Improving Query Translation for CLIR through Statistical Models

Jianfeng Gao[*], Jian-Yun Nie[**], Endong Xun[*],
Jian Zhang[#], 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. Dictionaries have often been used for query translation in 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 present a method that improves the previous dictionary-based query translation in two ways. First, as much as possible, phrases are recognized and translated. For this, we make use of statistical models for the detection of unknown phrases and translation patterns for their translation. Second, the best word translations are selected based on the cohesion of the translation words. Our experimental results on TREC-9 CLIR corpus 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.

1. Introduction

With the explosion of on-line documents in languages other than English, cross-language information retrieval (CLIR) systems 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 resolving ambiguity associated with translation. That is, how to select the best translation word(s) from the dictionary.

In this paper, we explore methods to improve query translation for English-Chinese CLIR. To deal with the ambiguity problem, we first focus on methods for translation of noun phrase (NP). Since phrases usually have fewer senses, the translation of multi-word concept as a phrase is more precise. In our approach, multi-word concepts that are not stored in the dictionary as NPs are identified automatically using a statistical model. We then develop a new statistical translation method to translate NPs. This method uses translation patterns and a language model. Second, we present a method based on co-occurrence statistics from corpora which can be used to further reduce the ambiguity associated with phrasal and term translation. To deal with the problem of dictionary completeness, we combine lexical sources including human compiled bilingual lexicons and a bilingual lexicon generated from parallel bilingual corpus automatically. We tested our methods using TREC-9 Chinese documents. Our results show that this method brings significant improvements over simple dictionary approaches, and achieve even better retrieval performance than a high-quality machine translation (MT) system.

In Section 2, we provide a brief survey on related work. In Section 3, we describe in detail our techniques for phrase identification and translation. In Section 4, we describe the method of term

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

disambiguation. In Section 5, we describe our work on lexical sources for dictionary coverage enhancement. In Section 6, experimental results are presented. Finally, we present our conclusion in Section 7.

2. Dictionary-based Query Translation

In CLIR, either documents or queries have to be translated. Methods for query translation, proposed recently, fall into three categories: (1) using MT systems, (2) using parallel corpora, and (3) using bilingual lexicons. The dictionary-based translation has been used in many approaches and experiments due to its simplicity and the availability of many online bilingual dictionaries.

However, previous work showed that English-Chinese CLIR using simple dictionary translation via machine-readable dictionary (MRD) yields a performance lower than 60% of the monolingual performance (Kowk, 2000; Gao et al., 2000a). The main problems observed on this simple method are: (1) the dictionary used may be incomplete; and (2) it is difficult to identify the correct translation of a word from the dictionary as several translations are usually given for a word in a dictionary.

For the first problem, a lot of efforts have been spent for collecting larger lexical resources either manually or automatically (Xu and Weischedel, 2000b; Nie et al., 1999; Kwok, 2000). These efforts contributed in increasing word coverage during translation. It is found that most words in TREC queries are covered by a reasonably large lexicon.

A technique to remedy the second problem - translation ambiguity - is to identify phrases in the query and translate them as a phrase using a phrase dictionary. It is shown that this technique has improves the performance of CLIR. (Hull and Grefenstette, 1996) showed that the performance achieved from manually translating phrases in queries is significantly better than that of word-by-word translation using a dictionary. (Davis and Ogden, 1997) showed that using a phrase dictionary built by extracting phrases from parallel sentences in French and English, the performance of CLIR is improved. (Ballesteros and Croft, 1998) performed phrase translation using information on phrases and words usage contained in the Collins MRD, and demonstrated that translation of multi-word concepts as phrases is 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 think about constructing a "complete" phrase dictionary. New phrases are constantly created. Therefore, we will always face with the problem of identification 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 (ore translation selection) is using word sense disambiguation. The value of disambiguation for CLIR is debatable. (Xu and Weischedel, 2000a) estimated an upper bound on CLIR performance, assuming that translation ambiguity were a solved problem, and concluded that it only led to very limited improvements. On contrast, (Ballesteros and Croft, 1998) presented a technique based on co-occurrence statistics from corpora. It helped reducing translation ambiguity. Other studies including (Hiemstra and Jong, 1999; Hull, 1997; Adrian, 2000) also used similar approaches to select the best translation(s). In this paper, by extending the work in (Ballesteros and Croft, 1998; Adrian, 2000), we present a disambiguation method, which can be combined with phrase translation and achieves further improvements.

