Published in Lecture Notes in Computer Science 3491, 233-244, 2005
which should be used for any reference to this work

Data Fusion for Effective European Monolingual Information Retrieval

Jacques Savoy

Institut interfacultaire d'informatique, Université de Neuchâtel, Pierre-à-Mazel 7, 2001 Neuchâtel, Switzerland
Jacques.Savoy@unine.ch

Abstract. For our fourth participation in the CLEF evaluation campaigns, our first objective was to propose an effective and general stopword list and a light stemming procedure for the Portuguese language. Our second objective was to obtain a better picture of the relative merit of various search engines when processing documents in the Finnish and Russian languages. Finally, based on the Z-score method we suggested a data fusion strategy intended to improve monolingual searches in various European languages.

1 Introduction

Making use of experiments we carried out in previous years [1], [2], we are now participating in the French, Finnish, Russian and Portuguese monolingual tasks without relying on dictionaries. Moreover, the IR approaches suggested are fully automatic and used freely available resources. This paper describes the information retrieval models we used in the monolingual tracks and is organized as follows: Section 2 describes our general approach to building stopword lists and stemmers for use with languages other than English. Section 3 evaluates two probabilistic models and five vector-space schemes using five different languages. Section 4 describes and evaluates various data fusion operators that will hopefully improve retrieval effectiveness. Finally, Section 5 depicts our official runs and presents a broad failure analysis.

2 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., Roma, police, minister, Chirac). 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.). We created a new one for the Portuguese language, adding it to last year's stopword lists [2] (these lists are available at www.unine.ch/info/clef/). For English we used the list provided by the SMART system (571 words), while for the other European languages, our

stopword list contained 463 words for the French language, 747 for Finnish, 420 for Russian and 392 for Portuguese.

Once high-frequency words were removed, an indexing procedure generally applied a stemming algorithm, in an attempt to conflate word variants into the same stem or root. In developing this procedure for the various European languages [3], we first wanted to remove only inflectional suffixes such as singular and plural word forms, and also feminine and masculine forms, such that they conflate to the same root. Our suggested stemmers also tried to remove various case markings (e.g., accusative or genitive) used in the Finnish and Russian languages. The Finnish language however involved additional morphological difficulties, given that this language frequently uses more than 12 cases. However, one of the real stemming problems with Finnish is the fact that the stem is often modified when suffixes are added. For example, "matto" (carpet in nominative singular form) becomes "maton" (in genitive singular form, with "-n" as suffix) or "mattoja" (in partitive plural form, with "-a" as suffix). Once we removed the corresponding suffixes, we were left with three distinct stems, namely "matto", "mato", and "matoj". Of course such irregularities also occur in other languages, usually introduced to make the spoken language flow better, such as "submit" and "submission". In Finnish however, these irregularities are more common, thus rendering the conflation of various word forms into the same stem more problematic. Thus, in order to index Finnish documents, some authors suggest using a morphological analyzer (based on a dictionary) [4].

More sophisticated schemes have already been proposed for the removal of derivational suffixes (e.g., "-ize", "-ably", "-ship" in the English language), as for example the stemmer developed by Lovins [5] (based on a list of over 260 suffixes), or that of Porter [6] (which looks for about 60 suffixes). For the French language only, we developed a stemming approach to remove some derivational suffixes (e.g., "communicateur" → "communiquer", "faiblesse" → "faible"). Our various stemming procedures can be found at `www.unine.ch/info/clef/`. Currently, it is not clear whether a stemming procedure removing only inflections from nouns and adjectives would result in better retrieval effectiveness, when compared to other stemming approaches that also consider verbs or remove both inflectional and derivational suffixes (e.g., the Snowball stemmers).

Diacritic characters are usually not present in English collections (with certain exceptions, such as "cliché"). For the Finnish, Portuguese and Russian languages, these characters were replaced by their corresponding non-accentuated letter. For the Russian language, we converted and normalized the Cyrillic Unicode characters into the Latin alphabet.

Finally, most European languages manifest other morphological characteristics, with compound word constructions being just one example (e.g., handgun, worldwide). In Finnish, we encounter similar constructions as such as "rakkauskirje" ("rakkaus" + "kirje" for love & letter) or "työviikko" ("työ" + "viikko" for work & week). Recently, Braschler & Ripplinger [7] showed that decompounding German words would significantly improve retrieval performance. In our experiments with the Finnish language, we used our decompounding al-

gorithm [2] (see also [8]), where both the compound words and their components were left in documents and queries.

