Stemming Approaches for East European Languages

Ljiljana Dolamic, Jacques Savoy

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

Abstract

In our participation in this CLEF evaluation campaign, the first objective is to propose and evaluate various indexing and search strategies for the Czech language in order to hopefully produce better retrieval effectiveness than that of the language-independent approach (*n*-gram). Based on our stemming strategy used with other languages, we propose two light stemmers for this Slavic language and a third one based on a more aggressive suffix-stripping scheme that removes some derivational suffixes. Our second objective is to obtain a better picture of the relative merit of various search engines in exploring Hungarian and Bulgarian documents. Moreover for the Bulgarian language we developed a new and more aggressive stemmer. To evaluate these solutions we use our various IR models, including the Okapi, *Divergence from Randomness* (DFR) and statistical language model (LM) together with the classical *tf·idf* vector-processing approach. Our experiments tend to show that for the Bulgarian language removing certain frequently used derivational suffixes may improve mean average precision. For the Hungarian corpus, applying an automatic decompounding procedure improves the MAP. For the Czech language, a comparison between a light (inflectional only) and a more aggressive stemmer that removes both inflectional and some derivational suffixes reveals small performance differences. For this language only, the performance difference between a word-based or a 4-gram indexing strategy is also rather small, while for the Hungarian or Bulgarian corpora, a word-based approach tend to produce better MAP.

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 East European Languages, Stemmer, Stemming Strategy, Czech Language, Hungarian Language, Bulgarian Language.

1 Introduction

During the last few years, the IR group at University of Neuchatel has been involved in designing, implementing and evaluating IR systems for various natural languages, including both European (Savoy & Abdou, 2007) and popular Asian (Savoy, 2005) (Abdou & Savoy, 2007a) languages (namely, Chinese, Japanese, and Korean). In this context our main objective is to promote effective monolingual IR in those languages. For our participation in the CLEF 2007 evaluation campaign we decided to review our stemming strategy by including some very frequently used derivational suffixes. When defining our stemming rules however we still focus only on nouns and adjectives.

The rest of this paper is organized as follows: Section 2 describes the main characteristics of the CLEF-2007 test-collections. Section 3 outlines the main aspects of our stopword lists and 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.

2 Overview of the Test-Collections

The corpora used in our experiments include newspaper articles, namely *Magyar Hirlap* (2002, Hungarian), *Sega* (2002, Bulgarian), *Standart* (2002, Bulgarian), *Novinar* (2002, a new Bulgarian sub-collection in CLEF 2007), *Mladná fronta Dnes* (2002, Czech), *Lidove Noviny* (2002, Czech). As shown in Table 1, the Bulgarian corpus is relatively large compared to the others, both in size and in the number of documents. As for average article length, the Czech corpus is longer (212.6), while for the Bulgarian (135.9) and Hungarian (152.3) languages the lengths are relatively similar. It is interesting to note that even though the Hungarian collection is the smallest (105 MB), it contains a larger number of distinct indexing terms (191,738 computed after stemming) when compared to the Bulgarian and Czech corpuses.

During the indexing process we retained only the following logical sections from the original documents: <TITLE>, <LEAD>, and <TEXT>. From the topic descriptions we automatically removed certain phrases such as "Relevant document report …", "Подходящ е всеки документ" or "Keressünk olyan cikkeket, amelyek …", etc. All our runs were fully automatic.

As shown in the Appendix 2, the available topics cover various subjects (e.g., Topic #409: "Bali Car Bombing," Topic #414: "Beer Festivals," Topic #436: "VIP Divorces," or Topic #443: "World Swimming Records"), including both regional (Topic #445: "Prince Harry and Drugs") and more international coverage.

	Bulgarian	Hungarian	Czech
Size (in MB)	261 MB	105 MB	178 MB
# of documents	87,281	49,530	81,735
# of distinct terms	169,394	191,738	194,500
Number of distinct indexing terms per document			
Mean	99.5	105.4	117.7
Standard deviation	93.86	91.08	105.79
Median	70	75	90
Maximum	1,193	1,284	2,350
Minimum	0	2	1
Number of indexing terms per document			
Mean	135.9	152.3	212.6
Standard deviation	143.58	145.86	193
Median	91	102	160
Maximum	2,837	6,008	4,846
Minimum	0	5	1
Number of queries	50	50	50
Number rel. items	1,012	911	762
Mean rel./ request	20.24	18.22	15.24
Standard deviation	14.23	14.08	12.08
Median	17.5	14	10.5
Maximum	62 (T#438)	66 (T#415)	47 (T#415)
Minimum	2 (T#419)	1 (T#411)	2 (T#411)

