Improved use of Contextual Information in Cross-language Information Retrieval

Ying Zhang, Phil Vines

School of Computer Science and Information Technology, RMIT University
GPO Box 2476V, Melbourne Victoria 3001, Australia

{ *yzhang, phil* }*@cs.rmit.edu.au*

Abstract

In this paper, we explore Dictionary based context-sensitive translation, a framework for query translation to reduce the translation ambiguity and improve the translation quality in English to Chinese cross-language information retrieval (CLIR). Our paper explores the effect of the context window size on translation effectiveness. We assume that the correct translations of the query key terms tend to co-occur together at a high frequency and incorrect translations do not. Our experimental results showed that when using a window size of 10, context-sensitive translation results in a dramatic improvement in retrieval performance, it brings about a 30% improvement compared to the results of previous Dictionary based approaches that used only immediately adjacent words for context.

Keywords Information Retrieval

1 Introduction

With the enormous increase in recent years in the number of multilingual text databases available online, there has been a growing interest in the research done in the area of Cross-Language Information Retrieval (CLIR). This paper is concerned with one type of CLIR, namely, issuing a query in English and retrieving a Chinese document.

In order to match a document and a query in two different languages, either the document or the query should be translated. By far the simplest approach is to convert queries to the document language and do the monolingual ranked retrieval. This approach is called query translation. Most research has concentrated on query translation [1, 4, 5, 6], as it is computationally less expensive than document translation, which requires a lot memory and processing capacity. Within the query translation framework, basic approaches to CLIR are: the Machine translation based approach, the Parallel corpus based approach and the Dictionary based approach. Regardless of the cross-language approach taken, the main hurdle to improved CLIR effectiveness is resolving ambiguity associated with translation. In our work, we aim to translate each English key term using both a dictionary lookup technique and the available context, into the most appropriate corresponding Chinese term. Rather than use only immediately adjacent words to determine the context, we experiment with a window size *w* of words around the key term.

The rest of the paper is organised as follows: section two presents the related work; in section three, we discuss the formula and the selection algorithm used in our context-sensitive translation approach; section four describes the experiment setup; section five presents and summarizes the evaluation results and provides our discussions, and section six concludes the paper.

2 Related work

Translation ambiguity is a major problem in CLIR and arises from the fact that many words have multiple possible translations. Some researchers in the past have used approaches such as First-Match and Every-Match [7]. More recent work has concentrated on adjacent words to provide the context and thus help select the appropriate translation. Ballesteros and Croft [1] describe a technique that employs co-occurrence statistics obtained from the corpus being searched to disambiguate dictionary translation. They measure the importance of co-occurrence of the elements in a set by using the *em* metric - the percentage of the occurrences of term *a* and term *b* which are net co-occurrences (co-occurrences minus expected co-occurrences).

Our work applies this technique to English to Chinese CLIR and extends this approach by using a larger window of contextual information.

Proceedings of the 7th Australasian Document Computing Symposium,
Sydney, Australia, December 16, 2002.

3 Context-sensitive translation

Given a set of n original query terms $\{e_1, e_2, ..., e_n\}$, we obtain a set C_i of all possible translations for each e_i through a bilingual dictionary lookup and then try to select the best translation that co-occurs with all other sets of translations in the same context window at the highest frequency, from each C_i.

3.1 Mutual information

Mutual information *(MI(x,y))* [2] is used to measure the correlation between all word pairs in the same context window in our application of query translation. In other words, the correct translation of the given query key term is not only determined by the immediately adjacent words but also the words that are in the same context window. Previous work has tended to look only at these immediately adjacent words. This is the major point of difference between our work and previous work [1] and has contributed significantly to our improved results.