3 Indexing and Searching Strategies

In order to obtain a broader view of the relative merit of various retrieval models, we represented each document (or request) by a set of weighted keywords. In order to define such weights, we would account for the term occurrence frequency (denoted tf_{ij} for indexing term t_j in document D_i), or we might also account for their frequency in the collection (or more precisely the inverse document frequency, denoted idf_j). However, we found that cosine normalization could prove beneficial, and in this case, each indexing weight could vary within the range of 0 to 1 (retrieval model notation: "doc=ntc, query=ntc" or "ntc-ntc"). Other variants might also be created. For example, the tf component could be computed as $0.5 + 0.5 \cdot$ [tf / max tf in a document] (retrieval model denoted "doc=atn"). We might also consider that a term's presence in a shorter document provides stronger evidence than it does in a longer document, leading to more complex IR models; for example, the IR model denoted by "doc=Lnu" [9], "doc=dtu" [10]. In Table 1, w_{ij} represents the indexing weight assigned to term t_j in document D_i, n indicates the number of documents in the collection, and nt_i the number of distinct indexing terms included in the representation of D_i.

Table 1. Weighting schemes

ntc	$w_{ij} = \frac{tf_{ij} \cdot idf_j}{\sqrt{\sum_{k=1}^{t} (tf_{ik} \cdot idf_k)^2}}$	atn	$w_{ij} = idf_j \cdot \left[\frac{0.5 + 0.5 \cdot tf_{ij}}{max\ tf_{i.}}\right]$
ltn	$w_{ij} = [ln(tf_{ij}) + 1] \cdot idf_j$	dtn	$w_{ij} = [ln(ln(tf_{ij}) + 1) + 1] \cdot idf_j$
Okapi	$w_{ij} = \frac{(k_1+1) \cdot tf_{ij}}{K + tf_{ij}}$ with $K = k_1 \cdot \left[(1-b) + b \cdot \frac{l_i}{avdl}\right]$		
dtu	$w_{ij} = \frac{\left[ln(ln(tf_{ij})+1)+1\right] \cdot idf_j}{(1-slope) \cdot pivot + (slope \cdot nt_i)}$		
Lnu	$w_{ij} = \frac{\frac{ln(tf_{ij})+1}{ln\left(\frac{l_i}{nt_i}\right)+1}}{(1-slope) \cdot pivot + (slope \cdot nt_i)}$		

In addition to the previous models based on the vector-space approach, we also considered probabilistic models. In this vein, we used the Okapi probabilistic model [11]. As a second probabilistic approach, we implemented the Prosit (or deviation from randomness) approach [12], [13] which is based on combining two information measures, formulated as follows:

$$\begin{aligned} w_{ij} &= Inf_{ij}^1 \cdot Inf_{ij}^2 = (1 - Prob_{ij}^1) \cdot -log_2\left[Prob_{ij}^2\right] \\ Prob_{ij}^1 &= tfn_{ij} \ / \ (tfn_{ij} + 1) \ \text{ with} \\ tfn_{ij} &= tf_{ij} \cdot \log_2\left[1 + ((C \cdot mean\ dl)/l_i)\right] \end{aligned}$$

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

where l_i indicates the number of indexing terms included in the representation of D_i, tc_j represents the number of occurrences of term t_j in the collection and n the number of documents in the corpus.

To measure the retrieval performance, we adopted the non-interpolated mean average precision (computed on the basis of 1,000 retrieved items per request by the TREC-EVAL program). To determine whether or not a given search strategy is better than another, a decision rule was required. To obtain this, we might apply statistical inference methods such as Wilcoxon's signed rank test, the Sign test [14] or the hypothesis testing based on bootstrap methodology [15]. In this paper, we based our statistical validation on the bootstrap approach because this methodology does not require that the underlying distribution of the observed data follow the normal distribution. Thus, in the tables found in this paper we have underlined statistically significant differences based on a two-sided non-parametric bootstrap test, based on those means having a significance level fixed at 5%.

