Domain Specific Retrieval: Back to Basics

Ray R. Larson
School of Information
University of California, Berkeley, USA
ray@sims.berkeley.edu

Abstract

In this paper we will describe Berkeley's approach to the Domain Specific (DS) track for CLEF 2006. This year we are *not* using the tools for thesaurus-based query expansion and de-compounding for German that were developed over the past many years and used very successfully in earlier Berkeley entries in this track. Our intent has been to incorporate those tools into the Cheshire system, but we were unable to complete the development in time for use in the officially submitted runs. This year Berkeley submitted 12 runs, including one for each subtask of the DS track. These include 3 Monolingual runs for English, German, and Russian, 7 Bilingual runs (3 X2EN, 1 X2DE, and 3 X2RU), and 2 Multilingual runs. For many DS sub-tasks our runs were the best performing runs, but sadly they were also the *only* runs for a number of sub-tasks. In the sub-tasks where there were other entries, our relative performance was above the mean performance in 2 sub-tasks and just below the mean in another.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval

General Terms

Algorithms, Performance, Measurement

Keywords

Cheshire II, Logistic Regression

1 Introduction

This paper discusses the retrieval methods and evaluation results for Berkeley's participation in the Domain Specific track. Our submitted runs this year are intended to establish a new baseline for comparision in future Domain Specific evaluations for the Cheshire system, and did not use the techniques of thesaurus-based query expansion or German decompounding used in previous years. This year we used only basic probabilistic retrieval methods for DS tasks. We hope, in future years, (assuming that the task will continue in future years) to be able to use those, or refinements of those techniques via the Cheshire II or Cheshire3 systems.

This year Berkeley submitted 12 runs, including one for each subtask of the DS track. These include 1 Monolingual run for each of English, German, and Russian for a total of 3 Monolingual runs, and 7 Bilingual runs (3 X2EN, 1 X2DE, and 3 X2RU), and 2 Multilingual runs.

This paper first very briefly describes the retrieval methods used, including our blind feedback method for text, which are discussed in greater detail in our ImageCLEF paper. We then describe our submissions for the various DS sub-tasks and the results obtained. Finally we present conclusions and discussion of future approaches to this track.

2 The Retrieval Algorithms

As we have discussed in our other papers for the ImageCLEF and GeoCLEF tracks in this volume, basic form and variables of the *Logistic Regression* (LR) algorithm used for all of our submissions were originally developed by Cooper, et al. [3]. To formally the LR method, the goal of the logistic regression method is to define a regression model that will estimate (given a set of training data), for a particular query Q and a particular document D in a collection the value $P(R \mid Q, D)$, that is, the probability of relevance for that Q and D. This value is then used to rank the documents in the collection which are presented to the user in order of decreasing values of that probability. To avoid invalid probability values, the usual calculation of $P(R \mid Q, D)$ uses the "log odds" of relevance given a set of S statistics, s_i, derived from the query and database, giving a regression formula for estimating the log odds from those statistics:

$$\log O(R \mid Q, D) = b_0 + \sum_{i=1}^{S} b_i s_i \tag{1}$$

where b_0 is the intercept term and the b_i are the coefficients obtained from the regression analysis of a sample set of queries, a collection and relevance judgements. The final ranking is determined by the conversion of the log odds form to probabilities:

$$P(R \mid Q, D) = \frac{e^{\log O(R|Q,D)}}{1 + e^{\log O(R|Q,D)}} \tag{2}$$

2.1 TREC2 Logistic Regression Algorithm

For all of our Domain Specific submissions this year we used a version of the Logistic Regression (LR) algorithm that has been used very successfully in Cross-Language IR by Berkeley researchers for a number of years[1] and which is also used in our GeoCLEF and Domain Specific submissions. For the Domain Specific track we used the Cheshire II information retrieval system implementation of this algorithm. One of the current limitations of this implementation is the lack of decompounding for German documents and query terms in the current system. As noted in our other CLEF notebook papers, the Logistic Regression algorithm used was originally developed by Cooper et al. [2] for text retrieval from the TREC collections for TREC2. The basic formula is:

$$\begin{aligned}
\log O(R|C,Q) &= log\frac{p(R|C,Q)}{1-p(R|C,Q)} = log\frac{p(R|C,Q)}{p(\overline{R}|C,Q)} \\
&= c_0 + c_1 * \frac{1}{\sqrt{|Q_c|}+1}\sum_{i=1}^{|Q_c|}\frac{qtf_i}{ql+35} \\
&+ c_2 * \frac{1}{\sqrt{|Q_c|}+1}\sum_{i=1}^{|Q_c|}\log\frac{tf_i}{cl+80} \\
&- c_3 * \frac{1}{\sqrt{|Q_c|}+1}\sum_{i=1}^{|Q_c|}\log\frac{ctf_i}{N_t} \\
&+ c_4 * |Q_c|
\end{aligned}$$

where C denotes a document component (i.e., an indexed part of a document which may be the entire document) and Q a query, R is a relevance variable,

