UniNE at CLEF 2006: Experiments with Monolingual, Bilingual, Domain-Specific and Robust Retrieval

Jacques Savoy, Samir Abdou

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

Abstract

For our participation in this CLEF evaluation campaign, the first objective was to propose and evaluate various indexing and search strategies for the Hungarian language in order to produce better retrieval effectiveness than language-independent approach (*n*-gram). Using both a new stemmer including some derivational suffixes removals, and a more aggressive automatic decompounding scheme, we were able to produce better retrieval effectiveness than corresponding 4-gram indexing scheme. Our second objective was to obtain a better picture of the relative merit of various search engines with the French, Brazilian/Portuguese and Bulgarian languages. To do so we evaluated these test-collections using the Okapi, *Divergence from Randomness* (DFR) and language model (LM) models together with nine vector-processing approaches. After pseudo-relevance feedback, either the DFR or the LM approach tends to produce the best IR performance. For the Bulgarian language, we also found that word-based indexing proposes usually better retrieval effectiveness than corresponding 4-gram indexing.

In the bilingual track, we evaluated the effectiveness of various machine translation systems to automatically translate a query submitted in English into the French and Portuguese languages. After blind query expansion, the MAP achieved by the best single MT system is around 95% of the corresponding monolingual search when French is the target language, or 83% with the Portuguese. Using the GIRT corpora (available in German and English), we investigated variations in retrieval effectiveness when facing with domain-specific collection composed of relatively short bibliographic notices. Finally, in the robust retrieval task we investigated different techniques in order to improve the retrieval performance of difficult topics. In this track, we found that both the mean average precision and the geometric mean are strongly correlated. Moreover, massive query expansion based on a search engine did not provide better retrieval effectiveness than Rocchio's approach.

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, Bilingual Information Retrieval, Digital Libraries, Hungarian Language, Bulgarian Language, Portuguese Language, French Language.

1 Introduction

During the last years, the IR group at University of Neuchatel is involved in designing, implementing and evaluating IR systems for various natural languages, including both European (Savoy 2004a) and popular Asian (Savoy 2005a) languages (namely, Chinese, Japanese, Korean). In this context, our first objective is to promote

effective monolingual IR in those languages. Our second aim is to design and evaluate effective bilingual search (using a query-based translation approach) and finally to propose effective multilingual IR systems.

During our participation in various evaluation campaigns, the first lesson learned is the fact that the best performing IR models often require relatively small amount of linguistic knowledge and are usually language-independent (when an IR model performs well with a given natural language, it tends to perform well with another language). Thus, statistical features present in each natural language seem to be appropriate in general to distinguish between relevant and non-relevant information items. As a second point, general language-independent indexing strategy (i.e., *n*-gram indexing (McNamee & Mayfield 2004)) presents a reasonable level of performance. For Korean and Chinese, such indexing strategy seems even to be the best choice. For European languages not previously studied, at least in the IR domain, such an indexing approach usually provides one of the best retrieval effectiveness (i.e., the best runs in both Bulgarian and Hungarian monolingual tracks at CLEF 2005 were obtained using such an indexing scheme (McNamee 2005)).

These two findings must be moderated. With a new language or a new corpus written in a known language or a new set of queries against a known test-collection, we are unable to predict *precisely* which IR model will achieve the best retrieval effectiveness. From our past experiments, the Okapi probabilistic model (Robertson *et al*. 2000) presents usually very good retrieval performance. As a second probabilistic approach, models derived from the *Divergence from Randomness* (DFR) family (Amati & van Rijsbergen 2002) provide also high retrieval effectiveness. On the other hand, various implementations based on the language model (Hiemstra 2000; 2002) may also produce the best retrieval performance. Finally, random variations may favor a given IR model for a given set of queries without producing an important performance difference that can be viewed as significant by a statistical test.

When considering the *n*-gram indexing strategy, the best choice for the parameter *n* seems to depend on the language. On the one hand, it seems that for popular Asian languages (i.e., Chinese, Japanese or Korean), the bigram indexing approach provides the best choice (sometimes combined with a character-based indexing strategy). For European languages, the situation is less clear. For example, it seems that the best value of *n* is 4 for the Bulgarian and Hungarian languages and 5 for the English, French and Portuguese languages (McNamee 2005). As usual, the performance differences are not always statistically significant.

Finally, various parameters like difference in stemming strategies, stopword lists, processing of diacritics and uppercase letters, indexing of noun-phrases, etc. differ from participant to participant. Thus comparisons between two runs imply comparing two IR systems with all their components. It is therefore difficult or even impossible to known precisely the impact of each single component when comparing runs provided by two participants.

The rest of this paper is organized as follows: Section 2 describes the main characteristics of the CLEF-2006 test-collections, Section 3 outlines the main aspects of our stopword lists and light stemming procedures. Section 4 analyses the principal features of different indexing and search strategies, and evaluates their use with the available corpora. The data fusion approaches adapted in our experiments are explained in Section 5, and Section 6 depicts our official results. Our bilingual experiments are presented and evaluated in Section 7 while Section 8 describes our experiments involving the domain-specific GIRT corpus. Section 9 presents the main results of our participation in the robust retrieval task, limited however to the French language.

2 Overview of the Test-Collections

The corpora used in our experiments include newspaper and news agency articles, namely *Le Monde* (1994-1995, French), *Schweizerische Depeschenagentur* (1994-1995, French), *Público* (1994-1995, Portuguese), *Folha de Sãn Paolo* (1994-1995, Brazilian), *Magyar Hirlap* (2002, Hungarian), *Sega* (2002, Bulgarian), *Standart* (2002, Bulgarian), *Los Angeles Times* (1994, English), *Glasgow Herald* (1995, English). As shown in Table 1, the English corpus (249.08 indexing terms / document) has a larger mean size article than the Portuguese collection (212.9). This mean value is a little bit lower for the French (178) and relatively similar for the Bulgarian (133.7) and Hungarian (142.1) languages. It is interesting to note that even though the Hungarian collection is the smallest (105 MB), it contains the largest number of distinct indexing terms (657,132), computed after stemming.

During the indexing process in our automatic runs, we retained only the following logical sections from the original documents: <TITLE>, <TEXT>, <LEAD>, <LEAD1>, <TX>, <LD>, <TI> and <ST>. From the topic descriptions we automatically removed certain phrases such as “Relevant document report …”, “Finde Dokumente, die über …”, “Keressünk olyan cikkeket, amelyek …” or “Trouver des documents qui …”, etc.

As shown in the Appendix, the available topics cover various subjects (e.g., "Consumer Boycotts," "Doping in Sports," "Theft of "The Scream," or "Grand Slam Winners"), and some of them may cover a relative large domain (e.g. Query #316: "Strikes," or Query #311 "Unemployment in Europe"), including both regional ("Hungarian-Bulgarian Relationships," "New Quebec Premier") or international coverage ("Energy Crises"). For the French, English and Portuguese collection, we had to use Topics #301 to #350 in which Topics#326 to #350 are covering more specifically the year 1994-95 (e.g., "Civil War in the Yemen," "Nixon's Death"). For the Hungarian and Bulgarian corpus, the query set is formed by Topics #301 to #325 and Topics #351 to #375 (covering more specifically the year 2002-2003 like "The Harry Potter Phenomenon," "Impact of September 11").

	French	Portuguese	Bulgarian	Hungarian	English
Size (in MB)	487 MB	564 MB	213 MB	105 MB	579 MB
# of documents	177,452	210,734	69,195	49,530	169,477
# of distinct terms	455,576	582,117	414,253	657,132	115,181
Number of distinct indexing terms / document					
Mean	127.8	153.5	102.7	107.9	156.874
Standard deviation	106.57	114.95	97.34	94.59	118.773
Median	92	129	72	77	129
Maximum	2,645	2,655	1,242	1,422	1,882
Minimum	1	1	1	2	2
Number of indexing terms / document					
Mean	178	212.9	133.7	142.1	249.08
Standard deviation	159.87	186.4	144.85	139.84	224.71
Median	126	171	88	95	191
Maximum	6,720	7,554	2,805	4,984	6,087
Minimum	1	1	1	2	2
Number of queries	49	50	50	48	49
Number rel. items	2,148	2,677	1,249	1,308	1,258
Mean rel./ request	43.8367	53.54	24.98	27.25	25.6735
Standard deviation	79.7528	52.2475	26.7051	25.2076	26.1784
Median	20	39	15.5	17	17
Maximum	521 (Q#316)	266 (Q#316)	158 (Q#316)	134 (Q#311)	118 (Q#316)
Minimum	1 (Q#336)	2 (Q#334)	2 (Q#301)	4 (Q#367)	1 (Q#306)

Table 1: CLEF 2006 test-collection statistics

3 Stopword Lists and Stemming Procedures

During this evaluation campaign, we mainly used the stopword lists and stemmers used in our CLEF 2005 participation (Savoy & Berger 2006). However, we corrected some errors in our Bulgarian stopword list (we removed words having a clear meaning and introduced by mistake in the suggested stopword list).

