# 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.** The University of California at Berkeley group two participated in the CLEF 2001 monolingual, bilingual, and multilingual retrieval tasks. In this paper, we present a German decompounding procedure and a method of combining multiple translation resources for translating Chinese into English. We also report on our experiments with three different approaches to multilingual retrieval.

## 1 Introduction

At CLEF 2001, we participated in the monolingual, bilingual, and multilingual tasks. Our main 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 component words. Our interest in cross-lingual retrieval is to experiment with techniques for combining translations from diverse translation 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. For 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.

Several official runs were submitted for the monolingual, bilingual, and multilingual tasks, and more unoffical runs were performed and evaluated locally. To differentiate the unofficial runs from the official ones, the IDs of the official runs are all in uppercase while the IDs of the unofficial runs are all in lowercase. The unofficial runs are those evaluated locally using the official relevance judgments for CLEF 2001.

## 2 Test Collection

The document collection for the multilingual IR task consists of documents in five languages: English, French, German, Italian, and Spanish. The collection has about 750,000 documents which are newspaper articles published in 1994

**Table 1.** Document collection for the multilingual task.

| Name | Language | No. of Documents | Size (MB) | Year(s) |
|---|---|---|---|---|
| Los Angeles Times | English | 113,005 | 425 | 1994 |
| Le Monde | French | 44,013 | 157 | 1994 |
| SDA French | French | 43,178 | 86 | 1994 |
| Frankfurter Rundschau | German | 139,715 | 320 | 1994 |
| Der Spiegel | German | 13,979 | 63 | 1994/95 |
| SDA German | German | 71,677 | 144 | 1994 |
| La Stampa | Italian | 58,051 | 193 | 1994 |
| SDA Italian | Italian | 50,527 | 85 | 1994 |
| EFE | Spanish | 215,738 | 509 | 1994 |

except that part of the *Der Spiegel* was published in 1995. The distribution of documents among the five document languages is presented in Table 1. A set of 50 topics was developed initially in Dutch, English, French, German, Italian, and Spanish. Then the topics were translated into Chinese, Finnish, Japanese, Swedish, Thai, and Russian. A topic has three parts: 1) *title*, a short description of information need; 2) *description*, a sentence-long description of information need; and 3) *narrative*, specifying document relevance criteria. The multilingual IR task at CLEF 2001 was concerned with searching the collection consisting of English, French, German, Italian, and Spanish documents for relevant documents, and returning a combined, ranked list of documents in any document language in response to a query. The bilingual IR task was concerned with searching a collection of documents using queries in a different language. We used the English and Chinese topic sets in our participation in the multilingual IR task; Chinese topics in the bilingual IR task; and German and Spanish topic sets in the monolingual IR task.

## 3 Document Ranking

We used a logistic regression-based document ranking formula developed at Berkeley [1] to rank documents in response to a query. The log odds of relevance of document $D$ with respect to query $Q$ , denoted by $\log O(R|D,Q)$, is given by

$$\begin{aligned}\log O(R|D,Q) &= log\frac{P(R|D,Q)}{P(\overline{R}|D,Q)} && (1)\\ &= -3.51 + 37.4 * x_1 + 0.330 * x_2 - 0.1937 * x_3 + 0.929 * x_4 && (2)\end{aligned}$$

where $P(R|D,Q)$ is the probability that document $D$ is relevant to query $Q$, $P(\overline{R}|D,Q)$ the probability that document $D$ is irrelevant 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 $i$th matching term, $dtf_i$ is the within-document frequency of the $i$th matching term, $ctf_i$ is the occurrence frequency in a collection of the $i$th matching term, $ql$ is query length (i.e., number of terms in a query), $dl$ is document length (i.e., number of terms in a document), and $cl$ is collection length (i.e., number of terms in a test collection). The relevance probability of document $D$ with respect to query $Q$ can be written as follows, given the log odds 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 the logistic regression model specified in Eqn. 2 to training data using a statistical software package. We refer readers to reference [1] for more details. All retrieval runs were performed without query expansion using blind (also called pseudo) relevance feedback technique.

