UniNE at Domain-Specific IR - CLEF 2008: Scientific Data Retrieval: Various Query Expansion Approaches

Claire Fautsch, Ljiljana Dolamic, Jacques Savoy

Computer Science Department
University of Neuchatel, Switzerland
{Claire.Fautsch, Ljiljana.Dolamic, Jacques.Savoy}@unine.ch

Abstract

Our first objective in participating in this domain-specific evaluation campaign is to propose and evaluate various indexing and search strategies for the German, English and Russian languages, in an effort to obtain better retrieval effectiveness than that of the language-independent approach (*n*-gram). To do so we evaluate the GIRT-4 test-collection using the Okapi, various IR models derived from the *Divergence from Randomness* (DFR) paradigm, the statistical language model (LM) together with the classical *tf·idf* vector-processing scheme.

Categories and Subject Descriptors

H.3.1 [**Content Analysis and Indexing**]: Indexing methods, Linguistic processing. I.2.7 [**Natural Language Processing**]: Language models. H.3.3 [**Information Storage and Retrieval**]: Retrieval models. H.3.4 [**Systems and Software**]: Performance evaluation.

General Terms

Experimentation, Performance, Measurement, Algorithms.

Additional Keywords and Phrases

Natural Language Processing with European Languages, Digital Libraries, German Language, Russian Language; Manual Indexing, Thesaurus.

1 Introduction

Domain-specific retrieval is an interesting task, one in which we access bibliographic notices (usually composed of a title and an abstract) extracted from two German social science sources and one Russian corpus. The records in these notices also contain manually assigned keywords extracted from a controlled vocabulary by librarians who are knowledgeable of the discipline to which the indexed articles belong. These descriptors should be helpful in improving document surrogates and consequently the extraction of more pertinent information, while also discarding irrelevant abstracts. Access to the underlying thesaurus would also improve retrieval performance.

The rest of this paper is organized as follows: Section 2 describes the main characteristics of the GIRT-4 (written in the German and English languages) and ISISS (Russian) test-collections. Section 3 outlines the main aspects of our stopword lists and light stemming procedures, along with the IR models used in our experiments. Section 4 explains different blind query expansion approaches and evaluates their use with the available corpora. Section 5 provides our official runs and results.

2 Overview of Test-Collections

In the domain-specific retrieval task, the two available corpora are composed of bibliographic records extracted from various sources in the social sciences domain. Typical records (see Figure 1 for a German example) in this corpus consist of a title (tag <TITLE-DE>), author name (tag <AUTHOR>), document language (tag <LANGUAGE-CODE>), publication date (tag <PUBLICATION-YEAR>) and abstract (tag <ABSTRACT-DE>). Manually assigned descriptors and classifiers are provided for all documents. An inspection of this German corpus reveals that all bibliographic notices consist of a title and 96.4% of them include an abstract. In addition to this information provided by the author, a typical record contains on average 10.15 descriptors

("<CONTROLLED-TERM-DE>"), 2.02 classification terms ("<CLASSIFICATION-TEXT-DE>"), and 2.42 methodological terms ("<METHOD-TEXT-DE>" or "<METHOD-TERM-DE>"). The manually assigned descriptors are extracted from the controlled list known as the "Thesaurus for the Social Sciences". Finally, associated with each record is a unique identifier ("<DOCNO>"). Kluck (2004) provides a more complete description of this corpus.

```
<DOC>
<DOCNO>  GIRT-DE19909343
<TITLE-DE>  Die sozioökonomische Transformation einer Region : Das Bergische Land von 1930 bis 1960
<AUTHOR>  Henne, Franz J.
<AUTHOR>  Geyer, Michael
<PUBLICATION-YEAR>  1990
<LANGUAGE-CODE>  DE
<CONTROLLED-TERM-DE>  Rheinland
<CONTROLLED-TERM-DE>  historische Entwicklung
<CONTROLLED-TERM-DE>  regionale Entwicklung
<CONTROLLED-TERM-DE>  sozioökonomische Faktoren
<METHOD-TERM-DE>  historisch
<METHOD-TERM-DE>  Aktenanalyse
<CLASSIFICATION-TEXT-DE>  Sozialgeschichte
<ABSTRACT-DE>  Die Arbeit hat das Ziel, anhand einer regionalen Studie die Entstehung des "modernen"
fordistischen Wirtschaftssystems und des sozialen Systems im Zeitraum zwischen 1930 und 1960 zu
beleuchten; dabei geht es auch um das Studium des "Sozial-imaginären", der Veränderung von Bewußtsein und
Selbst-Verständnis von Arbeitern durch das Erlebnis und die Erfahrung der Depression, des
Nationalsozialismus und der Nachkriegszeit, welches sich in den 1950er Jahren gemeinsam mit der
wirtschaftlichen Veränderung zu einem neuen "System" zusammenfügt.
<DOC>  …
```

