Improving Query Translation for Cross-Language Information Retrieval using Statistical Models

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

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

[**] Université de Montréal, Email: nie@iro.umontreal.ca

[#] Tsinghua University, China, Email: ajian@s1000e.cs.tsinghua.edu.cn

ABSTRACT

Dictionaries have often been used for query translation in cross-language information retrieval (CLIR). However, we are faced with the problem of translation ambiguity, i.e. multiple translations are stored in a dictionary for a word. In addition, a word-by-word query translation is not precise enough. In this paper, we explore several methods to improve the previous dictionary-based query translation. First, as many as possible, noun phrases are recognized and translated as a whole by using statistical models and phrase translation patterns. Second, the best word translations are selected based on the cohesion of the translation words. Our experimental results on TREC English-Chinese CLIR collection show that these techniques result in significant improvements over the simple dictionary approaches, and achieve even better performance than a high-quality machine translation system.

Keywords

Query translation, CLIR, Statistical model

1. INTRODUCTION

With the explosion of on-line non-English documents, cross-language information retrieval (CLIR) systems have become increasingly important in recent years.

Research in the area of CLIR has focused mainly on methods for query translation. In particular, dictionary-based translation has been a commonly used method because of its simplicity and the increasing availability of machine readable bilingual dictionaries. However, besides the problem of completeness of the dictionary, we are also faced with the problem of ambiguity in translation, i.e. the selection of the correct translation word(s) from the dictionary.

In this paper, we explore several methods to improve query translation for English-Chinese CLIR. First, we try to identify noun phrases (NP) in a query and translate them as units. Phrases usually have fewer senses, thus the translation of a multi-word concept as a phrase is more precise. In addition to the NPs stored in the dictionary, new multi-word NPs are identified automatically using a statistical model. They are translated using translation patterns and a language model. Second, to deal with the translation ambiguity problem, we propose a method based on statistics of co-occurrences. The method tries to select the best translation according to its coherence with the other translation words. Finally, to increase the coverage of the bilingual dictionary, additional words and translations are automatically generated from a parallel bilingual corpus. We tested our methods using TREC Chinese documents. Our results show that each of the methods can bring significant improvement over simple dictionary approaches. A combination of the methods achieves even better retrieval performance than a high-quality machine translation (MT) system.

The remainder of this paper is organized as follows. In Section 2, we provide a brief survey on related work. In Section 3, we describe in detail our techniques for NP identification and translation. In Section 4, we describe the method of translation selection. In Section 5, experimental results are presented. Finally, we present our conclusion in Section 6.

2. DICTIONARY-BASED QUERY TRANSLATION

Bilingual dictionaries have been used in several CLIR experiments. However, previous work showed that English-Chinese CLIR using simple dictionary translation yields a performance lower than 60% of the monolingual performance [14]. The main problems observed are: (1) the dictionary may have a poor coverage; and (2) it is difficult to select the correct translation of a word among all the translations provided by the dictionary.

For the first problem, much effort has been spent on collecting larger lexical resources either manually or automatically [14, 16, 20]. The coverage of the dictionary can be increased to some extent.

A technique often used to deal with the second problem - translation ambiguity - is to identify phrases in the query and translate them as a whole using a phrase dictionary. It has been shown that this technique can improve IR performance. Hull and Grefenstette [13] showed that the performance achieved by manually translating phrases in queries is significantly better than that of a word-by-word translation using a dictionary. Davis and Ogden [7] showed that by using a phrase dictionary

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

SIGIR'01, September 9-12, 2001, New Orleans, Louisiana, USA.

Copyright 2001 ACM 1-58113-331-6/01/0009…$5.00.

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

extracted from parallel sentences in French and English, the performance of CLIR is improved. Ballesteros and Croft [3] performed phrase translation using information on phrase and word usage contained in the Collins machine readable dictionary. They demonstrated that translations of multi-word concepts as phrases are more precise. However, a critical problem remains: if a phrase is not stored in a lexicon, how can one identify it in a query and translate it correctly? It is unrealistic to expect a "complete" phrase dictionary. New phrases are constantly created. Therefore, we will always face the problem of identification and translation of unknown phrases, no matter how complete a phrase dictionary may be. This problem is one of the foci of this paper.

Another possible solution to the problem of translation ambiguity is by using word sense disambiguation. The impact of disambiguation for CLIR is debatable. Xu and Weischedel [19] estimated an upper bound on CLIR performance. They concluded that even if the translation ambiguity were solved correctly, only limited improvement can be obtained. In contrast, Ballesteros and Croft [3] showed that using co-occurrence statistics from corpora can help with reducing translation ambiguity. Other studies including [1, 10, 12] also used similar approaches to select the best translation(s). In this paper, by extending the work in [1, 3], we present a disambiguation method, which can be combined with phrase translation and achieves further improvement.

Our query translation process may be summarized as follows:

- NPs are first identified in English (source language) queries using a statistical method;
- The translation of the identified phrases is determined by both a set of phrase translation patterns and probabilities of the translated phrases obtained from a Chinese language model;
- The remaining words in the query are translated as words. Our problem is to determine the best word translation among all that are stored in the dictionary.