Our query translation process may be summarized as follows:

- Noun phrases (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 their probabilities;
- Finally, 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

It has been demonstrated that the translation of multi-word phrases is usually more precise than a word-by-word translation. Unfortunately, many significant noun phrases (NPs) are not stored in the lexicon. For instance, more than 50% of phrases in TREC-9 CLIR queries are not in our dictionary.

In the previous IR research, NPs have been identified using a set of syntactic patterns (Fagin, 1988; Ballesteros and Croft, 1997): Sequences of nouns and adjective-noun pairs were taken as phrases. However, this simple method is more or less successful. Fagin reported a decrease in performance, while Ballesteros and Croft did not show a significant improvement over single words. One of the problems is that this simple approach often over-generates NPs: non-NPs are identified as NPs. This may negatively affect the monolingual IR performance (because of deformed distribution of occurrences of these items). In our approach, we will use a more sophisticated NP identification process. We first identify base NPs, i.e. the basic NPs that do not contain smaller NPs. The identification of base NPs is relatively accurate. Then some more complex NPs are identified through the use syntactic pattern and statistical models.

3.1 NP identification

Our NP identification process is carried out in bottom-up manner: we first identify base NPs, then more complex NP. A base NP is a simple NP, which does not contain other NP. Figure 1 gives some examples of base NP.

Base NP pattern	**Example base NP**
NNS	Measures
VBG NN	Manufacturing activity
DT JJ NNS	The overall measures

Figure 1: Examples of base NP.

The part-of-speech (POS) tags NNS (plural noun), IN (preposition), and VBG (verb-ing) etc. are those defined in (Marcus et al., 1993).

Identification of base NP

The identification of base NPs usually involves two steps: (1) the selection of N-best POS tagging given a sentence, and (2) the identification of base NPs given the N-best POS-sequences. These two steps are separated in classical approaches (Church, 1988). However, there is a strong relationship between them. Therefore, we integrate them into a unified statistical model.

The principle of our approach is as follows. 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 $B^*=\{b_1,\ldots,b_j\}$ ($j<=i$) is expressed as Equation (1).

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

One can see that the two steps are integrated: $P(T|E)$ determines the best POS tags for a sentence E, and $P(B|T,E)$ determines 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 considered in the first step (i.e. N=4 in our experiments).

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

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

where $P(E|T)$ and $P(T)$ may be approximated as follows:

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

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

In Viterbi algorithm of N-best search, $P(w|t)$ is called output probability of a word give the POS tag, and $P(t_i|t_{i-2},t_{i-1})$ is called transition probability whose value is given by the POS trigram.

In the second step, we seek the best B, given the N-best POS sequence. According to Bayes' rules, we have

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

For a given E and its T, we have $P(T)P(E|T)=P(E,T)=const$. By assuming words in E to be independent, we have $P(T|B)=1$. Thus, for a given T of a given E, according to Equation (1), the best B is estimated by

$$B^* = \arg\max_{B} (P(E \mid B, T) \times P(B)) \tag{6}$$

where

$$P(E \mid B, T) \approx \prod_{i=1}^{n} P(w_i \mid t_i, b_j) \tag{7}$$

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

Similarly, $P(w|t,b)$ is called output probability of a word give the POS tag and the base NP tag, and $P(b_i|b_{i-2},bi_{-1})$ is called transition probability whose value is given by the base NP trigram. Finally, substituting Equation (2) and Equation (6) in Equation (1), we have

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

In summary, for a given E, in the first step, Viterbi N-best searching algorithm is applied for POS tagging. Every resulting T is assigned with a probability P_t by Equation (2). In the second step, for each T, Viterbi algorithm is applied again to search for the best base NP sequence. Every resulting B^* is assigned with 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 (α = 2.4 in our experiments). When we determine the best base NP sequence for the given sentence E, we also determine the best POS sequence of E, which corresponds to the best base NP of E.

