TREC-9 Cross Language, Web and Question-Answering Track Experiments using PIRCS

K.L. Kwok, L. Grunfeld, N. Dinstl and M. Chan
Computer Science Department, Queens College, CUNY
Flushing, NY 11367

Abstract

In TREC-9, we participated in the English-Chinese Cross Language, 10GB Web data ad-hoc retrieval as well as the Question-Answering tracks, all using automatic procedures. All these tracks were new for us.

For Cross Language track, we made use of two techniques of query translation: MT software and bilingual wordlist lookup with disambiguation. The retrieval lists from them were then combined as our submitted results. One submitted run used wordlist translation only. All cross language runs make use of the previous TREC Chinese collection for enrichment. One MT run also employs pre-translation query expansion using TREC English collections. We also submitted a monolingual run without collection enrichment. Evaluation shows that English-Chinese crosslingual retrieval using only wordlist query translation can achieve about 70-75% of monolingual average precision, and combination with MT query translation further brings this effectiveness to 80-85% of monolingual. Results are well-above median.

Our PIRCS system was upgraded to handle the 10GB Web track data. Retrieval procedures were similar to those of the previous ad-hoc experiments. Results are well-above median.

In the Question-Answering track, we analyzed questions into a few categories (like ‘who’, ‘where’, ‘when’, etc.) and used simple heuristics to weight and rank sentences in retrieved documents that may contain answers to the questions. We used both the NIST-supplied retrieval list and our own. Results are also well-above median.

Two runs were also submitted for the Adaptive Filtering track. These were done using old programs without training because we ran out of time. Results were predictably unsatisfactory.

1 Introduction

By some coincidence, all the tasks that we participated in TREC-9 were to us either new or involve new processing of collections. We managed to complete three of the four tasks that we initially targeted with very good results. These are cross language information retrieval (Section 2), the 10GB web data retrieval (Section 3) and the question-answering track (Section 4). The adaptive filtering track (Section 5) was done with little preparation and the result was poor. Section 6 has the conclusions.

2 English-Chinese Cross Language IR

The aim of the task is to retrieve from a Chinese collection documents relevant to queries given in English. The collection consists of about 210 MB of text from three Hong Kong newspapers. Twenty-five queries (#55 to #79) were provided in both English and Chinese. We employed the query translation approach to CLIR by translating the English queries and retrieve in monolingual Chinese. The task is complicated by the fact that the Chinese collection is encoded in BIG5 while our translation resources are mainly GB-code oriented. Since no translation methodology is perfect, we rely on multiple (two) translation methods and retrieval combination technique to lessen wrong or null translations consequences and to provide better results than using one single methodology.

2.1 Query Translation Methodologies

The 25 English topics were first pre-processed by our system to remove some non-content introductory phrases. In addition, sentences that contain negation such as ‘not relevant’, ‘irrelevant’, ‘non-relevant’ are also discarded. We noticed that many narrative sections actually contain only one such sentence, and hence such topics would effectively contain only a title and a descriptive section only. The 25 queries have an average of 9.44 English terms.

The first translation method is based on commercial MT software. Such PC software for English-Chinese are quite common nowadays, costing between scores to about a thousand dollars for a single user license. We consider MT software as a poor man's way of gaining access to a bilingual dictionary with disambiguation technique built-in. For statistical IR, the output that counts is mainly the accuracy of content term translations; other factors such as style, word order, readability, etc. are not important. We tested several packages and finally decided on one called HuaJian (http://www.altlan.com) from Mainland China. It performs very well for the 54 long and short topics and 160MB Chinese collection of TREC 5&6. For example, its untouched translation output attains over 80% of monolingual results. This is used for TREC-9. An example of its quality is shown later in Section 2.2.

A second approach we used for translation is based on automatic dictionary lookup. Most bilingual dictionaries on the web or sold commercially are designed for consultation only. Downloadable dictionaries that can be accessed by program are rare. The LDC (Linguistics Data Consortium) however has compiled two fairly comprehensive English/Chinese wordlists of about 120K in size each, and are available for research purposes (http://www.morph.ldc.edu/Projects/Chinese). One is for English to Chinese, and the other the reverse, and is reported to have similar content. We studied both [Kwok00] and finally decided that the Chinese-to-English version ldc2ce is much more useful for translation purposes because of its dictionary structure. Example entries of the ldc2ce wordlist are shown below:

1	人性	/human/
2	人类	/humanity/human race/mankind/
3	人权	/human rights/
4	人权观察	/Human Rights Watch (organization)/
5	人体	/human body/
6	风土人情	/local conditions (human and environmental)/
7	最惠 国	/most-favored nation (trade status)/

It is seen that if a query has the word ‘human’, one can pick up several mappings that contain this English word in the explanations of lines 1-6. However, because of the wordlist structure, only one of them (line 1) has a precise translation – the other lines may have meaning (and their translation) being contaminated by the way ‘human’ is used in association with other words. Thus, we have a natural way of disambiguating these multiple translations. Moreover, if the word ‘human’ occurs as a phrase like ‘human rights’ in the query, one can also perform string matching in the explanations to pick up line 3 as the sole translation for the phrase instead of individual single word translations. Phrase translations generally are unambiguous and play an important role [BaCr97] for accurate cross language retrieval. Thus, the Chinese-English wordlist can be regarded as both a word and phrase dictionary.

