Phrase Discovery for English and Cross-language Retrieval at TREC-6

Fredric C. Gey and Aitao Chen
UC Data Archive & Technical Assistance (UC DATA)
gey@ucdata.berkeley.edu
aitao@sims.berkeley.edu
University of California at Berkeley, CA 94720

January 20, 1998

Abstract

Berkeley's experiments in TREC-6 center around phrase discovery in topics and documents. The technique of ranking bigram term pairs by their expected mutual information value was utilized for English phrase discovery as well as Chinese segmentation. This differentiates our phrase-finding method from the mechanistic one of using all bigrams which appear at least 25 times in the collection. Phrase finding presents an interesting interaction with stop words and stop word processing. English phrase discovery proved very important in a dictionary-based English to German cross language run. Our participation in the filtering track was marked with an interesting strictly Boolean retrieval as well as some experimentation with maximum utility thresholds on probabilistically ranked retrieval.

1 Introduction

Berkeley's participation in the TREC conferences has provided a venue for experimental verification of the utility of algorithms for probabilistic document retrieval. Probabilistic document retrieval attempts to place the ranking of documents in response to a user's information need (generally expressed as a textual description in natural language) on a sound theoretical basis. The approach is, fundamentally, to apply Bayesian inference to develop predictive equations for probability relevance where training data is available from past queries and document collections. Berkeley's particular approach has been to use the technique of logistic regression. Logistic regression has by now become a standard technique in the discipline of epidemiology for discovering the degree to which causal factors result in disease incidence [8, Hosmer and Lemeshow,89]. In document retrieval the problem is turned around, and one wishes to predict the incidence of a rare disease called 'relevance' given the evidence of occurrence of query words and their statistical attributes in documents.

In TREC-2 [3] Berkeley introduced a formula for ad-hoc retrieval which has produced consistently good retrieval results in TREC-2 and subsequent TREC conferences TREC-4 and TREC-5.

The logodds of relevance of document D to query Q is given by

$$\log O(R|D,Q) = -3.51 + \frac{1}{\sqrt{N}+1}\Phi + 0.0929 * N \tag{1}$$

$$\Phi \quad = \quad 37.4 \sum_{i=1}^{N} \frac{qtf_i}{ql+35} + 0.330 \sum_{i=1}^{N} \log \frac{dtf_i}{dl+80} - 0.1937 \sum_{i=1}^{N} \log \frac{ctf_i}{cf} \tag{2}$$

where

N is the number of terms common to both query and document,
qtf_i is the occurrence frequency within a query of the ith match term,
dtf_i is the occurrence frequency within a document of the ith match term,
ctf_i is the occurrence frequency in a collection of the ith match term,
ql is query length (number of terms in a query),
dl is document length (number of terms in a document), and
cf is collection length, i.e. the number of occurrences of all terms in a test collection.

The summation in equation (2) is carried out over all the terms common to query and document.

This formula has also been used, with equal success, in document retrieval with Chinese and Spanish queries and document collections of the past few TREC conferences. We utilized this identical formula for German queries against German documents in the cross-language track for TREC-6.

Berkeley's approach, in the past, has been to concentrate on fundamental algorithms and not attempt refinements such as phrase discovery or passage retrieval. However in doing further research in the area of Chinese text segmentation [2] we applied a technique from computational linguistics which seemed to show promise for rigorous discovery of phrases from statistical evidence based upon word frequency and word co-occurrence in document collections. Thus for TREC-6 we have begun the investigation of how to obtain and use phrases within the context of probabilistic document retrieval.

2 Phrase discovery using expected mutual information

The usual method at TREC (by many other groups) for choosing phrases has been to mechanistically choose all two word combinations which occur more than 'n' times in the collection (where n=25 has been the usual threshold). Other groups have used natural language processing techniques (rule and dictionary-based) to parse noun phrases. Berkeley's approach for TREC-6 was compute the mutual information measure between word combinations using individual and word co-occurrence frequency statistics:

$$MI(t_1, t_2) = \log_2 \frac{P(t_1, t_2)}{P(t_1)P(t_2)}$$

High values of this measure indicate a positive association between words. Near zero values indicate probabilistic independence between words. Values less than zero indicate a negative correlation between words (i.e. if one word occurs, the other word is not likely to occur next to it). Our experiments indicated that values of MI greater than 10 almost always identified proper nouns such as (for TREC topic 001 in routing) 'Ivan Boeski' and 'Michael Milkin'. This technique identifies important phrases such as 'unfriendly merger' which occur only 5 times in the collection. Berkeley used a cutoff of MI = 3.00. However, when both of the component words are commonly occurring words, the expected mutual information value will be a small value. In this case the mutual information technique may fail to identify high frequency phrases (such as 'educational standard' with MI = 1.70 which occurs 399 times in the 5 TREC disks).

