Back to Basics - Again - for Domain Specific Retrieval

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 2008. Last year we used *Entry Vocabulary Indexes* and Thesaurus expansion approaches for DS, but found in later testing that some simple text retrieval approaches had better results than these more complex query expansion approaches. This year we decided to revisit our basic text retrieval approaches and see how they would stack up against the various expansion approaches used by other groups. The results are now in and the answer is clear, they perform pretty badly compared to other groups' approaches.

All of the runs submitted were performed using the Cheshire II system. This year the Berkeley/Cheshire group submitted a total of twenty-four runs, including two for each subtask of the DS track. These include six Monolingual runs for English, German, and Russian, twelve Bilingual runs (four X2EN, four X2DE, and four X2RU), and six Multilingual runs (two EN, two DE, and two RU). The overall results include Cheshire runs in the top five participants for each task, but usually as the lowest of the five (and often fewer) groups.

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

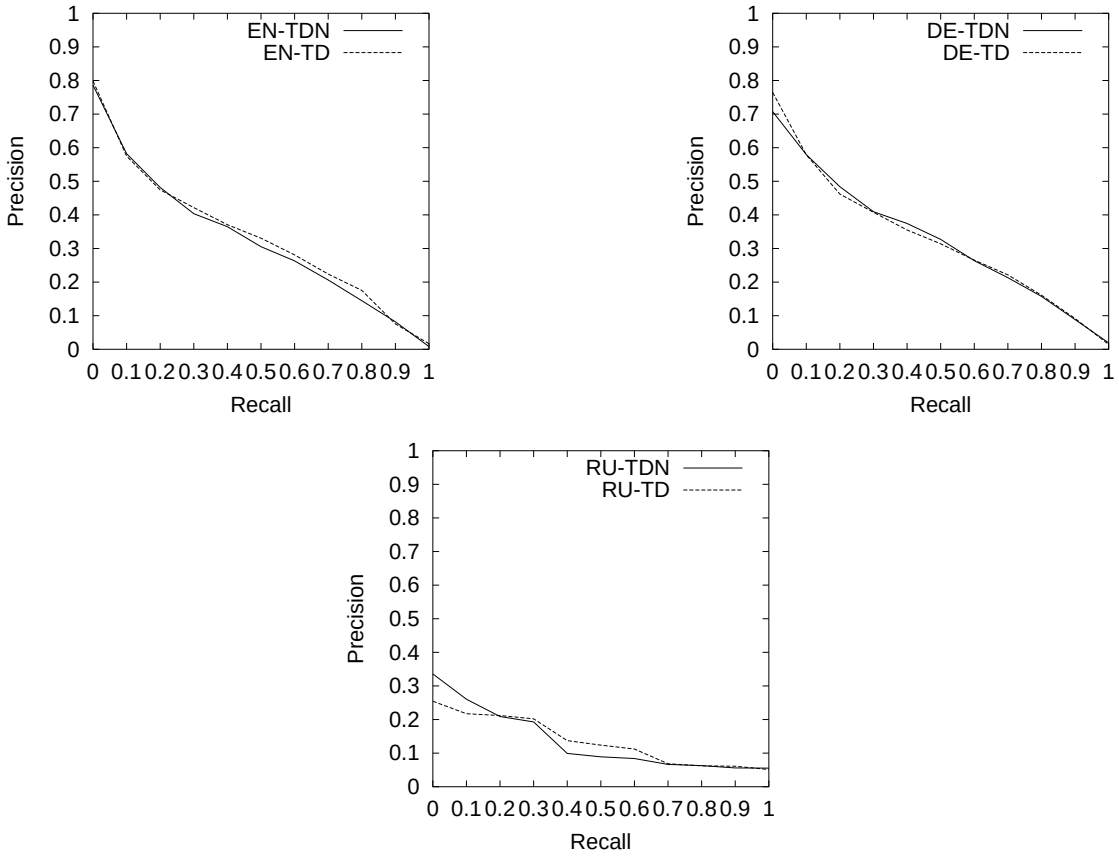
Keywords

Cheshire II, Logistic Regression, Entry Vocabulary Indexes

1 Introduction

This paper discusses the retrieval methods and evaluation results for Berkeley's participation in the CLEF 2008 Domain Specific track. In 2007 we focused on query expansion using Entry Vocabulary Indexes(EVIs)[4, 6], and thesaurus lookup of topic terms. Once the relevance judgements for 2007 were released we discovered that these rather complex method actually did not perform as well as basic text retrieval on the topics without additional query expansion. So, this year for the Domain Specific track we have returned to using a basic text retrieval approach using Probabilistic retrieval based on Logistic Regression with the inclusion of blind feedback, as used in 2006[5].