We extend the *MI* formula, which previously used to measure mutual information between immediately adjacent words, to measure the *MI* between a given word and every other word within a window size *w*.

$$MI(x, y) = \log_2 \frac{N f_w(x, y)}{f(x) f(y)}$$

Where $f_w(x, y)$ is the frequency that term *x* and term *y* co-occur within a window size *w*, $f(x)$ is the collection frequency of *x*, $f(y)$ is the collection frequency of *y*, and *N* is the total number of words in the document collection. A window size *w* is the context of the given term that consists of *w/2* words before and after the given term.

3.2 Selection algorithm

There are *n* terms in the English query, for each original English query term e_i $(i \in (1, n))$, we obtain a set C_i of all possible Chinese translations c_{ij} $(j \in (1, m))$ through a bilingual dictionary lookup:

```
For each set $C_i$, do
    For each translation $c_{ij}$ in $C_i$, do
        For each set $C_k$ $(k \in (1, n), k \neq i)$, do
            Compute the mutual information
            $MI(c_{ij}, C_k) = Max_{c_y \in C_k} MI(c_{ij}, c_y)$;
        Calculate the score of the translation $c_{ij}$:
        $S_{c_{ij}} = \sum_{k=1, k \neq i}^{n} MI(c_{ij}, C_k)$
    Select the $c_{ij}$ with the highest $S_{c_{ij}}$
```

Where n is the number of key terms e_i in the English query, m is the number of Chinese translations c_{ij} in the Chinese translation set C_i. For more detailed descriptions, see [8].

4 Experiments

We conducted the English to Chinese CLIR experiments in the Chinese collection of TREC 5 and TREC 6 [3]. There are 54 queries and 164,789 documents (170MB) of articles drawn from the People's Daily newspaper and the Xinhua newswire. In our experiments, the first 28 topics (CH01 to CH28) were processed as queries to retrieve the documents from the Chinese collection. The queries use all sections of the topic. In our experiments, we use a dictionary from the Linguist Data Consortium (LDC) - ldc2ec, which has 110,834 entries (http://www.morph.ldc.edu/ Projects/ Chinese). The Chinese stop list was manually selected from the statistical results we obtained from the Chinese document collection. The top 118 words with the highest collection frequency (CF) and relatively unimportant meanings are selected as stop-word and put into the Chinese stop list.

4.1 Monolingual IR experiment

We have carried out monolingual IR experiments using both character-based and word-based approaches. In word-based approach, we use the dictionary based method with greedy parsing, where a dictionary that contains 58,667 entries is employed. These experiments have shown that word-based approach is statistically better than character-based approach. The retrieval performance of the word-based monolingual run is in the second column in Table 1. This provides a benchmark for our CLIR results.

4.2 CLIR experiment

We explored two methods for disambiguating dictionary based query translation, the First-Match method and the context-sensitive translation.

To provide a baseline for CLIR results, we obtained a recall-precision average for the First-Match method. Only the first match translation per query term is retained instead of using all of the listed translations in lcd2ec when there is more than one translation for that term. When there is no exact matching for a single-word term, the term is skipped. The retrieval performance of the First-Match run is in the third column in Table 1.

Our context-sensitive translation scheme works in three stages: pre-processing, dictionary based query translation and translation disambiguation. At the first stage, English stop words are removed from

English topics. When an English word has a trailing "s", it is removed according to Porter's algorithm. Other steps of stemming are not done in order to avoid changes in meaning. The Chinese document collection is segmented into the words; and then Chinese stop words are removed from the Chinese document collection. The second stage does actual query translation based on ldc2ec dictionary look-up. For each Chinese translation, we applied a Chinese stop list filter. At the third stage, the $MI(x, y)$ statistics was used to determine whether the Chinese translations from different sets generated by the translation process are "compatible".

