

A New Hybrid Approach for Chinese-English Query Translation

Guo-Wei Bian and Hsin-Hsi Chen

*Department of Computer Science and Information Engineering
National Taiwan University*

Taipei, Taiwan, R.O.C.

Tel: +886-2-23625336 ext. 311; Fax: +886-2-23628167

Email: gwbian@nlg.csie.ntu.edu.tw, hh_chen@csie.ntu.edu.tw

Abstract

A new hybrid approach combining the dictionary-based and corpus-based approaches for Chinese-English cross-language information retrieval is proposed. The bilingual dictionary provides the translation equivalents of each query term. And the word co-occurrence information trained from a monolingual corpus can be used to disambiguate the translation. Further, we investigate the roles of phrase-level translation and short query by comparing the word-level translation and long query for different selection strategies. Our experiments have shown that the phrase-level translation is 14~31% more effective than the word-level translation. Most selection strategies perform better in the long queries than those in short ones. However, the simple select-all strategy has opposite result. The proposed method achieves 74.71% of the original monolingual English queries and 87.14% performance of the monolingual short version. Even though some multi-word concepts are not contained in the bilingual dictionary, our method still achieves near 70% monolingual effectiveness for different length of query at word-level translation.

1. Introduction

In recent years, the World Wide Web (WWW) breaks the boundaries of countries and very large online documents are available through the Internet. These multilingual textual resources have motivated the researches in Cross-Language Information Retrieval (CLIR) and online machine translation (MT) for the multilingual information accessing. A cross-language information retrieval system retrieves documents in a language that is different from the query language. Overall multilingual information accessing systems (Bian and Chen, 1997; David and Ogden, 1997) have been proposed to integrate the query translation of CLIR and MT technology to translate the queries and the retrieved documents on-the-flying.

Several approaches have been proposed for CLIR recently. There are four main approaches for query translation:

Dictionary-based approach (Ballesteros and Croft, 1997; David, 1996; Hull and Grefenstette, 1996; Kwok, 1997)

1. Corpus-based approach (David and Dunning, 1995; Landauer and Littman, 1990)
2. Hybrid approach (combined dictionary-based and corpus-based) (David, 1996)
3. Machine Translation based approach (MT-based) (Radwan, 1994)

The dictionary-based approach exploits the bilingual transfer dictionaries to select the target terms for source queries. The terms of query can be translated in two different levels of dictionary translation: word-level (word-by-word) and phrase-level translations. Different selection strategies: Select-All (Hull and

Grefenstette, 1996), Select-N (Ballesteros and Croft, 1996), and Select-Best-N (David, 1996; Hayashi, Kikui, and Susaki, 1997) may be adopted to translate the queries on word level. Further, Ballesteros and Croft (1997) have shown that the well-translated phrases can improve the effectiveness in phrase-level translation, but poorly translated phrases may negate the improvements.

Most of errors in dictionary-based approach are due to the following three factors: (Hull and Grefenstette, 1996; Ballesteros and Croft, 1997)

1. Missing terminology: the correct sense is not contained in the dictionary.
2. Translation ambiguity: the terms of dictionary translation are ambiguous and some extraneous terms are added to the query.
3. Failure in phrasal translation: failure to identify and translate the multi-term concepts (phrases) results in the ambiguities for the words of the multi-term concepts and reduces the performance.

Alternatively, the corpus-based approach uses parallel or comparable aligned corpora to disambiguate the word selection depending on the concurrence statistics between words. Landauer and Littman (1990) present a method based on Latent Semantic Indexing (LSI) for cross-language retrieval. This method uses the singular value decomposition of a parallel document collection to obtain the term vector representations, which are comparable across all the languages. Dumais, Littman, and Landauer (1997) use this method and extend testing for the dual English-French CLIR. David and Dunning (1995) used initial Spanish equivalents derived directly from a parallel corpus, and then the evolutionary programming methods are applied to refine Spanish translation of English queries by iteratively comparing the retrieval profiles of English and Spanish queries over a parallel corpus. But the results were comparatively poorer than the full transfer dictionary (Select-All) method under large-scale retrievals. And the evaluation optimization method was computationally expensive.

Generally, this corpus-based approach has four disadvantages. First, the parallel or comparable corpora are not always available. Second, the current available corpora tend to be relative small or cover only a small number of subjects. Third, the domain-dependent problem is involved between the query and the statistics of the corpora. Finally, the performance is dependent on how well the corpora are aligned.

