Cross-Language Retrieval for the CLEF Collections – Comparing Multiple Methods of Retrieval

Fredric C. Gey[1], Hailing Jiang[2], Vivien Petras[2] and Aitao Chen[2]

[1] UC Data Archive & Technical Assistance,
[2] School of Information Management and Systems
University of California
Berkeley, CA 94720 USA

Abstract. For our participation in CLEF, the Berkeley group participated in the monolingual, multilingual and GIRT tasks. To help enrich the CLEF relevance set for future training, we prepared a manual reformulation of the original German queries which achieved excellent performance, more than 110% better than average of median precision. The GIRT task performed English- German Cross-Language IR by comparing commercial machine translation with thesaurus lookup techniques and query expansion techniques. Combining all techniques using simple data fusion produced the best results.

1 Introduction

Unlike monolingual retrieval where the queries and documents are in the same language and where mechanistic techniques can be applied, Cross-language information retrieval (CLIR) must combine linguistic techniques (phrase discovery, machine translation, bilingual dictionary lookup) with robust monolingual information retrieval. The Berkeley Text Retrieval Research group has been using the technique of logistic regression from the beginning of the TREC series of conferences. Indeed our primary development has been a result of the U.S. TREC conferences and collections which provided the first large-scale test collection for modern information retrieval experimentation. In TREC-2 [2] we derived a statistical formula for predicting probability of relevance based upon statistical clues contained with documents, queries and collections as a whole. This formula was used for document retrieval in Chinese[3] and Spanish in TREC-4 through TREC-6. We utilized the identical formula for English queries against German documents in the cross-language track for TREC-6. In TREC-7 the formula was also used for cross-language runs over multiple European languages. During the past year the formula has proven well-suited for Japanese and Japanese-English cross-language information retrieval[7], even when only trained on English document collections. Our participation in the NTCIR Workshop in Tokoyo (http://www.rd.nacsis.ac.jp/ñtcadm/workshop/work-en.html)

led to different techniques for cross-language retrieval, ones which utilized the power of human indexing of documents to improve retrieval via bi-lingual lexicon development and a form of text categorization which associated terms in documents with humanly assigned index terms[1]. These techniques were applied to English-German retrieval for the GIRT-1 task and collection in the TREC-8 conference [5]

2 Logistic Regression for Document Ranking

The document ranking formula used by Berkeley in all of our CLEF retrieval runs was the TREC-2 formula [2]. The ad hoc retrieval results on the TREC test collections have shown that the formula is robust for long queries and manually reformulated queries. Applying the same formula (trained on English TREC collections) to other languages has performed well, as on the TREC-4 Spanish collections, the TREC-5 Chinese collection [6] and the TREC-6 and TREC-7 European languages (French, German, Italian) [4, 5]. Thus the algorithm has demonstrated its robustness independent of language as long as appropriate word boundary detection (segmentation) can be achieved. The logodds of relevance of document D to query Q is given by

$$\log O(R|D,Q) = log\frac{P(R|D,Q)}{P(\overline{R}|D,Q)} \tag{1}$$

$$= -3.51 + \frac{1}{\sqrt{N}+1}\Phi + .0929 * N \tag{2}$$

$$\Phi = 37.4\sum_{i=1}^{N}\frac{qtf_i}{ql+35} + 0.330\sum_{i=1}^{N}\log\frac{dtf_i}{dl+80}$$
$$-0.1937\sum_{i=1}^{N}\log\frac{ctf_i}{cl} \tag{3}$$

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. Details about the derivation of these formulae may be found in our NTCIR workshop paper [7]. It is to be emphasized that training has taken place exclusively on English documents but the matching has proven robust over seven other languages in monolingual retrieval, including Japanese and Chinese where word boundaries form an additional step in the discovery process.

3 Submissions for the CLEF main tasks

For CLEF we submitted 8 runs, 4 for the Monolingual (non-English) task and 4 for the Multilingual task.

In the following sections is a description for each run.