Table 2. Mean average precision of various single searching strategies (English, French & Portuguese language)

	Mean average precision					
Language	English	English	French	French	Portug.	Portug.
Query	T	TD	T	TD	T	TD
Model	42 queries	42 queries	49 queries	49 queries	46 queries	46 queries
Prosit	0.4638	0.5313	0.4111	0.4568	0.3824	0.4695
Okapi	**0.4763**	**0.5422**	**0.4263**	**0.4685**	**0.3997**	**0.4835**
Lnu-ltc	0.4435	0.4979	0.3952	0.4349	0.3633	0.4579
dtu-dtn	0.4444	0.5319	0.3873	0.4143	0.3620	0.4600
atn-ntc	0.4203	0.4764	0.3768	0.4210	0.3559	0.4454
ltn-ntc	0.3876	0.4602	0.3718	0.4035	0.3737	0.4319
ntc-ntc	0.3109	0.3706	0.3056	0.3309	0.2981	0.3708

We indexed the English, French, and Portuguese collections using words as indexing units. The evaluations of our two probabilistic models and five vector-space schemes are listed in Table 2 in which the best performance is listed in bold type. This best performance is used as a baseline for our statistical testing. The underlined results therefore indicate that the difference in mean average precision compared to the best system can be viewed as being statistically significant. As depicted in Table 2, the Okapi model presents the best IR model for all collections. For the Portuguese corpus five IR models produce statistically similar performance (Okapi, Prosit, "Lnu-ltc", "dtu-dtn", and "ltn-ntc"), and a similar conclusion can be drawn from the English collection. Moreover, the data in Table 2 shows that when the number of search terms increases (from T to TD), the retrieval effectiveness usually does also. When considering the five best

retrieval schemes (namely, Prosit, Okapi, "Lnu-ltc", "dtu-dtn" and "atn-ntc"), the improvement is around 24.4% when comparing title-only (or T) with TD queries for the Portuguese collection, 14.7% when comparing the English corpus or 10% for the French collection.

In order to represent Finnish and Russian documents and queries, we considered the n-gram, and word-based indexing schemes. The resulting mean average precision for these various indexing approaches is shown in Table 3 (Finnish word-based indexing with decompounding).

Table 3. Mean average precision of various single searching strategies (Finnish and Russian collection)

	Mean average precision				
Language	Finnish	Finnish	Finnish	Russian	Russian
Index	word	5-gram	4-gram	word	4-gram
Query	TD	TD	TD	TD	TD
Model	45 queries	45 queries	45 queries	34 queries	34 queries
Prosit	0.4620	0.4707	0.5357	0.3448	0.2879
Okapi	**0.4773**	0.4805	0.5385	**0.3800**	**0.2890**
Lnu-ltc	0.4643	0.4767	0.5022	0.3794	0.2852
dtu-dtn	0.4746	0.4629	0.5200	0.3768	0.2705
atn-ntc	0.4629	0.4735	**0.5428**	0.3422	0.2543
ltn-ntc	0.4580	**0.4824**	0.4880	0.3579	0.2137
ntc-ntc	0.3862	0.4472	0.4466	0.2716	0.1916

When looking at results for the Finnish language (Table 3), we can see that 4-gram indexing scheme usually performs better than both 5-gram indexing (e.g., with the TD queries, 4-gram: mean MAP of the five best IR models is 0.5278 vs. 0.4729 with 5-gram indexing approach, a performance difference of 11.6% in favor of the 4-gram model) or better than the word-based indexing model (mean of 5 best IR models of 0.4692, with a performance difference of 12.5% in favor of the 4-gram indexing approach). There are of course exceptions to this rule (e.g., for the "ntc-ntc" model, the 5-gram indexing scheme results in slightly better performance than the 4-gram strategy, or 0.4472 vs. 0.4466). Moreover, our statistical testing does not usually show any significant differences in mean average precision when comparing the best 6 IR models.

As illustrated in Table 3, the word-based indexing scheme used for the Russian language provides better retrieval performance than does the 4-gram schemes (based on the five best search models, the mean MAP of these five schemes is 0.3646 vs. 0.2774 for the 4-gram indexing scheme, a difference of 31.4%). Based on our statistical testing, we usually were not able to find any significant differences between 5 IR models.

It was observed that pseudo-relevance feedback (or blind-query expansion) seemed to be a useful technique for enhancing retrieval effectiveness. In this study, we adopted Rocchio's approach [9] with $\alpha = 0.75$, $\beta = 0.75$, and $\gamma = 0$,

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 probabilistic models and enlarged the query by the 10 to 30 terms provided by the 3 or 10 best-retrieved articles.

