Detection and Translation of OOV Terms Prior to Query Time

Ying Zhang Phil Vines
School of Computer Science and Information Technology
RMIT University, GPO Box 2476V, Melbourne, Australia, 3001.
{yzhang,phil}@cs.rmit.edu.au

ABSTRACT

Accurate cross-language information retrieval requires that query terms be correctly translated. Several new techniques to improve the translation of out of vocabulary terms in English-Chinese cross-language information retrieval have been developed. However, these require queries and a document collection to enable translation disambiguation. Although effective, they involve much processing and searching of the Web at query time, and may not be practical in a production web search engine. In this work, we consider what tasks maybe carried out beforehand, the goal being to reduce the processing required at query time. We have successfully developed new techniques to extract and translate out of vocabulary terms using the Web and add them into a translation dictionary prior to query time.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]

General Terms

Algorithms, Languages

Keywords

Cross-Language IR, OOV terms, query translation

1. INTRODUCTION

Many web queries that relate to topical events contain out of vocabulary (OOV) terms. When such terms are not translated correctly, the results returned are often meaningless. Our approach is to search the Web for potential OOV terms, and develop a validation procedure, so that such translations may be added to a dictionary that can then be used at query time. Other approaches to OOV term translation have included using parallel corpora [2], anchor text [1], and transliteration [3]. Parallel corpus based approaches attempt to locate parallel documents on the Web, and use these to build bilingual dictionaries. However, they often suffer from lack of sufficient high quality parallel texts. Lu et al. [1] exploited pages written in different languages with anchor text pointing to the same page. By applying statistical techniques, the top-ranked translation proved to be correct in 53% of cases. While this technique is useful it requires a web page relating to the OOV term, and sufficient interest to cause linking from a foreign language site, using the translation of that term as anchor text. This technique found a number of company names, but apparently did not find names of individuals, place names and other such terms that are unlikely to be the subject of a web page. Meng et al. [3] experimented with transliteration as a method of OOV terms translation. While this method is partially successful, it suffers from the problem that transliteration rules are not applied consistently, and that many terms are translated at least in part semantically.

Our first approach was to collect English text from the Web and exclude all terms that can be found in a translation dictionary. However the vast majority of such data was meaningless strings or spelling errors. Additionally, this method does not detect OOV phrases where regular words take on a special meaning when they occur together. We have observed that in Chinese web pages, when English terms occur, and especially when they occur within brackets, they almost invariably serve as the translation of an immediately proceeding Chinese term, for example 美国哥伦比亚广播公司(CBS), and thus Chinese web pages can be used as a source of potential English OOV terms. By contrast, we found that less than 1% of English terms that occur without brackets were accompanied by their Chinese translation.

SIGIR'04, July 25–29, 2004, Sheffield, South Yorkshire, UK.
ACM 1-58113-881-4/04/0007.

2. TRANSLATION EXTRACTION

In order not to rely on other search engines, it would be necessary to crawl all Chinese web pages that contain English text. However, purely for the purpose of testing our idea, we have relied on Google to pinpoint specific sites of interest and only crawled these. To extract some test data from the Web, we crawled some Chinese news web sites to obtain 2Gb of text and filtered to collect lines containing English terms that occurred within brackets. From this process, we found 1,168 distinct English terms including words and phrases, that fell into two categories.

First, some Chinese translations occur side by side within one of two types of Chinese quotation marks, for example “世界小姐” (Miss World) and 《内地与香港关于建立更紧密经贸关系的安排》(CEPA). We found 204 (17%) such instances and later evaluation showed that 98% of these were correct translation pairs. Second, some Chinese translations are followed by the English OOV terms without any quotation marks, for example 联合国下属的世界卫生组织(WHO). For this case, we developed a procedure to extract and then validate the translations. Since there is no white space between Chinese characters, it is not clear how many of the characters that proceed the OOV term constitute the translation. In our

	Number	Proportion	Out Of Vocabulary	In Dictionary	
				Existing Translation	New Translation
1. *Exactly correct*	713	61%	650	32	31
2. *Correct translation with 1 or 2 related extra words*	58	5%	54	3	1
3. *Correct translation with more than 2 related extra words*	78	7%	77	0	1
4. *Correct translation with any other misleading extra words*	71	6%	62	9	0
5. *Correct translation but incomplete*	157	13%	156	1	0
6. *Wrong translation*	91	8%	-	-	-
Total	1168	100%	999	45	33

Table 1: *Translation Quality Evaluation*

previous work, we developed a method that used all substrings together with a statistical technique to extract the most appropriate translation [4]. Where an OOV term had more than one meaning, we were able to successfully extract each of them. For example, we extracted both "国际电工委员会" and "国际金融公司" as the translations of "IEC".

