Exploiting a Chinese-English Bilingual Wordlist for English-Chinese Cross Language Information Retrieval

K.L. Kwok

Computer Science Department, Queens College,
City University of New York, Flushing, NY 11367, USA
Email: kwok@ir.cs.qc.edu

Abstract

We investigated using the LDC English/Chinese bilingual wordlists for English-Chinese cross language retrieval. It is shown that the Chinese-to-English wordlist can be considered as both a phrase and word dictionary, and is preferable to the English-to-Chinese version in terms of phrase translations and word translation selection. Additional techniques such as target corpus frequency-based term selection and weighting were employed. Experiments show that over 70% of monolingual effectiveness is achievable for the TREC Chinese corpus and retrieval environment with short queries of a few English words.

Keywords: cross language information retrieval; bilingual dictionaries.

1 Introduction

Cross language information retrieval (CLIR) concerns the research, development and implementation of methods and systems to facilitate users having one (source) language skill to do retrieval of collections written in another (target) language. In recent years the Internet has provided the hardware, software and logistics to allow users access websites around the globe easily, literally bringing worldwide information to one's desktop. However, the issue of language begins to surface. One could not make use of information written in a foreign language if one cannot search in that language or comprehend it effectively. Thus, there is a need for cross language retrieval, as well as usable document translation back to the source language for user comprehension [1,2,3]. In this paper, we focus on the retrieval aspect of this access problem. In particular, because English is practically the world language, and China is not only populous but also becoming a world power important in all walks of life, there is added significance in solving the English-Chinese CLIR issues.

Proceedings of the 5th International Workshop Information Retrieval with Asian Languages

Recently investigators in [4] have experimented with the query translation approach for CLIR using the Chinese resources in TREC (Text REtrieval Conference, see for example [5]). This experimental environment consists of 54 topics (queries) in both Chinese and English languages, about 170 MB of GB-encoded texts from Peoples' Daily and XinHua news articles, and the evaluated answer documents for each query that were judged manually. Kwok's approach was to convert the English queries to Chinese using an inexpensive COTS software (Transperfect, see http://www .otek.com.tw). Untranslated words were further looked-up in a small bilingual dictionary. The translated queries were then used for retrieval in our bilingual PIRCS IR system, and results compared to those of monolingual using the original Chinese queries. With short queries of a few words, it was shown that mean average precision from 53% to 62% (i.e. 0.237 to 0.277 vs 0.449) of monolingual can be achieved depending on whether multiple or unique mode translation is employed. Unique mode means the output has a unique translation for each English word or phrase, while multiple mode can have three alternatives to hedge for translation errors. With these short queries, it appears that multiple mode may introduce too much noise and is less effective. Further sophisticated techniques from IR such as pre-translation query expansion and retrieval combination can additionally bring results to over 70% of monolingual.

Using a machine translation package, though workable, has its drawbacks. Commercial software is generally a black box that one can use but not manipulate for experimentation or for fine turning. For example, their dictionary as well as parsing strategies is proprietary. With low cost versions, they do not come with an API, and are supported in a PC platform only. This makes interfacing with an existing IR system working under UNIX for example quite difficult. However, they do allow investigators to establish a basis for comparison using other resources. The purpose of this work is to investigate how bilingual wordlists may be employed to translate the queries, and to compare their CLIR effectiveness with the translation package and to the original Chinese monolingual results. Bilingual wordlists are beginning to be available (either by manual compilation, by statistical alignment based on existing parallel texts [6], or obtained through web mining for example [7]), and it is more under user control. Using wordlist look-up would not give a 'translation-style' output for the queries, but IR does not need accurate syntax, style nor readability -- good content term translation usually would suffice. Various investigators have also employed dictionaries and wordlists with variedly successful results, such as: [8] (Spanish-English), [9,10] (Finnish-English), [10] (Spanish-English), [11] (French-English), [12] (Chinese-English). This paper has focus on English-Chinese CLIR and is organized as follows: Section 2 describes properties of the wordlists we used; Section 3 describes disambiguation methods for wordlist translation, and Section 4 discusses our experiments and results. Section 5 has the conclusion.

