Published in Lecture Notes in Computer Science 5152, 37-44, 2008
which should be used for any reference to this work

Stemming Approaches for East European Languages

Ljiljana Dolamic and Jacques Savoy

Computer Science Department, University of Neuchatel, Rue Emile Argand 11, 2009 Neuchatel, Switzerland
Ljiljana.Dolamic@unine.ch, Jacques.Savoy@unine.ch

Abstract. During this CLEF evaluation campaign, the first objective is to propose and evaluate various indexing and search strategies for the Czech language that will hopefully result in more effective retrieval than language-independent approaches (n-gram). Based on the stemming strategy we developed for other languages, we propose that for the Slavic language a light stemmer (inflectional only) and also a second one based on a more aggressive suffix-stripping scheme that will remove some derivational suffixes. Our second objective is to undertake further study of the relative merit of various search engines when exploring Hungarian and Bulgarian documents. To evaluate these solutions we use various effective IR models. Our experiments generally 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 of a light and a more aggressive stemmer to remove both inflectional and some derivational suffixes, reveals only small performance differences. For this language only, performance differences between a word-based or a 4-gram indexing strategy are also rather small.

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 [1], [2] and popular Asian [3] languages (namely, Chinese, Japanese, and Korean). The main objective of our work has been to promote effective monolingual IR in these languages. For our participation in the CLEF 2007 evaluation campaign we thus decided to revamp our stemming strategy by including certain very frequently used derivational suffixes. When defining our stemming rules however we still focus on nouns and adjectives only. A description of the test-collections can be found in [4].

The rest of this paper is organized as follows: Section 2 outlines the main aspects of our stopword lists and stemming procedures. Section 3 analyses the principal features of different indexing and search strategies while Section 4 evaluates their use with the available corpora. Finally, Section 5 exposes our official results and Section 6 depicts our main findings.

2 Stemming Procedures

For the Hungarian language our suggested stemmer [5] mainly involves inflectional removal (gender, number and 23 grammatical cases, as for example in "házakat" → "ház" (house)) and also some pronouns (e.g., "házamat" (my house) → "ház") and a few derivational suffixes (e.g., "temetés" (burial) → "temet" (to bury)). Because the Hungarian language uses compound constructions (e.g., "hétvége" (weekend) = "hét" (week / seven) + "vég" (end)), we increase matching possibilities between search keywords and document representations by automatically decompounded Hungarian words. To do so we apply our decompounding algorithm, leaving both compound words and their component parts in documents and queries. All stopword lists (containing 737 Hungarian forms) and stemmers used in this experiment are freely available at `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. We also modified last year's stemmer, denoted as the light Bulgarian stemmer, by correcting an error and adapting it for the new transliteration scheme [2]. In this language, definite articles and plural forms are represented by suffixes and the general noun pattern is as follows: <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 stem's final consonant when followed by a suffix beginning with certain vowels), as is very common in most Slavic languages. We also proposed a new and more aggressive Bulgarian stemmer that removes some derivational suffixes (e.g., "страшен" (fearful) → "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 two 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 as follows: <stem> <possessive> <case> in which <case> ending includes both gender and number. Adjectives are declined to match the gender, case and number of 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 that control palatalization and certain vowel changes in the basic stem.

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 second more aggressive stemmer (denoted "derivational") we hope to improve mean average precision (MAP). Finally and unlike other languages, we do not remove the diacritic characters when building Czech stemmers.

3 Indexing and Searching Strategies

In order to obtain high MAP values, we considered adopting different weighting schemes for terms occurring in the documents or in the query. With this weighting we could 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 also considered normalize each indexing weight, using the cosine to obtain the classical $tf \cdot idf$ formulation.

In addition to this vector-space approach, we considered probabilistic models such as the Okapi [6] (or BM25). As a second probabilistic approach, we implemented three variants of the DFR (*Divergence from Randomness*) family of models suggested by Amati & van Rijsbergen [7]. Within this framework, indexing weights w_{ij} attached to term t_j in document D_i combine two information measures, expressed as follows:

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

As a first model, we implemented the GL2 scheme, defined as:

$$Prob^1_{ij} \;=\; \left[\frac{1}{1+\lambda_j}\right] \cdot \left[\frac{\lambda_j}{1+\lambda_j}\right]^{tfn_{ij}} \qquad \text{with } \lambda_j = \frac{tc_j}{n} \tag{2}$$

$$Prob^2_{ij} \;=\; \frac{tfn_{ij}}{tfn_{ij}+1} \;\text{ with } tfn_{ij} \;=\; tf_{ij} \cdot -\log_2\left[1 + \frac{c \cdot mean\ dl}{l_i}\right] \tag{3}$$

where df_j indicates the number of documents in which term t_j occurs, tc_j 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.