For the Hungarian language, our suggested stemmer removes only inflectional suffixes for this language. This year, we have tried to be more aggressive and we added 17 rules in our Hungarian stemmer to remove also some derivational suffixes (e.g., "jelent" (to mean) and "jelentés" (meaning), or "tánc" (to dance) and "táncol" (dance)). Moreover, the Hungarian language uses compound constructions (e.g., handgun, worldwide). In order to increase the matching between search keywords and document representations, we automatically decompounding Hungarian words using our decompounding algorithm (Savoy 2004b), leaving both compound words and their component parts in the documents and queries.

4 IR models and Evaluation

4.1. Indexing and Searching Strategies

In order to obtain a broader view of the relative merit of various retrieval models, we may first adopt a binary indexing scheme in which each document (or request) was represented by a set of keywords, without any weight. To measure the similarity between documents and requests, we computed the inner product (retrieval model denoted "doc=bnn, query=bnn" or "bnn-bnn"). In order to weight the presence of each indexing term in a document surrogate (or in a query), we took the term occurrence frequency into account (denoted tf_{ij} for

indexing term t_j in document D_i, and the corresponding retrieval model was denoted: "doc=nnn, query=nnn"). We might also account for their inverse document frequency (denoted idf_j). Moreover, we might normalize each indexing weight using different weighting schemes, as is described in the Appendix.

In addition to these models based on the vector-space paradigm, we also considered probabilistic models such as the Okapi model (or BM25) (Robertson *et al*. 2000). As a second probabilistic approach, we implemented 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 is based on the following equations:

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

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

For the second model called PL2, only the implementation of $Prob^1_{ij}$ is modified as:

$$Prob^1_{ij} = (e^{-\lambda j} \cdot \lambda^{tfij}) / tf_{ij}! \quad \text{with } \lambda_j = tc_j / n \tag{3}$$

For the third model called I(F)L2, only the implementation of $Prob^1_{ij}$ is modified as:

$$Prob^1_{ij} = \log_2[(tc_j+0.5) / (n+1)]^{tf_{ij}} \tag{4}$$

For the four model called PB2, the implementation of $Prob^1_{ij}$ is given by Equation 3, and

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

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, the constants b, k_1, *avdl*, *pivot*, *slope*, c and *mean dl* were fixed according to the values given in the Appendix).

Finally, we also considered an approach based on a language model (LM) (Hiemstra 2000; 2002), known as a non-parametric probabilistic model (the Okapi and DFR are viewed as parametric models). Probability estimates would thus not be based on any known distribution (as in Equation 2, 3 or 4) but rather be estimated directly, based on occurrence frequencies in document D or corpus C. Within this language model 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 ($P[t_j | C]$).

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

where λ_j is a smoothing factor (constant for all indexing terms t_j, and usually fixed at 0.35) and lc 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). Using this evaluation tool, some evaluation difference may occur with the values computed according to the official measure (the latter always takes account for 50 queries). 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.

Table 2 shows the MAP achieved by four probabilistic models and nine vector-space schemes using the French or the Portuguese collection and three different query formulations (title-only or T, TD, and TDN). In the last lines we have reported the MAP average over these 13 IR models, the average computed over the first ten IR models (ending with "doc=ltc, query=ltc"), and the percentage of improvement over the short (T) query formulation. From this data, we can see that the set of the best performing IR models correspond to the probabilistic one, more the "Lnu-ltc" and "dtu-dtn" vector-space models. As depicted in the last line, increasing the query improves the MAP, but the enhancement is lower than our prior estimation (+15%).

Table 3 reports the evaluations done with the Bulgarian and Hungarian languages (word-based indexing). In this table, the two last lines indicates the MAP average computed over the top-8 IR models (ending with "doc=ltc, query=ltc"), and the percentage of improvement over the short (T) query formulation. Mainly, the same conclusions can be drawn for these two languages. First the probabilistic models expose the best IR

performance and secondly the improvement over the short query formulation is usually greater than 10%. It could be a surprise to see that the vector-space model "dtu-dtn" produces the best performance for the Hungarian language. From a statistical point of view, the difference in performance with the Okapi model could be due to random variations and could be not statistically significant.

	Mean average precision					
Query Model \ # of queries	French T 49 queries	French TD 49 queries	French TDN 49 queries	Portuguese T 50 queries	Portuguese TD 50 queries	Portuguese TDN 50 queries
doc=Okapi, query=npn	**0.3601**	**0.4151**	**0.4489**	**0.3947**	**0.4333**	0.4388
DFR GL2	0.3352	0.3988	0.4457	0.3945	0.4033	**0.4406**
DFR PL2		0.4101			0.4147	
LM (λ=0.35)		0.3913			0.3909	
doc=Lnu, query=ltc	0.3156	0.3738	0.4059	0.3711	0.4212	0.4335
doc=dtu, query=dtn	0.3171	0.3781	0.3960	0.3713	0.4172	0.4257
doc=atn, query=ntc	0.3164	0.3808	0.4134	0.3475	0.3854	0.4043
doc=ltn, query=ntc	0.3051	0.3453	0.3577	0.3277	0.3567	0.3683
doc=ntc, query=ntc	0.2151	0.2606	0.2658	0.2664	0.2959	0.3114
doc=ltc, query=ltc	0.2115	0.2511	0.2703	0.2695	0.3112	0.3395
doc=lnc, query=ltc	0.2083	0.2602	0.2900	0.2702	0.3227	0.3498
doc=bnn, query=bnn	0.1594	0.1628	0.1258	0.1904	0.2037	0.1426
doc=nnn, query=nnn	0.1352	0.1475	0.1412	0.1109	0.1154	0.0929
Mean	0.2544	0.3061	0.3115	0.2920	0.3321	0.3307
Mean over top-10 models	0.2916	0.3495	0.3654	0.3355	0.3765	0.3888
% change over T		+18.05%	+26.41%		+10.68%	+15.29%

Table 2: Mean average precision of various IR models and query formulations (French & Brazilian/Portuguese languages)

	Mean average precision					
Query Model \ # of queries	Bulgarian T 50 queries	Bulgarian TD 50 queries	Bulgarian TDN 50 queries	Hungarian T 48 queries	Hungarian TD 48 queries	Hungarian TDN 48 queries
doc=Okapi, query=npn	**0.2661**	**0.2776**	**0.3079**	**0.2838**	0.3149	0.3256
DFR GL2	0.2578	0.2645	0.2964	0.2708	0.3043	0.3306
doc=Lnu, query=ltc	0.2413	0.2663	0.2848	0.2716	0.3107	0.3216
doc=dtu, query=dtn	0.2479	0.2527	0.2823	0.2817	**0.3224**	**0.3348**
doc=atn, query=ntc	0.2395	0.2520	0.2777	0.2810	0.3191	0.3320
doc=ltn, query=ntc	0.2289	0.2375	0.2492	0.2705	0.2956	0.3138
doc=ntc, query=ntc	0.1743	0.1898	0.2131	0.2297	0.2574	0.2677
doc=ltc, query=ltc	0.1788	0.1982	0.2271	0.2244	0.2654	0.2904
doc=lnc, query=ltc	0.1901	0.2089	0.2456	0.2154	0.2547	0.2836
doc=bnn, query=bnn	0.1258	0.1253	0.0662	0.1553	0.0942	0.0556
doc=nnn, query=nnn	0.0997	0.1042	0.1047	0.1346	0.1089	0.0900
Mean	0.2406	0.2161	0.2323	0.2381	0.2589	0.2678
Mean over top-8 models	0.2293	0.2423	0.2673	0.2642	0.2987	0.3146
% change over T		+5.67%	+16.56%		+13.07%	+19.07%

Table 3: Mean average precision of various IR models and query formulations (Bulgarian & Hungarian language, word-based indexing)

For this year, we tried to investigate more deeply the Hungarian language on the one hand and, on the other we implemented various IR models based on the Lucene (open source available at lucene.apache.org) search engine. In Table 4, we reported some experiments done with our new Lucene version using a word-based indexing scheme (TD queries) with a stemmer removing some derivational suffixes (see Section 3). In the third column, we automatically decompounded long words (composed by more than 8 characters) using our own algorithm (Savoy 2004b). In this experiment, both the compound words and their components were left in documents and queries. Finally, we have also reported the MAP achieved by the 4-gram approach (the best performance of last year, without stemming). The data depicted in Table 4 shows that, inside a given indexing strategy, all probabilistic approaches present similar level of performance. Compared to the word-based indexing strategy, the 4-gram indexing approach proposes, in average, an improvement of +6.91%, while the decompounded indexing method exposes the best performance (+8.08% better than the word-based approach).

