Berkeley at GeoCLEF: Logistic Regression and Fusion for Geographic Information Retrieval

Ray R. Larson[1], Fredric C. Gey[2], and Vivien Petras[1]

[1]*School of Information Management and Systems*
vivienp@sims.berkeley.edu / ray@sims.berkeley.edu
[2]*UC Data Archive & Technical Assistance*
gey@berkeley.edu
University of California, Berkeley, CA 94720 USA

This is an author's accepted manuscript version of a conference paper published in *International Conference of the Cross-Language Evaluation Forum for European Languages, CLEF 2005: Accessing Multilingual Information Repositories* within the Springer Lecture Notes in Computer Science book series (LNCS, volume 4022).

The final publisher's version is available online at:
https://doi.org/10.1007/11878773_108

Berkeley at GeoCLEF: Logistic Regression and Fusion for Geographic Information Retrieval

Ray R. Larson[1], Fredric C. Gey[2] and Vivien Petras[1]

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

Abstract. In this paper we will describe the Berkeley (groups 1 and 2 combined) submissions and approaches to the GeoCLEF task for CLEF 2005. The two Berkeley groups used different systems and approaches for GeoCLEF with some common themes. For Berkeley group 1 (Larson) the main technique used was fusion of multiple probabilistic searches against different XML components using both Logistic Regression (LR) algorithms and a version of the Okapi BM-25 algorithm. The Berkeley group 2 (Gey and Petras) employed tested CLIR methods from previous CLEF evaluations using Logistic Regression with Blind Feedback. Both groups used multiple translations of queries in for cross-language searching, and the primary geographically-based approaches taken by both involved query expansion with additional place names. The Berkeley1 group used GIR indexing techniques to georeference proper nouns in the text using a gazetteer derived from the World Gazetteer (with both English and German names for each place), and automatically expanded place names in topics for regions or countries in the queries by the names of the countries or cities in those regions or countries. The Berkeley2 group used manual expansion of queries, adding additional place names.

1 Introduction

For GeoCLEF 2005 the Berkeley IR research group split into two groups (Berkeley1 and Berkeley2). Berkeley2 used the same techniques as used in previous CLEF evaluations with some new query expansion experiments for GeoCLEF, while Berkeley1 tried some alternative algorithms and fusion methods for both the GeoCLEF and Domain Specific tasks. This paper will describe the results of both on the techniques used by the Berkeley1 group for GeoCLEF and the results of our official submissions, as well as some additional tests using versions of the algorithms employed by the Berkeley2 group. The main technique being tested is the fusion of multiple probabilistic searches against different XML components using both Logistic Regression (LR) algorithms and a version of the Okapi BM-25 algorithm. We also combine multiple translations of queries in

cross-language searching. Since this is the first time that the Cheshire II system has been used for CLEF, this approach can at best be considered a very preliminary base testing of some retrieval algorithms and approaches. This paper is organized as follows: In the next section we discuss the retrieval algorithms and fusion methods used by the Berkeley1 group for the submitted runs. We then discuss the Berkeley2 group algorithms. We will then discuss the specific approaches taken for indexing and retrieval in GeoCLEF and the results of the submitted runs for each group. We also compare our official submitted results to some additional runs with alternate approaches conducted later. Finally we present conclusions and discussion of lessons learned in GeoCLEF 2005.

2 Berkeley1 Retrieval Algorithms and Fusion Operators

The algorithms and fusion combination methods used by the Berkeley1 group are implemented as part of the Cheshire II XML/SGML search engine, as described in [7] and in the CLEF notebook paper[6]. The system also supports a number of other algorithms for distributed search and operators for merging result lists from ranked or Boolean sub-queries.

2.1 Logistic Regression Algorithm

The basic form and variables of the *TREC3 Logistic Regression* (LR) algorithm used by Berkeley1 was originally developed by Cooper, et al.[3]. It provided good full-text retrieval performance in the TREC3 ad hoc task and in TREC interactive tasks [4] and for distributed IR [5]. As originally formulated, the LR model of probabilistic IR attempts to estimate the probability of relevance for each document based on a set of statistics about a document collection and a set of queries in combination with a set of weighting coefficients for those statistics. The statistics to be used and the values of the coefficients are obtained from regression analysis of a sample of a collection (or similar test collection) for some set of queries where relevance and non-relevance has been determined.

