Report on CLEF-2005 Evaluation Campaign: Monolingual, Bilingual, and GIRT Information Retrieval

Jacques Savoy, Pierre-Yves Berger

Institut interfacultaire d'informatique
University of Neuchatel, Switzerland
Jacques.Savoy@unine.ch www.unine.ch/info/clef/

Abstract

For our fifth participation in the CLEF evaluation campaigns, the first objective was to propose an effective and general stopword list along with a light stemming procedure for the Hungarian, Bulgarian and Portuguese (Brazilian) languages. Our second objective was to obtain a better picture of the relative merit of various search engines when processing documents in those languages. To do so we evaluated our scheme using two probabilistic models and nine vector-processing approaches. In the bilingual track, we evaluated both the machine translation and bilingual dictionary approaches to automatically translate a query submitted in English into various target languages. This year we explored new freely available translation sources, together with a combined query translation approach in order to obtain a better translation of the user's information need. Finally, using the GIRT corpora (available in English, German and Russian), we investigated variations in retrieval effectiveness when including or excluding manually assigned keywords attached to bibliographic records (mainly comprising a title and an abstract).

Categories and Subject Descriptors

H.3.1 [**Content Analysis and Indexing**]: Linguistic processing. H.3.3 [**Information Storage and Retrieval**]: Retrieval models, Relevance feedback. 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

Since 2001 our research group has been investigating effective information retrieval (IR) techniques when handling a variety of natural languages (Savoy 2004a; 2005a) in order to improve both monolingual and bilingual searches. Continuing along this same stream, our participation in the CLEF 2005 evaluation campaign will target various objectives. First, our aim is to propose linguistic tools for less frequently spoken languages such as Bulgarian and Hungarian, to explore the underlying IR problems with closely related languages such as Portuguese and Brazilian, and to explore new alternatives when translating a query from one source language (English in this study) to other target languages (more precisely the French, Portuguese, Bulgarian and Hungarian languages). The domain-specific GIRT corpus presents other interesting features, namely questions related to digital libraries with a collection comprising a large number of bibliographic records.

In addition to these particular objectives, various interesting problems must be analyzed and resolved. All languages are not written with the same alphabet, and Bulgarian for example uses the Cyrillic alphabet. The presence of diacritics in others also raises certain questions that directly affect the effectiveness of IR systems. Can we simply ignore them? Do they have a real impact on mean average precision? Does the distinction between uppercase and lowercase letters really influence information retrieval systems or does this distinction need only be preserved when high search precision is required?

In our work we have assumed that the semantic content of documents (or requests) is mainly linked to nouns and adjectives, and thus an effective search system can be based on the use of an appropriate set of weighted keywords extracted from corresponding documents (or requests). Based on this assumption, we designed a set of stopword lists and light stemming procedures for certain European and Asian languages. Following our suggestion, these linguistic tools were designed to automatically remove the inflectional suffixes attached to nouns and adjectives linked to gender (masculine, feminine, neural), to number (singular or plural), and to case (nominative, dative, ablative, etc.). Needless to say we were also interested in other linguistic phenomena, such as compound constructions (does an effective IR system really need to decompound them and is this linguistic phenomenon really important for the retrieval of languages other than German?)

The rest of this paper is organized as follows: Section 2 describes the main characteristics of the CLEF-2005 test-collection, 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 four 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.

2 Overview of the Test-Collections

The corpora used in our experiments include newspaper and news agency articles, namely *Le Monde* (1994-1995, French), *SDA* (1994-1995, French), *Público* (1994-1995, Portuguese), *Folha* (1994-1995, Brazilian), *Magyar Hirlap* (2002, Hungarian), *Sega* (2002, Bulgarian), *Standart* (2002, Bulgarian). As shown in Table 1, the Portuguese corpus (212.9 indexing terms / document) has a larger mean size article than the French collection (178). This mean value is 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.

	French	Portuguese	Bulgarian	Hungarian
Size (in MB)	487 MB	564 MB	213 MB	105 MB
# of documents	177,452	210,734	69,195	49,530
# of distinct terms	455,366	582,117	414,253	657,132
Number of distinct indexing terms / document				
Mean	127.8	153.5	102.7	107.9
Standard deviation	106.57	114.95	97.34	94.59
Median	92	129	72	77
Maximum	2,645	2,655	1,242	1,422
Minimum	1	1	1	2
Number of indexing terms / document				
Mean	178	212.9	133.7	142.1
Standard deviation	159.87	186.4	144.85	139.84
Median	126	171	88	95
Maximum	6,720	7,554	2,805	4,984
Minimum	1	1	1	2
Number of queries	50	50	49	50
Number rel. items	2,537	2,904	778	939
Mean rel./ request	50.74	58.08	15.878	18.78
Standard deviation	45.349	50.415	16.233	17.616
Median	35.5	44	10	13
Maximum	185 (Q#253)	239 (Q#286)	69 (Q#295)	87 (Q#290)
Minimum	1 (Q#255)	2 (Q#258)	1 (Q#258)	1 (Q#272)

Table 1: CLEF 2005 test-collection statistics

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>. For this restriction we found 1,854 documents in the Bulgarian collection to have no indexable content (for example, they may correspond to articles containing only a picture with the tags <PICTURE>, <IMGTEXT> and <IMGAUTHOR>). 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., "Anti-Smoking Legislation",

"Football Refereeing Disputes", or "Lottery Winnings"), including both regional ("Swiss Referendums") or international coverage ("Anti-abortion Movements").

3 Stopword Lists and Stemming Procedures

In order to define general stopword lists, we first created a list of the top 200 most frequent words found in the various languages, from which some words were removed (e.g., police, minister, president, Magyar). From this list of very frequent words, we added articles, pronouns, prepositions, conjunctions or very frequently occurring verb forms (e.g., to be, is, has, etc.). Based on this scheme, we created a new list for the Bulgarian and Hungarian languages (these lists are available at www.unine.ch/info/clef/). Our final stopword list contained 463 words for the French language, 761 for Hungarian, 418 for Bulgarian and 400 for Portuguese-Brazilian (we added 8 Brazilian words to our Portuguese stopword list. These eight words are usually variants with or without accents, such as "vezes" in Portuguese and "vêzes" in Brazilian).