Figure 1: Berkeley Domain Specific Monolingual Runs for English (top left), German (top right), and Russian (lower)



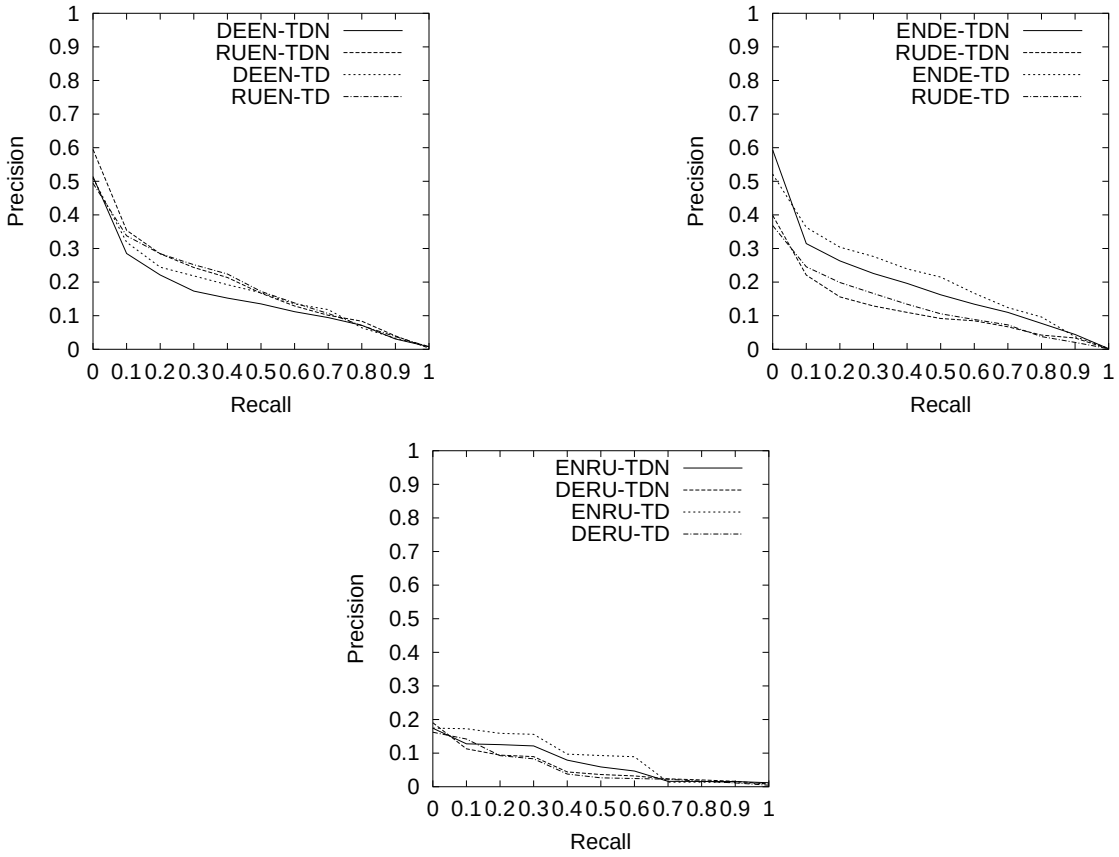
All of the submitted runs for this year's Domain Specific track used the Cheshire II system for indexing and retrieval.

This paper first very briefly describes the probabilistic retrieval methods used, including our blind feedback method for text, which are also discussed in our other notebook papers for this year. 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 Adhoc-TEL and GeoCLEF tracks, 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

Figure 2: Berkeley Domain Specific Bilingual Runs – To English (top left), to German (top right) and to Russian (lower)



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).