Figure 1: Example of record written in German

```
<DOC>
<DOCNO>  GIRT-EN19901932
<TITLE-EN>  The Socio-Economic Transformation of a Region : the Bergische Land from 1930 to 1960
<AUTHOR>  Henne, Franz J.
<AUTHOR>  Geyer, Michael
<PUBLICATION-YEAR>  1990
<LANGUAGE-CODE>  EN
<CONTROLLED-TERM-EN>  Rhenish Prussia
<CONTROLLED-TERM-EN>  historical development
<CONTROLLED-TERM-EN>  regional development
<CONTROLLED-TERM-EN>  socioeconomic factors
<METHOD-TERM-EN>  historical
<METHOD-TERM-EN>  document analysis
<CLASSIFICATION-TEXT-EN>  Social History
<DOC>  …
```

Figure 2: English translation of the record shown in Figure 1

```
<DOC>
<DOCNO>  ISISS-RAS-ECOSOC-20060324-41210
<AUTHOR-RU>  Мартынова, М.Ю.
<TITLE-RU>  Нормы и правила межличностного общения в культуре народов России
<KEYWORDS-RU>  Россия; межличностные отношения; межкультурные отношения; коммуникация
<DOC>  …
```

Figure 3: Example of a record extracted from the ISISS corpus

The above-mentioned German collection was translated into British English, mainly by professional translators whose native language was English. Included in all English records is a translated title (listed under "<TITLE-EN>" in Figure 2), manually assigned descriptors ("<CONTROLLED-TERM-EN>"), classification terms ("<CLASSIFICATION-TEXT-EN>") and methodological terms ("<METHOD-TERM-EN>"). Abstracts however were not always translated (in fact they are available for only around 15% of the English records).

In addition to this bilingual corpus, we may also access the GIRT thesaurus, containing 10,623 entries (all including both the <GERMAN> and <GERMAN-CAPS>) tags together with 9,705 English translations. It also contains 2,947 <BROADER-TERM> relationships and 2,853 <NARROWER-TERM> links. The synonym relationship between terms is expressed through <USE-INSTEAD> (2,153) links, <RELATED-TERM> (1,528) or <USE-COMBINATION> (3,263).

As a third language, we access bibliographic records written in the Russian language composed of the ISISS (Russian Economic and Social Science) bibliographic data collection (see Figure 3 for an example of a record extracted from the Russian collection). Using a pattern similar to that of the other two corpora, records include a title ("<TITLE-RU>" in Figure 3), sometimes an abstract ("<ABSTRACT-RU>"), and certain manually assigned descriptors ("<KEYWORDS-RU>").

Table 1 below lists a few statistics from these collections, showing that the German corpus has the largest size (326 MB), the English ranks second and the Russian third, both in size (81 MB) and in number of documents (145,802). The German corpus has the larger mean size (89.71 indexing terms/article), compared to the English collection (54.86), while for the Russian corpus the mean value is clearly smaller (18.77). The English corpus includes also the *CSA Sociological Abstracts* (20,000 documents, 38.5 MB).

During the indexing process, we retained all pertinent sections in order to build document representatives. Additional information such as author name, publication date and the language in which the bibliographic notice was written are of less importance, particularly from an IR perspective, and thus they will be ignored in our experiments.

As shown in Appendix 2, the available topics cover various subjects (e.g., Topic #206: "Environmental justice," Topic #209: "Doping and sports," Topic #221: "Violence in schools," or Topic #211: "Shrinking cities"), and some of them may cover a relative large domain (e.g. Topic #212: "Labor market and migration").

