Enhanced web-based translation extraction for English-Chinese CLIR

Chengye Lu Yue Xu Shlomo Geva
School of Software Engineering and Data Communications
Queensland University of Technology
Brisbane, QLD 4001, Australia
{c.lu,yue.xu,s.geva}@qut.edu.au

ABSTRACT

Dictionary based translation is a traditional approach in use by cross-language information retrieval systems. However, significant performance degradation is often observed when queries contain words that do not appear in the dictionary. This is called the Out of Vocabulary (OOV) problem. The common methods for translation selection for web-based translation always rely on word frequency calculation but the results are not always satisfactory. Our experiments show marked improvement in translation accuracy over other commonly used approaches.

1. INTRODUCTION

Dictionary based translation has often been used in cross-language information retrieval because bilingual dictionaries are widely available and dictionary approaches are easy to implement. This approach shows high efficiency in term and phrase translation, however, translation disambiguation and the out of vocabulary (OOV) problem challenge cross-language information retrieval systems. Translation disambiguation refers to finding the most appropriate translation from several choices in the dictionary. The OOV problem refers to the situation where translations of some words cannot be found in the dictionary at all. Even in the best of dictionaries this is to be expected of course. Very often the OOV terms are proper names or newly created words that carry the most information of the query. When it is missing in the translated query, it is most likely that the user will practically be unable to find any relevant documents at all.

2. PROPOSED APPROACH

Our approach is similar to the previous works[1][2][3] in terms of the web based translation approach which tries to find the OOV term's translation through web search engine. However, our approach differs in term ranking and selection strategy. The aim of our approach is to find the most appropriate translation from the word list regardless the term frequency.

The basic idea of our approach is to combine the translation disambiguation technology and the web-based translation extraction technology together. The web-based translation extraction process usually returns a list of words. As those words are all extracted from the results returned by the web search engine, it is reasonable to assume that those words are relevant to the English terms that were submitted to the web search engine. If we assume all those words are translations of the English terms, we can apply the translation disambiguation technique to select the most appropriate word as the translation of the English terms..

2.1 Results

Table 1 below gives the results from four runs.

Table 1 Retrieval performance

	Average precision	Percentage of Mono
Mono	0. 3526	100%
Ignore OOV	0.1290	36.5%
Previous	0.2302	65.3%
Propose	0.2576	73.1%

It is clearly that when using our proposed approach, we have highest retrieval performance. This result indicates that our translation approach has the highest effective. The precision of our approach is 174% comparing to the case of not processing OOV terms and it is 120% comparing to the case of the simulation of previous approaches.

3. CONCLUSION

In this paper, we have described an approach to tackling the OOV problem in English-Chinese information retrieval. By using web translation extraction based on co-occurrence model, the overall performance can boost to almost 174% comparing to the case of not processing OOV terms. 120% comparing to the simulation of previous approaches. This is a marked improvement in translation accuracy over other commonly used approaches.

4. Reference

[1] Cheng, P.-J., J.-W. Teng, et al. (2004). *Cross-language information retrieval: Translating unknown queries with web corpora for cross-language information retrieval.* Proceedings of the 27th annual international conference on Research and development in information retrieval.

[2] Gao, J., J.-Y. Nie, et al. (2001). *Improving query translation for cross-language information retrieval using statistical models*. SIGIR '01: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, New Orleans, Louisiana, United States, ACM Press.

[3] Zhang, Y., and Vines, P (2004). *Using the web for automated translation extraction in cross-language information retrieval*. Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval. Sheffield, United Kingdom, ACM Press.