## 4 Monolingual Retrieval Experiments

### 4.1 German Decompounding

We chose German topics for the monolingual task to study the effect of decomposing German compounds into their component words on the retrieval performance. We present an algorithm to break up German compounds into their component 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 5.1 to decompose German compound words. The German decompounding procedure consists of three steps. First, we create a German lexicon consisting of all the words, including compounds, found in the CLEF 2001 German collection for the multilingual task. The uppercase letters are changed to lower case. Second, we identify all possible ways to break up a compound into its component words found in the German lexicon. Third, we compute the probabilities for all possible ways to break up a compound into its component words, and choose the decomposition of the highest probability. For example, the German compound *Mittagessenzeit* has three decompositions: 1) *Mittag*, *Essen*, *Zeit*; 2) *Mittagessen*, *Zeit*; and 3) *Mittag*, *Essenzeit*. The probability of the first decomposition is computed as p(*Mittag*)*p(*Essen*)*p(*Zeit*), the probability of the second decomposition is p(*Mittagessen*)*p(*Zeit*), and the probability of the third decomposition is p(*Mittag*)*p(*Essenzeit*). If the second decomposition has the highest probability, then the compound is decomposed into *Mittagessen*, *Zeit*. As in Chinese word segmentation, the probability of a word is estimated by its relative frequency in the German document collection. That is,

$$p(w_i) = \frac{tf(w_i)}{\sum_{k=1}^{n} tf(w_k)},$$

where $tf(w_i)$ is the number of times word $w_i$ occurs in the collection, including the cases where $w_i$ is a component word in compounds; and $n$ is the number of unique words, including compounds, in the collection. When computing the occurrence frequency of a word in the collection for the purpose of decompounding, we not only count the cases where the word occurs alone, but also the cases where the word occurs as a component word of some compounds. Consider the example *Mittagessenzeit* again. It is considered that all of its component words, *Mittagessen*, *Mittag*, *Essen*, *Zeit*, *Essenzeit*, occur once in the compound. In decompounding, the component words consisting of three or fewer letters were not considered.

We submitted two German monolingual runs labeled BK2GGA1 and BK2GGA2, respectively, and two Spanish monolingual runs labeled BK2SSA1 and BK2SSA2, respectively. The first run for both languages used title, description, and narrative fields in the topics, while the second run for both languages used title and description only. The stopwords were removed from both documents and topics, compounds were split into their component words using the decompounding procedure described above, then words were stemmed using Muscat German stemmer[1]. The compounds and their component words were kept in both documents and topics indexing. The Spanish words were stemmed using Muscat Spanish stemmer. The monolingual retrieval evaluation results for the official runs and additional unofficial runs for other languages are presented in Table 2. The monolingual runs were performed using the original, untranslated topics. There are 50 topics for each document language. There are in total 856 relevant English documents, 1,212 relevant French documents, 1,246 relevant Italian documents, 2,130 relevant German documents, and 2,694 relevant Spanish documents. To provide a baseline for comparison, four additional German monolingual retrieval runs with different features were carried out. The results for the four unofficial runs whose IDs are in lower case together with one official run are presented in Table 3. For the three unofficial runs, bk2gga3,bk2gga4,bk2gga5, and our two official runs, BK2GGA1, BK2GGA2, the German compounds and their component words were retained in both query and document indexing. But for the run labeled bk2gga6, only the component words of German compounds were retained in the topic indexing while both compounds and their component words were retained in document indexing.

[1] http://open.muscat.com

**Table 2.** Monolingual IR performance.

| Run ID | Language | Topic Fields | Average Precision | Overall Recall |
|---|---|---|---|---|
| BK2GGA1 | German | T,D,N | 0.4050 | 1973/2130 |
| BK2GGA2 | German | T,D | 0.3551 | 1881/2130 |
| BK2SSA1 | Spanish | T,D,N | 0.5302 | 2561/2694 |
| BK2SSA2 | Spanish | T,D | 0.5225 | 2526/2694 |
| bk2eea1 | English | T,D,N | 0.5553 | 816/856 |
| bk2eea2 | English | T,D | 0.5229 | 820/856 |
| bk2ffa1 | French | T,D,N | 0.4743 | 1198/1212 |
| bk2ffa2 | French | T,D | 0.4676 | 1190/1212 |
| bk2iia1 | Italian | T,D,N | 0.4370 | 1194/1246 |
| bk2iia2 | Italian | T,D | 0.4527 | 1200/1246 |