-

Finally, some differences could appear between Table 3 (computed with our version of the SMART system) and evaluations computed with our Lucene system (Table 4).

	Mean average precision		
Query TD	Hungarian 48 queries word-based	Hungarian 48 queries decompounded	Hungarian 48 queries 4-gram
doc=Okapi, query=npn	0.3129	0.3392	**0.3496**
DFR GL2	0.3148	0.3396	0.3346
DFR PB2	**0.3233**	**0.3574**	0.3412
DFR PL2	0.3149	0.3399	0.3262
DFR I(F)L2	0.3157	0.3415	0.3441
LM (λ=0.3)	0.3142	0.3354	0.3329
LM (λ=0.35)	0.3132	0.3344	0.3330
Mean	0.3156	0.3411	0.3374
% change over word-based		+8.08%	+6.91%

Table 4: Mean average precision of best performing IR model (TD query formulation)

It was observed that pseudo-relevance feedback (PRF or blind-query expansion) seemed to be a useful technique for enhancing retrieval effectiveness. In this study, we adopted Rocchio's approach (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. To evaluate this proposition, we used four probabilistic models and enlarged the query by the 20 to 50 terms retrieved from the 10 best-ranked articles with the French corpus (Table 5) and Brazilian/Portuguese collection (Table 6).

	Mean average precision			
Query TD	French 49 queries	French 49 queries	French 49 queries	French 49 queries
IR Model / MAP	Okapi 0.4151	DFR PL2 0.4101	DFR GL2 0.3988	LM 0.3913
k doc. / *m* terms	10/20 0.4222	10/20 0.4255	10/20 0.4182	10/20 0.4388
	10/30 0.4269	10/30 0.4255	10/30 0.4282	10/30 0.4460
	10/40 **0.4296**	10/40 0.4307	10/40 0.4338	10/40 0.4508
	10/50 0.4261	10/50 **0.4311**	10/50 **0.4356**	10/50 **0.4509**

Table 5: Mean average precision using blind-query expansion (French collection)

	Mean average precision			
Query TD	Portuguese 50 queries	Portuguese 50 queries	Portuguese 50 queries	Portuguese 50 queries
IR Model / MAP	Okapi 0.4118	DFR PL2 0.4147	DFR GL2 0.4033	LM 0.3909
k doc. / *m* terms	10/20 0.4236	10/20 0.4412	10/20 **0.4141**	10/20 0.4273
	10/30 0.4361	10/30 **0.4414**	10/30 0.4129	10/30 **0.4286**
	10/40 0.4362	10/40 0.4386	10/40 0.4105	10/40 0.4266
	10/50 **0.4427**	10/50 0.4367	10/50 0.4101	10/50 0.4276

Table 6: Mean average precision using blind-query expansion (Brazilian/Portuguese collection)

For the French collection, the percentage of improvement varies from +3.5% (Okapi model, 0.4151 vs. 0.4296) to +15.2% (LM model, 0.3913 vs. 0.4509). For the Brazilian/Portuguese corpus, the enhancement raises from +2.7% (GL2 model, 0.4033 vs. 0.4141) to +9.6% (LM model, 0.3905 vs. 0.4286).

For the Bulgarian language, we have evaluated both a word-based indexing (top part of Table 7) and 4-gram indexing strategy (bottom part of Table 7). First it is interesting to note that both before and after blind query expansion, the word-based indexing approach provides always a better MAP than the corresponding 4-gram approach. When considering word-based indexing, the use of a blind query expansion improves the MAP from +9.2% (Okapi model, 0.2614 vs. 0.2854) to +21.7% (GL2 model, 0.2734 vs. 0.3327). With the 4-gram indexing approach, the enhancement goes from +12% (Okapi model, 0.2528 vs. 0.2831) to +33% (LM model, 0.2380 vs. 0.3165). Of course, the number of "terms" added with the 4-gram indexing scheme is usually greater than the corresponding run using a word-based indexing strategy.

-

	Mean average precision			
Query TD	Bulgarian 50 queries	Bulgarian 50 queries	Bulgarian 50 queries	Bulgarian 50 queries
IR Model word-based	Okapi 0.2614	DFR IFL2 0.2724	DFR GL2 0.2734	LM (λ=0.35) 0.2720
k doc. / *m* terms	3/50 0.2833 5/50 0.2798 10/50 0.2754 3/90 **0.2854** 5/90 0.2809 10/90 0.2789	3/10 0.2998 5/10 0.3223 10/10 **0.3323** 3/40 0.3163 5/40 0.3119 10/40 0.3171	3/10 0.2957 5/10 **0.3327** 10/10 0.3319 3/60 0.3225 5/60 0.3141 10/60 0.3161	3/20 0.3083 5/20 0.3167 10/20 **0.3305** 3/70 0.3228 5/70 0.3193 10/70 0.3208
IR Model 4-gram	Okapi 0.2528	DFR IFL2 0.2543	DFR GL2 0.2548	LM (λ=0.35) 0.2380
k doc. / *m* terms	3/50 0.2735 5/50 0.2699 10/50 0.2609 3/100 **0.2831** 5/100 0.2828 10/100 0.2768	3/40 0.2816 5/40 0.2925 10/40 **0.3050** 3/100 0.2926 5/100 0.2998 10/100 0.3009	3/60 0.2953 5/60 **0.3108** 10/60 0.3058 3/100 0.2924 5/100 0.3084 10/100 0.3018	3/70 0.3022 5/70 0.3141 10/70 0.2959 3/100 0.3019 5/100 **0.3165** 10/100 0.2953

Table 7: Mean average precision using blind-query expansion
(top part word-based indexing, bottom part 4-gram indexing, Bulgarian corpus)

	Mean average precision		
Query TD	English 49 queries	English 49 queries	English 49 queries
IR Model	Okapi 0.4006	DFR GL2 0.3999	LM (λ=0.35) 0.3862
k doc. / *m* terms	3/10 **0.4048** 5/10 0.4036 10/10 0.3989 3/15 0.3955 5/15 0.3876 10/15 0.3936	3/10 0.3965 5/10 0.3961 10/10 0.3923 3/40 0.4063 5/40 **0.4100** 10/40 0.3988	3/10 0.4059 5/10 0.4250 10/10 0.4067 3/15 0.3894 5/15 **0.4277** 10/15 0.4008

Table 8: Mean average precision using blind-query expansion
(English collection, TD query formulation)

4.3. Some Query-by-Query Analysis

In order to obtain some explanation of our failures, we decided to inspect the queries achieving a MAP below 0.1 for all IR models defined in Section 4.2. For the French collection, we found three such queries. The most difficult was Query #320 ("Energy Crises") for which the best MAP is 0.0368, obtained by DFR-GL2. In this case, terms in the query ("crises de l'énergie") cannot retrieved in a high position relevant documents where pertinent sentences included formulation like "crise énergétique" (stemming problem: "crise" or "crises" → "cris" but "énergie" → "energ" and "énergétique" → "energet") or "pénurie d'énergie" (synonyms, in the current context between "pénurie" and "crise").

With the French corpus, in the second position we found Query #336 ("NBA Labour Conflicts") for which the best MAP is 0.0667, obtained by LM model. In this case, the terms used in the query ("Labour Conflicts" or "conflits de travail") retrieved a lot of non-relevant items. The single relevant article for this query used mainly the term "strikes," "base-ball" and "ice-hockey" and the single word in common is "conflicts". Finally, we have Query #309 ("Hard Drugs") for which the best MAP is 0.0726, obtained by Okapi model with blind-query expansion (50 documents / 50 terms). In this case, the first relevant document appears in rank 20. This topic retrieves a relatively large number of documents (having more than one term in common with the query) describing however other non relevant aspects.