Much of our recent focus for the Cheshire II system has been on exploiting the structure of XML documents (including the GeoCLEF documents) as a tree of XML elements. We define a "document component" as an XML subtree that may include zero or more subordinate XML elements or subtrees with text as the leaf nodes of the tree. Naturally, a full XML document may also be considered a "document component". As discussed below, the indexing and retrieval methods we used take into account a selected set of document components for generating the statistics used in the search ranking process. Because we are dealing with not only full documents, but also document components, the algorithm that we use is geared toward estimating the probability of relevance for a given document component. The complete formal description of the algorithm used can be found in [7] or in the Berkeley1 GeoCLEF notebook paper[6].

We also use a version of the Okapi BM-25 algorithm in these experiments that is based on the description of that algorithm by Robertson [10], using parameters

from the TREC notebook proceedings [9]. As with the TREC3 LR algorithm, we have adapted the Okapi BM-25 algorithm to deal with document components.

The Cheshire II system also provides a number of operators to combine intermediate results of searches from different components or indexes. With these operators we have available an entire spectrum of combination methods ranging from strict Boolean operations to fuzzy Boolean and normalized score combinations for probabilistic and Boolean results. These operators are the means available for performing fusion operations between the results for different retrieval algorithms and the search results from different different components of a document. We will only describe two of these operators here, because they were the only types used in the GEOCLEF runs reported in this paper.

The MERGE_CMBZ operator is based on the "CombMNZ" fusion algorithm developed by Shaw and Fox [11] and used by Lee [8]. In our version we take the normalized scores, but then further enhance scores for components appearing in both lists (doubling them) and penalize normalized scores appearing low in a single result list, while using the unmodified normalized score for higher ranking items in a single list.

The MERGE_PIVOT operator is used primarily to adjust the probability of relevance for one search result based on matching elements in another search result. It was developed primarily to adjust the probabilities of a search result consisting of sub-elements of a document (such as titles or paragraphs) based on the probability obtained for the same search over the entire document. It is basically a weighted combination of the probabilities based on a "DocPivot" fraction, such that:

$$P_n = DocPivot * P_d + (1 - DocPivot) * P_s \quad (1)$$

where P_d represents the document-level probability of relevance, P_s represents the subelement probability, and P_n representing the resulting new probability. The "*DocPivot*" value used for all of the runs submitted was 0.64. Since this was the first year for GeoCLEF, this value was derived from experiments on 2004 data for other CLEF collections (we hope to do further testing to see if the was truly appropriate for the GeoCLEF data). This basic operator can be applied to either probabilistic results, or non-probabilistic results or both (in the latter case the scores are normalized using MINMAX normalization to range between 0 and 1).

In the following subsections we describe the specific approaches taken for our submitted runs for the GeoCLEF task. First we describe the indexing and term extraction methods used, and then the search features we used for the submitted runs.

2.2 Indexing and Term Extraction

For both the monolingual and bilingual tasks we indexed the documents using the Cheshire II system. The document index entries and queries were stemmed using the Snowball stemmer, and a new georeferencing indexing subsystem was used.

This subsystem extracts proper nouns from the text being indexed and attempts to match them in a digital gazetteer. For GeoCLEF we used a gazetteer derived from the World Gazetteer (http://www.world-gazetteer.com) with 224698 entries in both English and German. The indexing subsystem provides three different index types: verified place names (an index of names which matched the gazetteer), point coordinates (latitude and longitude coordinates of the verified place name) and bounding box coordinates (bounding boxes for the matched places from the gazetteer). All three types were created, but due to time constraints we only used the verified place names in our tests. Text indexes were also created for separate XML elements (such as document titles or dates) as well as for the entire document. It is worth noting that, although the names are compared against the gazetteer, it is quite common for proper name of persons and places to be the same and this leads to potential false associations between articles mentioning persons with such name and particular places.

Name	Description	Content Tags	Used
docno	Document ID	DOCNO	no
pauthor	Author Names	BYLINE, AU	no
headline	Article Title	HEADLINE, TITLE, LEAD, LD, TI	yes
topic	Content Words	HEADLINE, TITLE, TI, LEAD BYLINE, TEXT, LD, TX	yes yes
date	Date of Publication	DATE, WEEK	yes
geotext	Validated place names	TEXT, LD, TX	yes
geopoint	Validated coordinates for place names	TEXT, LD, TX	no
geobox	Validated bounding boxes for place names	TEXT, LD, TX	no

Table 1. Cheshire II Indexes for GeoCLEF 2005 (Berkeley1)

Table 1 lists the indexes created for the GeoCLEF database and the document elements from which the contents of those indexes were extracted. The "Used" column in the table indicates whether or not a particular index was used in the official Berkeley1 runs.