**Table 3.** German monolingual retrieval performance. Both compounds and their component words were retained in topic indexing for BK2GGA1, but only the component words of the compounds were retained in topic indexing for bk2gga6.

| Run ID | Topic Fields | Features | Overall Recall | Average Precision |
|---|---|---|---|---|
| BK2GGA1 | T,D,N | +stemming, +decompounding, –expansion | 92.63% | 0.4050 |
| bk2gga3 | T,D,N | +stemming, –decompounding, –expansion | 90.94% | 0.4074 |
| bk2gga4 | T,D,N | –stemming, +decompounding, –expansion | 89.81% | 0.3594 |
| bk2gga5 | T,D,N | –stemming, –decompounding, –expansion | 88.12% | 0.3673 |
| bk2gga6 | T,D,N | +stemming, +decompounding, –expansion | 94.60% | 0.4436 |

The results in Table 3 shows that decompounding degraded slightly the German monolingual retrieval performance when both the compounds and their component words were retained in topic indexing. However, when only the component words of the compounds were kept in topic indexing, decompounding improved the average precision by 8.88%.

### 4.2 More Experiments with German Decompounding

We present another German decompounding procedure which is a variant of the decompounding procedure described in section 4.1. The procedure is described as follows:

1. Create a German base dictionary consisting of single words only (compounds are excluded).
2. Decompose a compound into its components words found in the German base dictionary.
3. Choose the decomposition with the smallest number of component words. If there are two or more decompositions having the smallest number of component words, then choose the decomposition of the highest probability.

For example, when the German base dictionary contains *film, fest, fests, festspiele, piele, s* and others, the German compound *filmfestspiele* can be decomposed into component words with respect to the base dictionary in four different ways as shown in Table 4. The last decomposition has the smallest number of component words, so the German compound *filmfestspiele* is split into *film* and *festspiele.* Table 5 presents another example which shows the decompositions of German compound *hungerstreiks* with respect to a base dictionary containing *erst, hung, hunger, hungers, hungerst, reik, reiks, s, streik, streiks*, and other words. The compound *hungerstreiks* has six decompositions with respect to the base dictionary. Because two decompositions have the smallest number of component words, the rule of selecting the decomposition with the smallest number of component words cannot be applied here. We have to compute the probability of the decomposition for the decompositions with the smallest number of component words. The last column in Table 5 shows the log of the decomposition probability for all six decompositions. According to the rule of selecting the decomposition of the highest probability, the fourth decomposition should be chosen as the decomposition of the compound *hungerstreiks.* That is, the compound *hungerstreiks* should be split into *hunger* and *streiks.* Consider the decomposition of compound $c$ into $n$ component words, $c = w_1 w_2 \dots w_n$. The probability of a decomposition is computed as follows:

**Table 4.** Decompositions of compound *filmfestspiele.*

| | Decompositions | | | |
|---|---|---|---|---|
| 1 | film | fest | s | piele |
| 2 | film | fest | spiele | |
| 3 | film | fests | piele | |
| 4 | film | festspiele | | |

**Table 5.** Decompositions of compound *hungerstreiks.*

| | Decompositions | | | | log p(D) |
|---|---|---|---|---|---|
| 1 | hung | erst | reik | s | -55.2 |
| 2 | hung | erst | reiks | | -38.0 |
| 3 | hunger | streik | s | | -38.7 |
| 4 | hunger | streiks | | | -21.4 |
| 5 | hungerst | reik | s | | -52.1 |
| 6 | hungerst | reiks | | | -34.9 |

$$p(c) = p(w_1)p(w_2)\dots p(w_n) = \prod_{i=1}^{n} p(w_i)$$

where the probability of component word $w$ is computed as follows:

