanding (NLU) is far from the final goal, but the relevant NLU techniques can be of help in many applications. In machine translation, cross-lingual information retrieval (CLIR), question-answering (QA) etc, OOV can greatly influence the performance and the translation is also time-consuming and influential for quality. Therefore to take advantage of mature technology of word alignment must have important practical meaningfulness for improving the accuracy rate of OOV web mining.

Exploring abundant web page with bilingual or mixed languages then using some certain algorithm to obtain OOV term translation automatically is a hot topic. Such web-based approach to retrieve OOV words translation is easy to carry out and achieve the high accurate rate. However, the web mined OOV translations will always not be only one, they are commonly sorted by frequency. The order affects the correctness especially when the frequency of translation candidate is same. The right result is hardly discriminated or founded if the number of candidates is enormous. Making use of bilingual word alignment technology will alleviate the difficulty to improve the correct rate.

Related researches [5-7] are using web search engines to return to the first 100 web snippets for statistics. Researches [8-10] solve the issue how to obtain the effective pages that include both Chinese and English. However, these studies are basically using the frequency to sort the translations, while we will study on other effective features for in-depth study of translation mining.

Paper [11] utilizes the frequency plus the average semantic relevance to evaluate and sort the candidate translations. The four features are frequency, distribution in different pages, the distance between the source word and target candidates, and the key words, symbols and boundary information between the source word and target candidates. However, this method is more time-consuming for more computation.

This paper applies the word alignment technology to the web mining of translations. Experiments show that this approach can improve the accuracy of translation mining, especially when equal frequency of the candidates or when the occurrence is quite rare.

The paper is organized as follows. In section 2, a method based on the web search engine is first introduced for web mining of OOV translations. In section 3, a way based on word alignment technology is introduced for re-ordering of web mined translation candidates. Section 4 is the experiment evaluation, in which test suite for evaluation of the translation mining is constructed with technical terms from different areas. Section 5 is an analysis of the translation mining results, which aims to reveal the good and bad aspects of the algorithm. A conclusion is given and further study is proposed.

2. Web Mining of OOV Translation Based on Frequency Statistics

The approach is introduced in our previous work in [13], which in essence is to search for mixed-language web pages containing the OOV and its translation, and then mine the translation via statistical measures. To get the mixed-language web pages, we submit a query containing both the OOV and translation to the longest-subsequence. From the returned snippets we can find both the Chinese OOV and the segment translation, which is a strong hint for the OOV translation. Frequency, mutual information and other measures are used to mine the translation. The process is as follows: 1) Take the Chinese OOV as an initial query; 2) Expand the query with possible English segments based on a bilingual dictionary; 3) Submit the Chinese OOV and the expanded English words to the search engine; 4) Process the return snippets based on statistical measures and word alignment; 5) Output top k translation candidates. The algorithm is described in more details in the following sub-sections, and then we propose our improvement to it in section 3.

2.1. Query Expansion

This web-based translation mining process is composed of query expansion before submitting to the search engine, and translation mining after getting the information retrieval results.

The query involves two parts: the Chinese query (word, phrase or segment to be translated.) and 2) translation of its longest-possible sub-sequence. When a query mixing Chinese and English is submitted to the search engine, many of the returned snippets contain the query or its segments. Some may contain the whole translation of the Chinese query, but most of them are buried among many other English words.

Followed is an introduction to the query expansion algorithm.

ALGORITHM:

Query expansion for translation mining

INPUT: the Chinese query C_Query
OUTPUT: Expanded query composed of the Chinese query and the English query Exp_Query
BEGIN PROCEDURE
Sub_Seq = C_Query
LOOP UNTIL Sub_Seq is NULL
{
Sub_Seq = C_Query – first character of the C_Query
IF (Sub_Seq is in the dictionary)
Exp_Query = C_Query + translation of the Sub_Seq
RETURN Exp_Query
ENDIF
}
END PROCEDURE

Taking "聚合函数" as an example. Sub_Seq is sequentially "聚合函数", "合函数", "函数", the last of which is found in the dictionary with a series of translation of "function…（if there exist others）". The Exp_Query can be the C_Query plus any one of the translations. All the Exp_Query candidates are fed into the search engine to get the returned snippets. From the C_Query plus the various translations, we get an array of Exp_Query, and then an array of returned snippets. The translation mining is carried out on the snippets of the mixed languages.

2.2. Frequency-based Translation Candidates Mining

There may be multiple English expansion words, so we submit each expanded query, which is composed of the Chinese OOV and the expanded English item, to the search engine. For our present example, the pair is "聚合函数 function". Co-occurrence frequency is utilized for the translation mining from the returned snippets. After filtering out the stop words such as function words or some non-linguistic strings, count the frequency of all the English strings, and return the top N most frequent strings as the translation candidates.
In this case, the result is as in the table 1.