With the Brazilian/Portuguese corpus, we found four queries for which the best MAP is below 0.1. First we have Query #340 ("New Quebec Premier") for which the best MAP is 0.003, obtained by DFR-PL2 (with blind-query expansion; 10 documents / 50 terms). The second most difficult topic was Query #320 ("Energy Crises")

for which the best MAP is 0.0066, achieved by the language model. In the third position, we have Query #327 ("Earthquakes in Mexico City") for which the best MAP is 0.0372 (DFR-GL2 & blind-query expansion, 10 documents / 40 terms). Finally, we have Query #344 ("Brazil vs. Sweden World Cup Semifinals") with a best MAP of 0.0434 obtained by DFR-PL" model (with blind-query expansion, 10 documents / 20 terms).

With the Hungarian collection based on a 4-gram indexing scheme, we found six queries for which the best MAP is below 0.1 (Query #371, #357, #374, #317, #362, and #320). With our decompounding approach, the "difficult" queries formed a subset of the previous one (namely Query #357, #320, #374, and #371). If we consider also word-based indexing strategy, we can find a set of seven hard topics that includes Query #357, #374, and #371. These three topics are the most difficult queries for the Hungarian corpus, independently of the underlying indexing scheme.

5 Data Fusion

It is assumed that combining different search models should improve retrieval effectiveness, due to the fact that different document representations might retrieve different pertinent items and thus increase the overall recall (Vogt & Cottrell 1999). In this current study we combine two or three probabilistic models representing both the parametric (Okapi and DFR) and non-parametric (language model) approaches. To achieve this we evaluated various fusion operators (see Table 9 for a list of their precise descriptions). For example, the Sum RSV operator indicates that the combined document score (or the final retrieval status value) is simply the sum of the retrieval status value (RSV_k) of the corresponding document D_k computed by each single indexing scheme (Fox & Shaw 1994). Table 9 thus illustrates how both the Norm Max and Norm RSV apply a normalization procedure when combining document scores. When combining the retrieval status value (RSV_k) for various indexing schemes and in order to favor some more efficient retrieval schemes, we could multiply the document score by a constant α_i (usually equal to 1) reflecting the differences in retrieval performance.

Sum RSV	$SUM\ (\alpha_i \cdot RSV_k)$
Norm Max	$SUM\ (\alpha_i \cdot (RSV_k / Max^i))$
Norm RSV	$SUM\ [\alpha_i \cdot ((RSV_k - Min^i) / (Max^i - Min^i))]$
Z-Score	$\alpha_i\ .\ [((RSV_k - Mean^i) / Stdev^i) + \delta^i]$ with $\delta^i = [(Mean^i - Min^i) / Stdev^i]$

Table 9: Data fusion combination operators used in this study

In addition to using these data fusion operators, we also considered the round-robin approach, wherein we took one document in turn from all individual lists and removed any duplicates, retaining the most highly ranked instance. Finally we suggested merging the retrieved documents according to the Z-Score, computed for each result list. Within this scheme, for the *i*th result list, we needed to compute the average RSV_k value (denoted $Mean^i$ and the standard deviation (denoted $Stdev^i$). Based on these we could then normalize the retrieval status value for each document D_k provided by the *i*th result list by computing the deviation of RSV_k with respect to the mean ($Mean^i$). In Table 9, Min^i (Max^i) denotes the minimal (maximal) RSV value in the *i*th result list. Of course, we might also weight the relative contribution of each retrieval scheme by assigning a different α_i value to each retrieval model.

	Mean average precision (% of change)			
Query TD Model	French 49 queries	Portuguese 50 queries	Bulgarian 50 queries	Hungarian 48 queries
LM & PRF doc/term Okapi & PRF doc/term DFR & PRF doc/term	10/30 0.4460 10/60 0.4275	10/50 0.4276 10/80 0.4403	3/40 0.3201 4-gram GL2 10/90 0.2941	3/70 0.3815 4-gram Okapi 3/100 0.3870 PB2 5/40 0.3794
Round-robin	0.4480 (+0.45%)	0.4489 (+1.95%)	0.3104 (-3.03%)	0.4031 (+4.16%)
Sum RSV	0.4455 (-0.11%)	0.4556 (+3.47%)	0.3307 (+3.31%)	0.4216 (+8.94%)
Norm Max	0.4553 (+2.09%)	0.4556 (+3.47%)	0.3290 (+2.78%)	0.4246 (+9.72%)
Norm RSV	**0.4559** (+2.22%)	**0.4566** (+3.70%)	0.3298 (+3.03%)	0.4245 (+9.69%)
Z-Score	**0.4559** (+2.22%)	0.4560 (+3.57%)	**0.3326** (+3.91%)	**0.4308** (+11.32%)
Z-ScoreW	0.4553 (+2.09%)	0.4553 (+3.41%)	0.3314 (+3.53%)	0.4252 (+9.87%)

Table 10: Mean average precision using different combination operators (with blind-query expansion)

Table 10 depicts the evaluation of various data fusion operators, comparing them to the single approach using the language model (LM), Okapi or DFR probabilistic models. From this data, we can see that combining two or three IR models might improve retrieval effectiveness, slightly for the French collection, moderately for the Portuguese and clearly for the Hungarian corpus. When combining different retrieval models, the Z-Score scheme tended to perform the best. Compared to the best single search model, the performance achieved by the various data fusion approaches seems not to be statistically significant, except for the Hungarian corpus.

6 Official Results

Table 11 shows the exact specifications of our 12 official monolingual runs. These experiments were mainly based on the probabilistic models (Okapi, DFR and language model (LM)). All runs are fully automatic using the TD query formulation and using often a data fusion approach (often based on the Z-Score operator).

Run name	Language	Query	Index	Model	Query expansion	Single MAP	Comb MAP
UniNEfr1	French	TD	word	PL2	10 docs / 40 terms	0.4307	Z-ScoreW
		TD	word	LM	10 docs / 30 terms	0.4460	
		TD	word	Okapi	10 docs / 40 terms	0.4193	0.4549
UniNEfr2	French	TD	word	GL2	10 docs / 40 terms	0.4338	Z-ScoreW
		TD	word	Okapi	10 docs / 20 terms	0.4222	0.4430
UniNEfr3	French	TD	word	Okapi	10 docs / 60 terms	0.4275	Norm RSV
		TD	word	LM	10 docs / 30 terms	0.4460	**0.4559**
UniNEpt1	Portuguese	TD	word	LM	10 docs / 50 terms	0.4276	Z-ScoreW
		TD	word	Okapi	10 docs / 80 terms	0.4403	**0.4552**
UniNEpt2	Portuguese	TD	word	LM	10 docs / 40 terms	0.4266	Z-Score
		TD	word	GL2	10 docs / 40 terms	0.4105	
		TD	word	Okapi	10 docs / 30 terms	0.4361	0.4461
UniNEpt3	Portuguese	TD	word	LM	10 docs / 100 terms	0.4302	Z-ScoreW
		TD	word	Okapi	10 docs / 30 terms	0.4361	0.4495
UniNEbg1	Bulgarian	TD	4-gram	IFL2	5 docs / 50 terms	0.2924	Norm RSV
		TD	word	LM	3 docs / 70 terms	0.3300	
		TD	4-gram	Okapi	3 docs / 100 terms	0.2943	0.3129
UniNEbg2	Bulgarian	TD	word	LM	5 docs / 40 terms	0.3201	Z-ScoreW
		TTTD	4-gram	GL2	10 docs / 90 terms	0.2941	**0.3314**
UniNEbg3	Bulgarian	TTTD	4-gram	IFL2	5 docs / 50 terms	0.2924	Norm RSV
		TTTD	word	LM	3 docs / 70 terms	0.3300	
		TD	4-gram	LM	3 docs / 100 terms	0.2959	
		TD	word	Okapi	5 docs / 80 terms	0.3229	0.3298
UniNEhu1	Hungarian	TD	word	PB2	5 docs / 20 terms	0.3922	Norm RSV
		TD	4-gram	Okapi	3 docs / 90 terms	0.3927	0.4186
UniNEhu2	Hungarian	TTTD	wordDec	PL2	3 docs / 40 terms	0.3794	Z-Score
		TTTD	word	LM	3 docs / 70 terms	0.3815	
		TD	4-gram	Okapi	3 docs / 100 terms	0.3870	**0.4308**
UniNEhu3	Hungarian	TD	word	PB2	5 best docs / 20 terms	0.3922	0.3922

Table 11: Description and mean average precision (MAP) of our official monolingual runs

Run name	Language	Query	Index	Model	Query expansion	Single MAP	Comb MAP
UniNEtden	English	TD	word	GL2	3 docs / 10 terms	0.3965	Z-Score
		TD	word	Okapi	3 docs / 10 terms	0.4048	
		TD	word	LM	5 docs / 10 terms	0.4250	0.4367
UniNEtdnen	English	TDN	word	GL2	5 docs / 15 terms	0.4195	Sum RSV
		TDN	word	Okapi	3 docs / 10 terms	0.4273	
		TDN	word	LM	3 docs / 10 terms	0.4352	0.4444

Table 12: Description and mean average precision (MAP) of our unofficial monolingual runs for the English collection

Table 12 shows the exact specifications of two unofficial monolingual runs submitted to improve the pool for the English monolingual collection. These experiments are based on a combination of three probabilistic models (Okapi, DFR-GL2 and LM).