3. PHRASE IDENTIFICATION AND TRANSLATION

Although the translation of multi-word phrases is usually more precise than a word-by-word translation, many significant NPs are not stored in the dictionary. For instance, in TREC-9 queries, more than 50% of noun phrases, which can be detected by our method described in this section, are not in our dictionary.

In the previous IR research, NPs have been identified using a set of syntactic patterns [2, 8]: Sequences of nouns and adjective-noun pairs were taken as phrases. However, this simple method has not produced consistent improvement. Fagan [8] reported a decrease in performance, while Ballestero and Croft [2] did not obtain a significant improvement over single words. One of the problems is that this simple approach often over-generates NPs: non-NPs may be identified as NPs. This may negatively affect the monolingual IR performance (because of a deformed distribution of occurrences of these items). In addition, the identified phrases are still translated word-by-word in [2].

In our approach, we try to translate NPs as units as much as possible. To do so, we first identify English NPs, and translate them as units. If the translations have been used as document indexes (i.e. they are stored in our Chinese dictionary), then the translated NPs can directly match documents. Otherwise, these translations will be segmented into several words which can also match document indexes. So NP identification and translation are means to suggest possible long Chinese NPs. This is a query processing compatible with the longest-matching segmentation method used for document pre-processing.

Unlike previous methods, our approach uses a more sophisticated NP identification process. It is carried out in a bottom-up manner: we first identify base NPs, and then complex NPs. The reason to separate the process into two steps lies in the fact that base NPs can be identified with high accuracy, while the complex NPs cannot be. Therefore, we only use a small set of syntactic patterns in the second step in order to select sufficiently reliable complex NPs.

3.1 Identification of base NPs

3.1.1 Principle

A base NP is a simple noun phrase that does not contain other noun phrases recursively. For example, the elements within [...] in the example shown in Figure 1 are base NPs. The part-of-speech (POS) tags NNS (plural noun), IN (preposition), and VBG (verb-ing) etc. are those defined in [15].

[Measures/NNS] of/IN [manufacturing/VBG activity/NN] fell/VBD more/RBR than/IN [the/DT overall/JJ measures/NNS] ./.

Figure 1: An example sentence with base NP brackets.

The identification of base NPs usually involves two steps: (1) POS tagging, and (2) base NP chunking.

In classical statistical approaches [6, 17], these two steps have been separated. POS tagging often serves as a precursor, and noun phrase chunking uses POS patterns (e.g. DT-JJ-NNS) that are learnt from a tagged corpus.

By separating the two steps, the solution of the first step is used in the second step as if it is certain. The uncertainty involved in the first step is no longer taken into account in the second step. In fact, the correct solution of the first step may be ranked second, third, etc. This is particularly the case when the probabilities of these solutions are close to that of the first solution. Therefore, a too early selection in the first step may be an important source of error.

In our approach, we try to integrate the two steps and their uncertainties together, and use a unified statistical model to choose the globally optimal solution [22]: We keep the N-best (N>1) ranked POS assignments in the first step. Then, in the second step, we determine the best base NPs by considering both the probability of POS tagging and that of base NP pattern. The value of N is chosen empirically to obtain the optimum balance between efficiency and accuracy.

3.1.2 Mathematical formulation

Let us formulate the above two steps in mathematical terms. Given an English sentence $E=\{e_1,\ldots,e_n\}$, its POS tag sequence is denoted by $T=\{t_1,\ldots,t_n\}$. The most probable base NP sequence $B^*=\{b_1,\ldots,b_m\}$ ($m<=i$) is expressed as Equation (1).

$$B^* = \arg\max_B(P(B \mid E)) = \arg\max_B(P(T \mid E) \times P(B \mid T, \quad (1)$$

In this formula: $P(T|E)$ aims to determine the best POS tags for a sentence E, and $P(B|T,E)$ aims to determine the best base NP tag sequences from them. In practice, in order to reduce the search space, only N-best POS tagging of E are retained in the first step (N=4 in our experiments).

To determine the N-best tags from $P(T|E)$, we make use of Bayes' rule as follows:

$$T(N-best) = \arg\max_{T=T_1,\ldots,T_N}(P(T \mid E)) = \arg\max_{T=T_1,\ldots,T_N}(P(E \mid T) \times P(T)) \quad (2)$$

We now assume independence among the relationships between tags and English words; and we use a tag trigram model to approximate $P(T)$. Then $P(E|T)$ and $P(T)$ can be evaluated as follows:

$$P(E \mid T) \approx \prod_{i=1}^{n} P(e_i \mid t_i) \quad (3)$$

$$P(T) \approx \prod_{i=1}^{n} P(t_i \mid t_{i-2}, t_{i-1}) \quad (4)$$

In the Viterbi search algorithm, $P(e_i|t_i)$ is called the output probability of a word given the POS tag, and $P(t_i|t_{i-2},t_{i-1})$ is called the transition probability whose value is determined by a POS trigram model. Both probabilities can be estimated from a POS tagged corpus.

