

# Phrasal Translation for English-Chinese Cross Language Information Retrieval

Aitao Chen  
School of Information Management and Systems  
102 South Hall  
University of California at Berkeley, CA 94720, USA  
aitao@sims.berkeley.edu

## Abstract

*This paper introduces a simple and effective non-overlapping unigram and bigram segmentation method for both monolingual Chinese and English-Chinese cross language retrieval. It also describes English-Chinese cross language retrieval experiments involving 54 topics and some 164,000 documents. The translation of English queries to Chinese is done using a Chinese-English dictionary of about 120,000 entries. A technique for extracting noun phrases is presented and applied prior to query translation. The phrasal translation outperformed word translation by 23.6% even though most of the extracted noun phrases from the queries were not translated as phrase because of the limited coverage of the bilingual dictionary. The cross language retrieval achieved about 53% of the effectiveness of the monolingual retrieval, which suggests that there is lot of room for improvement. The two main limiting factors in English-Chinese retrieval performance are the limited coverage of the bilingual dictionary and the existence of multiple Chinese translation equivalents for many English words.*

## 1 Introduction

The increasingly large volume of Chinese textual data available online requires efficient and practical means to access the data in other languages. In particular, support is needed for people who are not proficient in Chinese but understand English. One such means of providing access to Chinese data in English is to couple Chinese information retrieval with English-Chinese machine translation system or some shallow translation techniques. To this date relatively little work on cross language information retrieval involving English and Chinese has been done. Part of the problem is the lack of data and linguistic resources to do research in this area. The availability of a bilingual wordlist and the TREC Chinese test collection made it possible for

us to experiment with English-Chinese cross language retrieval. We take the simple approach of translating English queries into Chinese. We also examine the effectiveness of phrasal translation and word translation in English-Chinese cross language retrieval.

The rest of this paper is organized as follows. In section 2 we propose an effective segmentation method for English-Chinese cross language retrieval. Section 3 describes the test collections and the bilingual dictionary for query translation. Section 4 reports the monolingual Chinese retrieval evaluation results and compares three segmentation methods. Section 5 describes English-Chinese cross language retrieval experiments using a bilingual dictionary. It also presents a technique for noun phrase extraction in topics. Section 6 discusses two factors that limit the performance of English-Chinese cross language retrieval. Section 7 reports the cross-language retrieval performance when the extracted noun phrases that are missing in the bilingual dictionary were translated by hand. Section 8 mentions related work. Section 9 concludes the paper and points out future work.

## 2 Word Segmentation

The documents and queries in most text retrieval systems are indexed by the words occurring in the text. For languages such as English in which words are separated by blank space, it is very simple to index text by words. To index Chinese text by words, however, one first needs to identify words in the text since the words are not explicitly marked in Chinese text. A Chinese sentence consists of a continuous string of Chinese characters. The process of breaking sentences into words is referred to as word segmentation. There is a large literature on Chinese word segmentation. We will not attempt to survey this field. Two recent papers on Chinese word segmentation are presented by Dai and Loh in [4] and Sun et al. in [10]. Both corpus-based statistical methods and dictionary-based methods have been developed to break a sentence into in-

dividual words. If one has a Chinese word dictionary, one could match the text against the dictionary and output as a word the (longest) sequence of characters that matches an dictionary entry. When a dictionary is not available, one could collect large amount of Chinese text and attempt to discover words by examing the occurrence patterns of the characters in the corpus.

A major problem with dictionary-based word segmentation methods is the dictionary coverage. The word segmentation accuracy may largely depend on the dictionary coverage. When the dictionary coverage is poor, the words that are missing in the dictionary may not be recognized.

The corpus-based or statistical methods can be easily applied to a new collection of Chinese text since they do not use word dictionaries. The overlapping bigram indexing is simple, efficient and effective as well [8]. One problem with bigram indexing is that the indexing file produced is two to three times as big as the size of the raw text. Here we refer to single Chinese characters as unigrams and two-character Chinese terms as bigrams.