Once high-frequency words were removed, our indexing procedure generally applied a stemming algorithm in an attempt to conflate word variants into the same stem or root. In developing such a procedure, we first wanted to remove only inflectional suffixes such as singular and plural word forms, and also feminine and masculine forms so that they would conflate to the same root.

Bulgarian involved additional morphological difficulties, given that in this language the definite article is usually represented by a suffix. For example, "mope" (sea) becomes "mopeto" (the sea) while "mopeta" (seas) becomes "mopetata" (the seas). The general noun pattern is as follows: <stem> <plural> <article>. Contrary to other Slavic languages (such as Russian), Bulgarian does not indicate grammatical cases by adding a suffix.

The Hungarian language shares certain similarities with the Finnish language (although both languages do not belong strictly to the same family, they can be viewed as cousins). Like Finnish, Hungarian has several number cases (usually 18) and each case has its own unambiguous form. For example, the noun "house" ("hàz") may appear as "hàzat" (accusative case, as in "(I see) the house"), "hàzakat" (accusative plural case, as in "(I see) the houses"), "hàzamat" ("… my house") or "hàzamait" ("… my houses"). In this language, the general construction used for nouns is as follows: <stem> <plural> <possessive marker> <case>. For example, for <hàz> a <m> a <t> in which the letter "a" is introduced to facilitate better pronunciation ("hàzmt" could be difficult to pronounce). From the IR point of view, certain linguistic aspects in Hungarian are viewed as good news. For example, a gender distinction is not attached to each noun (like in English) and adjectives are invariable, as in "… a szép hàzat" ("a beautiful house") or "… a szép hàzamat" ("my beautiful house"). Our suggested stemming procedures for these languages can be found at www.unine.ch/info/clef/.

Diacritic characters are usually not present in English collections (with certain exceptions, such as "résumé" or "cliché"). For the Hungarian, and Portuguese languages, these characters were replaced by their corresponding non-accentuated letter. Removing accents may however generate some semantic ambiguity (e.g., between "kor" ("age") and "kór" ("illness"), or "ver" ("hurt") and "vér" ("blood") in Hungarian language).

Finally, most European languages manifest other morphological characteristics, with compound word constructions being only one example (e.g., handgun, worldwide). Recently, Braschler & Ripplinger (2004) showed that decompounding German words could significantly improve retrieval performance, and in some experiments with Hungarian where we used our decompounding algorithm (Savoy 2004b), both compound words and their component parts were left in the documents and queries.

4 Indexing and Searching Strategies

In order to obtain a broader view of the relative merit of various retrieval models, we first adopted 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") or 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 (Robertson *et al*. 2000). As a second probabilistic approach, we implemented the Prosit approach, one member of a family of models suggested by Amati & van Rijsbergen (2002) and based on combining two information measures, formulated as follows:

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

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

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

where w_{ij} indicates the indexing weight attached to term t_j in document D_i, l_i the number of indexing terms included in the representation of D_i, where tc_j represents the number of occurrences of term t_j in the collection 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 listed in Table 2 (the German, English and Russian languages are used in the GIRT experiments).

	Okapi			Prosit	
Language	b	k_1	*avdl*	C	*mean dl*
French	0.7	1.5	600	1.25	182
Portuguese	0.7	1.5	700	1.7	250
Bulgarian	0.75	1.2	750	1.25	134
Hungarian	0.75	1.2	750	1.25	150
German	0.5	1.2	500	1.75	90
English	0.9	4	750	1.5	35
Russian	0.75	1.2	100	1.5	25

Table 2: Parameter settings for the various test collections

To measure the retrieval performance, we adopted non-interpolated mean average precision (MAP) (computed on the basis of 1,000 retrieved items per request by the new TREC-EVAL program). To statistically determine whether or not a given search strategy would be better than another, we applied the bootstrap methodology (Savoy 1997). Thus, in the tables included in this paper we underlined statistically significant differences using on a two-sided non-parametric bootstrap test, and based on the MAP difference with a significance level fixed at 5%.

	Mean average precision					
Query Model \ # of queries	French T 50 queries	French TD 50 queries	French TDN 50 queries	Portuguese T 50 queries	Portuguese TD 50 queries	Portuguese TDN 50 queries
Prosit	0.2895	0.3696	**0.3961**	0.2755	0.3438	0.3697
doc=Okapi, query=npn	**0.3029**	**0.3754**	0.3948	**0.2873**	**0.3477**	**0.3719**
doc=Lnu, query=ltc	0.2821	0.3437	0.3703	0.2611	0.3338	0.3517
doc=dtu, query=dtn	0.2726	0.3365	0.3633	0.2571	0.3221	0.3338
doc=atn, query=ntc	0.2809	0.3328	0.3507	0.2458	0.3076	0.3433
doc=ltn, query=ntc	0.2588	0.3066	0.3232	0.2149	0.2535	0.2740
doc=ntc, query=ntc	0.1862	0.2175	0.2335	0.1553	0.1868	0.2221
doc=ltc, query=ltc	0.1916	0.2363	0.2611	0.1625	0.2234	0.2543
doc=lnc, query=ltc	0.2050	0.2616	0.2953	0.1811	0.2475	0.2950
doc=bnn, query=bnn	0.1153	0.0937	0.0514	0.1309	0.1322	0.0900
doc=nnn, query=nnn	0.1148	0.0987	0.0748	0.0630	0.0639	0.0453

Table 3: Mean average precision of various single searching strategies (French & Portuguese languages)

We indexed the different collections using words as indexing units. The evaluations of our two probabilistic models and nine vector-space schemes are listed in Table 3 for the French and Portuguese corpus, and in Table 4 for the Bulgarian and Hungarian collection. In these tables, the best performance under given conditions (with the same indexing scheme and the same collection) is listed in bold type. Based on the best performance, this approach is also used as a baseline for our statistical testing. The underlined results therefore indicate that the difference in mean average precision can be viewed as statistically significant when compared to the best system value. As depicted in Table 3, the Okapi model was found to be the best IR model for French and Portuguese collection. For these two corpora however, the difference in MAP between the various IR models is usually statistically significant. As shown in Table 4 (and in Table A.4 in the Appendix) similar conclusions can be drawn for the Bulgarian and Hungarian collection. In this case the best performing system was the Prosit model for Bulgarian, and the Okapi probabilistic approach for Hungarian. Moreover five IR models were shown to have similar statistical performance levels (Okapi, Prosit, "doc=Lnu, query=ltc", "doc=dtu, query=dtn", "doc=atn, query=ntc").

