Domain-Specific IR for German, English and Russian Languages

Claire Fautsch, Ljiljana Dolamic, Samir Abdou, Jacques Savoy

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

Abstract

In participating in this CLEF evaluation campaign, our first objective is to propose and evaluate various indexing and search strategies for the Russian language, in order to obtain better retrieval effectiveness than that provided by the language-independent approach (*n*-gram). Our second objective is to more effectively measure the relative merit of various search engines when used for the German and to a lesser extent the English language. 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. We also evaluated different pseudo-relevance feedback approaches. For the Russian language, we find that word-based indexing with our light stemming procedure results in better retrieval effectiveness than does 4-gram indexing strategy (relative difference around 30%). Using the GIRT corpora (available in German and English), we examine certain variations in retrieval effectiveness that result from applying the specialized thesaurus to automatically enlarge topic descriptions. In this case, the performance variations were relatively small and usually non significant.

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, Manual Indexing, Digital Libraries, German Language, Russian Language, Thesaurus.

1 Introduction

In our domain-specific retrieval task we access the GIRT (German Indexing and Retrieval Test database) corpus, composed of bibliographic records. These are mainly extracted from two social science sources: SOLIS (social science literature) and FORIS[1] (current research in social science fields), covering Europe's German speaking countries (Germany, Austria, and Switzerland). This collection has grown from 13,000 documents in 1996 to more than 150,000 in 2005, and we are making a continuous effort to enhance the number of documents available, see Kluck (2004) for a more complete description of this corpus.

The fact that scientific documents may contain manually assigned keywords is of particular interest to us in our work. They are usually extracted from a controlled vocabulary by librarians who are knowledgeable of the domain to which the indexed articles belong. These descriptors should be helpful in improving document surrogates and thus the extraction of more pertinent information, and at the same time discarding irrelevant abstracts. Access to the underlying thesaurus would also improve the retrieval performance.

[1] See the Web sites http://www.gesis.org/Information/SOLIS/ and http://www.gesis.org/Information/FORIS/

The rest of this paper is organized as follows: Section 2 describes the main characteristics of the GIRT-4 and ISISS test-collections. Section 3 outlines the main aspects of our stopword lists and light stemming procedures. Section 4 analyses the principal features of various indexing and search strategies, and evaluates their use with the available corpora. Section 5 presents our official runs and results.

<DOC>
<DOCNO> GIRT-DE19908362
<TITLE-DE> Auswirkungen der Informationstechnologien auf die zukünftigen Beschäftigungs- und Ausbildungsperspektiven in der EG
<AUTHOR> Riedel, Monika
<AUTHOR> Wagner, Michael
<PUBLICATION-YEAR> 1990
<LANGUAGE-CODE> DE
<CONTROLLED-TERM-DE> EG
<CONTROLLED-TERM-DE> Informationstechnologie
<CONTROLLED-TERM-DE> Beschäftigungsentwicklung
<CONTROLLED-TERM-DE> Berufsbildung
<CONTROLLED-TERM-DE> Qualifikationsanforderungen
<METHOD-TERM-DE> beschreibend
<METHOD-TERM-DE> Aktenanalyse
<METHOD-TERM-DE> Interpretation
<CLASSIFICATION-TEXT-DE> Arbeitsmarkt- und Berufsforschung
<CLASSIFICATION-TEXT-DE> Arbeitsmarktforschung
<CLASSIFICATION-TEXT-DE> Berufsforschung, Berufssoziologie
<CLASSIFICATION-TEXT-DE > Bildungswesen quartärer Bereich
<ABSTRACT-DE> Veränderungen der Qualifikationsanforderungen an Beschäftigte im IT-Sektor und Zukunftsprojektionen. …

Figure 1: Record example written in German

<DOC>
<DOCNO> GIRT-EN19904041
<TITLE-EN > Impact of Information Technologies on Future Employment and Training Perspectives in the EC
<AUTHOR> Riedel, Monika
<AUTHOR> Wagner, Michael
<PUBLICATION-YEAR> 1990
<LANGUAGE-CODE> EN
<CONTROLLED-TERM-EN> EC
<CONTROLLED-TERM-EN> information technology
<CONTROLLED-TERM-EN> employment trend
<CONTROLLED-TERM-EN> vocational education
<CONTROLLED-TERM-EN> qualification requirements
<METHOD-TERM-EN> document analysis
<CLASSIFICATION-TEXT-EN> Employment Research
<CLASSIFICATION-TEXT-EN> Labor Market Research
<CLASSIFICATION-TEXT-EN > Occupational Research, Occupational Sociology
<CLASSIFICATION-TEXT-EN> Vocational Training, Adult Education …

