# TREC-9 CLIR Experiments at MSRCN

Jianfeng Gao[*], Jian-Yun Nie[**], Jian Zhang[#], Endong Xun[*],
Yi Su[#], Ming Zhou[*], Changning Huang[*]

[*] Microsoft Research China, Email: {jfgao, i-edxun, mingzhou, cnhuang} @microsoft.com

[**] Département d'informatique et de recherche opérationnelle, Université de Montréal,
Email: nie@iro.umontreal.ca

[#] Department of Computer Science and Technology of Tsinghua University, China,
Email: ajian@s1000e.cs.tsinghua.edu.cn

**Abstract**

In TREC-9, we participated in the English-Chinese Cross-Language Information Retrieval (CLIR) track. Our work involved two aspects: finding good methods for Chinese IR, and finding effective translation means between English and Chinese. On Chinese monolingual retrieval, we investigated the use of different entities as indexes, pseudo-relevance feedback, and length normalization, and examined their impact on Chinese IR. On English-Chinese CLIR, our focus was put on finding effective ways for query translation. Our method incorporates three improvements over the simple lexicon-based translation: (1) word/term disambiguation using co-occurrence, (2) phrase detecting and translation using a statistical language model and (3) translation coverage enhancement using a statistical translation model. This method is shown to be as effective as a good MT system.

## 1. Introduction

In TREC-9, Microsoft Research China (MSRCN), together with Prof. Jian-Yun Nie from University of Montreal, participated for the first time in the English-Chinese Cross-Language Information Retrieval (CLIR) track. Our work involved two aspects: Finding good methods for Chinese IR, and finding effective translation means between English and Chinese.

Finding a good monolingual IR method is a prerequisite for CLIR. On Chinese monolingual retrieval, we examined the problems such as using different entities as indexes, pseudo-relevance feedback, length normalization, as well as cutting documents done into passages. Each of these techniques gave some improvements to Chinese IR. The best combination of them is used for our Chinese monolingual IR.

On English-Chinese CLIR, our focus was put on finding effective ways for query translation. Large English-Chinese bilingual dictionaries are now available. However, beside the problem of completeness of the dictionary, we are also faced with the problem of selecting the best translation word(s) from the dictionary. To deal with this problem, we used an approach called, *improved lexicon-based query term translation*, which bring significant improvements over the simple approach based on bilingual lexicon. In this approach, we investigated the following three problems: (1) word/term disambiguation using co-occurrence, (2) phrase detecting using a statistical language model, and (3) translation coverage enhancement using a statistical translation model.

In section 2, we introduce briefly our work on finding the best indexing unit for Chinese IR. In section 3, we describe in detail the proposed method -- *improved lexicon-based query term translation*, and compare with the method using a machine translation (MT) system in CLIR. In section 4, we describe the use of query expansion techniques. In section 5, experimental results are presented. Finally, we present our conclusion in section 6.

## 2. Finding the Best Indexing Units for Chinese IR

It is well known that the major difference between Chinese IR and IR in European languages lies in the absence of word boundaries in sentences. Words have been the basic units of indexing in traditional IR. As

---

[** #] This work was done while these authors were visiting Microsoft Research China.

Chinese sentences are written as continuous character strings, a pre-processing has to be done to segment sentences into shorter units that may be used as indexes. Indexing units for Chinese IR may be of two kinds, words or *n*-grams [Nie, 2000].

When using words, several types of knowledge may be used: manually constructed dictionary that stores a set of known words, heuristic rules on word formation, or some statistical measures based on co-occurrences of characters. A dictionary-based segmentation is widely used to identify all occurrences of the dictionary words in a sentence. If there are word segmentation ambiguities, the longest-matching strategy is usually used to select the best choice. There are mainly two problems of this approach. The first is the loss in recall. A long word may contain several shorter words. In the longest matching, only the longest word is identified as an index, and all the included short words are ignored. For example, if 操作系统 (operating system) is identified as a word, 操作 (operating) and 系统 (system) will not. However, in practice, we also refer to an "operating system" by just "system". Although the word "system" is included in "operating system", it will not be considered as a completely independent index for IR. Therefore some relevant documents will not be retrieved. The second problem is the unknown word problem. Especially, many proper nouns, which play an important role in IR, are not in the dictionary, and are not considered as indexes.

Another kind of indexing units is *n*-grams. This method does not require any linguistic knowledge. Usually, one chooses n-grams of lengths 1 or 2 (uni-grams or bi-grams). Longer *n*-grams are rarely used due to the higher memory cost and their marginal improvement over bi-grams. In comparison with words, the advantage of bi-grams lies in its robustness to unknown words. For example, for proper nouns that are not in the dictionary, such as 大亚湾 (a place in southern China), word segmentation will segment the proper noun into three characters, i.e. 大, 亚, and 湾. When using bi-grams, we can still use part of the proper nouns as indexes, i.e. 大亚, 亚湾. If both bi-grams occur in the same document, there is a higher probability that the document concerns 大亚湾, than the documents where the three single characters occur. Political terms or abbreviations (e.g. 三乱 – three turmoils), and foreign names (e.g. 皮纳图博火山 - Mount Minatubo) are similar examples showing the advantage of using bi-grams.

