行政院國家科學委員會專題研究計畫研究成果報告

國際數位圖書館合作研究計畫—

IDLP 英中雙語資訊系統相關語言處理技術和資源整合之研究

計畫類別：☐ 個別型計畫　　☑ 整合型計畫
計畫編號：NSC 89-2750-P-002-016-ZZ
執行期間：89 年 8 月 1 日至 90 年 7 月 31 日

整合型計畫：總計畫主持人：項　潔
子計畫主持人：陳光華
子計畫共同主持人：陳信希

處理方式：☑ 可立即對外提供參考
☐ 一年後可對外提供參考
☐ 兩年後可對外提供參考
（必要時，本會得展延發表時限）

執行單位：國立台灣大學圖書資訊學系

中華民國 90 年 10 月 1 日

ABSTRACT

The purpose of this project is to investigate and propose a practical approach dealing with the increasingly important issue – cross-language information retrieval (CLIR), especially in the Internet environment. The traditional approaches of select-all, select-one, and select-top-n are investigated for performance in cross-language information retrieval. Two implicit problems in CLIR have also be made significant in this research: one the translation ambiguity; the other is target polysemy. New resolutions to the problems are proposed and a series of experiments are carried out. The results show that the proposed approaches are significant and promising.

ACKNOWLEDGMENT

We would like to thank research assistants for their efforts in making the research fruitful. Many thanks go to the National Science Council for the support to this research.

CONTENTS

ILLUSTRATIONS

TABLES

Chapter 1

Introduction

Internet and digital libraries make available heterogeneous collections in various languages. They provide many useful and powerful information dissemination services. Specially, the World Wide Web (WWW) breaks the boundaries of countries and provides a very large number of online documents (more than 10 million documents) in multiple languages. A number of search engines (e.g., AltaVista, Excite, Infoseek, Lycos, Yahoo, *etc.*) and information discovery systems (Bowman *et al.*, 1995; Selberg and Etzioni, 1995; Yuwono, 1995) have been introduced on the Internet for users to locate interesting and relevant information. However, about 80% of Web sites are in English and about 40% of Internet users do not speak English (Euro-Marketing Associates, 1999; Grimes, 1996; Hershman, 1998). Language barrier becomes the major problem for people to search, retrieve, and understand materials in different languages. That decreases the dissemination power of the WWW to some extent. To resolve the problem of language barrier, research in Cross-Language

Information Retrieval (CLIR) has been motivated by these multilingual textual resources to build the multilingual information access system.

1.1 Requirements of a Multilingual Information System

Several issues have to be addressed to design a multilingual information processing system. They cover the following basic operations for multilingual data management (Bian and Chen, 1998c).

1. Data Representation: character sets and coding systems
2. Data Input: input methods and transliterated input
3. Data Display and Output: font mapping
4. Data Manipulation: the application must be able to handle the different coding characters
5. Query Translation: to translate the information need of users
6. Document Translation using Machine Translation (MT): to translate documents

The first three requirements have been resolved by system applications in several computer operating systems. Some of applications and packages can

also handle both single-byte and multiple-byte coding systems for Indo-European and Eastern-Asian languages. However, the language barrier becomes the major problem for people to access the multilingual documents. How to incorporate the capability of language translation to meet the requirements 5 and 6 becomes indispensable for multilingual systems.

1.2 Previous Work

Traditionally, information retrieval (IR) system retrieves the relevant documents from the input query both in the same language. In recent years, a large number of multilingual documents are available on the World Wide Web. Cross language information retrieval (CLIR) (Oard and Dorr, 1996; Oard, 1997) deals with the use of queries in one language to access documents in another. Overall multilingual information access systems (Bian and Chen, 1997; David and Ogden, 1997) have been proposed to integrate query translation of CLIR and MT technology to translate the queries and the retrieved documents on-the-fly.

Due to the differences between source and target languages, query translation is usually employed to unify the language in queries and documents. Several approaches have been proposed for CLIR recently. There are four main approaches for query translation:

1. Dictionary-based approach (Ballesteros and Croft, 1997; David, 1996; Hull and Grefenstette, 1996; Kwok, 1997)
2. Corpus-based approach (David and Dunning, 1995; Landauer and Littman, 1990)
3. Hybrid approach (combined dictionary-based and corpus-based) (Ballesteros and Croft, 1998; Bian and Chen, 1998c; David, 1996; Kraaij and Hiemstra, 1997)
4. Machine Translation based approach (MT-based) (Radwan, 1994; Oard, 1998)

The dictionary-based approach exploits bilingual transfer dictionaries to select the target terms for source queries. The terms of a query can be translated on two different levels of dictionary translations: 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) show the importance of phrasal translation and query expansion techniques for query translation. The well-translated phrases can improve the effectiveness on 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 co-occurrence statistics among source words and target 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) use 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 the 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 a query term may be filtered out. Such a method will not be suitable for short queries, especially searching on WWW.

Alternatively, some new hybrid methods (Ballesteros and Croft, 1998; Bian and Chen, 1998c; Kraaij and Hiemstra, 1997) are proposed to employ dictionary-based translation and disambiguation using co-occurrence statistics trained from target language documents. The hypothesis is that the correct translations of query translation will co-occur in target language documents and incorrect translations should tend not to co-occur. (Ballesteros and Croft, 1998) Kraaij and Hiemstra use a Noun Phrase (NP) extraction to build a list of noun phrases from the retrieval TREC-6 collection. The query is tagged and NPs are extracted, then the translation of the NPs from the query is disambiguated using the collected target NP list. Ballesteros and Croft first identify phrasal units, and disambiguate the terms using the context of the source phrasal unit. The co-occurrence statistics is trained from a text window size of 250 terms.

Our previous work (Bian and Chen, 1998c) presents a new hybrid method combining the dictionary-based and corpus-based approaches for Chinese-English

cross-language information retrieval. A bilingual dictionary provides the translation equivalents of each query term. And the word co-occurrence information trained from target language text collection can be used to disambiguate the translation of query. This method considers the content around the translation equivalents to decide the best target word. The translation of a query term can be disambiguated using the co-occurrence of the translation equivalents of this term and other terms. We adopt mutual information (Church, *et al.*, 1989) to measure the strength. The mutual information table is trained using a window size 3 for adjacent words in an English text collection. This disambiguation method performs good translations even when the multi-term phrases are not found in the bilingual dictionary, or the phrases are not identified in the source language. Later, we will adopt this method to resolve translation ambiguity.

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. Oard (1998) compares the performance of the different dictionary-based query translation techniques, MT-based query translation method, and MT-based document translation method on TREC-6 collection. The MT-based document translation method performs better than other methods

on long query.

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 that 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.

Translation ambiguity and target polysemy are two major problems in CLIR.

Translation ambiguity results from the source language, and target polysemy occurs in target language. All the above approaches deal with the translation ambiguity problem in query translation. Few touch on translation ambiguity and target polysemy together.

1.3 Four-Layer Multilingual Information System (MLIS)

Figure 1.1 shows a four-layer multilingual information system. We put the different types of processing systems on the four layers: (Bian and Chen, 1998c)

Layer 1: Language Identification (LI)

Layer 2: Text Processing Systems

Level 3: Language Translation Systems

Level 4: User Interface (UI)

Because most of natural language processing techniques (e.g., lexical analysis, parsing, *etc.*) are dependent on the language of processed document, the layer 1 resolves language identification problem before text processing. The language identification system employs cues from the different character sets and coding systems of languages. At layer 2, the systems may perform information extraction, information filtering, information retrieval, text classification, text summarization, or other text processing tasks.

Some of the text processing systems may have interaction with another one. For example, the relevant documents retrieved by IR system can be summarized to users. Additionally, a multilingual text processing system should be able to handle the different coding characters to match the requirement 4 (data manipulation). Several search engines (e.g., AltaVista, Infoseek, *etc.*) have the ability to index the documents of multiple languages. The language translation systems at layer 3 are used to translate the information need of users for text processing systems and translate the resultant documents from text processing systems to users in their native languages. The user interface is the closest layer to users. It gets the user's information need (included parameters, query and user profile) and displays the resultant document to user.

