Using the Web for Translation Disambiguation

RMIT University at NTCIR-5 Chinese–English CLIR

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

RMIT University participated in the Chinese–English NTCIR-5 CLIR task. In previous work, researchers have relied on the test collection corpus to perform translation disambiguation in CLIR. In a production system, one would not be able to use a constrained test collection for disambiguation. Therefore we are interested to see how well our techniques perform when the web is used to provide context for disambiguation. Our experimental results show that when using the web, it is possible to achieve effectiveness comparable to that obtained with a test collection.

Keywords: *translation disambiguation, statistical model, web mining.*

1 Introduction

Dictionary-based query translation is a widely used approach in cross-language information retrieval (CLIR), not only because of its simplicity and the increasing availability of machine readable dictionaries, but also it has proved the most robust for the short queries that are typically entered by web users. In this approach, queries are translated by looking up terms in a bilingual dictionary and using some or all of the translated terms. By using simple dictionary translations, the ambiguity introduced by using all possible translations yields poor effectiveness [2, 3]. The translation ambiguity problem stems from the fact that many words do not have a unique translation, and sometimes the alternate translations have very different meanings. It is particularly severe when users enter short queries (often two or three words), a situation in which it may not be possible for even a human to determine the intended meaning from the available context. Previous web search engine log analysis revealed that the average query length for a web search was about 2.3 words in English [2] and 3.18 characters in Chinese [9]. The dictionary-based methods are error prone due to the high possibility of selecting the wrong translation of a term from among the translations provided. To reduce the ambiguity and errors introduced during query translation, various techniques utilizing statistics obtained from the test collection corpus, have been proposed [1, 3, 4, 5].

2 Query Translation

We use a dictionary to obtain all possible translations for each query term. An out-of-vocabulary (OOV) translation process is also employed to check for terms and phrases not in the dictionary [11]. The OOV translation extraction algorithms we developed employ co-occurrence statistics that combine length and frequency attributes. The Chinese OOV translation is the first step in the retrieval process. We add these terms into both a segmentation dictionary and a translation dictionary. Once candidate query term translations are collected, we use a disambiguation technique to determine the most appropriate English translation for each Chinese query term.

2.1 Chinese OOV translation

When looking for English translations of Chinese OOV terms, they need to be appropriately detected in the query. Many existing systems use a segmenter to determine Chinese word boundaries. However, if the Chinese OOV term is currently unknown, there is no information to indicate how it should be segmented.

The basis of our approach is the observation that most translated English terms tend to accompanied by the original English terms on the web, typically immediately after the Chinese text, but general terms do not. By mining the web to collect a sufficient number of such instances for any given word and applying statistical techniques, we