As a second model, we implemented the PB2 scheme, defined as:

$$Inf^1_{ij} \;=\; -\log_2\left[\frac{e^{-\lambda_j} \cdot \lambda_j^{tf_{ij}}}{tf_{ij}!}\right] \tag{4}$$

$$Prob^2_{ij} \;= 1 - \left[\frac{tc_j + 1}{df_j \cdot (tfn_{ij}+1)}\right] \tag{5}$$

We then implemented a third model called IneC2 as follows:

$$Inf^1_{ij} \;=\; tfn_{ij} \cdot \left[\frac{n+1}{n_e + 0.5}\right] \qquad \text{with } n_e \;=\; n \cdot \left[1 - \left(\frac{n-1}{n}\right)^{tc_j}\right] \tag{6}$$

$$Prob_{ij}^2 \;=\; 1 - \left[\frac{tc_j + 1}{df_j \cdot (tfn_{ij} + 1)}\right] \tag{7}$$

Finally, we considered an approach known as a non-parametric probabilistic model, based on a statistical language model (LM) [8]. As such, probability estimates would not be based on any known distribution (e.g., as in Equation 2), but rather be estimated directly, based on occurrence frequencies in document D_i or corpus C. Within this language model paradigm, various implementation and smoothing methods could be considered, although in this study we adopted a model proposed by Hiemstra [8], as described in Equation 8, combining an estimate based on document ($P[t_j|D_i]$) and on corpus ($P[t_j|C]$).

$$Prob[D_i|Q] \;=\; Prob[D_i] \cdot \prod_{tj \in Q} [\lambda_j \cdot Prob[t_j|D_i] + (1 - \lambda_j) \cdot Prob[t_j|C]] \tag{8}$$

$$Prob[t_j|D_i] \;=\; tf_{ij}/l_i \quad \text{and} \quad Prob[t_j|C] \;=\; df_j/lc \quad \text{with} \quad lc = \sum_k df_k \tag{9}$$

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

4 Evaluation

To measure the retrieval performance, we chose to use the mean average precision (MAP) obtained from 50 queries. In the following tables, the best performances under a given condition are listed in bold type. We then applied the bootstrap methodology [9] in order to statistically determine whether or not a given search strategy would be better than the performance depicted in bold. Thus, in the tables included in this paper we added an asterisk to indicate any statistically significant differences resulting from the use of a two-sided non-parametric bootstrap test ($\alpha = 5\%$).

Table 1 shows the MAP achieved by various probabilistic models using the Hungarian and Bulgarian collection, along with two different stemmers. An analysis of this data shows that the best performing IR model corresponds to the

Table 1. Evaluation of Hungarian and Bulgarian corpora

	Mean average precision			
Query Stemmer	Hungarian TD light	Hungarian TD + decomp.	Bulgarian TD light	Bulgarian TD derivat.
Okapi	0.3231*	0.3629*	0.3155*	0.3425*
DFR-GL2	0.3324*	0.3615*	0.3307	0.3541
DFR-IneC2	**0.3525**	**0.3897**	**0.3423**	**0.3606**
LM	0.3118*	0.3482*	0.3175*	0.3368*
tf idf	0.2344*	0.2532*	0.2103*	0.2143*

DFR-IneC2 model, with all stemming approaches and for both languages. For the Hungarian language, the best indexing strategy seems to be a word-based approach along with an automatic decompounding procedure. Using this strategy as a baseline, the average performance difference with an indexing strategy without a decompounding procedure is around 13% (DFR-IneC2: 0.3525 vs. 0.3897).

The evaluations done on the Czech language are depicted in Table 2. In this case, we compared two stemmers (light vs. derivational) and the 4-gram indexing approach (without stemming) [10]. The best performing IR model type is the DFR-IneC2 but the performance differences between the two DFR models are usually small. In the third column (labeled "no accent") we evaluated the light stemmer, with all diacritic characters removed, and thus slightly reduced retrieval performance. When comparing the stemmers, the best indexing strategy seem to be the word-based indexing strategy, using the light stemming approach. Moreover, the performance differences between the 4-gram and this light stemming approach seem to be statistically not significant.