David (1996) combines the POS disambiguation in the dictionary-based approaches and the corpus-based disambiguation to achieve 73.5% of performance of a monolingual system. At first, the system uses a part-of-speech (POS) tagger to select the Spanish potential equivalents from a bilingual lexicon for English query terms. A parallel corpus is then used to disambiguate the translated queries by choosing the Spanish terms that retrieve documents most like those retrieved for the English query. This combined method is more effective than previous ones.

However, this approach has the same problems as the corpus-based approach. Another problem is that the performance of a POS tagger is not good for short query in general. Because the queries tend to be very short (often only one or two words), the errors of tagging will decrease the performance of query translation. Because the POS tagging always produces errors for short queries, the target equivalents of query term may be filtered out. Such a method will not be suitable for short queries, especially searching on WWW.

Radman (1994) conducted experiments using the two methods: the term vector translation and the machine translation system (SYSTRAN). The experiments provide suggestion that the former method is more effective than machine translation.

Different language pairs for cross-language information retrieval have been evaluated. The language

pairs include: English-Spanish (David and Dunning, 1995; David, 1996; Ballesteros and Croft, 1997), English-French (Hull and Grefenstette, 1996), dual English-French (Dumais, Littman, and Landauer, 1997), German-Italian (Sheridan and Ballerini, 1996), Japanese-English (Hayashi, Kikui, and Susaki, 1997), and English-Chinese (Kwok, 1997). Most of the previous works are in the same Indian-European language family, and fewer ones are done for the different language families. Kwok (1997) evaluates an English-Chinese CLIR experiment which takes at most three translations of each word, one from each of the first three senses. If there are less than 3 senses, the synonyms are taken from the first, then the second until the system uses 3 translations or exhausts all definitions. The average precision of naive translation is 18.19%, and it is about 30% to 50% worse than good translation. For Japanese-English CLIR, Hayashi, Kikui, and Susaki (1997) proposed to search for a dictionary entry corresponding to the longest sequence of Japanese words from left to right. Then they choose the most frequently used word or phrase in a text corpus collected from WWW. But there is no report for this query translation approach.

In this paper, we will introduce a new hybrid approach and compare its performance with other selection strategies. A typical query translation for Chinese-English CLIR and different word selection methods are described in Section 2. The experiments using various methods on the word-level translation are discussed in Section 3. Section 4 touches on the phrasal translation to demonstrate the problems from missing multi-term concepts and failure in phrasal translation. In addition, the different selection strategies are evaluated with the short versions of queries. The overall experimental results are discussed in detail. Finally, Section 5 concludes the remarks.

2. Query Translation

The typical processing of query translation for Chinese-English CLIR consists of three major steps:

1. word segmentation: To identify the word boundary of the input stream of Chinese characters.
2. query translation: To construct the translated English query using the bilingual dictionary or the bilingual corpora. Translation disambiguation may be done using the monolingual corpus or the bilingual corpora.
3. monolingual IR: To search the relevant documents using the translated queries.

The segmentation and query translation use the same bilingual dictionary in this design. That speeds up the dictionary lookup and avoids the inconsistencies resulting from two dictionaries (i.e., segmentation dictionary and transfer dictionary). This bilingual dictionary has approximately 90,000 terms. The longest-matching method is adopted in Chinese segmentation. The segmentation processing searches for a dictionary entry corresponding to the longest sequence of Chinese characters from left to right. After identification of Chinese terms, the system selects some of the translation equivalents for each query term from the bilingual dictionary. The terms of query can be translated in two different levels of dictionary translations: word-level (word-by-word) and phrase-level translations. Those terms, missing from the bilingual dictionary, are passed unchanged to the final query.

When there is more than one translation equivalent in a dictionary entry, the following selection strategies are explored.

(1) Select-All (SA): The system looks up each term in the bilingual dictionary and constructs a translated query by concatenating of all the senses of the terms.

(2) Select-Highest-Frequency (SHF): The system selects the sense with the highest frequency in target language corpus for each term. Because the translation probabilities of senses for each term are unavailable without a large-scale word-aligned bilingual corpus, the translation probabilities are reduced to the probabilities of sense in the target language corpus. So, the frequently-used transferring sense of a term

is used instead of the frequently-translated sense.