Words and n-grams represent two different ways to represent a document – one relies on linguistic knowledge and the other on statistical information only. It is a common practice to combine different evidence to judge document relevance. So it is also reasonable to combine *n*-grams with words.

To sum up, we can create three possible representations for a document and a query as shown in figure 1, i.e. words, characters, and bi-grams. We also see that some correspondences may be created across representations, if different representations are integrated (for example, between words and characters).

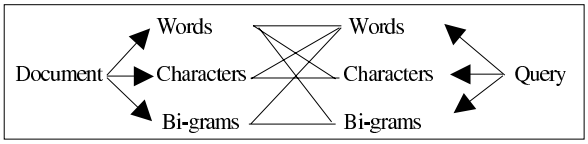


**Fig. 1.** Possible representations in Chinese IR

In order to determine the best indexing units, we conduct a series of test tests on TREC 5&6 Chinese data [Harman, 1996]. The documents in the collection are articles published in the People's Daily from 1991 to 1993, and a part of the news released by the Xinhua News Agency in 1994 and 1995. A set of 54 English queries (with translated Chinese queries) has been set up and evaluated by people in the NIST (National Institute of Standards and Technology).

Once Chinese sentences have been segmented into separate items, traditional IR systems may be used to index them. These separate items are called "terms" in IR. For our experiments, we used a modified version (the modifications are made in order to deal with Chinese) of the SMART system [Buckley, 1985].

The following methods have been compared:

1. using the longest matching with a small dictionary and with a large dictionary
2. combining the first method with characters
3. using full segmentation with or without adding characters
4. using bi-grams and characters
5. combining words with bi-grams and characters

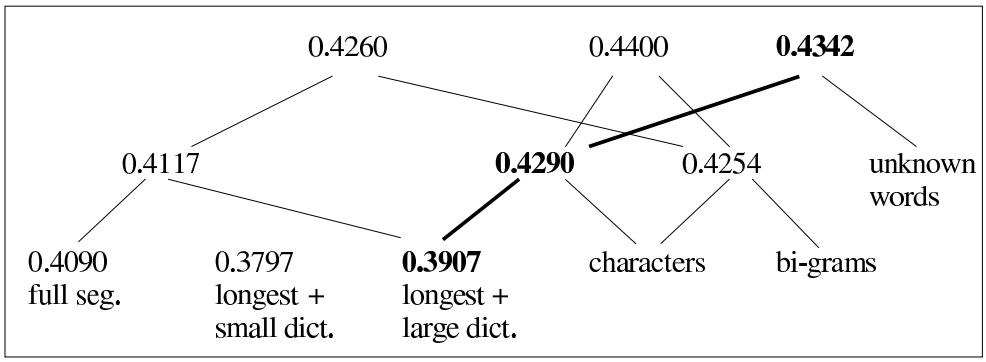


**Fig. 2.** Results of indexing units for Chinese IR

The results of this series of experiments are summarized in the figure 2, detailed description can be found in [Nie, 2000].

In order to examine the impact of dictionary in word segmentation, two different dictionaries are used. The small dictionary contains 65,502 entries. The large dictionary contains 220K entries, containing not only all entries in the small dictionary, but also a large number of phrase, including date expressions (e.g. 一九三四年 - year 1934), suffix structures (e.g. 使用者 - user), etc. The second dictionary is more complete. In both cases, we use the same forward longest-matching strategy. Using the first dictionary, we obtained an average precision of 0.3797. Using the second dictionary, the average precision is increased to 0.3907. We can see that a better dictionary can increase IR effectiveness to some extent.

To remedy the loss in recall caused by the use of the longest words, we complement the longest words by single characters. We obtain nontrivial improvements. In the case of large dictionary, we achieve an average precision of 0.4290 (9.8% improvement). It turns out that simply adding single characters is a more effective way to increase IR performance than increase the dictionary size. Another way to increase recall is to extract the short words implied in long words, called full segmentation. In this case, we obtain an average precision of 0.4090. Although the performance is better than using the longest words alone, it is worse than the method by adding single characters. One of the reasons might be due to the cross-word segmentation phenomenon; i.e. some words extracted in full segmentation are composed of parts of two different words. For example, from the string 开发油田 (exploit a oilfield), we not only extract the correct words 开发 (exploit) and 油田 (oilfield), but also 发油 (hair oil).

Previous studies [Kwok, 1997; Nie, 1999] show that when combining bi-grams with uni-grams, the IR performance is better. We repeat this experiment here, and obtained an average precision of 0.4254. This performance is comparable to the best performance we obtained using words. This is largely contributed to the robustness for dealing with unknown words by *n*-grams.