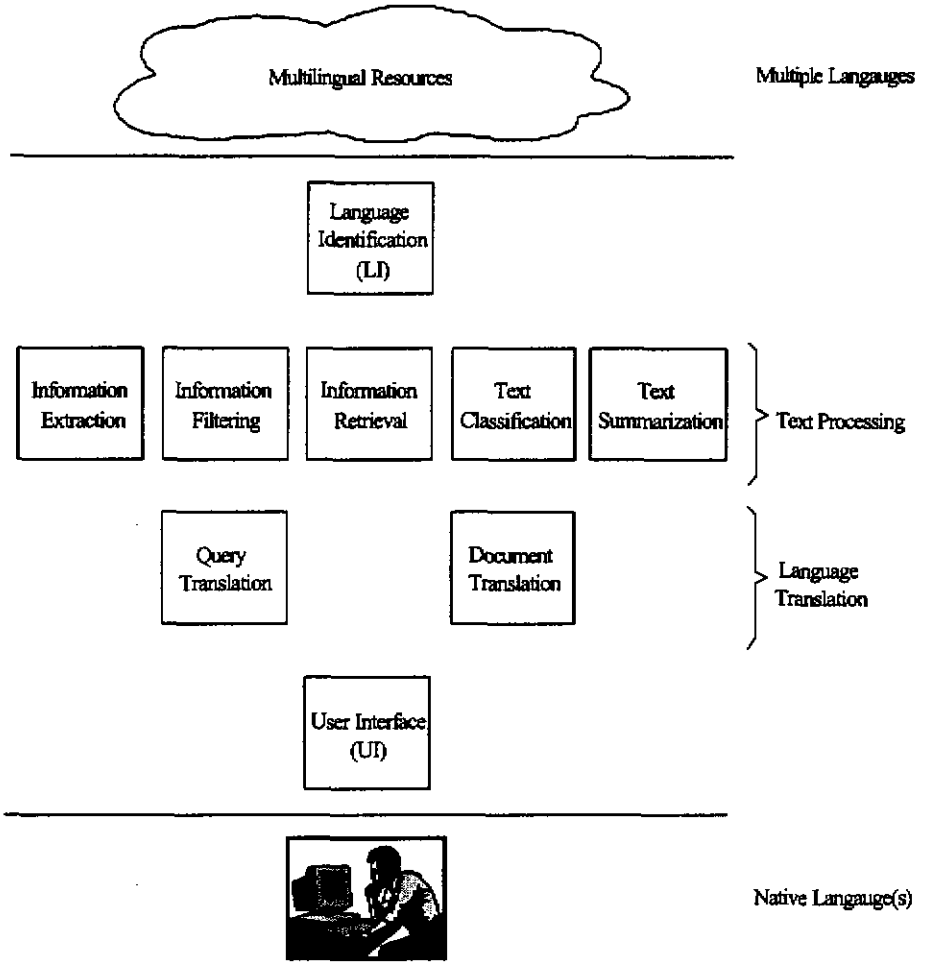


Figure 1.1 A Four-Layer Model of Multilingual Information System (MLIS)

Chapter 2

Query Translation: Translation Ambiguity Resolution

In this chapter, we present 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 each query term. And the word co-occurrence information trained from the retrieval document collection or a monolingual corpus can be used to disambiguate the translation. This new hybrid approach is evaluated and compared with other selection strategies. 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.

Section 2.1 introduces the different approaches of query translation. Section 2.2 describes a typical query translation for Chinese-English CLIR and different word selection methods. The experiments using various methods on the word-level translation are discussed in Section 2.3. Section 2.4 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. Section 2.5 discusses the overall experimental results in detail. Finally, Section 2.6 concludes the remarks.

2.1 Introduction

Cross language information retrieval (CLIR) deals with the use of queries in one language to access documents in another. Due to the differences between source and target languages, query translation is usually employed to unify the language in queries and documents. Several approaches have been proposed for query translation in CLIR recently. Most of errors in dictionary-based approach are due to the following three factors: missing terminology, translation ambiguity, and failure in phrasal translation (Hull and Grefenstette, 1996; Ballesteros and Croft, 1997). The ambiguities in translating words and multi-term concepts reduce the retrieval performance. The corpus-based and hybrid approaches use parallel or comparable aligned corpora to disambiguate the word selection depending on the concurrence statistics between words. Both approaches have the difficulty of obtaining large available bilingual corpora and the

domain-dependent problem, because the currently available corpora tend to be relative small or cover only a small number of subjects. David (1996) proposed the hybrid approach combining the POS disambiguation in the dictionary-based approaches and the corpus-based disambiguation. Because the queries tend to be very short (often only one or two words) and the POS tagging always produces errors for short queries, the target equivalents of query term may be filtered out. The errors of tagging will decrease the performance of query translation for short query.

In what follows, we will introduce a new hybrid approach and compare its performance with other selection strategies. This new method uses a monolingual corpus to solve the problems (translation ambiguity and failure in phrasal translation) in purely dictionary-based approach. Additionally, it avoids the difficulty of obtaining large available parallel or comparable corpora and the problems resulting in previous corpus-based and hybrid approaches.

2.2 Query Translation for Chinese-English CLIR

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

word segmentation: To identify the word boundary of the input stream of

Chinese characters.

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.

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 on 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 considers the content around the translation equivalents to decide the best target equivalent. (Bian and Chen, 1998a) The translation of a query term can be disambiguated using the co-occurrence of the translation equivalents of this term and other terms. We adopt mutual information (MI) (Church, *et al.*, 1989) to measure the strength.

$$MI(X,Y) = \log_2 \frac{P(X,Y)}{P(X)P(Y)}$$

MI is defined as follows:

where X and Y denote two terms,

P(X) and P(Y) are probabilities of X and Y,

P(X,Y) is their co-occurrence probability.

If MI(X,Y) >> 1, X and Y have strong relationship; if MI(X,Y) ≈ 0, X and Y have no relationship; and if MI(X,Y) << 0, X and Y are negatively correlated. The mutual information can be calculated from the retrieval document collection to prevent the domain shift problems in traditional corpus-based approach for query translation.

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. 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 obtaining 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 translations even when the multi-term phrases are not contained in the bilingual dictionary or the phrases are not identified in the source language.

2.3 Experiments

Figure 2.1 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 (Fox, 1990) is used to evaluate the performance of our approach and other selection methods. This collection contains 3204 texts and 64 queries both in English. Each query has the relevant judgements for evaluation. The average number of words in the query is approximately 20. We create the Chinese queries by translating the original English queries into Chinese ones manually. The Chinese queries are regarded as input queries. The mutual information is trained using a window size 3 for adjacent words in the text collection. Totally, there are 247,864 distinct word pairs.

The input Chinese query is segmented into several terms, and then translated to four possible representations using various selection methods. Table 2.1

illustrates an example for the different selection strategies. The Chinese concept '奇異值分解' (ji1yi4 zhi2 fen1jie3) and its phrase-level translation 'singular value decomposition' are employed. In Table 2.1(a), column 3 lists the translation equivalents in bilingual dictionary for the query terms at word-level. Four translated representations using different selection strategies on the word-level translation are shown.

