Computational Linguistics and Chinese Language Processing
Vol. 4, No. 2, August 1999, pp. 21-38

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

Hsin-Hsi Chen* Guo-Wei Bian* and Wen-Cheng Lin*

Abstract

This paper deals with translation ambiguity and target polysemy problems together. Two monolingual balanced corpora are employed to learn word co-occurrence for the purpose of translation ambiguity resolution and augmented translation restrictions for that of target polysemy resolution. Experiments show that the model achieves 62.92% monolingual information retrieval, which is 40.80% better than that of the select-all model. When target polysemy resolution is added, the retrieval performance represents approximately a 10.11% increase over that of the model which resolves translation ambiguity only.

Keywords: Cross-language information retrieval, Query translation, Translation ambiguity, Target polysemy, Augmented translation restriction

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 differences between the source and target languages, query translation is usually employed to unify the languages 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 the 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. In CLIR, however, the translated

---

+ Part of this work was presented at ACL'99

* 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

result is sent to a monolingual information retrieval system. The target polysemy adds extraneous senses and affects the retrieval performance.

Different approaches have been proposed for query translation. The 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. The dictionary provides translation candidates, and the corpus provides a context to fit the user's intention. Coverage of dictionaries, alignment performance and corpus domain shift are major problems with 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. In the conventional query translation task, the terms of the source query are translated into terms in the target language. However, the target terms may be polysemous. When the polysemous target terms are submitted to a monolingual IR system, the IR system may retrieve irrelevant documents that contain these ambiguous terms. These terms cause the retrieval performance to deteriorate even if they are translated correctly.

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. We use a hybrid approach to deal with translation ambiguity. 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 select the best one. For the target polysemy problem, a pseudo context that is derived from the source word is augmented to restrict the use of a target word. The contextual information is trained from a source language text collection. These two approaches are integrated to construct the target language query. This method is very suitable for translating very short queries. Even for a one-word query, which has no contextual information, the pseudo context trained from a source language text collection can be augmented to expand the query.

The rest of this paper is organized as follows. Section 2 discusses the effects of translation ambiguity and target polysemy in Chinese-English and English-Chinese information retrieval. Section 3 presents several models for resolving translation ambiguity and target polysemy problems. Section 4 gives experimental results and

compares the proposed models in terms of performance. Section 5 provides concluding 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 the target language. 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 query translation of different languages, we gathered 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.

Table 1. *Statistics from a Chinese and an English Thesaurus*

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

A Chinese thesaurus, i.e., 同義詞詞林 (tong2yi4ci2ci2lin2), [Mei, *et al.*, 1982] and an English thesaurus, i.e., Roget's thesaurus, were used to obtain statistics concerning the senses of words. On average, an English word has 1.687 senses, and a Chinese word has 1.397 senses. If the top 1000 high frequency words are considered, the English words have 3.527 senses, and the bi-character Chinese words only have 1.504 senses. In summary, Chinese words are comparatively unambiguous, so translation ambiguity is not difficult, but target polysemy is in CEIR. In contrast, an English word is usually ambiguous. 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 means of 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 other hand, when "bank" is submitted to an ECIR system, we must disambiguate its meaning first. If we find that its correct translation is " 銀行 " (yin2hang2), then the subsequent operation is very simple. That is, " 銀行 " (yin2hang2) is sent into an IR system, and then doc-

uments 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 recent works, Ballesteros and Croft (1998), and Bian and Chen (1998) employed dictionaries and co-occurrence statistics trained from target language documents to deal with translation ambiguity. We 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 context around the translation equivalents to decide on 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 employ mutual information [Church, *et al.*, 1989] to measure the strength. The mutual information MI(*x,y*) is defined as follow:

$$MI(x,y) = \log_2 \frac{p(x,y)}{p(x)p(y)} \qquad (1)$$

where *x* and *y* are words; *p(x)* and *p(y)* are probabilities of words *x* and *y*; *p(x,y)* is their co-occurrence probability.

When a Chinese query is submitted to the query translation system, it is segmented into Chinese words or phrases by looking up a dictionary, using the longest phrase first strategy. For each query term, the English translation equivalents are retrieved from a

bilingual dictionary. After that, MI values are used to select the best translation equivalent. Selection is carried out based on the order of the query terms. For a query term, we compare the MI values of all the translation equivalent pairs (x, y), where x is the translation equivalent of this term, and y is the translation equivalent of another query

Table 2. *Translation Equivalents of Each Term in "*奇異值分解*"*

Term	POS	Translation Equivalents
奇異 (jiyi)	N ADJ	Oddity; singularity Singular
值 (zhi)	N	Value; worth
分解 (fenjie)	N V XV	Decomposition; analysis; dissociation; cracking; disintegration Analyze; anatomize; decompose; decompound; disassemble; dismount; resolve Split up; break up

Table 3. *The Mutual Information Scores for Some Word Pairs*

			奇異 (jiyi)			值 (zhi)		分解 (fenjie)					
			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											

term within a sentence. The word pair (x_i, y_j) with the highest MI value is extracted, and the translation equivalent x_i is regarded as the best translation equivalent of this query term. All the selected translations comprise the final English query.