The results depicted in Table 4 (depicting our best results for the Okapi model) indicate that the optimal parameter setting seemed to be collection-dependant. Moreover, performance improvement also seemed to be collection dependant (or language dependant), with the Portuguese corpus showing an increase of 6% (from a mean average precision of 0.4835 to 0.5127), 5.2% for the English collection (from 0.5422 to 0.5704), 3.8% for the Russian collection (from 0.3800 to 0.3945), and 3.5% for the French corpus (from 0.4685 to 0.4851). For the Finnish corpus and the 4-gram indexing scheme, the query expansion approach did not improve the mean average precision. In Table 4, the baseline upon which we based our statistical testing is the mean average precision before automatically expanding the query. In this case, it is interesting to note that our statistical testing usually cannot detect a significant difference in mean average precision before and after blind query expansion.

Table 4. Mean average precision using blind-query expansion (Okapi model)

	Mean average precision				
TD queries Index Model	English word 42 queries	French word 49 queries	Finnish 4-gram 45 queries	Russian word 34 queries	Portug. word 46 queries
Okapi	0.5422	0.4685	**0.5385**	0.3800	0.4835
k doc. /m terms	3/10 0.5582 3/15 0.5581 5/10 **0.5704** 5/15 0.5587 10/10 0.5596 10/15 0.5596	3/10 **0.4851** 3/15 0.4748 5/10 0.4738 5/15 0.4628 10/10 0.4671 10/15 0.4547	3/10 0.5308 3/15 0.5296 5/10 0.5278 5/15 0.5213 10/10 0.5291 10/15 0.5297	3/15 0.3925 3/30 0.3678 5/15 0.3896 5/30 **0.3945** 5/40 0.3796 10/30 0.3912	3/10 0.5005 3/15 **0.5127** 3/20 0.5098 5/10 0.4465 5/15 0.5077 10/15 0.4806

Using the same query expansion technique (Rocchio in this case), various IR models have resulted in varying degrees of evolution when increasing the number of terms to be included in the expanded query. Figure 1 illustrates this phenomenon showing the evolution of the mean average precision of four different IR models (French corpus, and using the 3 best ranked documents). When we increased the number of terms to be included in the expanded query, the "dtu-dtn" model showed a small but constant improvement. With this IR model, each parameter setting produced a retrieval performance not that far from the best one. A similar evolution could also be seen with the "Lnu-ltc" model, yet with even greater improvement. When compared to the Okapi or Prosit models however, performance levels achieved were lower. For the Prosit model as well as for the Okapi scheme, the mean average precision increased, reaching a maximum point and then subsequently slowly decreasing (however with the Prosit model

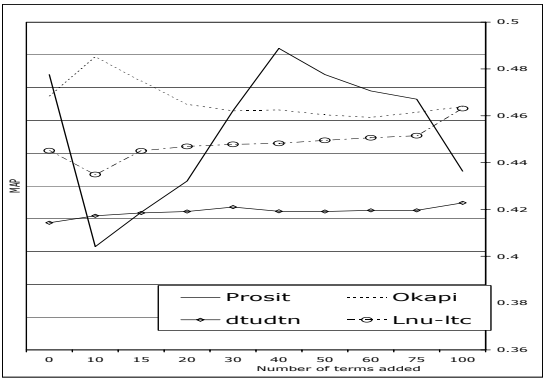

Fig. 1. Mean average precision using blind-query expansion within different retrieval models (French corpus, terms extracted from the 3 best ranked documents

showing greater variability). When a few terms were added to the original query however, the Prosit model usually performed at lower levels than did the Okapi.

4 Data Fusion

For each language studied, we may assume that different indexing and search models would retrieve different pertinent and non-relevant items, and that combining different search models would improve retrieval effectiveness. More precisely, when combining different indexing schemes we would expect to improve recall, due to the fact that different document representations might retrieve different pertinent items [16]. On the other hand, when combining different search schemes, we could suppose that these various IR strategies are more likely to rank the same relevant items higher on the list than they would the same non-relevant documents (viewed as outliers). Thus, combining them could improve retrieval effectiveness by ranking pertinent documents higher and ranking non-relevant items lower. In this study, we hope to enhance retrieval performance by making use of this second characteristic, while for the Finnish language our assumption would be that word-based and n-gram indexing schemes are distinct and independent sources of evidence regarding the content of documents. For this language only, we expect to improve recall due to the first effect described above.