Table 1: CLEF 2007 test-collection statistics

3 Stopword Lists and Stemming Procedures

During this evaluation campaign, our stopword list and stemmer for Hungarian were the same as that used in our CLEF 2006 participation (Savoy & Abdou, 2007). For this language our suggested stemmer mainly includes inflectional removals (gender, number and 23 grammatical cases, as for example in "házakat" → "ház" (house)) as well as some pronouns (e.g., "házamat" (my house) → "ház") and a few derivational suffixes (e.g., "temetés" (burial) → "temet" (to bury)). See Savoy (2007) for more information. Moreover, the Hungarian language uses compound constructions (e.g., "hétvégé" (weekend) = "hét" (week / seven) + "vég" (end)). In order to increase the matching possibilities between search keywords and document representations, we automatically decompounded Hungarian words using our decompounding algorithm (Savoy, 2004), leaving both compound words and their component parts in the documents and queries. The stopword list retained contains 737 words. The stemmer and stopword list are freely available www.unine.ch/info/clef.

For the Bulgarian language we decided to modify the transliteration procedure we used previously to convert Cyrillic characters into Latin letters. By correcting an error and adapting it for the new transliteration scheme, we modified last year's stemmer and denoted it the light Bulgarian stemmer. In this language, definite articles and plural forms are represented by suffixes and the general noun pattern is the following: <stem> <plural> <article>. Our light stemmer contains eight rules for removing plurals and five for removing articles. Additionally we applied seven grammatical normalization rules plus three others to remove palatalization (changing a stem's final consonant when followed by a suffix beginning with certain vowels), as is very common in most Slavic languages (see Appendix 3 for all the rules). We also proposed a new and more aggressive Bulgarian stemmer that also removes some derivational suffixes (e.g., "страшен" (fearfull) → "страх" (fear)). The stopword list used for this language contains 309 words, somewhat bigger than that of last year (258 items).

For the Czech language, we proposed a new stopword list containing 467 forms (determinants, prepositions, conjunctions, pronouns, and some very frequent verb forms). We also designed and implemented three Czech stemmers. The first one is a light stemmer that removes only those inflectional suffixes attached to nouns or adjectives in order to conflate to the same stem those morphological variations related to gender (feminine, neutral vs. masculine), number (plural vs. singular) and various grammatical cases (seven in the Czech language). For example, the noun "město" (city) appears as such in its singular form (nominative, vocative or accusative) but varies with other cases, "města" (genitive), "městu" (dative), "městem" (instrumental) or "městě" (locative). The corresponding plural forms are "města", "měst", "městům", "městy" or "městech". In the Czech language all nouns have a gender, and with a few exceptions (indeclinable borrowed words), they are declined for both number and case. For Czech nouns, the general pattern is the following: <stem> <possessive> <case> in which <case> ending includes both gender and number. Adjectives are declined to match the gender, case and number of the nouns to which they are attached. To remove these various case endings from nouns and adjectives we devised 52 rules, and then before returning the computed stem, we added five normalization rules in order to control palatalization and certain vowel changes in the basic stem (see Appendix 4 for all details).

Our second Czech stemmer denoted "light+" also includes rules for removing comparative forms from adjectives (e.g., "krásný", "krásnější", "nejkrásnější" → "krásn" (beautiful, more beautiful, the most beautiful)). We do not however expect this light stemmer variation to result in any significant changes in retrieval performance.

Finally, we designed and implemented a more aggressive stemmer that includes certain rules to remove frequently used derivational suffixes (e.g., "členství"(membership) → "člen"(member)). In applying this third more aggressive stemmer (denoted "derivational") we hope to improve mean average precision (MAP). Finally and unlike other languages, we do not remove the diacritics when building Czech stemmers.

4 IR models and Evaluation

4.1. Indexing and Searching Strategies

In order to obtain a high MAP values, we might adopt different weighting schemes applied to terms that occur in the documents or in the query. This weighting would allow us to account for term occurrence frequency (denoted tf_{ij} for indexing term t_j in document D_i), as well as their inverse document frequency (denoted idf_j). Moreover, we might normalize each indexing weight using the cosine to obtain the classical *tf·idf* formulation, rather than the more recent normalization approaches that account for document length.

In addition to this vector-space approach, we also considered probabilistic models such as the Okapi (or BM25) (Robertson *et al*. 2000). As a second probabilistic approach, we implemented three variants of the DFR (*Divergence from Randomness*) family of models 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:

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

As a first model, we implemented the PB2 scheme, defined by the following equations:

$$Inf^1_{ij} = -\log_2[(e^{-\lambda_j} \cdot \lambda_j^{tf_{ij}})/tf_{ij}!] \quad \text{with } \lambda_j = tc_j / n \tag{1}$$

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

where tc_j indicates the number of occurrences of term t_j in the collection, l_i the length (number of indexing terms) of document D_i, *mean dl* the average document length, *n* the number of documents in the corpus, and *c* a constant (the corresponding values are given in the Appendix 1).

