Are Thesauri Useful in Cross-Language Information Retrieval?

Vivien Petras[1] Natalia Perelman[1] and Fredric Gey[2]

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

1 Introduction

Digital libraries relating to particular subject domains have invested a great deal of human effort in developing metadata in the form of subject area thesauri. This effort has emerged more recently in artificial intelligence as ontologies or knowledge bases which organize particular subject areas. The purpose of subject area thesauri is to provide organization of the subject into logical, semantic divisions as well as to index document collections for effective browsing and retrieval. Prior to free-text indexing (i.e. the bag-of-words approach to information retrieval), subject area thesauri provided the only point of entry (or 'entry vocabulary') to retrieve documents. A debate began over thirty years ago about the relative utility of the two approaches to retrieval:

- to use index terms assigned by a human indexer, drawn from the controlled-vocabulary, or
- use automatic free-text indexing from the words or phrases contained in the document text.

This debate continues to this day and the evidence seems to have been mixed. In performance studies thesaurus-aided retrieval performs worse than free-text over a group of queries, while it performs better for particular queries [3].

It is an interesting question to evaluate what utility and performance can be obtained in cross-language information retrieval (CLIR) with the use of multilingual thesauri. The two domain-specific CLEF tasks, Amaryllis and GIRT, provide the opportunity to examine CLIR performance for such thesauri. The GIRT task provides a thesaurus for the social sciences in German, English, and (by translation) Russian, and Berkeley has studied it for three years. Amaryllis does not have a thesaurus per-se (i.e. it does not identify broader-terms, narrower-terms or related-terms), but it does have a specialized controlled-vocabulary for its domain of coverage in both the French and English languages.

In addition we have been investigating the viability of Russian as a query language for the CLEF collections and continue this research for the CLEF bi-lingual (Russian to German and Russian to French) main tasks and the GIRT task (Russian to German).

For monolingual retrieval the Berkeley group has used the technique of logistic regression from the beginning of the TREC series of conferences. In the TREC-2 conference [1] we derived a statistical formula for predicting probability of relevance based upon statistical clues contained within documents, queries and collections as a whole.

2 Amaryllis

The Amaryllis task consisted of retrieving documents from the Amaryllis collection of approximately 150,000 French documents which were abstracts of articles in a broad range of disciplines (e.g. biological sciences, chemical sciences, engineering sciences, humanities and social sciences, information science, medical sciences, physical and mathematics sciences, etc). There were twenty-five topics and the primary goal was French-French monolingual retrieval under multiple conditions (primarily testing retrieval with or without concept words from the Amaryllis controlled vocabulary. An auxiliary task was to test out English to French cross-language information retrieval.

For the Amaryllis task, we experimented with the effects of translation, inclusion of concept words and thesaurus matching. We indexed all fields in the document collection and used a stop-wordlist, the latin-to- lower normalizer and the Muscat French stemmer.

2.1 Amaryllis Thesaurus Matching

For the Amaryllis thesaurus matching task we first extracted individual words and phrases from the English topics. Phrases were identified by finding the longest matching word sequences in the Amaryllis vocabulary file that was used as a segmentation dictionary. This method identified phrases such as "air pollution" and "diesel engine" in the first topic. The individual words and phrases were then searched for in the Amaryllis vocabulary and if a match was found the words were replaced with their French equivalents.

2.2 Amaryllis Runs

Our Amaryllis results are summarized in Table 1. The runs are described below. The performance is computed over the top ranked 1000 documents for 25 queries.

Run Name	BKAMFF1	BKAMFF2	BKAMEF1	BKAMEF2	BKAMEF3
Retrieved	25000	25000	25000	25000	25000
Relevant	2018	2018	2018	2018	2018
Rel Ret	1935	1863	1583	1897	1729
Precision					
at 0.00	0.9242	0.8175	0.6665	0.8079	0.6806
at 0.10	0.8011	0.7284	0.5198	0.7027	0.6497
at 0.20	0.7300	0.6296	0.4370	0.6114	0.5874
at 0.30	0.6802	0.5677	0.3791	0.5612	0.5337
at 0.40	0.6089	0.5159	0.3346	0.5033	0.4983
at 0.50	0.5458	0.4722	0.2942	0.4489	0.4452
at 0.60	0.4784	0.4035	0.2481	0.3825	0.3848
at 0.70	0.4242	0.3315	0.1874	0.3381	0.3114
at 0.80	0.3326	0.2682	0.1251	0.2664	0.2414
at 0.90	0.2193	0.1788	0.0501	0.1888	0.1570
at 1.00	0.0596	0.0396	0.0074	0.0300	0.0504
Avg Prec.	0.5218	0.4396	0.2792	0.4272	0.4038

Table 1. Results of official Amaryllis runs for CLEF-2002.