We present a method that is equally efficient and effective as bigram indexing, but produces a much smaller index file than the overlapping bigram indexing. Our method is inspired by the model 1 of the IBM statistical machine translation [5] in which the target language sentence can be generated through different alignments from the source language sentence. Similarly a Chinese sentence can be generated from words combined through different segmentations. Our method is similar to but less general than the work presented by Ge et al. in [6]. Our method breaks a sentence into unigrams and bigrams by maximizing the probability of the sentence. Here we assume that unigrams and bigrams occur independently in the corpus. Let  $S = C_1C_2C_3 \dots C_n$  be a sentence consisting of  $n$  characters. The segmentation of a sentence of  $n$  characters can be represented by a  $n - 1$  dimensional vector  $K = (k_1, k_2, \dots, k_{n-1})$  since there are  $n - 1$  word boundary positions, where  $k_i$  indicates if there is a word boundary between character  $C_i$  and  $C_{i+1}$ . All  $k_i$ 's take on value 1 or 0. A value of 1 means there exists a word boundary and a value of 0 means otherwise. We call vector  $K$  the segmentation vector of sentence  $S$ . Each segmentation of a sentence can be represented by a unique segmentation vector. Since the  $k_i$ 's are binary, it is easy to see there are  $2^{n-1}$  possible ways to break up a sentence of  $n$  characters into words when a word can be arbitrarily long. For example, there are four ways to segment the sentence  $S = C_1C_2C_3$ . The different segmentations of the sentence are enumerated in Table 1. However, when the word length is limited to one or two characters, the number of possible ways to segment a sentence of  $n$  characters is given by the recurrence relation  $N(n) = N(n - 1) + N(n - 2)$ , where  $N(n)$  is the number of ways to break a sentence of  $n$  characters into one or

two-character words and  $N(0) = 0, N(1) = 1, N(2) = 2$ . For example, there are three ways to break up a sentence of three characters into one or two-character words as enumerated in the third column in Table 1.

No.	Segmented Sentences (word length unlimited)	Segmented Sentences (word length $\leq 2$ )
1.	$C_1/C_2/C_3$	$C_1/C_2/C_3$
2.	$C_1/C_2C_3$	$C_1/C_2C_3$
3.	$C_1C_2/C_3$	$C_1C_2/C_3$
4.	$C_1C_2C_3$	

**Table 1. All possible ways to break up a three-character sentence,  $S = C_1C_2C_3$ , into words when word length is unlimited (column 2) and when word length is restricted to 1 or 2 characters (column 3).**

For a segmented sentence  $S = w_1w_2 \dots w_m$ , if we assume words occur independently, then the probability of the sentence  $S$  can be expressed as follows:

$$\begin{aligned}
 P(S) &= P(w_1w_2 \dots w_m) & (1) \\
 &= P(w_1)P(w_2) \dots P(w_m) = \prod_{i=1}^m P(w_i) & (2)
 \end{aligned}$$

since we do not know how to break a sentence into words in advance, we will consider all possible ways of segmenting a sentence and estimate the probability of every segmentation given a sentence. We can then use the segmentation of the highest probability to break up the sentence into words. We denote by  $K$  a segmentation variable which represents the different ways of segmenting a sentence. Then the probability of a sentence can be expressed as the sum of the joint probability of a sentence with a particular segmentation. That is,

$$P(S) = \sum_k P(S, k) \quad (3)$$

where the summation is taken over all possible segmentations for the sentence. Given a sentence, we are looking for the most likely segmentation of a sentence.

$$P(k|S) = \frac{P(S, k)}{P(S)} \quad (4)$$

where  $P(k|S)$  is the probability of segmentation  $k$  given the sentence,  $P(S, k)$  is the probability of the sentence with segmentation  $k$ . The most likely segmentation for a sentence can be found by maximizing the probability  $P(S, k)$  as shown below:

$$k(S) = \operatorname{argmax}_k \frac{P(S, k)}{P(S)} = \operatorname{argmax}_k P(S, k) \quad (5)$$