2 Properties of LDC Wordlists

Although there are many web sites offering English-Chinese translation facility, actual bilingual dictionaries that are machine-readable for program access are rare. Recently LDC (Linguistic Data Consortium) has made available for research a fairly large bilingual wordlist (also referred to as dictionary) of about 120K records (http://www.morph.ldc.edu/Projects/Chinese). LDC actually provides two lists: one for Chinese-to-English version 2 (ldc2ce), and another from English-to-Chinese (ldc2ec), and it was reported that ldc2ec is obtained from the former. We looked at both lists, found some differences and eventually decided that for English-Chinese retrieval, it may be better to employ ldc2ce as is (without inversion as done in Davies for Spanish-English) rather than the obvious English-to-Chinese wordlist ldc2ec. Examples of both lists appear below:

ldc2ec:

1 case /场/诉讼/实例/病症/怪人/格/容器/框/一组/铅字盘/装入...内/加盖子/事件/盒子/案子/案情/橱/函/盒/例/例子/套/横/
2 cases /案例/
3 china /中国/瓷器/
4 China /中 国/中/
5 human /人的/人性的/通人情的/人间性的/人/人类/人性/
6 separate /分开/区别/使不亲/分居/断掉/各自的/脱离肉体的/分开的事物/套装/分别/间/另/另外/披/脱离/闲/
7 tee off /开 球/开 始/

ldc2ce:

8 人性 /human/
9 人类 /humanity/human race/mankind/
10 人权 /human rights/
11 人权观察 /Human Rights Watch (organization)/
12 人体 /human body/
13 风土人情 /local conditions (human and environmental)/

In ldc2ec, each English entry for lookup consists mainly of single words. Phrase translation is not supported except for about 4000 2 or 3-word entries like Line 7. For many words such as Lines 1 or 6: 'case', 'separate', etc., there are many Chinese translations that may be synonymous with each other or correspond to different senses of the English word. These are separated by slashes '/', but have no indication of which ones are more important. Thus, after picking up all translations for an English word, one has to provide more powerful phrase formation or disambiguation methods to select or weigh them in later processes before retrieval. Note that the two senses of 'china' are differentiated using the case for 'c'. Currently we do not make use of this property.

In the ldc2ce dictionary however, each Chinese term is followed by translations/explanations in English that are also separated by slashes. The translations also may be synonymous with each other, or correspond to the different senses of the Chinese term like line 9. Within a translation, there may be additional clarification or explanation enclosed by parenthesis (line 11 and 13). If we perform string matching between an English query and the translations in ldc2ce, we not only can locate single words but also many phrases of common usage. Thus, ldc2ce may be considered as a word as well as phrase dictionary. Moreover, the structure of the wordlist and the way how query strings are embedded in the translations can provide us with a

mechanism for ranking some translations over others for selection purposes. This report describes some of the methods used for exploiting ldc2ce, and their results for English-Chinese CLIR. The ldc2ce was indexed and a text database created for query translation. Thus, given an English word or phrase in a query, we are able to pull out all the entries that contain it. If the coverage is wide and the translation is good, the correct Chinese word(s) corresponding to the English should have a high probability of being among the entries pulled. The next issue is how to narrow them down.

3 Disambiguating Wordlist Translation

We investigated the 54 short (title section) TREC topics for retrieval. They average about 6.5 English words and the Chinese version averages to 12.5 characters per query. Short queries are more realistic and also harder to get good results. An example topic is shown below with the original Chinese. We assume they are correct translations of each other, and use them as the basis for our monolingual result.

Query 001
Original English:
U.S. to separate the most-favored-nation status from human rights issue in China.

Original Chinese:
美国决定将中国大陆的人权状况
与其是否给予中共最惠 国待遇分离

During query processing, each English word is first truncated of trailing-s according to the Porter's algorithm. Other steps of stemming are not done in order to avoid changes in meaning. A direct lookup of a query word in the wordlist generally would lead to a large number of records pulled because of multiple use, synonyms and the many senses of the word. Methods to disambiguate these entries are therefore essential. We investigated several methods in succession to narrow the number of translations: wordlist structure-based, phrase-based, corpus-based and translation weight for words.