For the second model called GL2, the implementation of $Prob^1_{ij}$ is given by Equation 3, and $Prob^2_{ij}$ is given by Equation 4, as follows:

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

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

where λ_j and tfn_{ij} were defined previously.

For the third model called IneC2, the implementation is given by the following two equations:

$$Inf^1_{ij} = tfn_{ij} \cdot \log_2[(n+1) / (n_e+0,5)] \quad \text{with } n_e = n \cdot [1 - [(n-1)/n]^{tc_j}] \quad (5)$$

$$Prob^2_{ij} = 1 - [(tc_j +1) / (df_j \cdot (tfn_{ij}+1))] \quad (6)$$

where *n*, tc_j and tfn_{ij} were defined previously, and df_j indicates the number of documents in with the term t_j occurs.

Finally, we also considered an approach based on a statistical 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 (e.g., as in Equation 1 or 3), but rather be estimated directly based on occurrence frequencies in document D_i or corpus C. Within this language model paradigm, various implementations and smoothing methods might be considered, although in this study we adopted a model proposed by Hiemstra (2002), as described in Equation 7, combining an estimate based on document ($P[t_j | D_i]$) and on corpus ($P[t_j | C]$).

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

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

4.2. Overall Evaluation

To measure retrieval performance, we adopted MAP values computed on the basis of 1,000 retrieved items per request as calculated with the new TREC-EVAL program. Using this evaluation tool, some evaluation differences may occur in the values computed according to the official measure (the latter always takes 50 queries into account while in our presentation we do not account for queries having no relevant items). 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.

	Mean average precision					
Query Stemmer / indexing unit Model \ # of queries	Bulgarian TD light / word 50 queries	Bulgarian TDN light / word 50 queries	Bulgarian TD deriv./word 50 queries	Bulgarian TDN deriv./word 50 queries	Bulgarian TD none/4-gram 50 queries	Bulgarian TDN none/4-gram 50 queries
Okapi	0.3155	0.3462	0.3425	0.3720	0.3022	0.3342
DFR GL2	0.3307	0.3653	0.3541	0.3909	0.3100	0.3250
DFR PB2	0.3266	0.3476	0.3394	0.3637	0.2960	0.3116
DFR IneC2	**0.3423**	**0.3696**	**0.3606**	**0.3862**	**0.3156**	**0.3409**
LM (λ=0.35)	0.3175	0.3580	0.3368	0.3782	0.2868	0.3294
tf · idf	0.2103	0.2264	0.2143	0.2293	0.2105	0.2271
Average	0.3265	0.3573	0.3467	0.3782	0.3021	0.3282
% change over TD		+9.4%		+9.09%		+8.6%
% change	-5.8%		baseline		-12.9%	

Table 2: MAP of various IR models and query formulations (Bulgarian language)

Table 2 shows the MAP achieved by various probabilistic models using the Bulgarian collection with two different query formulations (TD or TDN) and the two stemmers. The last two columns show the MAP achieved by using a 4-gram indexing scheme (without applying a stemming approach). An analysis of this data

shows that the best performing IR model corresponds to the DFR IneC2 model with all stemming approaches or query sizes.

In the last lines we reported the MAP average over these 5 IR models together with percentage of variation compared to the medium (TD) query formulation or to the derivational stemmer (TD query). As depicted in the last lines, increasing the query size improves the MAP (around +9%). According to the average performance, the best indexing approach seems to be a word-based approach using our derivational stemmer. In this case, the MAP with TD query formulation is, in average, 0.3467 vs. 0.3021 for the 4-gram approach, a relative difference of 12.9%. The performance difference with the light stemmer is smaller in average (0.3467 vs. 0.3265), a relative difference of 5.8%.

	Mean average precision					
Query Indexing unit Model \ # of queries	Hungarian TD decompound 50 queries	Hungarian TDN decompound 50 queries	Hungarian TD word 50 queries	Hungarian TDN word 50 queries	Hungarian TD 4-gram 50 queries	Hungarian TDN 4-gram 50 queries
Okapi	0.3629	0.3959	0.3255	0.3763	0.3445	0.3797
DFR GL2	0.3615	0.3994	0.3324	0.3809	0.3495	0.3702
DFR PB2	0.3799	0.4106	0.3428	0.3910	0.3355	0.3599
DFR IneC2	**0.3897**	**0.4271**	**0.3525**	**0.4031**	**0.3527**	**0.3828**
LM (λ=0.35)	0.3482	0.3921	0.3118	0.3669	0.3153	0.3555
tf · idf	0.2532	0.2887	0.2344	0.2806	0.2345	0.2506
Average	0.3492	0.3856	0.3166	0.3665	0.3220	0.3498
% change over TD		+10.4%		+15.8%		+8.6%
% change	baseline		-9.4%		-7.8%	

Table 3: MAP of various IR models and query formulations (Hungarian language)

Table 3 reports the evaluations done with the Hungarian language (word-based and 4-gram indexing) and with the classical *tf idf* vector-space scheme. For the most part the same conclusions can be drawn for this language as those shown for Bulgarian (Table 2). Firstly, the DFR In2C2 probabilistic model provides the best IR performance and secondly when compared to the TD query formulation the retrieval effectiveness is improved (around 11.6%). As depicted in the last three lines, the best indexing strategy seems to be a word-based approach with an automatic decompounding procedure. Using this strategy as baseline and with TD query formulation, the average performance difference with an indexing strategy without a decompounding procedure is around 9.4% (0.3492 vs. 0.3166), while a 4-gram indexing scheme depicts an average MAP of 0.3220 having a percentage of degradation of around 7.8%.

