English-Chinese Cross-Language IR using Bilingual Dictionaries

Aitao Chen*, Hailing Jiang*, and Fredric Gey†
*School of Information Management and Systems
†UC Data Archive & Technical Assistance (UC DATA)
University of California at Berkeley, CA 94720, USA
{aitao, hjiang1}@sims.berkeley.edu, gey@ucdata.berkeley.edu

Abstract

This report describes the English-Chinese cross-language retrieval experiments at Berkeley for TREC-9 Cross-Language Information Retrieval track. We present a simple and effective Chinese word segmentation method and compare the cross-language retrieval performance of two bilingual dictionaries for query translation.

1 Introduction

In TREC-9 we only participated in the English-Chinese cross-language information retrieval (CLIR) track. We performed one Chinese monolingual retrieval run and three English-Chinese cross-language retrieval runs. Our approach to the cross-language retrieval was to translate the English topics into Chinese by dictionary lookup. An English/Chinese bilingual wordlist compiled by Linguistic Data Consortium and an online English/Chinese bilingual dictionary were used in our cross-language retrieval experiments.

The four official runs we submitted are BRKCCA1, BRKECA1, BRKECA2, and BRKECM1. The BRKCCA1 is a monolingual run, the other three being English-Chinese cross-language runs. The BRKECA1 and BRKECA2 runs are automatic, while the BRKECM1 is manual.

For all four runs, the same document ranking algorithm based on logistic regression technique was used. The details on our ranking algorithm can be found in [2].

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 simple to index text by words. To index Chinese text by words, however, one first needs to identify words in the text since word boundaries are not explicitly marked in Chinese text. 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 [9]. Both corpus-based statistical methods and dictionary-based methods have been developed to break a sentence into individual 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 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 [7]. 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 similar to but less general than the work presented by Ge et al. in [5]. 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. For a segmented sentence $S = w_1 w_2 \ldots w_m$, if we assume words occur independently, then the probability of the sentence S can be expressed as follows:

$$P(S) = P(w_1 w_2 \ldots w_m) \quad (1)$$

$$= P(w_1)P(w_2)\ldots P(w_m) = \prod_{i=1}^{m} P(w_i) \quad (2)$$

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. The number of possible ways to break a sentence of n characters into words is 2^{n-1} when a word can be arbitrarily long. However, when a word is limited to one or two characters, the number of possible ways to segment a sentence of n characters can be expressed 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 $N(0) = 0, N(1) = 1, N(2) = 2$. 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:

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

where $S_{1,n} = C_1C_2 \ldots C_n$ 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, $S = C_1C_2C_3$, the probability of the sentence with the three different possible ways of segmentation are given, respectively, by

$$P(S,(1,1)) = P(C_1)P(C_2)P(C_3) \quad (3)$$
$$P(S,(1,0)) = P(C_1)P(C_2C_3) \quad (4)$$
$$P(S,(0,1)) = P(C_1C_2)P(C3) \quad (5)$$

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 test documents. When we use this method to segment topics, we assign 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.

3 Test Collection

The TREC-9 CLIR test collection consists of 25 new topics and 127,938 documents from three newspapers, namely the Hong Kong Commercial Daily, Hong Kong Daily News, and Takungpao. The topics are written in English with Chinese translations and contain *title*, *description*, and *narrative* fields.

One of the bilingual dictionaries 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.

The other bilingual dictionary used in our experiments is the online KingSoft dictionary at http://ciba.kingsoft.net/online/. It consists of a general dictionary and a set of 23 specialized dictionaries, such as ships, electricity, telecommunication, law, broadcasting, environment, chemistry, economy and trade, computer, medicine, and so on. The general dictionary contains about four million entries and the specialized dictionaries together contain about two million entries [6].

4 Monolingual Experiment

The Chinese documents and the Chinese translations of the English topics were indexed using the overlapping bigram technique. All three fields – title, description, and narrative – in the topics were used. The retrieval performance of the monolingual run BRKCCA1 is presented in the second column in table 1. The overall precision is 0.2936 and recall is .855.

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 recent survey on cross-language information retrieval in [8]. 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 pair (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 for cross-language retrieval.

For the English-Chinese cross language retrieval experiments 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.

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 description field in topic CH58.

> Are/VBP environmental/JJ protection/NN laws/NNS being/VBG enforced/VBN in/IN China/NNP and/CC Hong/NNP Kong/NNP ?/.

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 can be 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.

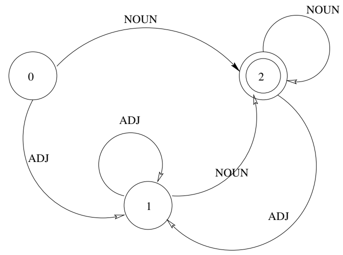

Figure 1. Simple noun phrase automaton

The noun phrases extracted from the above tagged text are *environmental protection laws* and *Hong Kong*. 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 a Chinese-English bilingual dictionary.