	German	English	Russian
Size (in MB)	326 MB	235 MB	81 MB
# of documents	151,319	171,319	145,802
# of distinct terms	10,797,490	6,394,708	40,603
Number of distinct indexing terms per document			
Mean	71.36	37.32	14.89
Standard deviation	32.72	25.35	7.54
Median	68	28	13
Maximum	391	311	74
Minimum	2	2	1
Number of indexing terms per document			
Mean	89.71	54.86	18.77
Standard deviation	44.5	42.41	9.32
Median	85	39	17
Maximum	629	534	98
Minimum	4	4	2
Number of queries	25	25	24
Number rel. items	2290	2133	292
Mean rel./ request	91.6	85.32	12.17
Standard deviation	90.85	59.95	17.45
Median	72	89	5
Maximum	431 (T #218)	206 (T #201)	73 (T #204)
Minimum	7 (T #204)	4 (T #218)	1 (T #215)

Table 1: CLEF GIRT-4 and ISISS test collection statistics

3 IR Models and Evaluation

3.1 Indexing and IR Models

For the English, German and Russian language, we used the same stopword lists and stemmers that we selected for our previous CLEF participation (Fautsch *et al*., 2008). Thus for English it was the SMART stemmer and stopword list (containing 571 items), while for the German we apply our light stemmer (available at http://www.unine.ch/info/clef/) and stopword list (603 words). For all our German experiments we also apply our decompounding algorithm (Savoy, 2004). For the Russian language, the stopword list contains 430 words and we apply our light stemming procedure (based on 53 rules to remove the final suffix representing gender (masculine, feminine, and neutral), number (singular, plural) and the six Russian grammatical cases (nominative, accusative, genitive, dative, instrumental, and locative)).

In order to obtain a broader view of the relative merit of various retrieval models, we may first adopt the classical *tf idf* indexing scheme. In this case, the weight attached to each indexing term in a document surrogate (or in a query) combines the term's occurrence frequency (denoted tf_{ij} for indexing term t_j in document D_i) and also the inverse document frequency (denoted idf_j).

In addition to this vector-processing model, we may also consider probabilistic models such as the Okapi model (or BM25) (Robertson *et al*., 2000). As a second probabilistic approach, we may implement four variants of the DFR (*Divergence from Randomness*) family suggested by Amati & van Rijsbergen (2002). In this framework, the indexing weight w_{ij} attached to term t_j in document D_i combines two information measures as follows.

$$w_{ij} = Inf^1_{ij} \cdot Inf^2_{ij} = -\log_2[Prob^1_{ij}(tf)] \cdot (1 - Prob^2_{ij}(tf))$$

The first model PB2 is based on the following equations:

$$Prob^1_{ij} = (e^{-\lambda j} \cdot \lambda^{tfij}) / tf_{ij}! \qquad \text{with } \lambda_j = tc_j / n \tag{1}$$

$$Prob^2_{ij} = 1- [(tc_j+1) / (df_j \cdot (tfn_{ij}+1))] \qquad \text{with } tfn_{ij} = tf_{ij} \cdot \log_2[1 + ((c \cdot mean\ dl) / l_i) \tag{2}$$

where tc_j represents the number of occurrences of term t_j in the collection, df_j the number of documents in which the term t_j appears, and n the number of documents in the corpus. Moreover, c and *mean dl* (average document length) are constants whose values are given in the Appendix 1.

The second model GL2 is defined as:

$$Prob^1_{ij} = [1 / (1+\lambda_j)] \cdot [\lambda_j / (1+\lambda_j)]^{tfn_{ij}} \tag{3}$$

$$Prob^2_{ij} = tfn_{ij} / (tfn_{ij} + 1) \tag{4}$$

For the third model I(n)B2, we still use Equation 2 to compute $Prob^2_{ij}$ but the implementation of Inf^1_{ij} is modified as:

$$Inf^1_{ij} = tfn_{ij} \cdot \log_2[(n+1) / (df_j+0.5)] \tag{5}$$

For the fourth model $I(n_e)C2$ the initial value of $Prob^2_{ij}$ is obtained from Equation 2 and for the value Inf^1_{ij} we use:

$$Inf^1_{ij} = tfn_{ij} \cdot \log_2[(n+1) / (n_e+0.5)] \qquad \text{with } n_e = n \cdot [1 - [(n-1) / n]^{tcj}] \tag{6}$$