Phrase discovery has an important interaction with stopword processing. For TREC-6 adhoc topic 340, the title query 'Land Mine Ban' processes to 'land' and 'ban' because 'mine' is a stopword. Interestingly this does not affect the Description Field for that topic which contains the phrase 'land mines' which stems to 'land mine'. Berkeley chose to identify phrases before stopword processing. This produces other interesting phrases such as 'for example' and 'e g', although they may not be particularly discriminating. Because we made this processing decision after examining the parsing of the title for topic 340, we did not submit a short title run for TREC-6. We do, however, include a short title result below for comparison purposes.

Another important question is whether to retain the individual word components of phrases or to remove them. Our experiments indicate that performance deteriorates upon removal of individual word components of phrases, at least for ad-hoc retrieval.

3 Ad-hoc Experiments

Berkeley's ad-hoc runs for TREC-6 utilized the new phrase discovery method as well as a new formula to incorporate phrases into probabilistic training. Our decision to modify the TREC-2 formula was based upon the observation that phrases have a very different pattern of occurrence in the collections than individual terms. The principle thrust of the change was to separate out a component which utilized the statistical clues for phrases as distinct from one which used single term statistical attributes. After training using logistic regression on relevance judgments for disks 1-4, the formula was as follows:

The logodds of relevance of document D to query Q is given by

$$\log O(R|D,Q) = -3.9912 + \frac{1}{\sqrt{N_t}+1}\Phi_t + 0.1281 * N_t + \frac{1}{\sqrt{N_p}+1}\Phi_p - 0.3161 * N_p \tag{3}$$

$$\Phi_t = 36.5904 \sum_{i=1}^{N_t} \frac{qtf_i}{ql_t + 35} + 0.3938 \sum_{i=1}^{N_t} \log \frac{dtf_i}{dl_t + 80} - 0.2147 \sum_{i=1}^{N_t} \log \frac{ctf_i}{cf_t} \tag{4}$$

$$\Phi_p = 6.5743 \sum_{i=1}^{N_p} \frac{qpf_i}{ql_p + 10} + 0.0959 \sum_{i=1}^{N_p} \log \frac{dpf_i}{dl_p + 25} - 0.1182 \sum_{i=1}^{N_p} \log \frac{cpf_i}{cf_p} \tag{5}$$

where

N is the number of terms common to both query and document,
qtf_i is the occurrence frequency within a query of the ith match term,
dtf_i is the occurrence frequency within a document of the ith match term ,
ctf_i is the occurrence frequency in a collection of the ith match term, ql_t is query length (number of single terms in a query),
dl_t is document length (number of single terms in a document), and
cf_t is collection length, i.e. the number of occurrences of all single terms in a test collection.
qpf_i is the occurrence frequency within a query of the ith match phrase,
dpf_i is the occurrence frequency within a document of the ith match phrase ,
cpf_i is the occurrence frequency in a collection of the ith match phrase, ql_p is query length (number of phrases in a query),
dl_p is document length (number of phrases in a document), and
cf_p is collection length, i.e. the number of occurrences of all phrases in a test collection.