For BRKECA1 run, a query term (noun phrase or single word) was looked up in the LDC bilingual wordlists. The top two Chinese translation equivalents that occur most frequently in the test document collection were retained as translations for an English term when there are more than two translations for that term. When there is no exact matching for a single-word term, that term is not translated. However when there is no exact matching 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 matching for the component words in the phrase. For example, if a three-word phrase $w_1w_2w_3$ is missing in the 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 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.

For BRKECA2 and BRKECM1 runs, the noun phrases and their constituent words were looked up in the online KingSoft Chinese/English dictionary. The first Chinese translation for a phrase or word was retained. The Chinese translation equivalents were segmented into words using the longest-matching method. These two runs used the word-based document index for retrieval.

5.3 Manual Query Reformulation

It has been the policy of the Berkeley group to attempt to create manual reformulations of TREC queries since TREC-2. Manual queries usually result in additional relevant documents found which enriches the value of the collection when used for machine learning in the future. Initially this manual reformulation was done without reference to the retrieval, i.e. by searching a comparable collection using the original topic terms. The first of these was the news title database available as part of the University of California's electronic library catalog. Later, as the TREC rules were relaxed to include manual relevant feedback, we have utilized that technique for finding words from top documents of an initial search or by manually marking particular documents as relevant. These techniques were used in our recent CLEF experiments for European languages.

For TREC-9 we created manual versions of the English queries by searching the WWW with topic words and taking pertinent text from the URLs found and inserting it to the manual version of the query. For example topic CH60 has description "Are China and Taiwan developing any types of laser weapons?" Using the words 'China', 'laser weapons' in a GOOGLE search returns the url: http://www.freerepublic.com/forum/a363ee3c93414.htm which has an initial sentence:

> China's People's Liberation Army is building lasers to destroy satellites and already has beam weapons capable of damaging sensors on space-based reconnaissance and intelligence systems, according to a Pentagon report.

which was incorporated into the manual version of that query. While the precision for our manual run BRKECM1 of .8875 was better than one automatic run BRKECA1 (precision 0.3821), it lagged our other run BRKECA2 (precision 0.9500).

One query for which manual augmentation worked well was topic CH67 "Tiananmen Anniversary on Mainland" which a www news archive provided the following additional sentences:

> On June 21, the SCMP reported the detentions on June 19 of 5 dissidents in Hangzhou. The 5 are ZHU LUFU, HAN SHENDAI, WANG RONGQING, MAO QINGXIANG, and LI BAGEN. The last three have been detained several times already over the the past month or two. The 5 are members of the China Democracy Party. Information Centre of Human Rights and Democratic Movement in China says that over 180 CDP members have been arrested in the past month, and 31 are still in detention and awaiting trial.

and

> the Free China Movement describing the arrest and sentencing of ZHOU YONGJUN. Zhou snuck into China in December to visit his parents. Zhou was jailed for two years after the 1989 Tiananmen massacre and subsequent crackdown. After his release 7 years ago he was exiled.

The performance of this manual query increased ten-fold to 0.2009 over the median precision of .0290 and our automatic run precisions of 0.0026 and 0.0378.

Another query, CH79, "Livestock in China". A GOOGLE search "China livestock" yielded a url at Cornell University: http://usda.mannlib.cornell.edu/data-sets/international/90014/ which offered statistical information on China's agriculture production. Its descriptive sentences:

> Comprehensive data on Chinese animal agriculture including production of red meats, milk, eggs, poultry meats, and honey by region and province. Also includes inventory data on cattle, hogs, sheep, goats, and draft animals.

were added to the manual query. The performance of BRKECM1 for topic CH79 was 0.1496, almost three times better than our best automatic run BRKECA2 (0.0545).

Overall, the precision of the manual run over 25 topics was 0.1869. This was 28 percent better than the average of medians for topics and 10.2 percent better than our best automatic run (BRKECA1, overall precision 0.1680).

The use of web searches and direct cut-and-paste transfer of new query words and sentences made manual reformulation quite fast. Our estimate is that an average of 10 minutes per query was spent on manual rewrite, or slightly more than four hours total.

5.4 Experimental Results

We performed three English to Chinese cross-language retrieval runs. The title, description, and narrative fields were used in all three runs. For BRKECA1, the queries were translated into English by LDC dictionary lookup. The Chinese translation equivalents were then segmented into non-overlapping bigrams and unigrams. The evaluation result for the BRKECA1 run is presented in the third column in table 1. The evaluation results for BRKECA2 and BRKECM1 are presented in in column 4 and 5 in table 1. The Chinese translation equivalents for these two runs were segmented into words using the longest-matching method. And the segmented Chinese queries were searched against the test document collection which was also segmented into words using the same method. The best automatic English-

recall level	BRKCCA1 (MONO)	BRKECA1 (CLIR)	BRKECA2 (CLIR)	BRKECM1 (CLIR)
at 0.00	0.7079	0.4296	0.3603	0.5624
at 0.10	0.4697	0.3325	0.2828	0.3561
at 0.20	0.4047	0.2655	0.2071	0.2900
at 0.30	0.3720	0.2306	0.1852	0.2264
at 0.40	0.3225	0.1763	0.1555	0.1878
at 0.50	0.2769	0.1586	0.1393	0.1523
at 0.60	0.2445	0.1338	0.1269	0.1261
at 0.70	0.2165	0.1062	0.1052	0.1042
at 0.80	0.1874	0.0664	0.0892	0.0946
at 0.90	0.1368	0.0526	0.0833	0.0851
at 1.00	0.1155	0.0417	0.0721	0.0748
average precision	0.2936	0.1680	0.1543	0.1869
relevant retrieved	567	465	384	451
% of mono		57.22%	52.55%	63.66%

Table 1. Evaluation results for one Chinese monolingual run and three English to Chinese cross-language retrieval runs.