Moreover, the data in these tables shows that when the number of search terms increases (from T, TD to TDN), retrieval effectiveness usually increases also (except for the "doc=bnn, query=bnn" or "doc=nnn, query=nnn" IR models). From an analysis of the five best retrieval schemes shown in Tables 3 and 4 (namely, Prosit, Okapi, "doc=Lnu, query=ltc", "doc=dtu, query=dtn" and "doc=atn, query=ntc"), the improvement is around 33.4% when comparing title-only (or T) with TDN queries for the Portuguese collection, 31.3% when comparing the French corpus, 21% for Hungarian (see Table A.4 in the Appendix), and 6.4% for the Bulgarian collection.

	Mean average precision					
Query Model \ # of queries	Bulgarian T 49 queries	Bulgarian TD 49 queries	Bulgarian TDN 49 queries	Hungarian TD 50 queries	Hungarian TD-decomp 50 queries	Hungarian TD-light 50 queries
Prosit	**0.2594**	**0.2953**	**0.2655**	0.3420	0.3390	0.3359
doc=Okapi, query=npn	0.2307	0.2704	0.2459	**0.3501**	**0.3391**	**0.3410**
doc=Lnu, query=ltc	0.2238	0.2679	0.2583	0.3301	0.3273	0.3249
doc=dtu, query=dtn	0.2255	0.2551	0.2364	0.3401	0.3341	0.3280
doc=atn, query=ntc	0.2277	0.2605	0.2411	0.3215	0.3179	0.3199
doc=ltn, query=ntc	0.1650	0.1999	0.1870	0.2853	0.2820	0.2856
doc=ntc, query=ntc	0.1758	0.1940	0.2052	0.2208	0.2099	0.2245
doc=ltc, query=ltc	0.2008	0.2323	0.2372	0.2484	0.2423	0.2482
doc=lnc, query=ltc	0.2036	0.2485	0.2445	0.2395	0.2424	0.2421
doc=bnn, query=bnn	0.0918	0.0689	0.0309	0.1424	0.1457	0.1432
doc=nnn, query=nnn	0.0774	0.0660	0.0354	0.0875	0.0824	0.1047

Table 4: Mean average precision of various single searching strategies (Bulgarian & Hungarian language)

With the Hungarian collection, 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 (under the label "TD-decomp" in Table 4). Using the TD queries and the Okapi model, we achieved a MAP of 0.3391, reflecting a degradation of -3.1% when compared to an indexing approach that did not use decompounding. Based on the five best retrieval schemes, the mean degradation is around -1.6%. Using a lighter stemmer (less rules) for the Hungarian language (retrieval performance depicted under the label "TD-light" in Table 4), the mean difference in MAP over the five best retrieval schemes is around 2% and in favor of a more complex stemming approach.

	Mean average precision			
Query TD Model	French 50 queries	Portuguese 50 queries	Bulgarian 49 queries	Hungarian 50 queries
Okapi	0.3754	0.3477	0.2704	0.3501
k doc. / *m* terms	3/10 0.3967	3/15 0.3656	3/10 0.2534	3/10 **0.3545**
	5/15 0.4034	5/15 **0.3668**	5/10 0.2626	5/10 0.3513
	10/15 **0.4099**	10/15 0.3626	5/20 0.2586	5/15 0.3490
	10/20 0.4075	10/20 0.3601	10/15 **0.2726**	10/15 0.3492

Table 5: Mean average precision using blind-query expansion (Okapi model)

	Mean average precision			
Query TD Model	French 50 queries	Portuguese 50 queries	Bulgarian 49 queries	Hungarian 50 queries
Prosit	0.3696	0.3438	0.2953	0.3420
k doc. / *m* terms	3/10 0.3898	3/20 0.3645	3/10 0.2897	3/50 **0.3940**
	3/15 0.3959	5/30 0.3818	3/15 0.3026	5/20 0.3649
	5/10 0.4004	5/50 0.3744	5/10 **0.3091**	5/50 0.3764
	5/50 0.3987	10/30 **0.3953**	5/15 0.2966	10/20 0.3530
	10/50 **0.4225**	10/50 0.3864	10/15 0.2852	10/30 0.3672

Table 6: Mean average precision using blind-query expansion (Prosit model)

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 the Okapi and the Prosit probabilistic models and enlarged the query by the 10 to 50 terms retrieved from the 3 to 10 best-ranked articles.

Table 5 depicts our best results using pseudo-relevance feedback technique for the Okapi model and demonstrates that the optimal parameter setting seemed to be collection-dependant. Moreover, performance improvement also seemed to be collection dependant (or language dependant), with the French corpus showing an increase of +9.2% (from a mean average precision of 0.3754 to 0.4099), +5.2% for the Portuguese collection (from 0.3477 to 0.3668), +1.3% for the Hungarian collection (from 0.3501 to 0.3545), and +0.8% for the Bulgarian corpus (from 0.2704 to 0.2726). Table 6 shows how similar conclusions can be drawn using the Prosit model. In this case however, the blind query expansion depicted a greater improvement for all collections (e.g., for the French corpus, an increase of +14.3%, from a mean average precision of 0.3696 to 0.4225). In both Tables 5 and 6, the baseline used for our statistical testing was the MAP, calculated before the query was automatically expanded. In this case, it is interesting to note that our statistical testing cannot always detect a significant difference in MAP before and after blind query expansion, specially for the Bulgarian and Hungarian collection.

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). On the other hand, when combining different search schemes, we might suppose that these various IR strategies are more likely to rank the same relevant items higher on the list than they would for non-relevant documents (viewed as outliers). Thus, combining them could improve retrieval effectiveness by ranking pertinent documents higher and ranking non-relevant items lower. Based on our previous studies (Savoy 2004b, 2005a), this expected positive effect does not always work.