Even with the above considerations, many single words still remain with a large number of mappings. To further disambiguate them, we rely on the retrieval corpus term statistics to help weed down this number. The hypothesis is that the larger the term’s occurrence in a corpus, the higher the probability that the term is a good translation. Thus, for a set of candidate translations of an English word, we keep only the top n most frequent (after ignoring stopwords). However, choosing the threshold n is problematic. Too small a number risks leaving out a correct translation, while too large a number means keeping too much noise. Interestingly, in [Pirkola 98] a method of weighting translations is introduced that allows one can to keep a larger number of translations without seeing the effect of noise. This method is to regard the candidate translations as a synonym set with each term having a collection frequency equal to the sum of the set. Thus, low occurrence frequency terms that are included would not unduly influence the resultant query. Our experiments allow a maximum of six candidate translations to be kept, and this has worked well with the TREC 5&6 Chinese collections in a cross language retrieval environment.

The ldc2ce wordlist discussed earlier is GB-coded, and historically it may have been derived from Mainland China documents. Since our target retrieval collection is in BIG5 and derived from newspapers in Hong Kong, there may be a mismatch in term usage. In the LDC website there is also an available parallel corpus whose content is Hong Kong government laws. Buried in the documents there are many content words or phrases that are followed with translations in parenthesis. We mined some 6000 such translation pairs, converted to GB code, and added to the ld2ce wordlist. This is our resultant translation wordlist.

For the 25 queries, 6 phrases (total 10 with repeats) are extracted. An example query translated via our wordlist is shown below. Numeric values show how many mappings are found for each English word (maximum 5 in this example). They are delimited by ^ as a group. For example, both ‘air’ and ‘pollution’ (first two words) are mapped into three Chinese terms. One phrase translation of ‘government organizations’ is correctly picked up. The word ‘auto’ was assigned two Chinese terms with different senses ‘automobile’ and ‘automatic’. The HuaJian MT translation is also shown, and it is seen that it picks up ‘air pollution’ correctly but misses out the ‘automobile’ sense of ‘auto’. Overall, both translations are quite adequate for CLIR.

Query #CH75 Original English

Air pollution in China .

China's efforts in reducing air pollution, including the
government organizations involved and their
effectiveness in dealing with air pollution in China.

All types of air pollution are relevant, including
industrial, auto emissions, and air pollution from private
sources. that reports a reduction or an increase in air
pollution in China is considered relevant.

Query #CH75 Translation using ldc2ce

^0.33 空气 0.33 样子 0.33 调 ^^0.33 污
0.33 败坏 0.33 弄脏 ^
IN ^0.50 中国 0.50 中华 ^.'
^0.50 中国 0.50 中华 ^^0.33 功夫 0.33 承诺
0.33 工夫 ^ IN
^1.0 减轻体重法^^0.33 空气 0.33 样子 0.33
调 ^^0.33 污 0.33 败坏 0.33 弄脏 ^^0.50
在内 0.50 包含 ^
THE ^1.0 政府机构^ ^1.0 紊^
AND ^0.50 其 0.50 亓 ^^0.33 有效 0.33 有力
0.33 效用 ^
IN ^0.25 处理 0.25 往来 0.25 生意 0.25 往还 ^
WITH
^0.33 空气 0.33 样子 0.33 调 ^^0.33 污
0.33 败坏 0.33 弄脏 ^
IN ^0.50 中国 0.50 中华 ^.
ALL ^0.25 类型 0.25 样式 0.25 试样 0.25 种 ^
OF
^0.33 空气 0.33 样子 0.33 调 ^^0.33 污
0.33 败坏 0.33 弄脏 ^ ARE
^0.50 有关 0.50 相应 ^^0.50 在内 0.50 包含 ^
^1.0 产 ^^0.50 汽车 0.50 自动 ^ ^1.0 排放^
AND
^0.33 空气 0.33 样子 0.33 调 ^^0.33 污
0.33 败坏 0.33 弄脏 ^
FROM
^0.50 私营 0.50 私人 ^^0.17 引起 0.17 来源
0.17 源头 0.17 本源 0.17 情报处 0.17 源点 ^.
THAT ^0.50 报告 0.50 汇报 ^^0.50 降价 0.50
裁减军备 ^
OR AN ^0.20 提高 0.20 增长 0.20 增添 0.20
益 0.20 茁 ^
IN ^0.33 空气 0.33 样子 0.33 调 ^^0.33 污
0.33 败坏 0.33 弄脏 ^
IN ^0.50 中国 0.50 中华 ^
IS ^0.50 考虑过 0.50 被尊重 ^^0.50 有关
0.50 相应 ^.