As bi-grams and words have their own advantages, we try to combine them to benefit from both. Theoretically, such a combination would result in a better precision (due to words) and an increased robustness for unknown words (due to n-grams). Unfortunately, the experimental result is not promising enough. We obtain slightly improvements of 2.6% (average precision 44.00%) over the uncombined case, whereas the space and the time of indexing are more than doubled.

After word segmentation, we noticed that some important proper nouns and noun phrases have not been recognized as words, but segmented into single characters, such as 皮纳图博火山 (Mount Minatubo). Therefore, we used NLPWin[1] to recognize multi-word phrases and unknown words. NLPWin first tags texts using a Chart-parser (with a dictionary). For unknown words, a category is guessed according to its context. Special rules have also been integrated to recognize proper nouns. As a consequence, most Chinese or English proper nouns can be tagged and recognized correctly. Some political terms and abbreviations (e.g. 中越 - Sino-Vietnam) can also be recognized. Using NLPWin, we created another set of words that is added

[1] The NLPWin system is a natural language processing system developed by Microsoft Research. The system converts text into a parse tree that represents the syntactic structure and then into its logical form that reflects the meaning of the text. These representations can then be used for tasks such as grammar checking, machine translation, and information retrieval.

| | riginal vg.P. | ew vg.P. | Impr. | New words added |
|---|---|---|---|---|
| 9 | 0.3648 | 0.4173 | 14.4% | 毒品买卖 (drug sale) |
| 23 | 0.3940 | 0.5154 | 30.8% | 联合国安理会 (Security committee of UN), 和平建议 (peace proposal) |
| 28 | 0.4824 | 0.5034 | 4.4% | 蜂窝式 (cellular), 交换网 (interchange network) |
| 46 | 0.3483 | 0.4192 | 20.4% | 中越 (Sino-Vietnam) |
| 47 | 0.5369 | 0.5847 | 8.9% | 皮纳图博火山 (Mount Minatubo), 臭氧层 (ozone layer) |
| 54 | 0.6778 | 0.7005 | 3.3% | F-16, 八．一七 (August 17) |

Table 1: Impact of unknown word recognition on some queries.

to our original dictionary. From the 54 queries, 80 new words have been recognized. Most of them are proper nouns or noun phrases. The addition of unknown words had positive impact for 10 queries out of 54, while the effectiveness is reduced for 4 queries. Table 1 contains some examples of queries for which the addition of new words has positive impacts. As we can see in Fig. 2, the global effect of adding an unknown word detection is positive.

We can see from figure 2 that as long as different kinds of indexes are combined the IR performance increases. The question now is whether the combination is worth the cost. In taking into account both effectiveness and cost, we think the combination should go in the direction represented by the bold lines in figure 2. For our experiments in TREC9, we will use the combination of the longest words, single characters and detected unknown words for Chinese IR.

# 3. Query Translation

The methods for query translation, proposed recently, fall into three categories: (1) using MT systems, (2) using parallel corpora, and (3) using bilingual lexicons. The third method is the simplest way to implement because of its simplicity and the increasing availability of machine-readable bilingual lexicons. Therefore, we decided to start with this method in TREC9 and try to improve it by adding other tools.

The main problems we observe on this simple method are: 1) the dictionary used may be incomplete; and (2) it is difficult to identify the correct word sense from the lexicon. To deal with these issues, we used an *improved lexicon-based query translation*. It tries to improve the lexicon-based translation through (1) word/term disambiguation using co-occurrence, (2) phrase detecting and translation using a statistical language model, and (3) translation coverage enhancement using a statistical translation model. In what follows, we will describe each of them in detail.

## 3.1 Word/term disambiguation

It is assumed that the correct translations of query terms tend to co-occur in target language documents and incorrect translations do not. Therefore, given a set of original English query terms, we select for each term the best translation term such that it co-occurs most often with other translation words in Chinese documents. Finding such an optimal set is computationally very costly. Therefore, an approximate algorithm is used. It works as follows. Given a set of $n$ original query terms $\{s_1,\ldots,s_n\}$, we first determine a set $T_i$ of translation words for each $s_i$ through the lexicon. Then we try to select the word in each $T_i$ that has the highest degree of cohesion with the other sets of translation words. The set of best words from each translation set forms our query translation.

The cohesion is based on term similarity calculated as follows. For terms $x$ and $y$, their similarity is:

$$SIM(x,y) = p(x,y) \times \log_2\left(\frac{p(x,y)}{p(x) \times p(y)}\right) - K \times \log_2 Dis(x,y) \quad (1)$$

where

$$p(x,y) = \frac{c(x,y)}{c(x)} + \frac{c(x,y)}{c(y)}$$

$$p(x) = \frac{c(x)}{\sum_x c(x)}$$

and $c(x,y)$ is the frequency that term $x$ and term $y$ co-occur in the same sentences in the collection, $c(x)$ is the number of occurrence of term $x$ in the collection,

$Dis(x,y)$ is the average distance (word count) between term $x$ and term $y$ in a sentence, and $K$ is a constant coefficient.