3.1 Wordlist Structure-based Disambiguation

In wordlist-based method, which was introduced in [Kwok 1999], a score is assigned to each entry extracted depending on the format and structure of the translation. If an English source word (e.g. 'human') exactly matches an entry translation like line 8, the translation is assigned a high 'single word' score of ws1: the correspondence in meaning between the English and Chinese word should be precise and uncontaminated by other uses of the same English word. When the source word is embedded in a string of text like /human race/, the translation could be fuzzier and a lesser score ws2 < ws1 is assigned. If the source word occurs in a string that is within a parenthesis pair, like line 13, it could be way off in meaning and is assigned the lowest score of ws3. When a word occurs several times in an entry, only the top score for this entry will be used. Out-of-vocabulary words not found are left as is un-translated. Words that occur more than n times (a threshold) in the wordlist are also left un-translated, and this takes care of most of the function words that are not useful for retrieval. The reason we do not use a stopword list is that some may combine with others to form phrases, and we do not want to miss them. This list of candidates is sorted and only the set of top scoring translations are kept for this English word. Finally, duplicates and sub-strings are also removed. In addition, we prefer short word translations (2 to 5 characters) compared to long ones. Long translations often bring in noise. After this wordlist-based elimination, our sample query becomes 'translated' as follows:

Query 001: by ldc2ce
^美国^
TO
^间 分别 另外 闲 披 脱离^
THE
^大 最 至 半 泰 顶 多数^
^受惠 的 幸运的 有利的^^国家^^
重要 地位 状态 待遇 身份 情形 身分^
FROM
^人性^^不 当 然 错 权利 呀 对头 右边 韪
^ ^是 问题 打 困难 内容 发行 颁布 颁发
纠纷 议题 散发^
IN
^中华 中国^.

Query 001: by ldc2ec
^美国^
TO
^的 不 区 使 间 断 装 亲 分别 套 掉 另外
各自 闲 披 脱离 分开 事物 分居 肉体^^这^
^的 其 量 至 极 半 佳 泰 顶 几乎 大多数
东西 最大限度 大概^^幸运的 受惠的
有利的^^国家 一种 民族 国民^^重要
地位 状况 状态 资格 待遇 身份 情形 身分^^
中 和 用 从 打 由于^^的 通 人类 人间
人情 人性^^的 地 不 方 情 然 实际
健康 正确 错 权利 整顿 适当 啊 平稳 呀
右翼 对头 韪^^是 问题 打 困难 内容
发行 颁布 颁发 纠纷 散发^
IN
^中国 瓷器^.

For ldc2ce translation, it is observed that there are many synonyms and ways of representing the word 'most', as well as the many sense translations (several inappropriate) for words like 'status', 'issue', etc. If we had not selected them based on scores, there will be many more.

The same query using ldc2ec mapping is also shown. Here, one has to deal with many more alternatives, although it is also true that ldc2ce does miss out a correct translation 分开 for the word 'separate'.

3.2 Phrase-based Disambiguation

It is well known that translating phrases using the individual single word components is often erroneous because the meaning of phrases can be non-compositional in nature. Even if it is compositional in meaning, such as 'nuclear power plant', the individual words may have many senses (e.g. plant), and word-based translations would be contaminated with erroneous senses. [Ballesteros & Croft 1998] have shown the importance of phrase translation in English-Spanish CLIR. Thus, to do effective CLIR, phrase translation appears important.

For each record that was pulled by a word in Section 3.1, we first screen for the presence of a phrase match: defined as consecutive words of a query matching on an English description in ldc2ce. If there is an exact match (like line 3 for 'human rights' for Query 001), the translation would be assigned the highest phrase score of ps1 and all single word translations for this phrase position would be ignored. Also longest phrase match overrules sub-phrase matches. There are 19 phrase matches affecting 17 of these 54 short queries, and they bring large improvements (over 10%) in retrieval. Currently we do not consider embedded phrase matches. Two phrases 'most-favored-nation' and 'human rights' in our sample have exact match in ldc2ce. The query translation result becomes:

Query 001:
^美国 ^
TO
^间 分别 另外 闲 披 脱离^
THE
^最惠 国 ^ ^重要 地位 状态 待遇 身份 情形 身分 ^
FROM
^人权 ^ ^是 问题 打 困难 内容 发行 颁 布 颁发 纠纷 议题 散发 ^
IN
^中国 中华 ^.