BKAMFF1, our monolingual run including the concepts in the queries (title, description and narrative) yielded the best results. Our second monolingual run, BKAMFF2, where we excluded the concepts from the query indexes (only title and description) resulted in a 20% drop in average precision. Blind feedback improved the performance for both runs.

In comparing thesaurus matching and translation, this year the translation runs yielded better results. As a baseline, we run the English Amaryllis queries (without concepts) against the French Amaryllis collection (BKAMEF1). As expected, average precision wasn't very high, but it is still greater than 50 percent of the best monolingual run. Using machine translation for the second bilingual run (BKAMEF2) improved precision over 50%. For translating the English topics, we used the Systran and L & H Power translator. By using only the Amaryllis thesaurus to match English words with French thesaurus terms (the BKAMEF3 run), we improved our average precision 44% compared to the baseline. For all runs, the query indexes only included the title and description fields, but we used blind feedback for BKAMEF2 and BKAMEF3.

3 GIRT task and retrieval

The GIRT collection consists of reports and papers (grey literature) in the social science domain. The collection is managed and indexed by 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. The GIRT multilingual thesaurus (German-English), which is based on the Thesaurus for the Social Sciences [2], provides the vocabulary source for the indexing terms within the GIRT collection of CLEF. There are 76,128 German documents in GIRT subtask collection. Almost all the documents contain manually assigned thesaurus terms. On average, there are about 10 thesaurus terms assigned to each document.

For the Girt task, we experimented with the effects of different thesaurus matching techniques and the inclusion of thesaurus terms. The German Girt collection was indexed using the German decompounding algorithm to split compounds. For all runs, we used our blind feedback algorithm to improve the runs' performance.

3.1 GIRT results and analysis

Our GIRT results are summarized in Table 2. We had five runs, two monolingual, and two with Russian topics, and one with English topics. The runs are described below. Only 24 of the 25 GIRT queries had relevant documents, so the performance is computed over the top ranked 1000 documents for 24 queries. Except for the second monolingual run (BKGRGG2), we indexed all

Run Name	BKGRGG1	BKGRGG2	BKGREG1	BKGRRG1	BKGRRG2
Retrieved	24000	24000	24000	24000	24000
Relevant	961	961	961	961	961
Rel. Ret	853	665	735	710	719
Precision					
at 0.00	0.7450	0.6227	0.5257	0.5617	0.5179
at 0.10	0.6316	0.4928	0.3888	0.3595	0.3603
at 0.20	0.5529	0.4554	0.3544	0.3200	0.3233
at 0.30	0.5112	0.3551	0.3258	0.2705	0.2867
at 0.40	0.4569	0.3095	0.2907	0.2275	0.2263
at 0.50	0.4034	0.2462	0.2345	0.1793	0.1932
at 0.60	0.3249	0.1953	0.2042	0.1451	0.1553
at 0.70	0.2753	0.1663	0.1432	0.0945	0.1105
at 0.80	0.2129	0.1323	0.1188	0.0679	0.0858
at 0.90	0.1293	0.0497	0.0713	0.0413	0.0606
at 1.00	0.0826	0.0216	0.0454	0.0256	0.0310
Avg Prec.	0.3771	0.2587	0.2330	0.1903	0.1973

Table 2. Results of official GIRT runs for CLEF-2002.

allowed fields (including the controlled terms) in the document collection.

Using all query fields and indexing the controlled terms resulted in a 45% improvement in average precision for the monolingual Girt runs BKGRGG1 compared to BKGRGG2 (which only indexed the Title and Description query fields). The positive effect of including the narrative fields for the query indexing was countered by the different thesaurus matching techniques for the Russian Girt run.

Although the BKGRRG1 run used all query fields for searching, its fuzzy thesaurus matching technique resulted in a 3% drop in average precision compared to the BKGRRG2 run, which only used the title and description topic fields for searching but used a different thesaurus matching technique. Both runs pooled 2 query translations (Systran and Promt) and the thesaurus matching results into one file.