The cohesion of a term $x$ with a set $T$ of other terms is the maximal similarity of this term with every term in the set, i.e.

$$Cohesion\ (x, T) = Max_{y \in T} SIM\ (x, y)$$

The greedy algorithm used to select the word translations is as shown in figure 2.

The term-similarity matrix is obtained via a statistical model, which is trained using a large Chinese corpus of MSRCN consisting of 1.6 billion characters.

==========

For each $s_i$ ($i$ = 1 to $n$), retrieve a set of senses $T_i$ from the lexicon;

For each set $T_i$ ($i$ = 1 to $n$), do

For each term $t_{ij}$ in $T_i$, do

For each set $T_k$ ($k$ = 1 to $n$ & $k$<>$i$), compute the cohesion $Cohesion(t_{ij}, S_k)$;

Compute the score of $t_{ij}$ as the sum of $Cohesion(t_{ij}, S_k)$ ($k$ = 1 to $n$ & $k$<>$i$);

Select the term $t_{ij}$ in $T_i$ with the highest score, and add the selected sense into the set $T$.

==========

**Fig. 3**. Greedy algorithm to find best translations

## 3.2 Phrase detecting and translation

The translation of multi-word phrases is usually more precise than a word-by-word translation [Ballesteros, 1998], since phrases usually have fewer senses. However, if a phrase is not stored in a lexicon, we usually can do nothing. Unfortunately, in TREC-9 query set, more than 50% phrases are not in our lexicon.

In our experiments, we try to incorporate some translation patterns between English and Chinese. For example, a (NOUN-1 NOUN-2) phrase is usually translated into the (NOUN-1 NOUN-2) sequence in Chinese, and a (NOUN-1 of NOUN-2) phrase is usually translated into the (NOUN-2 NOUN-1) sequence in Chinese. So if we can detect the English phrase of some patterns, we can guess the form(s) of the translation phrases. For instance, the translation of the multi-word phrase "drug sale" is 毒品(drug)/买卖(sale), and the translation of the multi-word phrase "security committee of UN" is 联合国(UN)/安理会(security committee).

To do this, we use again NLPWin to detect phrases in the English queries. We selected a set of 40 English patterns ($PAT_{Te}$ ) that are often used in phrases. For each of them, we estimate the probability of the order of translation words, $p(O_{Tc}|PAT_{Te})$. Then the best translation phrase is the one that maximizes the following function,

$$Tc = \text{argmax}\ p(O_{Tc}|PAT_{Te})\ p(Tc) \qquad (2)$$

where $p(Tc)$ is a priori probability whose value is given by the bigram language model. The bigram language model is trained using the same large Chinese corpus of MSRCN. For the moment, an approximate probability $p(O_{Tc}|PAT_{Te})$ is assigned by a linguist because of the lack of training data.

## 3.3 Using translation model

Translations stored in lexicons are always limited, no matter how complete they are. Parallel texts may contain additional translations. Therefore, we used a statistical translation model trained from a set of parallel texts as a complement of the previous methods.

Given a set of parallel texts in two languages, they are first aligned into parallel sentences. While the lexically based techniques use extensive online bilingual lexicons to match sentences [Chen 93], statistical techniques require almost no prior knowledge and are based solely on the lengths of sentences, i.e. length-based alignment method. We use a novel method that incorporates both approaches [Liu, 95]. First, the rough result is obtained by using the length-based method. Then anchors are identified in the text to reduce the complexity. An anchor is defined as a block that consists of $n=3$ successive sentences. Finally, a small, restricted set of lexical cues is applied to obtain further improvements.

Once a set of parallel sentences is obtained, word translation relations are estimated. Chinese sentences are first segmented into word strings by using a

dictionary, containing approximately 80 thousand words, in conjunction with an optimization procedure described in [Gao, 2000]. The bilingual training process employs a variant of the model in [Brown, 1993] and it is based on an iterative EM (expectation-maximization) procedure for maximizing the likelihood of generating the English given the Chinese portion. The output of the training process is a set of potential Chinese translations for each English word, together with the probability estimate for each translation.

The problem we often have with translation models is the unavailability of parallel texts for Chinese-English. To solve the problem, we conducted a text-mining project in the Web to find parallel texts automatically [Nie, 1999]. We select about 20,000 parallel document URLs, from which 870,414 pairs of sentences are selected for model training. The training data amounts to 74MB Chinese texts and 51MB English texts.