Run	Brkly21	Brkly22	Brkly23	Title	Words
Formula	TREC-6	TREC-6	TREC-2	TREC-2	TREC-2
Query	Description	Long	Manual	Title	Long
Phrase	Yes	Yes	Yes	Yes	No
Expansion	Yes	Yes	No	No	No
Overall	0.1376	0.2021	0.2282	0.2102	0.2054
0.00	0.5668	0.7105	0.6558	0.6001	0.7191
0.10	0.3298	0.4449	0.4885	0.4442	0.4416
0.20	0.2333	0.3456	0.3745	0.3337	0.3505
0.30	0.1802	0.2704	0.3128	0.2753	0.2747
0.40	0.1399	0.2218	0.2623	0.2425	0.2164
0.50	0.1160	0.1903	0.2182	0.2025	0.1721
0.60	0.0963	0.1398	0.1677	0.1619	0.1365
0.70	0.0611	0.1029	0.1193	0.1234	0.1070
0.80	0.0209	0.0472	0.0604	0.0843	0.0704
0.90	0.0070	0.0226	0.0221	0.0526	0.0195
1.00	0.0030	0.0119	0.0126	0.0475	0.0089
relevant	1615	2547	2583	2321	2382
5 docs	0.3680	0.4680	0.4880	0.3720	0.4400
10 docs	0.2940	0.4080	0.4320	0.3500	0.3880
15 docs	0.2680	0.3693	0.3813	0.3227	0.3520
20 docs	0.2450	0.3440	0.3510	0.3050	0.3340
30 docs	0.2160	0.3053	0.3140	0.2767	0.2960
100 docs	0.1286	0.1932	0.2112	0.1912	0.2000
200 docs	0.0878	0.1365	0.1433	0.1306	0.1385
500 docs	0.0502	0.0800	0.0816	0.0762	0.0786
1000 docs	0.0323	0.0509	0.0517	0.0464	0.0476
R-Precision	0.1675	0.2422	0.2612	0.2419	0.2500

Table 1: TREC-6 Adhoc Results

The summation in equations (4) and (5) is carried out over all the terms or phrases common between query and document.

The size of the training matrix produced was 3,812,933 observations. The normalization by collection length (single terms and phrases) was done by counting total occurrences of all single terms/pairs in the collection. These are:

158,042,364 single terms
34,018,769 pairs

Our official runs were Brkly21 (long topic run) Brkly22 (description field run) and Brkly23 (manual query reformulation). As can be seen from the table the description field run was significantly below the long topic run, continuing a pattern begun in TREC-5. Our unofficial run on the title field produced almost equivalent performance to the long field, attributable to the precision by which titles capture the essential meaning of the topics. We also ran a long query run using only the TREC-2 formula, and were dismayed to find that the phrase formula failed to improve upon single terms. It seems that phrases, which offer significantly more precise capture of topic meaning, have yet to be exploited properly by our probabilistic training.

4 Routing Experiments

Berkeley's routing runs for TREC-6 follow on the spirit of our routing runs of TREC-5. In all routing methodology the key problem is to choose additional terms to add to each query based upon documents found to be relevant in previous TREC runs. Several measures have been proposed to choose such terms, including the χ^2 measure which Berkeley used in TREC-3 and TREC-4. This measure ranks terms by the degree to which they are dependent upon relevance. In earlier TRECs, Berkeley did massive query expansion by choosing all terms associated with relevance at the 5 percent significance level. In TREC-5 this resulted in a variable number of terms per query from a minimum of 714 to a maximum of 3839 with a mean of 2032 terms over the 50 queries.

In TREC-5 Berkeley introduced the idea of using logistic on the term frequency in documents for the 15 most important terms in the ranking. This produced an approximately 20 percent improvement over the massive query expansion. Further investigations following TREC-5 showed equivalent performance improvements for the top 3 and 5 terms as well, [5] and that adding more terms achieved higher precision at the expense of total documents retrieved in the top 1000 documents.

As can be imagined, processing for 100,000 query terms over 50 documents becomes an i/o and cpu intensive task. Moreover, when we began a similar χ^2 selection for the 43 old queries of TREC-6, it produced 486,308 query terms, or 11,309 per query. The processing task for such queries seemed insurmountable for our limited resources. Thus we took to choosing a χ^2 cutoff at the 0.001 significance level.

At the same time we began investigating the U-Measure used by ETH in TREC-5 [4] also known as the Correlation Coefficient used in a text categorization study by Ng and others [9]. This measure is claimed to improved upon χ^2 by eliminating negative correlations between terms and relevance. Indeed our initial experiments showed that choice of the top 50 terms by u-measure ranking would produce results close to massive query expansion using χ^2. This was thus the method by which we choose terms for addition to the query, after retrieving all terms which satisfied a significance cutoff of 0.001 for the U-measure. We also performed logistic regression training on the term frequency in document for the top 5 and top 15 terms. These became our official runs BRKLY19 and BRKLY20.