$p(R|C,Q)$ is the probability that document component C is relevant to query Q,

$p(\overline{R}|C,Q)$ the probability that document component C is *not relevant* to query Q, which is 1.0 - $p(R|C,Q)$

$|Q_c|$ is the number of matching terms between a document component and a query,

qtf_i is the within-query frequency of the ith matching term,

tf_i is the within-document frequency of the ith matching term,

ctf_i is the occurrence frequency in a collection of the ith matching term,

ql is query length (i.e., number of terms in a query like $|Q|$ for non-feedback situations),

cl is component length (i.e., number of terms in a component), and

N_t is collection length (i.e., number of terms in a test collection).

c_k are the k coefficients obtained though the regression analysis.

More details of this algorithm and the coefficients used with it may be found in our ImageCLEF notebook paper where the same algorithm and coefficients were used. In addition to this primary algorithm we used a version that performs "blind feedback" during the retrieval process. The method used is described in detail in our ImageCLEF notebook paper. Our blind feedback approach uses the 10 top-ranked documents from an initial retrieval using the LR algorithm above, and selects the top 10 terms from the content of those documents, using a version of the Robertson and Sparck Jones probabilistic term relevance weights [5]. Those ten terms are merged with the original query and new term frequency weights are calculated, and the revised query submitted to obtain the final ranking.

3 Approaches for Domain Specific

In this section we describe the specific approaches taken for our submitted runs for the Domain Specific track. First we describe the indexing and term extraction methods used, and then the search features we used for the submitted runs.

3.1 Indexing and Term Extraction

Although the Cheshire II system uses the XML structure of documents and extracts selected portions of the record for indexing and retrieval, for the submitted runs this year we used only a single one of these indexes that contains the entire content of the document.

Name	Description	Content Tags	Used
docno	Document ID	DOCNO	no
title	Article Title	TITLE	no
topic	All Content Words	DOC	yes
date	Date	DATE	no
geoname	Geographic names	GEOGR-AREA, COUNTRY-CODE	no
subject	Controlled Vocabulary	CONTROLLED-TERM-x, CLASSIFICATION-TEXT-x	no

Table 1: Cheshire II Indexes for Domain Specific 2006

Table 1 lists the indexes created for the Domain Specific database and the document elements from which the contents of those indexes were extracted. The "Used" column in Table 1 indicates whether or not a particular index was used in the submitted Domain Specific runs. This year we

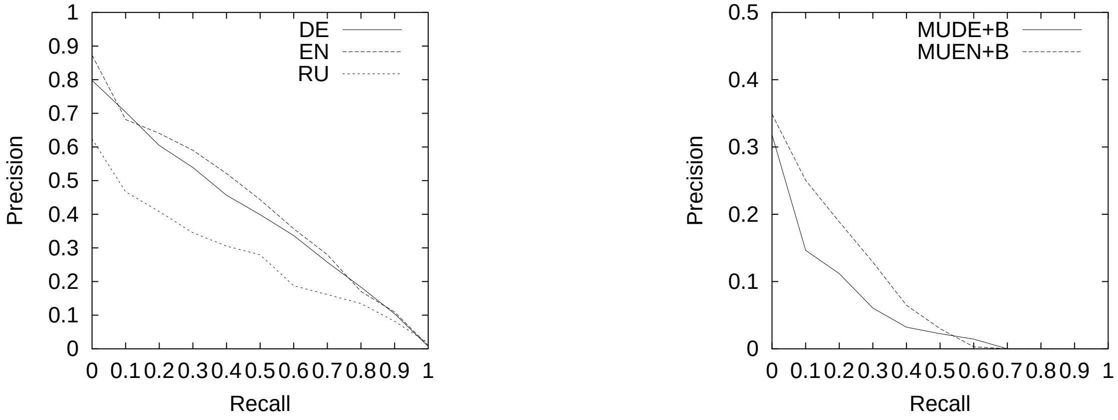

Figure 1: Berkeley Domain Specific Monolingual Runs(left) and Multilingual Runs (right)

did not use the Entry Vocabulary Indexes (search term recommender) that were used by Berkeley in previous years (see [4]), this certainly had an impact on our results, seen in a large drop in average precision when compared to similar runs using these query expansion strategies. We hope to be able to enable and test some of these strategies further using the new retrieval system in time for presentation at the meeting.

For all indexing we used language-specific stoplists to exclude function words and very common words from the indexing and searching. The German language runs, however, did *not* use decompounding in the indexing and querying processes to generate simple word forms from compounds (actually we tried, but there was a bug that failed to match any compounds in our runs). This is another aspect of our indexing for this year's Domain Specific task that reduced our results relative to last year.