	Mean average precision					
Query Stemmer Model \ # of queries	Czech TD light 50 queries	Czech TDN light 50 queries	Czech TD light+ 50 queries	Czech TD 4-gram 50 queries	Czech TD derivational 50 queries	Czech TDN derivational 50 queries
Okapi	0.3355	0.3616	0.3255	**0.3401**	0.3255	0.3669
DFR GL2	**0.3437**	**0.3678**	**0.3323**	0.3365	**0.3342**	**0.3678**
DFR PB2	0.3233	0.3434	0.3144	0.3188	0.3164	0.3472
LM (λ=0.35)	0.3263	0.3626	0.3182	0.3204	0.3109	0.3594
tf · idf	0.2050	0.2338	0.2105	0.2126	0.1984	0.2303
Average	0.3068	0.3338	0.3002	0.3057	0.2971	0.3343
% change over TD		+8.83%				+12.54%
% change	baseline		-2.14%	-0.35%	-3.16%	

Table 4: MAP of various IR models and query formulations (Czech language)

The evaluations done on the Czech language are depicted in Table 4. In this case, we compared three stemmers and the 4-gram indexing approach (without stemming). The best performing IR models corresponds to either the DFR GL2 or the Okapi probabilistic model. The performance differences between these two IR models are usually rather small.

As shown in the last three lines of Table 4, the best indexing strategy seems to be the word-based indexing strategy using the light stemming approach. As expected, performance differences between the "light" and "light+" stemmers are rather small (2.14% when using the TD query formulation). Moreover, the performance differences between the 4-gram and the light stemming approach seem to be statistically not significant (in

average, 0.3068 vs. 0.3057 with TD query formulation). As for the other corpora, increasing the query size improves the MAP (around +10%).

An analysis showed 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 (denoted "Roc") (Buckley *et al.*, 1996) with $\alpha = 0.75$, $\beta = 0.75$, whereby the system was allowed to add *m* terms extracted from the *k* best ranked documents from the original query. From our previous experiments we learned that this type of blind query expansion strategy does not always work well. More particularly, we believe that including terms occurring frequently in the corpus (because they also appear in the top-ranked documents) may introduce more noise, and thus be an ineffective means of discriminating between relevant and non-relevant items (Peat & Willett, 1991). Consequently we chose to also apply our *idf*-based query expansion model (denoted "idf" in Tables 9 and 10) (Abdou & Savoy, 2007b).

To evaluate these propositions, we applied certain probabilistic models and enlarged the query by the 20 to 150 terms (indexing words or *n*-grams) retrieved from the 3 to 10 best-ranked articles within the Bulgarian (Table 5), Hungarian (Table 6) and Czech corpora (Table 7).

	Mean average precision			
Query TD PRF using Rocchio	Bulgarian derivational	Bulgarian derivational	Bulgarian none / 4-gram	Bulgarian derivational
IR Model / MAP	Okapi 0.3425	DFR IneC2 0.3606	Okapi 0.3022	LM 0.3368
k doc. / *m* terms	10/50 **0.3574** 10/80 0.3548 10/100 0.3559 10/120 0.3565	10/50 0.3860 10/80 0.3865 10/100 0.3870 10/120 **0.3896**	3/80 0.3065 3/100 0.3121 3/120 **0.3177** 3/150 0.3169	10/50 **0.4098** 10/80 0.4043 10/100 0.4061 10/120 0.4004

Table 5: MAP using blind-query expansion (Bulgarian collection)

	Mean average precision			
Query TD PRF using Rocchio	Hungarian decompound	Hungarian decompound	Hungarian none / 4-gram	Hungarian decompound
IR Model / MAP	Okapi 0.3629	DFR IneC2 0.3897	Okapi 0.3445	LM 0.3921
k doc. / *m* terms	5/20 0.3909 5/50 0.3973 5/70 0.3983 5/100 **0.4010**	5/20 0.4193 5/50 0.4284 5/70 0.4283 5/100 **0.4298**	3/80 0.3654 3/100 0.3719 3/120 0.3752 3/150 **0.3785**	5/20 0.4309 5/50 0.4263 5/70 0.4315 5/100 **0.4323**

Table 6: MAP using blind-query expansion (Hungarian collection)

For the Bulgarian corpus (Table 5), enhancement increased from +1.47% (4-gram, Okapi, 0.3022 vs. 0.3065) to +21.7% (LM model, 0.3368 vs. 0.4098). For the Hungarian collection (Table 6), percentage improvement varied from +6.1% (4-gram, Okapi model, 0.3445 vs. 0.3654) to +10.1% (LM model, 0.3913 vs. 0.4323). For the Czech language (Table 7), the percentages of variation range from -2.6% (4-gram, Okapi model, 0.3401 vs. 0.3314) to +21.6% (DFR GL2 model, 0.3437 vs. 0.4179).