(3) Select-N-POS-Highest-Frequency (SNHF): This strategy selects the highest-frequent sense of each POS candidate of the term. If the term has N POS candidates, the system will select N translation senses. Compared to this strategy, the strategy (2) always selects only one sense for each term.

(4) Word co-occurrence (WCO): This method classifies words on the basis of their co-occurrence

$$I(x, y) = \log_2 \frac{P(x, y)}{P(x)P(y)}$$

with other words. The translation of a query term can be disambiguated with the co-occurrence of its translation equivalents and other words' equivalents. The mutual information (MI) of word pairs reflects the word association norms in one language. If two words x and y have probabilities $P(x)$ and $P(y)$, their mutual information (Church and Hanks, 1990) is defined to be

This method considers the content around the translation equivalents within the text collection to decide the best target equivalent. The mutual information of word pairs is trained using a window size 3 in the CACM text collection. Totally, there are 247,864 word pairs.

Table 1 illustrates an example for different translation. The Chinese concept ‘奇异值分解’ and its phrase-level translation ‘singular value decomposition’ are employed. Four translated representations using different selection strategies on the word-level translation is shown in Table 1 (a). Column 3 shows the translation equivalents in transfer dictionary for the query terms at word-level. Table 1 (b) lists the mutual information of some word pairs of translation equivalents. Most of word pairs have no co-occurrence relations. Considering the example, the equivalent ‘singular’ of the term ‘奇异’ (ji-yi) has the largest MI score with all translation equivalents of other two words.

Our work takes a new hybrid approach that exploits a bilingual dictionary and a monolingual English corpus. The translation equivalents of each query term are retrieved from the bilingual dictionary. Then the co-occurrence relationship between the terms' equivalents can be used to disambiguate the translation of a query term. The mutual information of word pairs reflects the word association norms in one language. This corpus-based disambiguation using the monolingual corpus resolves the problems (translation ambiguity and failure in phrasal translation) of dictionary-based approach. Additionally, it avoids the difficulty of getting large available parallel or comparable corpora and the problems resulting in previous corpus-based and hybrid approaches. The word co-occurrence disambiguation can perform good translation even when the multi-term phrases are not contained in the bilingual dictionary or the phrases are not identified in the source language.

3. Experiments

Figure 1 shows the basic architecture of the query processing for CLIR in our experiments. These experiments use the SMART information retrieval system (Salton and Buckley, 1988), which measures the similarity of the query and each document using the vector space model. The query weights are multiplied by the traditional IDF factor. The test collection CACM is used to evaluate the performance of our approach and other selection methods. This collection contains 3204 texts and 64 queries in English. Each query has relevance judgements. The average number of words in the query is approximately 20.

Table 1. Different translations of Chinese concept ‘奇异值分解’ (singular value decomposition)

Table 1(a). Translated representations based on different strategies

Term	POS	SA	SHF	SNHF	WCO
奇异	N	oddity singularity		singularity	
	ADJ	singular	singular	singular	singular
值	N	value worth	value	value	value
分解	N	decomposition analysis dissociation cracking disintegration		decomposition	decomposition
	V	analyze anatomize decompose decompound disassemble dismount resolve	analyze	analyze	
	XV	(split up) (break up)		(split up)	

Table 1(b). The mutual information for some word pairs

word	Equivalentents		奇异			值		分解						
			w11	w12	w13	w21	w22	w31	w32	w33	w34	w35	w36	
奇异	oddity	w11												
	singular	w12				6.099		4.115	6.669					
	singularity	w13												
值	value	w21		6.099				1.823	4.377					
	worth	w22												
分解	analysis	w31		4.115		1.823								
	decomposition	w32		6.669		4.377								
	analyze	w33												
	decompose	w34												
	decompound	w35												
	resolve	w36												