7 Bilingual Information Retrieval

Due to time constraint, we have limited our participation in the bilingual track to the French and Portuguese language. Moreover, we chose English as the language for submitting queries to be automatically translated into these two different languages, using ten different freely available machine translation (MT) systems, namely:

ALPHAWORKS	www.alphaWorks.ibm.com/
APPLIEDLANGUAGE	www.appliedLanguage.com/
BABELFISH	babelFish.altavista.com/
FREETRANSLATION	www.freetranslation.com/web.htm
GOOGLE	www.google.com/language_tools
INTERTRAN	www.tranexp.com/
ONLINE	www.online-translator.com/
REVERSO	webtranslation.paralink.com/
SYSTRAN	www.systranlinks.com/
WORLDLINGO	www.worldlingo.com/

Table 13 shows the mean average precision obtained using the various MT tools and the Okapi probabilistic model with blind query expansion (40 terms extracted from the first 10 retrieved items). Of course, all tools are not always available for each language and thus various entries are missing (as shown in Table 13, indicated by the label "N/A").

	Mean average precision (% of monolingual)	
Language Okapi (TD queries)	French 49 queries	Portuguese 50 queries
Manual & PRF (10/40)	0.4296	0.4389
AlphaWorks	0.3378 (78.6%)	N / A
AppliedLanguage	0.3726 (86.7%)	0.3077 (70.1%)
BabelFish (altavista)	0.3771 (87.8%)	0.3092 (70.4%)
FreeTranslation	0.3813 (88.8%)	0.3356 (76.5%)
Google	0.3754 (87.4%)	0.3070 (69.9%)
InterTrans	0.2761 (64.3%)	0.3343 (76.2%)
Online	0.3942 (91.8%)	**0.3677** (83.8%)
Reverso / Promt	**0.4081** (95.0%)	0.3531 (80.5%)
WorldLingo	0.3832 (89.2%)	0.3091 (70.4%)
Systran	N / A	0.3077 (70.1%)
WorldLingo	0.3832 (70.4%)	0.3091 (70.4%)

Table 13: Mean average precision of various machine translation systems (Okapi model with blind query expansion, TD queries)

From this data, we can see that for the French collection the best translation is obtained by Reverso (95% of the performance level achieved by a monolingual search) and for the Portuguese corpus by Online (84% of the performance level achieved by the corresponding monolingual search). From a more general point of view, both Reverso (Promt) and Online MT systems obtain satisfactory retrieval performances for both languages. For the French language, both the FreeTranslation and BabelFish present also an overall good performance. Starting with queries written in English, data depicted in Table 13 indicates also that the automatic translation process performs better with French as target language than with Portuguese.

Table 14 shows the retrieval effectiveness for various query translation combinations (concatenation of the translations produced by two or more MT systems) when using the Okapi probabilistic model with blind query expansion (40 terms extracted from the ten-best retrieved items). The top part of the table indicates the exact query translation combination used while the bottom part shows the MAP obtained with our combined query translation approach. The resulting retrieval performances depicted in Table 14 are never better than the best single translation scheme (row labeled "Best single") for the French language and usually slightly better for the Portuguese language.

	Mean average precision (% of change)	
Language Combination	French Okapi 49 queries	Portuguese Okapi 50 queries
Comb 1	BabelFish & Reverso	Promt & Free
Comb 2	BabelFish & Lingo	Promt& Free & Inter
Comb 3	Reverso & Online	Promt& Free & Online
Comb 4	BabelFish & Google	Promt & Online
Comb 5	BabelFish & Google & Free	
Best single	**0.4081**	0.3677
Comb 1 Comb 2 Comb 3 Comb 4 Comb 5	0.3925 (-3.82%) 0.3665 (-10.19%) 0.3971 (-2.70%) 0.3633 (-10.98%) 0.3682 (-9.78%)	**0.3815** (+3.75%) 0.3786 (+2.96%) 0.3741 (+1.74%) 0.3516 (-4.38%)

Table 14: MAP of various combined translation tools
(Okapi model with blind query expansion, TD queries)

From English to …	French 49 queries	Portuguese 50 queries
IR 1 (#docs/#terms)	PL2 (10/30)	I(n)L2 (10/40)
IR 2 (#docs/#terms)	LM (10/30)	LM (10/30)
Data fusion operator	Z-score	Round-robin
Translation tools	BabelFish & Reverso	Promt & Free & Online
MAP	**0.4278**	0.4114
Run name	UniNEBifr1	UniNEBipt2
IR 1 (#docs/#terms)	PL2 (10/30)	GL2 (10/40)
IR 2 (#docs/#terms)	Okapi (10/60)	Okapi (10/80)
IR 3 (#docs/#terms)	LM (10/50)	LM (10/40)
Data fusion operator	Z-scoreW	Z-Score
Translation tools	Reverso & Online	Promt & Free
MAP	0.4256	**0.4138**
Run name	UniNEBifr2	UniNEBipt1
IR 1 (#docs/#terms)	PL2 (10/30)	I(n)L2 (10/40)
IR 2 (#docs/#terms)	Okapi (10/60)	LM (10/30)
IR 3 (#docs/#terms)	LM (10/30)	
Data fusion operator	Z-score	Norm RSV
Translation tools	BabelFish & Google & Free	Promt & Online
MAP	0.4083	0.4062
Run name	UniNEBifr3	UniNEBipt3

Table 15: Description and MAP of our official bilingual runs

Finally, Table 15 lists the parameter settings used for our 6 official runs in the bilingual task. Each experiment uses queries written in English to retrieve documents in the other target languages. Before combining the result lists using a data fusion operator (see Section 5), we automatically expanded the translated queries using a pseudo-relevance feedback method (Rocchio's approach in the present case).

8 Monolingual Domain-Specific Retrieval: GIRT

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, see (Kluck 2004) for a more complete description of these corpora. A few statistics on these collections are given in Table 16. The English corpus is a manually translation of the German documents, the whole document is however not always fully translated.

A typical record in this collection is composed of a title, an abstract, and a set of manually assigned keyword. Additional information such as authors' name, publication date, or the language in which the bibliographic notice is written may of course be less important from an IR perspective but they are made available. As

-

depicted in the Appendix, the topics in this domain-specific collection cover a variety of themes (e.g., "Poverty", "Role of the father", "The computer in the everyday", or "Modernizing of public administration").

	German	English
Size (in MB)	326 MB	199 MB
# of documents	151,319	151,319
# of distinct terms	698,638	151,181
Number of distinct indexing terms / document		
Mean	70.83	107.9
Standard deviation	32.4	94.59
Median	68	77
Maximum	386	1,422
Minimum	2	2
Number of indexing terms / document		
Mean	89.61	142.1
Standard deviation	44.5	139.84
Median	84	95
Maximum	629	4,984
Minimum	4	2
Number of queries	25	25
Number rel. items	3,759	4,239
Mean rel./ request	150.36	169.56
Standard deviation	100.707	110.167
Median	144	142
Maximum	372 (Q#166)	381 (Q#161)
Minimum	11 (Q#169)	22 (Q#170)

Table 16: CLEF 2005 GIRT test collection statistics

Table 17 shows the MAP of various query formulations for the German and English. The best retrival models are usually the Okapi probabilistic model or the DFR-GL2. The language model (LM) achieves also a very good retrieval effectiveness with these test-collections. If we see a clear improvement from T query formulation to TD, the performance difference between TD and TDN query formulation is relatively small for both languages.