In this current study we combine only the two probabilistic models because they usually depict the best or one of the best retrieval performances (Savoy 2004b, 2005a). To achieve this we evaluated various fusion operators (see Table 7 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 7 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 γ_i (usually equal to 1) reflecting the differences in retrieval performance.

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

Table 7: 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 7, 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 γ_i value to each retrieval model.

Table 8 depicts the evaluation of various data fusion operators, comparing them to the single approach using the Okapi and the Prosit probabilistic models. From this data, we can see that combining two IR models might improve retrieval effectiveness. When combining two retrieval models, the Z-Score scheme tended to perform the best. In Table 8, under the heading "Z-ScoreW", we attached a weight of 1.5 to the Prosit model (depicted in bold in the second line), and 1 to the Okapi scheme. Using the Prosit performance as a baseline, our statistical testing was not usually able to detect any significant enhancement when combining two IR models.

	Mean average precision (% of change)			
Query TD Model	French 50 queries	Portuguese 50 queries	Bulgarian 49 queries	Hungarian 50 queries
Okapi & PRF doc/term	10/15 0.4099	5/15 0.3668	10/15 0.2726	3/10 0.3545
Prosit & PRF doc/term	10/50 **0.4225**	10/30 **0.3953**	5/10 **0.3091**	3/50 **0.3940**
Round-robin	0.4313 (+2.1%)	0.3938 (-0.4%)	0.2736 (-11.5%)	0.3858 (-2.1%)
Sum RSV	0.4319 (+2.2%)	0.4010 (+1.4%)	0.2775 (-10.2%)	**0.4034** (+2.4%)
Norm Max	0.4293 (+1.6%)	0.3984 (+0.8%)	0.2724 (-11.9%)	0.3778 (-4.1%)
Norm RSV	0.4323 (+2.3%)	**0.4018** (+1.6%)	0.2727 (-11.8%)	0.3779 (-4.1%)
Z-Score	0.4338 (+2.7%)	0.4005 (+1.3%)	0.2753 (-10.9%)	0.3900 (-1.0%)
Z-ScoreW	**0.4350** (+3.0%)	**0.4018** (+1.6%)	0.2765 (-10.6%)	0.3966 (+0.7%)

Table 8: Mean average precision using different combination operators (with blind-query expansion)

6 Official Results

Table 9 shows the exact specifications of our 12 official monolingual runs. These experiments were mainly based on the Okapi and the Prosit probabilistic models as well as the Z-Score data fusion operator.

Run name	Language	Query	Index	Model	Query expansion	Combined	MAP
UniNEfr1	French	TD TD	word word	Okapi Prosit	3 best docs / 10 terms 5 best docs / 50 terms	Z-ScoreW	**0.4207**
UniNEfr2	French	TD TD	word word	Prosit Okapi	3 best docs / 10 terms 3 best docs / 10 terms	Round-Robin	0.4051
UniNEfr3	French	TD	word	Prosit	5 best docs / 10 terms	n/a	0.4066
UniNEpt1	Portuguese	TD TD	word word	Okapi Prosit	10 best docs / 20 terms 10 best docs / 50 terms	Z-ScoreW	0.3825
UniNEpt2	Portuguese	TD TD	word word	Okapi Prosit	3 best docs / 15 terms 5 best docs / 60 terms	Z-ScoreW	**0.3875**
UniNEpt3	Portuguese	TD	word	Prosit	5 best docs / 75 terms	n/a	0.3665
UniNEbg1	Bulgarian	TD TD	word word	Okapi Prosit	5 best docs / 20 terms 5 best docs / 40 terms	Z-Score	0.2782
UniNEbg2	Bulgarian	TD TD	word word	Okapi Prosit	10 best docs / 15 terms 10 best docs / 50 terms	Z-ScoreW	0.2622
UniNEbg3	Bulgarian	TD	word	Prosit	5 best docs / 30 terms	n/a	**0.2839**
UniNEhu1	Hungarian	TD TD	word word	Okapi Prosit	10 best docs / 15 terms 10 best docs / 30 terms	Z-ScoreW	0.3699
UniNEhu2	Hungarian	TD	word	Okapi	5 best docs / 15 terms lighter stemmer	n/a	0.3395
UniNEhu3	Hungarian	TD	word	Prosit	5 best docs / 40 terms	n/a	**0.3889**

Table 9: Description and mean average precision (MAP) of our official monolingual runs

7 Bilingual Information Retrieval

For the bilingual track, we chose English as the language for submitting queries to be automatically translated into four different languages, using nine different machine translation (MT) systems and four bilingual dictionaries ("Babylon", "Ectaco", "Medios", and "Kerekes"). The following freely available translation tools were used in our experiments:

SYSTRAN	www.systranlinks.com/
GOOGLE	www.google.com/language_tools
FREETRANSLATION	www.freetranslation.com/web.htm
INTERTRAN	www.tranexp.com/
WORLDLINGO	www.worldlingo.com/
BABELFISH	babelFish.altavista.com/

PROMT	webtranslation.paralink.com/
ALPHAWORKS	www.alphaWorks.ibm.com/
APPLIEDLANGUAGE	www.appliedLanguage.com/
BABYLON	www.babylon.com
ECTACO	www.ectaco.co.uk/free-online-dictionaries/
MEDIOS	consulting.medios.fi/dictionary/ (only for Hungarian language)
KEREKES	www.cab.u-szeged.hu/cgi-bin/szotar (only for Hungarian language)

When using the different bilingual dictionaries to translate an English request word-by-word, usually more than one translation is provided, in an unspecified order. We decided to pick only the first translation available (labeled "Babylon 1" or "Ectaco 1"), the first two terms (e.g., "Babylon 2" or "Medios 2") or the first three available translations (labeled "Babylon 3").