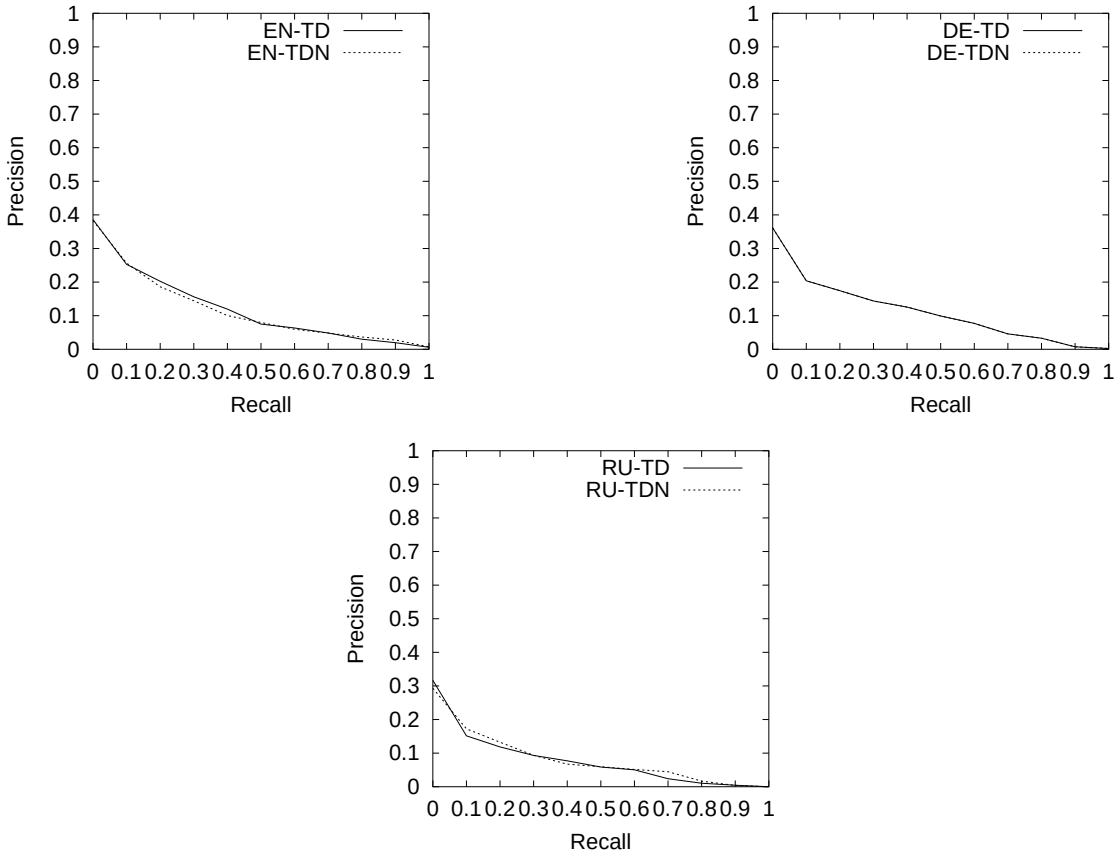
c_k are the k coefficients obtained though the regression analysis.

More details of this algorithm and the coefficients used with it may be found in our Adhoc-TEL 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 also described in detail in our Adhoc-TEL paper. Our blind feedback approach uses some number of top-ranked documents from an initial retrieval using the LR algorithm above, and selects some number of terms from the content of those documents, using a version of the Robertson and Sparck Jones probabilistic term relevance weights [7]. Those terms are merged with the original query and new term frequency weights are calculated, and the revised query submitted to obtain the final ranking. We used different numbers of documents and terms for different collections based on some tests run the 2007 data, varying these numbers to find the optimal point for the specific collection. For the German collection we selected 20 documents and the 35 topranked terms from those documents for feedback. For English we used 14 documents and 16 terms, and for Russian we used 16 documents and the topranked 10 terms.

3 Approaches for Domain Specific Retrieval

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

Figure 3: Berkeley Domain Specific Multilingual Runs – From English (top left), from German (top right), and from Russian (lower)



3.1 Database creation

We essentially used the same databases used in 2007. Although the Thesaurus and Classification Clusters created for last year were available, we did not use them this year.

3.2 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.

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.

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.

3.3 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

Table 1: Cheshire II Indexes for Domain Specific 2008

Name	Description	Content Tags	Used
docno	Document ID	DOCNO, DOCID	no
author	Author name	AUTHOR	no
title	Article Title	TITLE-DE, TITLE-EN, TITLE-RU, TITLE	no
topic	All Content Words	DOC	yes
date	Date	DATE, PUBLICATION-YEAR	no
subject	Controlled Vocabulary	CONTROLLED-TERM-EN CONTROLLED-TERM-DE, CLASSIFICATION-TEXT-EN, CLASSIFICATION-TEXT-DE, CLASSIFICATION, KEYWORDS, KEYWORDS-RU,	yes
geoname	Geographic names	GEOGR-AREA, COUNTRY-CODE	no

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 the LEC Power Translator PC-based program. Overall we have found that this translation program seems to generate good translations between any of the languages needed for this track, but we still intend to do some further testing to compare to previous approaches (which used web-based translation tools like Babelfish and PROMT). We suspect that, as always, different tools provide a more accurate representation of different topics for some languages, but the LEC Power Translator seemed to do pretty good (and often better) translations for all of the needed languages.