In order to improve the search performance, we have considered a pseudo-relevance feedback using the Rocchio's formulation (see Table 18 for the German corpus, and Table 19 for the English collection). Such query expansion clearly improves the mean average precision.

	Mean average precision				
Query TD Model \ # of queries	German T 25 queries	German TD 25 queries	German TDN 25 queries	English TD 25 queries	English TDN 25 queries
DFR GL2	0.3955	0.4339	0.4451	0.3532	**0.3724**
LM (λ=0.35)		0.4546		**0.3541**	0.3712
Okapi	**0.4236**	**0.4565**	**0.4509**	0.3516	0.3654
doc=Lnu, query=ltc	0.3745	0.3289	0.4079	0.3251	0.3324
doc=dtu, query=dtn	0.3754	0.3850	0.3392	0.3030	0.2793
doc=atn, query=ntc	0.3530	0.3756	0.3862	0.2970	0.3086
doc=ltn, query=ntc	0.3519	0.3760	0.3801	0.2555	0.2514
doc=ntc, query=ntc	0.2478	0.2665	0.2679	0.2126	0.2185
doc=ltc, query=ltc	0.2649	0.2329	0.3067	0.2148	0.2334
doc=lnc, query=ltc	0.2919	0.2626	0.3531	0.2554	0.2807
doc=bnn, query=bnn	0.2633	0.1209	0.0919	0.1054	0.0396
doc=nnn, query=nnn	0.1401	0.1317	0.1285	0.0819	0.0733
Mean	0.3165	0.3678	0.3730	0.2963	0.3036
Mean (top-9 best models)	0.3483	0.3678	0.3730	0.2963	0.3036
% change over T queries		+5.58%	+7.08%		

Table 17: Mean average precision of various single searching strategies (monolingual, GIRT corpus)

-

Our 6 official runs in the monolingual GIRT task are described in Table 20. Each run is built using a data fusion operator ("Z-ScoreW" in this case, see Section 5). For all runs, we automatically expanded the queries using a blind relevance feedback method (Rocchio in our experiments).

	Mean average precision		
Query TD	German 25 queries	German 25 queries	German 25 queries
IR Model / MAP	Okapi 0.4565	DFR GL2 0.4339	LM 0.4546
k doc. / *m* terms	3/10 0.4805 5/10 0.4878 10/10 0.4913 3/30 0.4837 5/30 0.4878 10/30 **0.5011**	3/30 0.4395 5/30 0.4601 10/30 0.4543 3/50 0.4377 5/50 **0.4628** 10/50 0.4617	3/10 0.4828 5/10 0.4878 10/10 0.4913 3/30 0.4837 5/30 0.4878 10/30 **0.5011**

Table 18: Mean average precision using blind-query expansion (German GIRT collection)

	Mean average precision		
Query TD	English 25 queries	English 25 queries	English 25 queries
IR Model / MAP	Okapi 0.3516	DFR GL2 0.3532	LM 0.3541
k doc. / *m* terms	3/30 0.3781 5/30 0.3910 10/30 0.3952 3/50 0.3757 5/50 0.3921 10/50 **0.3955**	3/50 0.3845 5/50 0.4073 10/50 0.4065 3/100 0.3884 5/100 **0.4144** 10/100 0.4106	3/10 0.3436 5/10 0.3541 10/10 0.4032 3/40 0.2948 5/40 0.2991 10/40 **0.4203**

Table 19: Mean average precision using blind-query expansion (English GIRT collection)

Run name	Language	Query	Index	Model	Query expansion	Single MAP	Comb. MAP
UniNEde1	German	TD	word	Okapi	10 best docs / 15 terms	0.4959	Z-ScoreW
		TD	word	DFR GL2	10 best docs / 100 terms	0.4677	0.5015
UniNEde2	German	TD	word	Okapi	5 best docs / 10 terms	0.4878	Z-ScoreW
		TD	word	DFR GL2	5 best docs / 10 terms	0.4420	
		TD	word	LM	10 best docs / 30 terms	0.5011	**0.5051**
UniNEde3	German	TDN	word	Okapi	5 best docs / 10 terms	0.4851	Z-ScoreW
		TDN	word	DFR GL2	5 best docs / 10 terms	0.4541	
		TDN	word	LM	10 best docs / 30 terms	0.4832	0.5159
UniNEen1	English	TD	word	LM	10 best docs / 20 terms	0.4160	Round-robin
		TD	word	DFR GL2	10 best docs / 150 terms	0.4113	0.4292
UniNEen2	English	TD	word	Okapi	10 best docs / 30 terms	0.3952	Z-scoreW
		TD	word	DFR GL2	10 best docs / 50 terms	0.4065	
		TD	word	LM	10 best docs / 10 terms	0.4032	**0.4303**
UniNEen3	English	TDN	word	Okapi	10 best docs / 10 terms	0.3981	ZscoreW
		TDN	word	DFR GL2	10 best docs / 50 terms	0.4410	
		TDN	word	LM	10 best docs / 20 terms	0.4291	0.4576

Table 20: Description and mean average precision (MAP) of our official GIRT runs

9 Robust Retrieval Track

The aim of this track consists to analyze and to improve IR systems when facing with "difficult" topics (Voorhees 2004). In the current context, difficult topics means queries having a poor mean average precision in previous evaluation campaigns. We also knew that topic difficulty depend on the underling collection (Voorhees 2005). The goal of the robust track is therefore to explore how one can build a search system that can perform "reasonably well" for all queries. This question is a main concern when evaluating real systems with users facing with unexpected search results or "stupid" answers returned by a search engine.

-

It is known that determining *a priori* whether a given topic is difficult of not, seems to be impossible (Voorhees 2004). Therefore, the queries created and evaluated during the CLEF-2001 (Query #41 - #90), CLEF-2002 (Query #91 - #140) and CLEF-2003 (Query #141 - #200) evaluation campaigns have been reused against mainly the same document collection. Moreover, the organizers divided arbitrary this query set into a training set (60 queries) and a test set (100 queries). In fact, in this latter set, 9 queries do not have any relevant items and thus the test set is formed by only 91 remaining queries. When analyzing this sample, we found that the mean number of relevant items per query was 24.066 (median: 14, min: 1, max: 177, standard deviation: 30.78).

When evaluating an IR system with previously created queries, we need to search into the same documents collection. With the French language, the CLEF-2001 and CLEF-2002 campaigns have used the newspaper *Le Monde* (1994), and *Schweizerische Depeschenagentur* (SDA, 1994) to generate a collection composed of 87,191 documents. During the CLEF-2003 campaign, 42,615 documents extracted from the SDA during the year 1995 have been added (the size of the final corpus is therefore of 129,806 articles). Thus for Query #41 to #140 (corresponding to 59 queries in the test set) we do not have any judgments against SDA 95 (and retrieved items not judged are assumed non relevant). If we remove all references to SDA95 in the result list, the MAP will change and increase. For example, with the Okapi model and TQ queries, the MAP is 0.4816 and after removing the SDA.95 references, the MAP increases to 0.5322 (+10.5%).

When using the MAP to measure the retrieval effectiveness, all observations (queries) have the same importance or weight. It is known that the average measure may hide irregularities among the observations (but owns, of course, the advantage to resume a large number of observations with a single number). Thus incorrect answers provided by the search engine are not really penalized by the arithmetic mean. Attaching the same importance to all topics has the following problem. If, for example, a search system improves its performance from 0.5 to 0.6 for one query, this enhancement is viewed as similar to a system improving its performance for a difficult query from 0.02 to 0.12. Replacing the arithmetic mean by the geometric mean, the second improvement in our example will have a greater impact than the first. Of course, other evaluation measure could be considered such as the median (Savoy 1997).

As a first experiment, we want to verify the retrieval effectiveness (both the mean average precision (or MAP) and the geometric mean (GMAP)) using short topic formulation (T), medium (TD) or long topic description (TDN). The results are depicted in Table 21. In the third line of this table, we have indicated the mean number of distinct terms in the three different query formulations (ranging from a mean value of 2.91 different terms for title-only formulation, to 16.32 for the TDN query formulation). In the last two lines we have indicated the overall mean and the mean when considering only the first 7 IR models (ending with the line labeled "doc=ltn, query=ntc"). The MAP presents always a higher value than the geometric mean but both evaluations are strongly correlated (r=0.9624). As a single IR system, the Okapi probabilistic model exposes always the best performance (either measured by the MAP or the GMAP).