For the Monolingual task we submitted:			
Run Name	Language	Run type	Priority
BKMOGGM1	German	Manual	1
BKMOFFA2	French	Automatic	2
BKMOGGA1	German	Automatic	3
BKMOIIA3	Italian	Automatic	4
For the Multilingual task we submitted:			
BKMUEAA1	English	Automatic	1
BKMUGAM1	German	Manual	2
BKMUEAA2	English	Automatic	3
BKMUGAA3	German	Automatic	4

Table 1. Summary of eight official CLEF runs.

3.1 Monolingual Retrieval of the CLEF collections

BKMOIIA3 (Berkeley Monolingual Italian against Italian Automatic Run 3) The original query topics in Italian were searched against the Italian collection (La Stampa). For indexing this collection, we used a stopwordlist, the Italian-to-lower normalizer and the Italian stemmer (from association dictionary) described in Section 4.

BKMOFFA2 (Berkeley Monolingual French against French Automatic Run 2)

The original query topics in French were searched against the French collection (Le Monde). For indexing this collection, we used a stopwordlist, the French-to-lower normalizer and the French stemmer (from association dictionary) described in Section 4.

BKMOGGA1 (Berkeley Monolingual German against German Automatic Run 1)

The original query topics in German were searched against the German collection (Frankfurter Rundschau and Der Spiegel). For indexing the collection, we used a stopwordlist that contained also capitalized versions of words and the German stemmer (from association dictionary) described in Section 3.4. We did not use a normalizer for this collection because all nouns in German are capitalized and hence this clue might be used in retrieval.

4. BKMOGGM1 (Berkeley Monolingual German against German Manual Run 1) The original query topics in German were extended with additional query terms obtained by searching the German CLEF collection (Frankfurter Rundschau and Der Spiegel) with the original German query topics and looking at the results for these original queries (with the help of Aitao Chen's Cross-language Text Retrieval System Web-interface). The additional query terms were obtained by either directly looking at the documents or looking at the top ranked document terms for the original query text. The searcher spent about 10 to 25 minutes per topic or query depending on familiarity with the context and meaningfulness of the returned documents and top ranked document terms. For indexing the collection, we used a stopwordlist that contained also capitalized

versions of words and the German stemmer (from association dictionary) built by Aitao Chen. We didn't use a normalizer for this collection.

3.2 Monolingual Performance

Our monolingual performance can be found in Table 2. While average of medians

Run ID	BKMOIIA3	BKMOFFA2	BKMOGGA1	BKMOGGM1
Retrieved	34000	34000	37000	37000
Relevant	338	528	821	821
Rel. Ret	315	508	701	785
Precision				
at 0.00	0.7950	0.7167	0.6342	0.6907
at 0.10	0.7617	0.6824	0.5633	0.6584
at 0.20	0.6601	0.5947	0.5173	0.6442
at 0.30	0.6032	0.5195	0.3999	0.6037
at 0.40	0.5756	0.4825	0.3687	0.5624
at 0.50	0.5336	0.4404	0.3181	0.5428
at 0.60	0.4189	0.3627	0.2731	0.4970
at 0.70	0.3098	0.2960	0.2033	0.4580
at 0.80	0.2417	0.2422	0.1704	0.4006
at 0.90	0.1816	0.1936	0.1364	0.2959
at 1.00	0.1533	0.1548	0.0810	0.2059
Brk. Prec.	0.4601	0.4085	0.3215	0.4968
Med. Prec.	0.4453	0.4359	0.3161	0.3161

Table 2. Results of four official CLEF monolingual runs.

cannot be considered a meaningful statistic from which inference can be made, we have found it useful to average the medians of all queries as sent by CLEF organizers. Comparing our overall precision to this average of medians gives us some fuzzy gauge of whether our performance is better, poorer, or about the same as the median performance. Thus the bottom two rows of the table present the Berkeley overall precision over all queries for which performance has been judged and, below it, the average of the median precision for each query over all submitted runs. From this we see that Berkeley's automatic runs are about the same as the overall 'average' while Berkeley's German-German manual run comes in at overall precision 57 percent better than Average of Median precisions for German-German monolingual runs. As we shall see in the next section, an improved German query set had an even greater impact on multilingual retrieval.