Because there was no explicit tagging of location-related terms in the collections used for GeoCLEF, we applied the above approach to the "TEXT", "LD", and "TX" elements of the records of the various collections. The part of news articles normally called the "dateline" indicating the location of the news story was not separately tagged in any of the GeoCLEF collections, but often appeared as the first part of the text for the story. (In addition, we discovered when writing these notes that the "TX" and "LD" were *not* included in the indexing in all cases, meaning that the SDA collection was not included in the German indexing for these indexes).

For all indexing we used English and German stoplists to exclude function words and very common words from the indexing and searching. For the runs reported here, Berkeley1 did not use any decompounding of German terms.

2.3 Berkeley1 Search Processing

For monolingual search tasks we used the topics in the appropriate language (English or German), for bilingual tasks the topics were translated from the source language to the target language using three different machine translation (MT) systems, the L&H Power Translator PC-based system, SYSTRAN (via Babelfish at Altavista), and PROMT (also via their web interface). Each of these translations were combined into a single probabilistic query. The hope was to overcome the translation errors of a single system by including alternatives.

We tried two main approaches for searching, the first used only the topic text from the title and desc elements, the second included the spatialrelation and location elements as well. In all cases the different indexes mentioned above were used, and probabilistic searches were carried out on each index, and the results combined using the CombMNZ algorithm, and by a weighted combination of partial element and full document scores. For bilingual searching we used both the Berkeley TREC3 and the Okapi BM-25 algorithm, for monolingual we used only TREC3. For one submitted run in each task we did no query expansion and did not use the location elements in the topics. For the other runs each of the place names identified in the queries were expanded when that place was the name of a region or country. For example when running search against the English databases the name "Europe" was expanded to "Albania Andorra Austria Belarus Belgium Bosnia and Herzegovina Bulgaria Croatia Cyprus Czech Republic Denmark Estonia Faroe Islands Finland France Georgia Germany Gibraltar Greece Guernsey and Alderney Hungary Iceland Ireland Isle of Man Italy Jersey Latvia Liechtenstein Lithuania Luxembourg Macedonia Malta Moldova Monaco Netherlands Norway Poland Portugal Romania Russia San Marino Serbia and Montenegro Slovakia Slovenia Spain Svalbard and Jan Mayen Sweden Switzerland Turkey Ukraine United Kingdom Vatican City", while for searches against the German databases "Europa" was expanded to "Albanien Andorra Österreich Weißrussland Belgien Bosnien und Herzegowina Bulgarien Kroatien Zypern Tschechische Republik Dänemark Estland Färöer-Inseln Finnland Frankreich Georgien Deutschland Gibraltar Griechenland Guernsey und Alderney Ungarn Island Irland Man Italien Jersey Lettland Liechtenstein Litauen Luxemburg Mazedonien Malta Moldawien Monaco Niederlande Norwegen Polen Portugal Rumänien Russland San Marino Serbien und Montenegro Slowakei Slowenien Spanien Svalbard und Jan Mayen Schweden Schweiz Türkei Ukraine Großbritannien Vatikan".

The indexes combined in searching included the headline, topic, and geotext indexes (as described in Table 1) for searches that include the location element, and the headline and topic for the searches without the locations element. For the bilingual tasks, three sub-queries, one for each query translation were run and then the results were merged using the CombMNZ algorithm. For Monolingual tasks the title and topic results were combined with each other using CombMNZ and the final score combined with an expanded search for place names in the topic and geotext indexes. However, There were some errors in the scripts used to generate the queries used in the official runs. These included things such as

including "Kenya" in the expansion for Europe, and including two copies of all expansion names, when a single copy should have been used. Also in some cases the wrong language form was used in expansions.

3 Berkeley2: Document Ranking, Collection and Query Processing and Translation

In all its CLEF submissions, the Berkeley2 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. The English stopword list used consists of 662 common English words whose origin is lost in the antiquities of the early TREC conference. Berkeley2s 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[1, 2], 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 [1, 2], 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 Berkeley2 results section below.