In the second step, we determine the best base NP sequence, given the N-best POS sequences. A similar approach to the first step is used. According to Bayes' rule, we have

$$P(B \mid T, E) = \frac{P(E \mid B,T) \times P(T \mid B) \times P(B)}{P(E \mid T) \times P(T)} \quad (5)$$

For a given E and its T, we have $P(T)P(E|T)=P(E,T)=constant$. By assuming words in E to be independent, we have $P(T \mid B) = \prod_{i=1}^{n} P(t_i, \ldots, t_j \mid b_{i,j}) = 1$. Thus, we have

$$P(B|T, E) \propto P(E|B, T) \times P(B) \quad (6)$$

The two elements on the right side can be estimated as follows

$$P(E \mid B,T) \approx \prod^{n} P(e_i \mid t_i, b_j) \quad (7)$$

$$P(B) \approx \prod_{-}^{m} P(b_i \mid b_{i-2}, b_{i-1}) \quad (8)$$

where b_j is the element in the base NP sequence corresponding to e_i.

Similarly, $P(e_i|t_i,b_j)$ is called the output probability of a word given the POS tag and the base NP tag, and $P(b_i|b_{i-2},b_{i-1})$ is called the transition probability whose value is determined by a base NP trigram model. Again, both probabilities can be estimated from a base NP tagged corpus.

Finally, substituting Equations (2) and (6) in Equation (1), we have

$$B^* = \arg\max_{B,T=T_1,\ldots,T_N}(P(T \mid E) \times P(T) \times P(E \mid B,T) \times P(B)) \quad (9)$$

In summary, for a given input English sentence E, in the first step, the Viterbi N-best searching algorithm is applied for POS tagging. Every resulting T is assigned a probability P_t by Equation (2). In the second step, for each T, the Viterbi algorithm is applied again to search for the best base NP sequence. Every resulting B^* is assigned another probability P_b by Equation (6). The final integrated probability of a base NP sequence is determined by $P_t^{\alpha}P_b$, where α is a normalization coefficient ($\alpha = 2.4$ in our experiments).

3.1.3 Model estimation and evaluation

The models used in this study are trained on Penn Treebank [15]. We used the section 20 as test data, while the other 24 sections are used as training data.

At first, all possible base NP patterns are extracted from the tagged training corpus. There are more than 6000 patterns in the Penn Treebank. After being filtered by linguistic rules described in [21], 1169 patterns are kept. Then, all parameters in our statistical model are estimated using MLE (maximum likelihood estimation) from the training corpus. These parameters are: (1)$P(t_i|t_{i-2},t_{i-1})$, (2)$P(e_i|t_i)$, (3)$P(b_i|b_{i-2},bi_{-1})$, and (4)$P(e_i|t_i,b_j)$. A detailed description can be found in [22].

Our tests showed that our integrated approach achieves 92.3% in precision and 93.2% in recall. The result is slightly better than the current state of the art.

3.2 Identification of complex NPs

Unlike base NP, there is not a widely accepted definition of complex NP. It is even worse in Chinese. In addition, a lot of complex NPs in English cannot be translated into Chinese as a unit. Therefore, with the help of a linguist, we selected 40 frequently used English NP patterns, which can be translated into Chinese as a unit. Some examples are shown in Figure 2. Any sequence of words or base NPs corresponding to one of the patterns is identified as a complex NP.

Complex NP patterns	**Examples**
Base NP *of* Base NP	[the sales] of [Chinese ships]
Base NP *in* Base NP	[human rights violations] in [China]
Base NP *and* Base NP	[China 's Panda bear population] and [research organizations]

Figure 2: Examples of complex NP patterns from TREC-9 CLIR queries

3.3 NP translation

3.3.1 Principle

The dictionary contains a certain number of NPs and their translations. However, many more NPs that we identified are not stored in the dictionary. Then, how can we translate these NPs better than a word-by-word method?

We observed that there are some translation patterns between English NPs and Chinese NPs. For example, a [NN-1 NN-2] phrase is usually translated into a [NN-1 NN-2] sequence in Chinese, and a [NN-1 *of* NN-2] phrase is usually translated into a [NN-2 NN-1] sequence in Chinese. So for an English phrase corresponding to such a pattern, if its translation is not stored in the dictionary, we can still generate its possible translation. For instance, we can derive the translation of the multi-word phrase "drug sale" as 毒品(drug)/ 买卖 (sale), and the translation of "security committee of UN" as 联合国(UN)/安理会(security committee).

Possible translation patterns can be extracted from a word-aligned bilingual corpus. We first used the NP identification method described above to tag POS, base NP, and complex NP for English sentences. Then, for each English NP pattern (*EPT*), we extracted its Chinese translation patterns (*CPT*). An example is shown in Figure 3. For an English sentence, each word is marked its POS tag and position. Elements within […] are base NPs, or complex NPs. The aligned Chinese sentence is segmented into a sequence of words, which are associated with their positions. A word alignment (*x*, *y*) indicates that an English word in position *x* is connected to a Chinese word in position *y*. When an English word is connected to no Chinese words, it is denoted by (*x*, **e**). Some examples of translation pattern we can extract are also illustrated in Figure 3. The estimation of the probability of translation patterns will be described in section 3.3.3.