Query #CH75 Translation using HuaJian MT

在中国的空气污染…
中国努力在 方面 降低 空气污染，包 括政府
组织 包含和他们的效率
在在中国处理空气污染方面。
全部种空气污染是相应的 ， 包括工业国，
自动发射物和空气污染从 私 人资本。那报告一
减少或 者在在中国的空气污染方面的一次增加
被认为相应。

2.2 Query Processing

Each English query was translated into GB-coded Chinese either by HuaJian MT or by our dictionary process. They were then converted into BIG5 for retrieval by a program developed in house that has accuracy similar to the NJSTAR Communicator (http://www.njstar.com). The GB version is also retained to select documents from the TREC 5&6 Chinese GB-encoded collection for the purpose of collection enrichment described in Section 2.4. These selected documents were later converted into BIG5.

2.3 Document Processing

Since the collection is BIG5 encoded, we have modified our document processing programs to support this new coding. Because the queries will be obtained via translation, we also decided to use the translation wordlist as part of our segmentation dictionary to insure correct matching between query and document terms. However, only short words of four or less characters are kept. Our final segmentation dictionary size is about 100K. This is in contrast to our previous work on Chinese retrieval where we derived our segmentation dictionary of about 43K in size from the collection itself. We also follow our tradition to truncate long documents into sub-documents of about 550 characters in size ending on a paragraph. There were 127,938 documents producing a total of 211,536 sub-documents. The master dictionary has 102,156 unique terms. After stopword removal based on a threshold of 20,000, it is reduced to 53,462 terms for retrieval.

2.4 Retrieval Methodologies

After query translation is done, retrieval will be monolingual ad-hoc. However, many techniques can be used to improve retrieval accuracy. Based on experience with the TREC 5&6 Chinese collection used for cross language retrieval, we adopted the following procedures:

Pre-translation query expansion:
This means using the English queries to do retrieval on an English collection and employ pseudo-relevance feedback to expand the queries with English terms. This often can bring highly related terms and more focus on the query topic for later translation. We used this expansion with 15 terms only for queries to be translated via MT. For dictionary translation, we are more cautious as the new expanded terms may bring more noise than signal after translation.

Pseudo-relevance feedback:
This is sometimes known as post-translation query expansion in a cross language retrieval setting. The idea is to use the documents resulting from a first stage retrieval to define the domain of the query and add more Chinese terms to it. This can often lead to substantial improvements of 10 to 30%. Our PIRCS system uses this 2-stage retrieval as a default. We have employed a

standard of 40 top documents for feedback and 70 terms for query expansion.

Collection enrichment:

Pseudo-relevance feedback works only if the first stage retrieval results in a document list that is rich in relevant or highly-related documents. Collection enrichment is the technique of adding an external collection to the target collection in order to improve the probability of acquiring more relevant documents in this first-stage retrieval. The only available Chinese collections we have for this purpose are those of TREC 5&6. However, the latter collection is in GB coding different from the target which is in BIG5. Thus code conversion is necessary. Moreover, the collections are from different years, and have cultural differences (the target collection is from Hong Kong while the enrichment collection is from Mainland China). Thus there is a risk that the procedure may not work.

We are cautious about pre-translation expansion and collection enrichment and only used the procedure for selected runs discussed in the next section.

2.5 Results and Discussion

We submitted one monolingual retrieval pir0Xori as our basis, and three CLIR runs named: pir0Xdin, pir0Xhnd and pir0XHxD. Our convention for pir0X means PIRCS for year 2000 crosslingual experiments, and the last 3 characters differentiate the runs: **'ori'** is the original query monolingual, **'din'** (also referred to as ldc6n) uses our enhanced ldc wordlist with collection enrichment, **'hnd'** combines HuaJian MT (with enrichment) and wordlist without enrichment, and **'HxD'** combines MT with pre-translation expansion and wordlist translation – all with enrichment.

	Rel.retr	Avg.Pre	P@10	P@20	P@30
*** ori**	**616 %**	**.285 %**	**.292 %**	**.236 %**	**.225 %**
hjx0	469 .76	.195 .68	.224 .77	.182 .77	.151 .67
hjx15	566 .92	.206 .72	.208 .71	.158 .67	.143 .63
ldc6	568 .92	.196 .69	.220 .75	.192 .81	.176 .78
orn	613 1.0	.297 1.04	.276 .95	.252 1.07	.231 1.03
hjx0n	469 .76	.223 .78	.252 .86	.184 .78	.153 .68
hjx15n	563 .91	.213 .75	.232 .79	.172 .73	.152 .67
*** din**	575 .93	.216 .76	.232 .79	.194 .82	.175 .78
hjd	509 .83	.221 .78	.236 .81	.196 .83	.169 .75
*** hnd**	507 .82	.240 .84	.252 .86	.206 .87	.179 .79
hndn	493 .80	.245 .86	.260 .89	.198 .84	.173 .77
*** HxD**	**568 .92**	**.245 .86**	**.260 .89**	**.188 .80**	**.169 .75**

