Multilingual Information Retrieval Using English and Chinese Queries

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

Abstract

We participated in the CLEF 2001 monolingual, bilingual, and multilingual tasks. Our interests in these tasks are to test the utility of applying Chinese word segmentation algorithms to German decompounding, to experiment with techniques for combining translations from diverse resources, and to experiment with different approaches to multilingual retrieval. This paper describes our retrieval experiments.

1 Introduction

At CLEF 2001, we participated in the monolingual, bilingual, and multilingual tasks. Our interest in monolingual task is to test the idea of treating the German decompounding problem as that of Chinese word segmentation and applying Chinese word segmentation algorithms to split German compounds into their constituent words. Our interest in cross-language is to experiment with techniques for combining translations from diverse resources. We are also interested in different approaches to the multilingual retrieval task and various strategies for merging intermediate results to produce a final ranked list of documents for a multilingual retrieval run. In our experiments, we used English and Chinese topics. In translating the topics into the document languages which are English, French, German, Italian, and Spanish, we used two machine translators, one bilingual dictionary, two parallel text corpora, and one Internet search engine.

We submitted several official runs to the multilingual, bilingual, and monolingual tasks and performed more unofficial runs. To differentiate the unofficial runs from the official ones, the IDs of the official runs are all in uppercase, and IDs of the unofficial runs are all in lowercase. The unofficial runs are those evaluated locally with the official release of the relevance judgments for CLEF 2001.

2 Document Ranking

The document ranking formula we used in all of our retrieval runs was Berkeley's TREC-2 formula [3]. The logodds of relevance of document D to query Q is given by

$$\log O(R|D,Q) = log\frac{P(R|D,Q)}{P(\overline{R}|D,Q)} = -3.51 + 37.4 * x_1 + 0.330 * x_2 - 0.1937 * x_3 + 0.929 * x_4$$

where $P(R|D,Q)$ is the probability of relevance of document D with respect to query Q, $P(\overline{R}|D,Q)$ is the probability of irrelevance of document D with respect to query Q. The four composite variables x_1, x_2, x_3, and x_4 are defined as follows: $x_1 = \frac{1}{\sqrt{n}+1}\sum_{i=1}^{n}\frac{qtf_i}{ql+35}$, $x_2 = \frac{1}{\sqrt{n}+1}\sum_{i=1}^{n}\log\frac{dtf_i}{dl+80}$, $x_3 = \frac{1}{\sqrt{n}+1}\sum_{i=1}^{n}\log\frac{ctf_i}{cl}$, $x_4 = n$, where n is the number of matching terms between a document and a query, qtf_i is the within-query frequency of the ith matching term, dtf_i is the within-document frequency of the ith matching term, ctf_i is the occurrence frequency in a collection of the ith matching term, ql is query length (number of terms in a query), dl is document length (number of terms in a document), and cl is collection length, i.e. the number of occurrences of all terms in a test collection. The relevance probability of document D with respect to query Q can be written as follows given the logodds of relevance. $P(R|D,Q) = \frac{1}{1+e^{-logO(R|D,Q)}}$ The documents are ranked in decreasing order by their relevance probability $P(R|D,Q)$ with respect to a query. The coefficients were determined by fitting training data to the logistic regression model using a statistical software package. We refer readers to reference [3] for more details.