Another observation to make is that of the skewedness of relevancy. More than twice as many relevant documents come from the German collection than the Italian collection. Thus a better German query set may have an impact on multilingual retrieval more than a better Italian query set.

3.3 Multilingual Retrieval of the CLEF collections

Several interesting questions have arisen in recent research on CLIR. First, is CLIR merely a matter of a marriage of convenience between machine translation combined with ordinary (monolingual) information retrieval? In our CLEF work we made use of two widely available machine translation packages, the SYSTRAN system found at the AltaVista site, and the Lernout and Hauspie Power Translator Pro Version 7.0. For the GIRT retrieval we made comparisions to Power Translator. For CLEF multilingual we combined translations and dictionary lookup from multiple sources, having found that different packages made different mistakes on particular topics. Second, what is the role of language specific stemming in improved performance? Our experience with the Spanish tracks of TREC have convinced us that some form of stemming will always improve performance. For this particular evaluation we chose to create a stemmer mechanistically from common leading substring analysis of the entire corpus. The impact of the stemmer on performance will be discussed at the end of the official results discussion. Third, is performance improved by creating a multilingual index by pooling all documents together in one index or by creating separate language indexes and doing monolingual retrieval for each language followed by data fusion which combines the individual rankings into a unified ranking independent of language? This was one of the major focuses of our experiments at CLEF.

1. BKMUEAA1 (Berkeley Multilingual English against all Automatic Run 1)

The original query topics in English were translated once with the Systran system
(http://babel.altavista.com/translate.dyn) and with L&H Powertranslator. The English topics were translated in French, German, and Italian. The two translated files for each languages were pooled together and then put together in one query file (the English original query topics were multiplied by 2 to gain the same frequency of query terms in the query file). The final topics file contained 2 English (original), French, German, and Italian versions (one Powertranslator and one Systran) for each topic. During the search, we divided the frequency of the search terms by 2 to avoid over-emphasis of equally translated search terms. The collection consisted of all languages. For indexing the English part of this collection, we used a stopwordlist, the default normalizer and the Porter stemmer. For indexing the French part of this collection, we used a stopwordlist, the French-to- lower normalizer and the French stemmer (from association dictionary in section 3.5). For indexing the German part of the collection, we used a stopwordlist that contained also capitalized versions of words and the German stemmer (from association dictionary) build by Aitao Chen. We didn't use a normalizer for this collection. For indexing the Italian part of this collection, we used a stopwordlist, the Italian-to- lower normalizer and the Italian stemmer (from association dictionary).

2. BKMUEAA2 (Berkeley Multilingual English against all Automatic Run 2)

Run ID	BKMUEAA1	BKMUEAA2	BKMUGAA2	BKMUGAM1
Retrieved	40000	40000	40000	40000
Relevant	2266	2266	2266	2266
Rel. Ret.	1434	1464	1607	1838
Precision				
at 0.00	0.7360	0.7460	0.7238	0.7971
at 0.10	0.5181	0.5331	0.5046	0.6534
at 0.20	0.4287	0.4465	0.4229	0.5777
at 0.30	0.3545	0.3762	0.3565	0.5032
at 0.40	0.2859	0.2929	0.3027	0.4373
at 0.50	0.2183	0.2290	0.2523	0.3953
at 0.60	0.1699	0.1846	0.1990	0.3478
at 0.70	0.1231	0.1454	0.1682	0.3080
at 0.80	0.1020	0.0934	0.1295	0.2238
at 0.90	0.0490	0.0480	0.0622	0.1530
at 1.00	0.0136	0.0081	0.0138	0.0474
Brk. Prec.	0.2502	0.2626	0.2654	0.3903
Med. Prec.	0.1843	0.1843	0.1843	0.1843

Table 3. Results of four official CLEF multilingual runs.