4 Berkeley1 Results for Submitted GeoCLEF Runs

The summary results (as Mean Average Precision) for the submitted bilingual and monolingual runs for both English and German are shown in Table 2, the Recall-Precision curves for these runs are also shown in Figures 1 (for monolingual) and 2 (for bilingual). In Figures 1 and 2 the name are abbrevated to the final letters and numbers of the full name in Table 2, and those beginning with "POST" are unofficial runs described in the next section.

Table 2 indicates some rather curious results that warrant further investigation as to the cause. Notice that the result for all of the English monolingual runs exceed the results for bilingual German to English runs - this is typical for cross-langauge retrieval. However, in the case of German this expected pattern

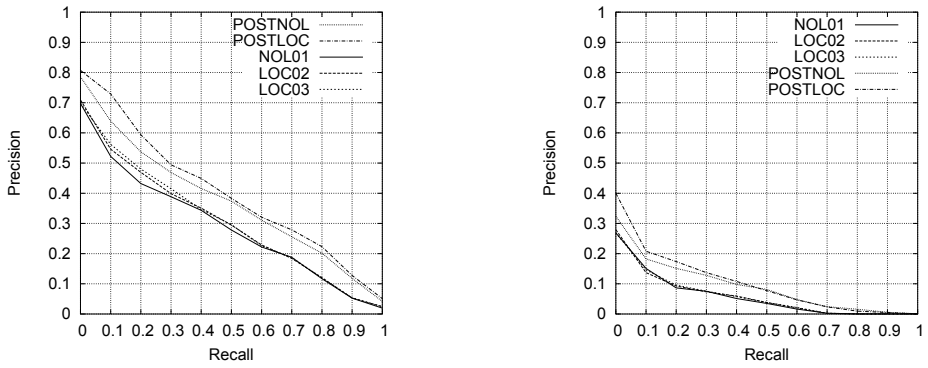

Fig. 1. Berkeley1 Monolingual Runs – English (left) and German (right)

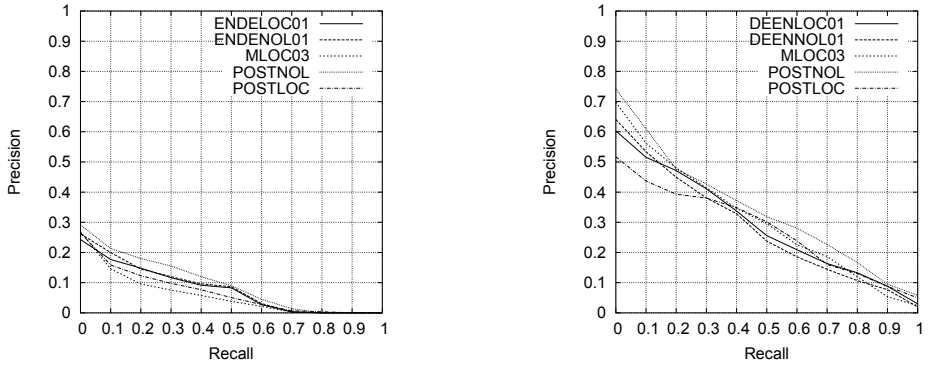

Fig. 2. Berkeley1 Bilingual Runs – English to German (left) and German to English (right)

is reversed, and the German monolingual runs *perform worse* than either of the bilingual English to German runs. We haven't yet determined exactly why this might be the case, but there are number possible reasons (e.g., since a combination of Okapi and Logistic Regression searches are used for the bilingual task this may be an indication that Okapi is more effective for German). Also, in the monolingual runs, both English and German, use of the location tag and expansion of the query (runs numbered LOC02 and LOC03 respectively) did better than no use of the location tag or expansion. For the bilingual runs the results are mixed, with German to English runs showing an improvement with location use and expansion (LOC01) and English to German showing the opposite. However, given the very low scores when compared to the Berkeley2 results below, we suspect that differences in stoplists, decompounding, etc. may have confused the effects.