3 Monolingual retrieval experiments

We present an algorithm to break up German compounds into their constituent words. We treat the German decompounding problem in the same way as the Chinese word segmentation problem which is to segment a string of characters into words. We applied the Chinese segmentation algorithm as described in section 4.1 to decompose German compound words. First, we created a base German word lexicon consisting of all the words, including compounds, found in the German collection for the multilingual task. The uppercase letters were changed to lower case. Second, we identify all possible ways to break up a compound into its constituent words found in the base German lexicon. Third, we compute the probabilities for all possible ways to break up a compound into its constituent words, and choose the segmentation of the highest probability. For example, a compound $c = a_1a_2a_3a_4a_5a_6$ may be split into either $c_1 = a_1a_2/a_3a_4/a_5a_6 = w_1w_2w_3$, or $c_2 = a_1a_2a_3/a_4a_5a_6 = w_4w_5$, where $w_1 = a_1a_2$, $w_2 = a_3a_4$, $w_3 = a_5a_6$, $w_4 = a_1a_2a_3$, and $w_5 = a_3a_4a_5$ are German words. The probability of splitting c into $w_1w_2w_3$ is computed as $p(c_1) = p(w_1w_2w_3) = p(w_1) * p(w_2) * p(w_3)$, and the probability of splitting c into w_4w_5 is estimated by $p(c_2) = p(w_4w_5) = p(w_4) * p(w_5)$. If $p(c_1)$ is larger than $p(c_2)$, then the compound c is split into the three words w_1, w_2, and w_3; otherwise it is split into the two words w_4 and w_5. As in Chinese word segmentation, the probability of a word is estimated by its relative frequency in the German document collection. That is, $p(w_i) = tf(w_i)/\sum_{k=1}^{n} tf(w_k)$, where $tf(w_i)$ is the number of times word w_i occurs in the collection, including the cases where w_i is a consituent word in compounds; and n is the number of unique words, including compounds, in the collection.

We submitted two official German monolingual runs labeled BK2GGA1 and BK2GGA2, and two official Spanish monolingual runs labeled BK2SSA1 and BK2SSA2. The first run used title, description, and narrative fields in the topics, while the second run used title and description only. The stopwords were removed from both documents and topics, compounds were split into their constituent words, then words were stemmed using the Muscat German stemmer. Both the compounds and their constituent words were kept in indexing. Both runs were carried out without query expansion. The results are in table 2. The monolingual runs for the other three languages were evaluated locally and the results are in table 1.

Run ID	bk2eea1	bk2ffa1	BK2GGA1	bk2iia1	BK2SSA1
Language	English	French	German	Italian	Spanish
Average Precision	0.5553	0.4743	0.4050	0.4370	0.5302
Overall Recall	95.33% (816/856)	98.84% (1198/1212)	92.63% (1973/2130)	95.83% (1194/1246)	95.06% (2561/2694)

Table 1. Monolingual IR performance.

Run ID	Topic Fields	Features	Overall Recall	Average Precision
BK2GGA1	T,D,N	+stemming, +decompounding	92.63%	0.4050
BK2GGA2	T,D	+stemming, +decompounding	88.31%	0.3683
bk2gga3	T,D,N	+stemming, –decompounding	90.94%	0.4074
bk2gga4	T,D,N	–stemming, +decompounding	89.81%	0.3594
bk2gga5	T,D,N	–stemming, –decompounding	88.12%	0.3673

Table 2. German monolingual retrieval performance. The total number of German relevant documents for 49 topics is 2130.

To provide a base for comparison, three additional runs whose labels are in lower case were carried out. The two official runs with three unofficial runs were summarized in table 2.

4 Bilingual retrieval experiments

In this section we will describe the pre-processing of the Chinese topics and translation of the Chinese topics into English.

4.1 Chinese topics preprocessing

We first break up a Chinese sentence into text fragments consisting of only Chinese characters. Generally there are many ways to segment a fragment of Chinese text into words. We segment Chinese texts in two steps. First, we examine all the possible ways to segment a Chinese text into words found in a Chinese dictionary. Second, we compute the probabilities of all the segmentations and choose the segmentation with the highest probability. The probability of a segmentation is the product of the probabilities of the words making up the segmentation. For example, let $S = C_1C_2 \ldots C_n$ be a fragment of Chinese text consisting of n Chinese characters. Suppose one of the segmentation for the Chinese text is $S_i = W_1W_2 \ldots W_m$, then the probability of this segmentation is computed as follows:

