Published in Lecture Notes in Computer Science 5152, 196-199, 2008
which should be used for any reference to this work

Domain-Specific IR for German, English and Russian Languages

Claire Fautsch, Ljiljana Dolamic, Samir Abdou, and Jacques Savoy

Computer Science Department, University of Neuchatel, Rue Emile Argand 11, 2009 Neuchatel, Switzerland
{Claire.Fautsch,Ljiljana.Dolamic,Samir.Abdou,Jacques.Savoy}@unine.ch

Abstract. In participating in this domain-specific track, our first objective is to propose and evaluate a light stemmer for the Russian language. Our second objective is to measure the relative merit of various search engines used for the German and to a lesser extent the English languages. To do so we evaluated the $tf \cdot idf$, Okapi, IR models derived from the *Divergence from Randomness* (DFR) paradigm, and also a language model (LM). For the Russian language, we find that word-based indexing using our light stemming procedure results in better retrieval effectiveness than does the 4-gram indexing strategy (relative difference around 30%). Using the German corpus, we examine certain variations in retrieval effectiveness after applying the specialized thesaurus to automatically enlarge topic descriptions. In this case, the performance variations were relatively small and usually non significant.

1 Introduction

In the domain-specific retrieval task we access the GIRT (German Indexing and Retrieval Test database) corpus, composed of bibliographic records extracted from two social science sources. This collection has grown from 13,000 documents in 1996 to more than 150,000 in 2005 (a more complete description of this corpus and the main results of this track can be found in [1]).

The manually assigned keywords contained in scientific documents are of particular interest to us, especially given that they are extracted from a controlled vocabulary by librarians. Through using this vocabulary and the corresponding thesaurus we hope to automatically enlarge the submitted queries and therefore improve retrieval performance.

2 Indexing and Searching Strategies

In order to obtain higher MAP values, we considered certain probabilistic models, such as the Okapi (or BM25). As a second probabilistic approach, we implemented variants of the DFR [2] (*Divergence from Randomness*) paradigm. We also examined an approach based on a statistical language model (LM) [3], also known as a non-parametric probabilistic model (a precise definition of these IR

models may be found at [4]). For comparison purpose, we also added the classical $tf \cdot idf$ model (with cosine normalization).

To measure retrieval performance, we adopted mean average precision (MAP) computed by `trec_eval`, based on 25 queries for the German and English corpora, and 22 for the Russian language. In the following tables, the best performance under a given condition is shown in bold type.

Table 1 lists evaluation results obtained using the Russian collection, combined with medium (TD) or long query formulations (TDN), along with two different indexing strategies (word-based using a light stemmer (inflectional only) and n-gram [5] scheme). An analysis of this data shows that the DFR model is the best performing of the IR models. This data also shows that the word-based approach uses the best indexing strategy. Taking this strategy as a baseline, the average performance difference for a 4-gram indexing strategy is around 29.5% (with TD query formulation) or 25% (with TDN queries).

Table 1. Evaluation of the Russian Corpus (22 queries)

	Mean average precision			
Query	TD	TD	TDN	TDN
Indexing	word+light	4-gram	word+light	4-gram
Okapi	0.1630	0.0917	0.2064	0.1277
DFR-GL2	0.1639	**0.1264**	**0.2170**	0.1498
DFR-I(n)B2	**0.1775**	0.1052	0.2062	0.1433
LM	0.1511	0.1246	0.1952	**0.1672**
tf idf	0.1188	0.0918	0.1380	0.1229

Evaluations done on the German and English GIRT corpora are depicted in Table 2. In this case, we compared two query formulations (TD vs. TDN) and automatically enlarged topic descriptions, using the GIRT thesaurus. To achieve this we considered each entry in the thesaurus as a document, and then for each query we retrieved the thesaurus entries. Given the relatively small number of retrieved entries, we simply added all of them to the query to form a new and enlarged one. Although certain terms occurring in the original query were repeated, the procedure added related terms in other cases. If for example the topic included the name "Deutschland", our thesaurus-based query expansion procedure might add the related term "BRD" and "Bundesrepublik". Thus, these two terms would usually be helpful in retrieving more pertinent articles.

The results shown in Table 2 indicate that the best performing IR approach was usually the DFR-I(n)B2 model. Enlarging the query with terms extracted from the thesaurus does not improve the MAP. Rather, the contrary tends to be true, for they slightly reduce retrieval performance. Moreover, performance differences between the TD and TDN query formulations seem to be around 11.3% (German corpus with a decompounding stage) or 6.2% (English collection).