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

2 Overview of Test-Collections

In the domain-specific retrieval task (called GIRT), 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 <TITLE-DE>), 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 have a title, and that 96.4% of them have 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 which is the "Thesaurus for the Social Sciences" (or GIRT Thesaurus). Finally, associated with each record is a unique identifier ("<DOCNO>"). Kluck (2004) provides a more complete description of this corpus.

The above-mentioned German collection was translated into British English, mainly by professional translators who are native English speakers. 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-TEXT-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 also have access to the GIRT thesaurus. Figure 3 shows some examples of four typical entries in this thesaurus. Each main entry includes the tag <GERMAN> followed by the descriptor written in the German language. Its corresponding uppercase form without diacritics or "ß" appears under the tag <GERMAN-CAPS>. The British English translation follows the label <ENGLISH-TRANSLATION>. The hierarchical relationships between the different descriptors are shown under the labels <BROADER-TERM> (a term having a broader semantic coverage) and <NARROWER-TERM> (a more specific term). The relationship <RELATED-TERM> is used to provide additional pertinent descriptors (similar to the relationship "see also …" often found in many controlled vocabularies). The tag <USE-INSTEAD> is used to redirected readers to another entry (usually a synonym of an existing entry or to indicate that an acronym exists). The tag <USE-COMBINATION> is sometimes used to indicate a possible decompounded or simplified term variant, or more generally a similar term. Usually however, the <USE-COMBINATION> is used like <USE-INSTEAD> to refer from a non-descriptor to a descriptor but having usually more than one descriptor that should be used in combination.

In the GIRT thesaurus are found 10,623 entries (all with both the tag <GERMAN> and <GERMAN-CAPS>) together with 9,705 English translations. Also found are 2,947 <BROADER-TERM> relationships and 2,853 <NARROWER-TERM> links. The synonym relationship between terms can be expressed through the <USE-INSTEAD> (2,153 links), <RELATED-TERM> (1,528) or <USE-COMBINATION> (3,263).

<ENTRY>
<GERMAN> Raumwahrnehmung
<GERMAN-CAPS> RAUMWAHRNEHMUNG
<BROADER-TERM> Wahrnehmung
<RELATED-TERM> Perspektive
<ENGLISH-TRANSLATION> spatial orientation
</ENTRY>

<ENTRY>
<GERMAN> Volksabstimmung
<GERMAN-CAPS> VOLKSABSTIMMUNG
<BROADER-TERM> direkte Demokratie
<NARROWER-TERM> Volksbegehren
<NARROWER-TERM> Volksentscheid
<ENGLISH-TRANSLATION> plebiscite
</ENTRY>
…

<ENTRY>
<GERMAN> Volksstamm
<GERMAN-CAPS> VOLKSSTAMM
<USE-INSTEAD> ethnische Gruppe
<ENGLISH-TRANSLATION> tribe
</ENTRY>

<ENTRY>
<GERMAN> Wachstumspolitik
<GERMAN-CAPS> WACHSTUMSPOLITIK
<USE-COMBINATION> Wirtschaftspolitik
<USE-COMBINATION> Wirtschaftswachstum
<ENGLISH-TRANSLATION> policy of economic growth
</ENTRY>

Figure 3: Example on four different entries in the GIRT thesaurus

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 (12 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 smaller (38.4). For the English corpus, we do not include the *CSA Sociological Abstracts* (20,000 documents, 38.5 MB) in our evaluation. The fact that the relevance assessments contain 1,032 items extracted from this sub-collection implies that our retrieval effectiveness measures for the English corpus are lower than expected.

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 in our experiments they will be ignored.

As shown in Appendix 2, the available topics cover various subjects (e.g., Topic #176: "Sibling relations," Topic #178: "German-French relations after 1945," Topic #196: "Tourism industry in Germany," or Topic #199: "European climate policy"), and some of them may cover a relative large domain (e.g. Topic #187: "Migration pressure").