Chinese cross-language retrieval performance is only about 57% of the monolingual retrieval performance. For 5 out of the 25 topics, the precision for the cross-language retrieval is higher than that for the monolingual retrieval. On the other hand, for 10 out of the 25 topics, the precision for the cross-language retrieval is much lower than that for the monolingual retrieval. The main reason is that some key concept terms in those topics were either not translated at all due to the limited coverage of the bilingual wordlist we used or improperly translated. For example, the monolingual precision is .5406 for topic CH78, but the cross-language precision is only 0.0037 for the same topic. Topic CH78 is about motor vehicle fatalities in China. A key concept term ‘fatalities’ was not translated because it is missing in the LDC dictionary we used. The term ‘silk’ in topic 74 was translated into 绸, instead of the more appropriate term " 丝绸 ". For topic CH63, the noun phrase ‘energy source (能源)’ was translated into two Chinese words, 能 (energy) and 源 (source). The main concept term ‘three-links (三通)’ in topic CH70 were translated word-by-word into 三 (three) and 连接 / 相连 (link). Not being able to translate the term ‘industrially’ and mistranslating the term ‘developed’ in topic CH72 resulted in very lower precision in cross-language retrieval. The precision per topic for the monolingual run and the three English-Chinese cross-language runs are presented in table 2.

6 Conclusions

In summary, we performed three English-Chinese cross-language information retrieval runs, one manual and two automatic. We took a simple approach of translating queries into document language by dictionary lookup in our cross-language retrieval experiments. Even though the dictionary used in the BRKECA2 run is much larger than the one used in the BRKECA1 run, the retrieval performance for BRKECA2 is slightly worse than that for BRKECA1. We believe the inferior performance can be attributed to the simple selection method and to the difference in word usages. The performance of the best automatic run is only about 57% of the monolingual performance. The main performance-limiting factor is the limited coverage of the dictionary used in query translation. Some of the key concepts were either not translated or improperly translated.

Topic No	BRKCCA1 (MONO)	BRKECA1 (CLIR)	BRKECA2 (CLIR)	BRKECM1 (CLIR)
CH55	0.2200	0.1757	0.0973	0.1382
CH56	0.2814	0.1270	0.2293	0.1928
CH57	0.2939	0.1348	0.1435	0.1386
CH58	0.0036	0.0022	0.0089	0.0059
CH59	0.0015	0.0000	0.0000	0.0000
CH60	1.0000	0.3821	0.9500	0.8875
CH61	0.0000	0.0124	0.0445	0.0115
CH62	0.5000	0.0909	0.0032	0.0019
CH63	0.3009	0.0001	0.0114	0.1118
CH64	0.5354	0.3196	0.3128	0.3840
CH65	0.1797	0.7058	0.0453	0.0133
CH66	1.0000	0.8333	1.0000	1.0000
CH67	0.1327	0.0378	0.0026	0.2009
CH68	0.1865	0.0165	0.0066	0.0738
CH69	0.1497	0.2916	0.0329	0.0531
CH70	0.1687	0.0057	0.0001	0.0025
CH71	0.2604	0.1456	0.0467	0.2768
CH72	0.2910	0.0314	0.0755	0.1174
CH73	0.1966	0.3311	0.0004	0.0789
CH74	0.2655	0.0004	0.5286	0.3772
CH75	0.1413	0.2922	0.1102	0.2883
CH76	0.5065	0.2140	0.1460	0.1495
CH77	0.0434	0.0263	0.0029	0.0043
CH78	0.5406	0.0037	0.0033	0.0145
CH79	0.1417	0.0188	0.0565	0.1496

Table 2. Precision per topic for the monolingual run and three English-Chinese cross-language runs.

7 Acknowledgements

This research was supported by DARPA (Department of Defense Advanced Research Projects Agency) under research grant N66001-00-1-8911 as part of the DARPA Translingual Information Detection, Extraction, and Summarization Program (TIDES).

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 coefficie nts 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] X. Ge, W. Pratt, and P. Smyth. Discovering Chinese Words from Unsegmented Text. In *SIGIR'99, Berkeley, August 1999*, pages 271–272, 1999.

[6] KingSoft. http://ciba.kingsoft.net/ciba2000/cidian.htm.

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

[8] D. Oard and A. Diekema. *Cross-Language Information Retrieval*, volume 33, pages 223–256. 1998.

[9] M. Sun, D. Shen, and C. Huang. CSeg&Tag1.0: A Practical Word Segmenter and POS Tagger for Chines e Texts. In *Proceedings of the Fifth Applied Natural Language Processing Co nference*, pages 119–126, 1997.