For example, the phrase " 奇異值分解 " is segmented into three terms: " 奇異 ", " 值 " and " 分解 ". The translation equivalents of each term are shown in Table 2. Table 3 lists the mutual information scores of some word pairs of translation equivalents. Consider the term " 奇異 " first. The translation equivalent pair with the highest MI score is <singular, decomposition>. Therefore, 'singular' is selected as the translation of the word " 奇異 ". Then we select the best translation equivalent of the words " 值 " and " 分解 ". In a similar way, 'value' and 'decomposition' are selected, respectively. Thus, the final translated phrase is 'singular value decomposition.' This disambiguation method performs translation well even when the multi-term phrases are not found in the bilingual dictionary or the phrases are not identified in the source language.

Now, we well shift to the problem of target polysemy. Before this discussion, we will present 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 integrate the knowledge from the source side to the 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 create a context for the target word. In our method, the contextual information is derived from the source word.

We collect 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 based on 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 the Chinese side. Consider the Chinese word " 銀行 " (yin2hang2). Some strongly co-related Chinese words in the 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, query translation is not so simple. A word C in a query Q may be ambiguous. Besides, the accompanying words CW_i (1£ i £ m) trained from a 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 have devised several models to consider the effects of augmented restrictions. Figure 1 shows the different models and the model refinement procedure. A Chinese query may go through the 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 lower corner). In the following, we will show how each module is operated independently, and how the two modules are combined.

Figure 1 *Models for Translation Ambiguity and Target Polysemy Resolution*

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

(a) select all (do-nothing)

This strategy does nothing in terms of translation disambiguation. All the English translation equivalents for the *n* Chinese words are selected and 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 context around the English translation equivalents to decide the best target equivalent.

For the target polysemy resolution part shown 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 $C_{s(1)}$, $C_{s(2)}$, ..., $C_{s(p)}$ (p £ n) have only one English translation. The restrictions are augmented to $C_{s(1)}$, $C_{s(2)}$, ..., $C_{s(p)}$ only. We can 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 a Chinese corpus. The accompanying words trained from the Chinese corpus may be translated into more than one English word. Here, translation ambiguity may occur when translating the restrictions. Three alternatives are considered. In the U1 (or A1) model, the terms without ambiguity are added. That is, we only consider those Chinese terms that have only one English translation in a Chinese-English dictionary, and the corresponding English translations are added as restrictions. In the UT (or AT) model, the terms with the same parts of speech (POSes) are added. That is, POS is used to select English words. In the UTT (or ATT) model, we use mutual information to select the top 10 accompanying terms of a Chinese query word, and use POS 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 the Co-Model, and EW_{i1}, EW_{i2}, ..., EW_{imi} are augmented by means of different target polysemy resolution models. Intuitively, E_i, EW_{i1}, EW_{i2}, ..., EW_{imi} should have different weights. To Ei is assigned a higher weight, and to the words EWi1, EW_{i2}, ..., EW_{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. In this formula, we assume that the sum of the weights of all the terms in the augmented translation results of C_1, ..., C_n is 1. The words E_1, ...,

E_n take n/(n+1) total weight. In this way, the weight of each E_i is 1/(n+1). The remaining 1/(n+1) is distributed to those equally EW_{ij}'s :

$$\text{weight}(E_i) = \frac{1}{n+1} \tag{2}$$

$$\text{weight}(EW_{ij}) = \frac{1}{(n+1)*\sum_{k=1}^{n} m_k} \tag{3}$$

Thus, six new models, i.e., A1W, ATW, ATTW, U1W, UTW and UTTW, are derived. Finally, we apply the 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 employed the TREC-6 text collection, TREC topics 301-350 [Harman, 1997], and Smart information retrieval system [Salton and Buckley, 1988]. The text collection contained 556,077 documents, and was about 2.2G bytes in size. Because the goal was to evaluate the performance of Chinese-English information retrieval on different models, we translated the 50 English queries into Chinese by hand. The translator is not one of the persons who conducted the query experiments. 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 tags <num>, <title>, <des>, and <narr> denote the topic number, title, description, and narrative fields. Narrative provides a complete description of document relevance for assessors. In our experiments, only the title and description fields were 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>

Table 4. *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 5. *Query Translation of Title Field of TREC Topic 332*

(a)Resolving Translaion Ambiguity Only

Original English query	income tax evasion
Chinese translation by hand	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 Translayion 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**)

Totally, there were 1,017 words (557 distinct words) in the title and description fields of the 50 translated TREC topics. Among these, 401 words had unique translations, and 616 words had multiple translation equivalents in our Chinese-English bilingual dictionary. Table 4 shows the degree of word sense ambiguity in English and in Chinese, respectively. On average, an English query term had 2.976 senses, and a Chinese query term had 1.828 senses only.