5 Results and Discussion

Using the TREC data and queries described earlier, we have gathered in Table 1 a comparison of the recall precision values for the experimental results. Columns two through seven correspond respectively to the word-based monolingual information retrieval, the First-Match method, and our context-sensitive translation method using a window size of 2, 8, 10 and 12. It is seen that when using a window size of 10, our context-sensitive translation disambiguation method successfully brings effectiveness to over 62% of monolingual retrieval. This is a considerable improvement over previous work (using adjacent information, i.e. $w = 2$) which yielded 33.5% of monolingual retrieval effectiveness. More detailed results and discussion are presented in [8].

Figure 1 plots the precision-recall curves for a comparison of the retrieval performance of the six runs. It is seen that the context-sensitive translation approach significantly outperforms the First-Match method, however, the gap between monolingual and cross-language retrieval is still large.

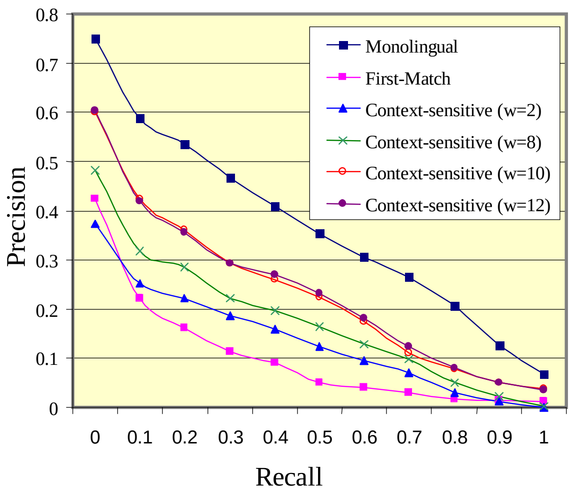

Figure 1: Recall-precision average of queries 01-28

Figure 2 shows the effect of the context window size on translation effectiveness. As the window size is enlarged from 2 to 10, the average precision increases from 0.1267 to 0.2354. There is no significant improvement for window size greater than 10.

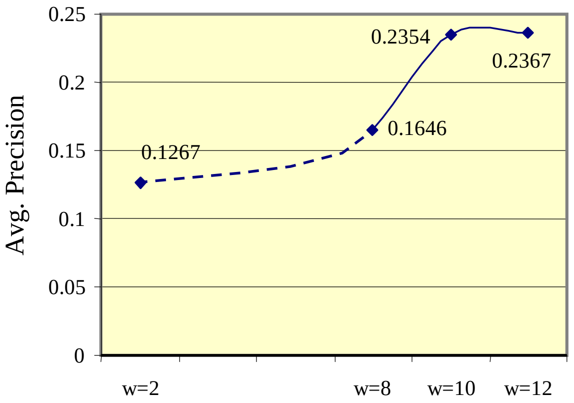

Figure 2: Comparison of the widow size effect on CLIR effectiveness (Avg. precision).

5.1 Factors affecting translation effectiveness

In this section, we examine several factors that may directly or indirectly have degraded the retrieval effectiveness of English to Chinese CLIR.

First, the translation of certain English query key terms cannot be found in the English-Chinese dictionary (ldc2ec) we used in the experiment, such as place names, person names and some terminologies. For example, Xinjiang (新疆, place name) and WTO (世界贸易组织, terminology). Out-of-vocabulary words not found are left as un-translated. This is one reason that our retrieval effectiveness is degraded.

Second, for some English query key terms, there is no exact equivalent Chinese translation, although some approximate equivalents are provided in the ldc2ec. For example, we obtained Chinese translations "/ 数据 / 数位 /" for English term "digital", however, the corresponding Chinese translation should be "数字". Therefore, we failed to select the best translation in this case.

Third, since ldc2ec does not support phrase translations, we obtained some inappropriate translations for phrases. For example, we translated peace talks (和平会议) into "和平会谈".

Fourth, some best Chinese translations are missing in the segmentation dictionary. Even when the correct Chinese translation is obtained through the lcd2ec lookup, it is not treated as a word in the segmentation dictionary and the worse thing is that the translation is missing in the segmentation dictionary. Therefore we cannot select it as the best translation. For example, the Chinese translation "海洛因" of English term "Heroin" is missing, so we can only select "白面儿" as the translation of the English term "heroin".