Comparing the Russian Girt runs (translation plus thesaurus matching) to the Russian to German bilingual runs with translation only (also: different collection), one can see a 36% and 70% improvement in average precision for the title and description only and the title, description and narrative runs, respectively.

Our final Girt run was BKGREG1 (Berkeley Girt English to German automatic run 1) where we used translation (L & H Power and the Systran translator) combined with our normalized

thesaurus matching technique. This run had better results than the Russian runs, but did not perform comparably (with respect to monolingual) to the bilingual main task English-to-German runs.

3.2 GIRT Thesaurus Matching

Similar to the Amaryllis thesaurus-based translation, we initially identified some phrases in the English GIRT topics by finding the longest matching entries in the English-German GIRT thesaurus. This method produced phrases such as "right wing extremism" and "drug abuse". Then individual topic words and phrases were matched against the thesaurus and replaced with their German translations.

For the thesaurus-based translation of Russian GIRT topics we first transliterated both Russian topics and Russian entries in the German-Russian GIRT thesaurus by replacing Cyrillic characters with their Roman alphabet equivalents. Then two different approaches were used to find matches in the thesaurus.

In one approach we identified phrases by finding the longest matching sequences of words from the topics in the thesaurus. We then used fuzzy matching method to match both phrases and individual words that were not identified as parts of the phrases. In this method, previously employed by our group in CLEF 2001, we identified thesaurus translations by determining Dice's coefficient of similarity between the topic words and phrases and the thesaurus entries.

Since fuzzy matching sometimes finds commonality between unrelated words, in our second approach, in order to deal with Russian inflectional morphology, we normalized Russian words by removing the most common Russian inflectional suffixes. Then we identified phrases as in the previous method and translated both phrases and individual words by finding their exact matches in the thesaurus.

4 Submissions for the CLEF main tasks

For the CLEF main tasks, we concentrated on French and German as the collection languages and English, French, German and Russian as the topic languages. We participated in 2 tasks: monolingual and bilingual for French and German document collections. We experimented with several translation programs, German decompounding and blind feedback. Two techniques are used almost universally:

Blind Feedback

For our relevance feedback algorithm, we initially searched the collections using the original queries. Then, for each query, we assumed the 20 top-ranked documents to be relevant and selected 30 terms from these documents to add to the original query for a new search.

German decompounding

To decompound the German compounds in the German and Girt collections, we first created a wordlist that included all words in the collections and queries. Using a base dictionary of component words and compounds, we then split the compounds into their components. During indexing, we replaced the German compounds with the component words found in the base dictionary.

4.1 Monolingual Retrieval of the CLEF collections

For CLEF-2002, we submitted monolingual runs for the French and German collections. Our results for the French bilingual runs were slightly better than those for the German runs. In both languages, adding the narrative to the query indexes improved average precision about 6% and 7% for the German and French runs, respectively.

BKMLFF1 (Berkeley Monolingual French against French Automatic Run 1). The original query topics (including all title, description and narrative) were searched against the French collection. We applied a blind feedback algorithm for performance improvement. For indexing the French collection, we used a stopwordlist, the latin-to- lower normalizer and the Muscat French stemmer.

BKMLFF2 (Berkeley Monolingual French against French Automatic Run 1). For indexing and querying the collections, we used the same procedure as in BKMLFF1. For indexing the topics, we only included the title and description.

BKMLGG1 (Berkeley Monolingual German against German Automatic Run 1). The query topics were searched against the German collection. For indexing both the document collection and the queries, we used a stopwordlist, the latin-to-lower normalizer and the Muscat German stemmer. We used Aitao Chen's decompounding algorithm to split German compounds in both the document collection and the queries. We applied our blind feedback algorithm to the results for performance improvement. All query fields were indexed.

BKMLGG2 (Berkeley Monolingual German against German Automatic Run 2). For this run, we used the same indexing procedure as for BKMLGG1. From the queries, only the title and description were search against the collections.

4.2 Bilingual Retrieval of the CLEF collections

We submitted 10 bilingual runs for search against the French and German collections. Overall, the Russian to German or French runs yielded decidedly worse results than the other language runs. Submitting English without any translation yielded much worse results than the same experiment in the Amaryllis collection – this was an error in processing where the French stop-word list and stemmer were applied to the English topic descriptions instead of the appropriate English ones. Correcting this error results in an overall precision of 0.2304 instead of the official result of 0.0513.

