Data Fusion for Effective European Monolingual Information Retrieval

Jacques Savoy

Institut interfacultaire d'informatique
Université de Neuchâtel, Switzerland
Jacques.Savoy@unine.ch Web site: www.unine.ch/info/clef/

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.

Introduction

Based on our experiments of the previous years (Savoy 2003; 2004a), we are participating in French, Finnish, Russian and Portuguese monolingual tasks without relying on a dictionary and using fully automated approaches. This paper describes the information retrieval models we used in the monolingual tracks and is organized as follows: Section 1 contains an overview of the test-collections built during this evaluation campaign while 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 nine vector-space schemes using five different languages. Finally, Section 4 describes and evaluates various data fusion operators, together with our official runs.

1. Overview of the Test-Collections

The corpora used in our experiments included newspaper and news agency articles, for example the *Glasgow Herald* (1995, English), *Le Monde* (1995, French), *SDA* (*Schweizerische Depeschenagentur*, 1995, French), *Aamulehti* (1994/95, Finnish), *Izvestia* (1995, Russian), and *Público* (1995, Portuguese). As shown in Table 1, these corpora are of various sizes, with the French collection being the biggest (244 MB) and the Portuguese, English and Finnish collections ranking second (around 150 MB). Finally the Russian collection ranks as the smallest, both in size (68 MB) and in number of documents (16,716). Across all the corpora the mean number of distinct indexing terms per document is relatively similar (around 130), but this number is a little bit larger for the Portuguese collection (180.94) and smaller for the Russian corpus (124.53). As for the mean number of indexing terms per article (listed in third part of Table 1), the Portuguese documents have the largest mean size (254.96), the English corpus ranks second (mean value: 200.72), and the Russian collection has the smallest mean document size (163.24). However this last corpus exhibits also the largest variability (standard deviation: 252.41) in terms of document length.

Table 1 (bottom part) also compares the number of relevant documents per request, with the mean always being greater than the median (e.g., for the English collection, the average number of relevant documents per query is 8.93 with the corresponding median being 4). These findings indicate that each collection contains numerous queries, yet only a rather small number of relevant items are found. For each collection, 50 queries were created. Relevant documents cannot however be found for each request and each language. For the French collection, Query #227 does not have any relevant items; for the English collection, these requests are #203, #220, #225, #227, #234, #243, #244 and #250; for the Finnish corpus: Queries #206, #227, #231, #240, #247; for the Russian corpus: Queries #204, #205, #206, #208, #217, #219, #222, #223, #229, #236, #240, #243, #246, #247, #248, #249, and for the Portuguese corpus: Queries #216, #220, #227, #240.

During the indexing process of our automatic runs, we retained only the following logical sections from the original documents: <TITLE>, <HEADLINE>, <TEXT>, <LEAD1>, <TX>, <LD>, <TI> and <ST>. From the topic descriptions we automatically removed certain phrases such as “Relevant document report …”, “Find documents …” or “Trouver des documents qui parlent …”.

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, president, 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 (Savoy 2003) (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 356 for Portuguese. To this last list, we recently added a few forms to obtain a Portuguese stopword list containing 392 words.

	English	French	Finnish	Russian	Portuguese
Size (in MB)	154 MB	244 MB	137 MB	68 MB	176 MB
# of documents	56,472	90,261	55,344	16,716	55,070
# of distinct terms	524,788	332,872	1,444,213	345,719	307,424
Number of distinct indexing terms / document					
Mean	136.45	127.10	128.25	124.53	180.94
Standard deviation	99.34	103.85	95.35	179.86	133.61
Median	116	92	101	41	154
Maximum	1,882	2,645	1,892	1,769	2,577
Minimum	5	1	2	1	1
Number of indexing terms / document					
Mean	200.72	176.47	183.68	163.24	254.96
Standard deviation	162.90	155.47	146.06	252.41	222.86
Median	162	125	150	49	204
Maximum	5,248	6,720	6,617	2,821	7,247
Minimum	6	1	2	1	1
Number of queries	42	49	45	34	46
Number rel. items	375	915	413	123	678
Mean rel./request	8.93	18.67	9.18	3.62	14.74
Standard deviation	10.28	22.16	10.15	3.81	30.00
Median	4	12	5	2.5	5
Maximum	41 (Q#232)	100 (Q#213)	49 (Q#212)	20 (Q#241)	189 (Q#229)
Minimum	1 (Q#210)	1 (Q#225)	1 (Q#209)	1 (Q#203)	1 (Q#215)

Table 1: CLEF 2004 test-collection statistics

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 various European languages (Sproat 1992), 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 case) used in the Finnish and Russian languages. The Finnish language however raised more morphological difficulties, because this language frequently uses 12 cases and also 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). When we simply removed the corresponding suffix, we were faced 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. For indexing Finnish documents, some authors therefore suggested using a morphological analyzer (using a dictionary) as well as word form normalization procedures (Hedlund *et al*. 2004).