	Mean average precision			
Query TD PRF using Rocchio	Czech light / word	Czech light / word	Czech none / 4-gram	Czech none / 4-gram
IR Model / MAP	Okapi 0.3355	DFR GL2 0.3437	Okapi 0.3401	LM 0.3204
k doc. / *m* terms	5/20 0.3560 5/50 0.3605 5/70 0.3614 5/100 **0.3636**	5/20 0.4131 5/50 0.4158 5/70 0.4154 5/100 **0.4179**	5/20 0.3314 5/50 0.3501 5/70 0.3672 5/100 **0.3710**	5/20 0.3457 5/50 0.3765 5/70 0.3754 5/100 **0.3823**

Table 7: MAP using blind-query expansion (Czech collection)

5 Data Fusion

It is assumed that combining different search models should improve retrieval effectiveness, due to the fact that each document representation might not retrieve the same pertinent items and thus increase the overall recall (Vogt & Cottrell, 1999). In this current study we combined three probabilistic models representing both the

parametric (Okapi and DFR) and non-parametric (language model or LM) approaches. On the other hand, we also combined both word-based and *n*-gram indexing strategies. To perform such combination we evaluated various fusion operators (see Table 8 for a detailed list of their descriptions). The "Sum RSV" operator for example 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 8 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 certain 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	$\alpha_i \cdot [((RSV_k - Mean^i) / Stdev^i) + \delta^i]$ with $\delta^i = [(Mean^i - Min^i) / Stdev^i]$

Table 8: 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 each individual list and removed any duplicates, retaining only the highest ranking occurrence. Finally we suggest merging the retrieved documents according to the Z-Score, computed for each result list. Within this scheme, for each *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 8, Min^i (Max^i) lists 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)			
Language / Query Model	Bulgarian TD 50 queries	Bulgarian TDN 50 queries	Hungarian TD 50 queries	Czech TD 50 queries
LM & PRF doc/term	Roc 10/50 **0.4098**	Roc 10/50 **0.4418**	Roc 5/70 0.4315	idf 5/20 0.4070
Okapi & PRF doc/term	Roc 3/150 0.3169	Roc 3/150 0.3406	idf 3/120 0.4233	Roc 5/70 0.3672
DFR & PRF doc/term	idf 5/60 0.3750	idf 5/60 0.4038	idf 5/100 **0.4376**	Roc 5/50 **0.4085**
Official run name	UniNEbg1	UniNEbg4	UniNEhu2	UniNEcz3
Round-robin	0.3747 (-8.6.%)	0.4038 (-8.6%)	0.4396 (+0.5%)	0.4136 (+1.2%)
Sum RSV	0.3841 (-6.3%)	0.4171 (-5.6%)	0.4677 (+6.9%)	0.3987 (-2.4%)
Norm Max	0.4076 (-0.5%)	0.4403 (-0.3%)	**0.4738** (+8.3%)	0.4131 (+1.1%)
Norm RSV	0.4069 (-0.7%)	0.4404 (-0.3%)	0.4726 (+8.0%)	0.4139 (+1.3%)
Z-Score	**0.4128** (+0.7%)	**0.4422** (+0.1%)	0.4716 (+7.8%)	**0.4225** (+3.4%)

Table 9: Mean average precision using different combination operators (with blind-query expansion)

Table 9 depicts the evaluation of various data fusion operators, comparing them to the single approach using the language model (LM), Okapi or the DFR probabilistic models (PB2 or GL2). From this data, we can see that combining three IR models might improve retrieval effectiveness, only slightly for the Bulgarian collection, moderately for the Czech and noticeably for the Hungarian corpus. When combining different retrieval models, the Z-Score scheme tended to perform the best, or at least it had one of the best performing MAP (e.g., for the Hungarian corpus). Except for the Hungarian corpus, when compared to the best single search model, the performance achieved by the various data fusion approaches did not seem statistically significant.