	German	English	Russian
Size (in MB)	326 MB	235 MB	12 MB
# of documents	151,319	151,319	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	38.4
Standard deviation	44.5	42.41	20.8
Median	85	39	32
Maximum	629	534	187
Minimum	4	4	4
Number of queries	25	25	22
Number rel. items	3,689	4,530	1,471
Mean rel./ request	147.56	181.2	66.86
Standard deviation	102.527	146.04	75.31
Median	99	127	51
Maximum	395 (T#195)	497 (T#187)	335 (T#187)
Minimum	35 (T#186)	18 (T#181)	1 (T#192)

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

3 Stopword Lists and Stemming Procedures

During this evaluation campaign, we used the same stopword lists and stemmers that we selected for our previous English and German language CLEF participation (Savoy, 2004a). Thus for English it was the SMART stemmer and stopword list (containing 571 items), while for the German we applied our light stemmer (available at http://www.unine.ch/info/clef/) and stopword list (603 words). For all our German experiments we applied our decompounding algorithm (Savoy, 2004b).

For the Russian language, we designed and implemented a new light stemmer that removes only inflectional suffixes attached to nouns or adjectives. This stemmer applies 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). The stemmer also applied three normalization rules in order to correct certain variations that occur when a particular suffix is attached to a noun or adjective. See Appendix 3 for a list all this new stemmer's rules.

4 IR Models and Evaluation

4.1. Indexing and Search Strategies

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) is composed by the term occurrence frequency (denoted tf_{ij} for indexing term t_j in document D_i) and 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 called GL2 was based on the following equations:

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

$$Prob^1_{ij} = [1 / (1+\lambda_j)] \cdot [\lambda_j / (1+\lambda_j)]^{tfn_{ij}} \quad \text{with } \lambda_j = tc_j / n \tag{2}$$

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. In our experiments, we fixed the constants values according to the values given in the Appendix 1.

For the second model called PL2, $Prob^2_{ij}$ was obtained from Equation 1, and $Prob^1_{ij}$ was modified as:

$$Prob^1_{ij} = (e^{-\lambda j} \cdot \lambda^{tfij}) / tf_{ij}! \tag{3}$$

For the third model called I(n)L2, we still used Equation 1 to compute $Prob^2_{ij}$ but the implementation of Inf^1_{ij} was modified as:

$$Inf^1_{ij} = tfn_{ij} \cdot \log_2[(n+1) / (df_j+0.5)] \tag{4}$$

For the fourth model called PB2, the implementation of $Prob^1_{ij}$ was obtained by Equation 3, and for evaluating $Prob^2_{ij}$ we used:

$$Prob^2_{ij} = 1- [(tc_j+1) / (df_j \cdot (tf_{ij}+1))] \tag{5}$$

For the fifth model called I(n)B2, the implementation of Inf^1_{ij} was obtained from Equation 4 while $Prob^2_{ij}$ was provided by Equation 5.

Finally, we also considered 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 probability estimates would not be based on any known distribution (as in Equations 2, or 3), but rather be estimated directly based on occurrence frequencies of document D or corpus C. Within this language model (LM) paradigm, various implementations and smoothing methods might be considered, and in this study we adopted a model proposed by Hiemstra (2002) as described in Equation 6, which combines an estimate based on document ($P[t_j | D_i]$) and on corpus ($P[t_j | C]$).

$$P[D_i | Q] = P[D_i] \cdot \prod_{t_j \in Q} [\lambda_j \cdot P[t_j | D_i] + (1-\lambda_j) \cdot P[t_j | C]]$$
$$\text{with } P[t_j | D_i] = tf_{ij}/l_i \quad \text{and } P[t_j | C] = df_j/lc \quad \text{with } lc = \sum_k df_k \tag{6}$$

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.

4.2. 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 performance under the given conditions (with the same indexing scheme and the same collection) is listed in bold type. For the English corpus, our evaluation measures are lower than expected due to the fact that our IR system does not take account for the CSA collection.

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 7 IR models (the average is computed without the *tf idf* scheme), and the percentage of change over the medium (TD) query formulation. The DFR I(n)B2 model for the German language or also the Okapi model when searching into the English corpus tends to produce the best retrieval performance.

Table 3 reports the evaluations done for the Russian language (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 percentage of change compared to the medium (TD) query formulation, and the percentage of change when comparing word-based and 4-gram indexing approaches.