The English to French runs yielded slightly better results than the English to German runs, whereas the French to German run did better than the German to French run.

4.3 Bilingual to French Documents

Our runs for the CLEF bilingual-to-French main task (as well as monolingual French runs) are summarized in Table 3.

Run Name	BKMLFF1	BKMLFF2	BKBIEF1	BKBIEF2	BKBIGF1	BKBIRF1
Retrieved	50000	50000	50000	50000	50000	50000
Relevant	1383	1383	1383	1383	1383	1383
Rel. ret.	1337	1313	1285	162	1303	1211
Precision						
at 0.00	0.8125	0.7475	0.6808	0.0840	0.6759	0.5686
at 0.10	0.7747	0.6990	0.6284	0.0795	0.6271	0.5117
at 0.20	0.6718	0.6363	0.5642	0.0695	0.5582	0.4726
at 0.30	0.5718	0.5358	0.5210	0.0693	0.4818	0.4312
at 0.40	0.5461	0.5068	0.4962	0.0672	0.4589	0.3841
at 0.50	0.5017	0.4717	0.4702	0.0669	0.4389	0.3312
at 0.60	0.4647	0.4332	0.4260	0.0612	0.3986	0.3022
at 0.70	0.3938	0.3752	0.3713	0.0481	0.3428	0.2656
at 0.80	0.3440	0.3301	0.3302	0.0411	0.2972	0.2283
at 0.90	0.2720	0.2666	0.2626	0.0242	0.2330	0.1674
at 1.00	0.1945	0.1904	0.1868	0.0188	0.1686	0.1093
Avg prec.	0.4884	0.4558	0.4312	0.0513	0.4100	0.3276

Table 3. Results of Berkeley Bilingual to French runs for CLEF-2002.

BKBIEF1 (Berkeley Bilingual English against French Automatic Run 1). We translated the English queries with two translation programs: the Systran translator (Altavista Babelfish) and L & H's Power translator. The translations were pooled together and the term frequencies of words occurring twice or more divided (to avoid overemphasis of terms that were translated the

same by both programs). The title and description fields of the topics were indexed and searched against the French collections. For indexing the collection, we used the same procedures as in the monolingual runs. For performance improvement, we applied our blind feedback algorithm to the query results.

BKBIEF2 (Berkeley Bilingual English against French Automatic Run 2). We submitted the English queries (all fields) without any translation to the French collections and used the blind feedback algorithm for performance improvement. Collection indexing remained the same.

BKBIGF1 (Berkeley Bilingual German against French Automatic Run 1). We translated the German queries with two translation programs: the Systran translator (Altavista Babelfish) and L & H's Power translator. The translations were pooled together and the term frequencies of words occurring twice or more divided (to avoid overemphasis of terms that were translated the same by both programs). The title and description fields of the topics were indexed and searched against the French collections. Again, a blind feedback algorithm was applied. Collection indexing remained the same.

BKBIRF1 (Berkeley Bilingual Russian against French Automatic Run 1). We translated the Russian queries with two translation programs: the Systran translator (Altavista Babelfish) and the Promt (http://www.translate.ru/) translator. The Promt translator translated the queries directly from Russian to French, whereas in the Systran translation, we used an intermediate step from the Russian translation to an English translation to than translate further to French (i.e. English is used as a pivot language). The translations were pooled and the title and description fields submitted to the collection. Our blind feedback algorithm was applied. Collection indexing remained the same.

4.4 Bilingual to German Documents

Our runs for the CLEF bilingual-to-German main task (as well as monolingual German runs) are summarized in Table 4.

