English-German Cross-Language Retrieval for the GIRT Collection – Exploiting a Multilingual Thesaurus

Fredric C. Gey and Hailing Jiang
UC Data Archive & Technical Assistance (UC DATA)
University of California, Berkeley
gey@ucdata.berkeley.edu, hjiang1@sims.berkeley.edu

Abstract

For TREC-8, the Berkeley experiments concentrated on the special GIRT collection. We utilized the GIRT thesaurus in multiple ways in working on English-German Cross-Language IR. Since the GIRT collection is truly multilingual (documents contain both German and English text), one would expect multilingual queries to achieve the best performance. This proved not to be the case.

1 Introduction

Successful cross-language information retrieval (CLIR) combines linguistic techniques (phrase discovery, machine translation, bilingual dictionary lookup) with robust monolingual information retrieval. The Berkeley group has been using the technique of logistic regression from the beginning of the TREC series of conferences. 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[4], 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].

2 The GIRT Collection

GIRT collection contains German documents (some have English sections inside) from the field of social science. It has some special features that make it ideal to try out different ideas. Among them are:

1. Each GIRT document was manually assigned controlled terms which are from the Social Science Thesaurus. Figure 1 shows a sample GIRT document.

On average there are about 10 terms given to a document. This offers an opportunity to explore how to utilize controlled vocabulary to enhance retrieval effectiveness.

```
<DOC>
<DOCNO>
GIRT950410185
</DOCNO>
<TITLE>
Ausländerinnen in der beruflichen qualifizierung - eine Handreichung
</TITLE>
<TITLE-ENG>
Female aliens in occupational qualification : a guide
</TITLE-ENG>
<AUTHOR>
Djafari, Nader; Brüning, Gerhild
</AUTHOR>
<DOCTYPE>
Sonstiges
</DOCTYPE>
<YEAR>
1994
</YEAR>
<PLACE>
Frankfurt am Main
</PLACE>
<CY>
DEU
</CY>
<ISBN>
3-88513-492-6
</ISBN>
<LANGUAGE>
DE
</LANGUAGE>
<CONTROLLED-TERM>
Ausländer; Frau; Beruf; Qualifikation; Bildungschance; Ausbildungssituation; Bundesrepublik Deutschland
</CONTROLLED-TERM>
<CLASSIFICATION>
Arbeitsmarkt- und Berufsforschung *
</CLASSIFICATION>
<METHOD>
Dokumentation
</METHOD>
<FREE-TERM>
GIRT
</FREE-TERM>
<CORPORATE-SOURCE>
Deutscher Volkshochschul-Verband e.V. Pädagogische Arbeitsstelle
</CORPORATE-SOURCE>
<TEXT>
"Die Handreichung gibt durch Hintergrundinformationen und Erfahrungsberichte Anregungen für Personen,
die in der beruflichen Erwachsenenbildung tätig sind, damit die Qualifizierungsmaßnahmen auch für Frauen
aus nicht-deutschen Kulturbereichen verstärkt geöffnet werden. Die dargestellten Praxisbereiche basieren auf
Modellversuchen und einzelnen innovativen Projekten. Sie sind punktuelle Impulsgeber in einer
Weiterbildungslandschaft, die für die Zielgruppe \'Ausländerinnen\' unzureichend ausgestattet ist."
(Autorenreferat, IAB-Doku)
</TEXT>
</DOC>
```


Figure 1: Sample GIRT Document, TREC-8 CLIR.

2. There are a total of 37637 documents in GIRT collection. Among them, about 27458 (73%) have both German and English titles. About 2714 (7%) have corresponding German and English text sections (abstracts). This feature would make it possible to apply some multilingual corpus techniques to create a specialized bilingual dictionary.

3 Approaches to GIRT Retrieval

3.1 Experimental setup

The GIRT collection was used in our experiments. Both German and English sections in a document were indexed. For the German sections, no stemmer was used. Single words were indexed except for Controlled-term section, in which the whole terms (phrases usually) were indexed in addition to the single word components. For the English sections of a document, the SMART stemmer was used. Single words were indexed, and no phrases were used.