In our experiments, the training and test data are derived from the 25 sections of Penn Treebank (Marcus et al., 1993). At first, all possible base NP patterns are extracted from the tagged training corpus. Some examples are shown in Figure 2. There are more than 6000 patterns in the Penn Treebank. After being filtered by linguistic rules described in (Xun, 1999), 1169 patterns are used. Then, all parameters in our statistical model are estimated using MLE (maximum likelihood estimation) from the training corpus. They are: (1)$P(t_i|t_{i-2},t_{i-1})$, (2)$P(w|t)$, (3)$P(b_i|b_{i-2},bi_{-1})$, and (4)$P(w|t,b)$. A detailed description can be found in (Xun et al., 2000). We compared our approach with other three approaches proposed in (Cardie, 1998; Ramshaw and Marcus, 1998; Argamon et al., 1998). The test data we used is available at http://www.cs.biu.ac.il/~yuvalk/MBSL. The training data is the 24 sections of Penn Treebank (excluding the section 20). The results are shown in Table 1. It turns out that our method achieves better performance then others.

	(Ramshaw and Marcus, 1998)	(Cardie, 1998)	(Argamon et al., 1998)	Our method
Precision (%)	91.8	89.0	91.6	92.3
Recall (%)	92.3	90.9	91.6	93.2
F-Measure (%)	92.0	89.9	91.6	92.7
$\frac{P+R}{2}$	92.1	90.0	91.6	92.8

Table 1: The comparison of our statistical method with three other approaches

Identification of some complex NP

Unlike base NP, there is not a widely accepted definition of complex NP. It is even worse in Chinese due to the big difference between Chinese and English languages. A lot of complex NPs in English are difficult to be translated into Chinese as a unit. Therefore, with the help of a linguistic, we selected 40 high frequently used English NP patterns, which can be translated into Chinese as a unit. Some examples are shown in Figure 2. In our experiments, after base NP identification, we applied a fuzzy pattern-matching algorithm to identify complex NPs.

Complex NP pattern	**Example complex NP**
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

3.2 NP translation

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 the [*NN-1 NN-2*] sequence in Chinese, and a [*NN-1 of NN-2*] phrase is usually translated into the [*NN-2 NN-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).

Translation patterns can be extracted from a word-aligned bilingual corpus. We first use 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 extract its Chinese translation patterns (*CPT*). An example is shown in Figure 3. For an English sentence, each word is attached with its POS tag and position. Elements within […] is a base NP, or a complex NP. The aligned Chinese sentence is segmented into a sequence of words, which are attached with their positions. For aligned word-pair, (*x,y*) indicates that an English word in position *x* is connected to a Chinese word in position *y*. When one English word is connected to no Chinese words, it is denoted by (*x*,**e**). Translation pattern illustrates that an English NP of a certain pattern can be translated into a Chinese phrase of a pattern with a probability *P*. The estimation of *P* will be described later.

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 [*BaseNP-1 at BaseNP-2*] → [*BaseNP-2 BaseNP-1*], *P*=0.67

Figure 3: English NP patterns and their Chinese translation patterns

We present a statistical translation model for NP translation. It uses translation patterns and a language model. Our method works as follows. Given an English NP, $ENP=\{e_1,\ldots,e_n\}$, with its NP pattern, *EPT*. For each English term *e* in *ENP*, we get all possible Chinese translations from the bilingual dictionary. For each *EPT*, we get all possible translation patterns *CPT*. Then the best Chinese translated phrase, $CNP^*=\{c_1,\ldots,c_m\}$, is the one that maximizes the Equation (10). Viterbi algorithm is used for efficient search.