Query	T	T	TD	TD	TDN	TDN
	MAP	GMAP	MAP	GMAP	MAP	GMAP
mean distinct terms/query	2.91	2.91	7.51	7.51	16.32	16.32
Model \ # of queries	91 queries	91 queries	91 queries	91 queries	91 queries	91 queries
doc=Okapi, query=npn	**0.3969**	**0.2121**	0.4816	**0.3534**	**0.5151**	**0.4146**
DFR GL2	0.3742	0.1833	0.4714	0.3316	0.5088	0.3961
LM (λ=0.35)	0.3611	0.1745	0.4535	0.3079	0.5003	0.3809
doc=Lnu, query=ltc	0.3669	0.1941	0.4518	0.3291	0.4958	0.3927
doc=dtu, query=dtn	0.3765	0.1735	0.4406	0.3015	0.4909	0.3662
doc=atn, query=ntc	0.3869	0.1952	0.4459	0.3143	0.5001	0.3855
doc=ltn, query=ntc	0.3705	0.1957	0.4328	0.3056	0.4636	0.3510
doc=ntc, query=ntc	0.2447	0.0944	0.2988	0.1606	0.3262	0.1901
doc=ltc, query=ltc	0.2540	0.0916	0.3193	0.1769	0.3581	0.2212
doc=lnc, query=ltc	0.2604	0.0964	0.3364	0.1917	0.3949	0.2513
doc=nnn, query=nnn	0.1572	0.0484	0.1379	0.0499	0.1465	0.0563
Mean	0.3227	0.1508	0.3882	0.2566	0.4273	0.3096
Mean over 7-best models	0.3761	0.1898	0.4539	0.3205	0.4964	0.3839

Table 21: Comparing the mean average precision (MAP) with the geometric mean (GMAP) with various query formulations and search models (French corpus)

For our investigations, the short query formulation (T) will represent the starting point. With this short query formulation, the three most difficult queries were Query#200 ("Innondationeurs en Hollande et en Allemagne") with the best performance for a single IR system is 0.0002 (Lnu-ltc), Query #91 ("AI en Amérique

latine", best MAP: 0.0012, Lnu-ltc), and Query #48 ("Forces de maintien de la paix en Bosnie", best MAP: 0.0077, ltn-ntc). We could note that the term "Innondationeurs" in not a French word and the right term must be "Innondations" that appears in the descriptive part of Query #200.

If the user may introduce more search terms (comparing T with TD performance), the overall IR performance measure by the geometric mean increases from 0.1898 to 0.3205 (or +69%). With this medium-size query description, the three most difficult queries are Query #48 (best MAP: 0.028, Lnu-ltc), Query #61 ("Catastrophe petrolière en Sibérie", best MAP: 0.0293, Lnu-ltc), and Query #148 ("Dommages à la couche d'ozone", best MAP: 0.0507, Okapi). It is a surprise to still encounter Query #48 in the top three most difficult queries. The descriptive part of this query adds new and related terms ("Nations Unies", "Kosovo") without providing a positive impact of the performance. The most difficult topic with T query (Query #200) is now in the 6th rank (MAP: 0.085). The inclusion of the correct term "Innondations" improves the search process but the performance is still relatively low. For Query #91 appearing in the second position with T queries, it now occurs in rank 7th (MAP: 0.0954).

When comparing short (T) with the longest query formulation (TDN), the IR performance measured by the geometric mean doubles (from 0.1898 to 0.3839, or +102%). With TDN formulation, the most difficult topics are Query #48 (best MAP: 0.048, atn-ntc), Query #148 (best MAP: 0.0507, Okapi), and Query #90 ("Les exportateurs de légumes", best MAP: 0.113, Okapi). The most difficult topic using T formulation (Query #200) appears now in rank 19th with a MAP of 0.353 while Query #91 (in the second position with short formulation) occurs in rank 6th (MAP: 0.1343).

As another way to improve the performance, we may employ a pseudo-relevance feedback procedure (Rocchio in this case). The performance indicated in Table 22 shows that the overall performance improves after adding a relatively small number of new terms (e.g., 15) extracted from the best 5 retrieved items.

Query Model \ # of queries	T MAP 91 queries	T GMAP 91 queries	TD MAP 91 queries	TD GMAP 91 queries	TDN MAP 91 queries	TDN GMAP 91 queries
doc=Okapi, query=npn	0.3969	0.2121	0.4816	0.3534	0.5151	0.4146
docs / terms 3 / 10	0.4058	0.2186	0.4936	0.3676	0.5226	0.4184
3 / 15	0.4014	0.2152	0.4993	0.3720	0.5224	0.4179
3 / 20	0.3981	0.2128	0.4974	0.3716	0.5220	0.4160
5 / 10	0.4123	0.2318	0.5015	0.3723	0.5348	0.4306
5 / 15	**0.4141**	**0.2324**	**0.5035**	**0.3755**	**0.5362**	**0.4319**
5 / 20	0.4120	0.2301	0.5011	0.3733	0.5353	0.4282
10 / 10	0.3945	0.1991	0.5015	0.3728	0.5305	0.4286
10 / 15	0.3984	0.2021	0.4953	0.3684	0.5284	0.4261
10 / 20	0.4059	0.2121	0.4893	0.3641	0.5255	0.4220

Table 22: MAP and geometric mean (GMAP) when applying pseudo-relevance feedback approach (Rocchio)

We may also apply a data fusion approach as described in Section 5. Such a procedure has been applied to form two of our official runs, namely UniNEfrr1 with TD query, and UniNEfrr2 with T query (complete description given in Table 23).

Run name	Index	Query	Model	Query expansion	MAP	comb. MAP	GMAP
UniNEfrr1	word	TD	Okapi	5 best docs / 15 terms	0.5035	Round-robin	
	word	TD	GL2	3 best docs / 30 terms	0.5014		
	word	TD	LM	10 best docs / 15 terms	0.5095	0.5227	**0.3889**
UniNEfrr2	word	T	Okapi	3 best docs / 10 terms	0.4058	Z-ScoreW	
	word	T	GL2	5 best docs / 30 terms	0.4029		
	word	T	LM	5 best docs / 10 terms	0.4137	0.4396	**0.2376**
UniNEfrr3	word	TD	GL2	3 best docs / 30 terms & Yahoo!.fr	0.4607		0.2935

Table 23: Description and evaluations (MAP and GMAP) of our official robust runs (91 queries)

With the last run (UniNEfrr3), we have exploited the Web, or more precisely the Yahoo! search engine. We have sent to this Web search engine the title of the queries. As an answer, we obtained a page with ten references with, for each, a short description. We have extracted these ten short textual descriptions and added them to the original query. The expanded query has been sent to our search model in order to hopefully obtain a better result list. When using TD query formulation, the mean number of distinct search terms was 7.51. When

including the first ten references retrieved by Yahoo.fr, this average value increased to 115.46 (meaning that we have added, in mean, 108 new search terms). Such a massive query expansion was not effective (see results depicted in Table 23) and we need to include a term selection procedure to hopefully improve the geometric mean.

In fact our first intent was to use a French or Swiss newspaper Web site to find related terms. We think that we need to use, of course the same language (French in this case, but differences in meaning exist between the French expressions in Montreal and in Geneva), but also to have similar cultural and regional coverage (e.g, news from Switzerland differ from those in Canada) together with comparable thematic coverage (e.g., a general vs. a business-oriented newspaper) and comparable writing style (e.g., all newspapers do not have the same clients like "The Sun" and "The Times" in England). However, the time difference could also be problematic (the searched corpus is composed of articles written during the year 1994-95).

10 Conclusion

In this seventh CLEF evaluation campaign, we proposed a more effective IR model for the Hungarian language. We have considered a more aggressive stemmer that tries to remove some frequent derivational suffixes for this language. We also investigated the relative merit of the word-based and 4-gram indexing scheme. We have also evaluated an automatic decompouding scheme for the Hungarian language. Combining different indexing and retrieval schemes for this language seems to be really effective but requires more processing time and disk space.

For the French, Brazilian/Portuguese and Bulgarian language, we used the same stopword lists and stemmers developed during the previous years. In order to enhance retrieval performance, we have implemented an IR model based on the language model and have suggested a data fusion approach based on the Z-Score after applying a blind query expansion. Such general search strategy seems also effective for the GIRT corpora (German and English).