3.2 Query translation

In CLIR, essentially either queries or documents or both need to be translated from one language to another. Query translation is usually selected for practical reasons of efficiency. In our GIRT experiments, we tried the following approach to translate the English query to German:

Thesaurus lookup. The 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[5].

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 took a simple approach to deal with this problem: use POS tagger to identify noun phrases.

c. Look up the single words and noun phrases in the English query in the English-German transfer dictionary. In our experiments, we found that in some cases mismatch was caused by the different formats of words used in the query and the dictionary. For example, "women" is not found in the dictionary, but "woman" is. "anti-semitism" is not in the dictionary, but "antisemitism" is. So we adopted some rules when looking up the dictionary, such as, If a word or phrase is not found in the dictionary, look up its base form. The base form of a word is obtained using WordNet. If a word with '-' inside is not found in the dictionary, replace the '-' with a space or remove the '-'.

In our experiments, over 60% of query words or phrases were found in the transfer dictionary. Those which were not found are mostly very general terms and may not have affected the retrieval result.

3.3 Query expansion

We tested two approaches to expand the translated query.

1. Use of thesaurus terms to expand queries. We tried a KNN-similarity method [6, 8] to assign German thesaurus terms to each English query and add the thesaurus terms to the the German query translated using the thesaurus lookup. First, we run the English query against the documents which have English titles and/or abstracts (using Berkeley TREC2 formula), then extract the thesaurus terms assigned to the top 30 retrieved documents, and rank them by the number of documents to which they are assigned. The top thesaurus terms that occur in at least 5 documents are chosen and added to the translated German query.

2. Use of the hierarchical relationship in the thesaurus to expand the query. For each German thesaurus term in the translated query, we add its narrow terms (NT) to expand the query.

4 GIRT Experiments - official runs

We submitted 5 official runs. The only difference between these runs is how the query was constructed to run against the collection:

BKCLGR01: English query translated to German using thesaurus lookup.

BKCLGR02: English query translated to German using thesaurus lookup and expanded by narrow terms in the thesaurus.

BKCLGR03: English query translated to German using thesaurus lookup and expanded by German thesaurus terms.

BKCLGR04: English query + German query translated using thesaurus lookup.

BKCLGR05: English query + German query translated using thesaurus lookup and expanded by German thesaurus terms.

The results of our five official runs are presented in Table 1.

Run ID	BKCLGR01	BKCLGR02	BKCLGR03	BKCLGR04	BKCLGR05
Retrieved	28000	28000	28000	28000	28000
Relevant	1294	1294	1294	1294	1294
Rel-ret	907	733	936	890	921
Precision					
at 0.00	0.6564	0.4124	0.6671	0.7111	0.7036
at 0.10	0.5326	0.3329	0.5402	0.5492	0.5325
at 0.20	0.4485	0.2826	0.4721	0.3388	0.3689
at 0.30	0.3858	0.2568	0.4087	0.2774	0.2854
at 0.40	0.2924	0.1971	0.3131	0.2068	0.2327
at 0.50	0.2640	0.1746	0.2864	0.1550	0.1768
at 0.60	0.2107	0.1175	0.2163	0.1103	0.1435
at 0.70	0.1674	0.0935	0.1726	0.0848	0.1159
at 0.80	0.1242	0.0774	0.1241	0.0572	0.0805
at 0.90	0.0412	0.0131	0.0314	0.0178	0.0319
at 1.00	0.0208	0.0018	0.0008	0.0008	0.0004
Avg prec.	0.2707	0.1667	0.2832	0.2049	0.2232

Table 1: Results of five official GIRT runs.

These results show that adding the original English terms to the queries reduced the overall precision of the results while modestly increasing the precision for the first few documents.

5 Other GIRT Experiments - unofficial runs

We continued to make other runs on the GIRT collection, exploring a variety of approaches and also creating some baseline monolingual runs against which to measure our cross-language techniques. In TREC-7 we made use of commercial machine translation software to do all runs and achieved better results than bilingual dictionary lookup. In addition to the 5 official runs, we also did these experiments:

1. SYSTRAN Machine Translation System [7]
2. SYSTRAN translation expanded by thesaurus terms
3. German monolingual
4. German monolingual expanded by thesaurus terms
5. English query directly run against the collection (without translation)

The results for these five experimental runs are shown in Table 2.