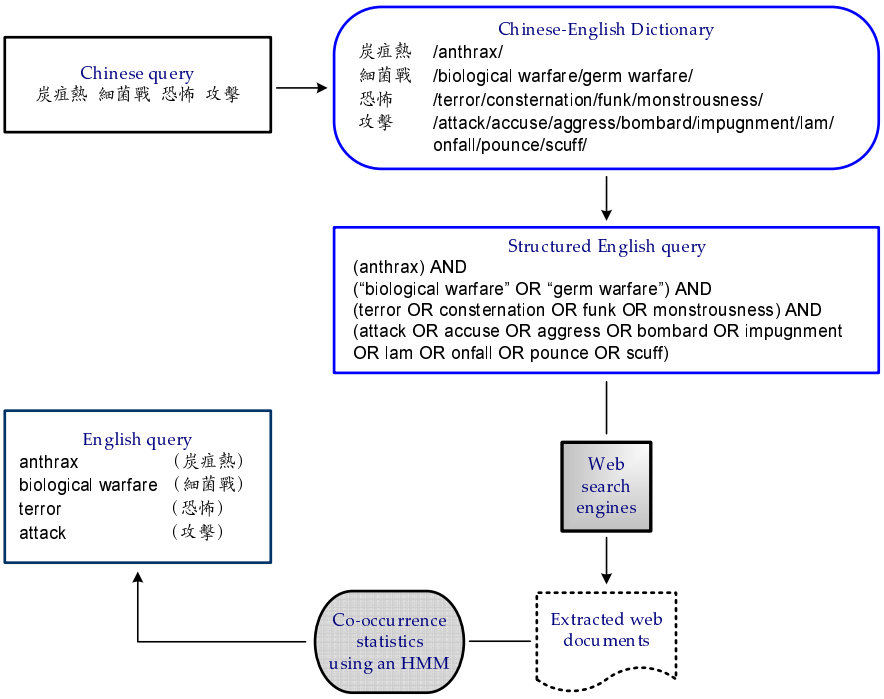

Figure 1. *The process of translation disambiguation using the web*

hypothesize that we are then able to infer an appropriate translation with reasonable confidence. In formulating our approach, we also considered English text that was not immediately adjacent to the Chinese query terms. However, we found that such text was only rarely a reliable translation. In some cases we found only a small number of Chinese-English co-occurrences, and our approach proved to be robust in such situations. The process is detailed as follows:

1. Use Google to fetch the top 300 Chinese documents, using the entire Chinese query. For each returned document, the title and the query-biased summary are extracted.

2. Where English text occurs, check the immediately preceding Chinese text to see if it is a substring of the Chinese query. Collect the frequency of co-occurrence of each distinct English string and all Chinese query substrings that appear immediately prior.

3. Select the English text e with the highest frequency, since the English text that occurs with the highest frequency is more likely to provide the correct translation compared to other text with the lower frequency.

4. For this English text e, select the associated Chinese query substring c with the highest co-occurrence frequency. In the event of a tie we use the length to discriminate.

5. If the selected Chinese query substring c cannot be found in the Chinese segmentation dictionary, we treat it as OOV term and add it into the Chinese segmentation dictionary and (c,e) into the translation dictionary.

A given Chinese term may have more than one English transliteration. However, this phenomenon is rare. Our methods tend to choose the most common form.

2.2 Translation disambiguation using the web

The disambiguation techniques we previously developed, based on the *Hidden Markov Model* (HMM) [6], have been shown to be effective in Chinese–English CLIR [10]. In NTCIR-5 Chinese–English CLIR task, instead of using the NTCIR-5 English collection, we experimented with English web documents extracted by a search engine as a corpus to disambiguate dictionary translation (see Figure 1).

2.2.1 Structured query construction

Rather than collect corpus statistics from the entire web, which is impractical in a test environment, our approach is to retrieve a set of documents that contain the candidate translation terms using a web search engine. The correct

translations are generally semantically related and tend to co-occur much more often in documents than incorrect translations.

Since the original Chinese query terms may have varying numbers of English translations, it is important to normalize the possible English translations in such a way that all possible English translations are treated equally for each Chinese query term. Not using normalization will make Chinese query terms with a lot of possible translations unintentionally more important than those having less possible translations. We use a structured query in order to prevent Chinese query terms with large numbers of English translations dominating the retrieval process.

As shown in Figure 1, a Chinese query is transformed into a structured English query using the Boolean operators AND and OR. The parentheses are used to indicate the order in which we want the search engine to interpret the operators. For example, suppose a query Q is composed of a sequence of Chinese terms $(c_1, c_2, ..., c_n)$. We obtain a set E_i of all possible English translations $\{e_{i1}, e_{i2}, ..., e_{im}\}$ for each query term c_i through a bilingual dictionary lookup, where $i \in [1, n]$. The structured query is constructed according to the following rules:

1. The translation sets E_i of all query terms are combined with the logical operator AND;
2. The candidate translations e_{ij} ($j \in [1, m]$) of a query term are enclosed in the parentheses and combined with the logical operator OR;
3. Phrases are enclosed in quotation marks as units.

2.2.2 Co-occurrence statistics using an HMM model

Boolean algebra queries are supported by most search engines, such as AltaVista, Yahoo, and Google. We used Google to fetch up to 300 top-ranked documents using the structured queries generated in Section 2.2.1. The retrieved documents are then filtered to remove HTML tags and metadata, leaving only the web text as the corpus to provide context for disambiguation.

