Resolving Translation Ambiguity and Target Polysemy in Cross-Language Information Retrieval

Hsin-Hsi Chen, Guo-Wei Bian and Wen-Cheng Lin
Department of Computer Science and Information Engineering
National Taiwan University
Taipei, TAIWAN, R.O.C.
E-mail: hh_chen@csie.ntu.edu.tw, {gwbian, denislin}@nlg2.csie.ntu.edu.tw

Abstract

This paper deals with translation ambiguity and target polysemy problems together. Two monolingual balanced corpora are employed to learn word co-occurrence for translation ambiguity resolution, and augmented translation restrictions for target polysemy resolution. 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, the retrieval performance is about 10.11% increase to the model resolving translation ambiguity only.

1. Introduction

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. Due to the differences between source and target languages, query translation is usually employed to unify the language in queries and documents. In 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.

Some different approaches have been proposed for query translation. Dictionary-based approach exploits machine-readable dictionaries and selection strategies like select all (Hull and Grefenstette, 1996; Davis, 1997), randomly select N (Ballesteros and Croft, 1996; Kwok 1997) and select best N (Hayashi, Kikui and Susaki, 1997; Davis 1997). Corpus-based approaches exploit sentence-aligned corpora (Davis and Dunning, 1996) and document-aligned corpora (Sheridan and Ballerini, 1996). These two approaches are complementary. Dictionary provides translation candidates, and corpus provides context to fit user intention. Coverage of dictionaries, alignment performance and domain shift of corpus are major problems of these two approaches. Hybrid approaches (Ballesteros and Croft, 1998; Bian and Chen, 1998; Davis 1997) integrate both lexical and corpus knowledge.

All the above approaches deal with the translation ambiguity problem in query translation. Few touch on translation ambiguity and target polysemy together. This paper will study the multiplication effects of translation ambiguity and target polysemy in cross-language information retrieval systems, and propose a new translation method to resolve these problems. Section 2 shows the effects of translation ambiguity and target polysemy in Chinese-English and English-Chinese information retrievals. Section 3 presents several models to revolve translation ambiguity and target polysemy problems. Section 4 demonstrates the experimental results, and compares the performances of the proposed models. Section 5 concludes the remarks.

2. 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. Take Chinese-English information retrieval (CEIR) and English-Chinese information retrieval (ECIR) as examples. The former uses Chinese queries to

Table 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

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 1 shows the degree of word sense ambiguity (in terms of number of senses) 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 CEIR. The Chinese word "銀行" (yin2hang2) is unambiguous, but its English translation "bank" has 9 senses (Longman, 1978). When the Chinese word "銀行" (yin2hang2) is issued, it is translated into the English counterpart "bank" by dictionary lookup 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, irrelevant documents will be reported. On the contrary, when "bank" is submitted to an ECIR system, 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, (2) exercise, (3) movement, (4) motion, (5) campaign, and (6) lobby. 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*; *the use of a power or right*; and so on. The multiplication effects of translation ambiguity and target polysemy make query translation harder.

3. Translation Ambiguity and Polysemy Resolution Models

In the recent works, Ballesteros and Croft (1998), and Bian and Chen (1998) employ dictionaries and co-occurrence statistics trained from target language documents to deal with translation ambiguity. We will follow our previous work (Bian and Chen, 1998), which combines the dictionary-based and corpus-based approaches for CEIR. A bilingual dictionary provides the translation equivalents of each query term, and the word co-occurrence information trained from a target language text collection is used to disambiguate the translation. 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. 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.

Before discussion, 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 9 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 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), *etc.* Thus the augmented translation restriction of "bank" is (rebate, show out, Lyons, negotiate, transfer, ...).

Unfortunately, the query translation is not so simple. A word C in a query Q may be ambiguous. Besides, the accompanying words CW_i ($1 \le i \le m$) 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 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:

(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 (Co-Model)

We adopt the strategy discussed previously for translation disambiguation (Bian and Chen, 1998). This method considers the content around the English translation equivalents to decide the best target equivalent.

For target polysemy resolution part in Figure 1, we also consider two alternatives. In the first alternative (called A model), we augment restrictions to all the words no matter whether they are ambiguous or not. In the second alternative (called U model), we neglect those Cs that have more than one English translation. Assume $C_{\sigma(1)}$, $C_{\sigma(2)}$, ..., $C_{\sigma(p)}$ ($p \le n$) have only one English translation. The restrictions are augmented to $C_{\sigma(1)}$, $C_{\sigma(2)}$, ..., $C_{\sigma(p)}$ only. We apply the above corpus-based method to find the restriction for each English word selected by the translation ambiguity resolution 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. Here, the translation ambiguity may occur when translating the restrictions. 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 (POSes) are added. That is, POS 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 POS is used to obtain the augmented translation restriction.