Let's assume that all multi-word phrases have been translated by equation (2). By combining translation model, we can arrive at the following equation of query phrase translation:

$$Tc = \arg\max \; p(Te \mid Tc)\, SIM(Tc) \qquad (3)$$

where $p(Te|Tc)$ is the translation probability of Chinese term $Tc$ to English term $Te$, and $SIM(Tc)$ is the sum of the maximum similarity score of the selected translation set $Tc$, which is estimated by algorithm in figure 2 and equation (1).

## 3.4 Tests of Query Translation on TREC 5&6

We carried out a series of tests to compare our improved method with the following four cases:

1. *monolingual*: retrieval using provided (manually translated) Chinese queries;
2. *simple translation*: retrieval using query translation obtained by looking up the bilingual lexicon;
3. *best-sense translation*: retrieval using query translation obtained by manually selecting the best senses among the senses in the bilingual lexicon for each query term;
4. *machine translation*: retrieval using translation queries obtained by the machine translation software system.

In our experiments, the English-Chinese bilingual lexicon we used comes from LDC (http://morph.ldc.upenn.edu/Projects/Chinese/). It contains 110,834 English entries as well as their corresponding Chinese translations. For each English entry, there are usually several Chinese translations. The *simple translation* works in two modes. One is u-mode that selects the most Chinese translation for each English term. The other is m-mode that selects the first three (if it contains no less than three translations) frequent-used Chinese translations.

For *best-sense translation*, we manually select one translation for each term in queries, for multi-word phrases not found in the lexicon, we translate it word-by word.

The *improve translation* makes use of the following tools described previously: (1) the term-similarity matrix for term disambiguation, (2) the language model for phrase translation, and (3) the translation model for lexicon coverage enhancement.

The use of an MT package is convenient for CLIR since it takes care of problems like word morphology, parsing, etc. On the other hand, its internal working scheme and dictionaries are proprietary, and one can only treat it as a black box and has to accept the output as is with little possibility of changing them. In our experiments, a commercial English to Chinese machine translation software system called IBM HomePage Dictionary™ 2000 is used. The system is released recently by IBM. It contains a 480K English-Chinese dictionary, which consists of both words, frequently used phrases (such as "information retrieval"), acronyms (such as "IBM"), and proper nouns (such as "Microsoft"). It can translate a word, phrase, sentence or whole document. According to our survey, this system is one of best machine translation product currently on the market. The result of query translation by the IBM system seems reasonable; less than 2% of the words are left untranslated, most phrases are translated as a whole, and the ambiguity problem of most words are solved successfully.

The results of this series of experiments on query translation are summarized in table 2. As can be expected, the simple translation methods are not very good. Their performances are lower than 60% of the monolingual performance.

The best-sense method improves the performance of the simple translation method. It achieves 73.05% of monolingual effectiveness. However, it is still worse than our improved translation method, which achieves a 75.40% performance of that of monolingual IR.

IBM HomePage Dictionary™ 2000 is a very powerful machine translation system. Using it for query translation, we can achieve 75.55% of monolingual effectiveness. On the other hand, the fact that the most

powerful commercial machine translation system performs almost the same as our improved method indicates the effectiveness of our query translation technique for CLIR.

The best performance is achieved by combining linearly two sets of translation queries obtained by machine translation method and the improved translation method. It is over 85% of monolingual effectiveness. The motivation of combination of different translation methods is that different translation systems would complement each other.

| | Translation Method | Avg.P. | % of Mono. IR |
|---|---|---|---|
| 1 | Monolingual | 0.5150 | * |
| 2 | Simple translation(m-mode) | 0.2722 | 52.85% |
| 3 | Simple translation(u-mode) | 0.3041 | 59.05% |
| 4 | Best-sense translation | 0.3762 | 73.05% |
| 5 | Improved translation | 0.3883 | 75.55% |
| 6 | Machine translation | 0.3891 | 75.40% |
| 7 | 5 + 6 | 0.4400 | 85.44% |

Table 2: Average retrieval precision of the English translation queries.

# 4. Query Expansion

## 4.1 Pre-translation & Post-translation Query Expansion

Earlier work showed that query expansion can greatly reduce the error associated with dictionary translation [Ballesteros, 1998]. A popular method of query expansion in TREC experiments is the 2-stage pseudo relevant feedback. At first, raw queries are used to retrieve a ranked list of documents. Then the set of *n* top-ranked documents is used for query expansion. Usually, we expand the initial query by adding *m* top-frequent terms from the *n* top-ranked documents. Through a preliminary experiment, we established the optimal values (with respect to our test collection) of *n* and *m*.

In CLIR, queries can be expanded prior to translation, after translation or both before and after translation. In English-Chinese CLIR, pre-translation query expansion means using a separate English collection for pre-translation retrieval in order to expand the English query with highly associated English terms. These terms may help focus on the query topic and bring more translated terms that together are useful for disambiguating the translation.