The two phrase matches: 最惠 国 and 人权 not only bring in the correct terms as used in the manual translation, but also eliminates a number of inappropriate words due to translations of the single word components. These are not present in ldc2ec.

3.3 Corpus-based Disambiguation

One of the functions of a dictionary is good coverage and often many obscure, infrequent Chinese words or phrases are captured in the list. These may not be useful for IR. When an English word has multiple translations, one simple way to select the correct sense(s) is according to the probability of usage. Since the translated words are to be used for retrieval in a target corpus, the frequency of word usage in that corpus could be useful for selection purposes. Thus, the list of translation output from Section 3.2 are segmented and sorted according to corpus occurrence frequency. For each list, preference is given to translation words of two to five characters first, and then single character followed by words of any length. Only the top n (say 3 or 5) highest frequency translations were kept. Long words of six or more characters may be too specific and may risk introducing large amount of noise when they are wrong. Thus, using the top three mappings, our example query translation is further reduced to:

Query 001:
^美国 ^
TO
^分别 另外 脱离 ^
THE
^最惠 国 ^ ^重要 地位 状态 ^
FROM
^人权 ^ ^ 问题 困难 内容 ^
IN
^中华 中国 ^.

Except for 'separate' and 'status', the mapping output is reasonable for this dozen-word English query.

3.4 Translation Weight for Single Words

When an English query word, E1, is translated into several Chinese counterparts, say C11, C12, .. C1n, one could consider them as individual words forming a query, and weigh them in the usual way. For example, the query-focused retrieval process in our PIRCS system leads to a 'tf*idf' type formula where 'tf' is a (sigmoid) function of dk/Ld, and 'idf' is approximately a log function of Nw/Fk (dk = term k frequency in a document of length Ld; Fk = collection term frequency of k and Nw = total

number of terms used (4). Thus, this query-focused retrieval status value (RSV)= S(dk/Ld)*log (Nw/Fk), and the log factor is the inverse collection term frequency.

One could imagine grouping the translated Chinese words into two sets: one (Ga) with members that are synonymous with each other and reflecting the meaning of E1 correctly. The other group Gb contains the rest that are wrong translations of E1. Either group could be empty. Ga plays the part of a synonymous thesaurus class and can be considered as one single high level term with collection term frequency = Σ Fk in Ga. For Gb, one can consider the words as a class of wrong translations and would like to set its 'tf' value to zero. Unfortunately one does not know how to make this division of good and bad translations. Some of these bad translation terms may have low collection term frequencies and lead to large values for their 'idf', adversely affecting retrieval. To minimize this effect, we assign a collection term frequency in the 'tf*idf' formula for every term in a translation set to Max Fk + tc*(ΣFk - Max Fk), where tc is a variable translation coefficient <=1, and the sum is over all terms in the set. When tc=1, the whole translation set is considered synonymous. Max Fk is chosen as an approximation to Fk in Ga only, but we do not know which term belongs to Ga. Weighting for a translation set has been introduced in the past [9,10].

When weighting is employed, we have relaxed the number of terms selected (discussed in Section 3.3) to the top five terms and experimented with different values of the translation coefficient tc.

4 Results and Discussion

Table 1 tabulates different experimental cross language results using 54 short queries and according to the TREC evaluation convention: Column 2: the number of relevant documents retrieved within the top-ranked 1000 documents; Column 3: the mean averaged non-interpolated precision value; Column 4: the mean R-precision; and Columns 5-7 mean precision values at top 10, 20 and 30 documents retrieved. Monolingual result is shown in the first (2) rows for comparison. Subsequent rows tabulate values obtained using different disambiguation methods, percentage of monolingual basis achieved, and the number of better performance vs. the number of unequal cases as compared to the 'single words' entry (ldc2ce) or to the 'single words+tc=0' entry (ldc2ec). These counts are shown only if the difference is significant at the 5% confidence level using the 1-tail sign test.