From this table, we can see that the best performing model when using word-based indexing strategy tends to be the DFR I(n)B2 or the DFR GL2 model. With the 4-gram indexing approach, we may also include the LM model in the set of the best performing schemes. The improvement over the medium query formulation (TD) is greater than 25%, a clear and important enhancement. As shown in the last line, when comparing word-based and 4-gram indexing system, we can see that the relative difference is rather large (around 30%) and favors the word-based approach.

	Mean average precision				
Query Model \ # of queries	German T 25 queries	German TD 25 queries	German TDN 25 queries	English TD 25 queries	English TDN 25 queries
DFR PB2	0.2375	0.2635	0.2820	0.3122	**0.3329**
DFR PL2	0.2288	0.2500	0.2885	0.2978	0.3114
DFR GL2	0.2363	0.2608	0.2905	0.2710	0.2852
DFR I(n)B2	**0.2605**	**0.2898**	0.2983	**0.3130**	0.3254
DFR I(n)L2	0.2469	0.2700	**0.3015**	0.2896	0.3254
LM (λ=0.35)	0.2271	0.2526	0.2993	0.2603	0.2929
Okapi	0.2432	0.2616	0.2927	0.2549	0.2501
tf idf	0.1704	0.1835	0.2019	0.1980	0.2091
Mean (top-7 best models)	0.2400	0.2640	0.2933	0.2855	0.3033
% change over TD queries	-9.09%		+11.1%		+6.2%

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

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 TD word / light 22 queries	Russian TDN word / light 22 queries	Russian TD 4-gram / none 22 queries	Russian TDN 4-gram / none 22 queries
DFR GL2	0.1639	**0.2170**	**0.1264**	0.1498
DFR I(n)B2	**0.1775**	0.2062	0.1052	0.1433
LM (λ=0.35)	0.1511	0.1952	0.1246	**0.1672**
Okapi	0.1630	0.2064	0.0917	0.1277
tf idf	0.1188	0.1380	0.0918	0.1229
Mean	0.1639	0.2062	0.1120	0.1470
% change over TD	baseline	+25.8%	baseline	+31.3%
over stemming	baseline	baseline	-31.7%	-28.7%

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

4.3. Blind-Query Expansion

	Mean average precision		
Query TD Rocchio' model IR Model / MAP *k* doc. / *m* terms	German 25 queries Okapi 0.2616 5/70 0.2872 10/100 0.3051 10/200 **0.3107**	German 25 queries DFR I(n)B2 0.2898 5/70 0.3298 10/100 0.3349 10/200 **0.3435**	German 25 queries LM 0.2526 5/70 0.3014 10/100 0.2973 10/200 **0.3076**

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

	Mean average precision		
Query TD	English	English	English
Rocchio' model	25 queries	25 queries	25 queries
IR Model / MAP	Okapi 0.2549	DFR PL2 0.2978	LM 0.2603
k doc. / *m* terms	10/50 0.2640	10/50 **0.3426**	10/50 0.2977
	10/100 0.2655	10/100 0.3390	10/100 0.3027
	10/200 **0.2667**	10/200 0.3237	10/200 **0.3077**

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

In an effort to improve search performance we examined pseudo-relevance feedback using Rocchio's formulation (denoted "Roc") (Buckley *et al.*, 1996) 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), enhancement increased from +9.8% (Okapi, 0.2616 vs. 0.2872) to +21.8% (LM model, 0.2526 vs. 0.3076). For the English collection (Table 5), Rocchio's blind query expansion improves the MAP from +3.6% (Okapi, 0.2549 vs. 0.2640) to +18.2% (LM model, 0.2603 vs. 0.3077). For the Russian language (Table 6), blind query expansion may hurt the MAP (e.g., -21.3% with the DFR InB2 model, 0.1775 vs. 0.1397) or improve the retrieval effectiveness (e.g., +8.9% with the LM model, 0.1511 vs. 0.1645). As another pseudo-relevance feedback technique we applied our *idf*-based approach (denoted "idf" in Table 8) (Abdou & Savoy, 2007).