Finally, we also consider an approach based on a statistical language model (LM) (Hiemstra 2000; 2002), known as a non-parametric probabilistic model (both Okapi and DFR are viewed as parametric models). Thus, the probability estimates would not be based on any known distribution (as in Equations 1, or 3), but rather be estimated directly based on the occurrence frequencies in document D or corpus C. Within this language model (LM) paradigm, various implementations and smoothing methods might be considered, and in this study we adopt a model proposed by Hiemstra (2002) as described in Equation 7, which combines an estimate based on document ($P[t_j \mid D_i]$) and on corpus ($P[t_j \mid C]$) (Jelinek-Mercer smoothing method).

$$P[D_i \mid Q] = P[D_i] \cdot \prod_{tj \in Q} [\lambda_j \cdot P[t_j \mid D_i] + (1-\lambda_j) \cdot P[t_j \mid C]]$$
$$\text{with } P[t_j \mid D_i] = tf_{ij}/l_i \text{ and } P[t_j \mid C] = df_j/lc \qquad \text{with } lc = \textstyle\sum_k df_k \tag{7}$$

where λ_j is a smoothing factor (constant for all indexing terms t_j, and usually fixed at 0.35) and lc an estimate of the size of the corpus C.

3.2 Overall Evaluation

To measure the retrieval performance, we adopted the mean average precision (MAP) (computed on the basis of 1,000 retrieved items per request by the new TREC-EVAL program). In the following tables, the best performances under the given conditions (with the same indexing scheme and the same collection) are listed in bold type.

Table 2 shows the MAP obtained by the seven probabilistic models and the classical *tf idf* vector-space model using the German or English collection and three different query formulations (title-only or T, TD, and TDN). In the bottom lines we reported the MAP average over the best 6 IR models (the average is computed without the *tf idf* scheme), and the percent change over the medium (TD) query formulation. The DFR I(n)B2 model for the German and also for the English corpus tend to produce the best retrieval performances.

	Mean average precision				
Query Model \ # of queries	German T 25 queries	German TD 25 queries	German TDN 25 queries	English T 25 queries	English TD 25 queries
DFR PB2	0.3877	0.4177	0.4192	0.2620	0.3101
DFR GL2	0.3793	0.4000	0.4031	0.2578	0.2910
DFR I(n)B2	**0.3940**	**0.4179**	0.4202	**0.2684**	**0.3215**
DFR I(n_e)C2	0.3935	0.4170	0.4121	0.2662	0.3191
LM (λ=0.35)	0.3791	0.4130	**0.4321**	0.2365	0.2883
Okapi	0.3815	0.4069	0.4164	0.2592	0.3039
tf idf	0.2212	0.2391	0.2467	0.1715	0.1959
Mean (top-6 best models)	0.3859	0.4121	0.4172	0.2584	0.3057
% change over TD queries	-6.37%		+1.24%	-15.48%	

Table 2: Mean average precision of various single searching strategies (monolingual, GIRT-4 corpus)

Table 3 lists the evaluations done for Russian (word-based indexing & *n*-gram indexing (McNamee & Mayfield, 2004)). The last three lines in this table indicate the MAP average computed for the 4 IR models, the percent change compared to the medium (TD) query formulation, and the percent change when comparing word-based and 4-gram indexing approaches.

From this table, we can see that when using word-based indexing, the DFR I(n_e)B2 or the LM models tend to perform the best. With the 4-gram indexing approach, the LM model always presents the best performing schemes. The short query formulation (T) tends to produce a better retrieval performance than medium (TD) topic formulation. As shown in the last line, when comparing the word-based and 4-gram indexing systems, the relative difference is seen to be rather short (around 4.6%) and favors the 4-gram approach.

Using our evaluation approach, evaluation differences occur when comparing with values computed according to the official measure (the latter always takes 25 queries into account).