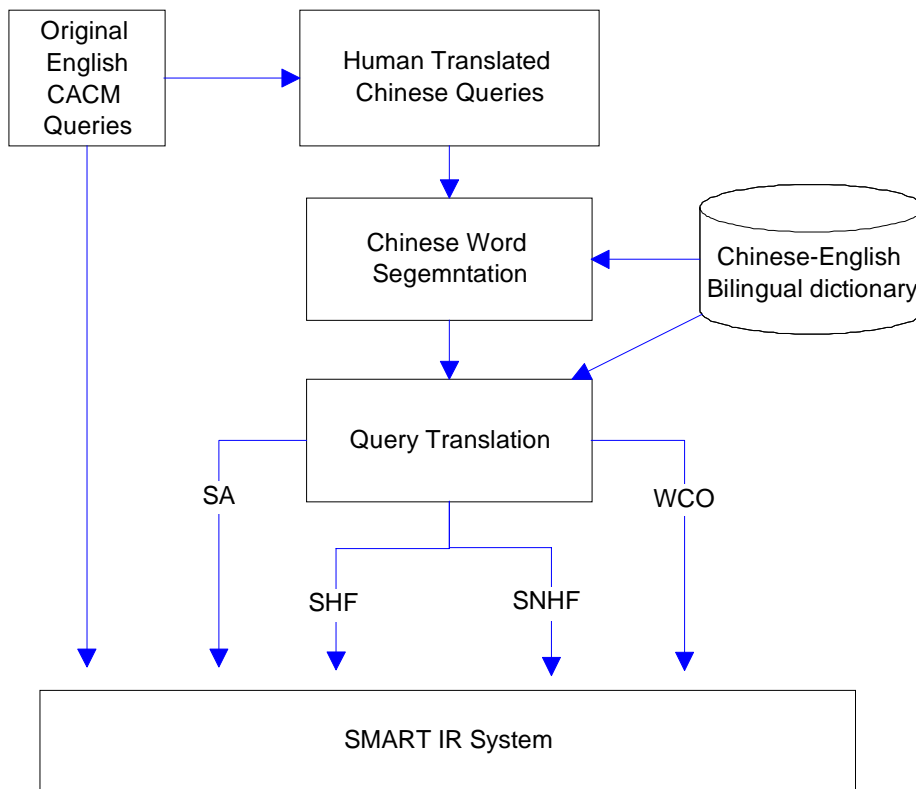


Figure 1. The diagram of query processing

In order to test the effectiveness of query translation, we create the Chinese queries by manually translating the original English queries to Chinese ones. The Chinese queries are regarded as the input queries later. Two examples of the original English queries and human translated Chinese ones are shown in Table 2. Our experiment compares the retrieval performance of the original English queries to the results of four translated versions of Chinese queries generated by the different selection methods. One example of the original English query, human translated Chinese version, and translated queries are shown in Table 3. It gives the segmented Chinese string and four automatically translated representations for the CACM Q1. Parentheses surround the English multi-term concepts and the brackets surround the translation equivalents of each term.

The performance of the various methods is shown in Table 4. The 11-point average precision values are listed by category. Rows 3 and 4 show the results compared with the monolingual retrieval and the simple Select-All method. The simple SA method can achieve 45.81% performance as well as the monolingual system. The SHF and SNHF methods achieve 61.18% and 54.02% respectively. The proposed WCO method achieves 65.18% performance of monolingual retrieval. It raises 42.28% effectiveness than the simple SA method does. On this word-level translation, some loss is due to the missing multi-term concepts in our bilingual dictionary.

Table 2. The original English and human translated Chinese queries of CACM Q1 and Q31

	Query String
Query 1	What articles exist which deal with TSS (Time Sharing System), an operating system for IBM computers? 那些文章是有關 TTS (分時系統), 一種 IBM 電腦的作業系統?
Query 31	I'd like to find articles describing the use of singular value decomposition in digital image processing. Applications include finding approximations to the original image and restoring images that are subject to noise. An article on the subject is H.C. Andrews and C.L. Patterson "Outer product expansions and their uses in digital image processing", American Mathematical Monthly, vol. 82. 我想要找敘述用於數位影像處理的奇異值分解的文章。應用包含尋找對於原來影像及有雜訊的影像修復的近似法。有關這主題的一篇文章是 H.C. Andrews 和 C.L. Patterson 發表在美國數學月刊第 82 卷上的 "外積的擴展及在數位影像處理上的使用".