Table 1. The candidate translations mined from search engine snippets, sorted in order of frequency , for the query of OOV phrase "聚合函数"

Candidate translations by web mining	Frequency
【 probefunc 】	5
【 functions 】	4
【 funcname 】	2
【 Func func 】	2
【 CSDNBlogfunc 】	1
【 Create or replace function func 】	1
【 create function func 】	1
【 call user funcfunc 】	1
【 Aggregating Functions 】	1
【fgetsphp fgetswhile fgetsc 】	1
【 echo fgets 】	1
【 aggregate functions 】	1

3. Word Alignment for Re-ordering of Translation Candidates

Word alignment is a base technology for machine translation, which involves mapping of the segments in the source language to those in the target language. In this section, we make study on utilization of word alignment to improve translation mining performance.

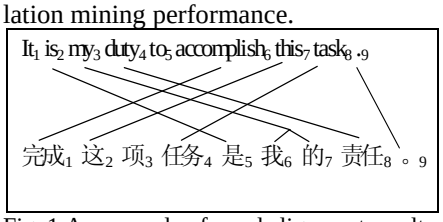

Fig. 1 An example of word alignment results

3.1. Word Alignment Technology

Word alignment refers to the process of discovering the correspondences between the source and target language components. Fig. 1 gives a graphical representa-

tion of the English-Chinese bilingual word alignment results. The right subscript of each word stand figures showing that position of the serial number in the sentence, the connection line between the sentences of bilingual words indicate alignment relations. Denote the English sentence as $ES = e_1, e_2, ..., e_m$, and $CS = c_1, c_2, ..., c_n$ is the sentence for the Chinese, where $e_i (1 \le i \le m)$ is ith word or phrase adjacent of English sentence, and $c_i (1 \le i \le n)$ is ith word in Chinese, while m is the length of the sentence in English, n is the length of sentence for the Chinese. An aggregate can be used to express the ES and CS alignment results, $A_i (1 \le i \le k)$ for each of these terms for a right match, with the unit doublet $(A_{ie} : A_{ic})$ A_{ie} is the serial number for the aggregate of English words location, A_{ic} is location of the serial number of Chinese words for the aggregate, express that A_{ie} the corresponding English words and the A_{ic} corresponding Chinese translation are the pairs to each other. A_{ie} 、 A_{ic} not empty at the same time. If, each of A_{ie} 、 A_{ic} is blank, then the match is empty words alignment. Fig. 1 in (1: NULL), (5: NULL), (NULL: 3) is empty alignment. To facilitate that, the paper assumes that only non-blank alignment is given. Fig. 1 Word alignment set is as follows:

It1 is2 my3 duty4 to5 accomplish6 this7 task8 .9

完成 1 这 2 项 3 任务 4 是 5 我 6 的 7 责任 8 。 9

{(2:5), (3:6,7), (4:8), (6:1), (7:2), (8:4), (9:9)}

Word alignment systems usually derive from either statistical approaches or linguistic ones (usually based on a bilingual dictionary), or a combination of both. Statistical methods involve co-occurrence measures and probability scores, and are especially effective on large corpora with high-frequency words but performance decreases with low-frequency occurrences. Linguistic ones make use of information such as syntactic parsing. They are less robust despite being able to deal with low-frequency words. Hybrid approaches seem to be a good compromise. Entirely based on bilingual dictionary, non-blank alignment aggregate are aligned with very high accuracy rate (96.16%). However the real translation in the context with the diversity and flexibility, according to the terms in dictionary, the relative alignment coverage is lower (55.37 percent). In order to increase the coverage, draw into the similarity of the vocabulary and semantic-based alignment methods.

Bilingual dictionary provides a reliable translation of information on vocabulary and make full use of information from the translation will help improve the precision of statistics. Put the known results from alignment into the calculation of statistical probability, and with word alignment, finally comes out the word alignment set in Fig. 1.

3.2. Re-order translation candidates

A pre-processing step is first implemented. The queries are submitted to a search engine. Then the key words are processed with web mining, i.e. sort the candidate translations by frequency. Then a further processing of re-ordering of translation candidates is implemented based on word alignment technology.

Implement word alignment to the expanded query and the candidate translations, then according to the number of non-empty alignment pairs, coupled with a certain weight (in this example, plus 30),

re-ordering the candidate translations. The results for the example in table 1 are improved as in table 2.

Table 2. The candidate translations mined from search engine snippets, sorted in order of frequency plus word alignment results, for the query of OOV phrase "聚合函数"

C1: Candidate translations by web mining
C2: Frequency
C3: Number of word alignment pairs
C4: Weight of the translation by combination of frequency and word alignment
C5: Final ranking

C1	C2	C3	C4	C5
probefunc	5	0	5	6
functions	4	1	34	3
funcname	2	0	2	7
Func func	2	0	2	8
CSDNBlogfunc	1	0	1	11
Create or replace function func	1	1	31	4
create function func	1	1	31	5
call user funcfunc	1	0	1	9
Aggregating Functions	1	2	61	2
fgetsphp fgetswhile fgetsc	1	0	1	10
echo fgets	1	0	1	12
aggregate functions	1	2	61	1

From the table 2, we can see that the ranking of the candidate translation is more reasonable than pure frequency method.