It is seen that employing ldc2ce, our disambiguation methods successively bring better effectiveness, improving the mean average precision (MAP) from 0.2303 (single-word translations based on word-list structure, Section 3.1), 0.2592 (add phrase translation, Section 3.2), 0.2974 (add translation weighting tc=0, Section 3.4), to .3199 (select top 5 single words based on corpus frequency Section 3.3), and to 0.3217 (setting tc=1). Most of the differences from the 'single word' approach are statistically significant: e.g. best approach with MAP 0.3217 has 35 out of 53 precision values better than 'single word' and 1 equal). This best MAP value is about 72% of monolingual, and is better than results based on an MT software [Kwok 1999] shown under the heading MT-based in the middle of Table 1.

Shown also in Table 1 are the results of using ldc2ec. This wordlist brings in so many translation alternatives that right at the beginning we need to use translation weighting (with tc=0) to help disambiguation and report a result of .2009 average precision. Selection based on corpus term frequency improves this value to 0.2320. At this point, one can also consider employing more powerful disambiguation techniques like term-term co-occurrence data to help form legitimate phrases or select terms. But co-occurrence data is expensive to obtain. Instead, we bring in the phrases from ldc2ce and manually remove 瓷器 (meaning chinaware that was not mapped in ldc2ce) in the translation for 'china', the effectiveness advances to 0.2947. Setting tc=1 brings the result of 0.3083, quite close but still less than using ldc2ce alone. The fact that setting tc=1 gives better results for both wordlists seem to suggest that translation words in a set are mostly synonyms. Thus, we have shown that it is preferable to employ the Chinese-to-English wordlist ldc2ce for English-Chinese CLIR use, rather than the English-to-Chinese version.

Fig.1 plots the precision-recall curves for the basis vs. selected disambiguation methods using the ldc2ce wordlist. It is seen that the gap between monolingual and cross language retrieval is still large.

Fig.2 shows the detailed comparison of the mean average precision for each query between the monolingual basis and the best query translation result. Out of 54 queries, CLIR has 39 performing worse than monolingual and 15 better. This is significant at the 2.5% confidence level according to the 1-tail sign test. Examples of failure include: no term mapping in ldc2ce ('reunification' in query 2, 'protecting' in query 51), wrong or correct but rare

	Rel.Retr		Avg.Prec		R-Prec		P@10		P@20		P@30	
monolingual	**4449**		**.4496**		**.4414**		**.6204**		**.5722**		**.5364**	
basis		100%		100%		100%		100%		100%		100%
					ldc2ce							
single wrds	**2505**		**.2303**		**.2421**		**.3537**		**.3278**		**.2963**	
(signtest basis)		56%		51%		55%		57%		57%		55%
words+ phr	**2838**		**.2592**		**.2739**		**.4385**		**.3671**		**.3266**	
	12/14	64%	14/17	58%		62%	10/13	71%	11/14	64%	11/12	61%
words+ phr	**3274**		**.2977**		**.3164**		**.4500**		**.4204**		**.3963**	
+ tc=0	27/42	73%		66%	26/37	71%	28/41	73%	32/41	73%	30/47	73%
5words+ph	**3438**		**.3199**		**.3269**		**.4741**		**.4435**		**.4062**	
+ tc=0	28/37	77%	34/53	71%	24/36	74%	25/37	76%	31/39	78%	29/43	76%
5words+ph	**3441**		**.3217**		**.3295**		**.4778**		**.4500**		**.4130**	
+ tc=1	29/37	77%	35/53	72%	25/37	75%	28/39	77%	32/40	79%	29/43	77%
					MT-based							
mt software	**2946**		**.2767**		**.2761**		**.3722**		**.3537**		**.3383**	
+20K wdlist		66%		62%		63%		60%		62%		63%
					ldc2ec							
words+tc=0	**2429**		**.2009**		**.2177**		**.3248**		**.2750**		**.2611**	
(signtest basis)		55%		45%		49%		52%		48%		49%
5words	**2770**		**.2320**		**.2443**		**.3500**		**.3185**		**.3000**	
+tc=0		62%		52%		55%		56%		56%	23/33	56%
5words	**3383**		**.2947**		**.3098**		**.4519**		**.4148**		**.3914**	
+tc=0+phr	26/40	76%		66%	22/30	70%	22/32	73%	26/37	72%	32/40	73%
5words	**3466**		**.3083**		**.3215**		**.4667**		**.4259**		**.4031**	
+tc=1+phr	26/40	78%	33/49	69%	23/30	73%	27/36	75%	28/39	74%	34/42	75%