	Mean average precision			
Query type Indexing / stemmer IR Model	Russian T word / light 24 queries	Russian TD word / light 24 queries	Russian T 4-gram 24 queries	Russian TD 4-gram 24 queries
DFR GL2	0.1515	0.1332	0.1617	0.1570
DFR I(n_e)B2	0.1470	**0.1468**	0.1402	0.1358
LM (λ=0.35)	**0.1528**	0.1337	**0.1688**	**0.1669**
Okapi	0.1418	0.1349	0.1499	0.1440
tf idf	0.1047	0.1089	0.1098	0.1132
Mean	0.1484	0.1372	0.1552	0.1509
% change over T	baseline	-7.5%	baseline	-2.72%
over stemming	baseline	baseline	+4.64%	+10.04%

Table 3: Mean average precision of various single search strategies (monolingual, ISISS corpus)

4 Blind-Query Expansion

To provide a better match between user information needs and documents, various query expansion techniques have been suggested. The general principle is to expand the query using words or phrases having similar meanings to, or related to those appearing in the original request. To achieve this, query expansion approaches consider various relationships between these words, along with term selection mechanisms and term weighting schemes. Specific answers regarding the best technique may vary, thus leading to a variety of query expansion approaches (Efthimiadis, 1996).

In our first attempt to find related search terms, we might ask the user to select additional terms to be included in an expanded query. This could be handled interactively through displaying a ranked list of retrieved items returned by the first query. As a second strategy, Rocchio (1971) proposed taking the relevance or non-relevance of top-ranked documents into account, as indicated manually by the user. In this case, a new query would then be built automatically in the form of a linear combination of the term included in the previous query and terms automatically extracted from both relevant (with a positive weight) and non-relevant documents (with a negative weight). Empirical studies have demonstrated that such an approach is usually quite effective.

As a third technique, Buckley *et al.* (1996) suggested that even without looking at them or asking the user, it could be assumed that the top-*k* ranked documents would be relevant. This method, denoted as the pseudo-relevance feedback or blind-query expansion approach does not require user intervention. Moreover, using the MAP as performance measure is a strategy that usually tends to enhance performance measures.

In the current context, we used Rocchio's formulation (denoted "Rocchio") with $\alpha = 0.75$, $\beta = 0.75$, whereby the system was allowed to add *m* terms extracted from the *k* best ranked documents from the original query. For the German corpus (Table 4, third column), such a search technique does not seem to enhance the MAP. For the English collection (Table 5, second and third column), Rocchio's blind query expansion may improve the MAP from +9.3% (DFR PB2, 0.3101 vs. 0.3392) or hurt the retrieval performance -8.72% (Okapi model, 0.3039 vs. 0.2774). For the Russian language (Table 6, second and forth column), blind query expansion improves the MAP (e.g., +28.98% with the Okapi model, 0.1740 vs. 0.1349 or +2.3% with the DFR I(n_e)B2 model, 0.1503 vs. 0.1468).

	Mean average precision			
Query TD	German	German	German	German
PRF model	idf	Rocchio	idf	idf
IR Model / MAP	PB2 **0.4177**	DFR I(n)B2 **0.4179**	DFR I(n)B2 **0.4179**	LM **0.4130**
k doc. / *m* terms	5/70 0.4149	5/70 0.3965	5/70 0.4120	5/70 0.3818
	10/100 0.4068	10/100 0.3965	10/100 0.4025	10/100 0.3879
	10/200 0.4078	10/200 0.3992	10/200 0.4104	10/200 0.3941

Table 4: Mean average precision using blind-query expansion (German GIRT-4 collection)

	Mean average precision			
Query TD	English	English	English	English
PRF model	Rocchio	Rocchio	idf	idf
IR Model / MAP	Okapi **0.3039**	DFR PB2 0.3101	DFR PB2 **0.3101**	LM **0.2883**
k doc. / *m* terms	10/50 0.2774	10/50 **0.3392**	10/50 0.3023	10/50 0.2672
	10/100 0.2776	10/100 0.3366	10/100 0.3032	10/100 0.2725
	10/200 0.2767	10/200 0.3324	10/200 0.3006	10/200 0.2746

Table 5: Mean average precision using blind-query expansion (English GIRT-4 collection)

	Mean average precision			
Query TD	Russian	Russian	Russian	Russian
PRF model	Rocchio	idf	Rocchio	idf
IR Model / MAP	Okapi 0.1349	Okapi 0.1349	DFR I(n_e)B2 0.1468	DFR I(n_e)B2 **0.1468**
k doc. / *m* terms	3/50 0.1737	3/50 **0.1612**	3/50 0.1457	3/50 0.1433
	5/70 **0.1740**	5/70 0.1245	5/70 0.1284	5/70 0.1366
	10/100 0.1733	10/100 0.1251	10/100 **0.1503**	10/100 0.1391

Table 6: Mean average precision using blind-query expansion (Russian, ISISS corpus)

Rocchio's query expansion approach however does not always significantly improve the MAP. Such a query expansion approach is based on term co-occurrence data and tends to include additional terms that occur very frequently in the documents. In such cases, these additional search terms will not always be effective in discriminating between relevant and non-relevant documents, and the final effect on retrieval performance could be negative.