Given a Chinese query $(c_1, c_2, ..., c_n)$, each English translation candidate set E is a sequence of words $(e_1, e_2, e_3, ..., e_n)$. We use a probability model $P(E) = P(e_1, e_2, e_3, ..., e_n)$ to estimate the maximum likelihood (ML) of each sequence of words. The English translations E with the highest $P(E)$ among all possible translation sets is selected.

The HMM model [6] is a general statistical modelling technique for sequence data and has been widely used in speech recognition applications for twenty years. We utilized a bigram HMM with a window size w to compute the probability of a sequence of words:

$$P(e_1, e_2, ..., e_n) = P(e_1) \prod_{a=2}^{n} P_w(e_a | e_{a-1})$$

where $P(e)$ is the probability of term e occurring within the corpus, and $P(e, e')$ is the probability of term e' occurring after term e within the corpus. The $P(e)$ can be calculated as follows:

$$P(e) = \frac{f(e)}{N}$$

where $f(e)$ is the corpus frequency of term e, N is the number of terms occurring in the corpus. The zero-frequency problem arises quite often in the context of probabilistic language models when the model encounters an event in a context where it has never been seen before. Smoothing provides a way to estimate and assign the probability to that unseen event. We calculated $P(e|e')$ by using the absolute discounting smoothing method described in [4]:

$$P(e|e') = max\left\{ \frac{f_w(e, e') - \beta}{N}, 0 \right\} + \beta P(e) P(e')$$

where $f_w(e, e')$ is the frequency of term e' occurring after term e within a window size w. The absolute discounting term β is equal to the estimate proposed by Ney et al. [7]:

$$\beta = \frac{n_1}{n_1 + 2n_2}$$

where n_k represents the number of terms with the collection frequency k.

The distance factor we employed in NTCIR-4 [10], that utilizes the term distance to discriminate between "strong" and "weak" term correlation, is completely omitted from the probability calculation. The reason for this is that the HMM model is a superior technique where sequence data is involved, and is not significantly improved by the addition of a decaying distance factor. Also the variation in window size used to collect word association information has a small effect on the outcome, with $w = 4$ producing the best results.

3 Experiments

The English document collection from the NTCIR–5 CLIR task contains 259,050 news articles from 2000 to 2001. There are 49 Chinese topics, each topic contains four parts: *title*, *description*, *narrative* and *key words*. The main issue we investigated was how well our techniques

