Experiments in Classification Clustering and Thesaurus Expansion for Domain Specific Cross-Language 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 2007. This year we are using forms of the *Entry Vocabulary Indexes* and Thesaurus expansion approaches used by Berkeley in 2005[10]. Despite the basic similarity of approach, we are using quite different implementations with different characteristics. We are not, however, using the tools for de-compounding for German that were developed over the past many years and used very successfully in earlier Berkeley entries in this track. All of the runs submitted were performed using the Cheshire II system. This year Berkeley submit a total of submitted 24 runs, including one for each subtask of the DS track. These include 6 Monolingual runs for English, German, and Russian, 12 Bilingual runs (4 X2EN, 4 X2DE, and 4 X2RU), and 6 Multilingual runs (2 EN, 2 DE, and 2 RU). Since the overall results were not available at the time this paper was due, we do not know how these results fared compared to other participants, so the discussion in this paper focuses on comparisions between our own runs.

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

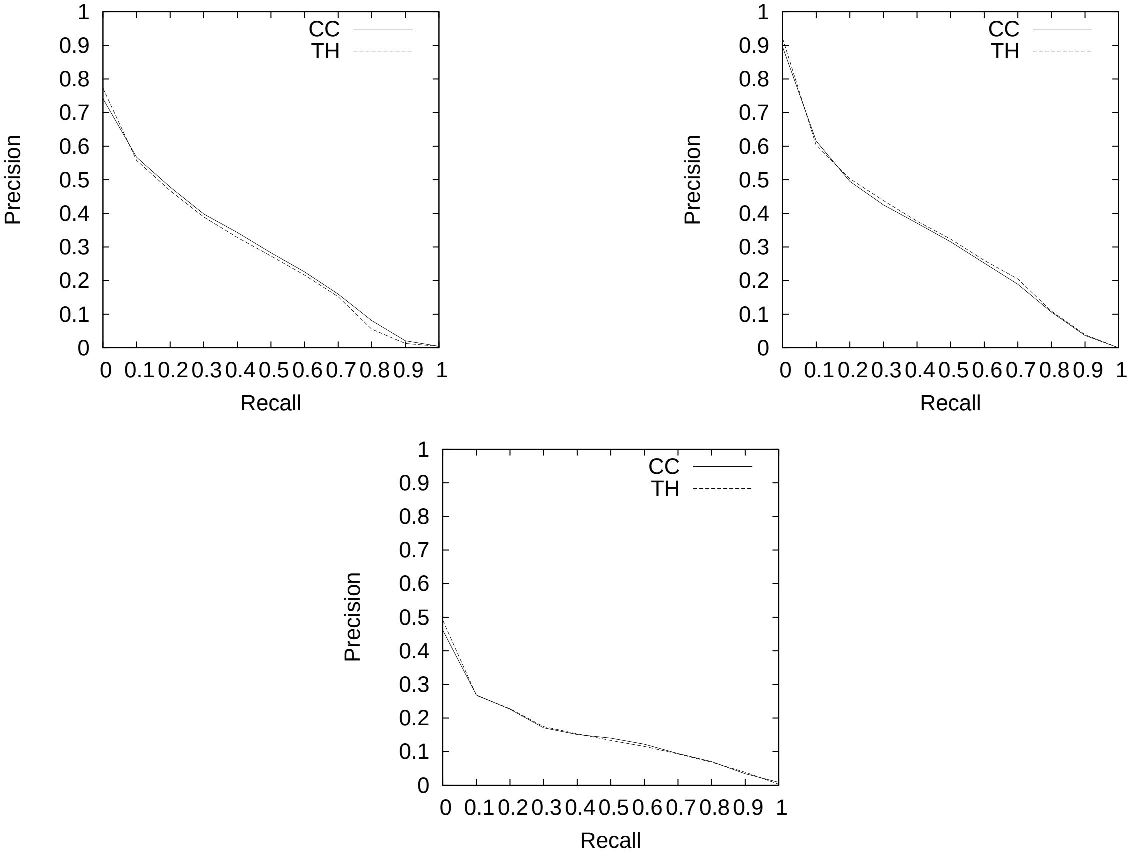
Cheshire II, Logistic Regression, Entry Vocabulary Indexes