As another pseudo-relevance feedback technique we may apply an *idf*-based approach (denoted “idf” in following tables) (Abdou & Savoy, 2008). In this query expansion scheme, the inclusion of new search terms is based on their *idf* values, tending to enlarge the query with more infrequent terms. Overall this *idf*-based term selection performs rather well and usually its retrieval performance is more robust.

For example, with the Russian language (Table 6, third and fifth column), this idf-based blind query expansion may also improve the MAP (e.g., +19.5% with the Okapi model, 0.1612) but, on the other hand, with the DFR $I(n_e)B2$ model, the MAP is slightly reduced (-2.3% from 0.1468 to 0.1433).

However, the *idf*-based query expansion tends to include rare terms, without considering the context. Thus among the top-*k* retrieved documents such a scheme may add terms appearing far away from where the search terms occurred. The single selection criterion is based only on *idf* values, not the position of those additional terms in the top-ranked documents. This year we investigated retrieval effectiveness when including a second criterion in the selection of terms to be included in the new expanded query. We considered it to be important to expand the query using terms appearing close to a search term (fixed at 10 indexing terms in the current experiments). This short window includes 10 terms to the right and 10 terms to the left of each query term. This type of query expansion method is denoted as “idf-window” in Table 7.

Finally, to find words or expressions related to the current request, we considered using commercial search engines (e.g., Google) or online encyclopedia (e.g., Wikipedia). In this case, we submitted a query containing the short topic formulation (T or title-only) to each information service. When using Google, we fetched the first two text snippets and added them as additional terms to the original topic formulation, forming a new expanded query. When using Wikipedia, we fetched the first returned article and added the ten most frequent terms (*tf*) contained in the extracted article.

	Mean average precision			
Query TD PRF model IR Model / MAP	German Rocchio Okapi **0.4069**	German idf Okapi **0.4069**	German idf + window Okapi 0.4069	German with Google Okapi 0.4096
k doc. / *m* terms	5/50 0.3801 10/50 0.3783 10/200 0.3822	5/50 0.3726 10/50 0.3696 10/200 0.3868	5/50 0.4110 10/50 0.4146 10/200 **0.4247**	**0.4196**

Table 7: Mean average precision using four blind-query expansions (German GIRT-4 collection)

The retrieval effectiveness of our two new query expansion approaches is depicted in Table 7 (German collection) and is compared to two other query expansion techniques. Compared to the performance before query expansion (0.4096), Rocchio's and the idf-based blind query expansion cannot improve the MAP. On the other hand, the variant “idf-window” presents a better retrieval performance (+4.9%, from 0.4069 to 0.4247). Using the first two text snippets returned by Google, we may also enhance slightly the MAP (from 0.4096 to 0.4196, or +2.4%). The MAP variation varied according to approaches and parameter settings, while the largest enhancement could be found using the idf+window technique (forth column in Table 7). Finally, using Google to find related terms or phrases implied that we required more processing time.

5 Official Results

Table 8 describes our 9 official runs in the monolingual GIRT task. In this case each run was built using a data fusion operator “Z-Score” (see (Savoy & Berger, 2005)). For all runs, we automatically expanded the queries using the blind relevance feedback method of Rocchio (denoted “Roc”), our IDFQE approach (denoted “idf”), or our new window-based approach (denoted “idf-win”). Finally Table 8 depicts the MAP obtained for the Russian collection when considering 24 queries and in parenthesis, the official MAP computed for 25 queries.

As a complementary search technique, we used two stemmers when defining the official run UniNEDSde3. In this case we first applied our light stemming approach and then a more aggressive one. If the same term was produced by the two stemmers, we only kept one occurrence. On the other hand, if the returned stem differed, we added the two forms to the query formulation.