English Sentence	[[The/DT/1 natural/JJ/2 language/NN/3 computing/NNP/4 group/NNP/5] at/IN/6 [Microsoft/NNP/7 Research/NNP/8 China/NNP9]] …
Chinese Sentence	微软/1 中国/2 研究院/3 自然/4 语言/5 计算/6 组/7 …
Aligned word-pair	(1,**e**) (2,4) (3,5) (4,6) (5,7) (6,**e**) (7,1) (8,3) (9,2) …
Translation Patterns	[*DT JJ NN NNP-1 NNP-2*] →[*JJ NN NNP-1 NNP-2*], *P*=0.48 [*NNP-1 NNP-2 NNP-3*] → [*NNP-1 NNP-3 NNP-2*], *P*=0.26 [*Base NP-1 at Base NP-2*] → [*Base NP-2 Base NP-1*], *P*=0.67

Figure 3: English NP patterns and their Chinese translation patterns

As we mentioned earlier, the obtained NP translations do not always correspond to document indexes. If they do not, the segmentation process will break them down into several words. Even in this case, we can still benefit from the word selection in this process that solves part of the translation ambiguity problem.

3.3.2 Mathematical formulation

Given an English NP, $ENP=\{e_1,\ldots,e_n\}$, with its NP pattern, *EPT*; for each English term e_i in *ENP*, we retrieve all the possible Chinese translations from the bilingual dictionary. We also get all the possible translation patterns *CPT* for *EPT*. Then the best Chinese translated phrase, $CNP^*=\{c_1,\ldots,c_m\}$, is the one that maximizes the Equation (10) below.

$$\begin{aligned} CNP^* &= \arg\max_{CNP} P(CNP \mid ENP) \\ &= \arg\max_{CNP} P(ENP \mid CNP) \times P(CNP) \end{aligned} \tag{10}$$

where $P(ENP|CNP)$ is the translation probability. $P(CNP)$ is a priori probability of words of the translated Chinese NP.

We consider an NP (*ENP* or *CNP*) as a set of words (*E* or *C*) assembled by an NP pattern (*EPT* or *CPT*). Assuming that the translation of words and NP patterns are independent, we have

$$\begin{aligned} P(ENP \mid CNP) &= P(E, EPT \mid C, CPT) \\ &= P(E \mid C, CNP) \times P(EPT \mid C, CNP) \\ &= P(E \mid C) \times P(EPT \mid CNP) \end{aligned} \tag{11}$$

Substituting Equation (11) in Equation (10), we have

$$CNP^* = \arg\max_{CNP} P(E \mid C) \times P(EPT \mid CPT) \times P(CNP) \tag{12}$$

where $P(E|C)$ is the translation probability from Chinese words *C* in *CNP* to English words *E* in *ENP*. $P(EPT|CPT)$ is the probability of the translation pattern *EPT* (i.e. the order of translation words), given the Chinese pattern *CNP*.

These probabilities are estimated as follows:

$$P(EPT \mid CPT) = \frac{C(EPT, CPT)}{C(CPT)} \tag{13}$$

where $C(CPT)$ represents the number of occurrences of *CPT* in the Chinese portion of the aligned bilingual corpus, and $C(EPT,CPT)$ represents the number of times *EPT* corresponds to *CPT* in the aligned sentences.

$P(CNP)$ is determined by the Chinese trigram language model as follows:

$$P(CNP) = P(c_1 \ldots c_n) = \prod_{i=1}^{n} P(c_i \mid c_{i-2}, c_{i-1}) \tag{14}$$

3.3.3 Model estimation

As described in section 3.3.2, for NP translation, there are three probabilities to be estimated: (1) $P(EPT|CPT)$, (2) $P(c_i|c_{i-2},c_{i-1})$, and (3) $P(E|C)$.

$P(EPT|CPT)$ is estimated from a bilingual corpus. In our cases, we used a word-aligned bilingual corpus containing approximately 100,000 English-Chinese sentence pairs. Translation patterns are first extracted automatically from the corpus, and then filtered by a linguist. The probability is then estimated according to Equation (13). For each Chinese NP pattern, there are 4.53 translation patterns on average.

The Chinese trigram language model is trained on a Chinese corpus consisting of approximately 1.6 billion Chinese characters. It contains documents of different domains, style, and period of time [9].

For $P(E|C)$, we simply assumed a uniform distribution on a word's translation in our experiments. If a Chinese word c has n translations in our bilingual dictionary , each of them will be assigned equal probability, i.e. $P(e|c)=1/n$. There are two reasons for this. First, there are no translation probabilities in our original bilingual dictionary. Second, we do not have enough parallel corpora for its accurate estimation.