6 Official Results

Table 10 shows the exact specifications of our 12 official monolingual runs, based mainly on the probabilistic models (Okapi, DFR and statistical language model (LM)). For all languages we submitted three runs with the TD query formulation and one with the TDN. All runs are fully automatic and the same data fusion approach (Z-score) was applied in all cases. For the Hungarian corpus however we sometimes applied our decompounding approach (denoted by "dec" in the "Index" column)

Run name	Query	Index	Stem	Model	Query expansion	Single MAP	Comb MAP
UniNEbg1	TD	4-gram	none	Okapi	Roc 3 docs / 150 terms	0.3169	Z-score
BG	TD	word	light	PB2	idf 5 docs / 60 terms	0.3750	**0.4128**
	TD	word	deriva.	LM	Roc 10 docs / 50 terms	0.4098	
UniNEbg2	TD	word	deriva.	LM	Roc 10 docs / 120 terms	0.4004	Z-Score
BG	TD	word	light	IneC2	idf 5 docs / 60 terms	0.3740	0.4108
UniNEbg3	TD	4-gram	none	LM	idf 3 docs / 120 terms	0.3336	Z-Score
BG	TD	word	light	LM	Roc 5 docs / 40 terms	0.3624	0.3999
	TD	word	deriva.	LM	idf 10 docs / 50 terms	0.4013	
UniNEbg4	TDN	4-gram	none	Okapi	Roc 3 docs / 150 terms	0.3406	Z-score
BG	TDN	word	light	PB2	idf 5 docs / 60 terms	0.4038	0.4422
	TDN	word	deriva.	LM	Roc 10 docs / 50 terms	0.4418	
UniNEhu1	TD	dec	stem	LM	Roc 5 docs / 100 terms	0.4323	Z-score
HU	TD	word	stem	GL2	Roc 5 docs / 70 terms	0.4375	0.4606
	TD	4-gram	none	PB2	idf 3 docs / 80 terms	0.3886	
UniNEhu2	TD	dec	stem	LM	Roc 5 docs / 70 terms	0.4315	Z-score
HU	TD	word	stem	GL2	idf 5 docs / 100 terms	0.4376	**0.4716**
	TD	4-gram	none	Okapi	idf 3 docs / 120 terms	0.4233	
UniNEhu3	TD	4-gram	none	LM	idf 3 docs / 120 terms	0.3842	Z-score
HU	TD	word	stem	GL2	Roc 5 docs / 100 terms	0.4379	0.4586
	TD	dec	stem	PB2	idf 5 docs / 20 terms	0.4366	
UniNEhu4	TDN	dec	stem	LM	Roc 5 docs / 100 terms	0.4604	Z-score
HU	TDN	word	stem	GL2	Roc 5 docs / 70 terms	0.4664	0.4773
	TDN	4-gram	none	PB2	idf 3 docs / 80 terms	0.4108	
UniNEcz1	TD	word	light+	Okapi	idf 5 docs / 20 terms	0.4013	Z-score
CZ	TD	word	deriva.	LM	Roc 5 docs / 50 terms	0.4002	0.4167
UniNEcz2	TD	word	light	Okapi	Roc 5 docs / 20 terms	0.3560	Z-score
CZ	TD	4-gram	none	GL2	idf 5 docs / 70 terms	0.3798	0.4134
	TD	word	light+	PB2	Roc 5 docs / 50 terms	0.3632	
UniNEcz3	TD	word	light	LM	idf 5 docs / 20 terms	0.4070	Z-score
CZ	TD	4-gram	none	Okapi	Roc 5 docs / 70 terms	0.3672	**0.4225**
	TD	word	light+	GL2	Roc 5 docs / 50 terms	0.4085	
UniNEcz4	TDN	word	deriva.	Okapi	Roc 5 docs / 20 terms	0.3627	Z-score
CZ	TDN	4-gram	none	LM	Roc 5 docs / 100 terms	0.3953	0.4242
	TDN	word	light+	GL2	idf 5 docs / 50 terms	0.4048	

Table 10: Description and mean average precision (MAP) of our official monolingual runs

7 Conclusion

In this eighth CLEF evaluation campaign we evaluated various probabilistic IR models using three different test-collections written in three different East European languages, namely the Hungarian, Bulgarian and Czech languages. We suggested a new stemmer for the Bulgarian language that removed some very frequent derivational suffixes. For the Czech language, we designed and implemented three different stemmers.

Our various experiments tend to demonstrate that the Okapi model or the IneC2 model derived from *Divergence from Randomness* (DFR) paradigm tend to produce the best overall retrieval performances (see Tables 2 to 4). The statistical language model (LM) used in our experiments usually results in retrieval performance inferior to that obtained with the Okapi or DFR approach.

For the Bulgarian language (Table 2), our new and more aggressive stemmer tends to produce a better MAP when compared to a light stemming approach (5.8% in relative difference) and better than the 4-gram indexing scheme (-12.9%). For the Hungarian language (Table 3), applying an automatic decompounding procedure seems to improve the MAP around 9.4% when compared to a word-based approach, or around 7.8% when compared to a 4-gram indexing scheme. For the Czech language however performance differences between a light (inflectional only) and a more aggressive stemmer removing both inflectional and some derivational suffixes were rather small (Table 4). Moreover, the performance differences were also small when compared to those achieved with a 4-gram approach. Pseudo-relevance feedback (Rocchio's model) improves the MAP

depending on the parameter settings (Tables 5 to 7). A data fusion strategy may clearly enhance the retrieval performance for the Hungarian language (Table 8) and slightly for the two other languages.