## 4.2 Sub-Documents[2]

The purpose of dividing a document into a sequence of subdocuments (or passages) of certain length is to create a length normalization effect. It is also hoped that each passage will concentrate on a specific topic, or at least on fewer topics than a complete document. Real documents can be very long (e.g. 2 MB) and very short (e.g. a few words). When such documents form the top-ranked pool, one would face a lot of noise during term selection. Using sub-documents have the advantage of being able to define a more specific domain that is less noisy for query expansion. In our experiments, the medium length of subdocument is set at 550 words. We used pivot normalization [Singhal, 1996] in Smart (the *ltu* weight scheme), given the old weight, *w*, of a term, the new weight, *w'*, can be written as:

$$w' = \frac{w}{(1.0 - slope) \times pivot + slope * nbTerm} \quad (4)$$

where *pivot* is the average numbers of terms in a documents, *nbTerm* is the actual number of terms in the current document, *slope* is a parameter determining the impact of document length normalization, and a typical setting is *slope* = 0.1.

## 4.3 Tests of Query Expansion on TREC 5&6

We conducted another series of experiments to measure the effectiveness of our query expansion techniques. The experimental results on monolingual IR are shown in table 4. The indexing units used in this case are the longest words and single characters. The query expansion was performed by adding the top 500 terms from the top 20 documents of the initial ranked documents. When using the SMART *ltc* weighting scheme, we obtained 9.1% improvement over the initial retrieval. Move improvements are obtained when we do retrieval and feedback using sub-documents of a certain size (550 words). The document length normalization, i.e. *ltu*, also leads to limited improvements. It is

[2] The idea of sub-document and its implementation details are introduced by Prof. K.L. Kwok during his one-month visit at MSRCN in June, 2000.

interesting to note that the best result is achieved when we use the *ltc* weighting scheme at the 2-stage retrieval, but keep the *ltu* at the 1-stage retrieval (last row of the table 3).

| Sub-doc | 1-stage – weighting | 2-stage – weighting | 1-stage: Avg.P. | 2-stage: Avg.P. |
|---|---|---|---|---|
| No | *ltc* | *Ltc* | 0.429 | 0.476 |
| Yes | *ltc* | *Ltc* | 0.435 | 0.485 |
| Yes | *ltu* | *Ltu* | 0.461 | 0.489 |
| Yes | *ltu* | *Ltc* | 0.461 | 0.515 |

Table 3: Average retrieval precision of the expanded queries for Chinese IR.

| Method | Avg.P. | % 1-stage |
|---|---|---|
| 1-stage retrieval | 0.3249 | * |
| 1+Post-translationQE | 0.4280 | 31.7% |
| 2+Pre-translation QE | 0.4400 | 35.4% |

Table 4: Average retrieval precision of the expanded queries.

The overall results of query expansion on CLIR are shown in table 4, which provides the average retrieval precision of 1-stage retrieval (without query expansion) as a baseline, as shown in row 2.

Post-translation expansion was performed by adding the top 500 terms from the top 20 documents of the initial ranked documents after query translation. It brings about 31.4% of improvements, as shown in row 3,

We experimented with pre-translation query expansion using the Foreign Broadcasting collections of TREC and used various levels of query expansion. An English query is first used to retrieve a set of documents from this collection. The top 10 English terms from the top 10 documents are used for query expansion before query translation. As shown in Table 4, the pre-translation QE brings an additional improvement of about 2.8% compared to not using it.

# 5. Experiments in TREC 9

## 5.1 Data

The documents in the TREC 9 Chinese collection are articles published in Hong Kong Commercial Daily, Hong Kong Daily News, and Takungpao. Some statistical data are shown in table 5. A set of 25 English queries (with translated Chinese queries) has been set up and evaluated by people in the NIST.

| Source | Dates | Size |
|---|---|---|
| Hong Kong Commercial Daily | 8/98-7/99 | ~100MB |
| Hong Kong Daily News | 2/99-7/99 | ~80MB |
| Takungpao | 9/98-9/99 | ~80MB |

Table 5: TREC 9 data.

## 5.2 Results

We submitted 5 runs, as shown in Table 6.

The monolingual run (MSRCN1) uses the longest words, single characters as well as automatically detected unknown words for indexing. The weighting scheme used is that *ltu* is used for the 1-stage retrieval and *ltc* for the 2-stage retrieval.