The original query topics in English were translated once with Systran and with L&H PowerTranslator. The English topics were translated into French, German, and Italian. The 2 translated versions for each language were pooled together in one query file (resulting in 3 topics files, one in German with the Systran and Powertranslator version, one in French with the Systran and Powertranslator version, and one in Italian accordingly). The original English topics file was searched against the English collection (Los Angeles Times). The pooled German topics file was searched against the German collection, the pooled French topics file was searched against the French collection, and the pooled Italian topics file was searched against the Italian collection. The frequency of the search terms was divided by 2 to avoid over-emphasis of equally translated search terms. This resulted in 4 result files with the 1000 top ranked records for each topic. These 4 result files were then pooled together and sorted by weight (rank) for each record and topic. The pooling method is described below. For a description of the collections see BKMOGGM1, BKMOFFA2, BKMOIIA3, BKMUEAA1.

3. BKMUGAA2 (Berkeley Multilingual German against all Automatic Run 2)

The original query topics in German were translated once with Systran and with Powertranslator. The German topics were translated into English, French, and Italian. The 2 translated versions for each language were pooled together in one query file. The original German topics file was multiplied by 2 to gain the same frequency of query terms in the query file searched. The final topics file contained 2 German (original), English, French, and Italian versions (one Powertranslator and one Systran) for each topic. During the search, we divided the

frequency of the search terms by 2 to avoid over-emphasis of equally translated search terms. For a description of the collection see BKMUEAA1.

4. BKMUGAM1 (Berkeley Multilingual German against all Manual Run 1)

The manually extended German query topics (see description from BKMOGGM1) were now translated with Powertranslator into English, French and Italian. These translations were pooled together with the German originals in one file. This topics file was searched against the whole collection including all 4 languages. For a description of the collection see BKMUEAA1.

3.4 Berkeley's CLEF Multilingual Performance

Our multilingual performance can be found in Table 3.

As contrasted from the average of medians for monolingual, the values in the last row of the table are the same for all columns. Our automatic runs performed almost identically at about 38 percent better than average of medians, while the run BKMUGAM1 at overall precision 0.39 is 112 percent greater than the average of multilingual query medians.

3.5 Building a Simple Stemmer for Cross-Language Information Retrieval

A stemmer for the French collection was created by first translating all the distinct French words found in the French collection into English using SYSTRAN. The English translations were normalized by reducing verbs to the base form, nouns to the singular form, and adjectives to the positive form. All the French words which have the same English translations after normalization were grouped together to form a class. A member from each class is selected to represent the whole class in indexing. All the words in the same class were replaced by the class representative in indexing.

The German stemmer and Italian stemmer were created alike.

We submitted four monolingual runs and four multilingual runs. These eight runs were repeated without the French, German, and Italian stemmers. The overall precision for each of the eight runs without stemming are shown in column 3 of table 4. Column 4 shows the overall precision with the French, German, and Italian stemmers. Column 5 shows the improvement in precision which can be attributed to the stemmers.

The overall precision for pooling queries and without stemming (the method we applied two years ago) for the multilingual run using English queries was .2335. With stemming and pooling documents, the overall precision for the same run was .2626, which is 12.46 percent better. This can be considered as additional evidence that adding a stemming capability will result in an improvement in automatic multilingual retrieval.

3.6 Data fusion or Monolingual Document Pooling

The second idea centers on pooling documents from monolingual retrieval runs. The brain-dead solution would be to simply combine the retrieval results from

RUN ID	TASK	RESULTS (unstemmed)	OFFICIAL RESULTS (stemmed)	Change Change
BKMUEAA1	multilingual	0.2335	0.2502	7.15pct
BKMUEAA2	multilingual	0.2464	0.2626	6.57pct
BKMUGAA3	multilingual	0.2524	0.2654	5.15pct
BKMUGAM1	multilingual	0.3749	0.3903	4.10pct
BKMOFFA2	monolingual	0.3827	0.4085	6.74pct
BKMOGGA1	monolingual	0.3113	0.3215	3.27pct
BKMOGGM1	monolingual	0.4481	0.4968	10.86pct
BKMOIIA3	monolingual	0.4054	0.4601	13.49pct