Moreover, the query terms could be preprocessed in order to obtain their part-of-speech (PoS) information (using www.ims.unistuttgart.de/projekte/corplex/TreeTagger/). Using this information, we could find the corresponding lemma and use it instead of the surface word before searching in the bilingual dictionaries. Once this lemmatizing procedure was done, we added the term "+ PoS" in the corresponding run label. Table 10 contains an example of this query preprocessing, showing how the plural form was removed (e.g., "disputes" into "dispute") and how various verb forms were transformed into their lexical forms (e.g., "made" into "make" or "refereeing" into "referee").

<num> C263 </num>
<title> Football Refereeing Disputes </title>
<desc> Find documents in which decisions made by a referee during a football match are criticised. </desc>
<narr> Relevant documents report on football (soccer) matches in which the referee made some disputable or disputed decision. </narr>

<num> C263 </num>
<title> Football referee dispute </title>
<desc> find document in which decision make by a referee during a football match be criticize. </desc>
<narr> relevant document report on football (soccer) match in which the referee make some disputable or disputed decision. </narr>

Table 10: Example of a query before (top) and after PoS processing (bottom) in which the modifications are underlined

Table 11 shows the mean average precision obtained using the various MT tools and the Okapi probabilistic model, while Table 12 depicts the same information when using bilingual dictionaries. Of course, all tools are not always available for each language and thus various entries are missing (as shown in Tables 11 and 12, indicated by the label "N/A"). As expected, only a few translation tools are available for translating from English to either Bulgarian or Hungarian languages.

	Mean average precision (% of monolingual search)			
Language Okapi (TD queries)	French 50 queries	Portuguese 50 queries	Bulgarian 49 queries	Hungarian 50 queries
Manual	0.3754	0.3477	0.2704	0.3501
Systran	0.3149 (83.9%)	0.1835 (52.8%)	N / A	N / A
Google	**0.3259** (86.8%)	0.1840 (52.9%)	N / A	N / A
FreeTranslation	0.2814 (75.0%)	0.2507 (72.1%)	N / A	N / A
InterTrans	0.1839 (49.0%)	0.2396 (68.9%)	0.0518 (19.2%)	0.1722 (49.2%)
WorldLingo	0.3095 (82.5%)	0.1836 (52.8%)	N / A	N / A
BabelFish	0.3149 (83.9%)	0.1836 (52.8%)	N / A	N / A
Promt	0.3066 (81.7%)	**0.2673** (76.9%)	N / A	N / A
AlphaWorks	0.2991 (79.7%)	N / A	N / A	N / A
AppliedLanguage	0.3149 (83.9%)	0.1835 (52.8%)	N / A	N / A

Table 11: Mean average precision of various machine translation systems (TD queries, Okapi model)

From this data, we can see that for the French collection the best translation is obtained by Google and for the Portuguese corpus by Promt. The FreeTranslation and Promt MT systems usually obtain satisfactory retrieval performances for these two languages (around 79.3% of the MAP obtained by the corresponding

monolingual search for the Promt system, and 73.6% for FreeTranslation). Other good translation systems found were the BabelFish, Systran and AppliedLanguage which worked well for French. For Bulgarian and Hungarian languages, we found only a few translation tools, and unfortunately their overall performance levels were not very good. As depicted in Table 12, we also found that lemmatizing the English queries (for both the Bulgarian or Hungarian languages at least) would improve mean average precision.

	Mean average precision							
Language Okapi (TD)	French 50 queries		Portuguese 50 queries		Bulgarian 49 queries		Hungarian 50 queries	
Manual	0.3754		0.3477		0.2704		0.3501	
		+ PoS		+ PoS		+ PoS		+ PoS
Babylon 1	0.2548	N / A	0.2087	N / A	0.0508	0.0539	0.1147	0.1551
Babylon 2	0.2236	N / A	0.1800	N / A	0.0658	0.0741	N / A	0.1578
Babylon 3	0.1930	N / A	N / A	N / A.	0.0754	0.0800	N / A	N / A
Ectaco 1	N / A	0.1767	N / A	N / A	N / A	0.0734	N / A	0.1822
Ectaco 2	N / A	0.1851	N / A	N / A	N / A	0.0746	N / A	0.1393
Ectaco 3	N / A	0.1601	N / A	N / A	N / A	0.0722	N / A	N / A
Medios 1	N / A	N / A	N / A	N / A	N / A	N / A	N / A	0.1357
Medios 2	N / A	N / A	N / A	N / A	N / A	N / A	N / A	0.1540
Medios 3	N / A	N / A	N / A	N / A	N / A	N / A	N / A.	0.0940
Kerekes 1	N / A	N / A	N / A	N / A	N / A	N / A	N / A	N / A
Kerekes 2	N / A	N / A	N / A	N / A	N / A	N / A	N / A	N / A
Kerekes 3	N / A	N / A	N / A	N / A	N / A	N / A	N / A	N / A

Table 12: Mean average precision of various bilingual dictionaries (TD queries, Okapi model)

Table 13 shows the retrieval effectiveness for various query translation combinations when using the Okapi probabilistic model. The top part of the table indicates the exact query translation combination used while the bottom part shows the mean average precision obtained with our combined query translation approach. When selecting which query translations were to be combined, we took our prior findings (Savoy 2005a) and feelings into consideration when selecting best translation tools. The resulting retrieval performances depicted in Table 13 are sometimes better than the best single translation scheme, as indicated in the row labeled "Best single" (e.g., the strategy "Comb 1" for French, or the strategies "Comb 3" or "Comb 5" for Portuguese, "Comb 2" for Bulgarian, and "Comb 5" for Hungarian). Statistically however these combined query translation approaches did not perform better than the best single translation tool (except "Comb 3" with the Portuguese corpus).