Table 2.1: Summary of Monolingual & Crosslingual Results

Internally we had many more runs, consisting of single translation methods: hjx0 and hjx15 (HuaJian MT without and with pre-translation expansion of 15 terms), hjx0n and hjx15n (same as before but with collection enrichment), ldc6 (wordlist only retaining maximum of 6 alternative translation), ldc6n which is also named pir0Xdin (ldc6 with collection enrichment), 'hjd' combines hjx0 with ldc6, and 'hndn' combines hjx0n with ldc6n. In addition, we had another monolingual run using collection enrichment called 'orn'. As discussed in Section 2.4, we do not know if enrichment using vastly different collections will work or not, and submitted the 'ori' monolingual run to be cautious. These results are shown in Table 2.1, where the * rows are our official submissions. The 'ori' row result is used as the basis (indicated by %) for measuring the various crosslingual retrievals. All our runs are automatic without human intervention.

It is surprising that the basic HuaJian MT (hjx0 – 68% monolingual in Avg.Pre) does not perform as well as for the TREC 5&6 environment (over 80% of monolingual). The basic wordlist (ldc6) approach performs as expected: 69% of monolingual in Avg.Pre and quite comparable to hjx0, with an edge for ldc6 – especially in the number of relevants-retrieved which attains an impressive 92% of monolingual. This is possibly due to the allowable 6 alternatives for each English word to be translated, while the MT software necessarily gives only one unique outcome. When pre-translation query expansion is used with MT (hjx15), this relevants-retrieved deficit is removed, but precision at low n suffers. Average precision however improves over both hjx0 and ldc6.

When the first 4 rows are compared with the next corresponding 4 that use collection enrichment, it is seen that this technique brings in 3 to 11% improvement among different measures except for two cases: hjx15n vs hjx15 where the relevants-retrieved practically remains unchanged, and orn vs ori where the precision at 10 documents declines by 5%. Otherwise, results show that collection enrichment works in the majority of cases even with such disparate collections. In particular, the monolingual run orn attains a 4% improvement over our official submission ori in average precision. Again, MT approach (hjx0n) shows good precision values but comparatively low relevants-retrieved. When pre-translation expansion is employed (hjx15n), this value is restored, but precision suffers. The 'din' (same as ldc6n) wordlist run attains good recall and precision in comparison. With collection enrichment, these cross language results now attain over 75% of 'ori' monolingual.

The final 4 rows show different combination runs. Results supports the fact that MT and wordlist approach seem to complement each other well, bringing average precision to 84 to 86% of monolingual. Collection enrichment seems to be an important factor to bring good results, as the 'hjd' row shows that plain hjx0 combined with ldc6 do not perform much better than their singleton runs with enrichment ('hjx0n' or 'din') and attains only about 78% of monolingual. Overall, the best result appears to be our submitted run HxD which combines MT with pre-translation query expansion, and wordlist approach and both with collection enrichment. For fairer comparison, we should use 'orn' (monolingual with enrichment) as the basis. In this case, HxD still attains over 82% of monolingual in average precision, and 93% in relevants-retrieved.

The next Table 2.2 shows how our submitted runs compare with others. For example, pir0XHxD has 17 better, 3 equal to median, and 5 worse for the Avg.Pre measure. pir0Xhnd also has 20 queries better or equal to median, and 5 worse. Of the 5, 1 query in 'hnd' is worst while HxD has 1 best among 17 better than median.

	pir0Xori			**pir0Xdin**			**pir0Xhnd**			**pir0XHxD**		
	>	=	<	>	=	<	>	=	<	>	=	<
AvgPrec	17,2	1	7	18,2	0	6,2	19	1	5,1	17,1	3	5
RR@100	19,6	3	3	16,6	5	4,2	16,5	6	3,2	15,5	8	2
RR@1K	20,6	3	2	18,10	5	2,2	18,11	3	4,1	19,11	3	3

Table 2.2 : Crosslingual Results: Comparing Submitted Runs with Median

We like to emphasize that these blind experimental results were achieved using publicly attainable resources.

3 10-GB Web Track

We participated in the Web Track the first time. The 10 GB represents a 5-fold increase in size from previous collections and is a challenge for our PIRCS system. From the raw text, we removed all the HTML tags like hypertext links, IMG elements, BACKGROUND, COLOR, WIDTH, HEIGHT and similar attributes. Heading and paragraph alignment attributes were replaced by a UNIX new line character. Entity or character references were also replaced by printable ASCII characters. Badly formed entity or character references were deleted. In order to reduce the inherent web data noise, we removed any contiguous strings that were longer than 32 characters. The data also contain many web pages in foreign languages like Spanish, German etc.; they were kept and not removed. To parse the text, we downloaded a C program written by Stephen M. Orth (Sorth@oz.net) and enhanced the program to fit our specific task.

As usual for our PIRCS processing, we broke long documents into approximately 3000 byte (instead of 550 words) long sub-documents ending at paragraph boundaries. This resulted in about 2.6 million sub-documents. After removing words that have a document frequency of less than 3 and more than 180,000, the resultant dictionary has 463K unique terms after stemming and stopword removal.