3.2 Search Processing

Searching the Domain Specific collection used Cheshire II scripts to parse the topics and submit the title and description elements from the topics to the "topic" index containing all terms from the documents. For the monolingual search tasks we used the topics in the appropriate language (English, German, or Russian), and for bilingual tasks the topics were translated from the source language to the target language using SYSTRAN (via Babelfish at Altavista.com) or PROMT via the PROMT web interface. We believe that other translation tools provide a more accurate representation of the topics for some languages (like the L&H P.C. translator used in our GeoCLEF entries) but that was not available to us for our official runs for this track this year. Wherever possible we used both of these MT systems and submitted the resulting translated queries as separate runs. However, some translations are only available on one system or the other (e.g., German *Rightarrow* Russian is only available in PROMT). All of our runs for this track used the TREC2 algorithm as described above with blind feedback using the top 10 terms from the 10 top-ranked documents in the initial retrieval.

4 Results for Submitted 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 Figure 1 (for monolingual and multilingual) and Figures 2 and 3 (for bilingual). In Figures 1, 2, and 3 the names are abbrevated to the letters and numbers of the full name in Table 2 describing the languages and translation system used. For example, in Figure 3 DERU+P corresponds to BERK_BI_DERU_T2FB_P in Table 2.

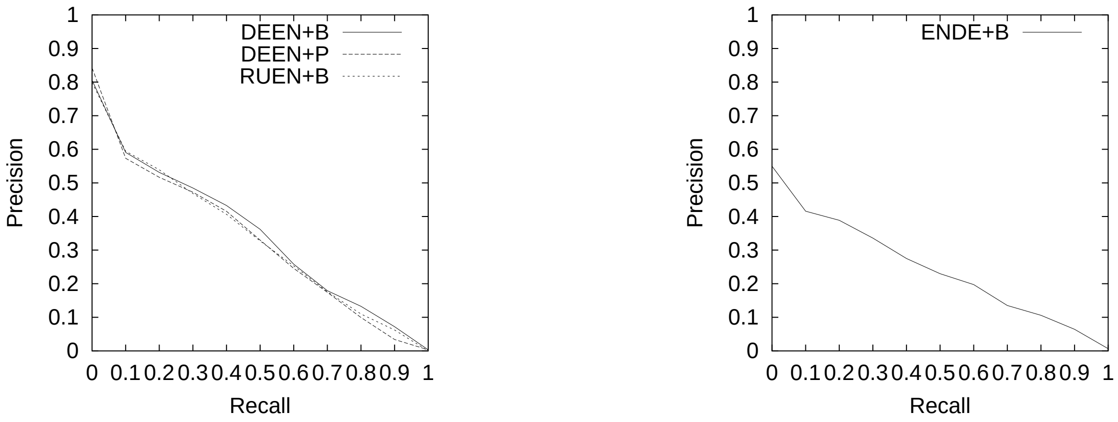

Figure 2: Berkeley Domain Specific Bilingual Runs – To English (left) and To German (right)

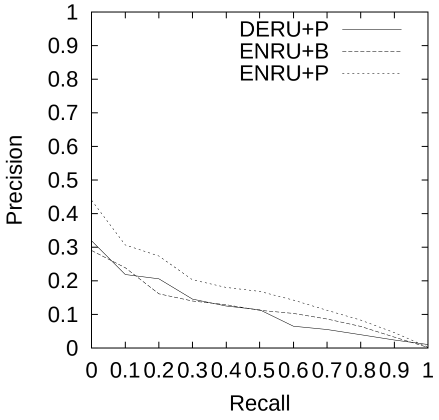

Figure 3: Berkeley Domain Specific Bilingual Runs – To Russian

Table 2 shows all of our submitted runs for the Domain Specific track. Precision and recall curves for the runs are shown in Figures 1 and 2 for the Monolingual and Multilingual and Bilingual tasks respectively. Although Berkeley's results for these tasks are high, compared to the overall average MAP for all participants, not much can be claimed for this since there were very few submissions for the Domain Specific track this year, and for many tasks (Multilingual, Monolingual Russian, Bilingual X⇒Russian, and Bilingual X⇒English) Berkeley was the *only* group that submitted runs. A few observations concerning translation are worth mentioning. First is that where we have comparable runs using different translations, all of the processing *except* for the topic translation was identical. Thus, it is obvious that for translations from German to English, Babelfish does a better job, and for English to Russian PROMT does a better job. We could not get PROMT (online version) to successfully translate the Russian topics to English apparently due to invalid character codes in the input topics (which were not detected by Babelfish).