	Mean average precision (% of change)			
Language Combination	French Okapi 50 queries	Portuguese Okapi 50 queries	Bulgarian Okapi 49 queries	Hungarian Okapi 50 queries
Comb 1	Systran + Promt	Promt + Babylon 1	Inter + all 2	Inter+Ecta 1 (PoS)
Comb 2	Lingo + Babylon 1	Promt + Inter	Ecta1+Baby 2 (PoS)	Inter+Baby 1 (PoS)
Comb 3	Free + Promt + Babylon 1	Promt + Free + Babylon 1		Baby 1 + Med 2 Ecta 1 (PoS)
Comb 4	Lingo+ Promt + Babylon 1	Promt + Inter + Babylon 1	Inter + Ecta 1 + Baby 2 (PoS)	Inter + Baby 1 + Ecta 1 (PoS)
Comb 5		Promt + Free + Inter + Babylon 1		Inter + Babylon 1 + Medi 2+Ecta 1 (PoS)
Best single	0.3259	0.2673	0.0800	0.1822
Comb 1	**0.3274** (+0.5%)	0.2849 (+6.6%)	0.0831 (+3.9%)	0.1845 (+1.6%)
Comb 2	0.3089 (-5.2%)	0.2749 (+2.8%)	**0.0962** (+20.2%)	0.1876 (+3.0%)
Comb 3	0.3246 (-0.4%)	0.2977 (+11.4%)		0.1966 (+7.9%)
Comb 4	0.3228 (-0.9%)	0.2955 (+10.6%)	0.0908 (+13.5%)	0.2005 (+10.0%)
Comb 5		**0.2978** (+11.4%)		**0.2183** (+19.8%)

Table 13: Mean average precision of various combined translation devices (TD queries, Okapi model)

From English to …	French 50 queries	Portuguese 50 queries	Bulgarian 49 queries	Hungarian 50 queries
IR 1 (#docs/#terms)	Okapi (5/10)	Okapi (10/30)	Okapi (3/30)	Okapi (0/0)
IR 2 (#docs/#terms)	Prosit (0/0)	Prosit (10/20)	Prosit (10/75)	Prosit (0/0)
Data fusion operator	Round-robin	Z-ScoreW	Z-ScoreW	Z-ScoreW
Translation tools	Comb4	Comb4	Comb4	Comb5
MAP	0.3357	**0.3404**	0.1062	0.2786
Run name	UniNEbifr1	UniNEbipt1	UniNEbibg1	UniNEbihu1
IR (#docs/#terms)	Okapi (10/10)	Prosit (10/50)	Lnu-ltc (3/50)	Prosit (10/10)
Translation tools	Comb3	Comb5	Comb4	Comb4
MAP	**0.3467**	0.3094	0.1200	0.2315
Run name	UniNEbifr2	UniNEbipt2	UniNEbibg2	UniNEbihu2
IR (#docs/#terms)	Okapi (5/10)	Okapi (10/30)	Prosit (3/50)	Prosit (3/50)
Translation tools	Comb4	Comb4	Comb3	Comb5
MAP	0.3444	0.3117	**0.1399**	**0.2882**
Run name	UniNEbifr3	UniNEbipt3	UniNEbibg3	UniNEbihu3

Table 14: Description and mean average precision of our official bilingual runs

Finally, Table 14 lists the parameter settings used for 12 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 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 three 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 15.

	German	English	Russian
Size (in MB)	326 MB	199 MB	65 MB
# of documents	151,319	151,319	94,581
# of distinct terms	698,638	151,181	131,231
Number of distinct indexing terms / document			
Mean	70.83	107.9	18.86
Standard deviation	32.4	94.59	26.8
Median	68	77	9
Maximum	386	1,422	567
Minimum	2	2	2
Number of indexing terms / document			
Mean	89.61	142.1	23.79
Standard deviation	44.5	139.84	41.48
Median	84	95	9
Maximum	629	4,984	1,111
Minimum	4	2	3
Number of queries	25	25	25
Number rel. items	2,682	2,105	831
Mean rel./ request	107.28	84.2	33.24
Standard deviation	91.654	69.109	41.95
Median	75	54	12
Maximum	318 (Q#150)	242 (Q#150)	138 (Q#139)
Minimum	8 (Q#129)	6 (Q#129)	1 (Q#146)

Table 15: CLEF 2005 GIRT test collection statistics

<DOC> <DOCNO> GIRT-EN19901932
<TITLE-EN> The Socio-Economic Transformation of a Region : the Bergische Land from 1930 to 1960
<AUTHOR> Henne, Franz J.
<AUTHOR> Geyer, Michael
<PUBLICATION-YEAR> 1990
<LANGUAGE-CODE> EN
<CONTROLLED-TERM-EN> Rhenish Prussia
<CONTROLLED-TERM-EN> historical development
<CONTROLLED-TERM-EN> regional development
<CONTROLLED-TERM-EN> socioeconomic factors
<METHOD-TERM-EN> historical
<METHOD-TERM-EN> document analysis
<CLASSIFICATION-TEXT-EN> Social History
<DOC> <DOCNO> GIRT-EN19902732
<TITLE-EN> Ethnic Politicians in Congress: German-American Case Studies on the Interaction of Ethnicity, Nationality and Democratic Government 1865-1930
<AUTHOR> Adams, Willi Paul
<PUBLICATION-YEAR> 1990
<LANGUAGE-CODE> EN
<CONTROLLED-TERM-EN> ethnic group
<CONTROLLED-TERM-EN> North America …

Table 16: Examples of two bibliographic notices written in English from originals available in German

In total theses collections contain 397,218 documents or about 590 MB, and for the most part are written in German. A typical record in this collection is composed of a title, an abstract, and a set of manually assigned keyword (see Table 16 for English examples and Table 17 for their corresponding German records). 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., “Electoral Behaviour”, “New Art”, “Soccer and Society”, or “Churches and Money”).