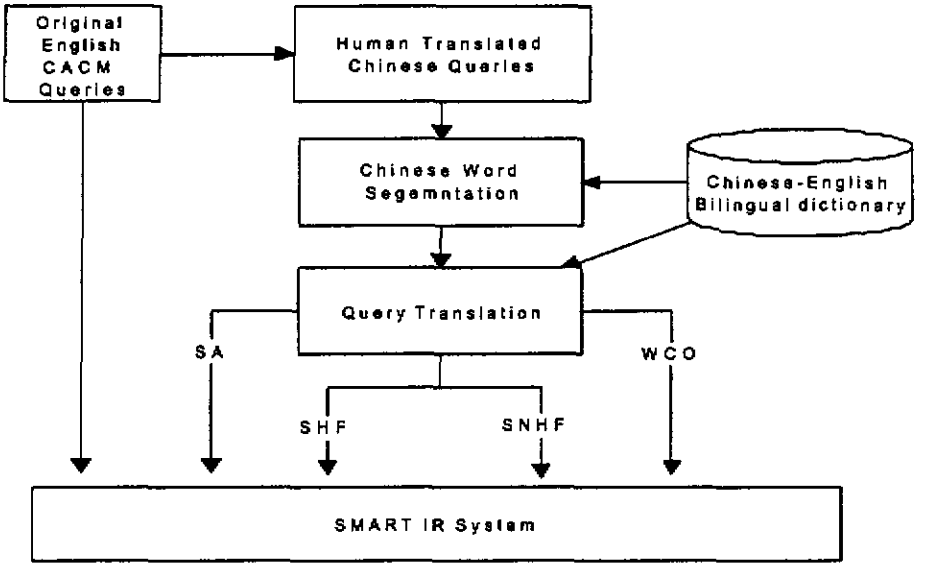


Figure 2.1 Diagram of Query Processing

Table 2.1(b) lists the mutual information of some word pairs of translation equivalents. The MI scores of word pairs 'singular value', 'singular analysis', 'singular decomposition', 'value analysis', and 'value decomposition' are 6.099,

4.225, 6.669, 1.823, and 4.377, respectively. Other word pairs have no co-occurrence relations in CACM text collection. Considering the example, the translation equivalent 'singular' of the term '奇異' (ji1yi4) has the largest MI score with all translation equivalents of the other two words.

Table 2.1 Different translations of Chinese concept '奇異值分解'

Table 2.1(a) Translated representations based on different strategies

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

Table 2.1(b) The mutual information for some word pairs

			奇異 (jiyi)			值(zhi)		分解(fenjie)					
word	Equivalents		w11	w12	w13	w21	w22	w31	w32	w33	w34	w35	w36
奇異 (jiyi)	oddity	w11											
	singular	w12				6.099		4.115	6.669				
	singularity	w13											
值 (zhi)	value	w21		6.099				1.823	4.377				
	worth	w22											
分解 (fenjie)	analysis	w31		4.115		1.823							
	decomposition	w32		6.669		4.377							
	analyze	w33											
	decompose	w34											
	decompound	w35											
	resolve	w36											

In order to test the effectiveness of query translation, 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 2.2. 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 2.3. 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.

Table 2.2 The English and 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 2.3 The Chinese query and four translations 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 2.4 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 2.4 shows the performance of the various methods. 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. The performance is 42.28% better than that of the simple SA method. On this word-level translation, some loss is due to the missing multi-term concepts in our bilingual dictionary.

2.4 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.

2.4.1 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., singular value decomposition (奇異值分解, ji1yi4 zhi2 fen1jie3), digital image processing (數位影像處理, shu4wei4 ying3xiang4 chu3li3), *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 '數位影像處理' (shu4wei4 ying3xiang4

chu3li3, digital image processing) will be segmented into three words '數位 影像 處理' (shu4wei4 ying3xiang4 chu3li3) if the concept is not stored in the bilingual dictionary. When the concept appears in the bilingual dictionary, it will be considered as a whole word '數位影像處理' (shu4wei4ying3xiang4chu3li3) instead.

Table 2.5 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.

2.4.2 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 design additional experiments to compare with the results of the original long queries. Three researchers helped 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 approximately 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 2.6 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 2.7. Table 2.8 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.

Table 2.5 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%

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.

Table 2.6 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 2.7 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 2.8 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%

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.

2.5 Overall Results

The overall results are shown in Figure 2.2. 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

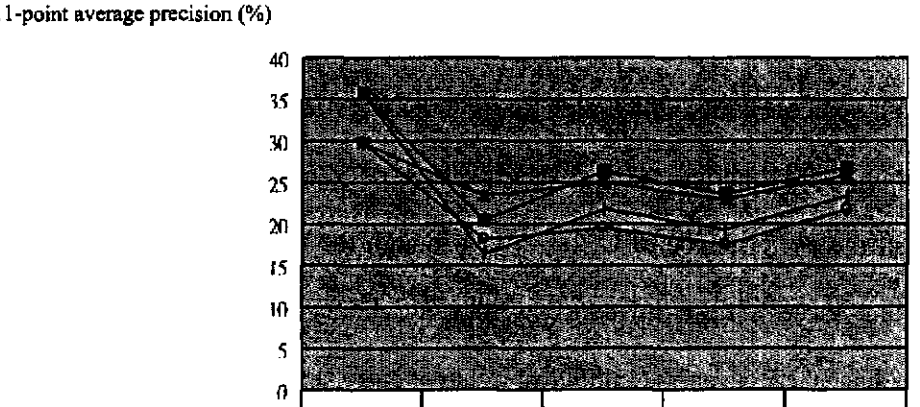
English version. In phrase-level, the WCO achieves 87.14% and 74.71% respectively. The SHF, SNHF, and WCO selection strategies perform better in the long queries than that in short ones. However, the simple SA strategy has opposite result. Because users give more specific terms in short queries, the SA strategy introduces less extraneous terms to the query.

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.

2.6 Feasibility and Portability

To test the feasibility and portability in other domain, we adopt the same methodologies on different document collection and queries. The TREC-6 text collection and TREC topics 301-350 (Harman, 1997) are used to evaluate the

performance. The text collection contains 556,077 documents, and is about 2.2G bytes. The collection is also employed to calculate the co-occurrence statistics using a context window size 3. Totally, there are 8,273,633 distinct word pairs. A TREC topic is composed of several fields. The fields of title and description are regarded as queries. Because the goal is to evaluate the performance of Chinese-English information retrieval on different models, we translate these 50 English queries into Chinese by human. Then the 50 human-translated Chinese queries are processed as input queries.



	Monolingual	SA	SHF	SNHF	WCO
word-level	35.78	16.39	21.89	19.33	23.32
phrase-level	35.78	20.45	26.41	23.62	26.73
word-level (short-query)	29.85	18.28	19.57	17.42	21.78
phrase-level (short query)	29.85	23.36	24.93	22.92	26.01

Figure 2.2 Retrieval Performances of Query Translations for the Long Queries and Short Queries on Different Levels of Translations

Table 2.9 Average Precision of Word-Level Query Translation on TREC-6

	Original English Query (Monolingual)	SA	SHF	SNHF	WCO
Average 11-point Precision	14.49%	6.52%	8.65%	8.57%	9.78%
% of Monolingual	baseline	45.00%	59.70%	59.14%	67.49%
% change		baseline	+32.67%	+31.44%	+50.00%

Table 2.9 shows the retrieval performance of different methods for the 50 queries. The 11-point average precision of the monolingual retrieval is 14.49%. The performance of SA, SHF, SNHF, and WCO are 6.52%, 8.65%, 8.57%, and 9.78%, respectively. The word co-occurrence (WCO) model is also the best of the four models, and achieves up to 67.49% of monolingual performance. It shows a 50.00% greater improvement than the simple select-all (SA) strategy. Thus the proposed WCO method is adopted for query disambiguation in Chinese-English information retrieval, and it has the feasibility and portability in different domains.

2.7 Summary

This chapter presents 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 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).

Recall that the mutual information table is trained from the retrieval text collection (e.g. TREC-6 text collection). However, users may retrieve relevant documents in different domains using various search engines on the Internet. How to obtain the mutual information table for translation disambiguation of query is a problem in cross-language information retrieval on the Internet. The next chapter will touch this problem and use a balanced corpus in target language to obtain a general-used table. Further, it will discuss another effects of ambiguities.