As before, the TREC-9 Web Track topics has several sections: title, description and narrative. This year we submitted five runs. Four are content-only while the fifth one tries to make use of the link information. The four content-only runs are named pir0Wt1, pir0Wtd2, pir0Wttd and pir0Watd. The prefix convention pir0W represents PIRCS runs year 2000 Web track. The last three characters differentiate the runs: t1 uses the title section only, td2 makes use of both the title and description, ttd is a combination of the retrieval lists from t1 and td1 (another title and description run that was not submitted; it differs from td2 in that the latter adds term variety to the query based on mutual information measure), and atd is a combination of the retrieval lists from pir0Wa1 and pir0Wtd1. a1 means using all sections of a topic.

The title, title-description, and all-section queries have 2.22, 5.32, and 9.12 unique terms respectively averaged over 50 queries. Our link-based run is called pir0WTTD and will be discussed in Section 3.3 while the content-based runs are discussed in Section 3.2.

3.1 General Methodology

We follow our TREC-8 ad-hoc approach by using four methods successively to produce a final retrieval list. These four methods [KwCh98] are: 1) average within-document term frequency to weight short query terms (avtf query term weighting); 2) variable high frequency Zipfian threshold dependent on query size; 3) collection enrichment to improve initial stage output relevant density; and 4) for td2 run only, enhancing term variety in raw queries by adding highly associated terms based on initial retrieval. For collection enrichment, we form a miscellaneous collection by retrieving the top 200 documents from the Question-Answering Track documents. This miscellaneous collection is used to enrich the top-ranked set of the initial stage retrieval. Second stage retrieval employs 25 top documents and 60 terms for pseudo-relevance feedback (long a1, and medium td queries). For short queries (t1) only 30 terms are added. Additionally, we use retrieval list combination to help improve effectiveness. The coefficients of combination are learnt from past results.

3.2 Content-based Retrieval

Our TREC-9 results are summarized in Table 3.1 and their nomenclature has been described previously. The title-description run is significantly better than that of title only run (td2 Avg.Pre 0.2164 vs. t1: avg. prec. 0.1750) -- an improvement of 24%. The lack-luster performance of the title run can be attributed to the fact that three of the queries have misspelled words. Query

464 ("nativityscenes"), query 487 ("angioplast7") and query 463 (“tartin”) produce zero-length queries in our system (we do not perform spell-check and correction). In addition, query 456 ("is the world going to end") and 474 (“how e-mail bennefits businesses”) also produce null queries (after stopword removal and stem conflation). They either contain high collection frequency terms like 'world', 'end', ‘businesses’ that are beyond our threshold and not retained in our dictionary or mis-spelling. We missed e-mail because it was not considered as a single word. Another query #475 (“the compostion of zirconium”) also returns null retrieval list because of the mis-spelling “compostion” that has a legitimate but different meaning after stemming. Even though our initial retrieval list managed to return some documents, they are ranked far lower than the top 25 ranking. This leads to a 2nd retrieval with zero relevants. Another query (#473) has only 1 relevant document, and our system missed it also. Instead of returning an empty ranked list for null queries, our PIRCS engine generates randomly a list of one thousand documents in such circumstances. These lists do not help, and the Avg.Pre values are all zero. Totally we have seven queries with zero Avg.Pre. Adding the description to the query removes these difficulties.

Because the title only run (t1) is not good, its combination with td1 resulting in ttd does not give much improvement over td1. Also, when a1 is combined with td1 resulting in atd, its result is actually worse than a1 by itself. For these web data and questions, it appears that the title run is too poor for combination to work. The best of our submitted runs is pir0Watd. The average precision 0.2209 is 26% better that that of title only. It also has a relevant-retrieved at 1000 documents of 2011, which is about 77% of the pooled documents that have been judged relevant (2617).

Comparisons with the all-sites median average-precision, precision at 100 and 1000 documents are given in Table 3.2. Our content-only runs are well above the median. For example, pir0Watd has avenge-precision better or equal to median in 36 instances with 2 queries achieving the best, and is worse than the median in 14 cases. For title only, the number of queries with precision better, equal or worse than the median are: 32:4:14. Out of the 32 that are above median, 5 have the best value. The medians for title only and non-title-only run are evaluated separately.

Figures for precision at 100 and 1000 documents may be complicated to interpret since the total does not add up to 50 (the number of queries). The reason is that quite a few values are equal to zero. For example, the best, median and worst values for query 473 in title only run for precision at 100 document are all zero. Therefore, our score of zero means that our query 473 achieves the best, median and worst result all at the same time. But it is not better than the median nor it is worse than the median.