<DOC> <DOCNO> GIRT-DE19909343
<TITLE-DE> Die sozioökonomische Transformation einer Region : Das Bergische Land von 1930 bis 1960
<AUTHOR> Henne, Franz J.
<AUTHOR> Geyer, Michael
<PUBLICATION-YEAR> 1990
<LANGUAGE-CODE> DE
<CONTROLLED-TERM-DE> Rheinland
<CONTROLLED-TERM-DE> historische Entwicklung
<CONTROLLED-TERM-DE> regionale Entwicklung
<CONTROLLED-TERM-DE> sozioökonomische Faktoren
<METHOD-TERM-DE> historisch
<METHOD-TERM-DE> Aktenanalyse
<CLASSIFICATION-TEXT-DE> Sozialgeschichte
<ABSTRACT-DE> Die Arbeit hat das Ziel, anhand einer regionalen Studie die Entstehung des "modernen" fordistischen Wirtschaftssystems und des sozialen Systems im Zeitraum zwischen 1930 und 1960 zu beleuchten; dabei geht es auch um das Studium des "Sozial-imaginären", der Veränderung von Bewußtsein und Selbst-Verständnis von Arbeitern durch das Erlebnis und die Erfahrung der Depression, des Nationalsozialismus und der Nachkriegszeit, welches sich in den 1950er Jahren gemeinsam mit der wirtschaftlichen Veränderung zu einem neuen "System" zusammenfügt.
<DOC> <DOCNO> GIRT-DE19909106
<TITLE-DE> Politiker einer ethnischen Gruppe im Kongreß: Deutsch-amerikanische Fallstudien zur Interaktion von Ethnizität, Nationalität und demokratischer Regierung, 1865-1930
<AUTHOR> Adams, Willi Paul
<PUBLICATION-YEAR> 1990
<LANGUAGE-CODE> DE
<CONTROLLED-TERM-DE> ethnische Gruppe
<CONTROLLED-TERM-DE> Nordamerika …

Table 17: Original of two bibliographic notices written in German

Query TD Model \ # of queries	Mean average precision German all 25 queries	German TI & AB 25 queries	English all 25 queries	English TI & AB 25 queries	Russian all 25 queries
Prosit	0.4249	**0.3659**	**0.4645**	**0.2948**	0.2270
doc=Okapi, query=npn	**0.4353**	0.3645	0.4604	0.2854	0.2742
doc=Lnu, query=ltc	0.3977	0.3307	0.4234	0.2712	0.2577
doc=dtu, query=dtn	0.3789	0.3236	0.3936	0.2738	**0.3003**
doc=atn, query=ntc	0.3914	0.3458	0.4102	0.2681	0.2695
doc=ltn, query=ntc	0.3724	0.3146	0.3448	0.2158	0.2636
doc=ntc, query=ntc	0.2765	0.2452	0.2859	0.2023	0.1393
doc=ltc, query=ltc	0.2926	0.2571	0.3095	0.1997	0.1248
doc=lnc, query=ltc	0.3364	0.2757	0.3446	0.1855	0.1416
doc=bnn, query=bnn	0.2025	0.1723	0.1790	0.0996	0.0826
doc=nnn, query=nnn	0.1373	0.1190	0.0951	0.0635	0.0984

Table 18: Mean average precision of various single searching strategies (GIRT corpus)

Based on the GIRT corpus we are therefore able to evaluate the impact of manually assigned descriptors as compared to an indexing scheme, based only on the information contained in the corresponding article's title and abstract sections. To tackle this question we evaluated the GIRT collection using all sections (denoted "all" in Table 18), or only using titles and abstracts from bibliographic records (under the label "TI & AB"). In related research using the Amaryllis French corpus, we found that the "TI & AB" indexing scheme presents a loss of around 45% in mean average precision (Savoy 2005b) when compared to the "all" approach. In our experiments, the decrease in mean average precision is around -14.4% for the German corpus and -36.5% for the English GIRT collection.

Our 12 official runs in the monolingual GIRT task are described in Table 19. For each language, we submitted the first run using a data fusion operator ("Z-ScoreW" in this case). For all runs, we automatically expanded the queries using a blind relevance feedback method (Rocchio in our experiments), hoping to improve retrieval effectiveness.

Run name	Language	Query	Index	Model	Query expansion	Combined	MAP
UniNEgde1	German	TD TD	word word	Okapi Prosit	5 best docs / 10 terms 10 best docs / 125 terms	Z-ScoreW	**0.4921**
UniNEgde2	German	TD	word	Prosit	10 best docs / 75 terms	N / A	0.4730
UniNEgde3	German	TD	word	Prosit	10 best docs / 150 terms	N / A	0.4597
UniNEgen1	English	TD TD	word word	Okapi Prosit	5 best docs / 10 terms 10 best docs / 50 terms	Z-ScoreW	**0.5065**
UniNEgen2	English	TD	word	Prosit	10 best docs / 60 terms	N / A	0.5043
UniNEgen3	English	TD	word	Prosit	5 best docs / 20 terms	N / A	0.4419
UniNEgru1	Russian	TD TD	word word	Okapi Prosit	5 best docs / 10 terms 10 best docs / 30 terms	Z-ScoreW	0.2491
UniNEgru2	Russian	TD	word	Okapi	5 best docs / 20 terms	N / A	**0.2774**
UniNEgru3	Russian	TD	word	Prosit	10 best docs / 50 terms	N / A	0.2477

Table 19: Description and mean average precision (MAP) of our official GIRT runs

9 Conclusion

In this sixth CLEF evaluation campaign, we proposed a general stopword list and a light stemming procedure (removing only inflections attached to nouns and adjectives) for the Bulgarian and Hungarian languages (see Table 4 and Table A.4). In order to enhance retrieval performance, we suggested using a data fusion approach based on the Z-Score in order to combine the two probabilistic IR models (see Table 8). The results of this evaluation campaign seem to indicate that for the French and Portuguese languages such an approach proved to be effective (Table 8). The use of this search strategy did however require the building of two inverted files and doubling the search time required. For both the Bulgarian and Hungarian languages, more experiments are needed to confirm our first evaluations (especially in the design of a light stemming procedure for the Hungarian language, see Table 4). For all languages however, the probabilistic models (either Okapi or Prosit) usually

result in better retrieval performances than do other vector-processing approaches (see Tables 3, 4, and 18 for the GIRT corpora), while the data fusion approach did not always improve mean average precision. The automatic decompounding of Hungarian words and its impact in IR remains an open question and our preliminary experiments did not provide a clear and precise answer (our decompounding scheme slightly decreased retrieval performance, as shown in Table 4).

As in previous evaluation campaigns we were able to confirm that pseudo-relevance feedback based on Rocchio's model usually did improve mean average precision statistics for the French and Portuguese language, even though this improvement is not always statistically significant. For the other languages (Bulgarian and Hungarian), this blind query expansion did not improve mean average precision from the statistics point of view (Tables 5 and 6).