Table 4. Results of Stemming Experiments

four monolingual retrieval runs and sort the combined results by the estimated probability of relevance. The problem with the simple combination approach is that when the estimated probability of relevance is biased toward one document collection (as the above statistics show for German), the documents from that collection will always appear in the top in the combined list of ranked documents. For our final run, we took a more conservative approach by making sure the top 50 documents from each of the four monolingual list of documents will appear in top 200 in the combined list of documents.

3.7 Failure Analysis

A query-by-query analysis can be done to identify problems. We have not had time to do this, but one query stands out. Query 40 about the privatization of the German national railway was one which seems to have given everyone problems (the median precision over all CLEF runs was 0.0537 for this query). As an American group, we were particularly vexed by the use of the English spelling 'privatisation' which couldn't be recognized by either of our machine translation softwares. The German version of the topic was not much better – in translation its English equivalent became 'de-nationalization' a very uncommon synonym for 'privatization,' and one which yielded few relevant documents. By comparison, our German manual reformulation of this query resulted in an average precision of 0.3749 for best CLEF performance for this query.

4 GIRT retrieval

A special emphasis of our current funding has focussed upon retrieval of specialized domain documents which have been assigned individual classification identifiers by human indexers. These classification identifiers come from what we call "domain ontologies", of which thesauri are a particular case. Since many millions of dollars are expended on developing these classification ontologies and applying them to index documents, it seems only natural to attempt to exploit the resources previously expended to the fullest extent possible to improve retrieval. In some cases such thesauri are developed with identifiers translated (or

provided) in multiple languages. This has been done in Europe with the GEMET (General European Multilingual Environmental Thesaurus) effort and with the OECD General Thesaurus (available in English, French, and Spanish). A review of multilingual thesauri can be found in [8].

The GIRT collection consists of reports and papers (grey literature) in the social science domain. The collection is managed and indexed the GESIS organization (http://www.social-science-gesis.de). GIRT is an excellent example of a collection indexed by a multilingual thesaurus, originally German-English, recently translated into Russian. We worked extensively with a previous version of the GIRT collection in our cross-language work for TREC-8 [5]

4.1 The GIRT Collection

There are 76128 German documents in GIRT subtask collection. Of them, about 54275 (72 percent) have English TITLE sections. 5317 documents (7 percent) have also English TEXT sections. Almost all the documents contain manually assigned thesaurus terms. On average, there are about 10 thesaurus terms assigned to each document.

In our experiments, we indexed only the TITLE and TEXT sections in each document (not the E-TITLE or E-TEXT). The CLEF rules specified that indexing any other field would need to be declared a manual run. For our CLEF runs this year we added a German stemmer similar to the Porter stemmer for the German language. Using this stemmer led to a 15 percent increase in average precision when tested using the GIRT-1 collection of TREC-8.

4.2 Query translation

In CLIR, essentially either queries or documents or both need to be translated from one language to another. Query translation is almost always selected for practical reasons of efficiency, and because translation errors in documents can propagate without discovery since the maintainers of a text archive rarely read every document.

For the CLEF GIRT task, our focus has been to compare the performance of different translation strategies. We applied the following three methods to translate the English queries to German: Thesaurus lookup, Entry Vocabulary Module (EVM), machine translation (MT). The resulted German queries were run against the GIRT collection.