At this stage, it is interesting to compare our translation method to the methods proposed in [2, 3, 4]. Ballesteros and Croft used a word-by-word strategy for phrase translation [2, 3]. It is based on two assumptions that are sub-optimal. First, they assume that there is a one-one mapping between words in English NP and words in Chinese N. However, in our experiments, we found that only 56% NP translation patterns have such one-one mappings. Second, they assume that the translation words in a phrase will remain in the same order as in the source language phrase. In our experiments, we found that 35% of translation patterns change word order. On the other hand, The IBM statistical models incorporate very little linguistic knowledge [4]. It is hard to capture non-local dependencies of the language with "local" models such as n-gram models. So even if the translation model generates the correct set of words, the language model will not assemble them correctly. In our method, we incorporate the language model with translation patterns. While the language model captures the "local" dependency, the translation patterns provide information on global dependency within a phrase. Although the method is not powerful enough for sentence-level translation, it performs well for NP translation.

4. THE SELECTION OF WORD TRANSLATION

Words that are not included in phrases are translated word-by-word. However, this does not mean that they should be translated in isolation from each other. Instead, while translating a word, the other words (or their translations) form a "context" that helps determine the correct translation for the given word. This is the principle of our translation selection process. Our assumption is that the correct translations of query words tend to co-occur in target language documents and incorrect translations do not. Therefore, given a set of original English query words, we select for each of them the best translation word 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 greedy 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 dictionary. 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. The EMMI weighting measure [18] has been successfully used to estimate the term similarity in [1, 3]. We take a similar approach. However, we also observe that EMMI does not take into account the distance between words. In reality, we observe that local context is more important for translation selection. If two words appear in the same document but at two distant places, it is unlikely that they are strongly dependent. Therefore, we add a distance factor in our calculation of word similarity. Formally, the similarity between terms x and y is

$$SIM(x,y) = p(x,y)\times\log_2\left(\frac{p(x,y)}{p(x)\times p(y)}\right) - K\times\log_2 Dis(x,y) \quad (14)$$

where

$$p(x,y) = \frac{c(x,y)}{c(x)} + \frac{c(x,y)}{c(y)} \quad (15)$$

$$p(x) = \frac{c(x)}{\sum_x c(x)} \quad (16)$$

$c(x,y)$ 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 terms x and y in a sentence, and K is a constant coefficient, which is chosen empirically. (K=0.8 in our experiments).

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

$$Cohesion\ (x, X) = Max_{\ y\in X}\ SIM\ (x, y) \quad (17)$$

Figure 4 depicts a left-to-right greedy algorithm for the selection of the best translation. Term-similarity is estimated using the same 1.6 billion character Chinese corpus mentioned earlier.

For each source query word s_i (i = 1 to n), retrieve a set of translations 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 & $k\neq i$), compute the cohesion $Cohesion(t_{ij}, T_k)$;

    Compute the score of t_{ij} as the sum of $Cohesion(t_{ij}, T_k)$ (k = 1 to n & $k\neq I$);

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

Figure 4. Greedy algorithm to find the best translations

5. EXPERIMENTAL RESULTS AND DISCUSSION

In this section, we present the results of our CLIR experiments on TREC Chinese corpora. The TREC-9 corpus contains articles published in Hong Kong Commercial Daily, Hong Kong Daily News, and Takungpao. They amount to 260MB. A set of 25 English queries has been set up and evaluated by people at NIST (National Institute of Standards and Technology). The TREC 5&6 corpus contains 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 at NIST.

Each of the TREC queries has three fields: title, description, and narratives. In our experiments, we used two versions of queries, short (only titles) and long (all the three fields).

The bilingual lexical resources we used include three human compiled bilingual lexicons and a bilingual lexicon generated from a parallel bilingual corpus automatically [16]. The resulting combined dictionary contains 401,477 English entries, including 109,841 words, and 291,636 phrases.

For our experiments, we used a slightly modified version of the SMART system [5]. We used the *ltc* weighting scheme. The main evaluation metric is interpolated 11-point average precision. Statistical t-test [11] and query-by-query analysis are also employed. To decide whether the improvement by method *X* over method *Y* is significant, the t-test calculates a p-value based on the performance data of *X* and *Y*. The smaller the p-value, the more significant is the improvement. Usually, if the p-value is small enough (p-value<0.05), we can conclude that the improvement is statistically significant.

We first carried out a set of preliminary experiments to investigate the impact of lexicon sources, phrase, and ambiguity on query translation. Our results confirmed our intuition. Results showed that larger lexicon sources, phrase translation, and disambiguation techniques improve CLIR performance significantly and consistently on TREC-9 corpus.

5.1 Impact of NP translation and translation selection

The following methods are compared to figure out the impact of NP translation and translation selection:

1. *Monolingual*: retrieval using the manually translated Chinese queries provided with the corpus.
2. *Simple translation*: retrieval using query translation obtained by look up the bilingual dictionary.
3. *Best-sense translation*: retrieval using query translation selected manually.
4. *Machine translation*: retrieval using translation queries obtained by a machine translation system.
5. Our methods that incorporate NP detection and translation, as well as word translation selection.

For *simple translation*, phrase entries in the dictionary are first used for phrase matching and translation, and then the remaining words are translated by their translations stored in the dictionary.