In the bilingual task, the freely available translation tools performed at a reasonable level for both the French and Portuguese languages (based on the best translation tool, the MAP compared to the monolingual search is around 95% for the French language and 83% for the Brazilian/Portuguese). Finally, in the robust retrieval task, we investigated the reasons and means to improve the retrieval effectiveness when facing with difficult topics.

Acknowledgments

The authors would like to also thank the CLEF-2006 task organizers for their efforts in developing various European language test-collections. The authors would also like to thank C. Buckley from SabIR for giving us the opportunity to use the SMART system, together with Pierre-Yves Berger for his help in translating the English topics and in using the Yahoo.fr search engine. This research was supported in part by the Swiss National Science Foundation under Grant #200020-103420.

References

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.

Fox, E.A. & Shaw, J.A. (1994). Combination of multiple searches. In *Proceedings TREC-2*, Gaithersburg: NIST Publication #500-215, 243-249.

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.

McNamee, P. (2005). Exploring new languages with HAIRCUT at CLEF 2005. In Working Notes, CLEF 2005, Vienna.

-

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. (1997). Statistical inference in retrieval effectiveness evaluation. *Information Processing & Management*, 33(4), 495-512.

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. (2005a). Comparative study of monolingual and multilingual search models for use with Asian languages. *ACM Transactions on Asian Languages Information Processing*, 4(2), 163-189.

Savoy, J. (2005c). Data fusion for effective European monolingual information retrieval. In Peters, P.D. Clough, G.J.F. Jones, J. Gonzalo, M. Kluck & B. Magnini (Eds.), *Multilingual Information Access for Text, Speech and Images*. LNCS #3491. Springer-Verlag, Berlin, 2005, 233-244.

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.

Vogt, C.C. & Cottrell, G.W. (1999). Fusion via a linear combination of scores. *IR Journal*, 1(3), 151-173.

Voorhees, E.M.. (2004). Overview of the TREC 2004 robust retrieval track. In *Proceedings TREC-2004*. Gaithersburg: NIST Publication #500-261, 70-79.

Voorhees, E.M.. (2006). The TREC 2005 robust track. ACM-SIGIR Forum, 40(1), 41-48.

Appendix: Weighting Schemes

To assign an indexing weight w_{ij} that reflects the importance of each single-term t_j in a document D_i, we might use the various approaches shown in Table A.1, where *n* indicates the number of documents in the collection, *t* the number of indexing terms, df_j the number of documents in which the term t_j appears, the document length (the number of indexing terms) of D_i is denoted by nt_i, and *avdl*, *b*, k_1, *pivot* and *slope* are constants. For the Okapi weighting scheme, *K* represents the ratio between the length of D_i measured by l_i (sum of tf_{ij}) and the collection mean noted by *avdl*.

bnn	$w_{ij} = 1$	nnn	$w_{ij} = tf_{ij}$
ltn	$w_{ij} = (\ln(tf_{ij}) + 1) \cdot idf_j$	atn	$w_{ij} = idf_j \cdot [0.5 + 0.5 \cdot tf_{ij} / \max tf_i]$
dtn	$w_{ij} = [\ln(\ln(tf_{ij}) + 1) + 1] \cdot idf_j$	npn	$w_{ij} = tf_{ij} \cdot \ln[(n-df_j) / df_j]$
Okapi	$w_{ij} = \left((k_1+1) \cdot tf_{ij}\right) \Big/ \left(K + tf_{ij}\right)$	Lnu	$w_{ij} = \dfrac{\left(1+\ln(tf_{ij})\right) \Big/ \left(\ln(\text{mean tf})+1\right)}{(1-\text{slope}) \cdot \text{pivot} + \text{slope} \cdot nt_i}$
lnc	$w_{ij} = \dfrac{\ln(tf_{ij})+1}{\sqrt{\sum_{k=1}^{t} (\ln(tf_{ik})+1)^2}}$	ntc	$w_{ij} = \dfrac{tf_{ij} \cdot idf_j}{\sqrt{\sum_{k=1}^{t} (tf_{ik} \cdot idf_k)^2}}$
ltc	$w_{ij} = \dfrac{\left(\ln(tf_{ij})+1\right) idf_j}{\sqrt{\sum_{k=1}^{t} \left((\ln(tf_{ik})+1) \cdot idf_k\right)^2}}$		
dtu	$w_{ij} = \dfrac{\left(\ln\left(\ln(tf_{ij})+1\right)+1\right) \cdot idf_j}{(1-\text{slope}) \cdot \text{pivot} + \text{slope} \cdot nt_i}$		

Table A.1: Weighting schemes

	Okapi			DFR	
Language	*b*	k_1	*avdl*	*c*	*mean dl*
French	0.7	1.5	600	1.25	182
English	0.8	2	800	1.5	167
Portuguese	0.7	1.5	700	1.7	250
Bulgarian	0.8	1.2	750	1.25	134
Hungarian	0.75	1.2	750	1.25	150
German GIRT	0.5	1.2	500	1.75	90
English GIRT	0.9	4	750	1.5	35

Table A.2: Parameter settings for the various test-collections

C301	Nestlé Brands	C338	Carlos' Extradition and Trial
C302	Consumer Boycotts	C339	Sinn Fein and the Anglo-Irish Declaration
C303	Italian paintings	C340	New Quebec Premier
C304	World Heritage Sites	C341	Theft of "The Scream"
C305	Oil Prices	C342	Four Weddings and a Funeral
C306	ETA Activities in France	C343	South African National Party
C307	Films Set in Scotland	C344	Brazil vs. Sweden World Cup Semifinals
C308	Solar Eclipse	C345	Cross-country Skiing at the Olympic Games
C309	"Hard" Drugs	C346	Grand Slam Winners
C310	Treatment of Industrial Waste	C347	Best Picture Oscar 1994
C311	Unemployment in Europe	C348	Yann Piat's Assassination
C312	Dog Attacks	C349	Nixon's Death
C313	Centenary Celebrations	C350	Ayrton Senna's Death
C314	Endangered Species	C351	Changes in Common Agricultural Policy
C315	Doping in Sports	C352	Peacekeeping in Afghanistan
C316	Strikes	C353	Forging of Euro
C317	Anti-cancer Drugs	C354	New Mobile Phone Functions
C318	Sex Education	C355	Greenhouse Effect Gases
C319	Global Opium Production	C356	New Heavenly Bodies
C320	Energy Crises	C357	Impact of September 11
C321	The Talibans in Afghanistan	C358	Wartime Looting
C322	Atomic Energy	C359	English Theatre
C323	Tightening Visa Requirements	C360	Water on Mars
C324	Supermodels	C361	Christian Pilgrimages
C325	Student Fees	C362	Dollar Euro Exchange Rate
C326	Emmy International Awards	C363	Winter Olympic Medallists
C327	Earthquakes in Mexico City	C364	German Chancellor Candidates
C328	Iraqi Kurds and Turkey	C365	US Economic Recession
C329	Consequences if Charles and Diana Divorce	C366	"Kursk" Submarine Tragedy
C330	Films with Keanu Reeves	C367	East Timor Independence
C331	Zedillo's Economic Policies	C368	Pharmaceutical Experiments
C332	Shooting of Tupac Shakur	C369	New Metro Lines
C333	Trial of Paul Touvier	C370	The Harry Potter Phenomenon
C334	Election of George W. Bush	C371	Broken Election Promises
C335	Labour after John Smith	C372	The Kashmir Crisis
C336	NBA Labour Conflicts	C373	Hungarian-Bulgarian Relationships
C337	Civil War in the Yemen	C374	Bin Laden's Associates
		C375	Kaliningrad Prospects

Table A.3: Query titles for CLEF-2006 ad-hoc test-collections

-

C151	Extreme right-wing parties in Germany	C164	The German school system
C152	Employment policy at the European level	C165	Street urchins
C153	Childlessness in Germany	C166	Poverty
C154	Modernizing of public administration	C167	Crime among women
C155	Domestic violence	C168	Anti-Semitism in Germany post 1945
C156	Illegal residency	C169	Genderspecific types of learning in elementary school
C157	Multilingualism among children	C170	Lean production in Japan
C158	Remigration and transmigration	C171	The computer in the everyday
C159	Role of the father	C172	Foreigners in elementary school
C160	Precarious working conditions	C173	Propensity towards violence among youths
C161	European social policy	C174	Poverty and homelessness in cities
C162	Motherhood and career development	C175	Parents' education level and children's school development
C163	Risk behavior		

Table A.4: Query titles for CLEF-2006 GIRT test-collections