The MSRCN2 run is the one in which our improved method is combined with the IBM MT system. No pre-translation QE is used.

| Run # | Avg.P. | % of mono. IR | Method |
|---|---|---|---|
| MSRCN1 | 0.2995 | * | Mono-lingual IR |
| MSRCN2 | 0.3083 | 102.9% | CLIR without pre-translation query expansion |
| MSRCN3 | 0.2974 | 99.3% | CLIR with pre-translation query expansion |
| MSRCN4 | 0.2677 | 89.4% | CLIR with *improved translation* only. |
| MSRCN5 | 0.2623 | 87.6% | CLIR with IBM MT system only |

Table 6: Average precision of the submitted runs

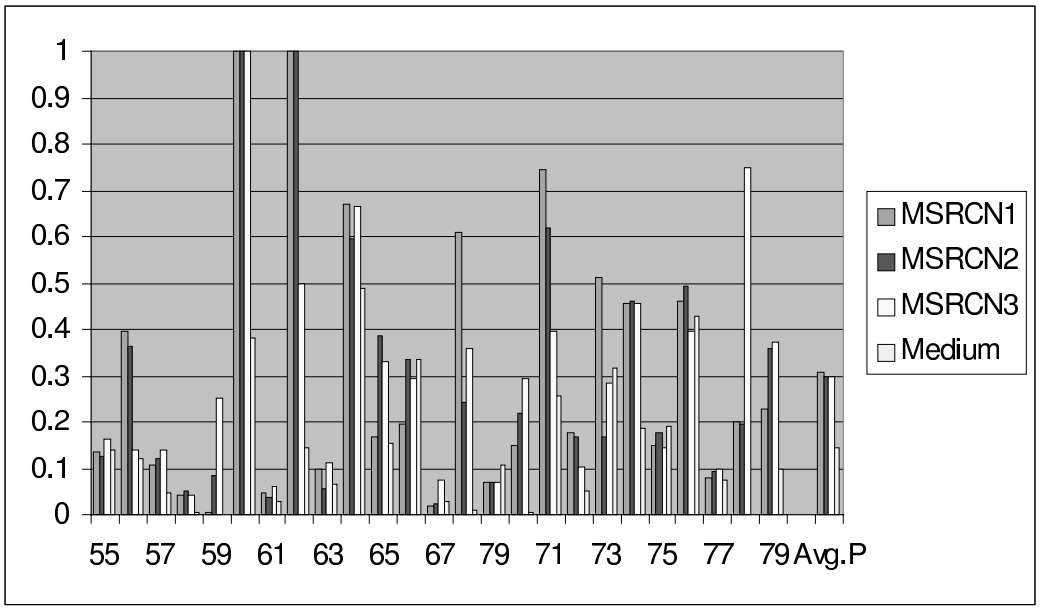


**Fig. 4.** TREC-9 results for 25 queries

| Method | ≥ Medium | < Medium |
|---|---|---|
| MSRCN1 | 20 | 5 |
| MSRCN2 | 19 | 6 |
| MSRCN3 | 20 | 5 |

Table 7. Comparison with medium

MSRCN3 run uses the same combination, but with a pre-translation QE.

MSRCN4 and MSRCN5 use respectively our improved method and the IBM MT system alone. Both pre-translation QE and post-translation QE are used in both cases.

As indicated in table 5, unlike the experimental results on TREC5&6, pre-translation QE does not obtain any improvements. The similar effectiveness of MSRCN4 and MSRCN5 shows again that our approach leads to almost the same effectiveness as the IBM MT system. It is also interesting to find that the best CLIR performance is over 100% of the monolingual. In order to analyze how good our query translation approach for CLIR, we display in Fig. 4 a comparison of the retrieval results for the 25 queries. Another comparison with the medium performance is given in Table 7.

Through our first analysis, the queries may be classified into three categories:

1) 5 queries that have both monolingual and CLIR result of Avg.P lower than 0.1. They are #58, 61, 67, 69, and 77. The bad effectiveness in these cases is not due to translation, but because the query topics are difficult for IR.

2) 11 queries with monolingual Avg.P lower than CLIR. There might be two possible reasons. The first is due to the multiple translations for some key words by combining different translation methods, i.e. our approach and IBM MT software. These multiple translations usually are exchangeable. Multiply translations act as the query expansion. Some examples are: "public key" in query 68# is translated to "公共密钥" as well as "公共密码", "Olympics" in query 70# to "奥林匹克" (Olympic) and "奥运会" (Olympic games), and "Panda bear" in query 76# to "大熊猫" and "大猫熊", etc. The second reason is due to better translations over the original ones. For example, "violation" in query #56 is translated to the more common "侵害" rather than "违反".

3) 9 queries with monolingual Avg.P higher than CLIR. Most of them are due to the bad translations of key concepts. For example, query 65# contains an important term "three-links" (三通), a political abbreviation. This term is not translated correctly. This situation is very similar to some cases observed in TREC5&6, where we encountered the terms such as "most-favor nation" (最惠国), "World Conference on Women" (世妇会), and "Project Hope" (希望工程).

Some domain specific composition phrases, which are not included in the lexicon, such as "stealth technology" (隐秘技术) and "stealth countermeasure" (反隐秘技术) in #59, "computer hacker" (电脑黑客) in #65, "synthetic aperture radar" (合成孔径雷达) in #66, "vehicle fatalities" (车祸) in #68 have special terminology in Chinese and are also not picked up, although every word in each phrase is given the correct sense. Other cases are due to the wrong translations of words, for example, "livestock" in #69 is translated to "牲畜", but the correct translation in this query should be "畜牧业", which is not included in the lexicon.