Unfortunately the uniform application of a 0.001 significance level adversely affected the new routing topics 10001-10004 for which there was limited training data. Thus our choice of cutoff produced less than 28 additional terms for each of these queries, including these ten terms for topic 10003 (Privatization in Peru) – 'span-feb', 'span', 'priv', 'editor-report', 'cop', 'editor', 'roundup', 'feb', 'through-febru', 'la' – hardly very discriminating terms. It is not surprising that our performance on this query was among the worst of our performances when compared to the median. Choice of a 5 percent significance level would surely have produced better queries.

Another problem which we immediately encountered in processing the routing data was massive document duplication in the initial files of FBIS2. For example a simple pattern search of headers H3 reveals over 50 copies of the document headed by

```
<H3> <TI>    Thomson-CSF, Thorn EMI Defense Link-Up </TI></H3>
```

Fortunately this massive duplication seems to be confined to the first 20 files of the collection, although a random selection of other files revealed a few duplicates. As far as results are concerned, we have not spent time examining for duplicate documents, but we have determined that our top ranked two documents

```
003 Q0 FB6-F144-0008        1    1.000000 Brkly20
003 Q0 FB6-F144-0025        2    1.000000 Brkly20
```

for the Brkly20 run for query 003 (Japanese joint ventures) are identical documents with different document ids.

5 Tracks

For TREC-6 Berkeley participated in the Filtering, Chinese, and Cross-language tracks. An independent effort was mounted for the interactive track which is summarized in a separate paper. Berkeley had participated in the Chinese track in TREC-5 but this was our first participation in the Filtering track. For Cross-language, Berkeley submitted runs for English queries against German documents.

5.1 Cross-language: English queries against German documents

Berkeley decided to participate in the cross-language track in order to once again test the robustness of our probabilistic algorithm for ad-hoc document retrieval which has performed so well for Chinese and Spanish retrieval [6]. Our German-German run used the TREC-2 algorithm unchanged from its English implementation. For both our German-German and English-German runs we recognized the importance of phrase discovery which Ballesteros and Croft [1] have found to be paramount in effective cross-language retrieval. In English to German this becomes paramount because of the propensity for German to form compounds of single words equivalent to phrases in English. For example, the phrase 'air pollution' of topic CL6 can become the word 'Luftverschmutzung' in German, whereas the words 'air' and 'pollution' submitted separately to a dictionary do not provide the same meaning. The choice in dictionary retrieval is between obtaining only individual words which have little relationship to the phrase or obtaining all possible compound variations of the particular individual words. The former course results in missing the particular compound, while the latter results in obtaining a large set of noise words.

Initially we were unable to obtain an English-German dictionary and discovered a WWW dictionary (http://www.bg.bib.de/ã2h6bu) We had to write a cgi script which submitted English words and phrases and captured the output of the German translation. Since the transmission was subject to timeout failures, several runs had to be pooled and duplicate entries removed to obtain a final query.

Unlike our processing of the main track documents and queries, we did not retain the individual word components of discovered phrases. Finally, when English words were not found in the dictionary we kept the English word in the German query under the assumption that proper names (Kurt Waldheim is a good example) would be the same in both languages. These principles guided our English to German automatic run BrklyE2GA.

Our manual run BrklyE2GM was produced by the same processing guidelines except that the English source was manually modified in much the same way as our main track manual modification. Phrases such as 'a relevant document will discuss' were removed (query reduction) while queries were also expanded to include reasonable specifics. In particular, topic CL13 on the Middle East peace process, specific country and place names such 'Israel', 'Egypt', 'Syria', 'west bank', 'golan heights' were added to the query. Unfortunately the dictionaries used did not contain translations for all geographic names so the value of the enhancement is unclear.

Our results are as follows: our German-German run (BKYG2GA) achieved average precision of .2845 over 21 judged topics (versus 0.3417 over the 13 topics judged before the conference), while our English-German automatic run had average precision of 0.1305 and the English-German manual run had average precision of 0.1822. Interestingly for topic CL24 on 'teddy bears', the precision