3.3 Link-based Retrieval

We tried one run, pir0WTTD, combining contend-based and link-based evidential information. The title-only run, pir0Wt1, retrieval list was used to perform the experiment using the link references in order to improve the retrieval ranking. We assume that a document referenced by many other documents in the output would indicate a higher relevance value compared to documents

			un-submitted						un-submitted					
	t1		**td1**		**td2**		**ttd**		**a1**		**atd**		**TTD**	
	value	% inc	value	% inc	value	% inc	value	% inc	value	% inc	value	% inc	value	% inc
Rel Retr	1518	0	2010	32	2010	32	2005	32	1915	26	**2011**	**32**	2005	32
Avg Prec	.1750	0	.2056	17	.2164	24	.2097	20	.2257	29	**.2209**	**26**	.1418	-19
Prec @ 10	.2180	0	.2960	36	.3020	39	.3180	46	.3320	52	**.2980**	**37**	.1800	-17
Prec @ 20	.1920	0	.2530	32	.2570	34	.2640	38	.2650	38	**.2750**	**43**	.1740	-9
Prec @ 30	.1773	0	.2307	30	.2393	35	.2327	31	.2360	33	**.2433**	**37**	.1680	-5
R-Precision	.1893	0	.2103	11	.2242	18	.2125	12	.2271	20	**.2275**	**20**	.1439	-24

Table 3.1: Automatic Web Track Results for the 50 Queries

	pir0Wt1	**pir0Wtd2**	**pir0Wttd**	**pir0Watd**	**pir0WTTD**
	> = <	> = <	> = <	> = <	> = <
Avg Prec	32,5 4 14	30,3 2 18	34,2 2 14	35,2 1 14	21 1 28
RR @ 100	29,11 15 6,10	30,9 12 8,3	31,10 12 7,2	37,11 8 5,3	23,6 12 15,2
RR @ 1K	32,24 10 8,7	33,23 14 3	35,26 13 2	36,27 12 2	35,26 13 2

Table 3.2: Web Track Results: Comparing Submitted Runs with Median

receiving less or no references, and that re-ranking the output based on this information will improve the result. We determined all incoming links for a document and calculated a link-value for that document (link-value = Ó(1 to 1000) (0.5 * log (1000 – source-rank)). A new rank was then calculated (new-rank = (old-rank + link-value)/2). The result was however disappointing. The table shows that this Avg.Pre value of 0.1418 (pir0WTTD) is considerably lower than the original content only result (pir0Wt1). Further investigation is necessary to determine the reason for the significantly lower results.

4 Query-Answering (QA) Track

4.1 Introduction

The QA Track involves 693 queries retrieving against a collection made up of: AP1-3, WSJ1-2, SJMN-3, FT-4, LA-5, and FBIS-5.

In [LeSJ96] Lewis and Sparck Jones contrast the promise of NLP retrieval systems to the basic statistical IR method. They observe, that while simple NLP strategies could improve text retrieval effectiveness, nevertheless statistical IR method 'has apparently picked some of the low-hanging fruit off the tree'. For example, statistical IR does not attempt word-sense disambiguation, yet 'when a document and a query match on several words, the individual matching words will have the same word sense'. Our QA system is constructed using the methods of classical IR, enhanced with some simple heuristics to pick off some more low-hanging fruit. Since our system lacks natural language understanding, the task is viewed as one of retrieving the best sentence, which is most likely to answer the query.

4.2 Components of our QA Approach

The simplest retrieval strategy seems to be 1) **coordinate matching,** a count of words in a document sentence matching the content words of the query. On top of this, we have added the following considerations:

2) **Stemming:** words are matched even if the are not exactly the same.
3) **Synonyms:** a hand created dictionary of some 300 terms. It contains unusual word forms, which are not handled well by stemming. Most of the entries were taken directly from Wordnet. More automatic use of Wordnet is contemplated for the future. There are four groups of synonym entries as shown in the sample Table 4.1.
4) **RSV:** the retrieval status values of the retrieval system. Given two sentences with the same score based on terms, preference is given to the one that is contained in a higher-ranking document.
5) **ICTF:** inverse collection term frequency gives more credit to less frequently occurring words. For practical reasons, the collection used to obtain the frequencies is the N top retrieved documents. This sometimes causes the system to misclassify the importance of a word. In the future we may want to use the statistics from the entire collection.
6) **Exact** important word: we give extra credit for words deemed important which must occur in the answer. At present, these are the superlatives: first, last, best, highest etc. However, one must be careful: 'best' is good but 'seventh best' is not.
7) **Proximity:** query words in close proximity in the sentence are likely to refer to the same concept as the query. This is currently done only, if all content query words are matched.

Description	Entry
Nationality	ROMAN ROME SPANISH SPAIN PORTUGUESE LUSITANIAN PORTUGAL SICILIAN SICILY FINNISH FINLAND SWEDISH SWEDEN DANISH DENMARK DANES BELGIAN BELGIUM LUXEMBOURGIAN LUXEMBOURG
Unusual Verb forms	KNEW KNOW KNOWN KNOWS LEND LENT LOST LOSE MISBECAME MISBECOME MISSPEND MISSPENT MISTOOK MISTAKE MISUNDERSTOOD MISUNDERSTAND MOLTEN MELT MOWN MOW MOWS
Noun synonyms	MALE MEN MAN FEMALE WOMEN WOMAN
Abbreviations	CAPT CAPTAIN UNITED STATES, US, USA, U.S. UNITED STATES OF AMERICA UNITED KINGDOM, UK U.K. UNITED NATIONS, UN U.N.