4. Experiment Design and Result Analysis

To evaluate the algorithm performance, we make experiments on large data sets. The experiment setup is as follows: 1) Phrases used in experiments are derive from "The Chinese Translation Forum" (http://www.chinatranslation.org) of the "English Zone" under the "bilingual information" section; and "the car bilingual glossary" and "computer bilingual glossary" in the first 10 and 20 terms as the test suite. We also select the first 20 names of key universities listed on the Ministry of Education of China as another test set.

The examples of translation mining results are attached in the appendix. Based on the experiments, a comparison of the frequency-based method and the alignment-based method is given in the table 3.

Table 3. The precision statistics of the frequency-based tranlsation mining vs. the method combining word-alignment technology

Technical domain of the OOV	TOP n	Precision	
		Sort by frequency	Sort by frequency and word-alignment
Vehicles	TOP1	80%	90%
	TOP2	80%	100%
Computer	TOP1	50%	75%
	TOP2	50%	80%
	TOP3	60%	85%
University names	TOP1	75%	85%
	TOP2	90%	95%
	TOP3	95%	100%

From the table 3, we can see that in [13] the approach in many fields are effective and the precision is high, at the same time after re-ordering by word alignment the precision is improved. In table 2, we define the precision metrics below:

$$Precision = \frac{\#correct\ translations}{\#all\ translations} \quad (1)$$

The precision in vehicles OOV translation mining result TOP 1 is improved from 80% to 90%; in computer OOV

translation mining TOP 1 is improved from 50% to 75%; in university name TOP 1 is improved from 75% to 85%.

From the experiment result, word alignment technology is very helpful to improve the correctness for OOV automatically translation of web mining.

Some research issues still exist for future research. Firstly, since word alignment is based on a bilingual dictionary, we need a high-quality dictionary which should cover various domains and possible translations. Secondly, this method is not applicable to abbreviation, transliteration, or free translation words. We will make further study on these issues.

5. Acknowledgements

The research project is supported by the Natural Science Foundation of Jiangsu Province (Contract No. BK2006539), the Natural Science Foundation for Higher Education in Jiangsu Province (Contract No. 06KJB520095)

6. References

[1] P. Resnik and N. A. Smith, The Web as a Parallel Corpus, *Computational Linguistics* 29(3), pp. 349-380, September 2003

[2] W. A. Gale and K. W. Church. 1991. Identifying Word Correspondences in Parallel Texts, *In Proc. Of DARPA Speech and Natural Language Workshop.*

[3] J. M. Kupiec. 1993. An algorithm for finding noun phrase correspondences in bilingual corpora. *In Proc. of ACL* 1993: 17-22.

[4] Zou Gang, Liu Yang, Liu Qun, Meng Yao, Yu Hao, Nishino Fumihito, Kang Shi-Yong, Internet-oriented Chinese New Words Detection, *Journal of Chinese Information Processing* 2004，18（6）：1-9

[5] M. Nagata, T. Saito, K. Suzuki. Using the web as a bilingual dictionary [A]. *Proc. ACL 2001 Workshop Data-Driven Methods in Machine Translation* [C]. 2001. 95-102.

[6] P.J. Cheng, J.W. Teng, R.C. Chen, et al. Translating unknown queries with web corpus for cross-language information retrieval [A]. *Proc. ACM SIGIR* [C]. 2004. 146-153.

[7] Y. Zhang, P. Vines, Using the web for automated translation extraction in cross-language information retrieval [A]. *Proc. ACM SIGIR* [C].2004 12-169

[8] Min-Shiang Shia. Improving Translation of Unknown Proper Names Using a Hybrid Web-based Translation Extraction Method. 2005

[9] P.-J. Cheng, Y.-C. Pan, W.-H. Lu, L.-F. Chien. 2004. Creating Multilingual Translation Lexicons with Regional Variations Using Web Corpora. In *Proc. of ACL 2004*: 535-542.

[10] Fei Huang, Ying Zhang and Stephan Vogel, Mining Key Phrase Translations from Web Corpora, *in the Proceedings of the Human Language Technologies Conference (HLT-EMNLP 2005)*, Vancouver, BC, Canada, October 2005.

[11] Fang Ga-Lin, Yu Hao, Meng Yao, Zou Gang. A Computer-Aided Chinese Reading System Based on Analysis Unit of CharactersJ]. *Journal of Chinese Information Processing. 2008*，22（2）：92-98

[12] Lü Ya-Juan. Research on Automatic Translation Knowledge Acquisition Based on Bilingual Corpus Alignment. April 2003:30~76

[13] Sun Jun .Web Mining of OOV Translations. *Journal of Information & Computational Science* 5: 1 (2008) 1-6