Berkeley2 at GeoCLEF: Cross-Language Geographic Information Retrieval of German and English Documents

Fredric Gey [1] and Vivien Petras[2]

[1] UC Data Archive & Technical Assistance
[2] School of Information Management and Systems
University of California, Berkeley, CA, USA
gey@berkeley.edu, vivienp@sims.berkeley.edu

ABSTRACT: Berkeley group 2 participated in the GeoCLEF pilot track of CLEF 2005 by performing English and German monolingual geographic information retrieval and bilingual geographic information retrieval using English topics to retrieve German documents and vice versa. Our results were substantially above the medians for all categories. We found that major improvements for German retrieval derived from successful blind feedback (between 53 and 72 percent improvement in MAP depending upon the run type) rather than decompounding of German compound words. Manual expansion of geographic references (e.g. Europe to individual country names) was disastrous to retrieval performance. Clearly, successful geographic information retrieval will require substantially more research to improve upon heuristic approaches.

Categories and Subject Descriptors
H.3 [Information Storage and Retrieval]: H.3.3 Information Search and Retrieval -- Query Formulation; H.3.4 Systems and Software -- Performance evaluation (efficiency and effectiveness); H.3.7 Digital Libraries

General Terms
Measurement, Performance, Experimentation

Keywords
Geographic Information Retrieval, Cross-language information retrieval

1 Introduction

GeoCLEF is a new track for the CLEF 2005 campaign and workshop. Geographic Information Retrieval (GIR) tries to exploit geographical references in text for targeted and improved retrieval. The topics presented in GeoCLEF differ from the usual ad-hoc topic layout by including additional tags customized for GIR. Each topic had a concept tag, a spatial relation tag and one or more location tags augmenting the usual title, description and narrative query structure. GeoCLEF provided English and German document collections and English, German, Portuguese and Spanish topic files. Berkeley group 2 participated in monolingual retrieval for both English and German as well as bilingual retrieval from English to German and vice versa. We experimented with the additional location-specific tags by including them into the query formulation and by manually manipulating the location tag to add more specific location names. For example, the location "Europe" was expanded to include the names of the individual countries in Europe. In all cases our retrieval was augmented with blind feedback.

2 Document Ranking, Collection and Query Processing and Translation

In all its CLEF submissions, the Berkeley 2 group used a document ranking algorithm based on logistic regression first used in the TREC-2 conference [1]. The document collections for GeoCLEF consisted of standard CLEF document collections from past CLEFs covering the time periods of 1994 and 1995. The English collections are the Los Angeles Times 1994 and the Glasgow Herald 1995. The German collections are the SDA Swiss news wire (1994 and 1995), Frankfurter Rundschau and Der Spiegel. Berkeley2's English stopword list consists of 662 common English words whose origin is lost in the antiquities of the early TREC conference. Berkeley2's German stopword list consists of 777 common German words developed over several CLEF evaluations. The stemmers used for GeoCLEF are the Muscat project stemmers for both English and German, also used in previous CLEF evaluations. Since Muscat is no longer open source and the English Muscat stemmer was developed by Martin Porter, very similar freely available stemmers may now be found among the SNOWBALL family: http://snowball.tartarus.org. In all official runs for GeoCLEF we utilized a blind feedback

algorithm developed by Aitao Chen [2,3], adding 30 top-ranked terms from the top 20 ranked documents of an initial ranking. Thus the sequence of processing for retrieval is: query → stopword removal → (decompounding) → stemming → ranking → blind feedback. For German runs, we used a decompounding procedure developed and also described by Aitao Chen in [2, 3], which has been shown to improve retrieval results. The decompounding procedure looks up document and query words in a base dictionary and splits compounds when found. We discuss the impacts of German decompounding and blind feedback in the results section below.

3 Runs and Results

3.1 Monolingual Retrieval

For monolingual retrieval, we submitted one title and description run, one run with title, description and narrative, one with title, description, concept and location tag and one with title, description, concept and the manually expanded location tag.

In English monolingual, adding the geographical tags (BKGeoE1) achieved the highest result with a MAP of 0.3936, but the manual expansion strategy did not improve the average precision (BKGeoE4 0.3550). The TDN run (BKGeoE3) outperforms the TD run (BKGeoE4) by 8% and improves from 0.3528 to 0.3840.

In German monolingual retrieval, 4 topics did not retrieve any relevant documents overall. Additionally, our runs failed to retrieve any relevant documents for 3 more of the remaining 21 queries. Manually adding location information lowered the average precision score considerably. The TDN run (BKGeoD3) achieved the highest MAP with 0.2042 followed by the TD run (BKGeoD2) with 0.1608. The manual expansion strategy (BKGeoD4) achieved the lowest MAP (0.1112), whereas adding the tags achieved a MAP of 0.1545. Because a significant proportion of topics retrieved very few relevant documents from the German collection, this might explain these low precision scores.