	BKYG2GA	BKYE2GM	BKYE2GA	XTGBL	XTETH
total rel	992	992	992	992	992
rel ret	675	422	295	452	425
avg prec	0.2845	0.1822	0.1305	0.1185	0.1831
0.00	0.6894	0.4647	0.3766	0.3789	0.4620
0.10	0.5259	0.3350	0.2390	0.2589	0.3377
0.20	0.4809	0.2871	0.1778	0.2035	0.2761
0.30	0.4402	0.2462	0.1598	0.1630	0.2321
0.40	0.3614	0.2237	0.1429	0.1279	0.2095
0.50	0.2933	0.1726	0.1337	0.1118	0.1706
0.60	0.2391	0.1356	0.1186	0.0886	0.1557
0.70	0.1556	0.1058	0.0977	0.0601	0.1277
0.80	0.0810	0.0812	0.0724	0.0460	0.0903
0.90	0.0269	0.0649	0.0274	0.0354	0.0598
1.00	0.0088	0.0059	0.0048	0.0263	0.0318

Table 2: TREC-6 Cross-Language Retrieval Results

of 0.8330 for our manual run exceeded the best precision of 0.7541 for the 10 German-German monolingual runs. This can be directly attributed to the process of query reduction.

On the other hand, the manual query for topic CL2 (marriages and marriage customs) had a disastrous reduction in precision from 0.1524 (BKYE2GA) to 0.0492 (BKYE2GM), which may be attributable to the addition of the word 'customs' (as in marriage customs) which produced numerous translations.

One question is to the degree of overlap between monolingual and crosslingual retrieval. We analyzed the overlap between our German-German and English-German automatic runs and found that 14,894 documents in common among the 25000 documents retrieved by each run. We did not examine the overlap in the top 50 documents.

Since the conference we purchased the GlobalLink web translation package and used it to translate the topics from English to German. This automatic run (XTGBL) produced a precision of 0.1185, worse than our dictionary based automatic run, while at the same time retrieving more relevant documents (452) than any other cross-language run. Paraic Sheridan of the ETH group kindly supplied their machine translation of the English topics which used the T1 text translator which incorporates the Langenscheidt Dictionary. This run (XTETH) achieved a precision of 0.1831, slightly better that Berkeley's manual run.

Table 2 provides a detailed comparison of all our experiments.

5.2 Filtering

TREC-6 was the Berkeley group's first participation in the filtering track. While our entry is a straightforward probabilistic ranking with threshold approach, some interesting twists appeared as we began to work on the problem. First, we used an approach to query development identical to our TREC-6 routing approach (basically query expansion using statistical measures of Chi Square and U-measure, as well as logodds of relevance) trained only on the FBIS disk5 training set. For some topics, important query terms proved to be identical to those for routing training, while for other queries a dramatically different set of terms emerged. In addition, we use logistic regression on the term frequencies of the 5 most important terms. Because of the paucity of training data for some queries, the regression would not converge for four of the 47 filtering topics, so we had to use

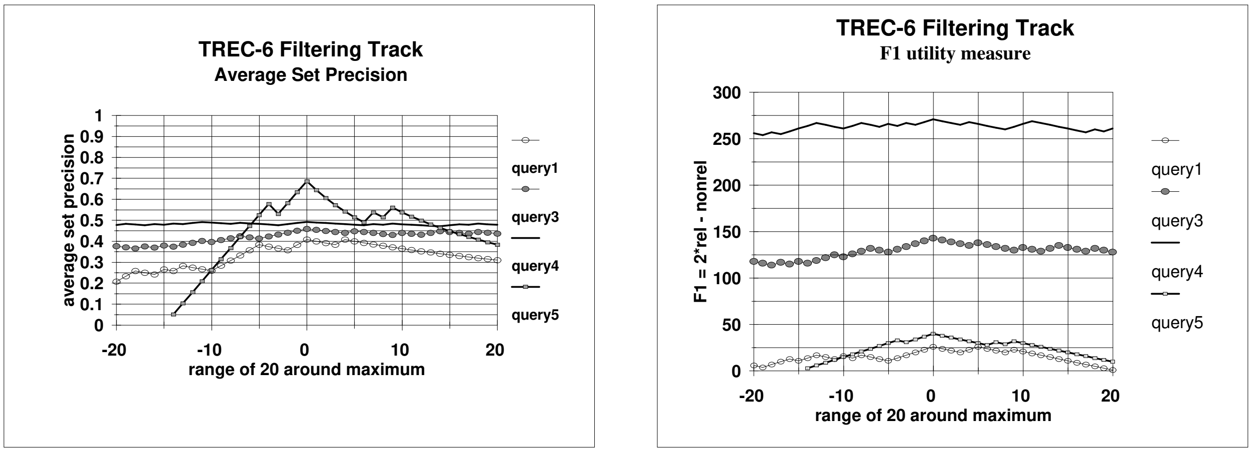

Figure 1: Filtering thresholds for ASP and F1.

a completely different thresholding mechanism for those four topics.