Thesaurus lookup The GIRT social science Thesaurus is a German-English bilingual thesaurus. Each German item in this thesaurus has a corresponding English translation. We took the following steps to translate the English query to German by looking up the thesaurus:

a. Create an English-German transfer dictionary from the Social Science Thesaurus. This transfer dictionary contains English items and their corresponding

German translations. This "vocabulary discovery" approach was taken by Eichmann, Ruiz and Srinivasan for medical information cross-language retrieval using the UMLS Metathesaurus[9].

b. Use the part-of-speech tagger LT-POS developed by University of Edinburgh

(http://www.ltg.ed.ac.uk/software/pos/index.html) to tag the English query and identify noun phrases in the English query. One problem with thesaurus lookup is how to match the phrasal items in a thesaurus. We have taken a simple approach to deal with this problem: use POS tagger to identify noun phrases.

For last year's GIRT task at the TREC-8 evaluation, we extracted an English-German transfer dictionary from the GIRT thesaurus and used it to translate the English queries to German. This approach left about 50 percent of English query words untranslated. After examining the untranslated English query words carefully, we found that most of them fell into the following two categories: one category contains general terms that are not likely to occur in a domain-specific thesaurus like GIRT. Examples are "country", "car", "foreign", "industry", "public", etc. The other category are terms that occur in the thesaurus but in a different format from the original English query words. For example, "bosnia-herzegovina" does not appear in the thesaurus, but "bosnia and herzegovina" does.

Fuzzy Matching for the Thesaurus To deal with the general terms in the first category, a general-purpose dictioanry was applied after thesaurus lookup. A fuzzy-matching strategy was used to address the problem for the second category. It counts the letter pairs that two strings have in common and uses Dice's coefficient as a means of accessing the similarity between the two strings. This fuzzy-matching strategy successfully recovered some query terms, for example,

original query terms	thesaurus terms
asylum policy	policy on asylum
anti-semitism	antisemitism
bosnia-herzegovina	bosnia and herzegovina
gypsy	gipsy
German Democratic Republic	German Democratic
Republic (gdr)	

Fuzzy matching also found related terms for some query terms which do not appear in the thesaurus at all, for example,

We tested this combined approach using last year's GIRT-1 data. The results showed about 18 percent increase as measured by average precision compared with simple thesaurus lookup.

original query terms	thesaurus terms
nature protection legislation	nature protection
violent act	violence
bosnia	bosnian

Entry Vocabulary Module (EVM) In the GIRT collection, about 72 percent of the documents have both German titles and English titles. 7 percent have also English text sections. This feature allows us to build a EVM which maps the English words appearing in English Title and text sections to German thesaurus terms. This mapping can then be used to translate the English queries. More details about this work can be found in [5].

Machine translation (MT) For comparison, we also applied the Lernout and Hauspie Power Translator product to translate the English queries into German.

Merging results While our CLEF Multilingual strategy focussed on merging monolingual results run independently on different subcollections, one per language, all our GIRT runs were done on a single subcollection, the German text part of GIRT. When analyzing the experimental training results, we noticed that different translation methods retrieved sets of documents that contain different relevant documents. This implies that merging the results from different translation methods may lead to better performance than of any one of the methods. Since we use the same retrieval algorithm and data collection for all the runs, the probability that a document is relevant to a query from different runs are commensurable. So, for each document retrieved, we used the sum of its probability from the different runs as its final probability to create the ranking for the merged results.

4.3 Results and analysis

Our GIRT results are summarized in Table 5. The runs can be described as follows: BKGREGA4 used our entry vocabulary method to map from query term to thesaurus term, the top ranked thesaurus term and its translation was used to create the German query. BKGREGA3 used the results of machine translation by the L&H Power Translator software. The run BKGREGA2 used thesaurus lookup of English terms in the query and a general purpose English German dictionary for not found terms as well as the fuzzy matching strategy described above. The final run BKGREGA1 pooled the merged results from the other three runs according to the sum of probabilities of relevance. Note that it performs significantly better than the other three runs, and about 61 percent better than the average of median precisions for the CLEF GIRT. One reason is that different individual runs performed much better on different queries. The three individual methods achieved best precision in eight of the 25 queries and the fusion run achieved best precision for another 4 queries.