More sophisticated schemes were already proposed for the removal of derivational suffixes (e.g., "-ize", "-ably", "-ship" in the English language), as for example the stemmer developed by Lovins (1968) (based on a list of over 260 suffixes), or that of Porter (1980) (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 inflectional suffixes from nouns and adjectives would result in better retrieval effectiveness than would other stemming approaches that also consider verbs or remove both inflectional and derivational suffixes (e.g., the Snowball stemmers available at http://snowball.tartarus.org/).

Diacritic characters are usually not present in English collections (with certain exceptions, such as "résumé" or "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 (the Perl script is available at www.unine.ch/clef/).

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 (2004) showed that decompounding German words would significantly improve retrieval performance. In our experiments, for the Finnish language we used our decompounding algorithm (Savoy 2003) (see also (Chen 2003)), 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 first adopted a binary indexing scheme in which each document (or request) was represented by a set of keywords, without any weight. To measure the similarity between documents and requests, we computed the inner product (retrieval model denoted "doc=bnn, query=bnn" or "bnn-bnn"). In order to weight the presence of each indexing term in a document surrogate (or in a query), we would account for the term occurrence frequency (denoted tf_{ij} for indexing term t_j in document D_i, and the corresponding retrieval model is denoted: "doc=nnn, query=nnn" or "nnn-nnn") or we might also account for their frequency in the collection (or more precisely the inverse document frequency, denoted by idf_j). Moreover, 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: "ntc-ntc"). In Table 3 w_{ij} represents the indexing weight assigned to term t_j in document D_i, n to indicate the number of documents in the collection and nt_i the number of distinct indexing terms included in the representation of D_i.

Other variants might also be created. For example, the tf component could be computed as 0.5 + 0.5 · [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" (Buckley *et al*. 1996), "doc=dtu" (Singhal *et al*. 1999).

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 (Robertson *et al*. 2000). As a second probabilistic approach, we implemented the Prosit (or deviation from randomness) approach (Amati & van Rijsbergen 2002) which is based on the combination of two information measures as follows:

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

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

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

where w_{ij} indicates the indexing weight attached to term t_j in document D_i, l_i the number of indexing terms included in the representation of D_i, tc_j represents the number of occurrences of term t_j in the collection and n the number of documents in the corpus. In our experiments, the constants b, k_1, *avdl*, *pivot*, *slope*, C and *mean dl* were fixed according to values listed in Table 2.

		Okapi			Prosit	
Language	Index	b	k_1	*avdl*	C	*mean dl*
English	word	0.8	2	750	1.8	136
French	word	0.7	1.5	600	1.25	182
Portuguese	word	0.75	1.2	750	1.25	250
Finnish	word	0.8	4	800	1.75	114
Finnish	4-gram	0.5	1.2	800	1.5	539
Finnish	5-gram	0.5	1.2	800	1.5	539
Russian	word	0.75	2	300	0.6	124
Russian	4-gram	0.75	0.8	1,000	2	468

Table 2: Parameter setting for the various test-collections

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

Table 3: Weighting schemes

To evaluate our approaches, we used the SMART system as a test bed running on an Intel Pentium III/600 (memory: 1 GB, swap: 2 GB, disk: 6 x 35 GB). 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). We indexed the English, French, and Portuguese collections using words as indexing units. The evaluation of our two probabilistic models and nine vector-space schemes are listed in Table 4 for the French and Portuguese corpus, and in Table 5 for the English 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 5 (Finnish word-based indexing with decompounding), in Table 6 (Finnish based on the 5-gram or the 4-gram indexing scheme) and in Table 7 (Russian corpus both word-based and 4-gram indexing). In these tables, we depicted in bold the best performance under given conditions (with the same indexing scheme and the same collection).

	Mean average precision					
Query Model \ # of queries	French T 49 queries	French TD 49 queries	French TDN 49 queries	Portuguese T 46 queries	Portuguese TD 46 queries	Portuguese TDN 46 queries
Prosit	0.4111	0.4568	**0.4857**	0.3824	0.4695	**0.4995**
doc=Okapi, query=npn	**0.4263**	**0.4685**	0.4852	**0.3997**	**0.4835**	0.4968
doc=Lnu, query=ltc	0.3952	0.4349	0.4666	0.3633	0.4579	0.4765
doc=dtu, query=dtn	0.3873	0.4143	0.4504	0.3620	0.4600	0.4735
doc=atn, query=ntc	0.3768	0.4210	0.4397	0.3559	0.4454	0.4579
doc=ltn, query=ntc	0.3718	0.4035	0.4238	0.3737	0.4319	0.4401
doc=ntc, query=ntc	0.3056	0.3309	0.3468	0.2981	0.3708	0.3751
doc=ltc, query=ltc	0.2822	0.3184	0.3433	0.2820	0.3571	0.3831
doc=lnc, query=ltc	0.3023	0.3463	0.3811	0.2911	0.3658	0.3977
doc=bnn, query=bnn	0.2262	0.2017	0.1460	0.1793	0.1834	0.1332
doc=nnn, query=nnn	0.2073	0.2104	0.2008	0.1714	0.1681	0.1578