since  $P(S)$  is the same for a sentence. When a sentence is short, one can easily enumerate all possible ways of segmenting the sentence and compute their associated probabilities, then choose the segmentation of the highest probability. But when a sentence is long, the number of possible segmentations is exponential, it is no longer practical to enumerate all possible ways of breaking the sentence and estimate their probabilities. However one can apply dynamic programming technique to find out the most likely segmentation efficiently without computing the probabilities of all possible segmentations of a sentence. The best way of breaking a sentence of  $n$  characters can be recursively expressed as follows:

$$P(S_{1,n}) = \text{MAX}(P(S_{1,n-1})P(C_n), P(S_{1,n-2})P(C_{n-1}C_n))$$

where  $S_{1,n} = C_1C_2 \dots C_n$  and  $P(S_{1,n})$  is the maximum probability of segmenting a sentence of  $n$  characters into one or two-character words. The probability of a one-character word (i.e., unigram) is estimated by  $P(C_i) = \frac{N(C_i)}{N}$ , and the probability of a two-character word (i.e., bigram) is estimated by  $P(C_iC_j) = \frac{N(C_iC_j)}{N}$ , where  $N(C_i)$  is the number of times that character  $C_i$  occurs in the corpus,  $N(C_iC_j)$  is the number of times that string  $C_iC_j$  occurs in the corpus and  $N$  is the total number of times that any single character terms and any two-character terms occurs in the corpus. A sentence is broken into one or two-character terms using the most likely segmentation. For example, for the sentence of three characters,  $S = C_1C_2C_3$ , shown in Table 1, the probability of the sentence with the three different possible ways of segmentation are given, respectively, by

$$P(S, (1, 1)) = P(C_1)P(C_2)P(C_3) \quad (6)$$

$$P(S, (1, 0)) = P(C_1)P(C_2C_3) \quad (7)$$

$$P(S, (0, 1)) = P(C_1C_2)P(C_3) \quad (8)$$

Assume that the second segmentation method ( $k = (1, 0)$ ) has the highest probability, then we break sentence  $S$  into  $C_1/C_2C_3$ . This is the method we used to break the Chinese sentences in the test collection into one or two-character terms. The probability of a one-character or two-character term is estimated using their occurrence statistics collected from the People’s Daily collection. The Chinese topics are segmented in the same way. When this method is applied to new text such as the topics, it is likely that there will be terms missing in the collection from which occurrence statistics are collected for all terms. The estimated probability for the new terms would be zero. Of course, having not seeing a term in the test collection does not necessarily mean this term will never occur in the future text. When we used this method to segment topics, we assigned a small probability to the terms missing in the test collection. The

estimated probability for a new term is one over the total number of unique unigrams and bigrams.

One problem with the non-overlapping unigram and bigram method is that the same two adjacent characters may be recognized as a bigram in some contexts but be separated in others since whether or not two-adjacent characters should be grouped as a bigram depends on the probabilities of all unigrams and bigrams found in the same sentence. This method also does not work well with words of three or more characters such as most of the personal names.

### 3 Test Collection

We used the TREC-5 [11] and TREC-6 [12] Chinese test collection in all of our Chinese monolingual and English-Chinese cross language retrieval experiments reported below. The test collection consists of 139,801 articles from the People’s Daily newspaper published between 1991 and 1993 and 24,988 news reports from Xinhua News Agency in 1994 and 1995. There are 54 topics, consisting of a title, description and narrative fields. The titles are usually phrases or clauses or short sentences. And a description consists of some main concept terms in a topic. The narratives states what makes a document relevant or non-relevant.

All topics come with English translations. For the English-Chinese cross language retrieval, we used the English translations as the English topics. There are 5140 documents in the test collection found relevant to the 54 topics.

The bilingual dictionary we used to translate English queries is the Chinese-to-English wordlist (version 2.0) compiled by Linguistic Data Consortium. We downloaded the bilingual wordlist from <http://morph ldc.upenn.edu/Projects/Chinese/>. The wordlist consists of a list of Chinese words, paired with a set of English words. The wordlist has some 128,000 entries.