Acknowledgments

The authors would like to also thank the CLEF-2007 task organizers for their efforts in developing various European language test-collections. The authors would also thank Samir Abdou for his help during the implementations of the different stemmers within the Lucene system. This research was supported in part by the Swiss National Science Foundation under Grant #200021-113273.

References

Abdou S. & Savoy J. (2007a). Monolingual experiments with Far-East Languages in NTCIR-6. In *Proceedings NTCIR-6*, Tokyo: NII publication (National Institute of Informatics), 52-59.

Abdou S. & Savoy J. (2007b). Searching in Medline: Stemming, query expansion, and manual indexing evaluation. Information Processing & Management, to appear.

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

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

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.

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

Peat, H. J. & Willett, P. (1991). The limitations of term co-occurrence data for query expansion in document retrieval systems. *Journal of the American Society for Information Science*, 42(5), 378-383

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

Savoy, J. (2005). 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. (2007). Searching strategies for the Hungarian language. *Information Processing & Management*, to appear.

Savoy J. & Abdou S. (2007). Experiments with monolingual, bilingual, and robust retrieval. In C. Peters, F.C. Gey, J. Gonzalo, H. Müller, G.J.F. Jones, M. Kluck, B. Magnini & M. de Rijke (Eds.). Lecture Notes in Computer Science. Berlin: Springer-Verlag, Berlin, to appear.

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

Appendix 1: Parameter Settings

	Okapi			DFR	
Language	*b*	k_1	*avdl*	*c*	*mean dl*
Czech	0.75	1.2	213	1.5	213
Bulgarian	0.85	1.2	135	1.5	135
Hungarian	0.75	1.2	152	1.5	152

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

Appendix 2: Topic Titles

C401	Euro Inflation	C426	9/11 Counterterrorism Measures
C402	Renewable Energy Sources	C427	Testimony against Milosevic
C403	Acting as a Cop	C428	Ecological Tourism
C404	NATO Summit Security	C429	Water Health Risks
C405	Childhood Asthma	C430	Cosmetic Procedures
C406	Animated Cartoons	C431	French Presidential Candidates
C407	Australian Prime Minister	C432	Zimbabwe Presidential Elections
C408	Human Cloning	C433	Child Abuse by Priests
C409	Bali Car Bombing	C434	Political Instability in Venezuela
C410	North Korea Nuclear Weapons Violation	C435	Causes of Air Pollution
C411	Best Picture Oscar	C436	VIP Divorces
C412	Books on Politicians	C437	Enron Auditing Irregularities
C413	Reducing Diabetes Risk	C438	Cancer Research
C414	Beer Festivals	C439	Accidents at Work
C415	Drug Abuse	C440	Winter Olympics Doping Scandal
C416	Moscow Theatre Hostage Crisis	C441	Space Tourists
C417	Airplane Hijacking	C442	Queen Mother's Funeral
C418	Bülent Ecevit's Statements	C443	World Swimming Records
C419	Nuclear Waste Repositories	C444	Brazil World Soccer Champions
C420	Obesity and Ill-health	C445	Prince Harry and Drugs
C421	Kostelic Olympic Medals	C446	Flood damage to cultural heritage
C422	Industrial and Business Closures	C447	Pim Fortuyn's Politics
C423	Alternatives to Flu Shots	C448	Nobel Prizes for Chemistry
C424	Internet Banking Increase	C449	Civil Wars in Africa
C425	Endangered Species	C450	Failed Assassination Attempts