Table 4: Mean average precision of various single searching strategies (French & Portuguese language)

From an analysis of these results, it can be seen that when the number of search terms increases (from T, TD to TDN), so usually does retrieval effectiveness (except for "bnn-bnn" or "nnn-nnn" IR models). When considering the five best retrieval schemes (namely, Prosit, Okapi, "Lnu-ltc", "dtu-dtn" and "atn-ntc"), Tables 4 and 5 show that the improvement is around 29% when comparing title-only (or T) with TDN queries for the Portuguese collection, or of 22.1% with the English corpus or 16.6% for the French collection. When considering the Finnish language (Table 6 and right part of Table 5), 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 TD queries and "ntc-ntc" model, the 5-gram indexing scheme results in slightly better performance than the 4-gram strategy, 0.4472 vs. 0.4466). As illustrated in Table 7, for the Russian language the word-based indexing scheme provides better retrieval performance than do the 4-gram schemes (based on the five best search models, for TD queries the mean MAP of the five best retrieval is 0.3646 vs. 0.2774 for the 4-gram indexing scheme, a difference of 31.4%).

	Mean average precision					
Query Model \ # of queries	English T 42 queries	English TD 42 queries	English TDN 42 queries	Finnish(wd) T 43 queries	Finnish(wd) TD 45 queries	Finnish(wd) TDN 45 queries
Prosit	0.4638	0.5313	0.5652	0.3237	0.4620	0.4697
doc=Okapi, query=npn	**0.4763**	**0.5422**	**0.5707**	**0.4190**	**0.4773**	0.4820
doc=Lnu, query=ltc	0.4435	0.4979	0.5470	0.4187	0.4643	0.4961
doc=dtu, query=dtn	0.4444	0.5319	0.5372	0.4152	0.4746	**0.4989**
doc=atn, query=ntc	0.4203	0.4764	0.5245	0.4019	0.4629	0.4819
doc=ltn, query=ntc	0.3876	0.4602	0.5072	0.4054	0.4580	0.4801
doc=ntc, query=ntc	0.3109	0.3706	0.4006	0.3485	0.3862	0.3960
doc=ltc, query=ltc	0.3072	0.3915	0.4028	0.3511	0.3964	0.4172
doc=lnc, query=ltc	0.3342	0.4108	0.4326	0.3451	0.4176	0.4354
doc=bnn, query=bnn	0.3177	0.3005	0.2090	0.2226	0.1859	0.1394
doc=nnn, query=nnn	0.1937	0.1846	0.1570	0.1817	0.1318	0.1200

Table 5: Mean average precision of various single searching strategies (English & Finnish language)

	Mean average precision					
Finnish Query Model \ # of queries	word & CC TD 45 queries	5-gram TD 45 queries	5-gram TDN 45 queries	4-gram T 45 queries	4-gram TD 45 queries	4-gram TDN 45 queries
Prosit	0.4445	0.4707	0.4666	0.4953	0.5357	0.5166
doc=Okapi, query=npn	0.4564	0.4805	0.4855	**0.4987**	0.5386	0.5151
doc=Lnu, query=ltc	0.4466	0.4767	0.4805	0.4731	0.5022	0.5138
doc=dtu, query=dtn	**0.4565**	0.4629	0.4615	0.4806	0.5200	0.5143
doc=atn, query=ntc	0.4187	0.4735	**0.5104**	0.4900	**0.5427**	**0.5465**
doc=ltn, query=ntc	0.4466	**0.4824**	0.4907	0.4553	0.4880	0.4688
doc=ntc, query=ntc	0.3747	0.4472	0.4709	0.4000	0.4466	0.4472
doc=ltc, query=ltc	0.3897	0.4290	0.4398	0.3766	0.4284	0.4693
doc=lnc, query=ltc	0.4005	0.4177	0.4592	0.3989	0.4345	0.4893
doc=bnn, query=bnn	0.2373	0.2616	0.1631	0.3146	0.2387	0.1185
doc=nnn, query=nnn	0.1694	0.2038	0.1668	0.2028	0.1781	0.1354

Table 6: Mean average precision of various single searching strategies (Finnish collection)

For the Finnish language, we also indexed documents and the queries using words and "words" composed only of consonants. With this indexing scheme, the term "rakkaus" is indexed under both "rakkaus" and "rkks". In this experiment, before removing all vowels, we applied our Finnish stemming stemmer. The mean average precision achieved by this indexing strategy was always lower than the corresponding word-based approach (see second column of Table 6 under the label "word & CC"). We must recognize that the Finnish language, with its rich inflectional morphology and its frequent irregularities, resulted in many difficulties for our simple stemming approach.