Chapter 3

Query Translation: Target Polysemy Resolution

This chapter deals with the target polysemy problem in cross language information retrieval. As discussed in Chapter 1, several different approaches have been proposed to deal with the translation ambiguity problem for query translation. However, few touch on translation ambiguity and target polysemy together. This chapter will study the multiplication effects of translation ambiguity and target polysemy in cross-language information retrieval, and propose a new translation method to resolve these problems. Section 3.1 shows the effects of translation ambiguity and target polysemy in Chinese-English and English-Chinese information retrievals. Section 3.2 presents several models to revolve translation ambiguity and target polysemy problems. Section 3.3 demonstrates the experimental results, and compares the performances of these models. Finally, Section 3.4 concludes the remarks.

3.1 Effects of Ambiguities

Translation ambiguity and target polysemy are two major problems in CLIR. Translation ambiguity results from the source language, and target polysemy occurs in target language. For query translation, translation ambiguity is a basic problem to be resolved. A word in a source query may have more than one sense. Word sense disambiguation identifies the correct sense of each source word, and lexical selection translates it into the corresponding target word. The above procedure is similar to lexical choice operation in a traditional machine translation (MT) system. However, there is a significant difference between the applications of MT and CLIR. In MT, readers interpret the translated results. If the target word has more than one sense, readers can disambiguate its meaning automatically. Comparatively, the translated result is sent to a monolingual information retrieval system in CLIR. The target polysemy adds extraneous senses and affects the retrieval performance (Chen, Bian, and Lin, 1999).

Table 3.1 Statistics of Chinese and English Thesaurus

	Total Words	Average # of Senses	Average # of Senses for Top 1000 Words
English Thesaurus	29,380	1.687	3.527
Chinese Thesaurus	53,780	1.397	1.504

Take Chinese-English information retrieval (CEIR) and English-Chinese information retrieval (ECIR) as examples. The former uses Chinese queries to retrieve English documents, while the later employs English queries to retrieve Chinese documents. To explore the difficulties in the query translation of different languages, we gather the sense statistics of English and Chinese words. Table 3.1 shows the degree of word sense ambiguity (in terms of number of senses of a word) in English and in Chinese, respectively. A Chinese thesaurus, i.e., 同義詞詞林(tong2yi4ci2ci2lin2), (Mei, *et al.*, 1982) and an English thesaurus, i.e., Roget's thesaurus, are used to count the statistics of the senses of words. On the average, an English word has 1.687 senses, and a Chinese word has 1.397 senses. If the top 1000 high frequent words are considered, the English words have 3.527 senses, and the bi-character Chinese words only have 1.504 senses. In summary, Chinese word is comparatively unambiguous, so that translation ambiguity is not serious but target polysemy is serious in CEIR. In contrast, an English word is usually ambiguous. The translation disambiguation is important in ECIR.

Consider an example in Chinese-English information retrieval (CEIR), which uses Chinese queries to retrieve English documents. The Chinese word "銀行" (Yin2Hang2) is unambiguous, but its English translation "bank" has many senses

shown below[1]:

(1) Sense Definition: land along the side of a river, lake, *etc.*

Chinese Translation: 岸；堤

(2) Sense Definition: earth which is heaped up in a field or garden, often making a border or division

Chinese Translation: 田埂

(3) Sense Definition: a mass of snow, clouds, mud, *etc.*

Chinese Translation: 一堆；一團

(4) Sense Definition: a slope made at bends in a road or race-track, so that they are safer for cars to go round

Chinese Translation: 邊坡

(5) Sense Definition: SANDBANK

Chinese Translation: 沙洲

(6) Sense Definition: a row, esp. of OARs in an ancient boat or KEYs on a TYPEWRITER

Chinese Translation: 一排

(7) Sense Definition: (a person who keeps) a supply of money or pieces for payment or use in a game of chance

[1] The sense definitions are selected from Longman English-Chinese Dictionary of Contemporary English (ECLDOCE).

Chinese Translation: 莊家

(8) Sense Definition: a place in which money is kept and paid out on demand, and where related activities go on

Chinese Translation: 銀行

(9) Sense Definition: (usu. in comb.) a place where something is held ready for use, esp. ORGANIC products of human origin for medical use

Chinese Translation: 儲存所; 庫

(10) Sense Definition: (of a car or aircraft) to move with one side higher than the other, esp. when making a turn

Chinese Translation: 傾斜轉彎

(11) Sense Definition: to put or keep (money) in a bank

Chinese Translation: 存（款）於銀行

(12) Sense Definition: to keep one's money (esp. in the stated bank)

Chinese Translation: 存款

When the Chinese word "銀行" is issued, it is translated into the English counterpart "bank" without difficulty, and then "bank" is sent to an IR system. The IR system will retrieve documents that contain this word. Because "bank" is not disambiguated, many irrelevant documents will be reported.

On the contrary, when "bank" is submitted to an English-Chinese information

retrieval (ECIR) system (employing English queries to retrieve Chinese documents), we must disambiguate its meaning at first. If we can find that its correct translation is "銀行" (yin2hang2), the subsequent operation is very simple. That is, "銀行" (yin2hang2) is sent into an IR system, and then documents containing "銀行" (yin2hang2) will be presented. In this example, translation disambiguation should be done rather than target polysemy resolution.

The above examples do not mean translation disambiguation is not required in CEIR. Some Chinese words may have more than one sense. For example, "運動" (yun4dong4) has the following meanings (Lai and Lin, 1987):

(1) sport: an outdoor or indoor game

(2) exercise: use of any part of the body or mind to strengthen and improve it

(3) movement: a group of people who make united efforts for a particular purpose

(4) motion: state of moving

(5) campaign: to lead, take part in, or go on a campaign

(6) lobby: to meet to persuade him/ her to support one's actions and needs

Each corresponding English word may have more than one sense. For example, "exercise" may mean *a question or set of questions to be answered by a pupil for practice* ("練習" (lian4xi2); "習題" (xi2ti2)); *the use of a power or right* ("力量" (li4liang4)), "權力的運用" (quan2li4de0yun4yong4)); and so on. The

multiplication effects of translation ambiguity and target polysemy make query translation harder.

3.2 Target Polysemy Resolution Models

We take Chinese-English information retrieval as an example to explain our methods. Consider the Chinese query "銀行" (yin2hang2) to an English collection again. The ambiguity grows from none (source side) to 12 senses (target side) during query translation. How to incorporate the knowledge from source side to target side is an important issue. To avoid the problem of target polysemy in query translation, we have to restrict the use of a target word by augmenting some other words that usually co-occur with it. That is, we have to make a context for the target word. In our method, the contextual information is derived from the source word.

We collect the frequently accompanying nouns and verbs for each word in a Chinese corpus. Those words that co-occur with a given word within a window are selected. The word association strength of a word and its accompanying words is measured by mutual information. For each word C in a Chinese query, we augment it with a sequence of Chinese words trained in the above way. Let these words be CW_1, CW_2, ..., and CW_m. Assume the corresponding English

translations of C, CW_1, CW_2, ..., and CW_m are E, EW_1, EW_2, ..., and EW_m, respectively. EW_1, EW_2, ..., and EW_m form an *augmented translation restriction* of E for C. In other words, the list (E, EW_1, EW_2, ..., EW_m) is called an *augmented translation result* for C. EW_1, EW_2, ..., and EW_m are a *pseudo English context* produced from Chinese side. Consider an example of the Chinese word "銀行" (yin2hang2). Some strongly co-related Chinese words in ROCLING balanced corpus (Huang, *et al.*, 1995) are: "貼現" (tie1xian4), "領出" (ling3chu1), "里昂" (li3ang2), "押匯" (ya1hui4), "匯兌" (hui4dui4), and so on. Thus the augmented translation restriction of the English word "bank" is (rebate, show out, Lyons, negotiate, transfer, .).