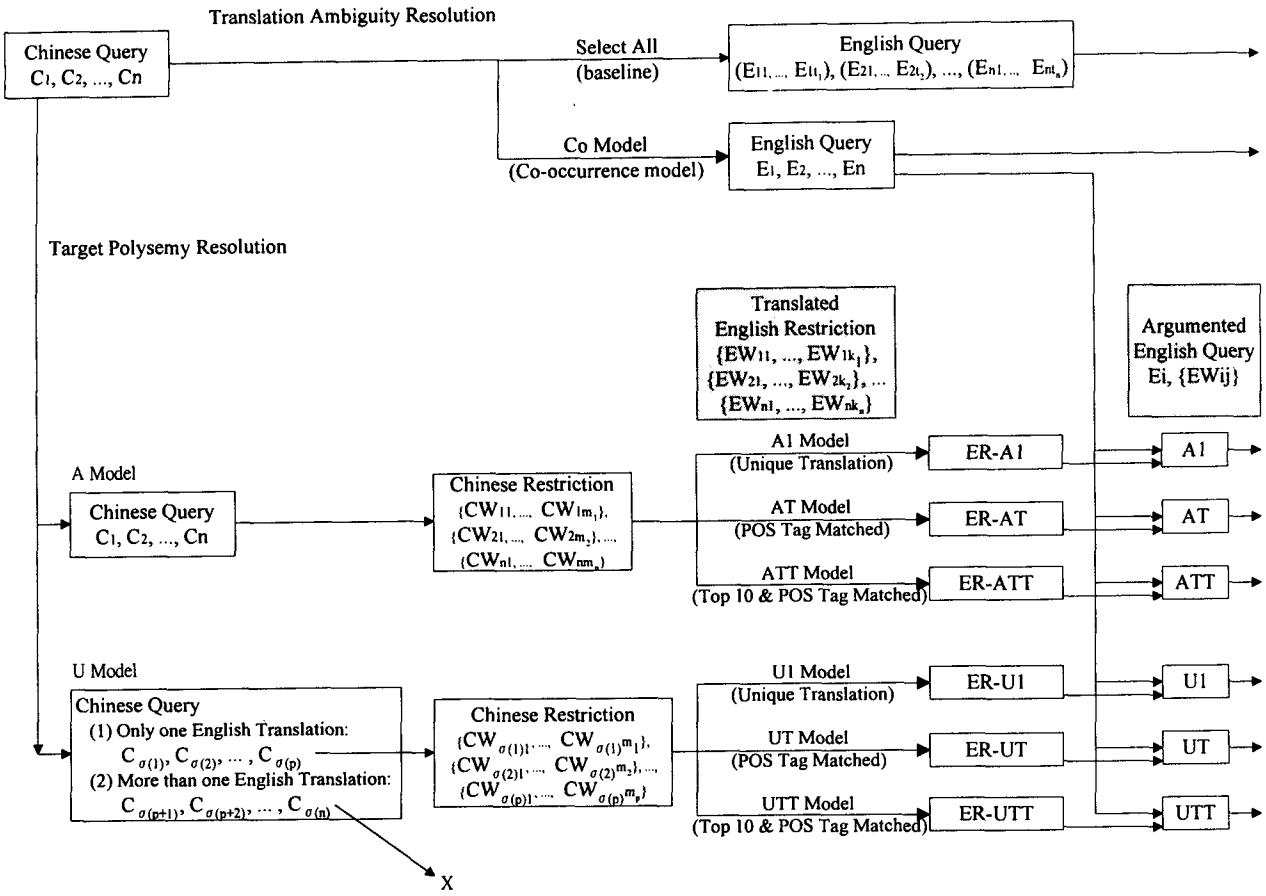

Figure 1. Models for Translation Ambiguity and Target Polysemy Resolution

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}, ..., E_{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 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.

4. 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. Tags <num>, <title>, <des>, and <narr> denote topic number, title, description, and narrative fields. Narrative provides a complete description of document relevance for the

assessors. In our experiments, only the fields of title and description are used to generate queries.

<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 2 shows the degree of word sense ambiguity 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 shows the query translation of TREC topic 332. For the sake of space, only title field is shown. In Table 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(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), 遏止_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 "逃漏" (tao2luo4) has only one translation and "所得" (suo3de2) and "稅" (sui4) have more than one translation. Similarly, the augmented translation restrictions are omitted in the other U-models. Now we consider AT model. The Chinese restrictions, which have the matching POSes, are listed below:

逃漏 (**evasion**):
稅捐_N (N: poundage), 租稅_N (N: scot), 遏止_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)

Table 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. 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) 税 (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**)

Those English words whose POSes 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 POSes, i.e., V and N, so only "avoid", "elude", "wangle", "welch", and "welsh" are chosen. The other terms are added in the similar way. Recall that we use mutual information to select the top 10 accompanying terms of a Chinese query term in ATT model. The 5th row shows that the augmented translation restrictions for "所得" (suo3de2) and "税" (sui4) are removed because their top 10 Chinese accompanying terms do not have English translations of the same POSes. Finally, we consider ATWCO model. The words "tax", "land", and "surtax" are selected from the three lists in 3rd row of Table 3(b) respectively, by using word co-occurrences.

Figure 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 2 is an example.