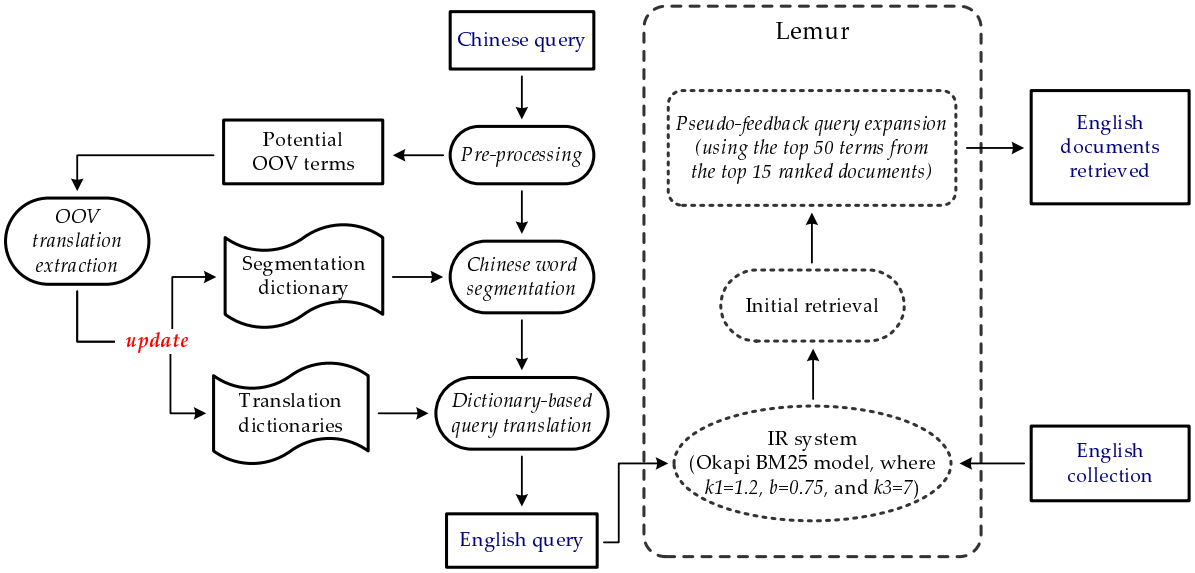

Figure 2. *The process of cross-language information retrieval*

perform when the web is used to provide context for disambiguation.

We used two dictionaries in our experiments: ce3 from the Linguistic Data Consortium[1], and the CEDICT Chinese-English dictionary[2] to translate Chinese queries into English. The basic method of query translation is the same we used in NTCIT-4. First, we detect Chinese OOV terms and add them into a Chinese segmentation dictionary for later use. In the dictionary-based query translation phase, we used the updated Chinese segmentation dictionary to segment the queries and replace each Chinese query term using a set of English translations through a bilingual translation dictionary lookup. Our CLIR process is shown in Figure 2.

3.1 Collection processing

English stop words were removed from the English document collection. We used a stop list that contains 477 entries and the Porter stemmer [8] to reduce words to stems.

3.2 Query processing

The Chinese queries were processed as follows:

Pre-processing. In title runs, each Chinese query was represented as a list of comma separated query key terms. Our assumption is that each query key term is a Chinese word. We found 80 out of 128 query key terms cannot be found in the translation dictionaries. We treat all of these as potential Chinese OOV terms, although some of them can be correctly translated word by word after segmentation. In description runs, each Chinese query was segmented in the first step. If any part of the query was segmented into single character sequences of length larger than 2, then the sequences and the tokens immediately adjacent (left and right) were extracted as potential Chinese OOV terms.

OOV translation extraction. Using these 80 query key terms as queries, we applied our OOV translation extraction technique, then added extracted Chinese OOV terms into a Chinese segmentation dictionary and Chinese–English translation pairs into the translation dictionary.

Chinese word segmentation. We compiled a segmentation dictionary using the two translation dictionaries and updated it at run time. A dictionary-based word segmentation method with greedy-parsing was employed to segment the Chinese queries.

Dictionary-based query translation. The translation dictionaries were used to replace each query term by all possible English translations. Our translation disambiguation technique, as described in section 2, was used to select the most appropriate translation for each Chinese query.

[1] http://www.ldc.upenn.edu/
[2] http://www.mandarintools.com/cedict.html

RunID	MAP		Recall		P@10	
	Rigid	*Relax*	*Rigid*	*Relax*	*Rigid*	*Relax*
T–*mono*	*0.4564*	*0.5066*	*2818*	*3748*	*0.5286*	*0.6265*
C–E–T–*collection*	**0.3702**	**0.4130**	2611	3466	0.4388	0.5347
C–E–T–*web*	0.3428	0.3855	2650	3520	0.4143	0.5204
D–*mono*	*0.4391*	*0.4897*	*2852*	*3757*	*0.5429*	*0.6367*
C–E–D–*collection*	0.3917	0.4379	2674	3463	0.4959	0.5980
C–E–D–*web*	**0.4042**	**0.4496**	2738	3530	0.4857	0.5837

Table 1. *NTCIR-5: Results of Official Runs*

3.3 Experimental design

Our CLIR retrieval experiments consist of four official runs. In C–E–T–*runs*, we have used the *title* of the Chinese topics as queries, and in C–E–D–*runs* the *description* fields are used as queries to retrieve the documents from the English document collection. We established a monolingual reference (T–*mono* and D–*mono*) by which we can measure our CLIR results. We then tested the translation disambiguation using the test collection (see C–E–T–*collection* and C–E–D–*collection*) and the English web documents extracted by a search engine, (see C–E–T–*web* and C–E–D–*web*), respectively.