Our probability threshold was chosen for each utility measure based upon maximizing the utility over the training data. However examination of the distribution of utilities around the maximum showed quite different behavior patterns for different topics-some maxima were quite crisp while others were fuzzy or uncertain. Furthermore, for crisp thresholds (ones where the maximum utility is significantly higher than the surrounding utilities), it is unclear whether to choose that threshold or to lower the threshold in the direction of the next-highest values.

Figure 1 plots the values of average set precision for 20 document ranks on either side of the maximum value for the first four trec queries. As can be seen the maximum is crisp only for TREC query 005. This query is also the only one where the maximum is achieved before 20 documents have been ranked. On the other hand query 001 has a very fuzzy threshold, achieving close to the maximum at document ranks well beyond the actual maximum. It is unclear what value should have been used for thresholding for this query. The choice of thresholds from ranked retrieval appears to be a fundamental research problem.

Finally Berkeley decided to submit a pure Boolean run which consisted of those documents which contained all 5 most important query terms for each topic. We submitted this run (BKYT6BOOL) to be evaluated by all three evaluation measures. The number of documents retrieved by this method was dramatically different from the probability threshold results. By all measures (when averaged over 47 queries) the Boolean retrieval performed much worse than probabilistic retrieval with thresholding. Interestingly enough, however, the retrieval of 52 documents for topic 001 scored the maximum for all three performance measures. For that topic the five terms used for coordination retrieval were 'commit' 'fair trad' 'trad' 'fair' 'ftc'.

5.3 Chinese

Because Chinese text is delivered without word boundaries, automatic segmentation of text into imputed word components is a prerequisite to retrieval. One group of word segmentation methods are based on dictionary. Berkeley believes that the coverage of the dictionary over the collection to index can have significant impact on the retrieval effectiveness of a Chinese text retrieval system that uses a dictionary to segment text. In TREC-5 [7], we combined a dictionary found on the web and entries consisting of words and phrases extracted from the TREC-5 Chinese collections to create a dictionary of about 140,000 entries and we used the dictionary to segment the Chinese collection. This dictionary certainly is not small in size, yet we found that the dictionary did not include many of the proper names such as personal names, transliterated foreign names, company names, university and college names, research institutions and so on. Our focus in Chinese track of TREC-6 was on automatic and semi-automatic augmentation of the Chinese dictionary which we used to segment the Chinese collection.

Based on the observations that personal names are often preceded by title names and followed by a small group of verbs such as *say*, *visit*, *suggest* et al, and the first name, middle name and the last name of a transliterated foreign name are separated by a special punctuation mark, we constructed a set of pattern rules by hand to extract any sequence of characters in the text that matches any pattern rule. We then went through the list by hand to remove the entries that are not personal names.

In Chinese text, the items (such as names) in a list are uniquely marked by a special punctuation mark. We wrote a simple program to take out any sequence of characters flanked by the special punctuation mark. The technique seems to be quite productive for it produced over 10,000 entries from the TREC-5 Chinese collection. There are, of course, some entries that are not meaningful. The appendix contains a sample text excerpt and the names (country names and company names) that were extracted from the excerpt.

Berkeley submitted two runs, named BrklyCH3 and BrklyCH4 respectively, for the Chinese track. BrklyCH3 is the run using the original long queries with automatic query expansion and BrklyCH4 is the run based on the manually reformulated queries. For both runs, the collection was segmented using the dictionary-based maximum matching method. For BrklyCH3, an initial retrieval run was carried out to produce a ranked list of documents, then 20 new terms were selected from the top 10 ranked documents for each query. The selected terms are those that occur most frequently in the top 10 documents in the initial ranked list. The chosen terms were added to the original long queries to form the expanded queries. A final run was carried out using the automatically expanded queries to produce the results in BrklyCH3. For both runs, the documents were ranked by the probability of relevance estimated using the Berkeley's TREC-2 adhoc retrieval formula. For BrklyCH4, we spent about 40 minutes per query to manually reformulate each query by 1) removing non-content words from the original queries; 2) adding new words found in the collection to the original queries; and 3) adjusting the weights assigned to each term in the queries.