$$CNP^* = \arg\max_{CPT,CNP} P(CPT \mid EPT)P(CNP)P(E \mid C) \tag{10}$$

where $P(CPT|EPT)$ is the probability of the translation pattern *CPT* (i.e. the order of translation words), given the English NP patterns. It is estimated by

$$P(CPT \mid EPT) = \frac{C(CPT, EPT)}{C(EPT)} \tag{11}$$

where $C(EPT)$ represents the number of times *EPT* occurs in the English portion of the aligned bilingual corpus, and $C(CPT,EPT)$ represents the number of times *CPT* occurs in the Chinese portion of the corpus if and only if its corresponding English NP is of the pattern form *EPT*. In our experiments, 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 linguistic. For each English NP pattern, there are 4.53 translation patterns on average.

$P(CNP)$ in Equation (10) is a priori probability of words of the translated Chinese NP, whose value is given by the trigram language model. It is estimated by

$$P(CNP) = P(w_1 \ldots w_n) = \prod_{i=1}^{n} P(w_i \mid w_{i-2}, w_{i-1}) \tag{12}$$

To obtain the trigram language model, the training corpus we used consists of approximately 1.6 billion Chinese characters, containing documents with different domains, style, and time (Gao et al., 2000b).

$P(E|C)$ in Equation (10) is the translation probability from Chinese words in *CNP* to English words in *ENP*. We simply assumed a uniform distribution on a word's translation in our experiments. If a Chinese word *c* has *n* translations $e_1,\ldots,e_n$, each of them will be assigned equal probability, i.e. $P(e|c)=1/n$. We have three reasons. First, there are no translation probabilities in the bilingual dictionary. Second, we have no enough parallel corpora for estimation. Finally and more importantly, in our previous study, we used the translation model of the form $P(C)P(E|C)^{\alpha}$, and used a held-out data to estimate α. We obtained a very large value, $\alpha>8$. It indicates that in English-Chinese translation, the selection of the Chinese translation heavily depends on other translations in context. Therefore, Equation (10) can be simplified as follows

$$CNP^* = \arg\max_{CPT,CNP} P(CPT \mid EPT)P(CNP) \tag{13}$$

It is interesting to compare our method to methods in (Ballesteros and Croft, 1997; 1998) and (Brown et al., 1993). Ballesteros and Croft used a word-by-word strategy for phrase translation. It is based on two assumptions that we see are sub-optimal. First, it assumes that there is a one-one mapping between words in English NP and words in Chinese NP. But, in our experiments, only

56% NP translation patterns have such one-one mappings. Second, it assumes that the translation words in a phrase will remain the same order to that in a phrase of source language. But in case of English-Chinese NP translation, 35% translation patterns are in different orders in our experiments. Brown's translation model is of the form $P(C)P(E|C)$. A detailed analysis in (Brown et al., 1993) showed that the main problem with the model is that it incorporates very little linguistic knowledge. It is hard to capture non-local dependencies of the language with 'local' models like *n*-gram models. So even if the translation model generates the right 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 the information of global dependency within a phrase. Although the method is not powerful enough for sentence-level translation, we believe it is enough for NP translation. We will show the effectiveness of query translation using our method in Section 6.

4. Translation selection

Many words in queries are not included in phrases. These words have to be translated word-by-word. However, this is not to say 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 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. EMMI (van Rijsbergen, 1979) has been successfully used to estimate the term similarity in (Ballesteros and Croft, 1998; Adrian, 2000). We take a similar approach. In addition, we also observe that EMMI does not take into account the distance between words. If both words occur within the allowed window, they are considered to co-occur. In some cases, the window is even relaxed to the whole document. 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 new factor of distance in our calculation of word similarity. Formally, the similarity between terms *x* and *y* is

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

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 (which is 0.8 in our experiments).

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.

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

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

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 best translations

5. Lexical sources

The bilingual lexical sources we used include three human compiled bilingual lexicons and a bilingual lexicon generated from parallel bilingual corpus automatically.