For *best-sense translation*, we manually disambiguated the queries in order to get an upper bound of performance using dictionary look-up and disambiguation. In this method, a native Chinese speaker selected one translation from the dictionary for each English word or phrase. If no translation is correct, the first one is randomly chosen.

For *machine translation*, a commercial English-Chinese machine translation system - IBM HomePage Dictionary™ 2000 - is used. This system was released recently by IBM. It contains an English-Chinese dictionary with 480,000 entries, including words, frequently used phrases (such as "information retrieval"), acronyms (such as "IBM"), and proper nouns (such as "Microsoft"). According to our survey, this system is one of the best machine translation products currently on the market. The result of query translation by the IBM system seems reasonable; less than 3% of the words are left untranslated, most phrases are translated as a whole, and the translation ambiguity problem is solved to some degree for most of the words.

The average precision of this series of experiments on query translation is summarized in Table 1. The precision-recall (P-R) curves using short and long queries are shown in Figures 5 and 6. It is interesting to compare the results of our NP translation method (in row 4 in Table 1) with that of phrase translation using dictionary look-up (in row 2). It turns out that by using NP identification and translation, we obtained better performance. For example, in short query retrieval, in 25 queries, only 11 multi-word phrases are stored in the dictionary, and translated as a phrase, while using our method, 26 NPs are identified and translated. It thus results in a 102.6% improvement, which is statistically significant (p-value = 0.015).

	Translation method	Short queries		Long queries	
		Avg. P.	% Mono. IR	Avg. P.	% Mono. IR
1	Monolingual	0.2684		0.3099	
2	Simple translation	0.1174	43.74%	0.1823	58.83%
3	Best-sense translation	0.1611	60.02%	0.2618	84.48%
4	2 + NP translation	0.2379	88.64%	0.2398	77.38%
5	4 + Translation selection	0.2468	91.95%	0.2956	95.38%
6	Machine translation	0.1303	48.55%	0.2466	79.57%
7	5 + 6	0.2395	89.23%	0.3280	105.84%

Table 1: Average retrieval precision results, using TREC-9 short queries and long queries

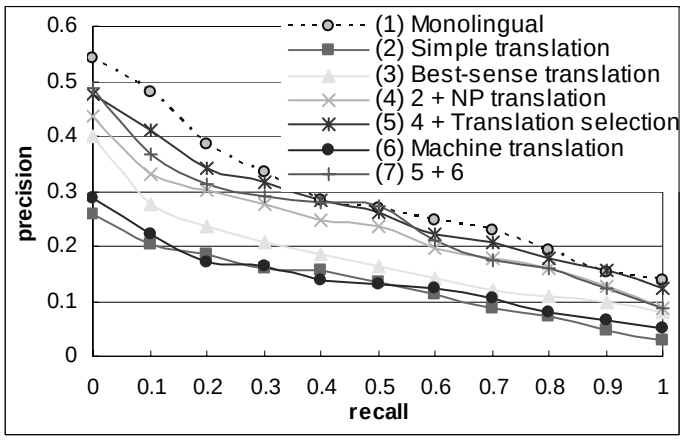

Figure 5: P-R curves, using TREC-9 short queries

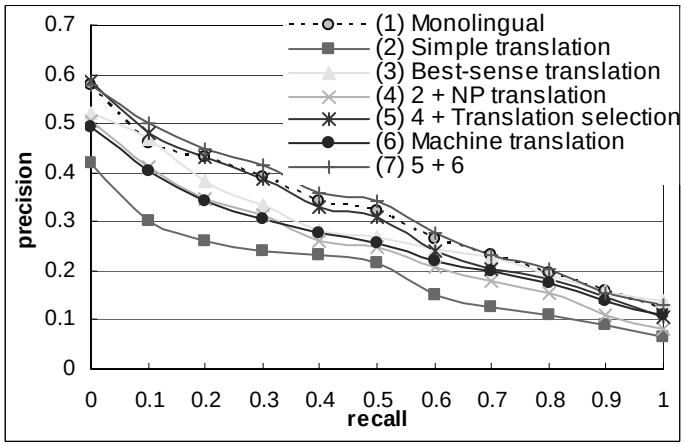

Figure 6: P-R curves, using TREC-9 long queries

Row 5 shows that further improvement can be obtained using our translation selection technique. We obtain 23.3% improvement for long queries and 3.7% improvement for short queries. It is not surprising the improvement in the first case is statistically significant (p-value = 0.05), but the improvement in the second case is not. The reason is that long queries provide much richer context information for translation disambiguation/selection.

Our results in row 5 are even better than *best-sense* translation results in row 3 (although it is not statistically significant), which are regarded as the upper bound on performance of any disambiguation techniques. This shows again the positive impact of our NP identification and translation methods.

Row 6 shows that the using the MT system, we can achieve 79.57% of monolingual effectiveness for long queries. This performance is comparable to those reported by others using an MT system. We can see that our method outperforms the MT system. This confirms that there may be better ways for query translation than MT systems.