The query translation is not so simple. A word C in a query Q may be ambiguous. Besides, the accompanying words CW_i ($1 \leq i \leq m$) trained from Chinese corpus may be translated into more than one English word. An augmented translation restriction may add erroneous patterns when a word in a restriction has more than one sense. Thus we devise several models to discuss the effects of augmented restrictions. Figure 3.1 shows the different models and the model refinement procedure. A Chinese query may go through translation ambiguity resolution module (left-to-right), target polysemy resolution module (top-down), or both (i.e., these two modules are integrated at the right corner). In the following, we will show how each module is operated independently, and

how the two modules are combined.

For a Chinese query which is composed of n words C_1, C_2, ..., C_n, find the corresponding English translation equivalents in a Chinese-English bilingual dictionary. To discuss the propagation errors from translation ambiguity resolution part in the experiments, we consider the following two alternatives:

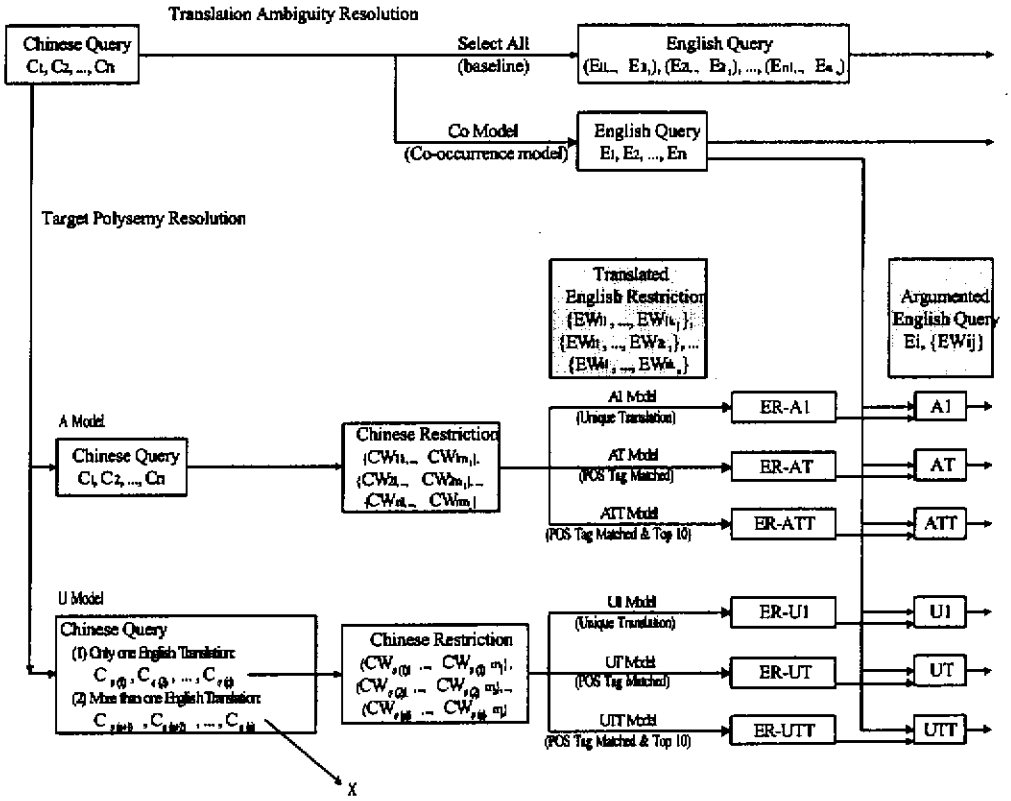


Figure 3.1 Models for Translation Ambiguity and Target Polysemy Resolution

(a) select all (do-nothing)

The strategy does nothing on the translation disambiguation. All the English translation equivalents for the n Chinese words are selected, and are submitted to a monolingual information retrieval system.

(b) co-occurrence model (abbreviated as Co-Model)

We adopt the strategy discussed previously for translation disambiguation. This method considers the content around the English translation equivalents to decide the best target equivalent. The translation of a query term can be disambiguated using the co-occurrence of the translation equivalents of this term and other terms. We adopt mutual information to measure the strength.

For target polysemy resolution part in Figure 3.1, we also consider two alternatives. The first alternative (called A model) is that we augment restrictions to all the words no matter whether they are ambiguous or not. The second alternative (called U model) is that we neglect those Cs that have more than one English translation. Assume $C_{\sigma(1)}$, $C_{\sigma(2)}$, ..., $C_{\sigma(p)}$ ($p \leq n$) have only one English translation. The restrictions are augmented to $C_{\sigma(1)}$, $C_{\sigma(2)}$, ..., $C_{\sigma(p)}$ only. We apply the corpus-based method mentioned above to find the restriction for each English word selected by the translation ambiguity resolution model (i.e., Co-Model). Recall that the restrictions are derived from Chinese corpus. The

accompanying words trained from Chinese corpus may be translated into more than one English word. Three alternatives are considered. In U1 (or A1) model, the terms without ambiguity, i.e., Chinese and English words are one-to-one correspondent in a Chinese-English bilingual dictionary, are added. In UT (or AT) model, the terms with the same parts of speech are added. That is, part of speech is used to select English word. In UTT (or ATT) model, we use mutual information to select top 10 accompanying terms of a Chinese query word, and part of speech is used to obtain the augmented translation restriction.

In the above treatment, a word C_i in a query Q is translated into (E_i, EW_{i1}, EW_{i2}, ..., EW_{imi}). E_i is selected by Co-Model, and EW_{i1}, EW_{i2}, ..., EW_{imi} are augmented by different target polysemy resolution models. Intuitively, E_i, EW_{i1}, EW_{i2}, ..., EW_{imi} should have different weights. E_i is assigned a higher weight, and the words EW_{i1}, EW_{i2}, ..., $_{imi}$ in the restriction are assigned lower weights. They are determined by the following formula, where n is number of words in Q and m_k is the number of words in a restriction for E_k.

$$\text{weight}(E_i) = \frac{1}{n+1}$$

$$\text{weight}(EW_{ij}) = \frac{1}{(n+1)*\sum_{k=1}^{n} m_k}$$

Thus six new models, i.e., A1W, ATW, ATTW, U1W, UTW and UTTW, are

derived. Finally, we apply the co-occurrence model (i.e., Co-model) again to disambiguate the pseudo contexts and devise six new models (A1WCO, ATWCO, ATTWCO, U1WCO, UTWCO, and UTTWCO). In these six models, only one restriction word will be selected from the words EW_{i1}, EW_{i2}, ..., EW_{imi} via disambiguation with other restrictions.

3.3 Experimental Results

To evaluate the above models, we employ TREC-6 text collection, TREC topics 301-350 (Harman, 1997), and Smart information retrieval system (Salton and Buckley, 1988). The text collection contains 556,077 documents, and is about 2.2G bytes. Because the goal is to evaluate the performance of Chinese-English information retrieval on different models, we translate the 50 English queries into Chinese by human. The topic 332 is considered as an example in the following. The original English version and the human-translated Chinese version are shown. A TREC topic is composed of several fields. The tag <num> denotes the topic number; <title> and <C-title> are English and Chinese titles of a topic; <des> and <C-des> form description part of a topic; <narr> and <C-narr>, which are narrative fields, provides a complete description of document relevance for the assessors. Only the fields of title and

description are used to generate queries in our experiments.

<top>

<num> Number: 332

<title> Income Tax Evasion

<desc> Description:

This query is looking for investigations that have targeted evaders of U.S. income tax.

<narr> Narrative:

A relevant document would mention investigations either in the U.S. or abroad of people suspected of evading U.S. income tax laws. Of particular interest are investigations involving revenue from illegal activities, as a strategy to bring known or suspected criminals to justice.