Table 3. The Chinese query and four automatically translated representations for CACM Q1

Original Query	What articles exist which deal with TSS 'Time Sharing System', an operating system for IBM computers?
Chinese Query	那些文章是有關 TTS (分時系統), 一種 IBM 電腦的作業系統?
1 Segmentation	那些文章是有關 TTS '分時系統', 一種 IBM 電腦的作業系統?
2.1 SA	those article [be yes yah yep] about TTS '[minute cent apportion deal dissever sharing] time [formation lineage succession system]', [a ace mono] [class seed] IBM [computer computing] of [(operating system) (operation system) OS]
2.2 SHF	those article be about TTS 'deal time system', a class IBM computer of (operating system)
2.3 SNHF	those article [be yes] about TTS '[minute deal] time system', [a mono] class IBM computer of [(operating system) OS]
2.4 WCO	those article be about TTS 'sharing time system', a class IBM computer of (operating system)

Table 4. Average precision of word-level query translation

	Original English Query (Monolingual)	SA	SHF	SNHF	WCO
Average 11-point Precision	35.78%	16.39%	21.89%	19.33%	23.32%
% of Monolingual	baseline	45.81%	61.18%	54.02%	65.18%
% change		baseline	33.56%	17.94%	42.28%

Table 5. Average precision of phrase-level query translation

	Original English Query	SA	SHF	SNHF	WCO
Average 11-point Precision	35.78%	20.45%	26.41%	23.62%	26.73%
% of Monolingual	baseline	57.15%	73.81%	66.01%	74.71%
% change		baseline	29.14%	15.50%	30.71%

4. Phrasal Translation and Short Query

The experimental results (Hull and Grefenstette, 1996; Ballesteros and Croft, 1997) have shown that recognizing and translating multi-word expressions is crucial for CLIR. The reason is that the individual components of phrases often have different senses in translation. But the entire phrase has the distinct meaning for translation disambiguation. In this section, we discuss the comparison of performances based on word-level and phrase-level translations. In addition, the short queries are created to evaluate the behaviors of our strategies for the real queries.

4.1 Phrasal Translation

With the dictionary-based approach, three problems result in the major loss in effectiveness of 40-60% below that of monolingual IR (Hull and Grefenstette, 1996; Ballesteros and Croft, 1997). These factors are: (1) missing terminology, (2) translation ambiguity, and (3) the identification and translation of multi-term concepts (multi-word expressions) as phrases.

Among these factors, the correct identification and translation of multi-word expressions (MWE) make the biggest difference in average performance (Hull and Grefenstette, 1996). Although dictionaries contain a number of phrasal entries, there are many lexical phrases that are missing. These are typically the technical concepts and the terminology in specific domain. To compare the performances of the word-level translation and phrase-level translation, the CACM English queries are manually checked to find the multi-term concepts that are not contained in our bilingual dictionary. These phrases and their translations are added into the bilingual dictionary for the phrase-level experiments. Totally, 102 multi-word concepts (e.g., remote procedure call (遠端程序呼叫), singular value decomposition (奇異值分解), *etc.*) are identified in the CACM queries.

By the longest-matching method, the segmentation can handle the identification of these multi-word concepts easily within the string of Chinese characters. For example, the string '視窗管理器' (window manager) will be segmented as three words of string '視窗 管理 器' if the concept is not stored in the bilingual dictionary. When the concept appears in the bilingual dictionary, it will be segmented as a whole word '視窗管理器'.

Table 5 lists the performance of phrase-level query translation. The simple SA method can achieve 57.15% performance as well as the monolingual system. The SHF, SNHF, and WCO methods achieve 73.81%, 66.01%, and 74.71% respectively. The WCO method raises 30.71% effectiveness than the simple SA method does. The difference between WCO and SA methods is less than that in word-level experiments, because the translations of multi-term concepts are fixed in phrasal experiments. The phrase-level translation raises 24.77%, 20.65%, 22.19%, and 14.62% of performance for SA, SHF, SNHF, and WCO respectively. The WCO method obtains less from the phrasal translations than other methods do, because the WCO method can disambiguate some translations of multi-term concepts in word-level experiment. On the average, the phrase-level translation raises near 20% than the word-level translation.

4.2 Short Query

Researchers have recognized that most real queries are only a few words long (Hull and Grefenstette, 1996). Many previous works (Pinkerton, 1994; Fitzpatrick and Dent, 1997) have also shown this phenomenon in searching on WWW. Over a wide range of operational environments, the average terms of user-supplied queries are 1.5 ~ 2 words and rarely more than 4 words. Hull and Grefenstette (1996) work with the short versions of queries (average length of seven words) from French to English in TREC experiments. But no comparison of the short and long queries is available. To evaluate the behavior of user's short queries, we make additional experiments to compare with the results of the original long queries.