	Mean average precision		
Query TD	Russian	Russian	Russian
PRF Rocchio's model	22 queries	22 queries	22 queries
IR Model / MAP	Okapi 0.1630	DFR InB2 **0.1775**	LM 0.1511
k doc. / *m* terms	5/50 0.1709	5/50 0.1397	5/50 0.1515
	10/20 **0.1712**	10/20 0.1462	10/20 0.1614
	10/60 0.1709	10/60 0.1477	10/10 **0.1645**

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

4.4. Query Expansion Using a Specialized Thesaurus

The GIRT collection has certain interesting aspects from an IR perspective. Each record has manually assigned descriptors (see examples given in Figures 1 and 2) in order to provide more information on the semantic contents of each bibliographic record. Additionally, descriptors from the specialized thesaurus are accessed (see entry examples depicted in Figure 3).

	Mean average precision	
Query Model \ # of queries	German TD 25 queries without thesaurus	German TD 25 queries with thesaurus
DFR PB2	0.2635	0.2470
DFR PL2	0.2500	0.2347
DFR GL2	0.2608	0.2599
DFR I(n)B2	**0.2898**	**0.2877**
DFR I(n)L2	0.2700	0.2664
LM (λ=0.35)	0.2526	0.2336
Okapi	0.2616	0.2610
tf idf	0.1835	0.1805
Mean (top-7 models)	0.2640	0.2558
% change		-3.1%

Table 7: Mean average precision of various IR models with and without using the specialized thesaurus (monolingual, GIRT-4 corpus)

In an effort to improve the mean average precision, we used the GIRT thesaurus to automatically enlarge the query. To achieve this, we considered each entry in the thesaurus as a document and then indexed it. We then took each query in turn and used it to retrieve the thesaurus entries. Since the number of retrieved thesaurus entries was relatively small, we simply added all these thesaurus entries to the query, forming a new and enlarged one. Although certain terms occurring in the original query were repeated, in other cases this procedure added related terms. If for example the topic included the country name "Deutschland", our thesaurus-based query expansion

procedure might add the related term "BDR" and "Bundesrepublik". Thus, these two terms would usually be helpful to retrieve more pertinent articles.

Using the TD query formulation, MAP differences were relatively small (around -3.1%, in average). We believe that one possible explanation for this relatively small difference was that a query might be expanded with frequently used terms that would not be really effective in discriminating between the relevant and irrelevant items.

5 Official Results

Run name	Language	Query	Index	Model	Query expansion	Single MAP	Comb. MAP
UniNEde1	German	TD	dec	PL2	Roc 10 docs / 120 terms	0.3383	Z-score
		TD	dec	PL2	Roc 10 docs / 120 terms	0.3383	0.3403
		TD	dec	InB2		0.2898	
		TD	dec	PB2	idf 10 docs / 150 terms	0.3261	
UniNEde2	German	TD	dec	PB2	idf 10 docs / 150 terms	0.3261	Z-score
		TD	dec	InL2	Roc 10 docs / 100 terms	0.3525	0.3531
		TD	dec	LM	Roc 10 docs / 230 terms	0.3367	
UniNEde3	German	TD	dec	PL2	Roc 10 docs / 120 terms	0.3383	Z-score
		TD	dec	InB2		0.2898	**0.3535**
thesaurus		TD	dec	PL2	Roc 10 docs / 120 terms	0.3431	
		TD	dec	InB2	idf 10 docs / 230 terms	0.3444	
UniNEde4	German	TDN	dec	PL2	Roc 10 docs / 120 terms	0.3973	Z-score
		TDN	dec	InB2		0.2983	0.3604
		TDN	dec	PB2	idf 10 docs / 150 terms	0.2995	
UniNEen1	English	TD	word	GL2	Roc 10 docs / 100 terms	0.3080	Z-score
		TD	word	PB2	Roc 10 docs / 150 terms	0.3165	**0.3472**
		TD	word	InB2		0.3130	
UniNEen2	English	TD	word	LM2	Roc 10 docs / 150 terms	0.3054	Z-score
		TD	word	Okapi	idf 10 docs / 150 terms	0.2734	0.3038
UniNEen3	English	TD	word	PL2	Roc 10 docs / 50 terms	0.3426	Z-score
		TD	word	PL2		0.2978	0.3399
		TD	word	InB2	idf 10 docs / 150 terms	0.2901	
UniNEen4	English	TDN	word	GL2	Roc 10 docs / 100 terms	0.3182	Z-score
		TDN	word	PB2	Roc 10 docs / 150 terms	0.3361	0.3534
		TDN	word	InB2		0.3254	
UniNEru1	Russian	TD	wd/light	Okapi	idf 10 docs / 20 terms	0.1552	Z-score
		TD	wd/light	GL2	Roc 10 docs / 20 terms	0.1445	0.1560
		TD	wd/light	LM	idf 10 docs / 60 terms	0.1557	(0.1372)
UniNEru2	Russian	TD	wd/light	GL2	idf 5 docs / 50 terms	0.1410	Z-score
		TD	4-gram	GL2	Roc 5 docs / 50 terms	0.1335	0.1520
		TD	wd/light	GL2	Roc 5 docs / 50 terms	0.1476	(0.1338)
UniNEru3	Russian	TD	wd/light	Okapi	idf 5 docs / 50 terms	0.1579	Z-score
		TD	4-gram	LM	Roc 5 docs / 50 terms	0.1331	**0.1648**
		TD	wd/light	LM	Roc 10 docs / 60 terms	0.1645	(0.1450)
		TD	4-gram	GL2	Roc 5 docs / 50 terms	0.1335	
UniNEru4	Russian	TDN	wd/light	GL2	Roc 5 docs / 50 terms	0.2028	Z-score
		TDN	wd/light	LM	idf 5 docs / 50 terms	0.1789	0.2240
		TDN	4-gram	GL2	Roc 10 docs / 60 terms	0.1686	(0.1971)