# 6. Conclusion

In this paper, we described our work in the TREC-9 evaluation in the English-Chinese Cross-Language Information Retrieval (CLIR) track. It involved two aspects: finding good methods for Chinese IR, and finding effective translation means between English and Chinese.

On Chinese monolingual retrieval, we examined the problems such as using different entities as indexes, pseudo-relevance feedback, length normalization, as well as cutting documents done into passages. Each of these techniques gave some improvements to Chinese IR. The best combination of them is used for our Chinese monolingual IR.

On English-Chinese CLIR, our focus was put on finding effective ways for query translation. We have a large English-Chinese bilingual dictionary from LDC. However, beside the problem of completeness of the dictionary, we are also faced with the problem of selecting the best translation word(s) from the dictionary. To address this problem, the following complementary tools have been used: (1) word/term disambiguation using co-occurrence, (2) phrase detecting and translation using language model, and (3) translation coverage enhancement using translation model.

The experimental results we obtained are very encouraging. On Chinese monolingual IR, we obtained 51.50% for TREC5 and 6 Chinese data. This is favorably comparable to the best effectiveness achieved in the previous Chinese TREC experiments.

On English-Chinese CLIR of TREC5 and TREC6, we obtained 75.55% of monolingual effectiveness using our approach. To compare with an MT system, we also tested the IBM MT system, which, when used alone, leads to the same effectiveness (75.40%). When our approach is combined with IBM MT system, we obtained over 85% of monolingual effectiveness. This shows that some translation tools specially designed for query translation may be as suitable as a high-cost MT system, and even if a high-quality MT system is available, our approach can still lead to additional improvements.

# Acknowledgement.

The authors would like to thank Prof. K.L. Kwok for his helpful suggestions, and Aitao Chen for his comments on the paper.

# Reference

[Ballesteros, 1998] L. Ballesteros, and W. B. Croft. Resolving ambiguity for cross-language retrieval. In *Proceedings of the 21st International Conference on Research and Development in Information Retrieval*. Melbourne, Australia, 1998.

[Brown, 1993] Brown, P. F., S. A. Della Pietra, V.J. Della Pietra, and R.L. Mercer. 1993. The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, 19(2): 263-311

[Buckley, 1985] Buckley, C. *Implementation of the SMART information retrieval system*, Technical report, #85-686, Cornell University, 1985.

[Chen 93] Chen, Stanley F.(1993). Aligning sentences in bilingual corpora using lexical information. In *Proceedings of the 31st Annual Conference of the Association for Computational Linguistics, 9-16, Columbus, OH.*

[Gao, 2000] Jianfeng Gao, Han-Feng Wang, Mingjing Li, and Kai-Fu Lee, 2000. A Unified Approach to Statistical Language Modeling for Chinese. In *IEEE, ICASPP2000.*

[Harman, 1996] Harman, D. K. and Voorhees, E. M., Eds. *Information Technology: The Fifth Text REtrieval Conference (TREC-5)*, NIST SP 500-238. Gaithersburg, National Institute of Standards and Technology, 1996.

[Kowk, 1999] K.L. Kowk, English-Chinese cross-language retrieval based on a translation package. In *Proceedings of the 22st International Conference on Research and Development in Information Retrieval.* 1999.

[Kwok, 1997] Kwok, K. L. Comparing representations in Chinese information retrieval. *Conference on*

*Research and Development in Information Retrieval, ACM-SIGIR*, 1997, pp. 34-41.

[Liu, 95] Xin Liu, Ming Zhou, Shenghuo Zhu, Changning Huang (1998), Aligning sentences in parallel corpora using self-extracted lexical information, *Chinese Journal of Computers (in Chinese), 1998, Vol. 21 (Supplement):151-158.*

[Nie, 1999] Nie, J.-Y., Ren, F. Chinese information retrieval: using characters or words? *Information Processing and Management*, 1999, 35: 443-462.

[Nie, 2000] Jian-Yun Nie, Jianfeng Gao, Jian Zhang, and Ming Zhou. On the use of words and n-grams for Chinese information retrieval. In the *Fifth International Workshop on Information Retrieval with Asian Languages, IRAL-2000.* Hong Kong, September 30 to October 1, 2000.

[Salton, 1983] Gerard Salton and M. J. McGill. Introduction to modern information retrieval. McGraw Hill Book Co., New York, 1983.

[Singhal, 1996] Amit Singhal, Chris Buckley and Mandar Mitra. Pivoted document length normalization. In *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, 1996, Pages 21 – 29.

[Zhang, 2000] Jian Zhang, Jianfeng Gao, Ming Zhou. Extraction of Chinese compound words – an experimental study on a very large corpus. The second Chinese Language Processing Workshop, Hong Kong, October 8, 2000.