$$p(S_i) = p(W_1W_2 \ldots W_m) = \sum_{j=1}^{m} p(W_j) \tag{1}$$

and

$$p(W_j) = \frac{tf(W_j)}{\sum_{k=1}^{N} tf(W_k)} \tag{2}$$

where $tf(W_j)$ is the number of times the word W_j occurs in a Chinese corpus, and N is the number of unique words in the corpus. $p(W_j)$ is just the maximum likelihood estimate of the probability that the word W_j occurs in the corpus. For a Chinese text, we first enumerate all the possible segmentations with respect to a Chinese dictionary, then we compute the probability for each segmentation. The segmentation of the highest probability is chosen as the final segmentation for the Chinese text. We used the Chinese corpus of the English-Chinese CLIR track at TREC-9 for estimating word probabilities. The Chinese corpus is about 213 MB in size and consist of about 130,000 newspaper articles.

A commonly used Chinese segmentation algorithm is the longest-matching method which repeatly chops off the longest initial string of characters that appears in the segmentation dictionary until the end of the sentence. A major problem with the longest-matching method is that a mistake often leads to multiple mistakes immediately after the point where the mistake is made. All dictionary-based segmentation methods suffer from the out-of-vocabulary problem. When a new word is missing in the segmentation dictionary, it is often segmented into a sequence of single or two-character words. Based on this observation, we combine the consecutive single-character terms into one word after removing the stopwords from the segmented Chinese topics.

4.2 Chinese topics translation

The segmentation and de-segmentation of the Chinese topics result in a list of Chinese words for each topic. We translate the Chinese topic words into English using three resources: 1) a Chinese/English bilingual dictionary, 2) two Chinese/English parallel corpora, 3) a Chinese Internet search engine. First, we look up each Chinese word in a Chinese-English bilingual wordlist prepared by the Linguistic Data Consortium and publicly available from http://morph.ldc.upenn.edu/Projects/Chinese/. The wordlist has about 128,000 Chinese words, each paired with a set of English words. If a Chinese word has only one, two or three English translations, we retain them all, otherwise we choose the three translations that occur most frequently in the *Los Angeles Times* collection which is part of the document collections for the CLEF 2001 multilingual task.

We created a Chinese-English bilingual lexicon from two Chinese/English parallel corpora, the *Hong Kong News corpus* and the *FBIS corpus*. The Hong Kong News corpus consists of the daily Press Release of the Hong Kong Government in both Chinese and English during the period of from April, 1998 through March, 2001. The source Chinese documents and English documents are not paired. So for each Chinese document, we have to identify the corresponding English document. We first aligned the Hong Kong News corpus at the document level using the LDC bilingual wordlist. Then we aligned the documents at the sentence level. Unlike the Hong Kong News corpus, the Chinese documents and their English translations are paired in the FBIS corpus. The documents in the FBIS corpus are usually long, so we first aligned the parallel documents at the paragraph level, then at the sentence level. We adapted the length-based alignment algorithm proposed by Gale and Church [5] to align parallel English/Chinese text. We refer interested readers to the paper in [1] for more details.

From the aligned pairs of Chinese/English sentences, we created a Chinese/English bilingual lexicon based on co-occurrence of word pairs across the aligned sentences. We used the maximum likelihood ratio measure proposed by Dunning [4] to compute the association score between a Chinese word and an English word. The bilingual lexicon takes as input a Chinese word and returns a ranked list of English words. We looked up each Chinese topic word in this bilingual Chinese/English lexicon, and retained the top two English words.