Table 2. Evaluation of the Czech Corpus

	Mean average precision			
Query Stemmer	TD light	TD no accent	TD derivat.	TD 4-grams
Okapi	0.3355	0.3306*	0.3255*	0.3401*
DFR-GL2	0.3437	0.3359	0.3342	0.3365
DFR-IneC2	**0.3539**	**0.3473**	**0.3437**	**0.3517**
LM	0.3263*	0.3174*	0.3109*	0.3304*
tf idf	0.2050*	0.2078*	0.1984*	0.2126*

A query-by-query analysis reveals that our various search strategies encountered some serious problems. For example with the Hungarian corpus, Topic #436 "VIP divorces" resulted in an average precision of 0.0003 because the term "VIP" is unknown in the collection and thus the query is composed of only a single and frequent word. With the Bulgarian corpus, Topic #429 "Water Health Risks" can be used to show the difference between our two stemming strategies. The search term "Health" is translated as "здравето" in the topic's title, and we found the following forms in the relevant documents: "здравен", "здравна" or "здравното". When using our derivational stemmer, all these forms were conflated to the same stem ("здрав") which was also the same stem for the word appearing in the query. With the light stemmer, the forms used in the relevant document were indexed under "здравн" which differs from the form appearing in the query ("здрав"). For the Czech corpus, we encountered a problem with spelling variations. With Topic #411 "Best picture Oscar", the award name appears with two distinct spellings. In the Czech query however, the form used was "Oskar" (with a "k") while in the relevant documents we found the form "Oscar". The different search models were not able to find a match for the two forms.

Table 3. MAP Before and After Blind-Query Expansion

	Mean average precision			
Query TD Stemmer	Hungarian decompound	Hungarian decompound	Bulgarian derivation.	Czech light
Model	IneC2	Okapi	LM	Okapi
Before	0.3897	0.3629	0.3368	0.3355
k docs/ m terms	5/20 0.4193*	5/20 0.3909*	10/50 **0.4098***	5/20 0.3557*
	5/50 0.4284*	5/50 0.3973*	10/80 0.4043*	5/50 0.3610*
	5/70 0.4283*	5/70 0.3983*	10/100 0.4061*	5/70 **0.3702***
	5/100 **0.4298***	5/100 **0.4010***	10/120 0.4004*	5/100 0.3685*

We found that pseudo-relevance feedback (PRF or blind-query expansion) could be a useful technique for enhancing retrieval effectiveness. In this study, we adopted Rocchio's approach [11] 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 three IR models and enlarged the query by the 20 to 120 terms extracted from the 5 to 10 best-ranked articles (see Table 3).

For the Hungarian collection, percentage improvement varied from +7.6% (IneC2 model, 0.3897 vs. 0.4193) to +10.5% (Okapi model, 0.3629 vs. 0.4010). For the Bulgarian corpus, enhancement increased from +18% (LM model, 0.3368 vs. 0.4004) to +21.7% (LM model, 0.3368 vs. 0.4098). For the Czech language, the variation percentages ranged from 6.0% (Okapi model, 0.3355 vs. 0.3557) to +10.3% (0.3355 vs. 0.3702). As shown in Table 3, the performance differences before and after query expansion were always statistically significant.

5 Data Fusion and Official Results

It is usually assumed that combining result lists computed by different search models (data fusion) should improve retrieval effectiveness, for three reasons [12]. This first is a skimming process, in which only the m top-ranked items retrieved from each ranked list are considered. In this case, we would combine the best answers obtained from various document representations. The second is the chorus effect, by which different retrieval schemes would retrieve the same item, and as such provide stronger evidence that the corresponding document is indeed relevant. The third is an opposite or dark horse effect, which may also play a role. A given retrieval model may provide unusually high and accurate estimates of a document's relevance. Thus, a combined system could possibly return more pertinent items by accounting for documents obtaining a relatively high score.