The relevance judgements provided by NTCIR are at two levels — strictly relevant documents known as *rigid* relevance, and likely relevant documents, known as *relax* relevance. We report results for each of these. Our experiments used the Lemur IR system [3] developed by the Computer Science Department at the University of Massachusetts and the School of Computer Science at Carnegie Mellon University.

3.4 Results and discussion

The results of our experiments are shown in Table 1. The MAP values of C–E–T–*collection* and C–E–D–*web* runs provide the best results among all participants. The *rigid* relevance assessment MAP values for those two runs were 0.3702 and 0.4042, respectively representing 67.3% and 92% of monolingual retrieval effectiveness. The *relax* relevance assessment MAP values were 0.4130 and 0.4496, respectively representing 81.5% and 91.8% of monolingual retrieval effectiveness.

We used the Wilcoxon ranked signed test to test the statistical significance of our results. Our results show that using our new technique, we were able to perform translation disambiguation using the web, rather than a specific collection, with no significant loss of effectiveness.

[3] http://www.lemurproject.org/

4 Conclusions

In this paper, we presented our experience in the Chinese–English CLIR task in NTCIR-5. Rather than employ the English test collection, we make use of English web documents extracted by a search engine as a corpus to disambiguate dictionary translation. Our experimental results show that when using the web, it is possible to achieve effectiveness comparable to that obtained with a test collection.

References

[1] Mirna Adriani. Using statistical term similarity for sense disambiguation in cross-language information retrieval. *Information Retrieval*, Volume 2, Number 1, pages 67–68, 2000.

[2] Mohammed Aljlayl and Ophir Frieder. Effective arabic-english cross-language information retrieval via machine-readable dictionaries and machine translation. In *Proceedings of the 10th International Conference on Information and Knowledge Management*, pages 295–302, Atlanta, Georgia, USA, 2001. ACM Press.

[3] Lisa Ballesteros and W. Bruce Croft. Resolving ambiguity for cross-language retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 64–71, Melbourne, Australia, 1998. ACM Press.

[4] Marcello Federico and Nicola Bertoldi. Statistical cross-language information retrieval using n-best query translations. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 167–174, Tampere, Finland, 2002. ACM Press.

[5] Jianfeng Gao, Ming Zhou, Jian-Yun Nie, Hongzhao He and Weijun Chen. Resolving

query translation ambiguity using a decaying co-occurrence model and syntactic dependence relations. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 183–190, Tampere, Finland, 2002. ACM Press.

[6] David R. H. Miller, Tim Leek and Richard M. Schwartz. A hidden markov model information retrieval system. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 214–221, Berkeley, California, United States, 1999. ACM Press.

[7] Hermann Ney, Ute Essen and Reinhard Kneser. On structuring probabilistic dependencies in stochastic language modelling. *Computer Speech and Language*, Volume 8, Number 3, pages 1–38, 1994.

[8] Martin F. Porter. An algorithm for su#x stripping. *Automated Library and Information Systems*, Volume 14, Number 3, pages 130–137, 1980.

[9] Hsiao-Tieh Pu, Shui-Lung Chuang and Chyan Yang. Subject categorization of query terms for exploring web users' search interests. *Journal of the American Society for Information Science and Technology*, Volume 53, Number 8, pages 617–630, 2002.

[10] Ying Zhang and Phil Vines. RMIT Chinese-English CLIR at NTCIR-4. In *Working Notes of the Fourth NTCIR Workshop Meeting (NTCIR4)*, pages 60–64, Tokyo, Japan, 2004. National Institute of Informatics.

[11] Ying Zhang and Phil Vines. Using the web for automated translation extraction in cross-language information retrieval. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 162–169, Sheffield, UK, 2004. ACM Press.