Run ID	BKGREGA1	BKGREGA2	BKGREGA3	BKGREGA4
Retrieved	23000	23000	23000	23000
Relevant	1193	1193	1193	1193
Rel. Ret.	901	772	563	827
at 0.00	0.7013	0.5459	0.6039	0.6139
at 0.10	0.5610	0.4436	0.3662	0.4482
at 0.20	0.4585	0.4172	0.2881	0.3583
at 0.30	0.4203	0.3576	0.2633	0.3292
at 0.40	0.3774	0.3165	0.2486	0.2465
at 0.50	0.3454	0.2856	0.2266	0.2004
at 0.60	0.2938	0.2548	0.1841	0.1611
at 0.70	0.2025	0.1816	0.1107	0.1477
at 0.80	0.1493	0.1439	0.0663	0.1252
at 0.90	0.0836	0.0829	0.0575	0.0612
at 1.00	0.0046	0.0075	0.0078	0.0003
Brk. Prec.	0.3119	0.2657	0.2035	0.2299
Med. Prec.	0.1938	0.1938	0.1938	0.1938

Table 5. Results of four official GIRT English-German runs.

5 Summary and Acknowlegments

Berkeley's participation in CLEF has enabled us to explore refinements in Cross-language information retrieval. Specifically we have explored two data fusion methods – for the CLEF multilingual we developed a technique for merging from monolingual, language specific rankings which ensured representation from each constituent language. For the GIRT English-German task, we obtained improved retrieval by fusion of the results of multiple methods of mapping from English queries to German. A new stemming method was developed which maps classes of words to a representative word in both English and the targeted languages of French, German, and Italian. For future research we are creating a Russian version of the GIRT queries to test strategies for Russian-German retrieval via a multilingual thesaurus.

This research was supported in part by the Information and Data Management Program of the National Science Foundation under grant IRI-9630765 from the Information and Data Management program of the Computer and Information Science and Engineering Directorate. Major support was also provided by DARPA (Department of Defense Advanced Research Projects Agency) under research grant N66001-00-1-8911, Mar 2000-Feb 2003 as part of the DARPA Translingual Information Detection, Extraction, and Summarization Program (TIDES).

References

1. F Gey A Chen and H Jiang. Applying text categorization to vocabulary enhancement for japanese-english cross-language information retrieval. In S. Annandiou,

editor, *The Seventh Machine Translation Summit, Workshop on MT for Cross-language Information Retrieval, Singapore*, pages 35–40, September 1999.

2. W Cooper A Chen and F Gey. Full text retrieval based on probabilistic equations with coefficients fitted by logistic regression. In D. K. Harman, editor, *The Second Text REtrieval Conference (TREC-2)*, pages 57–66, March 1994.
3. A Chen J He L Xu F 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.
4. F. C. Gey and A. Chen. Phrase discovery for english and cross-language retrieval at trec-6. In D. K. Harman and Ellen Voorhees, editors, *The Sixth Text REtrieval Conference (TREC-6), NIST Special Publication 500-240*, pages 637–647, August 1998.
5. F. C. Gey and H. Jiang. English-german cross-language retrieval for the girt collection – exploiting a multilingual thesaurus. In Ellen Voorhees, editor, *The Eighth Text REtrieval Conference (TREC-8), draft notebook proceedings*, pages 219–234, November 1999.
6. 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.
7. A Chen F Gey K Kishida H Jiang and Q Liang. Comparing multiple methods for japanese and japanese-english text retrieval. In N. Kando, editor, *The First NTCIR Workshop on Japanese Text Retrieval and Term Recognition, Tokoyo Japan*, pages 49–58, September 1999.
8. J. Purat. *The World of Multilingual Environmental Thesauri.* www.sims.berkeley.edu/research/metadata/papers/purat98, 1998.
9. D Eichmann M Ruiz and P Srinivasan. Cross-language information retrieval with the umls metathesaurus. In W B Croft A Moffat C J van Rijsbergen R Wilkinson and J Zobel, editors, *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Melbourne, Australia*, pages 72–80, August 1998.