All searches were submitted using the TREC2 Algorithm with blind feedback described above. This year we did no expansion of topics or use of the thesaurus or the classification clusters created last year. The differences in the runs for a given language or language pair (for bilingual) in Table 2 are primarily whether the topic title and description only (TD) or title, description and narrative (TDN).

4 Results for Submitted Runs

The summary results (as Mean Average Precision) for all of our submitted runs for English, German and Russian are shown in Table 2, the Recall-Precision curves for these runs are also shown in Figure 1 (for monolingual), Figure 2 (for bilingual) and Figure 3 (for multilingual). 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 query expansion approach used. For example, in Figure 2 DEEN-TD corresponds to run BRK-BI-DEEN-TD in Table 2.

We observe that for the vast majority of our runs, using the narrative tends to degrade instead of improve performance. (We observed the same in other tracks as well.)

It is worth noting that the approaches used in our submitted runs provided the best results when testing with 2007 data and topics when compared to our official 2007 runs. In fact we may have over-simplified for this track. Although at least one Cheshire run appeared in the top five runs of the overall summary results available on the DIRECT system, none of them were top-ranked and for many tasks there appeared to be fewer than five participants.

Table 2: Submitted Domain Specific Runs

Run Name	Description	Exp.	MAP
BRK-MO-DE-TD	Monolingual German	TD auto	0.3155
BRK-MO-DE-TDN	Monolingual German	TDN auto	0.3111
BRK-MO-EN-TD	Monolingual English	TD auto	0.3200
BRK-MO-EN-TDN	Monolingual English	TDN auto	0.3095
BRK-MO-RU-TD	Monolingual Russian	TD auto	0.1306
BRK-MO-RU-TDN	Monolingual Russian	TDN auto	0.1260
BRK-BI-ENDE-TD	Bilingual English⇒German	TD auto	0.1982
BRK-BI-ENDE-TDN	Bilingual English⇒German	TDN auto	0.1726
BRK-BI-RUDE-TD	Bilingual Russian⇒German	TD auto	0.1188
BRK-BI-RUDE-TDN	Bilingual Russian⇒German	TDN auto	0.1087
BRK-BI-DEEN-TD	Bilingual German⇒English	TD auto	0.1668
BRK-BI-DEEN-TDN	Bilingual German⇒English	TDN auto	0.1454
BRK-BI-RUEN-TD	Bilingual Russian⇒English	TD auto	0.1765
BRK-BI-RUEN-TDN	Bilingual Russian⇒ English	TDN auto	0.1748
BRK-BI-DERU-TD	Bilingual German⇒Russian	TD auto	0.0515
BRK-BI-DERU-TDN	Bilingual German⇒Russian	TDN auto	0.0550
BRK-BI-ENRU-TD	Bilingual English⇒Russian	TD auto	0.0857
BRK-BI-ENRU-TDN	Bilingual English⇒Russian	TDN auto	0.0662
BRK-MU-DE-TD	Multilingual German	TD auto	0.0984
BRK-MU-DE-TDN	Multilingual German	TDN auto	0.0984
BRK-MU-EN-TD	Multilingual English	TD auto	0.1057
BRK-MU-EN-TDN	Multilingual English	TDN auto	0.1034
BRK-MU-RU-TD	Multilingual Russian	TD auto	0.0662
BRK-MU-RU-TDN	Multilingual Russian	TDN auto	0.0701

5 Conclusions

Since we have not yet had a chance to test alternative approaches on the 2008 topics and relevance judgement, we don't yet have much to report on ways forward. Given that the re-introduction of fusion approaches in our GeoCLEF entry led to very good results, we suspect that the application of selected fusion approaches for this task may also prove valuable.

We are much more curious to see what approaches the other groups in this task used this year, since some very strong results (at least compared to our own) appeared in the overall summary data.

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] Fredric Gey, Michael Buckland, Aitao Chen, and Ray Larson. Entry vocabulary – a technology to enhance digital search. In *Proceedings of HLT2001, First International Conference on Human Language Technology, San Diego*, pages 91–95, March 2001.

[5] Ray R. Larson. Domain specific retrieval: Back to basics. In *Evaluation of Multilingual and Multi-modal Information Retrieval – Seventh Workshop of the Cross-Language Evaluation Forum, CLEF 2006*, LNCS, page to appear, Alicante, Spain, September 2006.

[6] 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.

[7] 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.