Recall	Monolingual	First-Match	Context-sensitive Translation			
			$(w=2)$	$(w=8)$	$(w=10)$	$(w=12)$
0.00	0.7489	0.4245	0.3745	0.4815	0.5996	0.6023
0.10	0.5884	0.2224	0.2518	0.3173	0.4251	0.4188
0.20	0.5355	0.1609	0.2221	0.2858	0.3607	0.3569
0.30	0.4672	0.1139	0.1876	0.221	0.2919	0.2939
0.40	0.4096	0.0906	0.1586	0.1975	0.2588	0.2697
0.50	0.3521	0.0504	0.1240	0.1647	0.2235	0.2325
0.60	0.3055	0.0414	0.0967	0.1277	0.1742	0. 1829
0.70	0.2645	0.0303	0.0698	0.0972	0.1120	0.1240
0.80	0.2063	0.0178	0.0297	0.0502	0.0782	0.0812
0.90	0.1266	0.0157	0.0121	0.0232	0.0513	0.0513
1.00	0.0675	0.0116	0.0008	0.003	0.0368	0.0349
Avg. precision	0.379	0.089	0.1267	0.1646	0.2354	0.2367
% Monolingual	100	23.51	33.47	43.48	62.18	62.52

Table 1: Evaluation Results

6 Conclusion

As explained above, there are several factors that degrade the English to Chinese CLIR effectiveness. In this paper, we have looked in detail at only one of these problems, which is the translation ambiguity problem, and shown how significant improvement may be obtained. We adopted an improved context-sensitive translation method to improve dictionary based query translation in English to Chinese CLIR. We translate each English key term using both a dictionary lookup technique and the available context, into the most appropriate corresponding Chinese term. Through our experiments, we showed that our approach significantly outperforms the previous work that used only immediately adjacent words for context, i.e. a window size of 2, and leads to a relatively high effectiveness.

References

[1] Ballesteros, L & Croft, W.B. Resolving Ambiguity for Cross-Language Retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and* Development in Information Retrieval, August 24-28 1998, Melbourne, Australia. ACM 1998. pp. 64-71.

[2] Church, K. W. and Hanks, P. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 1990, 16(1), pp. 22-29.

[3] E. Voorhees and D. Harman. Overview of the Sixth Text REtrieval Conference. In *E.M. Voorhees and D.K. Harman, editors, Proceedings of the Sixth Text REtrieval Conference* (TREC-6), pages 1 -- 24, Nov 1997.

[4] Jianfeng Gao, Jian-Yun Nie, Endong Xun, Jian Zhang, Ming Zhou, Changning Huang. Improving Query Translation for Cross-Language Information Retrieval using Statistical Models. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, September 9-13, 2001, New Orleans, Louisiana, USA. ACM 2001.

[5] Hull, D. A., and Grefenstette, G. Querying Across Languages: A dictionary-based approach to Multilingual Information Retrieval. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 49-57). Zurich, Switzerland: ACM Press.

[6] Jang, Myung-Gil Jang, Sung Hyon Myaeng and Se Young Park. Using Mutual Information to Resolve Query Translation Ambiguities and Query Term Weighting. *In ACL-99*. College Park, Maryland, pp. 223--229.

[7] Mohammed Aljlayl and Ophir Frieder. Effective Arabic-English cross-language information retrieval via machine-readable dictionaries and machine translation. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, September 9-13, 2001, New Orleans, LA USA. ACM 2001. pp. 295 - 302 J.

[8] Zhang, Ying. Multilingual Querying for Information Retrieval. A Minor thesis submitted in partial fulfillment of the requirements for the degree of Masters of Applied Science, School of Computer Science and Information Technology, RMIT University, pp. 16-17.