1 Introduction

This paper discusses the retrieval methods and evaluation results for Berkeley's participation in the CLEF 2007 Domain Specific track. Last year for this track we used a baseline approach using text retrieval methods only[7] without query expansion or use of the Thesaurus. This year we have focused instead on query expansion using Entry Vocabulary Indexes(EVIs)[4, 10], and thesaurus lookup of topic terms. We continue to use probabilistic IR methods based on logistic regression.

All of the submitted runs for this year's Domain Specific track used the Cheshire II system for indexing and retrieval. The "Classification Clustering" feature of the system was used to

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



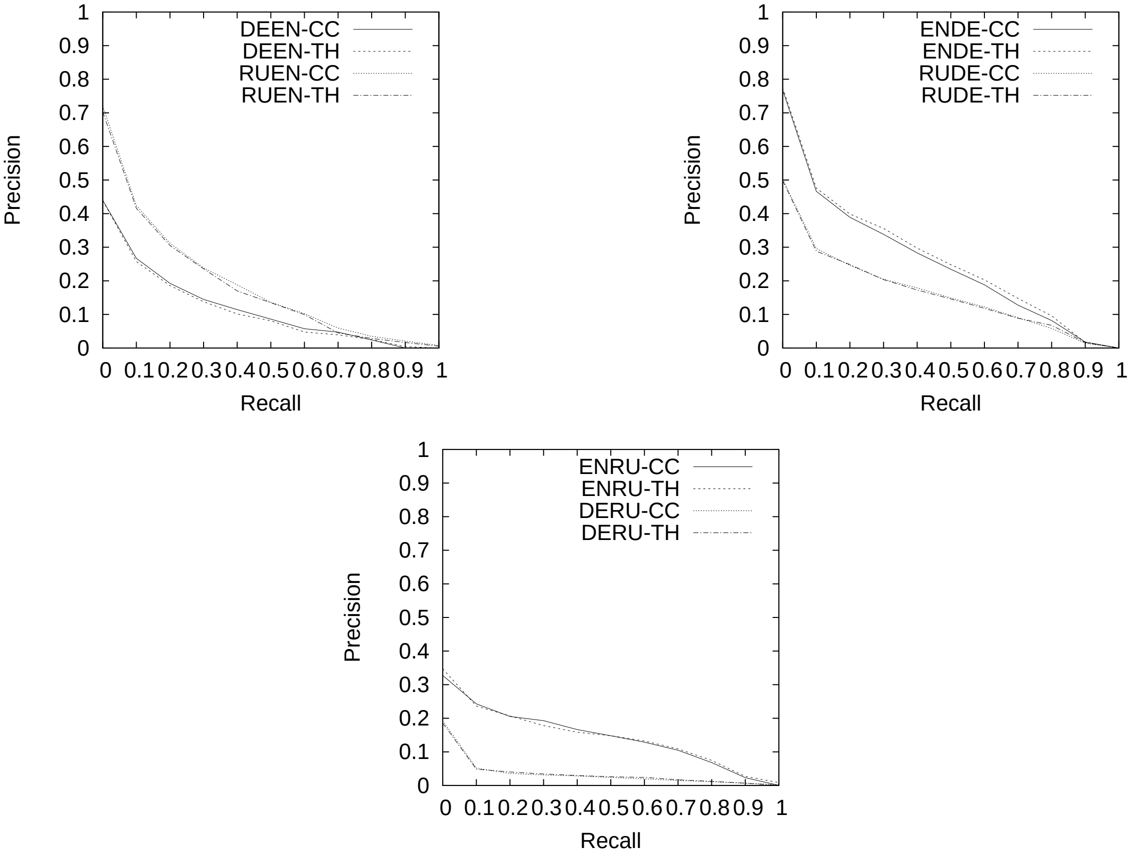
generate the EVIs used in query expansion. The original approach for Classification Clustering was in searching was described in [5] and [6]. Although the method has experienced considerable changes in implementation, the basic approach is still the same: topic-rich elements extracted from individual records in the database (such as titles, classification codes, or subject headings) are merged based on a normalized version of a particular organizing element (usually the classification or subject headings), and each such *classification cluster* is treated as a single "document" containing the combined topic-rich elements of all the individual documents that have the same values of the organizing element. The EVI creation and search approach taken for this research is described below in Section 3.3.