The best performance is achieved by combining linearly two sets of translation queries obtained by machine translation method and our method. While using long query retrieval, it is over 105% of monolingual effectiveness. The intuition of combination of different translation methods is that different translation systems would complement each other. This result confirms the intuition. It also shows that monolingual performance is not necessarily the upper bound of CLIR performance. An important reason for this is that there is an implicit query expansion effect during translation because related words/phrases may be added.

In summary, the improvement by using NP translation for short queries is statistically significant (p-value = 0.015). The addition of translation selection is also statistically significant for long queries (p-value = 0.05). The improvement obtained with the combination of both approaches (i.e. NP translation and translation selection) are statistically significant for both short queries (p-value = 0.03) and long queries (p-value = 0.001). The comparison with the MT approach shows that at least for short queries, the improvement brought by our methods is statistically significant (p-value = 0.02).

Due to the limited number of the TREC-9 queries, we also tested our methods on TREC 5&6 Chinese collection. The results are similar, as shown in Table 2. This further confirms our conclusions made above.

5.2 Analysis

In order to analyze the effectiveness and remaining problems of our query translation approach, for the 25 queries of TREC 9, we display in Figure 7 a comparison of the long query retrieval results of row 1 and row 5 in Table 1. We observe that the queries may be classified into three categories:

1) 5 queries that have both monolingual and CLIR result of average precision lower than 0.1 (#58, #61, #67, #69, and #77). The bad effectiveness in these cases is not due to translation, but to the high difficulty of query topics.

No.	Translation Method	Avg.P.	% Mono. IR
1	Monolingual	0.5150	
2	Simple translation	0.2722	52.85%
3	Best-sense translation	0.3762	73.05%
4	2 + NP translation + translation selection	0.3883	75.55%
5	Machine translation	0.3891	75.40%
6	4 + 5	0.4400	85.44%

Table 2: Average retrieval precision results, using TREC-5&6 long queries

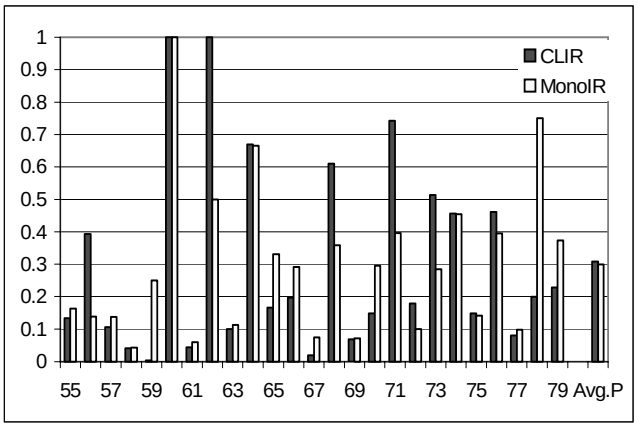

Figure 7: TREC-9 results for 25 queries: monolingual IR vs. CLIR using long queries

2) 11 queries with monolingual average precision lower than CLIR. There might be two possible reasons. One is that by accepting several translations for a key concept, we are in fact making 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 #71 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 that sometimes, the translations obtained are more natural expressions than those given in the original Chinese queries. For example, "violation" in query #56 is translated to a more common word 侵害 rather than 违反 in the manually translated Chinese query (provided with the corpus).

3) 9 queries with monolingual average precision higher than CLIR. Most of them are due to the bad translations of key concepts, which are not stored in the dictionary as a phrase. We divide NP into two types: compositional NP and non-compositional NP.

Compositional NP is the phrase whose translation can be assembled by translations of words within the phrase, such as "computer hacker" (电脑黑客), "public key" (公共密钥), and "environmental protection laws" (环境保护法), etc. Generally speaking, our method is good for compositional NP translation. For some domain-specific NPs, it failed. For example, "stealth technology" (隐秘技术) and "stealth countermeasure" (反隐秘技术) in #59, and "synthetic aperture radar" (合成孔径雷达) in #66 have special terminology in Chinese and are not translated correctly.

A non-compositional NP is a phrase whose translation cannot be assembled by translations of its component words. Our method is unable to deal with the translation of non-compositional NPs. Examples include "three-links" (三 通) in #65, "vehicle fatalities" (车祸) in #68, "most-favored nation" (最惠国), and "World Conference on Women" (世妇会), etc. A large portion of non-compositional NPs in queries are political abbreviations. If these NPs are not stored in the dictionary, they are most likely to be translated incorrectly. This indicates that the coverage of the dictionary is still an important problem to be solved to improve the performance of CLIR.

6. CONCLUSION

Dictionary-based query translation has been widely used in CLIR because of its simplicity and the increasing availability of machine-readable bilingual lexicons. However, besides the problem of completeness of the lexicon, we are also faced with the problem of selecting the best translation word(s) from the dictionary.

In this paper, we proposed several approaches to improve dictionary-based query translation for CLIR. We focused on translation of phrases, which has been demonstrated to be one of most effective ways to obtain more accurate translations. We presented a method to identify and translate unknown NPs. English NPs in queries are first identified statistically, and then translated into Chinese phrases using a new method that combines translation patterns and a Chinese language model. We also presented a method of translation selection based on the cohesion among translation words.