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

Table 8 describes our 12 official runs in the monolingual GIRT task. In this case each run was built using a data fusion operator "Z-Score" (see (Savoy & Berger, 2006)). For all runs, we automatically expanded the queries using a blind relevance feedback method, Rocchio (denoted "Roc") or our IDFQE approach (denoted "idf"). In order to obtain more relevant documents in the pool, we also submitted three runs using the TDN queries (UniNEde4, UniNEen4, and UniNEru4). For the English collection the runs retrieved only documents from the

GIRT-4 collection and thus we have ignored the CSA corpus. The MAP values achieved for this language are therefore clearly below the expected performance. Finally for the Russian collection, Table 8 depicts the MAP achieved when considering 22 queries and in parenthesis, the official MAP computed with 25 queries.

6 Conclusion

For our participation in this domain-specific evaluation campaign, we propose a new light stemmer for the Russian language. The resulting MAP (see Table 3) shows that for this Slavic language our approach may produce better MAP than a 4-gram approach (relative difference around 30%). For the German corpus, we try to exploit the specialized thesaurus in order to improve the resulting MAP. The retrieval effectiveness difference is rather small and we still need to analyze the reasons for obtaining so little difference (see Table 7). We believe that a more specific query enrichment procedure is needed, one able to take the various different term-term relationships into account, along with the occurrence frequencies of the potential new search terms.

When comparing the various IR models (see Table 2), we found that the I(n)B2 model derived from the *Divergence from Randomness* (DFR) paradigm tends usually to result in the best performance. When analyzing blind query expansion approaches (see Tables 4 to 6), we find that this type of automatic query expansion can enhance MAP but there is clearly larger improvement when using the LM model. Finally for the Russian corpus, this search strategy produces less improvement than for the English or German collections.

Acknowledgments

The authors would like to also thank the GIRT - CLEF-2007 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., (2007). Searching in Medline: Stemming, query expansion, and manual indexing evaluation. Information Processing & Management, to appear.

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), 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, 25-48.

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, 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, 376-390.

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

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

Savoy, J. (2004a). Combining multiple strategies for effective monolingual and cross-lingual retrieval. *IR Journal*, 7(1-2), 121-148.

Savoy, J. (2004b). 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, 322-336.

Savoy, J. & Berger, P.-Y. (2006). Monolingual, Bilingual and GIRT Information Retrieval at CLEF 2005. In C. Peters, P. Clough, J. Gonzalo, G.J.F. Jones, M. Kluck & B. Magnini (Eds.), *Multilingual Information Access for Text, Speech and Images*. Springer-Verlag, Berlin, 2006, to appear.