When working with a specific collection, we can filter out the translations that never occurred in the collection. Even if some of these translations are valid or meaningful, the fact that they do not occur in the collection means that they will not effect retrieval. However when we crawl the Web, all the translations we obtained occurred more than once. Some of these translations may be valid while others are meaningless. Thus we developed a new technique that uses the Web instead of specific collection to validate translations.

2.1 Translation Validation

We use a three stage process to validate the Chinese translations:

1. For each OOV term (e_{oov}) that has more than one translation for each meaning we use Google to fetch the top 100 Chinese documents using the Chinese string S that immediately proceeds e_{oov}, as the query. For each document returned, only the title and the query-biased summary are extracted and filtered to remove HTML tags and metadata, leaving only the web text.

2. We scan for the occurrence of e_{oov} and check the immediately proceeding Chinese text to see if it is a substring of S. Then we collect the frequency f_{s_j} of all substrings of S.

3. We select the Chinese translation s_{final} where $f_{s_{final}} = max(f_{s_j})$. In the event of a tie we use length to discriminate.

3. RESULTS AND CONCLUSION

We used a native Chinese speaker to evaluate the translations we extracted from the Web. The person was given the web pages, the English OOV terms, and the candidate translations. Results are shown in Table 1.

We can see that in terms of translation accuracy, 61% of the translations found were strictly correct. However, an additional 12% of translations contained the correct translation, with some additional related information, such as the person's title or the organization's location. When used for retrieval, this additional information may arguably improve effectiveness. A further 6% of translations were correct but included some misleading terms, or terms commonly associated with the OOV term. For example "第二届美国偶像大赛" is extracted as the translation of "American Idol", where term "the second" (第二届) is commonly associated with the OOV term "American Idol" (美国偶像大赛) in web documents. In 13% of cases involving translation of personal names, only the surname was found, for example "克莱维茨" is extracted as the translation of "Lenny Kravitz". Arguably in both of these cases such translations would find relevant documents although the precision may be lower. Of the 1,077 correct or partially correct translations, only 78 of these were already in one of the three translation dictionaries [1] we used in our experiments, although in 33 cases we found new additional meanings. Only 8% of translations were assessed as wrong.

Our initial results are encouraging. Of the 1,168 OOV terms we extracted, 61% of translations were strictly correct and only 8% were completely wrong. We have shown that it is possible to search the Web for potential OOV terms and use a procedure to validate the translations. Such translation pairs may be added to both English-to-Chinese and Chinese-to-English dictionaries, that can latter be used at query time, thus improving the efficiency of query processing. Another important by-product of collecting OOV terms in the way is that we could develop a procedure to periodically crawl the Web to add additional translations to the translation dictionary. This will be the subject of our further research.

4. REFERENCES

[1] W. Lu, C. Tung, L. Chien, and H. Lee. Translation of web queries using anchor text mining. *ACM Transactions on Asian Language Information Processing*, 2(1):159 – 172, 2002.

[2] C. J. A. McEwan, I. Ounis, and I. Ruthven. Building bilingual dictionaries from parallel web documents. In *Proceedings of the Twenty-Fourth European Colloquium on Information Retrieval Research*, pages 303–323, Glasgow, Scotland, UK, 2002.

[3] H. Meng, B. Chen, S. Khudanpur, G. Levow, W. Lo, D. Oard, P. Schone, K. Tang, H. Wang, and J. Wang. Mandarin-English Information (MEI): investigating translingual speech retrieval. In *Computer Speech and Language*, 2003.

[4] Y. Zhang and P. Vines. Using the Web for Automated Translation Extraction in Cross-Language Information Retrieval. In *Proc. ACM-SIGIR International Conference on Research and Development in Information Retrieval*, 2004.

[1] http://www.mandarintools.com/cedict.html
http://www.ldc.upenn.edu/Projects/Chinese/docs/ldc_ec_dict.2.0.txt
http://www.ldc.upenn.edu/Projects/Chinese/docs/ldc_ce_dict.2.0.txt