Mining Translations of OOV Terms from the Web through Cross-lingual Query Expansion

Ying Zhang
joy+@cs.cmu.edu

Fei Huang
fhuang+@cs.cmu.edu

Stephan Vogel
vogel+@cs.cmu.edu

Language Technologies Institute
Carnegie Mellon University
5000 Forbes Ave. Pittsburgh, PA 15217, U.S.A.

ABSTRACT

Translating out-of-vocabulary (OOV) terms is a great challenge for the Cross-lingual Information Retrieval and Data-driven Machine Translation systems. Several approaches have been proposed to mine translations for OOV terms from the web, especially from pages containing mixed languages. In this paper, we propose a novel approach to automatically translate OOV terms on the fly through cross-lingual query expansion. The proposed approach does not require any web crawling and has achieved an inclusion rate of 95% and overall translation accuracy of 90%, outperforming state-of-the-art OOV translation techniques.

Categories and Subject Descriptors: H.3.3 [Information Search and Retrieval]:

General Terms: Algorithms, Language, Experimentation, Performance.

Keywords: Cross-lingual IR, OOV terms, automatic translation, query expansion.

1. INTRODUCTION

Out-of-vocabulary(OOV) terms cause serious problems for Cross-lingual Information Retrieval and Data-driven Machine Translation systems. Given the enormous amount of information on the web, many approaches have been developed to mine the translations of the OOV terms from the web. STRAND system [4] searches the web for parallel text and [3] extracts translations pairs among anchor texts pointing together to the same webpage. However they all suffered from the lack of such bilingual resources available on the web. On the other hand, great amount of bilingual information exist on the web in the form of tentative translation or references, such as "中文片名*:* 廊桥遗梦*,* 英文片名*: The Bridges of Madison County* 导演*:* 克林特·伊斯特伍德 *Director: Eastwood, Clint*". As observed by Zhang and Vines [5], when English terms occur in Chinese web pages, and especially when they occur within brackets, they are very likely to be translations of an immediately preceding Chinese term. Cheng et al. [1] observed that if a Chinese term occurs in an English web page, its translation usually exists in the same page too. They queried Google for English pages containing the Chinese OOV term and extract the translations from the returning snippets. These two approaches all intended to extract translation candidates from the web pages containing both Chinese and English text. The former method does not restrict the search space, which means lots of web pages have to be crawled to get one containing the English translation. On the other hand, the latter method restricts too strongly and the search space is too small. According to our analysis, only 1/45 of the pages containing both the OOV term and its English translation are identified by Google as English pages. In this paper, we propose a new approach to retrieve web pages of mixed languages which might contain the translations for the OOV term by expanding the Chinese query with an English hint word. Throughout this paper, we consider Chinese as the source language and English as target although the proposed method is language independent.

For a Chinese OOV term f, we want to find its translation e. Let's assume that there is a Chinese term f' which is relevant to f, and can be translated to e' using the existing bilingual lexicon. We observe that when f and e exist in a web page, f' and e' are also very likely to exist in the same page. Thus to search for pages containing f and e, we search for pages containing f and e', where e' is a *hint* word generated by cross-lingual query expansion. For example, to find web pages which might contain translations for "列夫·托尔斯泰"(Leo Tolstoy), we expand the query to "列夫·托尔斯泰*+war+peace*" since "战争与和平"(War and Peace) is very relevant to "列夫·托尔斯泰" and we know its translation.

2. CROSS-LINGUAL QUERY EXPANSION

To propose a "good" English hint e' for f, we first need to find a Chinese term f' that is relevant to f. Because f is an OOV term, it is unlikely to obtain much information from the existing Chinese monolingual corpora. Instead, we queried Google for web pages containing f. From the returning snippets we select Chinese terms f' based on the following criteria: First, f' should be reliably translated into English noun or noun phrases given the available bilingual resources (e.g. LDC Chinese-English dictionary or word/phrase lexicons trained from parallel corpus). Secondly, f' should be one of the most relevant words to f, where the relevance is estimated in terms of its frequency amongst the snippets. The corresponding translations e' for each f' were then used as the hint words for each f. For example, for $f = r,y$ (*Faust*), the top candidate f's are "歌德", "简介", "文学"and "悲剧". We expanded the original query 浮士德 to 浮士德*+goethe*, 浮士德*+introduction*, 浮士德*+literature*, 浮士德*+tragic* and searched Google again.

SIGIR'05, August 15–19, 2005, Salvador, Brazil.
ACM 1-59593-034-5/05/0006.

3. EXTRACTING TRANSLATIONS

When a query and its expanded English hint words are sent to Google, snippets containing the query and possibly its English translation are returned. We apply preprocessing on snippet text by filtering out HTML tags, punctuation marks and non-query source words. We extract the English translation from the processed top-N snippets, and provide confidence scores for each translation candidates. The translation extraction features include a transliteration cost, a translation cost and their frequency-distance weights [2]. The transliteration model measures the pronunciation similarity between a source phrase and a English candidate, while the translation cost indicates their semantic equivalence, which is calculated from a bilingual lexicon. The frequency-distance feature shows how relevant the English candidate is with regard to the source query, based on the observation that correct translation pairs co-occur more often and closer within a snippet. According to the confidence scores of different models, we output the top-5 translation hypotheses for evaluation.