Appendix 1: Parameter Settings

	Okapi			DFR	
Language	*b*	k_1	*avdl*	*c*	*mean dl*
German GIRT	0.55	1.2	500	1.5	200
English GIRT	0.55	1.2	500	2.5	55
Russian ISISS	0.9	4	750	1.5	38

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

Appendix 2: Topic Titles

C176	Sibling relations	C190	Mortality rate
C177	Unemployed youths without vocational training	C191	Economic elites in Eastern Europe and Russia
C178	German-French relations after 1945	C192	System change and family planning in East Germany
C179	Multinational corporations		
C180	Partnership and desire for children	C193	Gender and career chances
C181	Torture in the constitutional state	C194	Ecological standards in emerging or
C182	Family policy and national economy	C193	Gender and career chances
C183	Women and income level	C195	Integration policy
C184	Lifestyle and environmental behaviour	C196	Tourism industry in Germany
C185	Unstable employment situations	C197	Promoting health in the workplace
C186	Value change in Eastern Europe	C198	Economic situations of families
C187	Migration pressure	C199	European climate policy
C188	Quality of life of elderly persons	C200	Economic support in the East
C189	Class-specific leisure behaviour		

Table A.2: Query titles for CLEF-2007 GIRT test-collections

Appendix 3: The Russian Stemmer

```
RussianStemmer (word)  {
  RemoveCase (word);
  Normalize (word);
  return;
  }

Normalize(word)  {
  if (word ends with “-ь”) then remove “-ь” return;
  if (word ends with “-и”) then remove “-и”return;
  if (word ends with “-нн”) then replace by “-н” return;
  return;

RemoveCase (word)  {
  if (word ends with “-иями”) then remove “-иями” return;
  if (word ends with “-оями”) then remove “-оями” return;
  if (word ends with “-оиев”) then remove “-оиев” return;
  if (word ends with “-иях”) then remove “-иях” return;
  if (word ends with “-иям”) then remove “-иям” return;
  if (word ends with “-ями”) then remove “-ями” return;
```

```
if (word ends with “-оям”) then remove “-оям” return;
if (word ends with “-оях”) then remove “-оях” return;
if (word ends with “-ами”) then remove “-ами” return;
if (word ends with “-его”) then remove “-его” return;
if (word ends with “-ему”) then remove “-ему” return;
if (word ends with “-ери”) then remove “-ери” return;
if (word ends with “-ими”) then remove “-ими” return;
if (word ends with “-иев”) then remove “-иев” return;
if (word ends with “-ого”) then remove “-ого” return;
if (word ends with “-ому”) then remove “-ому” return;
if (word ends with “-ыми”) then remove “-ыми” return;
if (word ends with “-оев”) then remove “-оев” return;
if (word ends with “-яя”) then remove “-яя” return;
if (word ends with “-ях”) then remove “-ях” return;
if (word ends with “-юю”) then remove “-юю” return;
if (word ends with “-ая”) then remove “-ая” return;
if (word ends with “-ах”) then remove “-ах” return;
if (word ends with “-ею”) then remove “-ею” return;
if (word ends with “-их”) then remove “-их” return;
if (word ends with “-ия”) then remove “-ия” return;
if (word ends with “-ию”) then remove “-ию” return;
if (word ends with “-ие”) then remove “-ие” return;
if (word ends with “-ий”) then remove “-ий” return;
if (word ends with “-им”) then remove “-им” return;
if (word ends with “-ое”) then remove “-ое” return;
if (word ends with “-ом”) then remove “-ом” return;
if (word ends with “-ой”) then remove “-ой” return;
if (word ends with “-ов”) then remove “-ов” return;
if (word ends with “-ые”) then remove “-ые” return;
if (word ends with “-ый”) then remove “-ый” return;
if (word ends with “-ым”) then remove “-ым” return;
if (word ends with “-ми”) then remove “-ми” return;
if (word ends with “-ою”) then remove “-ою” return;
if (word ends with “-ую”) then remove “-ую” return;
if (word ends with “-ям”) then remove “-ям” return;
if (word ends with “-ых”) then remove “-ых” return;
if (word ends with “-ея”) then remove “-ея” return;
if (word ends with “-ам”) then remove “-ам” return;
if (word ends with “-ее”) then remove “-ее” return;
if (word ends with “-ей”) then remove “-ей” return;
if (word ends with “-ем”) then remove “-ем” return;
if (word ends with “-ев”) then remove “-ев” return;
if (word ends with “-я”) then remove “-я” return;
if (word ends with “-ю”) then remove “-ю” return;
if(word ends with “-й”) then remove “-й” return;
if (word ends with “-ы”) then remove “-ы” return;
if (word ends with “-[аеиоу]”) then remove “-[аеиоу]” return;
}
```

Table A.3: Our light stemmer for the Russian language