</top>

<top>

<num> Number: 332

<C-title>

逃漏所得稅。

<C-desc> Description:

這個查詢要找出針對美國所得稅逃漏稅者的調查。

<C-narr> Narrative:

相關文件提到對美國國內或國外有逃漏美國所得稅企圖的人的調查。對於來自非法活動的收入的稅收，這是一種把罪犯訴諸正法的另一種方法。

</top>

Totally, there are 1,017 words (557 distinct words) in the title and description fields of the 50 translated TREC topics. Among these, 401 words have unique translations and 616 words have multiple translation equivalents in our Chinese-English bilingual dictionary. Table 3.2 shows the degree of word sense ambiguity (in terms of number of senses of a word) in English and in Chinese, respectively. On the average, an English query term has 2.976 senses, and a Chinese query term has 1.828 senses only. In our experiments, LOB corpus is employed to train the co-occurrence statistics for translation ambiguity resolution, and ROCLING balanced corpus (Huang, *et al.*, 1995) is employed to train the restrictions for target polysemy resolution. The mutual information tables are trained using a window size 3 for adjacent words.

Table 3.2 Statistics of TREC Topics 301-350

	# of Distinct Words	Average # of Senses
Original English Topics	500 (370 words found in our dictionary)	2.976
Human-translated Chinese Topics	557 (389 words found in our dictionary)	1.828

Table 3.3 Query Translation of Title Field of TREC Topic 332

(a) Resolving Translation Ambiguity Only

original English query	income tax evasion
Chinese translation by human	逃漏 (tao2luo4) 所得 (suo3de2) 稅 (sui4)
by select all model	(evasion), (earning, finance, income, taking), (droit, duty, geld, tax)
by co-occurrence model	evasion, income, tax

(b) Resolving both Translation Ambiguity and Target Polysemy

by A1 model	(**evasion**, poundage, scot, stay), (**income**, quota), (**tax**, evasion, surtax, surplus, sales tax)
by U1 model	(**evasion**, poundage, scot, stay), (**income**), (**tax**)
by AT model	(**evasion**; **poundage**; **scot**; stay; droit, duty, geld, tax; custom, douane, tariff; avoid, elude, wangle, welch, welsh; contravene, infract, infringe), (**income**; impose; assess, put, tax; Swiss, Switzer; minus, subtract; quota; commonwealth, folk, land, nation, nationality, son, subject), (**tax**; surtax; surplus; sales tax; abase, alight, debase, descend; altitude, loftiness, tallness; comprise, comprize, embrace, encompass; compete, emulate, vie)
by UT model	(**evasion**; poundage, scot, stay, droit, duty, geld, tax, custom, douane, tariff, avoid, elude, wangle, welch, welsh, contravene, infract, infringe), (**income**), (**tax**)
by ATT model	(**evasion**, poundage, scot, stay, droit, duty, geld, tax, custom, douane, tariff), (**income**), (**tax**)
by UTT model	(**evasion**, poundage, scot, stay, droit, duty, geld, tax, custom, douane, tariff), (**income**), (**tax**)
by ATWCO model	(**evasion**, tax), (**income**, land), (**tax**, surtax)
by UTWCO model	(**evasion**, poundage), (**income**), (**tax**)
by ATTWCO model	(**evasion**, tax), (**income**), (**tax**)
by UTTWCO model	(**evasion**, poundage), (**income**), (**tax**)

Table 3.3 shows the query translation of TREC topic 332. For the sake of space, only title field is shown. In Table 3.3(a), the first two rows list the original English query and the Chinese query. Rows 3 and 4 demonstrate the English translation by select-all model and co-occurrence model by resolving translation ambiguity only. Table 3.3(b) shows the augmented translation results using different models. Here, both translation ambiguity and target polysemy are resolved. The following lists the selected restrictions in A1 model.

逃漏(evasion): 稅捐_N (N: poundage), 租稅_N (N: scot), 過止_V (V: stay)

所得(income): 限額_N (N: quota)

稅(tax): 逃漏_V(N: evasion), 附加稅_N (N: surtax), 盈餘_N (N: surplus), 營業稅_N (N: sales tax)

Augmented translation restrictions (poundage, scot, stay), (quota), and (evasion, surtax, surplus, sales tax) are added to "evasion", "income", and "tax", respectively. From Longman dictionary, we know there are 3 senses, 1 sense, and 2 senses for "evasion", "income", and "tax", respectively. Augmented restrictions are used to deal with target polysemy problem. Compared with A1 model, only "evasion" is augmented with a translation restriction in U1 model. This is because the Chinese term "逃漏" (tao2luo4) has only one English translation and the other two terms "所得" (suo3de2) and "稅" (sui4) have more than one translation. Similarly, the augmented translation restrictions are omitted in the other U-models (i.e., UT, UTT, UTWCO, and UTTWCO) to avoid possible error propagation from ambiguous Chinese terms (e.g., "所得" (suo3de2) and "稅" (sui4)). Now we consider AT model. The Chinese restrictions, which have the matching parts of speech in our Chinese-English dictionary, are listed below:

逃漏 (evasion):

稅捐_N (N: poundage),

租稅_N (N: scot),

過止_V (V: stay)

稅_N (N: droit, duty, geld, tax),

關稅_N (N: custom, douane, tariff),

逃避_V (V: avoid, elude, wangle, welch, welsh; N: avoidance, elusion, evasion, evasiveness, miss, runaround, shirk, skulk),

違反_V (V: contravene, infract, infringe; N: contravention, infraction, infringement, sin, violation)

所得 (**income**):

課_V (V: impose; N: division),

課稅_V (V: assess, put, tax; N: imposition, taxation),

瑞士人_N (N: Swiss, Switzer),

減去_V (V: minus, subtract),

限額_N (N: quota),

國民_N (N: commonwealth, folk, land, nation, nationality, son, subject)

稅 (**tax**):

附加稅_N (N: surtax),

盈餘_N (N: surplus),

營業稅_N (N: sales tax)

降_V (V: abase, alight, debase, descend),

高_N (N: altitude, loftiness, tallness; ADJ: high; ADV: loftily)

含_V (V: comprise, comprize, embrace, encompass),

爭_V (V: compete, emulate, vie; N: conflict, contention, duel, strife)

Those English words whose parts of speech are the same as the corresponding Chinese restrictions are selected as augmented translation restriction. For example, the translation of "逃避"_V (tao2bi4) has two possible parts of speech, i.e., V (avoid, elude, wangle, welch, welsh) and N (avoidance, elusion, evasion, evasiveness, miss, runaround, shirk, skulk). In this way, only "avoid", "elude", "wangle", "welch", and "welsh" are chosen. The other terms are added in the similar way. Recall that an extra condition, i.e., to use mutual information to select top 10 accompanying terms of a Chinese query word, is employed in ATT model. The 5th row shows that the augmented translation restrictions for the words "所得" (suo3de2) and "稅" (sui4) are removed because their top 10 Chinese accompanying terms do not have English translations of the same parts of speech. Finally, we consider ATWCO model. The words "tax", "land", and "surtax" are selected from the three lists in 3rd row of Table 3.3(b) respectively, by using word co-occurrences.

Figure 3.2 shows the number of relevant documents on the top 1000 retrieved documents for Topics 332 and 337. The performances are stable in all of the +weight (W) models and the enhanced CO restriction (WCO) models, even there

are different number of words in translation restrictions. Especially, the enhanced CO restriction models add at most one translated restriction word for each query term. They can achieve the similar performance to those models that add more translated restriction words. Surprisingly, the augmented translation results may perform better than the monolingual retrieval. Topic 337 in Figure 3.2 is an example.