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

Appendix 3: Bulgarian Stemmer

```
BulgarianStemmer (word)  {
  RemoveArticle(word);
  RemovePlural(word);
  Normalize(word);
  Palatalization(word)
  return;
  }

RemoveArticle(word)  {
  if (word ends with "-ът") then remove "-ът" return;          # masculine
  if (word ends with "-ят") then                               # masculine
        if (word ends with " V+ят") then replace by "-й"       # V –any vowel
              else remove "-ят" return;
  if (word ends with "-то") then remove "-то" return;          # neutral
  if (word ends with "-те") then remove "-те" return;          # neutral
  if (word ends with "-та") then remove "-та" return;          # feminine
  return;
  }

RemovePlural(word)  {
  if (word ends with "-ища") then remove "-ища" return;        # for adjectives
  if (word ends with "-ище") then remove "-ище" return;        # for adjectives
  if (word ends with "-овци") then replace by "-о" return;     # for adjectives
  if (word ends with "-евци") then replace by "-е" return;     # for adjectives
```

```
  if (word ends with "-ове") then remove "-ове" return;          # masculine
  if (word ends with "-еве") then                                # masculine
        if (word ends with " V+ еве") then replace by "-й"
              else remove "-еве" return;
  if (word ends with "-та") then remove "-та" return;            # feminine
  if (word ends with "-..е.и") then replace by "-.я." return;    # rewriting rule
  return;                                                        # with . any character
  }


Normalize(word)  {
  if (word ends with "-еи" or "-ии") then remove "-еи" or "-ии";
  if (word ends with "-я") then                                  # normalize
        if (word ends with " V+ я") then replace by "-й"         # adjectives
              else remove "-я";
  if (word ends with "-[аой]") then remove "-[аой]";
  if (word ends with "-[еи]") then remove "-[еи]";
  if (word ends with "-йн") then replace by "-н" return;         # rewriting rule
  if (word ends with "-LeC") then replace by "-LC";              # L-any letter
  if (word ends with "-LъL") then replace by "-LL";              # C-any consonant
  return;
  }

Palatalization(word)  {
  if (word ends with "-ц" or "-ч") then replace by "-к" return;
  if (word ends with "-з" or "-ж") then replace by "-г" return;
  if (word ends with "-с" or "-ш") then replace by "-х" return;
  return;
  }
```

Table A.3: Our new light Stemmer for the Bulgarian language

Appendix 4: Czech Stemmer

```
CzechStemmer (word)  {
  RemoveCase (word);
  RemovePossessives (word);
  Normalize (word);
  return;
  }

RemovePossessives(word)  {
  if (word ends with "-ov") then remove "-ov" return;
  if (word ends with "-in") then remove "-in" return;
  if (word ends with "-ův") then remove "-ův" return;
  return;
  }

Normalize(word)  {
  if (word ends with "čt") then replace by  "ck" return;
  if (word ends with "št") then replace by  "sk" return;
  if (word ends with "c" or "č") then replace by  "k" return;
  if (word ends with "z" or "ž") then replace by  "h" return;
  if (word ends with ".ů.") then replace by ".o." return;
  return;
  }
```

```
RemoveCase(word)  {
  if (word ends with “-atech”) then remove “-atech” return;
  if (word ends with “-ětem”) then remove “-ětem” return;
  if (word ends with “-etem”) then remove “-etem” return;
  if (word ends with “-atům”) then remove “-atům” return;
  if (word ends with “-ech”) then remove “-ech” return;
  if (word ends with “-ich”) then remove “-ich” return;
  if (word ends with “-ích”) then remove “-ích” return;
  if (word ends with “-ého”) then remove “-ého” return;
  if (word ends with “-ěmi”) then remove “-ěmi” return;
  if (word ends with “-emi”) then remove “-emi” return;
  if (word ends with “-ému”) then remove “-ému” return;
  if (word ends with “-ěte”) then remove “-ěte” return;
  if (word ends with “-ete”) then remove “-ete” return;
  if (word ends with “-ěti”) then remove “-ěti” return;
  if (word ends with “-eti”) then remove “-eti” return;
  if (word ends with “-ího”) then remove “-ího” return;
  if (word ends with “-iho”) then remove “-iho” return;
  if (word ends with “-ími”) then remove “-ími” return;
  if (word ends with “-ímu”) then remove “-ímu” return;
  if (word ends with “-imu”) then remove “-imu” return;
  if (word ends with “-ách”) then remove “-ách” return;
  if (word ends with “-ata”) then remove “-ata” return;
  if (word ends with “-aty”) then remove “-aty” return;
  if (word ends with “-ých”) then remove “-ých” return;
  if (word ends with “-ama”) then remove “-ama” return;
  if (word ends with “-ami”) then remove “-ami” return;
  if (word ends with “-ové”) then remove “-ové” return;
  if (word ends with “-ovi”) then remove “-ovi” return;
  if (word ends with “-ými”) then remove “-ými” return;
  if (word ends with “-em”) then remove “-em” return;
  if (word ends with “-es”) then remove “-es” return;
  if (word ends with “-ém”) then remove “-ém” return;
  if (word ends with “-ím”) then remove “-ím” return;
  if (word ends with “-ům”) then remove “-ům” return;
  if (word ends with “-at”) then remove “-at” return;
  if (word ends with “-ám”) then remove “-ám” return;
  if (word ends with “-os”) then remove “-os” return;
  if (word ends with “-us”) then remove “-us” return;
  if (word ends with “-ým”) then remove “-ým” return;
  if (word ends with “-mi”) then remove “-mi” return;
  if (word ends with “-ou”) then remove “-ou” return;
  if (word ends with “-[aeiouyáéíýě]”) then remove “-[aeiouyáéíýě]” return;
  return;
  }
```

Table A.4: Our light+ stemmer for the Czech language