Three researchers help us to create the English and Chinese versions of short queries from the original English queries of CACM. On the average, the short query has near 4 words, including single-word terms and multi-term concepts. The short version of English queries is regarded as the baseline to compare the results of translated queries of the short Chinese queries. Table 6 shows the four versions of CACM query 31.

The Chinese query can be translated on phrase-level and word-level. The equivalents of query term are selected by four different selection strategies. The performance of word-level translation for short version of CACM queries is listed in Table 7. Table 8 shows the performance of phrase-level translation. The 11-point average precision of the monolingual short English queries is 29.85%. It achieves the 83.42% performance of the original English queries. The WCO strategy gets 72.96% performance of the monolingual English short version on word-level translation and 87.14% performance on phrase-level translation. The simple SA method achieves 61.24% and 78.25% respectively. The differences between various selection methods of query translation in short queries are less than those in long queries. In other words, the simple SA method combining phrase-level translation is an acceptable approach of CLIR if the query is short and the domain-dependent concepts are included in bilingual dictionary.

In the experiments for the short queries, all of the selection strategies perform better to obtain higher performances of the monolingual results than the long ones. This is because the users often give more specific terms in short queries. However, the query translation of long query adds more extraneous terms to the query.

The phrase-level translation raises 27.78%, 27.39%, 31.57%, and 19.42% of performance for SA, SHF, SNHF, and WCO respectively than the word-level translation does. On the average, the phrase-level translation raises near 26% performance of the word-level translation. Compared with those experiments for the long queries, the phrase-level translation plays more important role in short queries.

4.3 Overall Results

The overall results are shown in Figure 2. The 11-point average precision of the monolingual short English queries is 29.85%. It achieves the 83.42% performance of the original English queries. In word-level experiments, the best WCO (word co-occurrence) strategy gets the 72.96% performance of the monolingual English short version and 65.18% of the monolingual original English version. In phrase-level, the WCO achieves 87.14% and 74.71% respectively. The SHF, SNHF, and WCO selection strategies perform better in the long queries than that in short ones. However, the simple SA strategy has opposite result. Because users give more specific terms in short queries, the SA strategy introduces less extraneous terms to the query.

Table 6. Four versions of CACM query 31: Original, Short English, Chinese, Short Chinese

Type	Query
Original	I'd like to find articles describing the use of singular value decomposition in digital image processing. Applications include finding approximations to the original image and restoring images that are subject to noise. An article on the subject is H.C. Andrews and C.L. Patterson "Outer product expansions and their uses in digital image processing", American Mathematical Monthly, vol. 82.
Chinese	我想要找敘述用於數位影像處理的奇異值分解的文章。應用包含尋找對於原來影像及有雜訊的影像修復的近似法。有關這主題的一篇文章是 H.C. Andrews 和 C.L. Patterson 發表在美國數學月刊第 82 卷上的 "外積的擴展及在數位影像處理上的使用"。
Short English	singular value decomposition, digital image processing, noise.
Short Chinese	奇異值分解, 數位影像處理, 有雜訊的影像修復。

Table 7. Average precision of word-level translation for short query

	Short English Query	SA	SHF	SNHF	WCO
Average 11-point Precision	29.85%	18.28%	19.57%	17.42%	21.78%
% of Monolingual	baseline	61.24%	65.56%	58.36%	72.96%
% change		baseline	7.06%	-4.70%	19.15%

Table 8. Average precision of phrase-level translation for short query

	Short English Query	SA	SHF	SNHF	WCO
Average 11-point Precision	29.85%	23.36%	24.93%	22.92%	26.01%
% of Monolingual	baseline	78.25%	83.52%	76.78%	87.14%
% change		baseline	6.72%	-1.88%	11.34%

11-point average precision (%)