For each topic, one thousand top-ranked documents are retrieved from the document collection. The documents are ranked by their probability of relevance to a topic. And the relevance probability of a document to a topic is estimated using the ranking algorithm developed by Cooper et al. [2].

### 4 Monolingual Experiments

We indexed the test collection using the word dictionary-based longest matching method, overlapping bigram and non-overlapping unigram and bigram method presented in section 2. The wordlist used in segmentation are extracted from the Chinese-to-English dictionary mentioned above. The following table shows the size of the index files and the dictionary files produced from three segmentation methods. The source document collection has 167 MB text.

recall level	word	overlapping bigram	non-overlapping unigram and bigram
average precision	0.4395	0.4446	0.4385
relevant retrieved	4423	4593	4505

**Table 2. The performances of three segmentation (indexing) methods.**

segmentation method	index file size (MB)	dictionary size (terms)
word	146	48,111
overlapping bigram	318	1,390,046
non-overlapping unigram and bigram	159	281,591

The retrieval performance of all three indexing methods are comparable as the results in Table 2 show. The retrieval performances of the three different indexing methods are summarized in Table 2. The retrieval performance of all three indexing methods are comparable as the results in Table 2 show.

## 5 Cross-Language Retrieval

There are a number of ways to perform the task of cross-language information retrieval in which a query posed in one language is searched against a collection of documents written in a different language. Oard and Diekema provide a survey on cross-language information retrieval in [9]. It is obvious that any retrieval method based on matching a query in one language against documents in a different language would fail when there are no cognates between this language pairs (e.g., Chinese and English). For matching-based retrieval algorithms to work, both the documents and queries need to be expressed in the same language or conceptual space as in the latent semantic indexing. A common approach to cross-language information retrieval is to couple translation with monolingual information retrieval. One can translate users' queries into the document language, or translate documents into the query language, or translate both the queries and documents into a third language. One can translate queries or documents using a machine translation system. When such resource is not available, one can use bilingual dictionaries if available to do word translation or phrase translation, or one can resort to parallel or comparable bilingual corpora from which to mine translation dictionary or to build a common conceptual space as in latent semantic indexing method for cross-language retrieval.

For the English-Chinese cross language retrieval experi-

ments reported below, we take the simple approach of translating queries to the document language, that is, we translate the English queries into Chinese. We then apply the monolingual retrieval ranking algorithm to rank Chinese documents by their estimated probability of relevance to the translated Chinese queries.

The English queries are translated into Chinese by looking up English phrases and words in the Chinese-to-English bilingual wordlist. All Chinese equivalents of an English word or phrase are retained in the translation. We do not rank the Chinese equivalents of an English word and then choose the most appropriate Chinese translation when there are more than one equivalents. We translate the English phrases in the queries first whenever there are matching phrase entries in the bilingual dictionary to ameliorate the problem of resolving translation disambiguities. In general, the phrases are more specific and thus less ambiguous than single words. Therefore overall one would expect that phrases have fewer translation equivalents than single words.

### 5.1 Topics Preprocessing

The topics were processed in three steps to generate the queries before translation. First, the topics were tagged using Brill's part-of-speech tagger [1]. Second, noun phrases are extracted from the tagged topics. Third, the single-word terms and phrases are normalized using a morphological analyzer. The following text shows the tagged text of the narrative field in topic 23.

---

```
</NARR>
a/DT relevant/JJ document/NN discusses/VBZ the/DT
Soviet/NNP Union/NNP ' /POS s/PRP mediation/NN in/IN
the/DT Gulf/NNP War/NNP ,/, including/VBG
communication/NN with/IN Iraq/NNP ,/, cease-fire/NN
resolution/NN to/TO the/DT UN/NNP Security/NNP
Council/NNP and/CC their/PRP$ peace/NN proposal/NN
for/IN withdrawal/NN of/IN multi-national/JJ
troops/NNS ,/, etc/FW ./.
```

---

Each word is followed by its part-of-speech tag. The tags NN and NNS represent singular nouns and plural nouns, respectively; NNP represents the proper name, and JJ represents adjective. Then the tagged text is passed to a noun phrase recognizer for noun phrase extraction. The recognizer detects simple noun phrases based on the pattern of the tags. The noun phrase patterns we used to extract noun phrases is concisely specified in a three-state automaton as shown in Figure 1. The initial state is 0 and the final state is 2. Any words tagged with part-of-speech tags NN, NNS, NNP, NP and NPS are represented by the label NOUN, and words tagged with JJ, JJR, and JJS, which are the positive,

comparative and superlative form of an adjective, are represented by the label ADJ. Any sequence of words whose part-of-speech tags completes a path from the initial state to the final state will be extracted as a noun phrase, excluding the single-word nouns.