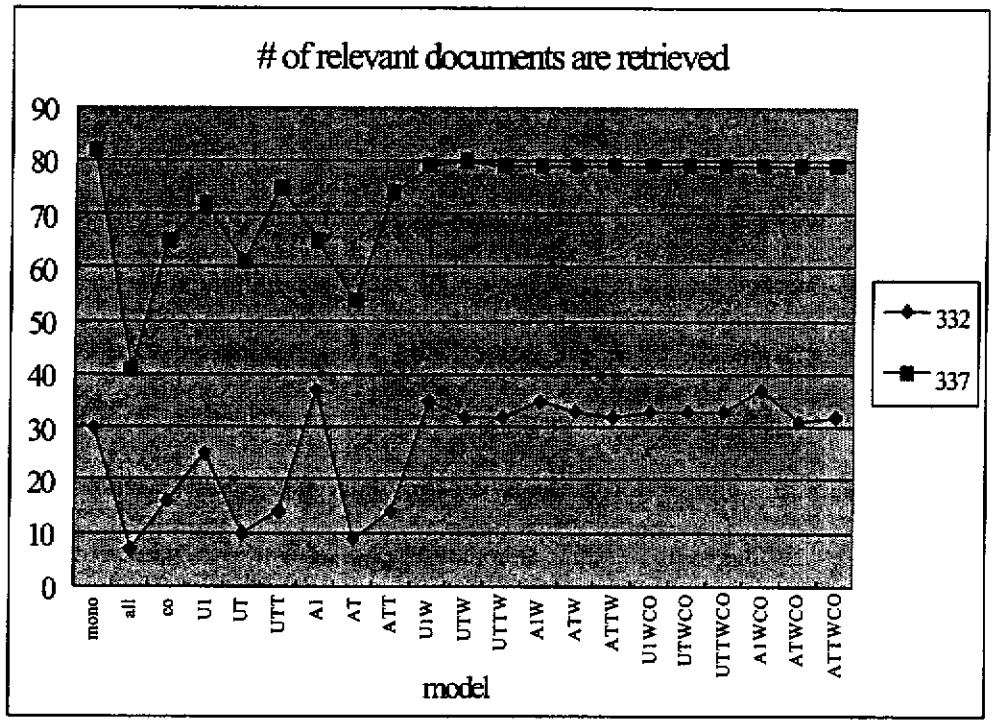


Figure 3.2 The Retrieved Performances of Topics 332 and 337

Table 3.4 Performance of Different Models (11-point Average Precision)

Monolingual IR	Resolving Translation Ambiguity		Resolving Translation Ambiguity and Target Polysemy					
	Select All	English Co Model	Unambiguous Words			All Words		
			U1	UT	UTT	A1	AT	ATT
			0.0797 (54.63%)	0.0574 (39.34%)	0.0709 (48.59%)	0.0674 (46.20%)	0.0419 (28.72%)	0.0660 (45.24%)
			+ Weight			+ Weight		
			U1W	UTW	UTTW	A1W	ATW	ATTW
0.1459	0.0652 (44.69%)	0.0831 (56.96%)	0.0916 (62.78%)	0.0915 (62.71%)	0.0914 (62.65%)	0.0914 (62.65%)	0.0913 (62.58%)	0.0914 (62.65%)
			+ Weight, English Co Model for Restriction Translation			+ Weight, English Co Model for Restriction Translation		
			U1WCO	UTWCO	UTTWCO	A1WCO	ATWCO	ATTWCO
			0.0918 (62.92%)	0.0917 (62.85%)	0.0915 (62.71%)	0.0917 (62.85%)	0.0917 (62.85%)	0.0915 (62.71%)

Table 3.4 shows the overall performance of 18 different models for 50 topics. Eleven-point average precision on the top 1000 retrieved documents is adopted to measure the performance of all the experiments. The monolingual information retrieval, i.e., the original English queries to English text collection, is regarded as a baseline model. The performance is 0.1459 under the specified environment. The select-all model, i.e., all the translation equivalents are passed without disambiguation, has 0.0652 average precision. About 44.69% of the performance of the monolingual information retrieval is achieved. When co-occurrence model is employed to resolve translation ambiguity, 0.0831 average precision (56.96% of monolingual information retrieval) is reported.

Compared to do-nothing model, the performance is 27.45% increase.

Now we consider the treatment of translation ambiguity and target polysemy together. Augmented restrictions are formed in A1, AT, ATT, U1, UT and UTT models, however, their performances are worse than Co-model (translation disambiguation only). The major reason is the restrictions may introduce errors. That can be found from the fact that models U1, UT, and UTT are better than A1, AT, and ATT. Because the translation of restriction from source language (Chinese) to target language (English) has the translation ambiguity problem, the models (U1 and A1) introduces the unambiguous restriction terms and performs better than other models. Controlled augmentation shows higher performance than uncontrolled augmentation.

When different weights are assigned to the original English translation and the augmented restrictions, all the models are improved significantly. The performances of A1W, ATW, ATTW, U1W, UTW, and UTTW are about 10.11% addition to the model for translation disambiguation only. Of these models, the performance change from model AT to model ATW is drastic, i.e., from 0.0419 (28.72%) to 0.0913 (62.58%). It tells us the original English translation plays a major role, but the augmented restriction still has a significant effect on the performance.

We know that restriction for each English translation presents a pseudo English

context. Thus we apply the co-occurrence model again on the pseudo English contexts. The performances are increased a little. These models add at most one translated restriction word for each query term, but their performances are better than those models that adding more translated restriction words. It tells us that a good translated restriction word for each query term is enough for resolving target polysemy problem. U1WCO, which is the best in these experiments, gains 62.92% of monolingual information retrieval, and 40.80% increase to the do-nothing model (select-all).

3.4 Summary

This chapter deals with translation ambiguity and target polysemy at the same time. We utilize two monolingual balanced corpora to learn useful statistical data, i.e., word co-occurrence for translation ambiguity resolution, and augmented translation restrictions for target polysemy resolution. Aligned bilingual corpus or special domain corpus is not required in this design. Experiments show that the model achieves 62.92% of monolingual information retrieval, and is 40.80% addition to the select-all model. Combining the target polysemy resolution in cross-language information retrieval, the retrieval performance is about 10.11% increase to the model resolving translation ambiguity only.

We also analyze the two factors: word sense ambiguity in source language (translation ambiguity), and word sense ambiguity in target language (target polysemy). On the average, an English word has 1.687 senses, and a Chinese word has 1.397 senses. If the top 1000 high frequent words are considered, the English words have 3.527 senses, and the bi-character Chinese words only have 1.504 senses. The statistics of word sense ambiguities have shown that target polysemy resolution is critical in Chinese-English information retrieval.

This treatment is very suitable to translate very short query on Web. According to the papers (Pinkerton, 1994; Fitzpatrick and Dent, 1997), the queries on Web are 1.5-2 words on the average. Because the major components of queries are nouns, at least one word of a short query of length 1.5-2 words is noun. Besides, most Chinese nouns are unambiguous, so that translation ambiguity is not relatively serious, but target polysemy is critical in Chinese-English Web retrieval. The translation restrictions, which introduce pseudo contexts, are helpful for target polysemy resolution.

Bibliography

Baker, K., et al. (1994) "Coping with Ambiguity in a Large-scale Machine Translation System." In Proceedings of COLING-94, 1994, pp. 90-94.

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.B. (1997) "Phrasal Translation and Query Expansion Techniques for Cross-Language Information retrieval." In *Proceedings of ACM SIGIR'97*, pp.84-91, 1997.

Ballesteros, L and Croft, W.B. (1998) "Resolving Ambiguity for Cross-Language Retrieval." In *Proceedings of ACM SIGIR 98*, pp. 64-71, 1998.

Bian, G.W. and Chen, H.H. (1998a) "A New Hybrid Approach for Chinese-English Query Translation." In *Proceedings of First Asia Digital Library Workshop*, Hong Kong, 1998, pp. 156-167.

Bian, G.W. and Chen, H.H. (1998c). "Integrating Query Translation and Document Translation in a Cross-Language Information Retrieval System." In *Proceedings of AMTA Conference on Machine Translation (AMTA-98)*, Langhorne, PA, USA, October 28-31, 1998, pp.250-265.

Chen, H.H.; Bian, G.W.; and Lin, W.C. (1999) "Resolving Translation Ambiguity and Target Polysemy in Cross-Language Information Retrieval", In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL '99)*, Maryland, USA, June 20-26, 1999, pp. 215-222.

Church, K. and Hanks, P. (1990) "Word Association Norms, Mutual Information and Lexicography." *Computational Linguistics*, 16(1), 1990, pp.22-29.