The human compiled bilingual lexicons include: HIT lexicon (provided by Harbin Institute of Technology, China), LDC lexicon (ftp://ftp.cis.upenn.edu/pub/ldc_www/hpage.html), and a new E-C Dictionary. They respectively contain 77183, 110834, and 105601 English entries. For each English entry, there are usually several Chinese translations.

Parallel texts may contain additional translations that are not stored in a manually compiled dictionary. Therefore, we extract an additional set of translation relationships from parallel texts. The parallel texts are parallel Web pages gathered by an automatic text mining system (Nie et al., 1999). They amount to 20,000 parallel texts, 870,414 pairs of sentences or 74MB Chinese texts and 51MB English texts. These texts are processed as follows: Parallel texts are first aligned into parallel sentences (Liu et al., 1995). Then word translation relations are estimated. The bilingual training process employs a variant of the model in (Brown et al., 1993) and it is based on an iterative EM (expectation-maximization) procedure for maximizing the likelihood of generating the English given the Chinese portion. Chinese sentences are first segmented into words by using a dictionary, containing approximately 60 thousand words. 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 combination of the above resources results in a large dictionary contains 401477 English entries, including 109841 words, and 291636 phrases.

6. Experimental Results and Discussion

In this section, we present the results of our main experiments of CLIR. Most of our experiments are conducted on TREC 9 Chinese corpora of size approximately 260 MB. The documents in the collection are articles published in Hong Kong Commercial Daily, Hong Kong Daily News, and Takungpao. A set of 25 English queries has been set up and evaluated by people in the NIST (National Institute of Standards and Technology).

Each of the TREC queries, also called topics, has three fields: title, description, and narratives, plus manually translations.

For our experiments, we used a slightly modified version of the SMART system (Buckley, 1985).

6.1 Impact of lexical sources, phrase, and ambiguity on query translation

To investigate the impact of lexical sources, we compared the IR performance using our combined dictionary with that using only LDC lexicon. In order to exam the impact of phrase translation, dictionaries are used in two modes: word-only and word-phrase. In word-only mode, only word entries in the dictionary are used, while in word-phrase mode, phrase entries are first used for phrase detecting/matching and translation before word entries are used. Since the LDC E-C lexicon contains almost no phrase entries, we extracted 18777 phrase entries from the LDC C-E lexicon as a complement. To get an upperbound on performance of any disambiguation techniques, we manually disambiguated the queries, called the *best-sense* method in the following description. 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.

In our experiments, we used two versions of queries, short and long. Short queries only use the words in the title field of the topics. Long queries use title, description, and narrative. The results are shown in Table 2. Several conclusions are suggested by the experiments.

First, as expected, combined dictionary achieves consistently better retrieval results by 12% to 35% improvements. It indicates that lexical completeness is an important factor. For example, as shown in row 3, in phrase-word mode, we detected 58 phrases from the queries using combined dictionary, and only 40 phrases using LDC lexicon. When using *best-sense* method, as shown in row 4, all correct translations can be selected using combined dictionary, while using LDC, more than 10% terms have no correct translation or remain un-translated.

Second, not surprisingly, the more we detect phrases, the better the performance we obtained. This conclusion may be drawn by comparing results in rows 2 and 4: the improvements of combined lexicon over LDC lexicon is much larger in phrase-word mode (25% for short queries and 60% for long queries) than in word-only mode (14% for short queries, and for 15% long queries).

Finally, our results show that manual disambiguation improves performance significantly and consistently on TREC-9. The largest improvement is 112% in short query retrieval when using combined lexicon in phrase-word mode, while the minimum improvement is still more than 43% in long query retrieval when using combined lexicon in phrase-word mode. This result clearly indicates the great benefit of a correct translation selection.