This paper first very briefly describes the probabilistic retrieval methods used, including our blind feedback method for text, which are discussed in greater detail in our ImageCLEF notebook paper[8]. 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

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



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).

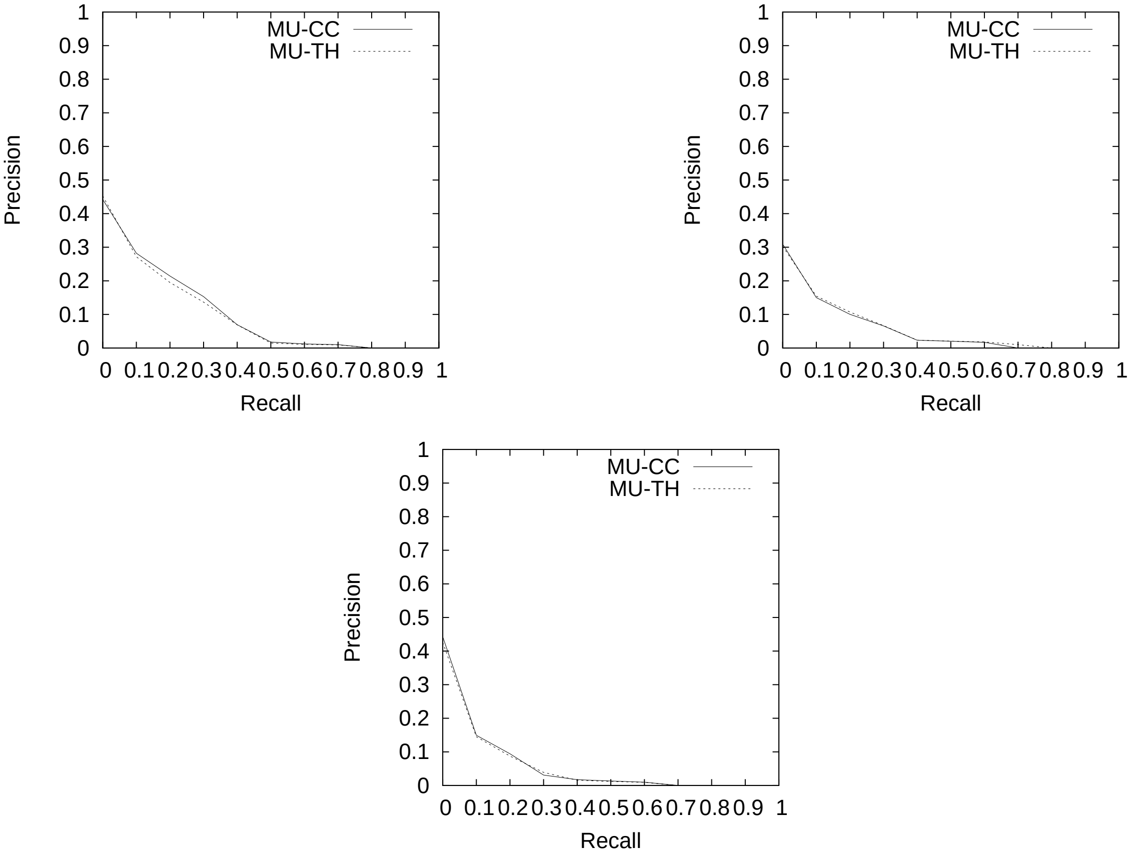
c_k are the k coefficients obtained though the regression analysis.