In our experiments, the LOB corpus was employed to train the co-occurrence statistics for translation ambiguity resolution, and the ROCLING balanced corpus [Huang, *et al*., 1995] was employed to train the restrictions for target polysemy resolution. The mutual information tables were trained using a window size of 3 for adjacent words.

Table 5 shows the query translation of TREC topic 332. For the sake of brevity, only the title field is shown. In Table 5(a), the first two rows list the original English query and the Chinese query. Rows 3 and 4 list the English translation obtained by the select-all model and by the co-occurrence model by resolving translation ambiguity only. Table 5(b) shows the augmented translation results obtained using different models. Both translation ambiguity and target polysemy were resolved.

The following lists the selected restrictions in the A1 model. The first Chinese term is a query term; the bold English word is its translation equivalent selected by the Co-Model; the words in italics are the augmented restrictions; and the capital letters indicate POS.

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*)

The augmented translation restrictions (poundage, scot, stay), (quota), and (evasion, surtax, surplus, sales tax) were added to "evasion," "income," and "tax," respectively. The terms in the above Chinese augmented restrictions had only one English translation in our dictionary. From the Longman dictionary, we know there are 3 senses, 1 sense, and 2 senses for "evasion," "income," and "tax," respectively. Augmented restrictions were used to deal with target polysemy problem. Compared with the A1 model, only "evasion" was augmented with a translation restriction in the U1 model. This is because " 逃漏 " (tao2luo4) has only one translation while " 所得 " (suo3de2) and " 稅 " (sui4) have more than one translation. Similarly, the augmented translation restrictions were omitted in the other U-models. Next we considered the AT model. The Chinese restrictions which had 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)

Those English words whose POSes were the same as the corresponding Chinese restrictions were selected as augmented translation restrictions. For example, the translation of " 逃避 "_V (tao2bi4) had two possible POSes, i.e., V and N, so only "avoid," "elude," "wangle," "welch," and "welsh" were chosen. The other terms were added in a similar way.

Recall that we use mutual information to select the top 10 accompanying terms of a Chinese query term in the ATT model. The 5th row shows that the augmented translation restrictions for " 所得 " (suo3de2) and " 稅 " (sui4) were removed because their top 10 Chinese accompanying terms did not have English translations of the same POSes. Finally, we considered the ATWCO model. The words "tax," "land," and "surtax" were selected from the three lists in the 3rd row of Table 5(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 performance was stable for all of the +weight (W) models and the enhanced CO restriction (WCO) models, even when there were different numbers of words in translation restrictions. Further more, the enhanced CO restriction models added at most one translated restriction word for each query term. They could achieve performance similar to that of the models that added more translated restriction words. Surprisingly, the augmented translation results may be better than those of monolingual retrieval. Topic 332 shown in Figure 2 is an example.