6 Conclusions and Acknowledgments

In our TREC-6 experiments for the main tasks and tracks, Berkeley worked primarily on extending our probabilistic document retrieval methods to incorporate two word phrases found using the ranking provided by expected mutual information measure. While these methods did not result in performance improvements for English retrieval, they were central in obtaining reasonable performance in English queries against German documents in the crosslingual track. Our first foray into the Filtering task obtained reasonable results for precision by using threshold computations to truncate a ranked retrieval and obtain a pool of unranked documents. Clearly finding the proper threshold in transforming from ranked retrieval to document sets is a research problem which will require considerably more study.

We acknowledge the assistance of Jason Meggs who indexed and ran the German document collection and Lily Tam and Sophia Tang, computer science undergraduates who provided programming assistance and who helped in the manual reformulation of Chinese queries. This research was supported by the National Science Foundation under grant IRI-9630765 from the Database and Expert Systems program of the Computer and Information Science and Engineering Directorate.

References

[1] Lisa Ballesteros and W. Bruce Croft. Phrasal Translation and Query Expansion Techniques for Cross-Language Information Retrieval. In Nicholas J. Belkin, A. Desai Narasimhalu, and Peter Willett, editors, *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Philadelphia*, pages 84–91, 1997.

[2] A. Chen, J. He, L. Xu, F. C. Gey, and J. Meggs. Chinese Text Retrieval Without Using a Dictionary. In A. Desai Narasimhalu Nicholas J. Belkin and Peter Willett, editors, *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Philadelphia*, pages 42–49, 1997.

[3] W. S. Cooper, A. Chen, and F. C. Gey. Full Text Retrieval based on Probabilistic Equations with Coeffici ents fitted by Logistic Regression. In D. K. Harman, editor, *The Second Text REtrieval Conference (TREC-2)*, pages 57–66, March 1994.

[4] Ballerini et al. SPIDER Retrieval System at TREC-5. In D. K. Harman and Ellen Voorhees, editors, *The Fifth Text REtrieval Conference (TREC-5), NIST Special Publication 500-238*, pages 217–228, November 1997.

[5] F. C. Gey and A. Chen. Term importance in routing retrieval . In *Submitted for publication*, December 1997.

[6] F. C. Gey, A. Chen, J. He, L. Xu, and J. Meggs. Term importance, Boolean conjunct training, negative terms, and foreign language retrieval: probabilistic algorithms at TREC-5. In D. K. Harman and Ellen Voorhees, editors, *The Fifth Text REtrieval Conference (TREC-5), NIST Special Publication 500-238*, pages 181–190, November 1997.

[7] J. He, L. Xu, , A. Chen, J. Meggs, and F. C. Gey. Berkeley Chinese Information Retrieval at TREC-5: Technical Report. In D. K. Harman and Ellen Voorhees, editors, *The Fifth Text REtrieval Conference (TREC-5), NIST Special Publication 500-238*, pages 191–196, November 1996.

[8] David W. Hosmer and Stanley Lemeshow. *Applied Logistic Regression*. John Wiley & Sons, New York, 1989.

[9] H-T Ng, W-B Goh, and K-L Low. Feature Selection, Perceptron Learning, and a Useability Case Study for Text Categorization. In Nicholas J. Belkin, A. Desai Narasimhalu, and Peter Willett, editors, *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in In formation Retrieval, Philadelphia*, pages 67–73, 1997.

Appendix

An excerpt from a news article in the Xin Hua News collection.

——国外大公司投资活跃 去年年初，山东半岛各地就把吸引国外大公司、大商社投资做为重点，积极谋求与国际上资本雄厚、管理一流的大公司进行合作，先后有来自美国、德国、法国、瑞士、日本、韩国、香港的三十多个大公司前来投资办厂。

如美国道化学公司、戴诺润滑油公司、德国西门子公司、德固萨公司、瑞士汽巴嘉基公司、日本丰田公司、英国商业集团公司、韩国大宇公司、世运工业公司、香港中策公司、中银集团等。

Names extracted from the above paragraph include:

德国German
法国(France)
瑞士(Switzerland)
日本(Japan)
韩国(German)
戴诺润滑油公司
德国西门子公司
德固萨公司
瑞士汽巴嘉基公司
日本丰田公司
英国商业集团公司
韩国大宇公司
世运工业公司
香港中策公司