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[8]. 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 [11]. 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 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

For the purposes of this research we combined the GIRT German/English thesaurus along with the English and Russian mappings for the CSASA and ISISS databases to produce a multilingual thesaurus where elements from each of the original sources, as well as transliterations and capitalizations and the conversion of all data to UTF-8 encoding (this was also performed on the databases themselves before indexing). An example entry from this thesaurus is shown below:

```
<entry>
<german>Absatz</german>
<german-caps>ABSATZ</german-caps>
<scope-note-de>nicht im Sinne von Vertrieb</scope-note-de>
<english-translation>sale</english-translation>
<german_utf8>Absatz</german_utf8>
<russian>
```

сбыт

```
</russian>
<translit>sbyt </translit>
<mapping>
    <original-term>Absatz</original-term>
    <mapped-term>Sales</mapped-term>
</mapping>
<mapping>
    <original-term>sale</original-term>
    <mapped-term>Sales</mapped-term>
</mapping>
</entry>
```

Note that the spacing around the Russian cyrillic term was inserted in the paper formatting process and was not in the original data.

Because not all of the terms had mappings, or equivalent Russian terms those parts are not present for all of the thesaurus entries.

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: Cheshire II Indexes for Domain Specific 2007

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

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 used the Entry Vocabulary Indexes (search term recommenders) that were used in somewhat different form by Berkeley in previous years (see [10]), without overall data on the track performance this year it is difficult to say whether this approach improved upon, or degraded, the text-retrieval baseline we established last year. Given the changes in the collections used (the addition of the CSASA English collection and elimination of the Russian SocioNet data), it is not possible to directly compare MAP or other evaluation measures across years. The implementation of the Classification Cluster -based EVIs will be discussed in the next section.

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.

3.3 Entry Vocabulary Indexes

As noted above earliest versions of Entry Vocabulary Indexes were developed to facilitate automatic classification of library catalog records, and first used in searching in [6]. Those used a simple frequency-based probabilistic model in searching, but a primary feature was that the "Classification clusters", were treated as documents and the terms associated with top-ranked clusters were combined with the original query, in a method similar to "blind feedback", to provide an enhanced second stage of search.

Our later work with EVIs used a maximum likelihood weighting for each term (word or phrase) in each classification. This was the approach described in [4] and used for Cross-language Domain-Specific retrieval for CLEF 2005. One limitation of that approach is that the EVI can produce maximum likelihood estimates for only a single term at a time, and alternative approaches needed to be explored for combining terms (see [10] for the various approaches).

Although the method has experienced considerable changes in implementation, the basic approach for "Classification Clustering" in Cheshire II is still the same. Various topic-rich elements are extracted from individual records in the database (such as titles, classification codes, or subject headings) and are merged into single records based on a normalized version of a particular organizing element (usually the classification or subject headings, e.g., one record is created for each unique classification or subject heading). Each of these *classification clusters* is treated as a single "document" containing the combined topic-rich elements of all the individual documents that have the same values of the organizing element. In place of the simpler probabilistic model used in the early research, we use the same logistic regression based algorithm that is used for text retrieval. In effect, we just search the "Classification Clusters" as if they were documents using the TREC2 algorithm with blind feedback described above, then take some number of the top-ranked terms and use those to expand the query for submission to the normal document collection. Testing with the 2006 data showed that just taking the single top-ranked term performed better than using multiple terms for this approach, so only the single top-ranked recommended term was used in the experiments reported here.

Two separate EVIs were built for the databases in each target language. The first used the contents of the "CONTROLLED-TERM-??" (or "KEYWORD" for Russian) fields as the organizing element. The second EVI used the contents of the "CLASSIFICATION-??" fields. Both of these EVIs were used in query expansion. One problem was that some records included multiple controlled terms in a single field instead of as separate fields. This was particularly common for the Russian "KEYWORD" terms. For this year we just ignored this problem rather than attempting to fix it, but we will be examining the effects in our analysis of the results.

3.4 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 the LEC Power Translator PC-based program. Our original testing of LEC Power Translator seemed to show a good translations between any of the languages needed for the track, but we 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.