Run Name	Description	Location	MAP
BERK1BLDEENLOC01	Bilingual German⇒English	yes	0.2753
BERK1BLDEENNOL01	Bilingual German⇒English	no	0.2668
BERK1BLENDELOC01	Bilingual English⇒German	yes	0.0725
BERK1BLENDENOL01	Bilingual English⇒German	no	0.0777
BERK1MLDELOC02	Monolingual German	yes	0.0535
BERK1MLDELOC03	Monolingual German	yes	0.0533
BERK1MLDENOL01	Monolingual German	no	0.0504
BERK1MLENLOC02	Monolingual English	yes	0.2910
BERK1MLENLOC03	Monolingual English	yes	0.2924
BERK1MLENNOL01	Monolingual English	no	0.2794

Table 2. Berkeley1 Submitted GeoCLEF Runs

4.1 Additional Runs

After the official submission we used a version of the same Logistic Regression algorithm as the Berkeley2 group (the "TREC2" algorithm), which incorporates blind feedback (which is currently lacking in the "TREC3" algorithm used in the official runs). This version of the TREC2 algorithm was implemented as another option of the Cheshire II system. The parameters used for blind feedback were 13 documents and the top-ranked 16 terms from those documents added to the original query. This is essentially an identical algorithm to that defined in [1]. The results from the bilingual and monolingual runs for both English and German using this algorithm, but with the remaining processing components the same as in the Berkeley1 official runs, are shown in Table 3, the Recall-Precision curves for these runs are also included in Figures 1 (for monolingual) and 2 (for bilingual). In Figures 1 and 2 the names abbrevated to the final letters of the full name in Table 3, prefixed by "POST". These are unofficial runs to test the difference in the algorithms in an identical runtime environment.

Run Name	Description	Location	MAP
POSTBLDEENEXP	Bilingual German⇒English	yes	0.2636
POSTBLDEENNOL	Bilingual German⇒English	no	0.3205
POSTBLENDEEXP	Bilingual English⇒German	yes	0.0626
POSTBLENDENOL	Bilingual English⇒German	no	0.0913
POSTMLDELOC	Monolingual German	yes	0.0929
POSTMLDENOL	Monolingual German	no	0.0861
POSTMLENEXP	Monolingual English	yes	0.2892
POSTMLENLOC	Monolingual English	yes	0.3879
POSTMLENNOL	Monolingual English	no	0.3615

Table 3. Berkeley1 Additional Post-Submission GeoCLEF Runs

As can be seen by comparing Table 3 with Table 2, all of the comparable runs show improvement in results with the TREC2 algorithm with blind feedback. We have compared notes with the Berkeley2 group and except for minor differences to be expected given the different indexing methods, stoplists, etc. used, the

English monolingual and German⇒English results are comparable to theirs as shown in the tables below.

The queries submitted in these unofficial runs were much simpler than those used in the official runs. For monolingual retrieval only the "topic" index was used and the geotext index was not used at all, for the bilingual runs the same pattern of using multiple query translations and combining the results was used as in our official runs. This may actually be detrimental to the performance, since the expanded queries perform worse than the unexpanded queries - the opposite behaviour observed in the official runs.

In the monolingual runs there appears to be similar behavior, using the topic titles and description along with the location tag provided the best results, but expanding the locations as in the official runs (the English ML run ending in EXP) performed considerably worse than the the unexpanded runs. Also, as in the offical runs the German monolingual and English⇒German bilingual had very poor results. We believe that this indicates a significant processing problem for German (in addition to the lack of decompounding).

5 Berkeley2 Runs and Results

5.1 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. (Note that in the tables a dagger † indicates the official Berkeley2 results).

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 4. Berkeley2 GeoCLEF English Monolingual

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.

GeoCLEF Run Name	Type	MAP BF decomp.	MAP no BF decomp.	MAP BF no decomp	MAP no BF no decomp
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 5. Berkeley2 GeoCLEF German Monolingual

5.2 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.

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

Table 6. Berkeley2 GeoCLEF German→English Bilingual

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

Table 7. Berkeley2 GeoCLEF English → German Bilingual (with decompounding)

5.3 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 4 (monolingual) and 6 (bilingual German → English):

There is however, considerably greater impact of blind feedback on German monolingual and bilingual results, as Tables 5 and 7 show, on the order of 53 to 72 percent improvement.

5.4 Source of Improvement when using Blind Feedback

To try to understand how blind feedback produced such stunning improvement in results (for both groups), we need to make a more detailed examination of improvement produced for each topic. Table 8 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%).

GeoCLEF Topic ID	Best Overall Monoling.	Median Overall Monoling.	BKGeoD2 TD	TD decomp No BF	BKGeoD1 TD+CL	BKGeoD3 TDN	BKGeoD4 TD 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

Table 8. GeoCLEF German monolingual runs with no blind feedback comparison

‡ GeoCLEF topics with no relevant German documents

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.