Table 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

Figure 2. The Retrieved Performances of Topics 332 and 337

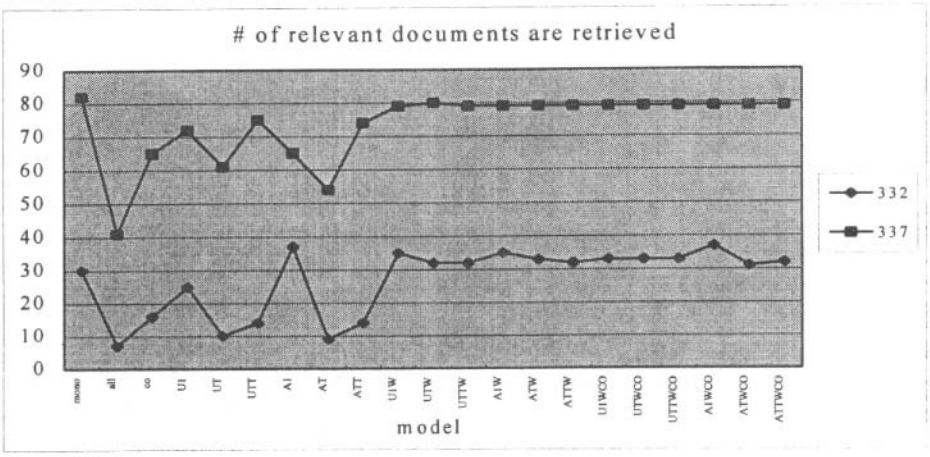

Table 4. Performance of Different Models (11-point Average Precision)

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%)

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) introduce the unambiguous restriction terms and perform 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).

5. Concluding Remarks

This paper 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 translation restrictions for target polysemy resolution. Aligned bilingual corpus or special domain corpus is not required in this design. Experiments show that resolving both translation ambiguity and target polysemy gains about 10.11% performance addition to the method for translation disambiguation in cross-language information retrieval. We also analyze the two factors: word sense ambiguity in source language (translation ambiguity), and word sense ambiguity in target language (target polysemy). 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. The queries on Web are 1.5-2 words on the average (Pinkerton, 1994; Fitzpatrick and Dent, 1997). 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 of the Chinese nouns are unambiguous, so that translation ambiguity is not serious comparatively, but target polysemy is critical in Chinese-English Web retrieval. The translation restrictions, which introduce pseudo contexts, are helpful for target polysemy resolution. The applications of this method to cross-language Internet searching, the applicability of this method to other language pairs, and the effects of human-computer interaction on resolving translation ambiguity and target polysemy will be studied in the future.

References

Ballesteros, L. and Croft, W.B. (1996) "Dictionary-based Methods for Cross-Lingual Information Retrieval." *Proceedings of the 7th International DEXA Conference on Database and Expert Systems Applications*, 791-801.

Ballesteros, L. and Croft, W.B. (1998) "Resolving Ambiguity for Cross-Language Retrieval." *Proceedings of 21st ACM SIGIR*, 64-71.

Bian, G.W. and Chen, H.H. (1998) "Integrating Query Translation and Document Translation in a Cross-Language Information Retrieval System." *Machine Translation and Information Soup*, Lecture Notes in Computer Science, No. 1529, Spring-Verlag, 250-265.

Church, K. *et al.* (1989) "Parsing, Word Associations and Typical Predicate-Argument Relations." *Proceedings of International Workshop on Parsing Technologies*, 389-398.

Davis, M.W. (1997) "New Experiments in Cross-Language Text Retrieval at NMSU's Computing Research Lab." *Proceedings of TREC 5*, 39-1~39-19.

Davis, M.W. and Dunning, T. (1996) "A TREC Evaluation of Query Translation Methods for Multi-lingual Text Retrieval." *Proceedings of TREC-4*, 1996.

Fitzpatrick, L. and Dent, M. (1997) "Automatic Feedback Using Past Queries: Social Searching. " *Proceedings of 20th ACM SIGIR*, 306-313.

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

Hayashi, Y., Kikui, G. and Susaki, S. (1997) "TITAN: A Cross-linguistic Search Engine for the WWW." *Working Notes of AAAI-97 Spring Symposiums on Cross-Language Text and Speech Retrieval*, 58-65.

Huang, C.R., *et al.* (1995) "Introduction to Academia Sinica Balanced Corpus. " *Proceedings of ROCLING VIII*, Taiwan, 81-99.

Hull, D.A. and Grefenstette, G. (1996) "Querying Across Languages: A Dictionary-based Approach to Multilingual Information Retrieval." *Proceedings of the 19th ACM SIGIR*, 49-57.

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

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

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

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

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*, 131-139.

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. http://www.ee.umd.edu/medlab/filter/papers/mlir.ps.

Pinkerton, B. (1994) "Finding What People Want: Experiences with the WebCrawler." *Proceedings of WWW.*

Salton, G. and Buckley, C. (1988) "Term Weighting Approaches in Automatic Text Retrieval." *Information Processing and Management*, Vol. 5, No. 24, 513-523.

Sheridan, P. and Ballerini, J.P. (1996) "Experiments in Multilingual Information Retrieval Using the SPIDER System." *Proceedings of the 19th ACM SIGIR*, 58-65.