Because all of our submitted runs this year used some form of query expansion, each required a 2-phase search process. The first phase involved a search in the EVI or the merged thesaurus, and the second phase combined some of the results of first phase search with the original query and used the expanded query to search the collections in the target language.

Table 2: Submitted Domain Specific Runs

Run Name	Description	Exp.	MAP
Berk_M_DE_CC_p15	Monolingual German	EVI	0.3150
Berk_M_DE_TH_p7	Monolingual German	Thes	0.3199
Berk_M_EN_CC_p15	Monolingual English	EVI	0.2814
Berk_M_EN_TH_p7	Monolingual English	Thes	0.2733
Berk_M_RU_CC_p15	Monolingual Russian	EVI	0.1390
Berk_M_RU_TH_p7	Monolingual Russian	Thes	0.1401
Berk_B_DEEN_CC_p15	German⇒English	EVI	0.1096
Berk_B_DEEN_TH_p7	German⇒English	Thes	0.1043
Berk_B_DERU_CC_p15	German⇒Russian	EVI	0.0269
Berk_B_DERU_TH_p7	German⇒Russian	Thes	0.0285
Berk_B_ENDE_CC_p15	English⇒German	EVI	0.2412
Berk_B_ENDE_TH_p7	English⇒German	Thes	0.2514
Berk_B_ENRU_CC_p15	English⇒Russian	EVI	0.1348
Berk_B_ENRU_TH_p7	English⇒Russian	Thes	0.1341
Berk_B_RUDE_CC_p15	Russian⇒German	EVI	0.1520
Berk_B_RUDE_TH_p7	Russian⇒German	Thes	0.1501
Berk_B_RUEN_CC_p15	Russian⇒English	EVI	0.1757
Berk_B_RUEN_TH_p7	Russian⇒English	Thes	0.1701
BerkMUDEp15	Multiling. from German	EVI	0.0468
BerkMUDETHp7	Multiling. from German	Thes	0.0486
BerkMUENp15	Multiling. from English	EVI	0.0884
BerkMUENTHp7	Multiling. from English	Thes	0.0839
BerkMURUp15	Multiling. from Russian	EVI	0.0414
BerkMURUTHp7	Multiling. from Russian	Thes	0.0400

3.4.1 EVI Searches

For the monolingual and bilingual EVI searches (all those indicated in Table 2 with "EVI" in the "Exp." or expansion column) the first search phase used all terms included in the "title" and "desc" fields of the topics (or the tranlated version of these fields). These terms were searched using the TREC2 algorithm with blind feedback to obtain a ranked result of classification clusters from the EVIs. The main or "organizing term" phrases for the top-ranked two clusters from the results for the "CONTROLLED-TERM" EVI, and the single top-ranked result phrase for the "CLASSIFICATION" EVI were extracted for use in the second phase.

For example, Topic #190 was searched using "mortality rate : find information on mortality rates in individual european countries" and the two EVIs yielded the following terms: "child mortality : infant mortality : demography and human biology; demography (population studies)".

For the second phase search the original query was searched using the initial title+desc from the topic using the "topic" index and the expansion terms were searched in the "subject" index, these searches were merged using a weighted sum for items in both lists that is based on the "Pivot" method described by Mass and Mandelbrod[9] to combine the results of different document components. In our case the probability of relevance for a component is a weighted combination of the initial estimate probability of relevance for the subject search and the probability of relevance for the entire document. Formally this is:

$$P(R \mid Q, C_{new}) = (X * P(R \mid Q, C_{subj})) + ((1 - X) * P(R \mid Q, C_{doc})) \tag{3}$$

Where X is a "pivot value" between 0 and 1, and $P(R \mid Q, C_{new})$, $P(R \mid Q, C_{subj})$ and $P(R \mid Q, C_{doc})$ are the new weight, the original subject search weight, and document weight for a given query. We found that a pivot value of 0.15 was most effective for CLEF2006 data when combining EVI and search queries.