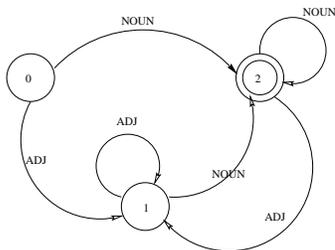


Figure 1. Simple noun phrase automaton

The noun phrases extracted from the above tagged text are presented here:

```

relevant document
Soviet Union
Gulf War
cease-fire resolution
UN Security Council
peace proposal
multi-national troops
  
```

A total of 367 multi-word noun phrases were extracted from the 54 English topics. The words appearing in the stoplist were removed and then the remaining single words and noun phrases are normalized using a morphological analyzer [3], which reduces plural nouns to their singular form and verbs to their base form. Also, all words and phrases are converted to lower case. The normalized single words and the simple noun phrases constitute the English queries before translation.

## 5.2 Query Translation

After the preprocessing of the English topics, each query now is comprised of single words and noun phrases. We translate each query by looking up every single word and noun phrase in the Chinese-English bilingual dictionary. If there is an exact match for a single-word term or a noun phrase in the bilingual dictionary, all the Chinese equivalents from the bilingual dictionary are retained in the translated query. When there is no exact match for a single-word term, that term is not translated. However when there is no exact match for a noun phrase, we proceed to match the sub-phrases against the dictionary until there are some matches. If all sub-phrases matching fails, we then look for exact matches for the component words in the phrase. For example, if a three-word phrase  $w_1w_2w_3$  is missing in the

English words/phrases	Chinese translations
soviet union	苏联
mediation	调解 / 调停 / 仲裁
gulf war	海湾战争
communication	交通 / 联络 / 通讯 / 通讯会话
iraq	伊拉克 / 伊拉克共和国
cease-fire resolution	停火
un security council	安理会
peace proposal	安 / 和 / 和平 / 虞 / 只 动议 / 方案 / 建议 / 提案 提议 / 议案
withdrawal	No translated
multi national troops	多 / 多工 / 多种 / 复选 国立 / 国民 / 侨民 / 全国性 兵力 / 兵员 / 部 / 部队 / 队伍

Table 3. This table shows the English words/phrases in the first column and their Chinese translations in the second column.

dictionary, we will search the sub-phrases  $w_1w_2$  and  $w_3$ ; and if there is no match for  $w_1w_2$ , we will search  $w_1$  and  $w_2w_3$  in the dictionary. If none of the sub-phrases is found in the dictionary, we translate this phrase word-by-word by looking up each component word in the dictionary, and take the Chinese translations of all the component words in the phrase as the translation of the phrase.

The Chinese translation equivalents for the single words and noun phrases extracted from the above narrative text are presented in table 3. Only the phrases “Soviet Union”, “Gulf War”, and “UN Security Council” are translated into Chinese as a whole since they are present in the Chinese-English bilingual dictionary. The other three phrases “cease-fire resolution”, “peace proposal” and “multi-national troops” are missing in our dictionary. The phrases “peace proposal” and “multi-national troops” are translated word-by-word, resulting in many Chinese words as shown in table 3. The single term “withdrawal” is not translated because it is missing in our dictionary.