3.2 Bilingual Retrieval

For bilingual retrieval, we used the L&H Power Translator Pro to translate the topics from English to German and vice versa. In bilingual retrieval, adding the concept and location information improved the average precision score modestly. For English → German, adding the concept and location tag improved precision from 0.1685 to 0.1788, a performance that is better than the same strategy in monolingual retrieval! For German → English, adding the tags improved the average precision from 0.3586 (this TD run is even slightly better than the monolingual one) to 0.3715 in average precision.

3.3 Impact of Blind Feedback and German Decompounding

Since our best results were considerably above an average of medians for both English and German monolingual and bilingual runs, we ran an additional set of experiments to see if we might isolate the effects of blind feedback and (for German) decompounding. What we found was that there was little effect of blind feedback on the English monolingual and German → English bilingual results. Without blind feedback, English monolingual title-description (TD) run mean average precisions are virtually indistinguishable, while blind feedback title-description plus concept-location is about 5% better (0.3936 versus 0.3737). The blind feedback results for English are summarized in Tables 1a (monolingual) and 1b (bilingual German → English):

Table 1a: Berkeley2 GeoCLEF English Monolingual

Run Name	Type	MAP- blind feedback (BF)	MAP-no BF
BKGeoE1	TD+Concept/Locat. (CL)	0.3936† (+5.3%)	0.3737
BKGeoE2	TD	0.3528† (-0%)	0.3536
BKGeoE3	TDN	0.3840† (+3.8%)	0.3701
BKGeoE4	TD+CL manual	0.3550† (+7.6%)	0.3348

Table 1b: Berkeley2 GeoCLEF German→ English Bilingual

Run Name	Type	MAP-BF	MAP-no BF
BKGeoDE1	TD	0.3586† (+8.8%)	0.3296
BKGeoDE2	TD+CL	0.3715† (+12.6%)	0.3298

† Official run

There is however, considerably greater impact of blind feedback on German monolingual and bilingual results, as Tables 2a and 2b show, on the order of 53 to 72 percent improvement:

Table 2a: Berkeley2 GeoCLEF German Monolingual

Run Name	Type	MAP- blind feedback with decompounding	MAP-no BF, with decompounding	MAP- blind feedback no decompounding	MAP-no BF, no decom-pounding
BKGeoD1	TD+CL	0.1545† (+65.1%)	0.0936 (0%)	0.1547 (+65.1%)	0.0937
BKGeoD2	TD	0.1608† (+71.6%)	0.0937 (0%)	0.1613 (+72.1%)	0.0937
BKGeoD3	TDN	0.2042† (+53.5%)	0.1330 (0%)	0.2012 (+53.1%)	0.1330
BKGeoD4	TD+CL manual	0.1112† (+56.1%)	0.0711 (0%)	0.1116 (+56.7%)	0.0712

Table 2b: Berkeley2 GeoCLEF English → German Bilingual (with decompounding)

Run Name	Type	MAP-BF	MAP-no BF
BKGeoED1	TD	0.1685† (+52.6%)	0.1104
BKGeoED2	TD+CL	0.1788† (+57.3%)	0.1137

† Official run

3.4 Source of Improvement when using Blind Feedback

To try to understand how blind feedback produced such stunning improvement in results, we need to make a more detailed examination of improvement produced for each topic. Table 3 presents MAP of our German monolingual runs for each topic, with Median, official TD and TD without blind feedback highlighted
The four queries, where query expansion through blind feedback achieved the most improvement were 10 (Hochwasser in Holland und Deutschland, BF strategy improves by 1400%), 14 (Umweltschädigende Vorfälle in der Nordsee, BF improves by 650%) and 19 (Golfturniere in Europa, BF improves by 285%) and 13 (Besuche des amerikanischen Präsidenten in Deutschland, 168%). Query 12 is an example where blind feedback has a negative effect on the average precision scores (Kathedralen in Europa, -67%).

The blind feedback algorithm adds 30 terms to the query, which are weighted half compared to the original query terms in retrieval. “Good” terms to be added are terms that are relevant to the query and add new information to the search, for example synonyms of query terms but also proper names or word variations. The most improved queries seem to add mostly proper names and word variations and very few irrelevant words that won’t distort the search towards another direction.

For query 10, some of the words added by blind feedback were Hochwassergebiet (flooded area), Waal, Maas (rivers in Holland), Deich (levee) and Flut (flood) – all words that didn’t occur in the title and description tags of the original query but are eminently important words for the search.

For query 12, only a few original query words (after stopword removal) were fed into the blind feedback algorithm: Kathedrale, Europa, Artikel and einzeln, of which the last two don’t add relevant information to the search. Consequently, the suggested blind feedback terms don’t really fit the query (e.g. Besucherinnen (female visitors), kunstvoll (artful), Aussöhnung (reconciliation), Staatsbesuch (state visit), Parade).

The more words are used to feed the blind feedback algorithm and the more distinctive they are in terms of occurrence in the collection and connectedness to a certain concept, the better the blind feedback algorithm will work. For example, the word Golfturnier doesn’t occur very frequently in the collection but it always co-occurs with articles that are related to golf, whereas Besucherinnen will be used in more frames (concepts) than just the European cathedrals.