For the Chinese words that are missing in the two bilingual lexicons, we submitted them one by one to Yahoo!China, a Chinese Internet search engine at http://chinese.yahoo.com. Each entry in the search result pages has one or two sentences that contain the Chinese word searched. For each Chinese word, we downloaded all the search result pages if there are fewer than 20 result pages, or the first 20 pages if there are more than 20 result pages. Each result page contains 20 entries. From the downloaded result pages for a Chinese word, we extracted the English words in parentheses that follow immediately after the Chinese word. If there are English words found in the first step, we keep all the English words as the translations of the Chinese word. And if the first step failed to extract any English words, we extracted the English words appearing after the Chinese words. If there are more than 5 different English translations extracted from the result pages, we keep the top three most frequent words in the translations. Otherwise we keep all English translations. We refer interested readers to the paper in [2] for more details. This technique is based on the observation that the original English proper nouns sometimes appear in parentheses immediately after the Chinese translation. This technique should work well for proper nouns which are often missing in dictionaries. For many of the proper nouns in the CLEF 2001 Chinese topics missing in both the LDC bilingual dictionary and the bilingual dictionary created from parallel Chinese/English corpora, we extracted their English translations from the Yahoo!China search results. The last step in translating Chinese words into English is to merge the English translations obtained from the three resources mentioned above and weight the English translation terms. We give an example to illustrate the merging and weighting of the English translation terms. If a Chinese word has three English translation terms e_1, e_2, and e_3 from the LDC bilingual dictionary; and two English translation terms e_2 and e_4 from the bilingual dictionary created from the parallel texts. Then the set of words e_1, e_2, e_3, e_2, e_4 constitute the translation of the Chinese word. There is no translation terms from the third resource because we submit a Chinese word to the search engine only when the Chinese word is not found in both bilingual dictionaries. Next we normalize the weight of the translation terms so that the sum of their weights is one unit. For the example, the weights are distributed among the four unique translation terms as follows: $e_1 = .2$, $e_2 = .4$, $e_3 = .2$, and $e_4 = .2$. Note that the weight for the term e_2 is twice of that for the other three terms because it came from both dictionaries. We believe a translation term appearing in both dictionaries are more likely to be the appropriate translation than the ones appearing in only one of the dictionaries. Finally we multiply the weight by the frequency of the Chinese word in the original topic. So if the Chinese word occurs three times in the topic, the final weights assigned to the English translation terms of the Chinese word are $e_1 = .6$, $e_2 = 1.2$, $e_3 = .6$, and $e_4 = .6$.

The English translations of the Chinese topics were indexed and searched against the LA Times collection. We submitted two Chinese-to-English bilingual runs, one using all three topics fields, and the other using title and description only. Both runs were carried out without pre-translation or post-translation query expansion. The documents and English translations were stemmed using the Muscat English stemmer. The performance of these two runs are summarized in table 3. The results of the cross-language runs from English to the other four languages are in table 4, and the results of the cross-language runs from Chinese to all five document languages are in table 5.

Run ID	Topic Fields	Translation Resources	Overall Recall	Average Precision
BK2CEA1	T,D,N	dictionary, parallel texts, search engine	755/856	0.4122
BK2CEA2	T,D	dictionary, parallel texts, search engine	738/856	0.3683

Table 3. Chinese to English bilingual retrieval performance.

Run ID	Topic Fields	Topic Language	Document Language	Translation Resources	Overall Recall	Average Precision	% Monolingual Performance
bk2efa1	T,D,N	English	French	Systran+L&H Power	1186/1212	0.4776	100.7%
bk2ega1	T,D,N	English	German	Systran+L&H Power	1892/2130	0.3789	93.56%
bk2eia1	T,D,N	English	Italian	Systran+L&H Power	1162/1246	0.3934	90.02%
bk2esa1	T,D,N	English	Spanish	Systran+L&H Power	2468/2694	0.4703	88.70%

Table 4. Bilingual IR performance.

Run ID	Topic Fields	Topic Language	Document Language	Overall Recall	Average Precision	%Monolingual Performance
BK2CEA1	T,D,N	Chinese	English	755/856	0.4122	74.23%
bk2cfa1	T,D,N	Chinese	French	1040/1212	0.2874	60.59%
bk2cga1	T,D,N	Chinese	German	1605/2130	0.2619	64.67%
bk2cia1	T,D,N	Chinese	Italian	1004/1246	0.2509	57.41%
bk2csa1	T,D,N	Chinese	Spanish	2211/2694	0.2942	55.49%

Table 5. Bilingual IR performance.