	Translation method	Short queries		Long queries	
		LDC	Combined	LDC	Combined
1	Word-only	0.0861	0.0981	0.0962	0.1106
2	1 + *best-sense*	0.1562	0.1821	0.1611	0.1938
3	Phrase-word	0.0938	0.1174	0.1140	0.1823
4	3 + *best-sense*	0.1611	0.2499	0.1733	0.2618

Table 2: Average retrieval precision of the English translation queries

6.2 Impact of NP translation

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

1. *Monolingual*: retrieval using provided (manually translated) Chinese queries.
2. *Simple translation*: retrieval using query translation obtained by looking up the bilingual dictionary (as the method shown in row 3 of Table 2).

3. *Best-sense translation*: retrieval using query translation obtained by manually selecting the best translation among the translations in the bilingual dictionary for each query term (as the method shown in row 4 of Table 2).
4. *Machine translation*: retrieval using translation queries obtained by the machine translation software system.
5. Our method that incorporates NP detection and translation, as well as word translation selection.

The combined dictionary (in phrase-word mode) is used in all these experiments. For *machine translation*, a commercial English-Chinese machine translation system called IBM HomePage Dictionary™ 2000 is used. The system is released recently by IBM. It contains a 480K English-Chinese dictionary, including 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 3% of the words are left un-translated, most phrases are translated as a whole, and the translation ambiguity problem of most words are solved to some degree.

The results of this series of experiments on query translation are summarized in Table 3. It is interesting to compare the results of our NP translation method (in row 4) with that of phrase translation using dictionary looking-up (in row 2). It turns out that by using NP identification and translation, we obtained significant improvements. For example, in short query retrieval, in 25 queries, only 11 multi-term concepts are stored in the dictionary, and translated as a phrase, while using our method, 26 NPs are identified and translated. It thus results in 102.6% improvements.

Row 5 shows that further improvements can be obtained using translation selection technique. It is not surprising that we achieved much larger improvements in long query retrieval (i.e. 23.3% improvements) than in short query retrieval (i.e. 3.7% improvements), because long query provides much richer context information for translation disambiguation/selection. Our results in row 5 are even better than *best-translation* results in row 3, which are regarded as the upperbound on performance of any disambiguation techniques. It shows again the effectiveness of our NP identification and translation methods.

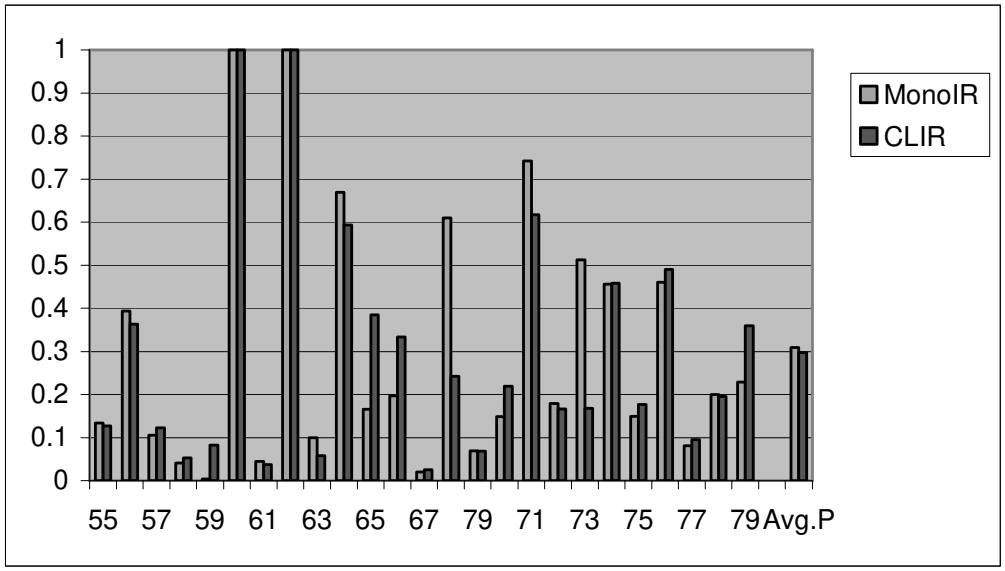
IBM HomePage Dictionary™ 2000 is a powerful machine translation system for query translation. As shown in row 6, it achieves 79.57% of monolingual effectiveness for long queries retrieval. This performance is comparable to those reported by others using an MT system. We can see that our method largely the MT system. This confirms that there may be better way 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 upperbound 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.