It is perhaps more interesting to compare our results with last year's result, where Berkeley was also quite successful (see [4]). This year we see considerable improvement in all tasks when compared to the Berkeley1 submissions for 2005 (which used the same system as this year, but with different algorithms). However, none of this years runs for German Monolingual or Bilingual tasks approached the performance seen for last year's Berkeley2 runs. Because we did no decompounding for German, nor did we do query expansion using EVMs this year, it seems a reasonable assumption

Run Name	Description	translation	MAP
BERK_BI_ENDE_T2FB_B	Bilingual English⇒German	Babelfish	0.23658
BERK_BI_DEEN_T2FB_B	Bilingual German⇒English	Babelfish	0.33013*
BERK_BI_DEEN_T2FB_P	Bilingual German⇒English	PROMT	0.31763
BERK_BI_RUEN_T2FB_B	Bilingual Russian⇒English	Babelfish	0.32282*
BERK_BI_DERU_T2FB_P	Bilingual German⇒Russian	PROMT	0.10826*
BERK_BI_ENRU_T2FB_B	Bilingual English⇒Russian	Babelfish	0.11554
BERK_BI_ENRU_T2FB_P	Bilingual English⇒Russian	PROMT	0.16482**
BERK_MO_DE_T2FB	Monolingual German	none	0.39170
BERK_MO_EN_T2FB	Monolingual English	none	0.41357
BERK_MO_RU_T2FB	Monolingual Russian	none	0.25422**
BERK_MU_DE_T2FB_B_CMBZ	Multilingual from German	PROMT and Babelfish	0.04674
BERK_MU_EN_T2FB_B_CMBZ	Multilingual from English	Babelfish	0.07534**

Table 2: Submitted Domain Specific Runs

Task Description	2005	2005	2006	Pct. Diff	Pct. Diff
	Berk1	Berk2	MAP	from Berk1	from Berk2
Bilingual English⇒German	0.1477	0.4374	0.2366	+60.19	-45.91
Bilingual German⇒English	0.2398	0.4803	0.3301	+37.66	-31.27
Bilingual Russian⇒English	0.2358	–	0.3228	+36.90	
Bilingual German⇒Russian	0.1717	0.2331	0.1083	-36.92	-53.54
Bilingual English⇒Russian	01364	0.1810	0.1648	+20.82	-8.95
Monolingual German	0.2314	0.5144	0.3917	+69.27	-23.85
Monolingual English	0.3291	0.4818	0.4136	+25.68	-14.16
Monolingual Russian	0.2409	0.3038	0.2542	+5.52	-16.33
Multilingual from German	0.0294	–	0.0467	+58.84	–
Multilingual from English	0.0346	–	0.0753	+117.63	–

Table 3: Comparison of MAP with 2005 Berkeley1 and Berkeley2

that the lack of those is what led to this relative worse performance.

The results of these comparisons with our 2005 Domain Specific results are shown in Table 3. The only exception to the large percentage improvements seen by our best 2006 runs over the Berkeley1 2005 runs is found in Bilingual German⇒Russian. We suspect that there were translation problems for the German topics using PROMT, but we lack sufficient language skills in each language to understand what all these problems were. But we did observe that a large number of German terms were not translated, and that many spurious additional characters appeared in the translated texts. Another difference worth noting is that the Berkeley2 group used the L&H PowerTranslator software for it's English to X translations, while this year we used only Babelfish and PROMT.

5 Conclusions

Given the small number of submissions for the Domain Specific track this year, we wonder about the viability of the track for the future. Berkeley's runs this year did not build on the successes of the Berkeley2 group from 2005, but instead worked only to establish a new baseline set of results for future for retrieval processing without query expansion using EVMs or thesaurus information.

This obviously hurt our comparable overall performance in tasks with submissions from elsewhere. We did, however, see a marked improvement in performance for our specific system (Cheshire II was also used for Berkeley1 last year) with the improvements to our probabilistic retrieval algorithms developed after last year's submissions. We suspect that with the further addition of decompounding for German, and the use of EVMs and thesaurus expansion we can match or exceed the performance of the Berkeley2 runs last year.

References

[1] Aitao Chen and Fredric C. Gey. Multilingual information retrieval using machine translation, relevance feedback and decompounding. *Information Retrieval*, 7:149–182, 2004.

[2] W. S. Cooper, A. Chen, and F. C. Gey. Full Text Retrieval based on Probabilistic Equations with Coefficients fitted by Logistic Regression. In *Text REtrieval Conference (TREC-2)*, pages 57–66, 1994.

[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] Vivien Petras, Fredric Gey, and Ray Larson. Domain-specific CLIR of english, german and russian using fusion and subject metadata for query expansion. In *Cross-Language Evaluation Forum: CLEF 2005*, pages 226–237. Springer (Lecture Notes in Computer Science LNCS 4022), 2006.

[5] S. E. Robertson and K. Sparck Jones. Relevance weighting of search terms. *Journal of the American Society for Information Science*, pages 129–146, May–June 1976.