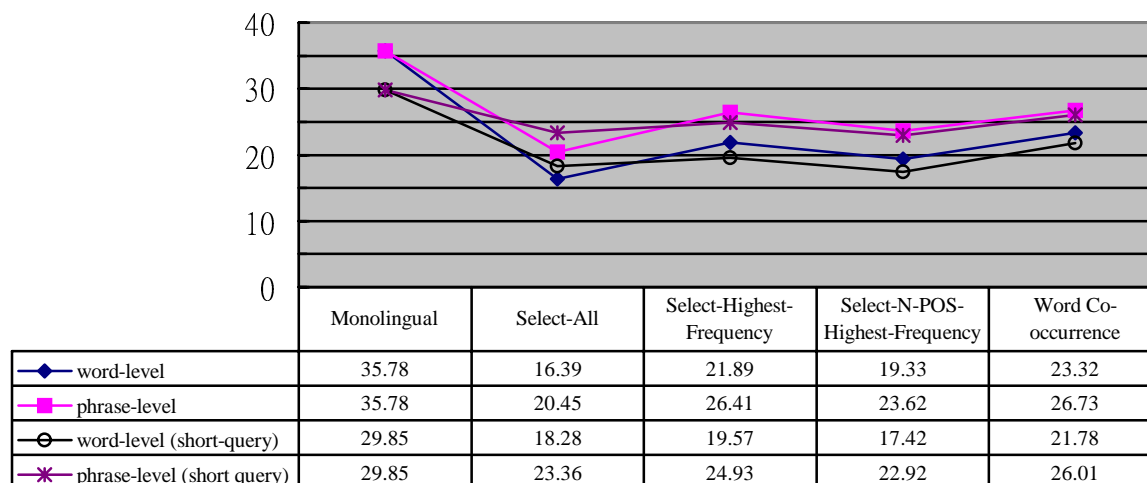


Figure 2. The comparison of query translations at different levels and the short queries

Alternatively, the phrase-level translation raises 14~31% performance than the word-level translation does in Chinese-English CLIR. Combining the phrase dictionary and co-occurrence disambiguation can bring CLIR performance up to 74.71% of monolingual retrieval in long query and 87.14% of monolingual retrieval in short query. Recall that the multi-word concepts and their translations are added to the dictionary in our experiments after domain expert has examined the queries. Hence the coverage of bilingual phrasal dictionary will affect the performance of CLIR. Even though the bilingual dictionary does not contain these multi-word concepts, the WCO method still achieves near 70% monolingual effectiveness for different length of query at word-level translation.

5. Conclusion and Future Work

This paper proposes a new hybrid approach combining the dictionary-based and corpus-based approaches for Chinese-English cross-language information retrieval. The bilingual dictionary provides the translation equivalents of query term. And the word co-occurrence information trained from a monolingual corpus can be used to disambiguate the translation. Further, we investigate the roles of phrase-level translation and short query by comparing the word-level translation and long query. The average length of query is

approximately 20 words in the original CACM query and 4 words in human created short query.

Our experiments have shown that the phrase-level translation is 14~31% more effective than the word-level translation in Chinese-English CLIR. This result illustrates that the multi-word concepts play the important role in CLIR. In other way, the SHF, SNHF, and WCO selection strategies perform better in the long queries than that in short ones. However, the simple SA strategy has opposite result.

The proposed WCO strategy combining the phrase dictionary can achieve 74.71% of the original monolingual English queries and 87.14% performance of the monolingual short version. The experimental results have shown the effectiveness of our approach in query translation, especially for short query. Even though some multi-word concepts are not contained in the bilingual dictionary, the WCO method still achieves near 70% monolingual effectiveness for different length of query at word-level translation. Our experiments have illustrated the WCO method can be adopted for CLIR to retrieve information in English using Chinese query with the search engines on WWW, because the real queries are always short (average length of two words).

Some problems of CLIR for different pairs of language families are not covered in this paper. Especially, the identification and translation of proper names play important roles in CLIR between Eastern-Asian and Western-European language families. In the languages of the same Western-European family, the proper nouns are often the loan words from another language. (David, 1996) However, the proper nouns in the languages of Western-European family are usually transliterated in Eastern-Asian languages. For example, the company name 'Intel' will be transliterated as '英代爾' (ying1-dai4-er3) or '英特尔' (ying1-te4-er3) in Chinese. How to identify the proper nouns is important for Chinese word segmentation. Our previous works (Chen and Lee, 1996; Chen and Bian; 1997) present various strategies to identify and classify Chinese proper nouns and transliterated names. However, a proper noun may have different transliterated equivalents by different users. Some of the transliterated equivalents could not be found in the bilingual dictionary. How to match other equivalents with the same or similar pronunciation for the transliterated names; and how to translate those unfound transliterated names play important roles for CLIR in different families. We will investigate these problems in retrieving English documents by Chinese query in the future.