Table 1: 54 Short-Query CLIR Results based on ldc2ce and ldc2ec Wordlist

translations ('women' 姬 query #12, 'cellular phone' 汽车电话 #28, 'bubble economy' (沤, 气泡, 水泡) 经济制度 #37, 'red cross' 赤, 彤, 红色, 丹) (交叉, 十字架, 错, 渡) query 45). One could modify the wordlist to include correct translations of terms incrementally.

5 Conclusion

We show that the English/Chinese bilingual wordlists of medium size (~120K) from LDC are reasonable for CLIR investigations under the TREC environment. However, the structure of the Chinese-to-English version can be employed as a phrase dictionary as well as providing some word translation selection criteria, and appears more useful than the English-to-Chinese version. Using this wordlist alone can bring effectiveness to over 70% of monolingual, which is better than using a COTS translation software under similar conditions. Employing a wordlist is more flexible and gives a user better control. We have made use of corpus term and collection frequency for selection and weighting as a means of diminishing translation ambiguity. One could also use term co-occurrence data for this purpose as done by others. However, term-term co-occurrence data is time consuming to generate on the fly, and is less flexible if calculated in advance. One could also merge multiple bilingual wordlists to improve translation coverage.

6 Acknowledgments

This work was partially sponsored by the Space and Naval Warfare Systems Center San Diego, under grant No. N66001-00-1-8912.

References

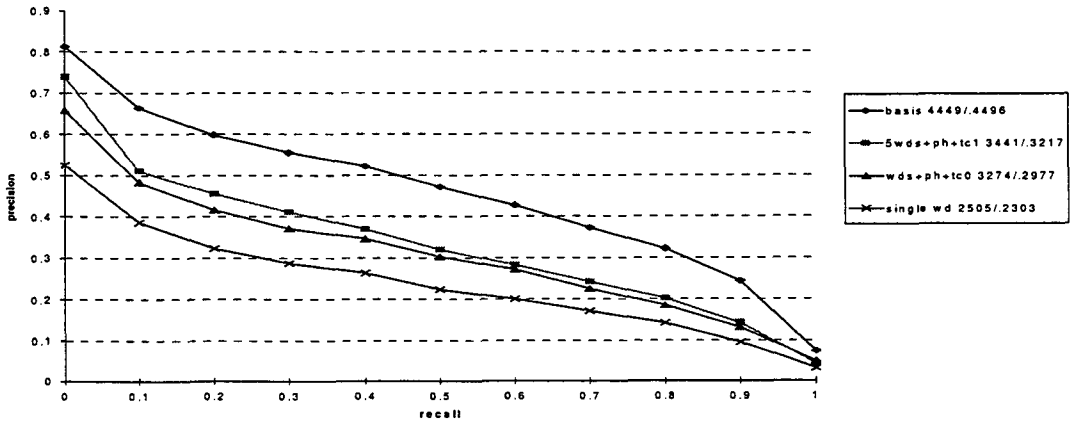
1 G. Grefenstette (ed.) Cross language Information Retrieval. Kluwer, 1998.