We translated all 54 English queries into Chinese. The Chinese translation equivalents were then segmented into one or two-character words using the segmentation method as described above. The documents in the collection were segmented into one or two-character words as well. The probability of the unigrams and bigrams were estimated from the People’s Daily collection only. As in the monolingual retrieval experiments, the Chinese documents are ranked by their probability of relevance to a Chinese query. And the top-ranked 1000 documents are retrieved and eval-

recall level	word-by-word translation	phrasal translation
average precision	0.1874	0.2316
relevant retrieved	3023	3414

**Table 4. The performances of English-Chinese cross language retrieval for word-by-word translation (col 2) and phrasal translation (col 3). The Chinese translations were segmented into non-overlapping one or two-character words.**

uated.

We also translated the queries into Chinese by looking up each English word in the bilingual dictionary. If there is an exact match, all Chinese translation equivalents are retained. In this experiment, phrases are translated word-by-word.

### 5.3 Results

We carried out two runs of cross-language retrieval using word-by-word translation and phrasal translation. After query translation, the Chinese equivalents were then segmented into non-overlapping one or two-character words. The cross-language retrieval performance using the word-by-word translation method is presented in the second column in Table 4. Column 3 presents the evaluation results for the phrasal translation experiment. The query translation by phrase achieved an average precision of 0.2316, which is 23.6% better than the average precision of the word-by-word translation even though 322 noun phrases were translated word-by-word or in sub-phrases because they are missing in the dictionary. The average precision should be even higher if all noun phrases found in the topics can be translated as a whole. However, the average precision of the English-Chinese cross language retrieval using phrase translation is only about 52.8% of the average precision achieved in the Chinese monolingual retrieval. The difference between the performances of Chinese monolingual and English-Chinese cross retrieval suggests that there is still a lot of room for improvement for English-Chinese cross language retrieval.

We tested the same three segmentation methods on the translated Chinese queries from the original English queries by phrasal translation. The experiments of using three indexing methods in the monolingual Chinese retrieval show the overlapping bigram indexing achieved marginally better precision than the other two indexing methods. However, the results presented in Table 5 for the English-Chinese

	word	overlapping bigram	non-overlapping unigram and bigram
average precision	0.2093	0.1887	0.2316
relevant retrieved	3274	3208	3414

**Table 5. The performances of the three word segmentation (or indexing) methods in English-Chinese cross language retrieval. The English queries were translated into Chinese by phrasal translation.**

cross language retrieval show that the overlapping bigram indexing performed substantially worse than the other two methods. The inferior performance of overlapping bigrams in comparison to the other two methods could be attributed to the poor translation quality of the English queries. Most of the query words/phrases were translated word-by-word into Chinese. And most of the English words have multiple translation equivalents in Chinese in our bilingual dictionary. We did not rank and select the most appropriate Chinese translation for each English word, which may have resulted in the poor translation quality. For example, the noun phrase "multi-national troops" produced 13 Chinese words as shown in table 3. Many of the translations for "multi-national troops" were not appropriate. When overlapping bigrams are extracted from the translated Chinese words, even more inappropriate words will be included in indexing, which will further degrade the performance of overlapping bigrams. When the performances in both monolingual and cross language retrieval are considered, the non-overlapping unigram and bigram indexing method is more effective than the other two.

## 6 Discussions

When an English query is translated into Chinese, one should attempt to identify the noun phrases and the proper nouns and then look them up in a bilingual dictionary. One should translate them word-by-word into Chinese only when they are missing in the dictionary. While the component words in a noun phrase or proper noun may have several translation equivalents, in general a proper noun or noun phrase has only one translation. For example, the following proper noun and noun phrases all found in the test topics have one Chinese translation: "human rights", "intellectual property rights", "World Trade Organization", "UN Security Council", "Olympic Games", "China Red Cross". Since the proper nouns and noun phrases are unambiguous and thus have only one translation, there is no need to choose the most appropriate translation. To see the ad-

vantages of translating queries by phrases over the word-by-word translation, we will look at two examples: “human rights” and “China Red Cross”. The noun phrase “human rights” is “人权” in Chinese. However, when the phrase is translated word-by-word, we get “/ 人的 / 人性的 / 通人情的 / 人间性的 / 人 / 人类 / 人性 /” for the word “human” and “权利” for the word “rights”. Even though one of the combination “人的权利” is an appropriate translation in meaning, the form is different from the word “人权” which occurs in the documents. Some semantic analysis will be needed to reduce the term “人的权利” to the proper form “人权”. Otherwise the translated term will not match the proper form “人权”. Here the phrase as a whole has one meaning only, whereas the word-by-word translation resulted in eight individual words.