In this current study we limited the number of IR schemes to be combined to two. To achieve this, we evaluated various fusion operators, and their precise descriptions are listed in Table 5. 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 [17]. We can thus see from Table 5 that 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, we may multiply the document score by a constant α_i (usually equal to 1) in order to favor the ith more efficient retrieval scheme.

Table 5. Data fusion combination operators used in this study

Sum RSV	$\alpha_i \cdot RSV_k$
Norm Max	$\alpha_i \cdot (RSV_k / Max^i)$
Norm RSV	$\alpha_i \cdot [(RSV_k - Min^i)/(Max^i - Min^i)]$
Z-Score	$\alpha_i \cdot [((RSV_k - \mu^i)/\sigma^i) + \delta^i]$, with $\delta^i = [(\mu^i - Min^i)/\sigma^i]$

In addition to using these data fusion operators, we also considered the round-robin approach, wherein we take one document in turn from all individual lists and remove any duplicates, keeping 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 ith result list, we needed to compute the average of the RSV_k (denoted μ^i) and the standard deviation (denoted σ^i). Based on these values, we would then normalize the retrieval status value for each document D_k provided by the ith result list by computing the deviation of RSV_k with respect to the mean (μ^i). In Table 5, Min^i (Max^i) denotes the minimal (maximal) RSV value in the ith result list.

Table 6 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 could see that combining two IR models might improve retrieval effectiveness. When combining two retrieval models, the Z-score scheme tended to produce the best performance. In Table 6, under the heading "Z-scoreW", we attached a weight of 1.5 to the best performing model (depicted in bold in the first two lines), and 1 to the other. Using the best single IR as a

Table 6. Mean average precision using different combination operators (with blind-query expansion)

	Mean average precision				
Query TD Index Model	English word 42 queries	French word 49 queries	Finnish 4-gram 45 queries	Russian word 34 queries	Portug. word 46 queries
Okapi-PRF	5/10 0.5704	3/10 **0.4851**	0/0 0.5385	5/30 **0.3945**	3/15 0.5127
Prosit-PRF	3/30 **0.5742**	10/20 0.4643	3/40 **0.5684**	10/15 0.3736	5/75 **0.5230**
Round-robin	0.5790	0.4824	0.5643	0.3900	0.5251
Sum RSV	0.5837	0.4792	0.5500	0.4041	0.5153
Norm Max	0.5789	0.4851	0.5696	0.4081	0.5396
Norm RSV	0.5752	0.4864	0.5692	0.4130	0.5348
Z-Score	0.5818	0.4906	0.5718	**0.4160**	**0.5399**
Z-ScoreW	**0.5854**	**0.4933**	**0.5754**	0.4145	0.5359

baseline, our statistical testing was not able to detect a significant enhancement when combining two IR models.

5 Official Results and Analysis

Finally, in Table 7 we show the exact specifications of our 12 official monolingual runs. These experiments were based on different data fusion operators (mainly the Z-score and the round-robin schemes). Although we expected that combining the Okapi and the Prosit probabilistic models would provide good retrieval effectiveness, for some languages (e.g., French or Russian), we also considered other IR models (e.g., "dtu-dtn" or "Lnu-ltc"). We also sent some runs with longer queries formulations (TDN) in order to increase the number of relevant documents found for each language. In the "UniNEfi1" run, we filter all documents appearing in the year 1994 out before returning the final list (in order to search all newspaper articles that described events occurring in the year 1995. However, 66 (over 413) relevant items had been published in year 1994). This was not a good strategy. If we keep the articles appearing in the year 1994, we may achieve a MAP of 0.5340 (instead of 0.4967 obtained by the "UniNEfi1" run).

For both the Portuguese and French languages and compared to other experiments done during this CLEF evaluation campaign, it is our opinion that the IR approach we used produces very good results. Even though our statistical tests did not detect significant enhancement, we would still suggest automatically expanding the query and following this step, combining both the Okapi and Prosit probabilistic models.