Through our experiments, we showed that each of the above methods leads to some improvement, and that the combined approach significantly improves CLIR performance. The fact that our approach outperformed one of the best commercial MT systems indicates that some specific translation tools designed for query translation in CLIR may be better than on-the-shelf MT systems. The combination of our approach with the MT system leads to a high effectiveness of 105% of that of monolingual IR. This shows that even if a high-quality MT system is available, our approach can still lead to additional improvement.

Though our method shows very promising improvement in experiments, we are faced with some unsolved problems. First, the lack of large amount of word/phrase-aligned parallel corpus prevents us from extracting more reliable translation patterns. Second, to translate queries in a specific domain, it would be better to use a domain-specific translation and language model. This could help with selecting the correct domain-specific translations. Then there come the problems of building domain-specific translation/language models, and activating the corresponding model when a specialized query is submitted. These are some of the topics of our future work.

ACKNOWLEDGEMENTS

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

REFERENCES

[1] Adriani, M. (2000). Using statistical term similarity for sense disambiguation in cross-language information retrieval. *Information Retrieval*. 2, 69-80.

[2] Ballesteros, L., and Croft, W. B. (1997). Phrasal translation and query expansion techniques for cross-language information retrieval. In: *Proceedings of the 20th International Conference on Research and Development in Information Retrieval.* 84-91.

[3] Ballesteros, L., and Croft, W. B. (1998). Resolving ambiguity for cross-language retrieval. In *Proceedings of the 21st International Conference on Research and*

Development in Information Retrieval. Melbourne, Australia.

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

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

[6] Church, K., (1988) *A stochastic parts program and noun phrase parser for unrestricted text.* In Proceedings of the Second Conference on Applied Natural Language Processing, pages 136-143. Association of Computational Linguistics.

[7] Davis, M. W., and Ogden, W. C. (1997). Free resources and advanced alignment for cross-language text retrieval. In: *The Sixth Text Retrieval Conference (TREC-6).* NIST, Gaithersbury, MD.

[8] Fagan, J. (1988). Experiments in Automatic Phrase Indexing for Document Retrieval: A Comparison of Syntactic and Non-Syntactic methods. Computer Science, Cornell University.

[9] Gao, J., Wang, H. F., Li, M. and Lee, K. F. (2000). A unified approach to statistical language modeling for Chinese. *ICASSP-2000*, Istanbul, Turkey, June.

[10] Hiemstra, D., and Jong, F. (1999). Disambiguation strategies for cross-language information retrieval. In: *Proceedings of the third European Conference on Research and Advanced Technology for Digital Libraries.* 274-293.

[11] Hull, D. (1993). Using statistical testing in the evaluation of retrieval experiments. In: *Conference on Research and Development in Information Retrieval, ACM-SIGIR*, 1993.

[12] Hull, D. A. (1997). Using structured queries for disambiguation in cross-language information retrieval. In: *AAAI Symposium on Cross-Language Text and Speech Retrieval*.

[13] Hull, D. A., and Grefenstette, G. (1996). Querying across languages: a dictionary-based approach to multilingual information retrieval. *Research and Development in Information Retrieval*, pp46-57.

[14] Kowk, K. L. (2000). Exploiting a Chinese-English bilingual wordlist for English-Chinese cross language information retrieval. In: *Fifth International Workshop on Information Retrieval with Asian Languages, IRAL-2000.* Hong Kong, September 30 to October 1, 2000.

[15] Marcus, M., Marcinkiewicx, M., and Santorini, B. (1993). Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics*, 19(2): 313-330.

[16] Nie, J. Y., Simard, M., Isabelle, P., and Durand, R. (1999). Cross-language information retrieval based on parallel texts and automatic mining of parallel texts from the web. In: *Conference on Research and Development in Information Retrieval, ACM-SIGIR*, 1999.

[17] Ramshaw, L. A., and Marcus, M., (1998). *Text chunking using transformation-based learning.* In Natural Language Processing Using Very large Corpora. Kluwer. Originally appeared in The second workshop on very large corpora WVLC'95, pp.82-94.

[18] van Rijsbergen, D. J. (1979). Information retrieval, 2nd ed. Butterworths, London.

[19] Xu, J., and Weischedel, R. (2000a). Cross-lingual information retrieval using Hidden Markov models. In: Proceeding of the 2000 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora. Hong Kong, October 7-8, 2000.

[20] Xu, J., and Weischedel, R. (2000b). TREC-9 cross-language retrieval at BBN. In: *The Ninth Text Retrieval Conference (TREC-9).* NIST, Gaithersbury, MD.

[21] Xun, E. (1999). Incremental English parsing using combination of statistic and learning methods. Ph.D. thesis. Harbin Institute of Technology, China.

[22] Xun, E., Zhou, M., and Huang, C. (2000). A unified statistical model for the identification of English base NP. In: *The 38th Annual Meeting of the Association for Computational Linguistics*, Hong Kong 3-6 October.