Table 4.1 Samples from Synonym Table

8) **Heading:** query words in the headline tag will receive credit if they do not occur in the sentence.
9) **Phrases:** if consecutive words in the query occur in consecutive order in the sentence.
10) **Caps:** capitalized query words.
11) **Quoted:** quoted query words.

A query-analyzer was built to recognize a number of specialized queries. 'Who', 'Where', 'What name' queries are processed by the capitalized answer module. 'When', 'How many', 'How much' and 'What number' are processed by the numerical answer module.

Name Answer Module: we included some simple heuristics to identify the following:

- Persons: Capitalized word not preceded by 'the'
- Places: Capitalized words preceded by 'on', 'in' and 'at'
- Capitalized words. When no other clues are available.

Numerical Answer Module:

- Units: there are classes of queries, which require units. Our system recognizes five types of units: length, area, time, currency and people. See Table 4.2 below
- Dates: There are some queries that have a date year in the question. This date must occur in the sentence or within the date tag.
- Numbers. When no other clues are available.

Type	Entry
Length	METER KM KILOMETER MILE KM CM FEET FT INCH FOOT MM MILIMETER
Area	SQ SQUARE ACRE
Time	MIN MINUTE DAY WEEK YEAR SECOND MONTH
Currency	DOLLAR $ YEN POUND
Population	PEOPLE INHABITANT POPULATION

Table 4.2 Units Recognized

These heuristics are of course not foolproof. For example we assume that a 'Where' question requires an upper case answer, which is not always the case. In particular the following queries have lower case answers:

227. Where does dew come from?
258. Where do lobsters like to live?
385. Where are zebras most likely found?

Selecting 50-byte answer from the top retrieved answer is quite a challenge. We used proximity to query words criterion for selection, and it misses many answers.

The contribution made by each of these components is illustrated by showing their performance for the 198 TREC-8 questions shown in Table 4.3. The results shown are for the long answer (250 bytes) task. The documents used are the top 30 retrieved by the ATT system, which was made available to the participants. Since 28 of the queries have no answer in the top 20, the best possible score is .859.

1)	Term matching	0.439
2)	Stemming	0.470
3)	Synonym	0.478
4)	RSV	0.498
5)	ICTF	0.509
6)	Exact	0.506
7)	Prox	0.515
8)	Head	0.515
9)	8)+Name heuristics	0.566
10)	8)+Numerical heuristics	0.584
11)	8)+Name+Numerical	0.616
12)	8)+Others	0.500
13)	8)+Others+Name+Num	0.589

Table 4.3 QA System for TREC-8 198 Queries

Until Line 8, there were steady improvements in the score when we augment the system with a new component. Line 12 shows, that when Others (Phrases Caps and Quoted described in number 9 10 and 11) are added in to the previous 8, overall performance is actually harmed. Unfortunately this was discovered too late and they were included in the official run.

4.3 Results and Discussions

Four runs named pir0qa[sl][12] were submitted. The s or l indicates short (50byte) or long (250 byte) answers. The submitted runs ending with 1 utilized the top 50 retrievals of the ATT system; the runs ending with 2 used the top 300 sub-documents retrieved by our PIRCS system. PIRCS preprocesses the original documents and returns sub-documents of about 300-550 words in size. Tag information such as heading and date are lost, which may result in small degradation of the final score. Table 4.4 compares the submitted runs to the TREC overall median result using 'strict' MRR evaluation. It seems to indicate that using more documents in the retrieval list

TREC long average	0.350	base
pir0qal1	0.433	+23.77%
pir0qal2	0.464	+32.73%
TREC short average	0.218	base
pir0qas1	0.263	+20.82%
pir0qas2	0.284	+30.65%

Table 4.4. MRR Comparison with TREC-9 Median

helps a lot (pir0q?2 vs pir0q?1). Our simple strategy returns results 20 –33% better than median.

We attempted to analyze our results to see what are the difficulties in QA in general.

Easy questions we missed

The queries may be ranked by their overall performance from all the participants. It is instructive to look at some easy queries that we missed. We comment mainly on pir0qal1, which uses the ATT retrieval list.

207. What is Francis Scott Key best known for?

This is a failure to recognize meta-words, words that are instructions to the query engine rather than real content words. We gave too much credit for matching best and known.

265. What's the farthest planet from the sun?

Our system returned Neptune, which at that time was the farthest. The high-scoring sentence ‘Pluto, the farthest planet from the sun’ from AP901116-0022 was not returned by the ATT retrieval within 30 documents. PIRCS returned this sentence, and pir0qal2 got full credit.

447. What is anise?

In this query, the name Anisi was confused with anise. Since this is a one-word query, the ranking was decided by the document RSV. Perhaps more credit should be given to exact match than stemmed match, or don’t stem proper names at all.