For the Finnish language, it seems that a deeper morphological analysis will improve the retrieval effectiveness. Moreover, a better decompounding algorithm will clearly enhance the mean average precision. For example, Tomlinson [18] indicates that we may enhance the mean average precision from 0.469 to 0.561 (+ 19.6% for the Finnish collection, Title-only queries) when including a good decompounding approach. Moulinier & Williams [19] used a commercial morphological analyzer for Finnish and also obtained good overall retrieval performance levels with this language. On the other hand, an analysis of our IR system shows that we failed to decompound important search terms due to the fact that our decompounding strategy was too conservative.

For the Russian language, we were not able to draw any definitive conclusions due to the small size of the corpus (composed of 16,716 documents) and also due to the fact that for numerous queries the number of relevant items was rather small. For example, for ten queries out of a total of 34, we found only one relevant document in the corpus (and seven other queries found only two pertinent items in the collection). This fact may therefore only favor a given IR system by chance, and this to the detriment of another. For example, if a given system retrieves the single pertinent item in the first rank, it will obtain a precision of 1.0 for this query, and if this pertinent item is only retrieved in the 2nd position, it will only obtain a precision of 0.5. If we repeat this swapping between the first and second extracted document for the ten requests having only one relevant item,

Table 7. Description and mean average precision (MAP) of our official runs

Run name	Lan.	Query	Index	Model	Query exp.	Combined	MAP
UniNEfr1	FR	TD	word	dtu-dtn	5 d. / 40 t.		
		TD	word	Prosit	10 d. /30 t.	RR	0.4437
UniNEfr2	FR	TD	word	Prosit	10 d. / 30 t.		
		TD	word	Okapi	3 d. / 10 t.	Z-Score	**0.4849**
UniNEfr3	FR	TDN	word	Prosit	5 d. / 20 t.		
		TDN	word	dtu-dtn	10 d. / 30 t.	Z-ScoreW	0.4785
UniNEfi1	FI	TD	4-gram	Prosit	3 d. / 40 t.		
		TD	word	Prosit	3 d. / 20 t.	Z-ScoreW	0.4967
UniNEfi2	FI	TD	4-gram	Prosit	3 d. / 40 t.		
		TD	word	Prosit	3 d. / 20 t.		
		TD	4-gram	Okapi	3 d. / 20 t.	Sum RSV	**0.5453**
UniNEfi3	FI	TDN	4-gram	Prosit	3 d. / 30 t.		
		TDN	word	Prosit	3 d. / 20 t.	Z-ScoreW	0.5454
UniNEru1	RU	TD	word	Prosit			
		TD	word	Lnu-ltc	3 d. / 20 t.	RR	**0.3546**
UniNEru2	RU	TD	word	Prosit			
		TD	word	Okapi		Z-score	0.3545
UniNEru3	RU	TDN	word	Prosit	10 d. / 15 t.		
		TDN	word	Okapi	5 d. / 15 t.	RR	0.4070
UniNEpt1	PT	TD	word	Okapi	5 d. / 15 t.		
		TD	word	Prosit	10 d. / 10 t.	Norm RSV	0.5004
UniNEpt2	PT	TD	word	Prosit	5 d. / 30 t.		
		TD	word	Lnu-ltc	10 d. / 15 t.	Z-score	0.5105
UniNEpt3	PT	TD	word	Okapi	10 d. / 20 t.		
		TD	word	Prosit	10 d. / 50 t.	Norm RSV	**0.5188**

the mean average precision over 34 queries between these two systems will be 0.147 (or $(0.5 \cdot 10) / 34$).

6 Conclusion

In this fifth CLEF evaluation campaign, we proposed a general stopword list and a light stemming procedure (removing only inflections attached to nouns and adjectives) for the Portuguese language. In order to enhance the retrieval performance, we suggest using a data fusion approach based on the Z-score in order to combine two probabilistic IR models. The results of this evaluation campaign seem to indicate that such an approach is effective for the French and Portuguese languages.

However, we also found that pseudo-relevance feedback based on Rocchio's model usually does not statistically improve mean average precision, even though mean precision following query expansion usually the shows a better value. Similarly, combining two retrieval models based on the same indexing strategy usually does not statistically enhance retrieval performance.

Acknowledgments. The author would like to also thank the CLEF-2004 task organizers for their efforts in developing various European language test-collections. The author would also like to thank C. Buckley from SabIR for giving us the opportunity to use the SMART system. This research was supported by the Swiss National Science Foundation under Grant #21-66 742.01.