These results, when compared with the official runs, show that the vocabulary provided by the GIRT Thesaurus supplied considerable improvement over general machine translation unaugmented by a specialized dictionary. Most surprizingly, we found that the general purpose SYSTRAN translation did not perform as well as the untranslated English query.

Run ID	SYSTRAN	SYSTRAN W/Expansion	German Monolingual	Monolingual W/Expansion	English Only
Retrieved	28000	28000	28000	28000	28000
Relevant	1294	1294	1294	1294	1294
Rel-ret	644	818	838	918	545
Precision					
at 0.00	0.4057	0.5302	0.8463	0.7966	0.6802
at 0.10	0.2407	0.3244	0.6554	0.6067	0.4234
at 0.20	0.2078	0.2885	0.5601	0.4856	0.1509
at 0.30	0.1531	0.2504	0.4415	0.4327	0.1238
at 0.40	0.1033	0.1878	0.3037	0.3637	0.1141
at 0.50	0.0849	0.1642	0.2398	0.3058	0.0690
at 0.60	0.0635	0.1144	0.1480	0.2159	0.0289
at 0.70	0.0454	0.0836	0.0842	0.1409	0.0216
at 0.80	0.0228	0.0424	0.0558	0.0741	0.0078
at 0.90	0.0061	0.0159	0.0228	0.0237	0.0000
at 1.00	0.0000	0.0000	0.0122	0.0084	0.0000
Avg. prec.	0.1063	0.1654	0.2860	0.2960	0.1211

Table 2: Results of Other GIRT runs

It is interesting to note that while overall precision of the German monolingual run with query expansion (0.2960) is better than that of our best official run BKCLGR03, the official run finds more relevant documents (936 versus 918) in the top 1000 than the monolingual run.

6 Conclusions and Acknowledgments

There are many document collections available in the growing digital library world which have been humanly indexed from a controlled vocabulary. Retrieval techniques which exploit this indexing to improve retrieval are in their infancy. The TREC-8 GIRT collection provides an interesting example of how such indexing may be utilized for cross-language information retrieval if indexing is done from a multi-lingual thesaurus. We conclude that exploiting the special vocabulary features of the thesaurus can more than double retrieval precision over general purpose machine translation. We also find that using a multilingual query to search multilingual documents may not achieve the best possible performance. Furthermore we find that query expansion using narrower terms from a thesaurus may degrade performance. This is probably because the extra terms seem to add noise documents to the retrieved set. It remains to be seen whether the inherent structure of the thesaurus can be successfully utilized to improve retrieval performance.

For future research it would be useful to take the GIRT German/English titles and align them to create a bilingual lexicon and see how that would perform against the multilingual thesaurus approach. We are also working on applying promising text categorization techniques, which have worked in Japanese-English CLIR, for query expansion [1].

This research was supported 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. Partial support was also provided by DARPA (Department of Defense Advanced Research Projects Agency) under research contract N66001-97-C-8541, AO-F477.

We thank Aitao Chen for much helpful advice and some programming support.

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] 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.

[5] 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 De velopment in Information Retrieval, Melbourne, Australia*, pages 72–80, August 1998.

[6] S Dumais J Platt D Heckerman M Sahamiu. Inductive learning algorithms and representations for text categorization. In G Gardarin J French N Pissinou K Makki and L Bouganim, editors, *Proceedings of CIKM98: The Seventh Internation Conference on Informatio and Knowledge Management, Nov 3-7, 1998, Bethesda MD*, pages 148–155, November 1998.

[7] SYSTRAN:. *http://babelfish.altavista.digital.com/.*

[8] Y. Yang and X. Lin. A Re-examination of Text Categorization Methods. In Fredric Gey Marti Hearst and Richard Tong, editors, *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Berkeley*, pages 42–49, 1999.