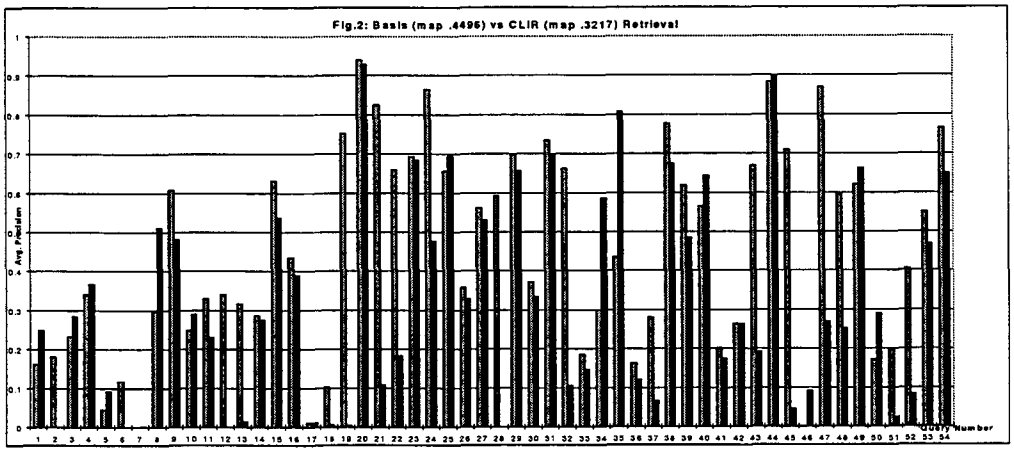
2 Oard, D & Diekema, A. Cross-language information retrieval. *In: Annual Review of Information Science and Technology*, Vol. 33. 1998, pp.223-256.

3 Schauble, P and Sheridan, P. Cross-language information retrieval (CLIR) track overview. *In: Information Technology: The Sixth Text REtrieval Conference (TREC-6)*. Eds. E.M. Voorhees & D.K. Harman, NIST SP 500-240. 1998, pp.31-44.

4 Kwok, K.L. English-Chinese cross language retrieval based on a translation package. *In: MT Summit VII Workshop: Machine Translation for Cross Language IRI.* 1999, pp.8-13.

Fig.1: Various Disambiguation Results vs Basis using Idc2ce Wordlist





5 Voorhees, E.M. & Harman, D.K (1998). Overview of the Sixth Text REtrieval Conference (TREC-6). *In: Information Technology: The Sixth Text REtrieval Conference (TREC-6).* Eds. E.M.Voorhees & D.K. Harman, NIST SP 500-240. 1998, pp.1-24.

6 Wu, DeKai & Xia, X.Y. Large-scale automatic extraction of an English-Chinese translation lexicon. *Machine Translation*, 1995, 9:3-4, pp.285-313.

7 Nie, J-Y, Simard, M., Isabelle, P & Durand, R (1999). Cross-language information retrieval based on parallel texts and automatic mining of parallel texts from the web. *In: Proc. of 22nd Ann. Intl. ACM SIGIR Conf. on R&D in IR.* 1999, pp.74-81.

8 Davies, M. New experiments in cross-language text retrieval at NMSU's Computing Research Lab. *In: Information Technology: The Fifth Text REtrieval Conference (TREC-5). Eds. E.M. Voorhees & D.K. Harman, NIST SP 500-238. 1997, pp.447-453.*

9 Pirkola, A, Keskustalo, H & Jarvelin, K. The effects of conjunction, facet structure and dictionary combinations concept-based cross-language retrieval. *Information Retrieval*, 1999, 1(3), pp.217-250.

10 Pirkola, A. The effects of query structure and dictionary setups in dictionary-based cross-language information retrieval. . *In: Proc. of 21th Ann. Intl. ACM SIGIR Conf. on R&D in IR.* 1998, pp.55-63.

11 Ballesteros, L and Croft, W.B. Resolving ambiguities for cross-language retrieval. *In: Proc. 21th Ann. Intl. ACM SIGIR Conf. on R&D in IR.* 1998, pp.64-71.

12 Eichmann. D., Ruiz, M.E. & Srinivasan, P. Cross-language information retrieval with the UMLS metathesaurus. *In: Proc. of 21th Ann. Intl. ACM SIGIR Conf. on R&D in IR.* 1998, pp.72-80.

13 Bian, G-W and Chen, H.H. Cross-language information access to multilingual collections on the internet. *Journal of the American Society for Information Science*, 51(3), 2000, pp.281-296.