Table 3: GeoCLEF German monolingual runs with no blind feedback comparison

GeoCLEF Topic ID	Best Overall Monolingual	Median Overall Monolingual	BKGeoD2 TD	TD,decomp, No Blind Feedback	BKGeoD1 TD+CL	BKGeoD3 TDN	BKGeoD4 TD with manual
‡1	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
2	0.1506	0.0018	0.008	0.0188	0.0141	0.0067	0.0000
3	0.6713	0.2336	0.2942	0.2902	0.3145	0.3579	0.0491
4	0.6756	0.0627	0.0335	0.0324	0.0626	0.6756	0.0005
5	0.5641	0.0988	0.095	0.1599	0.0988	0.4705	0.0988
6	0.3929	0.0066	0.0000	0.0000	0.0000	0.0001	0.0000
7	0.1907	0.0539	0.1033	0.0879	0.1405	0.0581	0.0005
8	0.5864	0.0003	0.0000	0.0010	0.0000	0.0005	0.0000
9	0.6273	0.5215	0.523	0.4684	0.5413	0.6273	0.5413
10	0.7936	0.0782	0.6349	0.0452	0.614	0.7936	0.6140
11	0.2342	0.0041	0.0000	0.0000	0.0000	0.0000	0.0000
12	0.2956	0.1007	0.0457	0.1387	0.0759	0.1003	0.1237
13	0.5682	0.2466	0.5682	0.3377	0.4554	0.525	0.4554
14	0.7299	0.0717	0.7299	0.1121	0.3665	0.452	0.3665
15	0.3630	0.235	0.1787	0.1345	0.2130	0.1479	0.2130
16	0.4439	0.0939	0.0651	0.0902	0.0930	0.0821	0.0930
17	0.2544	0.0421	0.0211	0.0555	0.0633	0.2499	0.0633
18	0.1111	0.0087	0.0128	0.0026	0.0139	0.0200	0.0139
19	0.6488	0.1271	0.6014	0.2108	0.6488	0.3972	0.0000
‡20	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
21	0.1123	0.0744	0.0961	0.1324	0.1046	0.1038	0.1046
‡22	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
23	0.1682	0.0000	0.0006	0.0055	0.0023	0.0000	0.0023
24	0.0410	0.0086	0.0086	0.0181	0.0396	0.0364	0.0396
‡25	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Average	0.3449	0.0828	0.1608	0.0937	0.1545	0.2042	0.1112

‡ GeoCLEF topics with no relevant German documents

The combined queries 10, 13, 14, 19 account for almost all of the improvement in the average precision score between the run without blind feedback and the run with blind feedback. This is a thought provoking fact because for the rest of the queries the impact of the blind feedback terms in precision for each query centers around zero. We have found over and over again that blind feedback improves precision, but it seems to do so for only a particular kind of query.

4 Failure Analysis

Manual expansion of general geographic regions to individual country names was a clear losing strategy. For topics 2 and 4, expanding the German location name "Europa" to Frankreich, Griechenland, Grossbritannien, Holland, Italien, Irland, Malta, Österreich, Portugal, Russland, Skandinavien, Dänemark, Finnland, Norwegen, Schweden, Spanien, Tschechien, Ungarn, Zypern, Belgien, Bulgarien, Deutschland, Großbritannien, Kroatien, Liechtenstein, Litauen, Monaco, Niederlande, Polen, Schweiz and Slowenien turned reasonable retrievals go to zero precision for those topics. Similarly poor results were obtained from equivalent English monolingual expansion of "Europe." For topic 3, "Latin America" was expanded to 42 country names with equally dismal results. This does not bode well for using a geographic thesaurus to automatically obtain such expansions.

5 Discussion

Berkeley Group 2 participated in the GeoCLEF track with a focus on the German and English languages for both documents and topics. We utilized standard information retrieval techniques of blind feedback and German complex word decompounding. Query translation used commonly available machine translation software. Blind feedback was particularly impressive in improving German monolingual and bilingual English → German results. Our venture into geographic location resolution by manually expansion of the general terms "Europe"

and “Latin America” into a list of individual country names resulted in a considerably diminished performance effectiveness. It seems that successful geographic expansion will only occur in the context of requiring the concept (e.g. “Golf Tournaments”) to also be present in the documents. This may argue for a generalized Boolean approach to retrieval.

6 Acknowledgement

Thanks to Aitao Chen for implementing and permitting the use of the logistic regression formula for probabilistic information retrieval as well as German decompounding and blind feedback in his MULIR retrieval system.

7 References

1. Chen, A, W Cooper, and F Gey. 1994. Full text retrieval based on probabilistic equations with coefficients fitted by logistic regression. In *The Second Text Retrieval Conference (TREC-2)*, edited by D. K. Harman.
2. Chen, Aitao. 2003. *Cross-Language Retrieval Experiments at CLEF 2002*. *Lecture Notes in Computer Science* 2785, Springer 2003.
3. Chen, A, and F Gey. 2004. Multilingual Information Retrieval Using Machine Translation, Relevance Feedback and Decompounding. *Information Retrieval* 7 (1-2):149-182.