500. What city in Florida is Sea World in?

We had Orlando in our answer, but it was judged incorrect.

504. Who is the founder of the Wal-Mart stores?

Our system did retrieve the correct sentence, but it was long and the correct phrase was not returned. Strangely pir0qas1, the 50-byte answer found the correct phrase.

683. What do river otters eat?

Oops, we again retrieved a correct sentence and filtered out the correct phrase.

688. What country are Godiva chocolates from?

Our system tries to match the word ‘country’.

715. What could I see in Reims?

This is a difficult question.

Difficult questions

There are a number of queries for which NLP is required. Consider the following:

679. What did Delilah do to Samson's hair?

The answer to this can be found in the following three sentences: “Samson, whose story is told in the Book of Judges, was known for feats of enormous strength, such as slaying 1,000 Philistines with the jawbone of a mule. But he was stopped by Delilah, who was sent by the Philistines. She seduced him, learned that the secret to his strength was his hair and cut it off while he was sleeping.” Impressively some systems were able to resolve the references and find the correct answer.

Some queries like:
208. What state has the most Indians?
375. What ocean did the Titanic sink in?
581. What flower did Vincent Van Gogh paint?
688. What country are Godiva chocolates from?

seek a specific class of objects. A good NLP system would make use of knowledge bases, listing states, countries, flowers and oceans. A naive retrieval system like ours, only matches the words state, flower, country and ocean.

Another difficult query is
471. What year did Hitler die?

The answer is in strings like ‘the Nazi leader committed suicide April 30, 1945’ and ‘Hitler killed himself in 1945’, which requires the knowledge that suicide and killed are a form of death.

The two senses of who

The word ‘who’ in a query has two meanings. Consider the queries:
209. Who invented the paper clip?
269. Who was Picasso?

The first question seeks a person, while the second is looking for a description. Our system assumes the first case. Table 4.5 shows that while this does not harm the long answer, it is disastrous for the short. Apparently, other participants had fewer problems with this. At any rate, this illustrates the dangers of applying highly specific heuristics.

	Num of queries	TREC long	TREC short	pirc0qal1	pirc0qas1
who/1	90	0.42	0.30	0.52	0.44
who/2	20	0.51	0.22	0.60	0.08

Table 4.5. Two Types of "Who"

5 Adaptive Filtering Track

This year, by some coincidence, all experiments we participated involve either new programs or heavy extensions to old programs. Moreover, we also took part in other cross language experiments that have deadline quite close to the filtering track. We found ourselves overextended both in time and resources. Some formatting of the OHSU collection for our system was done earlier, but at the end we found no time to do any training or testing. Finally, we decided to use our old programs from TREC-7 & 8 as is without change, and just release them on the OHSU data – to see how bad it gets without training at all. The parameters of the program were trained on newspaper type of documents, while the OHSU data is of course medical documents. One thing we did try to tailor to the new environment was to use the topic descriptions to do retrieval on OHSU87 documents, and expand the queries in a pseudo-relevance feedback fashion, but with the two given relevant documents included. Our filtering runs were supposed to target for utility values rather than precision. The resulting mean T9U score of –55.7and –69.14 were bad. Apparently, expanding the query at the beginning and running a system without training is not a good idea.

6 Conclusion

Our query translation approach to cross language retrieval by combining MT software and bilingual wordlist lookup with disambiguation seems to work quite well – at over 80% of monolingual effectiveness. This is because the topics do not carry too many names or proper nouns that are not translatable by our resources. There were only 25 queries for this experiment. More query types as well as document genre need to be experimented with in the future.

We have succeeded in extending our PIRCS system to handle 10 GB web data. This is done by aggressively screening away a lot of non-textual data. Results were well above median. For topics of a few words, it is necessary to devise ways to handle null queries – either due to spelling errors, or due to terms being filtered out due to high document frequencies.

We presented a QA system based on classical IR methods for sentence retrieval, enhanced with simple heuristics. It achieved above average results that can serve as a baseline. There is much room for future improvement. More heuristics, increased use of knowledge bases, exploring part-of-speech information and more careful query analysis may be employed to attain better performance.

Acknowledgment

This work was partially supported by the Space and Naval Warfare Systems Center San Diego, under grant No. N66001-1-8912. Peter Deng implemented a BIG5/GB conversion program for our use.

References

[BaCr98] 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. pp. 64-71, 1998.

[LeSJ96] "Natural language processing for information retrieval", Comm. of ACM 39, pp.92-101, 1996.

[Kwok00] Kwok, K.L. "Exploiting a Chinese-English bilingual wordlist for English-Chinese cross language information retrieval' Proc. 5th Intl. Workshop on Information Retrieval with Asian Languages (IRAL'00), pp.173-179, 2000.

[KwCh98] Kwok, K.L & Chan, M "Improving two-stage ad-hoc retrieval for short queries." Proc. 21st Ann. Intl. ACM SIGIR Conf. on R&D in IR. pp.250-256, 1998.

[Pirk98] 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. pp. 55-63, 1998.