5 Multilingual retrieval

We participated in the multilingual task using both English and Chinese topics. Our main approach was to translate the source topics into the document languages which are English, French, German, Italian, and Spanish, perform retrieval runs separately for each language, then merge the individual results for all five document languages into one ranked list of documents. We created a separate index for each of the five document collections by language. The stopwords were removed, words were stemmed using Muscat stemmers, and all uppercase letters were changed to lower case. The topics were processed in the same way.

For the multilingual retrieval experiments using English topics, we translated the English topics directly into French, German, Italian, and Spanish using both Systran translator and L&H Power translator. The topic translations of the same language from both translators were combined by topic, and then searched against the document collection of the same language. So for each multilingual retrieval run, we had five ranked list of documents, one for each document language. The five ranked lists of documents were merged to produce the final ranked list of documents for each multilingual run.

Our merging strategy is to combine all five intermediate runs and rank the documents by adjusted weights. Before we merge the intermediate runs, we made two adjustments to the estimated probability of document relevance in the intermediate runs. First, we reduced the estimated probability of document relevance by 20% (i.e, multiplying the original probability by .8) for the English documents retrieved using the un-translated English source topics. Then we added a value of 1.0 to the estimated probability of relevance for the top-ranked 50 documents in all monolingual runs. After these two adjustments to the estimated probability, we combined all five intermediate runs, sorted the combined results by adjusted probability of relevance, then took the top-ranked 1000 documents for each topic to create the final ranked list of documents. The aim of making the first adjustment is to make the estimated probability of relevance for all document languages comparable. Since translating topics from the source language to a target language probably introduces information loss to some degree, the estimated probability of relevance for the same topic may be slightly underestimated for the target language. In order to make the estimated probabilities for the documents retrieved using the original topics and using the translated topics comparable, the estimated probabilities for the documents retrieved using the original topics should be slightly lowered. The intention of making the second adjustment is to make sure that the top-ranked 50 documents in each of the intermediate results will be among the top-ranked 250 documents in the final ranked list.

For the multilingual retrieval experiments using Chinese topics, we translated the Chinese topics word by word into English, French, German, Italian, and Spanish in two stages. First, we translated the Chinese topics into English using three resources: 1) a bilingual dictionary, 2) two parallel corpora, and 3) one Chinese search engine. The procedure of translating Chinese topics into English was described in section 4. The English translations from the source Chinese topics consist of not sentences but words. Second, we translated the English words into French, German, Italian, and Spanish using both Systran translator and L&H translator for lack of resources to directly translate the Chinese topics into these languages. The rest is the same as for multilingual experiments using English topics.

We submitted four official multilingual runs, two using English topics and two using Chinese topics. The official runs are summarized in table 6. The multilingual run labeled BK2MUEAA1 was produced by combining the monolingual run bk2eea1 (.5553), and four cross-language runs bk2efa1 (.4776), bk2ega1 (.3789), bk2eia1 (.3934), bk2esa1 (.4703). The multilingual run labeled BK2MUCAA1 was produced by combining five cross-language runs, BK2CEA1, bk2cfa1, bk2cga1, bk2cia1, and bk2csa1. The performance of these five cross-language runs using Chinese topics is presented in table 5.