4. EXPERIMENTS AND CONCLUSION

We collected 310 Chinese OOV terms from 12 categories, including movie titles, book titles, organization names, product brands, sci & tech. terms, specie names, person names, location, military terms, medical terms, musical terms and sports terms. On average 13.2 snippets were used to identify the relevant Chinese terms f' for each OOV term f. Top-5 f's were used to generate hint words e'. Snippets containing both f and e' were then used to extract translations for f. Figure 1 shows the inclusion rate vs. the number of snippets used for three mixed language page searching strategies. Inclusion rate is defined as the number of OOV terms which have correct translations included in the returning snippets. Searching web pages containing f without any language constraints as did in [5] resulted in a relatively low inclusion rate. Constraining the search to English pages only [1] resulted in a much higher inclusion rate. But such English pages are limited. On average, only 45 unique snippets could be found for each f which resulted in a maximum inclusion rate of 85.8%. In the case of the cross-lingual query expansion, the search spaces was much larger and we achieved a high inclusion rate of 89.7% when using 32 snippets and 95.2% when using 165 snippets. However, more snippets also introduced more noises. Table 1 showed the translation extraction accuracy using different features. With more returned snippets, the extraction accuracy is decreased when using transliteration (Trl) and translation model (Trans) alone. However, adding frequency-distance model (Freq-Dis) boosts the accuracy significantly.

Features	Trl	Trans	Freq-Dis	All
No Hints	51.45	51.45	53.62	65.94
With Hints	17.97	40.98	73.22	86.73

Table 1: Extraction Accuracies Using Different Features

Overall translation quality figures are listed in Table 2 and compared with the LiveTrans system [1] and Systran[1]. Reference translations for each OOV terms are provided by a human translator. A hypothesized translation is consider "correct" when it can be matched exactly with one of the reference translations.

[1]www.systransoft.com

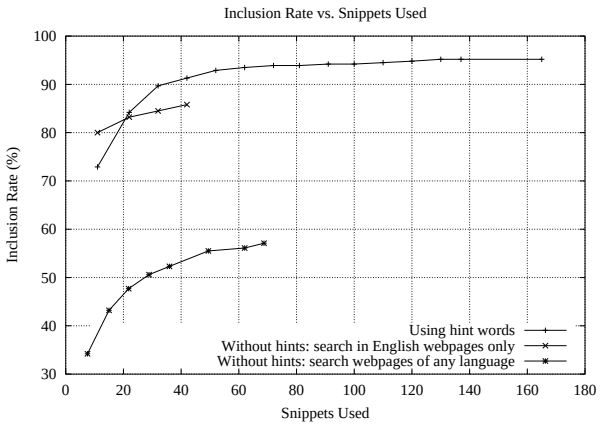

Figure 1: Inclusion Rate vs. Snippets Used

Snippets Used	Accuracy of the Top-N Hyp. (%)				
	Top1	Top2	Top3	Top4	Top5
10	46.1	55.2	59.0	61.3	62.3
20	57.4	64.2	69.7	72.6	72.9
50	63.2	74.5	77.7	79.7	80.6
100	75.2	84.5	85.8	87.4	87.4
165	81.0	86.5	89.0	90.0	90.0
Systran	31.3	N/A	N/A	N/A	N/A
$LiveTrans^{Fast}$	20.6	30.0	36.8	41.9	45.2
$LiveTrans^{Smart}$	30.0	41.9	48.7	51.0	52.9

Table 2: Overall Translation Accuracy

Our initial results have shown that cross-lingual query expansion fetches snippets with very high inclusion rate, and various similarity and relevancy features ensure high accuracy translation extraction. As a whole, these result in high quality translations for OOV terms. This approach is fast and language independent. We will apply this method to the Cross-lingual Information Retrieval, Machine Translation and Question Answering systems in our future research and test it on other language pairs.

5. REFERENCES

[1] P.-J. Cheng, J.-W. Teng, R.-C. Chen, J.-H. Wang, W.-H. Lu, and L.-F. Chien. Translating unknown queries with web corpora for cross-language information retrieval. In *SIGIR '04*, pages 146–153. ACM Press, 2004.

[2] F. Huang, S. Vogel, and A. Waibel. Automatic extraction of named entity translingual equivalence based on multi-feature cost minimization. In *Proceeding of the 41st ACL, Workshop on Multilingual and Mixed-Language Named Entity Recognition*, Sapporo, Japan, July 2003.

[3] W.-H. Lu, L.-F. Chien, and H.-J. Lee. Translation of web queries using anchor text mining. *ACM TALIP*, 1(2):159–172, 2002.

[4] P. Resnik and N. A. Smith. The web as a parallel corpus. *Comput. Linguist.*, 29(3):349–380, 2003.

[5] Y. Zhang and P. Vines. Detection and translation of oov terms prior to query time. In *SIGIR '04*, pages 524–525. ACM Press, 2004.