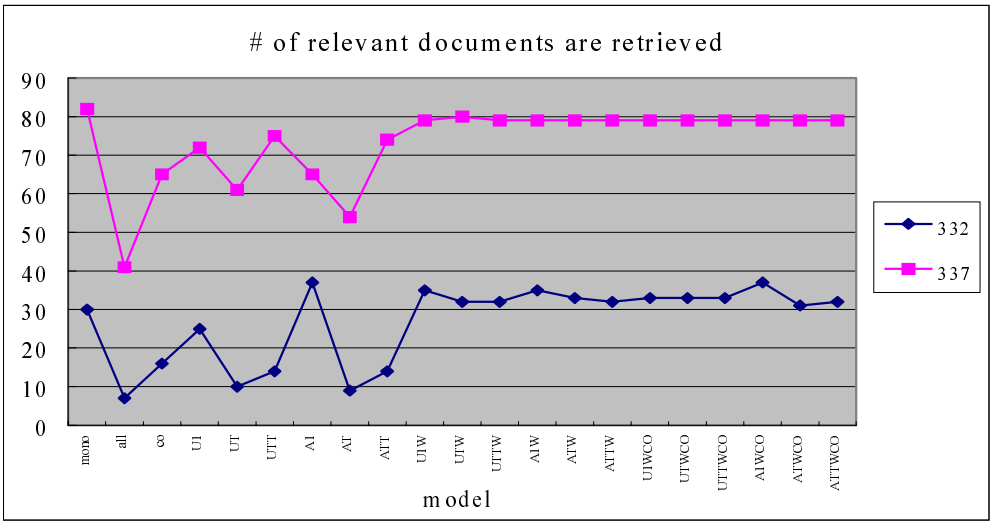

Figure 2 *The Retrieved Performance of Topics 332 and 337*

Table 6. *Performance of Different Models (11-point Average Precision)*

Table 6 shows the overall performance of 18 different models for 50 topics. Eleven-point average precision [Berthier and Ricardo, 1999] for the top 1000 retrieved documents was adopted to measure the performance in all the experiments. Monolingual information retrieval, i.e., with the original English queries to the English text collection, was regarded as a baseline model. The performance was 0.1459 under the specified environment. The select-all model, i.e., where all the translation equivalents are passed without disambiguation, had 0.0652 average precision. About 44.69% performance in monolingual information retrieval was achieved. The performance of monolingual information retrieval is shown in parentheses for comparison. When the co-occurrence model was employed to resolve translation ambiguity, 0.0831 average precision, i.e., 56.96% monolingual information retrieval, was achieved. Compared to the do-nothing model, the performance increased by 27.45% increase.

Now we will consider treatment with translation ambiguity and target polysemy together. Augmented restrictions were formed in the A1, AT, ATT, U1, UT and UTT models, but their performance was worse than that of the Co-model (translation disambiguation only). The major reason was that the restrictions could introduce errors. This can be seen from the fact that the models U1, UT, and UTT were better than A1, AT, and ATT. Because the translation of restrictions from the source language (Chinese) to the target language (English) led to the translation ambiguity problem, incorrect senses would introduce irrelevant documents if the senses were not disambiguated correctly. Thus, the models (U1 and A1) that introduced unambiguous restriction terms performed better than the other models. The performance of model AT was the worst because it augmented more ambiguous terms than the other models did. Controlled augmentation achieved better performance than uncontrolled augmentation.

When different weights were assigned to the original English translation and the augmented restrictions, all the models improved significantly. The performance of A1W, ATW, ATTW, U1W, UTW, and UTTW was about 10.11% better than that of the model with translation disambiguation only. Of these models, the performance change from model AT to model ATW was drastic, i.e., from 0.0419 (28.72%) to 0.0913 (62.58%). This tells us that the original English translation played a major role, but that the augmented restriction still had a significant effect on the performance. In the models A1, AT, ATT, U1, UT and UTT, the weights of the augmented restriction terms were the same as that of the original English translation, so the importance of the original English translation decreased. In this situation, an irrelevant document that only contained some augmented restriction terms may be proposed by our IR system. In the +weight (W) models, to the original English translation was assigned a higher weight than was the augmented terms. The overall performance improved.

We know that a restriction for each English translation is used as a pseudo English context. Thus, we applied the co-occurrence model again to the pseudo English contexts. The performance improved a little. These models added at most one translated restriction word for each query term, but their performance was better than that of the models that added more translated restriction words. This tells us that a good translated restriction word for each query term is sufficient to resolve the target polysemy problem. The performance of U1WCO, which was the best in these experiments, was 62.92% monolingual information retrieval, and increased 40.80% to the do-nothing model (select-all).

5. Concluding Remarks

This paper deals with translation ambiguity and target polysemy at the same time. We have utilized two monolingual balanced corpora to obtain useful statistical data, i.e., word co-occurrence for translation ambiguity resolution, and translation restrictions for target polysemy resolution. Aligned bilingual corpus nor a special domain corpus is required in this design. Experiments show that a gain in performance of about 10.11% can be achieved in resolving both translation ambiguity and target polysemy compare to the method which performs translation disambiguation only in cross-language information retrieval. We also have analyzed two factors: word sense ambiguity in the source language (translation ambiguity), and word sense ambiguity in the target language (target polysemy). Statistics for word sense ambiguities have shown that target polysemy resolution is critical in Chinese-English information retrieval.

This treatment is very suitable for translating very short queries on the Web. Queries on the Web are 1.5-2 words in length on average [Pinkerton, 1994; Fitzpatrick and Dent, 1997]. Because the major components of queries are nouns, at least one word in a short query 1.5-2 words in length is noun. Furthermore, most Chinese nouns are unambiguous, so translation ambiguity is not very difficult, but target polysemy is critical in Chinese-English Web retrieval. Translation restrictions which introduce pseudo contexts are helpful in target polysemy resolution. The application of this method in 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.

Berthier, R.N. and Ricardo, B.Y. (1999) *Modern Information Retrieval.* ACM Press.

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.

Chen, H.H., Bian, G.W. and Lin, W.C. (1999) "Resolving Translation Ambiguity and Target Polysemy in Cross-Language Information Retrieval. " *Proceedings of 37th Annual Meeting of the Association for Computational Linguistics*, 215-222.

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.