Run name	BKMLGG1	BKMLGG2	BKBIFG1	BKBIFG2	BKBIEG1	BKBIEG2	BKBIRG1	BKBIRG2
Retrieved	50000	50000	50000	50000	50000	50000	50000	50000
Relevant	1938	1938	1938	1938	1938	1938	1938	1938
Rel Ret.	1705	1734	1798	1760	1628	1661	1351	1260
Precision								
at 0.00	0.7686	0.7670	0.8141	0.8122	0.7108	0.6625	0.5638	0.5051
at 0.10	0.6750	0.6161	0.7345	0.6959	0.6190	0.6011	0.5055	0.4029
at 0.20	0.6257	0.5836	0.6959	0.6219	0.5594	0.5595	0.4565	0.3779
at 0.30	0.5654	0.5352	0.5947	0.5565	0.5207	0.5075	0.4141	0.3417
at 0.40	0.5367	0.4983	0.5490	0.5174	0.4741	0.4642	0.3761	0.3202
at 0.50	0.5018	0.4753	0.4851	0.4596	0.4358	0.4359	0.3408	0.2923
at 0.60	0.4722	0.4426	0.4465	0.4226	0.4090	0.4105	0.3122	0.2685
at 0.70	0.4239	0.4027	0.3833	0.3637	0.3647	0.3588	0.2687	0.2375
at 0.80	0.3413	0.3406	0.3084	0.3010	0.2972	0.3061	0.2253	0.1906
at 0.90	0.2642	0.2445	0.2289	0.2191	0.2204	0.2172	0.1659	0.1366
at 1.00	0.1681	0.1451	0.1271	0.1256	0.1441	0.1140	0.0927	0.0720
Bky Avg.	0.4696	0.4404	0.4722	0.4448	0.4150	0.4060	0.3254	0.2691

Table 4. Results of official Bilingual to German runs for CLEF-2002.

BKBIEG1 (Berkeley Bilingual English against German Automatic Run 1). We translated the English queries with two translation programs: the Systran translator (Altavista Babelfish) and L & H's Power translator. The translations were pooled together and the term frequencies of words occurring twice or more divided (to avoid overemphasis of terms that were translated the same by both programs). We used the German decompounding procedure to split compounds in

the collections and the queries. All query fields were indexed and searched against the German collections. A blind feedback algorithm was applied.

BKBIEG2 (Berkeley Bilingual English against German Automatic Run 1). This resembles BKBIEG1, except that we only submitted the title and description fields of the topics to the German collections.

BKBIFG1 (Berkeley Bilingual French against German Automatic Run 1). We used the same procedures as for the BKBIEG1 run.

BKBIFG2 (Berkeley Bilingual French against German Automatic Run 2). We used the same procedures as for the BKBIEG2 run.

BKBIRG1 (Berkeley Bilingual Russian against German Automatic Run 1). We translated the Russian queries with two translation programs: the Systran translator (Altavista Babelfish) and the Prompt translator. The Prom translator translated the queries directly from Russian to German, whereas we used an intermediate step from the Russian translation to an English translation to translate further to German. The translations were pooled and the topics (all fields) submitted to the collection. As before, we used German decompounding for indexing the collections and blind feedback to improve our results.

BKBIRG2 (Berkeley Bilingual Russian against German Automatic Run 2). This resembles BKIRG1, except that we only submitted the title and description fields of the topics to the German collections.

5 Summary and Acknowledgments

For CLEF-2002, the Berkeley group one concentrated on two document languages, French and German, and three document collections, Amaryllis, GIRT and CLEF main (French and German newspapers). We worked with four topic languages, English, French, German and Russian. For the three tasks where we worked with Russian as a topic language (GIRT, bilingual Russian to French, and bilingual Russian to German) Russian bilingual consistently underperformed other bilingual topic languages. Why this is the case needs further in-depth investigation. Interestingly enough in the bilingual-to-German documents task, our French topics slightly outperformed our monolingual German runs, retrieving considerably more relevant documents in the top 1000.

Another major focus of our experimentation was to determine the utility of controlled vocabulary and thesauri in cross-language information retrieval. We did experiments with both the Amaryllis and GIRT collections utilizing thesaurus matching techniques. Our results do not show any particular advantage to thesaurus matching over straight translation when machine translation is available; however a preliminary look at individual queries shows that thesaurus matching can be a big win sometimes. We are beginning a detailed analysis of individual queries in the CLEF tasks.

This research was supported by research grant number N66001-00-1-8911 (Mar 2000-Feb 2003) from the Defense Advanced Research Projects Agency (DARPA) Translingual Information Detection Extraction and Summarization (TIDES) program. within the DARPA Information Technology Office. We thank Aitao Chen supplying us with his German decompounding software.

References

1. 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.
2. Hannelore Schott (ed.). *Thesaurus for the Social Sciences. [Vol. 1:] German-English. [Vol. 2:] English-German. [Edition] 1999.* InformationsZentrum Sozialwissenschaften Bonn, 2000.
3. William Hersh, Susan Price, and Larry Donohoe. Assessing Thesaurus-Based Query Expansion Using the UMLS Metathesaurus. In *Proceedings of the 2000 American Medical Informatics Association (AMIA) Symposium*, 2000.