$$p(w_i) = \frac{tfc(w_i)}{\sum_{j=1}^{N} tfc(w_j)}$$

where $tfc(w_i)$ is the number of occurrences of word $w_i$ in a collection, $N$ is the number of unique words, including compounds, in the collection. The occurrence frequency of a word is the number of times the word occurs alone in the collection. The frequency count of a word does not include the cases where the word is a component word of some compounds. Also, the base dictionary does not contain any words that are three-letter long or shorter except for the letter *s*. We created a German base dictionary of about 780,000 words by combining a lexicon extracted from Morphy, a german morphological analyzer [2], German wordlists found on the Internet, and German words in the CLEF-2001 German collection. In our implementation, we considered only the case where a compound is the concatenation of component words, including the single-letter word *s*. A component word may change its form when it is combined with other component words to create a compound word. For example, when the word *Erde* is combined with the word *Atmoshpäre* to create a compound, the compound is not *Erdeatmoshpäre*, but *Erdatmoshpäre*. The final letter *e* of the word *Erde* is removed from the compound. Note that the number of possible decompositions of a compound is determined by what is in the base dictionary. For example, when the word *mittagessen* is not in the base dictionary, the compound *mittagessenzeit* would be split into three component words *mittag*, *essen*, and *zeit*. We performed six German monolingual retrieval runs using three German test collections, CLEF-2001, CLEF-2000, and combined TREC-6/7/8. The CLEF-2001 test collection consists of 50 topics and about 225,000 documents, CLEF-2000 consists of 40 topics and about 154,000 documents, and combined TREC-6/7/8 consists of 73 topics with at least one relevant document and about 252,000 doc-

**Table 6.** German decompounding and monolingual retrieval performance. Both compounds and their component words were retained in document indexing, but only the component words were retained in query indexing. The numbers given in parenthesis are overall recall values.

| Test collections | -decompounding<br>-stemming<br>-query expansion | +decompounding<br>-stemming<br>-query expansion | Change |
|---|---|---|---|
| CLEF-2001 | .3673<br>(1877/2130) | .4314<br>(1949/2130) | +17.45% |
| CLEF-2000 | .3189<br>(673/821) | .4112<br>(770/821) | +28.94% |
| TREC-6/7/8 | .2993<br>(1907/2626) | .3368<br>(2172/2626) | +12.53% |

uments. Three of the runs were performed without German decompounding and the other three with German decompounding. All three topic fields were used in all six runs. Both the compounds and their component words were retained in document indexing, but only the component words of the compounds found in the topics were retained in topic indexing. The evaluation results for all six runs were presented in Table 6. German decompounding brought an improvement in overall precision of 17.45% for CLEF-2001 collection, 28.94% for CLEF-2000 collection, and 12.53% for combined TREC-6/7/8 collection.

The decompounding procedure differs from the one described in section 4.1 in two aspects. First, the base dictionary in the new procedure contains German words only (the compounds are excluded), while the base dictionary for the previous procedure contains both German words and German compounds. Second, the occurrence frequency of a word includes only the cases where the word occurs alone in the new procedure, while the occurrence frequency for the previous procedure consider both the cases where the word occurs alone and the cases where the word occurs as a component word of some compounds. The results presented in this and previous sections show that the new procedure is more effective.

## 5 Bilingual Retrieval Experiments

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

### 5.1 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 $s = c_1c_2 \dots c_n$ be a fragment of Chinese text consisting of $n$ Chinese characters. Suppose one of the segmentation for the Chinese text is $s_i = w_1w_2 \dots w_m$, where $m$ is the number of words resulted from the segmentation, then the probability of this segmentation is computed as follows:

$$p(s_i) = p(w_1w_2 \dots w_m) = \sum_{j=1}^{m} p(w_j) \tag{3}$$

and

$$p(w_j) = \frac{tf(w_j)}{\sum_{k=1}^{N} tf(w_k)} \tag{4}$$

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 consists 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. We will call this process the de-segmentation of the segmented text.

### 5.2 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, and 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.[2] The wordlist has about 128,000 Chinese words, each paired with a set of English words. If a Chinese word has three or fewer 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 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

[2] The wordlist is publicly available at http://morph.ldc.upenn.edu/Projects/Chinese/.

documents at the paragraph level, then at the sentence level. We adapted the length-based alignment algorithm proposed by Gale and Church [4] to align parallel English/Chinese text. We refer readers to the paper in [5] for more details.