3.4.2 Thesaurus-based searches

The basic steps for the searched doing thesaurus lookup is the same for EVIs, but the search structure is different. For the first phase search the topic title is searched among the language-appropriate main terms of the thesaurus, and the description is searched among all terms in the thesaurus entry. These intermediate results are combined using the pivot merger method described about with a pivot weight of 0.55. The top two results are used, and both the language-appropriate main term, and the appropriate mapping terms are used for the query expansion. In the second phase the full topic title and desc fields are searched as topics, and the thesaurus terms are also searched as topics. These searches are combined using the pivot merge with a pivot weight of 0.07.

For topic #190 the first part of the query (i.e., the topic title and desc terms) is the same as for the EVI searches, but the second part of the search uses the terms yielded by the thesaurus search: "mortality : Infant mortality" (only a single thesaurus entry was retrieved in the search).

For multilingual searches, we combined the various translations of the topic title and desc fields produced by the LEC Power Translator for each source language and searched those combined translations in each target language. The results for each language were merged based on the MINMAX normalized score for each resultset. Within each language the same approaches were used as for EVI and Thesaurus-based expansion of bilingual and monolingual searches.

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-CC corresponds to run Berk_B_DEEN_CC_p15 in Table 2.

Since summary information on the scores for all submissions were not available at the time this paper was written, we have no idea of how our result stack up against other approaches for the same data. We can, however, compare the results for EVIs versus Thesaurus lookup.

Since our experiments were conducted using the same topics, database, translation tools, and basic combination approaches for both EVIs and Thesaurus-based expansion, we were hoping to find a clear benefit for one approach versus the other. Unfortunately, the results are not at all clear. While EVIs seem to best results when English is the target language, the opposite is true for German and Russian targets. As always our multilingual results are significantly lower than monolingual or bilingual results for a given source language, with the exception of German⇒Russian, which is the lowest MAP of any of the runs.

As the precision/recall graphs show, (Figures 1, 2, and 3) there is very little difference in the curves for the EVI and Thesaurus-based expansion in a given source/target language set. Although we did not run significance testing of the results, we suspect that there is no statistically significant different between these runs.

It is worth noting that the approaches used in our submitted runs provided the best results when testing with 2006 data and topics. However, as we discovered after the 2007 qrels were made available, some simpler approaches worked as well or better than the more complex methods described above. For example a simplified version of English monolingual search using only the topic title and desc fields, and searching each of those in the topic and subject indexes, and merging the results using a pivot value of 0.15 obtained a MAP result of 0.2848, compared to the 0.2814 obtained in our best submitted monolingual run. Further simplification to individual index searches does not, however provide results approaching those of the pivot-merged results.

We suspect that the range of MAP scores for the track is different from previous years, or else our results are much worse than we thought they would be with the 2007 databases and topics.

5 Conclusions

We cannot say, overall, how effective query expansion by EVI or Thesaurus are relative to other approaches for this task. We can assume that there is very little difference in the effectiveness of the two methods, and that both seem to perform better than simple single-index "bag of words" searches of the collection contents.

We plan to conduct further runs to test whether modifications and simplifications, as well as combinations, of the EVI and Thesaurus-based approaches will provide can provide improved performance for the Domain Specific tasks.

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. Classification clustering, probabilistic information retrieval, and the online catalog. *Library Quarterly*, 61(2):133–173, 1991.

[6] Ray R. Larson. Evaluation of advanced retrieval techniques in an experimental online catalog. *Journal of the American Society for Information Science*, 43(1):34–53, 1992.

[7] 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.

[8] Ray R. Larson. Linked relevance feedback for the imageclef photo task. In *CLEF 2007 - Notebook Papers*, page to appear, Budapest, Hungary, September 2007.

[9] Yosi Mass and Matan Mandelbrod. Component ranking and automatic query refinement for xml retrieval. In *Advances in XML Information Retrieval: Third International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX2004*, pages 73–84. Springer (LNCS #3493), 2005.

[10] 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.

[11] 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.