In the bilingual task, the freely available translation tools performed at a reasonable level for both the French and Portuguese languages (based on the three best translation tools, the MAP compared to the monolingual search is around 85% for the French language and 72.6% for the Portuguese). For less frequently used languages such as Bulgarian and Hungarian, the freely available translation tools (either the bilingual dictionary or the MT system) did not perform well. The mean average precision decreased by more than 50% (for Hungarian) to 80% (for Bulgarian), when compared to a monolingual search.

In the GIRT task (Table 18), we were able to measure the retrieval effectiveness by assigning keywords manually, and the presence of this information improved MAP by around 36.5% for the English corpus and 14.4% for the German collection.

Acknowledgments

The authors would like to also thank the CLEF-2005 task organizers for their efforts in developing various European language test-collections, and C. Buckley from SabIR for giving us the opportunity to use the SMART system. The first author is not able to thank the computing services at UniNE, because they consistently made no effort to be cooperative during this project. This research was supported in part by the Swiss National Science Foundation under Grant #21-66 742.01.

References

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

Braschler, M. & Ripplinger, B. (2004). How effective is stemming and decompounding for German text retrieval? *IR Journal*, 7(3-4), 291-316.

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

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

Kluck, M. (2004). The GIRT data i 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.

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). 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. (2005b). Bibliographic database access using free-text and controlled vocabulary: An evaluation. *Information Processing & Management*, 41(4), 873-890.

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

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	$w_{ij} = \frac{((k_1 + 1) \cdot tf_{ij})}{(K + tf_{ij})}$	Lnu	$w_{ij} = \frac{1 + \ln(tf_{ij}) / \ln(\text{mean } tf) + 1}{(1 - \text{slope}) \cdot \text{pivot} + \text{slope} \cdot nt_i}$
lnc	$w_{ij} = \frac{\ln(tf_{ij}) + 1}{\sqrt{\sum_{k=1}^{t} (\ln(tf_{ik}) + 1)^2}}$	ntc	$w_{ij} = \frac{tf_{ij} \cdot idf_j}{\sqrt{\sum_{k=1}^{t} (tf_{ik} \cdot idf_k)^2}}$
ltc	$w_{ij} = \frac{(\ln(tf_{ij}) + 1) \cdot idf_j}{\sqrt{\sum_{k=1}^{t} ((\ln(tf_{ik}) + 1) \cdot idf_k)^2}}$		
dtu	$w_{ij} = \frac{(\ln(\ln(tf_{ij}) + 1) + 1) \cdot idf_j}{(1 - \text{slope}) \cdot \text{pivot} + \text{slope} \cdot nt_i}$		

Table A.1: Weighting schemes

C251	Alternative Medicine	C276	EU Agricultural Subsidies
C252	Pension Schemes in Europe	C277	Euthanasia by Medics
C253	Countries with Death Penalty	C278	Transport for Disabled
C254	Earthquake Damage	C279	Swiss Referendums
C255	Internet Junkies	C280	Crime in New York
C256	Creutzfeldt-Jakob Disease	C281	Radovan Karadzic
C257	Ethnic Cleansing in the Balkans	C282	Prison Abuse
C258	Brain-Drain Impact	C283	James Bond Films
C259	Golden Bear	C284	Space Shuttle Missions
C260	Anti-Smoking Legislation	C285	Anti-abortion Movements
C261	Fortune-telling	C286	Football Injuries
C262	Benefit Concerts	C287	Hostage / Terrorist Situations
C263	Football Refereeing Disputes	C288	US Car Imports
C264	Smuggling of Radioactive Material	C289	Falkland Islands
C265	Deutsche Bank Takeovers	C290	Oil Price Fluctuation
C266	Discrimination against European Gypsies	C291	EU Illegal Immigrants
C267	Best Foreign Language Films	C292	Rebuilding German Cities
C268	Human Cloning and Ethics	C293	China-Taiwan Relations
C269	Treaty Ratification	C294	Hurricane Force
C270	Microsoft Competitors	C295	Money Laundering
C271	Gay Marriages	C296	Public Performances of Liszt
C272	Czech President's Background	C297	Expulsion of Diplomats
C273	NATO Expansion	C298	Nuclear Power Stations
C274	Unexploded World War II Bombs	C299	UN Peacekeeping Risks
C275	Smoking-related Diseases	C300	Lottery Winnings

Table A.2: Query titles for CLEF-2005 test collection

C126	New Art	C139	Health Economics
C127	Electoral Behaviour	C140	Oil and Politics
C128	Life Satisfaction	C141	Street Children
C129	Sexuality and Disability	C142	Advertising and Ethics
C130	Water Shortage	C143	Giving up Smoking
C131	Bilingual Education	C144	Radio and Internet
C132	Sexual Abuse of Children	C145	Poverty and Wealth
C133	Churches and Money	C146	Diabetes Mellitus
C134	Russian-Chinese Economic Relations	C147	Soccer and Society
C135	Pensions in Post-Soviet Countries	C148	Russian Germans and their Language
C136	Ecological Waste Economics	C149	Anti-Semitism in the Soviet Union
C137	Honour in Society	C150	Television Behaviour
C138	Insolvent Companies		

Table A.3: Query titles for CLEF-2005 GIRT test collection

	Mean average precision		
Query Model \ # of queries	Hungarian T 50 queries	Hungarian TD 50 queries	Hungarian TDN 50 queries
Prosit	0.2964	0.3420	**0.3600**
doc=Okapi, query=npn	**0.3076**	**0.3501**	0.3511
doc=Lnu, query=ltc	0.2868	0.3301	0.3421
doc=dtu, query=dtn	0.2900	0.3401	0.3571
doc=atn, query=ntc	0.2755	0.3215	0.3517
doc=ltn, query=ntc	0.2567	0.2853	0.3212
doc=ntc, query=ntc	0.2079	0.2208	0.2708
doc=ltc, query=ltc	0.2183	0.2484	0.2870
doc=lnc, query=ltc	0.2153	0.2395	0.2931
doc=bnn, query=bnn	0.1782	0.1424	0.1013
doc=nnn, query=nnn	0.1256	0.0875	0.0744

Table A.4: Mean average precision of various single searching strategies (Hungarian language)