Run name	Language	Query	Index	Model	Query expansion	MAP	Comb.MAP
UniNEDSde1	German	TD	dec	I(n)B2	Roc 10 docs / 200 terms	0.3992	Z-score
		TD	dec	LM	Google	0.4265	**0.4537**
		TD	dec	PB2	idf-win 10 docs / 150 terms	0.4226	
UniNEDSde2	German	TD	dec	PB2	idf 5 docs / 200 terms	0.4151	Z-score
		TD	dec	I(n)B2		0.4179	0.4399
		TD	dec	I(n)B2	idf-win 10 docs / 200 terms	0.4248	
UniNEDSde3	German	T	dec	I(n)B2		0.3940	Z-score
	special	TD	dec	I(n)B2	idf-win 10 docs / 200 terms	0.4319	0.4251
		TD	dec	I(n_e)C2		0.4170	
UniNEDSde4	German	TD	dec	Okapi	idf-win 5 docs / 50 terms	0.4110	Z-score
		TD	dec	IneC2		0.4170	0.4343
		TD	dec	PB2	idf 10 docs / 200 terms	0.4078	
UniNEDSen1	English	TD	N-stem	InB2	Roc 10 docs / 100 terms	0.3140	Z-score
		TD	N-stem	InB2		0.3562	**0.3770**
		TD	N-stem	LM	Roc 5 docs / 150 terms	0.3677	
UniNEru1	Russian	TD	word/light	I(n_e)B2	Roc 3 docs / 50 terms	0.1457	Z-score
		TD	word/light	I(n_e)B2	idf 5 docs / 70 terms	0.1366	0.1594
							(0.1531)
UniNEru2	Russian	TD	word/light	I(n_e)B2	idf 5 docs / 70 terms	0.1366	Z-score
		TD	word/light	I(n_e)B2	Roc 5 docs / 70 terms	0.1284	0.1628
		TD	word/light	Okapi	Roc 3 docs / 50 terms	0.1737	(0.1563)
UniNEru3	Russian	TD	4-gram	I(n_e)B2	Roc 5 docs / 150 terms	0.1164	Z-score
		TD	word/light	I(n_e)B2	idf 5 docs / 70 terms	0.1366	**0.1655**
		TD	word/light	I(n_e)B2	Roc 5 docs / 70 terms	0.1284	(**0.1589**)
UniNEru4	Russian	TDN	4-gram	I(n_e)B2	Roc 3 docs / 150 terms	0.1129	Z-score
		TDN	word/light	I(n_e)B2	Roc 5 docs / 70 terms	0.1652	**0.1890**
		TDN	word/light	I(n_e)B2	idf 3 docs / 70 terms	0.1739	(**0.1815**)

Table 8: Description and mean average precision (MAP) of our official GIRT runs

5 Conclusion

For our participation in this domain-specific evaluation campaign, we evaluated different probabilistic models using the German, English and Russian languages. For the German and Russian languages we applied our light stemming approach and stopword list. The resulting MAP (see Tables 2 and 3) show that the DFR I(n)B2 or the LM model usually provided in the best retrieval effectiveness. The performance differences between Okapi and the various DFR models were usually rather small.

In our analysis of the blind query expansion approaches (see Tables 4 to 6), we find that this type of automatic query expansion we used can sometimes enhance the MAP. Depending on the collection or languages however, this approach will not provide the same degree of improvement or can sometimes hurt the retrieval effectiveness. For example this search strategy results in less improvement for the English corpus than it does for the Russian collection. For the German collection however, this search strategy clearly hurt the MAP.

This year we suggest two new query expansion techniques. The first, denoted "idf-window", is based on co-occurrence of relatively rare terms in a close context (within 10 terms from the occurrence of a search term in a retrieved document). As a second approach, we add the first two text snippets found by Google to expand the query. Compared to the performance before query expansion (e.g., with Okapi the MAP is 0.4096), Rocchio's and the idf-based blind query expansion cannot improve this retrieval performance. On the other hand, the variant "idf-window" presents a better retrieval performance (+4.9%, from 0.4069 to 0.4247). Using the first two text snippets returned by Google, we may also enhance slightly the MAP (from 0.4096 to 0.4196, or +2.4%).

Acknowledgments

The authors would like to also thank the GIRT - CLEF-2008 task organizers for their efforts in developing domain-specific test-collections. This research was supported in part by the Swiss National Science Foundation under Grant #200021-113273.

References

Abdou, S., & Savoy, J. (2008). Searching in Medline: Stemming, query expansion, and manual indexing evaluation. *Information Processing & Management*, 44(2), p. 781-789.