Upon looking at some queries more carefully, we can see when and why our search strategy fails to place pertinent articles at the top of the returned list. For

Table 2. Evaluation of German and English Corpora (25 queries)

	Mean average precision				
Language Query Indexing	German TD word	German TD + thesaurus	German TDN word	English TD word	English TDN word
Okapi	0.2616	0.2610	0.2927	0.2549	0.2501
DFR-GL2	0.2608	0.2599	0.2905	0.2710	0.2852
DFR-I(n)B2	**0.2898**	**0.2877**	0.2983	**0.3130**	**0.3254**
LM	0.2526	0.2336	**0.2993**	0.2603	0.2929
tf idf	0.1835	0.1805	0.2019	0.1980	0.2091

the German corpus, using the GIRT thesaurus, our system automatically added the term "Osterweiterung" related to the query term "Europäisch". In general a relationship exists between these two terms but not in the context of Topic #199 ("Europäische Klimapolitik"). Generally, specific search terms would not have an entry in the GIRT thesaurus, yet for more frequent and less important words we might find some related terms in the thesaurus. Adding such terms did not help us find more relevant items.

From our observations we noted that another source of failure was the use of different word phrases to express the same concept. For Topic #171 ("Sibling relations") there were two relevant items using the term "семейные" (family) but not the word "братьями" ("brothers") or "сестрами" ("sisters") used in the Russian topic formulation. Finally our search system encountered a real problem with Topic #192 ("System change and family planning in East Germany"). In this case, the only term common to the query formulations and the single relevant article was the frequently appearing noun "Germany"

3 Official Results

To define our official runs as described in Table 3, we first applied a pseudo-relevance feedback using Rocchio's formulation [6] with $\alpha = 0.75$, $\beta = 0.75$, whereby the system was allowed to add m terms extracted from the k best ranked documents (the exact values used in our experiments are listed in Table 3).

In a second step, we combined three or four probabilistic models, representing both the parametric (Okapi and DFR) and non-parametric(LM) approaches. All runs were fully automatic and in all cases we applied the same data fusion approach (Z-score [4]). For the German corpus however we applied our decompounding approach (denoted by "dec." in the "Index" column). For the English corpus our data fusion strategy clearly enhanced retrieval performance, but for the German or Russian, we obtained only slight improvements.

For our participation in this domain-specific evaluation campaign, we proposed a new light stemmer for the Russian language. The resulting MAP (see Table 1) shows that for this Slavic language our approach may produce better MAP than a 4-gram approach (relative difference around 30%). For the German

Table 3. Description and MAP Results for Our Best Official Monolingual Runs

Language	Index	Query	Model	Query exp.	MAP	comb. MAP
German	dec.	TD	PL2	10 docs/120 terms	0.3383	Z-score
UniNEde3	dec.	TD	InB2		0.2898	**0.3535**
	dec.	TD	PL2	10 docs/120 terms	0.3431	
	dec.	TD	InB2	10 docs/230 terms	0.3444	
English	word	TD	GL2	10 docs/100 terms	0.3080	Z-score
UniNEen1	word	TD	PB2	10 docs/150 terms	0.3165	**0.3472**
	word	TD	InB2		0.3130	
Russian	word	TD	Okapi	5 docs/50 terms	0.1579	Z-score
UniNEru3	4-gram	TD	LM	5 docs/50 terms	0.1331	**0.1648**
	word	TD	LM	10 docs/60 terms	0.1645	(0.1450)
	4-gram	TD	GL2	5 docs/50 terms	0.1335	

corpus, we tried to exploit the specialized thesaurus in order to improve the resulting MAP, yet retrieval effectiveness differences are rather small. We thus believe that a more specific query enrichment procedure is needed, one that is able to take the various different term-term relationships into account, along with the occurrence frequencies for the potential new search terms. Upon comparing the various IR models (see Table 1), we found that the I(n)B2 model derived from the *Divergence from Randomness* (DFR) paradigm would usually provide the best performance.

Acknowledgments. This research was supported in part by the Swiss National Science Foundation under Grant #200021-113273.

References

1. Petras, V., Baerisch, S., Stempfhuber, M.: The Domain-Specific Track at CLEF 2007. In: Peters, C., et al. (eds.) CLEF 2007. LNCS, vol. 5152, pp. 160–173. Springer, Heidelberg (2008)
2. Amati, G., van Rijsbergen, C.J.: Probabilistic Models of Information Retrieval Based on Measuring the Divergence from Randomness. ACM Transactions on Information Systems 20, 357–389 (2002)
3. Hiemstra, D.: Using Language Models for Information Retrieval. PhD Thesis (2000)
4. Dolamic, L., Savoy, J.: Stemming Approaches for East European Languages. In: Peters, C., et al. (eds.) CLEF 2007. LNCS, vol. 5152, pp. 37–44. Springer, Heidelberg (2008)
5. McNamee, P., Mayfield, J.: Character N-gram Tokenization for European Language Text Retrieval. IR Journal 7, 73–97 (2004)
6. Buckley, C., Singhal, A., Mitra, M., Salton, G.: New Retrieval Approaches Using SMART. In: Proceedings TREC-4, Gaithersburg, pp. 25–48 (1996)