The problem of merging multiple runs into one is closely related to the problem of calibrating the estimated probability of document relevance and the problem of estimating the number of relevant documents with respect

Run ID	Topic Language	Topic Fields	Overall Recall	Average Precision
BK2MUEAA1	English	T,D,N	5953/8138	0.3424
BK2MUEAA2	English	T,D	5686/8138	0.3029
BK2MUCAA1	Chinese	T,D,N	4738/8138	0.2217
BK2MUCAA2	Chinese	T,D	4609/8138	0.1980

Table 6. Multilingual retrieval performance.

to a given query in a collection. If the estimated probability of document relevance is well calibrated, that is, the estimated probability is close to the true probability of relevance, then it would be trivial to combine multiple runs into one, since all one needs to do will be to combine the multiple runs and re-rank the documents in the estimated probability of relevance. If the number of relevant documents with respect to a given query could be well estimated, then one could take the number of documents from each individual run that is proportional to the number of estimated relevant documents in each collection. Unfortunately neither one of the problems is easy to solve.

Since merging multiple runs is not an easy task, an alternative approach to this problem is to work on it indirectly, that is, transform it into another problem that may be easier to solve. There are two alternative approaches to the problem of multilingual information retrieval. The first method works by translating the source topics into all document languages, combining the source topics and their translations in document languages, and then searching the combined, multilingual topics against a single index of documents in all languages. The second method works by translating all documents into the query language, then performing monolingual retrieval against the translated documents which are all in the same language as that of the query.

We applied the first alternative method to the multilingual IR task. We translated the source English topics directly into French, German, Italian, and Spanish using both Systran translator and L&H Power translator. Then we combined the English topics with the other four translations of both translators into one set of topics. The within-query term frequency is reduced by half. We used the multilingual topics for retrieval against a single index of all documents. The performance of this run labeled bk2eaa4 is shown in table 7. For lack of resources, we were not able to apply the second alternative method. Instead, we experimented with the method of translating the French, Italian, German, and Spanish documents retrieved in the intermediate runs back into English, and then carring out a monolingual retrieval run. We did not use Systran translator or L&H Power translator to translate the retrieved documents into English. We compiled a wordlist from the documents retrieved, then submitted the wordlist into Systran. The translation results of the wordlist were used to translate word by word the retrieved documents into English. The overall precision is .3648 for this run labeled bk2eaa5.

Run ID	Topic Language	Topic Fields	Overall Recall	Average Precision
bk2eaa3	English	T,D,N	5551/8138	0.3126
bk2eaa4	English	T,D,N	5697/8138	0.3648

Table 7. Multilingual IR performance.

6 Conclusion

We have tested the idea of treating the German decompounding problem in the same way as the Chinese word segmentation problem. The decompounding of German compound words did not improve precision. We believe the problem is that the decompounding algorithm failed to consistently decompose German compounds into their consitituent words. We observed that multi-word compounds are sometimes split into single words and shorter compounds. We also presented a method for combining translations from three different translation resources which seems to work well. We experimented with three approaches to multilingual retrieval. The method of translating the documents retrieved in the intermediate runs back into the language of the source topics, and then carring out monolingual retrieval achieved better precision than the other two methods.

7 Acknowledgements

This research was supported by DARPA (Department of Defense Advanced Research Projects Agency) under research contract N66001-97-8541; AO# F477: Search Support for Unfamiliar Metadata Vocabularies within the DARPA Information Technology Office.

References

[1] A. Chen, F. Gey, and H. Jiang. Alignment of english-chinese parallel corpora and its use in cross-language information retrieval. In *19th International Conference on Computer Processing of Oriental Languages*, pages 251–257, Seoul, Korea, May 14-16 2001.

[2] A. Chen, H. Jiang, and F. Gey. Combining multiple sources for short query translation in chinese-english cross-language information retrieval. In *Proceedings of the Fifth International Workshop on Information Retrieval with Asian Languages*, pages 17–23, Hong Kong, Sept. 30-Oct 1 2000.

[3] 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.

[4] T. Dunning. Accurate methods for the statistics of surprise and coincidence. *Computational linguistics*, 19:61–74, March 1993.

[5] W. A. Gale and K. W. Church. A program for aligning sentences in bilingual corpora. *Computational linguistics*, 19:75–102, March 1993.