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 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. 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.

Davis, M.W. and Ogden, W.C. (1997) "Implementing Cross-Language Text Retrieval Systems for Large-scale Text Collections and the World Wide Web." *Working Notes of the AAAI-97 Spring Symposium on Cross-Language Text and Speech Retrieval*, 1997, pp. 9-17.

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.

Fox, E., ed.: Virginia Disk One, Blacksburg: Virginia Polytechnic Institute and State University, 1990.

Gachot, D.A.; Lange, E. and Yang, J. (1996) "The SYSTRAN NLP Browser: An Application of Machine Translation Technology in Multilingual Information

Retrieval." In *Proceedings of Workshop on Cross-Linguistic Information Retrieval*, 1996, pp. 44-54.

Grimes, Barbara F. (Editor) (1996) *Ethnologue: Languages of the World*, 13th Edition, Summer Institute of Linguistics, Inc., Dallas, Texas, 1996. URL: http://www.sil.org/ethnologue/

Harman, D.K. (Editor) (1997) *TREC-6 Proceedings*, Gaithersburg, Maryland, 1997.

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.

Johansson, S. (1986) *The Tagged LOB Corpus: Users' Manual*. Bergen: NORWEGIAN Computing Centre for the Humanities, 1986.

Kraaij, W. and Hiemstra, D. (1997) "Cross Language Retrieval with the Twenty-One System." In *Proceedings of the Sixth Text Retrieval Evaluation Conference (TREC-6)*, Gaithersburg, MD, National Institute of Standards and Technology, 1997.

Kwok, K.L. (1997) "Evaluation of an English-Chinese Cross-Lingual Retrieval Experiment." In *Working Notes of the AAAI-97 Spring Symposium on Cross-Language Text and Speech Retrieval*, 1997, pp. 110-114.

Lai, M. and Lin, T.Y. (1987) *The New Lin Yutang Chinese-English Dictionary*. Panorama Press Ltd, Hong Kong, 1987.

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.

Longman (1978) *Longman Dictionary of Contemporary English*. Longman Group Limited, UK, 1978.

Mei, *et al.* (1982) *tong2yi4ci2ci2lin2*. Shanghai Dictionary Press, 1982

Nagao, M. (1984) "A Framework of Mechanical Translation between Japanese and English by Analogy Principle." *Artificial and Human Intelligence*, 1984, pp. 173-180.

Nirenburg, S., *et al.* (1993) "Multi-purpose Development and Operations Environments for Natural Language Applications." In *Proceedings of Applied Language Processing*, Trento, Italy, 1993.

Oard, D.W. (1997) "Alternative Approaches for Cross-Language Text Retrieval." *Working Notes of AAAI-97 Spring Symposiums on Cross-Language Text and Speech Retrieval*, pp. 131-139, 1997.

Oard, D.W. (1998) "A Comparative Study of Query and Document Translation for Cross-Language Information Retrieval." In *Proceedings of AMTA Conference on Machine Translation (AMTA-98)*, Langhorne, PA, USA, October 28-31, 1998, pp.472-483.

Oard, D.W. and Dorr, B.J. (1996) *A Survey of Multilingual Text Retrieval.* Technical Report UMIACS-TR-96-19, University of Maryland, Institute for Advanced Computer Studies, USA, 1996.
URL: http://www.ee.umd.edu/medlab/filter/papers/mlir.ps.

Salton, G. and Buckley, C. (1988) "Term Weighting Approaches in Automatic Information Processing and Management 5(24): 513-523, 1988.

Selberg, E. and Etzioni, O. (1995) "Multi-Service Search and Comparison Using the MetaCrawler." In *Proceedings of the Fourth International World Wide Web Conference*, Boston, MA., Dec. 1995.
URL: http://www.w3.org/pub/Conferences/WWW4/Papers/169/

Sheridan, P. and Ballerini, J.P. (1996) "Experiments in Multilingual Information Retrieval Using the SPIDER system." In ,
pp.58-65, 1996.

Yuwono, B.; Lam, S.L.Y.; Ying, J.H.; and Lee, D.L. (1995) "A World Wide Web Resource Discovery System." In *Proceedings of the Fourth International World Wide Web Conference*, Boston, MA., Dec. 1995.
URL: http://www.w3.org/pub/Conferences/WWW4/Papers/66/

國際數位圖書館合作研究計畫(IDLP)出國心得報告

陳信希

國立台灣大學資訊工程學系

1. 考察活動簡介

國際數位圖書館合作研究計畫(International Digital Library Program Project，簡稱 IDLP)，整合台灣(台灣大學、清華大學、中央研究院)、大陸(北京大學、北京清華大學、上海交通大學)、和美國(Simmons 等大學)三地的研究人員，共同合作開發數位圖書館關鍵技術。本次訪問擬考察大陸相關單位的發展情況，並建立未來合作的模式。

2. 考察活動經過

本次訪問由台灣大學資訊工程學系項潔教授和陳信希教授、圖書資訊學系陳雪華主任、以及故宮博物院資訊中心蔡順慈主任和賴鼎聲博士等組成。報告人與圖資系陳主任於 5 月 27 日，搭乘澳門航空 NX611 班機，在澳門轉 NX002 班機赴北京，住清華大學附近北京西郊賓館。

5 月 28 日早上在故宮蔡主任的安排下，赴北京中國歷史博物館訪問，由李季副館長和谷長江副館長接待，並由肖飛先生(IT 主管)作簡

報。主要是探討中國歷史博物館文物數位化的狀況，由於人力編制的關係，該館目前只有兩位人員專門負責網站的維護，整體數位博物館建製較無長遠的規畫。下午安排到北京故宮訪問，由外事辦公室楊森先生接待。台大與故宮在國科會數位博物館計畫下，合作設計故宮數位博物館，此次訪問擬觀摩北京故宮在數位化方面的進展。由於當天並沒有安排好，因此只作室外導覽，數位博物館合作部份，另外安排在 5 月 31 日早上再繼續討論。

5 月 29 日早上參與 NIT2001 國際會議的開幕式(由 IDLP 計畫美方主持人陳劉清智教授主持)。下午訪問中國科學技術館，由王渝生館長接待，並由前館長李象益教授親至導覽。這個館去年才開幕，整體很新穎，但由於經費的關係，科技館裡很多設備因為怕壞掉並沒有完全開放，不過兩位前後館長的觀念算蠻開放。

5 月 30 日在 NIT2001 報告國科會故宮文物之美數位博物館計畫，討論有關數位博物館跨語檢索系統建製相關問題，並與 IDLP 大陸合作單位座談，討論未來之發展，與合作項目。5 月 31 日參觀清華大學，以及清華大學附設的"育成中心"(學研樓)。

6 月 1 日在 NIT2001 主辦單位安排下，遊覽長城及頤和園兩處。6 月 2 日自由活動。6 月 3 日訪問北京大學計算語言學研究所，和俞士(水文)副所長談中文計算語言學的發展，並討論由北大引進現代漢語

語法信息辭典的可行性。6 月 3 日下午搭澳航 NX001 班機，至澳門轉澳航 NX608 回台北。

3. 結語

北京大學、北京清華大學、和上海交通大學，是本項數位圖書館計畫大陸地區的合作伙伴。在中文數位圖書館的研究，這三所學校扮演非常重要的角色。北京大學、北京清華大學、和上海交通大學是大陸最著名的三所學府，在人力和設備上都擁有最好的競爭力。透過研究訪問，交換不同的心得，所學到的分工整合，更是難得的經驗，對國內數位圖書館的推展與整合有助益。

4. 帶回的資料

(1) 中國歷史博物館網站簡介

(2) 中國科學技術館簡介，科學隧道，中國古代科學展覽等

(3) 高教出版信息

(4) 現代漢語語法信息辭典詳解

(5) Global Digital Library Development in the Millennium 專書一冊

(6) NIT2001 Program