From the aligned Chinese/English sentence pairs, we created a Chinese/English bilingual lexicon based on co-occurrence of word pairs in the aligned sentences. We used the maximum likelihood ratio measure proposed by Dunning [6] 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 publicly accessible 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 readers to the paper in [3] 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 $e_1, e_2, e_3, e_2, e_4$ constitutes 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 above 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 topic 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 7. The results of the cross-language runs from English to the other four languages are presented in Table 8, and the results of the cross-language runs from Chinese to all five document languages are in table 9. The translations of both the English topics and the Chinese topics into French, German, Italian, and Spanish are described in the next section.

**Table 7.** Chinese to English bilingual retrieval performance.

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

**Table 8.** Bilingual IR performance.

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

**Table 9.** 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% |

## 6 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 the combined topics were 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 by topic to produce the final ranked list of documents for each multilingual run.

The documents in the intermediate runs were ranked by their relevance probability estimated using Eqn. 2. Our merging strategy is to combine all five intermediate runs and rank the documents by adjusted relevance probability. Before we merge the intermediate runs, we made two adjustments to the estimated probability of relevance in the intermediate runs. First, we reduced the estimated probability of relevance by 20% (i.e, multiplying the original probability by .8) for the English documents retrieved using the original 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 by topic, 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 closely 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 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 5. 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 power translator for lack of resources to directly translate the Chinese topics into these languages. The merging strategy was the same as the one in 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 10. 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 9.

**Table 10.** Multilingual retrieval performance. The document languages are English, French, German, Italian, and Spanish.

| 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 |

The problem of merging multiple runs into one is closely related to the problem of calibrating the estimated probability of relevance and the problem of estimating the number of relevant documents with respect to a given query in a collection. If the estimated probability of 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 by 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 does not involve merging documents and so may be easier to solve. We describe 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 was reduced by half. We used the multilingual topics for retrieval against a single index of all documents. The performance of this run labeled bk2eaa3 is shown in Table 11. 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 translator and L&H Power translator. The translation results of the wordlist were used to translate word by word the retrieved documents in the intermediate runs into English. The overall precision is .3648 for this run labeled bk2eaa4.

**Table 11.** Multilingual IR performance.

| 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 |

## 7 Conclusions

We have presented a German decompounding procedure in two different versions. German decompounding did not improve overall precision when both the compounds and their component words were retained in topic indexing. However, when only the component words of the compounds found in the topics were retained in topic indexing, the German monolingual retrieval performance on the CLEF 2001 German collection improved 8.88% using the first version of the German decompounding procedure. The second version of the German decompounding improved the overall precision on the CLEF 2001 German collection by 17.45%. A German base dictionary of words, but not compounds, is required for applying the second version. We also presented a method for translating Chinese topics into English by 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 the best precision.

## 8 Acknowledgements

This research was supported by DARPA under research contract N66001-97-8541; AO# F477: Search Support for Unfamiliar Metadata Vocabularies , PI: Michael Buckland. The author would like to thank Fred Gey for translating the English topics into other document languages using L&H Power Translator, Hailing Jiang for downloading the Muscat stemmers, and Michael Buckland for helpful comments.

## References

1. Cooper, W. S., Chen, A., Gey F.: Full Text Retrieval based on Probabilistic Equations with Coefficients fitted by Logistic Regression. In: Harman, D. K. (ed.): The Second Text REtrieval Conference (TREC-2) (1994) 57-66.
2. Lezius, W., Rapp R., Wettler M.: A Freely Available Morphological Analyzer, Disambiguator and Context Sensitive Lemmatizer for German. In: Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (COLING-ACL'98), Montreal, Canada, August 10-14, 1998. pp.743-748.
3. Chen, A., Jiang H., Gey, F.: 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, Hong Kong, Sept. 30-Otc 1, 2000. pp.17-23.
4. Gale, W. A., Church, K. W.: A Program for Aligning Sentences in Bilingual Corpora. In: Computational linguistics. **19** (1993) 75-102.
5. Chen, A., Gey, F., Jiang H.: Alignment of English-Chinese Parallel Corpora and its Use in Cross-Language Information Retrieval. In: 19th International Conference on Computer Processing of Oriental Languages, Seoul, Korean, May 14-16, 2001. pp. 251-257.
6. Dunning, T.: Accurate Methods for the Statistics of Surprise and Coincidence. In: Computational linguistics. **19** (1993) 61-74.