Amati, G. & van Rijsbergen, C.J. (2002). Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Transactions on Information Systems*, 20(4), p. 357-389.

Buckley, C., Singhal, A., Mitra, M. & Salton, G. (1996). New retrieval approaches using SMART. In *Proceedings of TREC-4*, Gaithersburg: NIST Publication #500-236, p. 25-48.

Efthimiadis, E.N. (1996). Query expansion. *Annual Review of Information Science and Technology*, 31, p. 121-187.

Fautsch, C., Dolamic, L., Savoy, J., (2008). Domain-Specific IR for German, English and Russian Languages. In C. Peters, P. Clough, F.C. Gey, J. Karlgen, B. Magini, D.W. Oard, M. de Rijke & M. Stempfhuber (Eds.), *8th Workshop of the Cross-Language Evaluation Forum*. LNCS #5152, Springer-Verlag, Berlin, p. 196-199.

Hiemstra, D. (2000). Using language models for information retrieval. CTIT Ph.D. Thesis.

Hiemstra, D. (2002). Term-specific smoothing for the language modeling approach to information retrieval. In *Proceedings of the ACM-SIGIR*, The ACM Press, Tempere, p. 35-41.

Kluck, M. (2004). The GIRT data in the evaluation of CLIR systems - from 1997 until 2003. In C. Peters, J. Gonzalo, M. Braschler, M. Kluck (Eds.), *Comparative Evaluation of Multilingual Information Access Systems*. LNCS #3237. Springer-Verlag, Berlin, 2004, p. 376-390.

McNamee, P. & Mayfield, J. (2004). Character *n*-gram tokenization for European language text retrieval. *IR Journal*, 7(1-2), p. 73-97.

Robertson, S.E., Walker, S. & Beaulieu, M. (2000). Experimentation as a way of life: Okapi at TREC. *Information Processing & Management*, 36(1), p. 95-108.

Rocchio, J.J.Jr. (1971). Relevance feedback in information retrieval. In G. Salton (Ed.): *The SMART Retrieval System*. Prentice-Hall Inc., Englewood Cliffs (NJ), p. 313-323.

Savoy, J. (2004). Report on CLEF-2003 monolingual tracks: Fusion of probabilistic models for effective monolingual retrieval. In C. Peters, J. Gonzalo, M. Braschler, M. Kluck (Eds.), *Comparative Evaluation of Multilingual Information Access Systems*. LNCS #3237. Springer-Verlag, Berlin, 2004, p. 322-336.

Savoy, J., & Berger, P.-Y. (2005): Selection and merging strategies for multilingual information retrieval. In: Peters, C., Clough, P., Gonzalo, J., Jones, G.J.F., Kluck, M., Magnini, B. (Eds.): Multilingual Information Access for text, Speech and Images. Lecture Notes in Computer Science: Vol. 3491. Springer, Heidelberg, p. 27-37.

Appendix 1: Parameter Settings

	Okapi			DFR	
Language	*b*	k_1	*avdl*	*c*	*mean dl*
German GIRT	0.55	1.2	200	1.5	200
English GIRT	0.55	1.2	53	4.5	53
Russian word	0.55	1.2	19	1.5	19
Russian 4-gram	0.55	1.2	113	1.5	113

Table A.1: Parameter settings for the various test-collections

Appendix 2: Topic Titles

C201	Health risks at work	C213	Migrant organizations
C202	Political culture and European integration	C214	Violence in old age
C203	Democratic transformation in Eastern Europe	C215	Tobacco advertising
C204	Child and youth welfare in the	C216	Islamist parallel societies in Western Europe
	Russian Federation	C217	Poverty and social exclusion
C205	Minority policy in the Baltic states	C218	Generational differences on the Internet
C206	Environmental justice	C219	(Intellectually) Gifted
C207	Economic growth and environmental		
	destruction	C220	Healthcare for prostitutes
C208	Leisure time mobility	C221	Violence in schools
C209	Doping and sports	C222	Commuting and labor mobility
C210	Establishment of new businesses after		
	the reunification	C223	Media in the preschool age
C211	Shrinking cities	C224	Employment service
C212	Labor market and migration	C225	Chronic illnesses

Table A.2: Query titles for CLEF-2008 GIRT test-collections