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.

6 Failure Analysis

Manual expansion of general geographic regions to individual country names was a clear losing strategy. For topics 2 and 4, the German location name "Europa" was expanded using a similar list to that used by Berkeley1, which turned reasonable retrievals go to zero precision for those topics. Similarly poor results were obtained from equivalent English monolingual expansion of "Europe"or topic 3, and "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.

7 Discussion and Conclusions

Berkeley groups participated in the GeoCLEF track with a focus on the German and English languages for both documents and topics. Berkeley2 utilized

standard information retrieval techniques of blind feedback and German complex word decompounding, while Berkeley1 used multiple algorithm fusion approaches and combinations of different document elements in searching. Query translation used commonly available machine translation software. Blind feedback was particularly impressive in improving the Berkeley2 German monolingual and bilingual English→German results and the Berkeley1 "POST" runs. The Berkeley2 venture into geographic location resolution by manual expansion of the general terms "Europe" and "Latin America" into a list of individual country names resulted in a considerably diminished performance effectiveness, which was also seen in the Berkeley1 "POST" runs. However, the message is not entirely unmixed. Expansion appeared to help in cases where fusion of multiple search elements was used. It remains for future experimentation to see whether this was an anomaly, or whether it is a useful property of the fusion algorithms. It does seem clear, however, 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 be something that the combinations of operators and algorithms available in the Cheshire II system can test.

Analysis of these results (and cross analysis of the two groups' results) is still ongoing. There are a number of, as yet, unexplained behaviors in some of our results. We plan to continue working on the use of fusion, and hope to discover effective ways to combine highly effective algorithms, such as the TREC2 algorithm, as well as working on adding the same blind feedback capability to the TREC3 Logistic Regression algorithm.

One obvious conclusion that can be drawn is that basic TREC2 is a highly effective algorithm for the GeoCLEF tasks, and the fusion approaches tried in these tests are most definitely *not* very effective (in sprite of their relatively good effectiveness in other retrieval tasks such as INEX).

Another conclusion is that, in some cases, query expansion of region names to a list of names of particular countries in that region is modestly effective (although we haven't yet been able to test for statistical significance). In other cases, however it can be quite detrimental. However we still need to determine if the problems with the expansion were due the nature of the expansion itself, or errors in how it was done.

Acknowledgements

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 for the Berkeley2 runs.

References

1. Aitao Chen. *Cross-Language Retrieval Experiments at CLEF 2002*, pages 28–48. Springer (LNCS #2785), 2003.

2. Aitao Chen and Fredric C. Gey. Multilingual information retrieval using machine translation, relevance feedback and decompounding. *Information Retrieval*, 7:149–182, 2004.
3. William S. Cooper, Fredric C. Gey, and Daniel P. Dabney. Probabilistic retrieval based on staged logistic regression. In *15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Copenhagen, Denmark, June 21-24*, pages 198–210, New York, 1992. ACM.
4. Ray R. Larson. TREC interactive with cheshire II. *Information Processing and Management*, 37:485–505, 2001.
5. Ray R. Larson. A logistic regression approach to distributed IR. In *SIGIR 2002: Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 11-15, 2002, Tampere, Finland*, pages 399–400. ACM, 2002.
6. Ray R. Larson. Cheshire II at GeoCLEF: Fusion and query expansion for GIR. In *CLEF 2005 Notebook Papers*. DELOS Digital Library, 2005.
7. Ray R. Larson. A fusion approach to XML structured document retrieval. *Information Retrieval*, 8:601–629, 2005.
8. Joon Ho Lee. Analyses of multiple evidence combination. In *SIGIR '97: Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, July 27-31, 1997, Philadelphia*, pages 267–276. ACM, 1997.
9. Stephen E. Robertson, Stephen Walker, and Micheline M. Hancock-Beauliee. OKAPI at TREC-7: ad hoc, filtering, vlc and interactive track. In *Text Retrieval Conference (TREC-7), Nov. 9-1 1998 (Notebook)*, pages 152–164, 1998.
10. Stephen E. Robertson and Steven Walker. On relevance weights with little relevance information. In *Proceedings of the 20th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 16–24. ACM Press, 1997.
11. Joseph A. Shaw and Edward A. Fox. Combination of multiple searches. In *Proceedings of the 2nd Text REtrieval Conference (TREC-2), National Institute of Standards and Technology Special Publication 500-215*, pages 243–252, 1994.