The Chinese name for “China Red Cross” is “中国红十字会”. Again when the English name is translated word-by-word into Chinese as was done in our query translation since it is not in the dictionary we used, we get “华 / 中 / 中国 / 中华 /” for “China”, “赤 / 丹 / 红 / 红的 / 红热的 / 红色 /” for “Red”, and “叉 / 错 / 交叉 / 十字架 / 小十字形 /” for “Cross”. None of the combinations from the individual translations will match the Chinese name “中国红十字会”. Each component word in the English name has multiple Chinese equivalents, which compound the problem of choosing the proper ones from all translation equivalents.

The above two examples clearly illustrate that the proper nouns and noun phrases should be translated as a whole rather than as a set of component words.

We believe that the two major factors that may have degraded the performance of the English-Chinese cross language retrieval are 1) the limited coverage of the bilingual dictionary; and 2) the existence of multiple Chinese translation equivalents for many of the English words. Some other factors include misspelling of words in the topics and dictionary and inappropriate translations of English words.

Many important words or phrases in the test topics are missing in our dictionary. Among the missing ones are “reunification”, “most-favored-nation status”, “China Red Cross”, “peace-keeping troops”, “Asian-Pacific”, “Mid-East”, “Information Super Highway” “hijackings”, “oil fields”, “acid rain”, “South-African”, “Pinatubo”, “Minatubo”, “Sino-American”, “Sino-Vietnamese”, “joint communicate”, “campuchea”, “underground unclear tests”, and “non-proliferation treaty”. Some of the mismatches are the result of misspellings occurred in the original topics or in the dictionary or the result of inconsistency. For example, “South-Africa”, “Asia-Pacific”, and “Middle East” are all in our dictionary. The inappropriate translation of a term presents another problem. For example, the term “cellular phone” is translated into “汽车电话” in our dictionary,

	word	overlapping bigram	non-overlapping unigram and bigram
average precision	0.3811	0.3873	0.3593
relevant retrieved	4102	4357	4125

**Table 6. The performances of the three word segmentation methods in cross-language retrieval. The noun phrases missing in the dictionary were manually translated.**

which is too narrow. More importantly that is not the Chinese term used in the documents which is “移动电话”.

The existence of multiple translation equivalents of an English word as shown in the translation of the narrative text in topic 23 and the word-by-word translations of “human rights” and “China Red Cross” is another source of degrading the performance of the cross language retrieval. Clearly when there are multiple translation equivalents, one should attempt to resolve translation ambiguities.

## 7 Manual Phrasal Translation

Because of the limited coverage of our bilingual dictionary, the majority of the noun phrases extracted from the original English topics were not translated as phrases. We were interested in finding out 1) what the English-Chinese cross-language retrieval performance would be if all the noun phrases were translated as phrase, and 2) if the different segmentation methods make any difference when all noun phrases were translated as phrase. We translated the queries into Chinese by looking the phrases and single-word terms in the bilingual dictionary. If there is a match for a phrase or a single-word term, we keep the Chinese translations. We discarded the single-word terms that are missing in the dictionary. However we manually translated all the noun phrases that are missing in the dictionary into Chinese as phrases. We then segmented the Chinese translations into words using 1) word segmentation, 2) overlapping bigram, 3) non-overlapping unigram and bigram. The segmented queries were used to search against the Chinese collection. The retrieval performances were presented in table 6

The results in table 6 shows that the non-overlapping unigram and bigram segmentation method no longer has an advantage over the other two methods. The results also show that good retrieval performance in English-Chinese cross-language retrieval can be achieved if the bilingual dictionary used to translate queries has wide coverage in noun phrases.