References

1. Savoy, J.: Combining Multiple Strategies for Effective Monolingual and Cross-Lingual Retrieval. IR Journal, **7** (2004) 121–148
2. Savoy, J.: Report on CLEF-2003 Monolingual Tracks: Fusion of Probabilistic Models for Effective Monolingual Retrieval. In: Peters, C., Braschler, M., Gonzalo, J., Kluck, M. (Eds.): Advances in Cross-Language Information Retrieval. Lecture Notes in Computer Science. Springer, Heidelberg (2004), to appear
3. Sproat, R.: Morphology and Computation. The MIT Press, Cambridge (1992)
4. Hedlund, T., Airio, E., Keskustalo, H., Lehtokangas, R., Pirkola, A., Järvelin, K.: Dictionary-Based Cross-Language Information Retrieval: Learning Experiences from CLEF 2000-2002. IR Journal, **7** (2004) 99–119
5. Lovins, J.B.: Development of a Stemming Algorithm. Mechanical Translation and Computational Linguistics **11** (1968) 22–31
6. Porter, M.F.: An Algorithm for Suffix Stripping. Program **14** (1980) 130–137
7. Braschler, M., Ripplinger, B.: How Effective is Stemming and Decompounding for German Text Retrieval? IR Journal, **7** (2004) 291–316
8. Chen, A.: Cross-Language Retrieval Experiments at CLEF 2002. In: Peters, C., Braschler, M., Gonzalo, J., Kluck, M. (Eds.): Advances in Cross-Language Information Retrieval. Lecture Notes in Computer Science: Vol. 2785. Springer, Heidelberg (2003), 28–48
9. Buckley, C., Singhal, A., Mitra, M., Salton, G.: New Retrieval Approaches Using SMART. In Proceedings TREC-4. NIST Publication #500-236, Gaithersburg (1996) 25–48
10. Singhal, A., Choi, J., Hindle, D., Lewis, D.D., Pereira, F.: AT&T at TREC-7. In Proceedings TREC-7. NIST, Publication #500-242, Gaithersburg (1999) 239–251
11. Robertson, S.E., Walker, S., Beaulieu, M.: Experimentation as a Way of Life: Okapi at TREC. Information Processing & Management, **36** (2000) 95–108
12. Amati, G., Carpineto, C., Romano, G.: Italian Monolingual Information Retrieval with PROSIT. In: Peters, C., Braschler, M., Gonzalo, J., Kluck, M. (Eds.): Advances in Cross-Language Information Retrieval. Lecture Notes in Computer Science: Vol. 2785. Springer, Heidelberg (2003), 257–264
13. Amati, G., van Rijsbergen, C.J.: Probabilistic Models of Information Retrieval Based on Measuring the Divergence from Randomness. ACM Transactions on Information Systems, **20** (2002) 357–389
14. Hull, D.: Using Statistical Testing in the Evaluation of Retrieval Experiments. In Proceedings of the ACM-SIGIR'93. The ACM Press, New York (1993) 329–338
15. Savoy, J.: Statistical Inference in Retrieval Effectiveness Evaluation. Information Processing & Management, **33** (1997) 495–512
16. Vogt, C.C., Cottrell, G.W.: Fusion via a Linear Combination of Scores. IR Journal, **1** (1999) 151–173
17. Fox, E.A., Shaw, J.A.: Combination of Multiple Searches. In Proceedings TREC-2. NIST Publication #500-215, Gaithersburg (1994) 243–249

18. Tomlinson, S.: Finnish, Portuguese and Russian Retrieval with Hummingbird SearchServer^{TM}at CLEF 2004. In: Peters, C., Clough, P., Gonzalo, J., Jones, G., Kluck, M., Magnini, B. (Eds.): Advances in Cross-Language Information Retrieval. Lecture Notes in Computer Science. Springer, Heidelberg (in print)
19. Moulinier, I., Williams, K.: Report on Thomson Legal and Regulatory Experiments at CLEF 2004. In: Peters, C., Clough, P., Gonzalo, J., Jones, G., Kluck, M., Magnini, B. (Eds.): Advances in Cross-Language Information Retrieval. Lecture Notes in Computer Science. Springer, Heidelberg (in print)