To present the official runs described in Table 4 we combined three probabilistic models, representing both the parametric (Okapi and DFR) and non-parametric (LM) probabilistic approaches. All runs were fully automated and in

Table 4. Description and MAP of Our Best Official Monolingual Runs

Language	Index	Query	Model	Query exp.	MAP	comb. MAP
Hungarian	dec.	TD	LM	5 docs/70 terms	0.4315	Z-score
UniNEhu2	word	TD	GL2	5 docs/100 terms	0.4376	**0.4716**
	4-gram	TD	Okapi	3 docs/120 terms	0.4233	
Bulgarian	4-gram	TD	Okapi	3 docs/150 terms	0.3169	Z-score
UniNEbg1	word	TD	PB2	5 docs/60 terms	0.3750	**0.4128**
	word	TD	LM	10 docs/50 terms	0.4098	
Czech	word	TD	LM	5 docs/20 terms	0.4070	Z-score
UniNEcz3	4-gram	TD	Okapi	5 docs/70 terms	0.3672	**0.4225**
	word	TD	GL2	5 docs/50 terms	0.4085	

all cases applied the same data fusion approach (Z-score [13]). For the Hungarian corpus however we occasionally applied our decompounding approach (denoted by "dec" in the"Index" column). As shown in Table 4, for a data fusion strategy retrieval performance in clearly better for the Hungarian language, moderate for the Bulgarian and only slightly better for the Czech language.

6 Conclusion

In this eighth CLEF evaluation campaign we analyze various probabilistic IR models using three different test-collections written in three East European languages (Hungarian, Bulgarian and Czech). We suggest a new stemmer for the Bulgarian language that removes some very frequently appearing derivational suffixes. For the Czech language, we design and implement two different stemmers.

Our various experiments demonstrate that the IneC2 model derived from *Divergence from Randomness* (DFR) paradigm tends to produce the best overall retrieval performances (see Tables 1 or 2). The statistical language model (LM) used in our experiments usually provides inferior retrieval performance to that obtained with the Okapi or DFR approach.

For the Bulgarian language (Table 1), our new and more aggressive stemmer tends to produce better MAP compared to a light stemming approach (around +6% in relative difference). For the Hungarian language (Table 1), applying an automated decompounding procedure improves the MAP around +10.8% when compared to a word-based approach. For the Czech language however performance differences between a light and a more aggressive stemmer removing both inflectional and some derivational suffixes are rather small (Table 2). Moreover, performance differences are also small when compared to those achieved with a 4-gram approach. The pseudo-relevance feedback may improve the MAP, depending on the parameter settings used (Table 3).

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

References

1. Savoy, J.: Combining Multiple Strategies for Effective Monolingual and Cross-Lingual Retrieval. IR Journal 7, 121–148 (2004)
2. Savoy, J., Abdou, S.: Experiments with Monolingual, Bilingual, and Robust Retrieval. In: Peters, C., Clough, P., Gey, F.C., Karlgren, J., Magnini, B., Oard, D.W., de Rijke, M., Stempfhuber, M. (eds.) CLEF 2006. LNCS, vol. 4730, pp. 137–144. Springer, Heidelberg (2007)
3. Savoy, J.: Comparative Study of Monolingual and Multilingual Search Models for Use with Asian Languages. ACM Transactions on Asian Languages Information Processing 4, 163–189 (2005)
4. Di Nunzio, G.M., Ferro, N., Mandl, T., Peters, C.: CLEF 2007 Ad Hoc Track Overview. In: Peters, C., et al. (eds.) CLEF 2007. LNCS, vol. 5152, pp. 13–32. Springer, Heidelberg (2008)
5. Savoy, J.: Searching Strategies for the Hungarian Language. Information Processing & Management 44, 310–324 (2008)
6. Robertson, S.E., Walker, S., Beaulieu, M.: Experimentation as a Way of Life: Okapi at TREC. Information Processing & Management 36, 95–108 (2002)
7. Amati, G., van Rijsbergen, C.J.: Probabilistic Models of Information Retrieval Based on Measuring the Divergence from Randomness. ACM Transactions on Information Systems 20, 357–389 (2002)
8. Hiemstra, D.: Using Language Models for Information Retrieval. PhD Thesis (2000)
9. Savoy, J.: Statistical Inference in Retrieval Effectiveness Evaluation. Information Processing & Management 33, 495–512 (1997)
10. McNamee, P., Mayfield, J.: Character N-gram Tokenization for European Language Text Retrieval. IR Journal 7, 73–97 (2004)
11. Buckley, C., Singhal, A., Mitra, M., Salton, G.: New Retrieval Approaches Using SMART. In: Proceedings TREC-4, Gaithersburg, pp. 25–48 (1996)
12. Vogt, C.C., Cottrell, G.W.: Fusion via a Linear Combination of Scores. IR Journal 1, 151–173 (1999)
13. Savoy, J., Berger, P.-Y.: Monolingual, Bilingual, and GIRT Information Retrieval. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M. (eds.) CLEF 2005. LNCS, vol. 4022, pp. 131–140. Springer, Heidelberg (2006)