## 8 Related Work

The only work on English-Chinese retrieval using a large test collection set that we are aware of is that presented by Kwok [7]. The same test collection and the topics are used as in our work. While the best cross language retrieval performance in Kwok's study is 71% of the monolingual retrieval for the long queries, and 79% for the short queries, our cross language retrieval performance is about 52.8% of the monolingual retrieval. A machine translation package augmented with a bilingual word dictionary is used to translate queries into Chinese in Kwok's study. Furthermore, query expansion prior to translation and combination of retrieval results seem to help achieve better results in Kwok's study. In our work, we only used a bilingual dictionary to translate queries with no pre-translation or after-translation query expansion. Also, it seems that many more query terms in our work are not translated as a result of limited coverage of our bilingual dictionary.

## 9 Conclusions

We have presented a statistical word segmentation method which is efficient and effective for both monolingual Chinese retrieval and English-Chinese cross language retrieval. We have also presented a noun phrase recognizer based on the part-of-speech tags assigned to words. The recognizer is used to extract noun phrases from the topics before translation. The queries are translated into Chinese using a bilingual wordlist of some 120,000 entries. Then we have compared the performance between phrasal translation and word-by-word translation. The phrasal translation achieved superior performance than word-by-word translation as a result of the fact that the noun phrases and proper nouns often have one translation only while an English word may have multiple Chinese translation equivalents. Our cross language retrieval effectiveness is about 53% of the monolingual retrieval. The low performance in cross language retrieval may be attributed to the limited coverage of the bilingual dictionary, especially the absence of many important proper nouns and noun phrases, and the existence of multiple translation equivalents for many of the English words. Our future work will focus on augmenting the bilingual dictionary and techniques for resolving translation ambiguity.

## 10 Acknowledgements

We would like to thank the reviewers for helpful comments. This research was supported by DARPA contract "Search Support for Unfamiliar Metadata Vocabularies" N66001-97-C-8541; AO-F477.

## References

- [1] Eric Brill. A simple rule-based part of speech tagger. In *Proceedings of the Third Conference on Applied Natural Language Processing*, 1992.
- [2] W. S. Cooper, A. Chen, and F. C. Gey. Full text retrieval based on probabilistic equations with coefficients fitted by logistic regression. In D. K. Harman, editor, *The Second Text REtrieval Conference (TREC-2)*, pages 57–66, March 1994.
- [3] M. Zaidel D. Karp, Y. Schabes and D. Egedi. A freely available wide coverage morphological analyzer for english. In *Proceedings of COLING*, 1992.
- [4] Y. Dai and T. Loh. A New Statistical Formula for Chinese Text Segmentation Incorporating Contextual Information. In *SIGIR'99, Berkeley, August 1999*, pages 82–89, 1999.
- [5] P. Brown et al. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19:313–330, June 1993.
- [6] X. Ge, W. Pratt, and P. Smyth. Discovering Chinese Words from Unsegmented Text. In *SIGIR'99, Berkeley, August 1999*, pages 271–272, 1999.
- [7] K.L. Kwok. English-chinese cross language retrieval based on a translation package. In *Machine Translation Summit VII workshop on Machine Translation for Cross Language Information Retrieval*, Kent Ridge Digital Laboratories, Singapore, 1999.
- [8] J. Nie and F. Ren. Chinese information retrieval: using characters or words? *Information Processing and Management*, 35:443–462, 1999.
- [9] D. Oard and A. Diekema. *Cross-Language Information Retrieval*, volume 33, pages 223–256. 1998.
- [10] M. Sun, D. Shen, and C. Huang. CSeg&Tag1.0: A Practical Word Segmenter and POS Tagger for Chinese Texts. In *Proceedings of the Fifth Applied Natural Language Processing Conference*, pages 119–126, 1997.
- [11] E. M. Voorhees and D. K. Harman, editors. *The Fifth Text Retrieval Conference (TREC-5)*, National Institute of Standards and Technology, Gaithersburg, MD, 1997.
- [12] E. M. Voorhees and D. K. Harman, editors. *The Sixth Text Retrieval Conference (TREC-6)*, National Institute of Standards and Technology, Gaithersburg, MD, 1998.