References

- Ballesteros, L. and Croft, W.C. (1996) "Dictionary-Based Methods for Cross-Lingual Information Retrieval." In *Proceedings of the 7th International DEXA Conference on Database and Expert Systems Applications*, pp. 791-801. 1996.
- Ballesteros, L. and Croft, W.C. (1997) "Phrasal Translation and Query Expansion Techniques for Cross-Language Information retrieval." In *Proceedings of ACM SIGIR'97*, pp.84-91, 1997.
- Bian, G.W. and Chen, H.H. (1997) "An MT Meta-Server for Information Retrieval on WWW." *Working Notes of the AAAI Spring Symposium on Natural Language Processing for the World Wide Web*, Palo Alto, California, USA, March, 1997, pp.10-16.
- Chen, H.H. and Bian, G.W. (1997) "Proper Name Extraction from Web Pages for Finding People in Internet." In *Proceedings of the 10th Research on Computational Linguistics (ROCLING X) International Conference*, 1997, pp. 143-158.
- Chen, H.H. and Lee, J.C. (1996) "Identification and Classification of Proper Nouns in Chinese Texts." *Proceedings of 15th International Conference on Computational Linguistics*, 1996, pp. 222-229.
- David, M.W. and Dunning, T. (1995) "A TREC Evaluation of Query Translation Methods for Multi-Lingual Text Retrieval." In *Proceedings of the Fourth Text Retrieval Evaluation Conference (TREC-4)*, Gaithersburg, MD, National Institute of Standards and Technology.
- David, M.W. (1996) "New Experiments in Cross-Language Text Retrieval at New Mexico State University's Computing Research Laboratory." In *Proceedings of the Fifth Text Retrieval Evaluation Conference (TREC-5)*, Gaithersburg, MD, National Institute of Standards and Technology.

- David, M.W. and Ogden, W.C. (1997) "QUILT: Implementing a Large-Scale Cross-Language Text Retrieval System." In *Proceedings of ACM SIGIR'97*, 1997, pp.92-98.
- Dumais, S.T., Littman, M.L., and Landauer, T.K. (1997) "Automatic Cross-Language Retrieval Using Latent Semantic Indexing." *Working Notes of the AAAI-97 Spring Symposium on Cross-Language Text and Speech Retrieval*, 1997, pp. 18-24.
- Fitzpatrick, L. and Dent, M. (1997) "Automatic Feedback Using Past Queries: Social Searching?" In *Proceedings of ACM SIGIR'97*, 1997, pp.306-313.
- Hayashi, Y., Kikui, G., and Susaki, S. (1997) "TITAN: A Cross-Linguistic Search Engine for the WWW." *Working Notes of the AAAI-97 Spring Symposium on Cross-Language Text and Speech Retrieval*, 1997, pp. 58-65.
- Hull, D.A. and Grefenstette, G. (1996) "Querying Across Languages: A Dictionary-Based Approach to Multilingual Information Retrieval." In *Proceedings of ACM SIGIR'96*, pp.49-57, 1996.
- Kwok, K.L. (1997) "Evaluation of an English-Chinese Cross-Lingual Retrieval Experiment." *Working Notes of the AAAI-97 Spring Symposium on Cross-Language Text and Speech Retrieval*, 1997, pp. 110-114.
- Landauer, T.K. and Littman, M.L. (1990) "Fully Automatic Cross-Language Document Retrieval." In *Proceedings of the Sixth Conference on Electronic Text Research*, pp. 31-38, 1990.
- Pinkerton, B. (1994) "Finding What People Want: Experiences with the WebCrawler." In *Proceedings of WWW '94*, 1994.
- Radwan, K. (1994) *Vers l'Acces Multilingue en Langage Naturel aux Baess de Donnees Textuelles*. PhD Thesis, Universite de Paris-Sud, Centre d'Orsay. 1994.
- Salton, G. and Buckley, C. (1988) "Term Weighting Approaches in Automatic Text Retrieval." *Information Processing and Management* 5(24): 513-523, 1988.
- Sheridan, P. and Ballerini, J.P. (1996) "Experiments in Multilingual Information Retrieval Using the SPIDER system." In *Proceedings of ACM SIGIR'96*, pp.58-65, 1996.