	Translation method	Short queries		Long queries	
		Avg. P.	% of Mono. IR	Avg. P.	% of 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 + term disambiguation	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 3: Average retrieval precision of the English translation queries



In order to analyze the effectiveness and remaining problems of our query translation approach, for the 25 queries, we display in Figure 5 a comparison of the long query retrieval results of row 1 and row 5 in Table 3. 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.

2) 11 queries with monolingual average precision 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. These multiple translations usually are exchangeable. Multiply translations act as 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 word '侵害' rather than '违反' in the manually translated Chinese query (provided in the TREC 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 at compositional NP translation. But 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. There maybe two reasons. The first is due to the problem of dictionary incompleteness.

The second is due to the domain bias of the training corpus we used for trigram language model training. That is, we have not enough training corpus on military domain.

Non-compositional NP is the phrase, which cannot be assembled by translations of words within the phrase. Our method can do nothing for 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 the most likely to be translated incorrectly. This indicates that the completeness of the dictionary is still one of the bottlenecks to improve the performance of CLIR.

Figure 5. TREC-9 results for 25 queries: monolingual IR vs. CLIR using long queries

7. 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 improving dictionary-based query translation for CLIR. We focused on translation of phrases, which has been demonstrated to be one of most effective way to disambiguate dictionary 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 disambiguation/selection based on the cohesion among translation words.

Through our experiments, we showed that each of the above methods leads to improvements, 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-shelve MT systems. The combination of our approach with the MT system led 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 improvements.

Though our method shows very promising improvements 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 and more reliable translation patterns. Second, to translate queries in a specific domain, it would be better to use a domain-specific language translation and language model. This could help selecting the correct domain-specific translations. Then there comes the problem of building domain-specific translation/language models, and the activation of the corresponding model when a specialized query is submitted. These are some of the topics of our future work.

Reference

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

Argamon, S., Dagan, I., and Krymolowski, Y. (1998). A memory-based approach to learning shallow language patterns. In *Proceedings of the 36th International Conference on Computational Linguistics*, 67-73. COLING-ACL'98

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.

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.

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

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

Cardie and Pierce, D. (1998). Error-driven pruning of treebank grammas for baseNP identification. In: *Proceedings of the 36th International Conference on Computational Linguistics*, 218-224. COLING-ACL'98

Chen, S. 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.*

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.

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

Franz, M., McCarley, J., and Roukos, S. (1999). Ad hoc and multilingual information retrieval at IBM. In *The 8th Text Retrieval Conference (TREC-8).*

Gao, J., Nie, J. Y., Zhang, J., Xun, E., Su, Y., Zhou, M., and Huang, C. (2000a). TREC-9 CLIR experiments at MSRCN. In: *The Ninth Text Retrieval Conference (TREC-9).* NIST, Gaithersbury, MD.

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

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

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.

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

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.

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.

Kwok, K. L., and Chan, M. (1998). Improving two-stage ad-hoc retrieval for short queries. In: *Proceedings of the 21st International Conference on Research and Development in Information Retrieval*.

Liu, X, Zhou, M., Zhu, S., and Huang, C. (1998). Aligning sentences in parallel corpora using self-extracted lexical information, *Chinese Journal of Computers (in Chinese), 1998, Vol. 21 (Supplement):151-158.*

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

Nie, J. Y. Gao, J. Zhang, J., and Zhou, M. (2000) On the use of words and n-grams for Chinese information retrieval. In: *Fifth International Workshop on Information Retrieval with Asian Languages, IRAL-2000.* Hong Kong, September 30 to October 1, 2000.

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.

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

Xu, J., and Croft, W. B. (1996). Querying expansion using local and global document analysis. In: *Proceedings of the 19th International Conference on Research and Development in Information Retrieval.* 4-11.

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.

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

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

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