It was observed that pseudo-relevance feedback (blind-query expansion) seemed to be a useful technique for enhancing retrieval effectiveness. In this study, we adopted Rocchio's approach (Buckley *et al.* 1996) with $\alpha = 0.75$, $\beta = 0.75$ whereby the system was allowed to add *m* terms extracted from the *k* best ranked documents from the original query. To evaluate this proposition, we used the Okapi and the Prosit probabilistic models and enlarged the query by the 10 to 40 terms provided by the 3 or 10 best-retrieved articles.

The results depicted in Table 8 (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 4-gram indexing scheme, the query expansion approach did not improve the mean average precision, while with word-based indexing scheme, the best improvement was of 4.4% (0.4773 vs. 0.4984). Using the Prosit model (see Table 9), similar conclusions can be drawn. In this case however, the blind query expansion improves the mean average precision for all collections.

	Mean average precision					
Russian Query Model \ # of queries	word T 34 queries	word TD 34 queries	word TDN 34 queries	4-gram T 34 queries	4-gram TD 34 queries	4-gram TDN 34 queries
Prosit	0.3130	0.3448	0.3598	0.2268	0.2879	0.2734
doc=Okapi, query=npn	0.3566	**0.3800**	**0.3944**	0.2367	**0.2890**	0.2800
doc=Lnu, query=ltc	0.3409	0.3794	0.3900	**0.2425**	0.2852	**0.3109**
doc=dtu, query=dtn	**0.3802**	0.3768	0.3894	0.1851	0.2705	0.2923
doc=atn, query=ntc	0.3264	0.3422	0.3650	0.2325	0.2543	0.2173
doc=ltn, query=ntc	0.3272	0.3579	0.3241	0.2014	0.2137	0.1697
doc=ntc, query=ntc	0.2541	0.2716	0.2581	0.1690	0.1916	0.1862
doc=ltc, query=ltc	0.2341	0.2362	0.2451	0.1134	0.1430	0.1290
doc=lnc, query=ltc	0.1850	0.1598	0.2014	0.1032	0.1303	0.1167
doc=bnn, query=bnn	0.1680	0.1512	0.1055	0.1437	0.0373	0.0061
doc=nnn, query=nnn	0.1130	0.1023	0.0967	0.0537	0.0408	0.0229

Table 7: Mean average precision of various single searching strategies (Russian corpus)

	Mean average precision					
Query TD Model	English word 42 queries	French word 49 queries	Finnish 4-gram 45 queries	Finnish word 45 queries	Russian word 34 queries	Portuguese word 46 queries
Okapi	0.5422	0.4685	**0.5386**	0.4773	0.3800	0.4835
k doc. /□*n* 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 0.5277 5/15 0.5213 10/10 0.5291 10/15 0.5297	3/10 0.4687 3/20 0.4628 5/10 0.4799 5/20 **0.4984** 5/30 0.4758 10/20 0.4461	3/15 0.3925 3/30 0.3678 5/15 0.3896 5/30 **0.3945** 5/40 0.3796 10/30 0.3913	3/10 0.5005 3/15 **0.5127** 3/20 0.5098 5/10 0.5005 5/15 0.5077 10/15 0.4806

Table 8: Mean average precision using blind-query expansion (Okapi model)

	Mean average precision					
Query TD Model	English word 42 queries	French word 49 queries	Finnish 4-gram 45 queries	Finnish word 45 queries	Russian word 34 queries	Portuguese word 46 queries
Prosit	0.5313	0.4568	0.5357	0.4620	0.3448	0.4695
k doc. /□*n* terms	3/20 0.5571 3/30 **0.5742** 3/40 0.5608 5/20 0.5339 5/30 0.5272 10/20 0.5395	3/10 0.4463 3/20 0.4503 5/20 0.4401 10/10 0.4367 10/20 **0.4643** 10/30 0.4483	3/30 0.5635 3/40 **0.5684** 3/50 0.5627 5/40 0.5460 10/30 0.5345 10/40 0.5307	3/30 0.4802 3/40 0.4768 5/30 0.4805 5/40 **0.4853** 5/50 0.4718 10/40 0.4812	3/15 0.2956 5/15 0.3410 5/20 0.3527 10/10 0.3593 10/15 **0.3736** 10/20 0.3707	3/50 0.4995 5/60 0.5091 5/75 **0.5230** 5/100 0.5137 10/60 0.4998 10/75 0.5076

Table 9: Mean average precision using blind-query expansion (Prosit model)

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. To illustrate this phenomenon, Figure 1 depicts 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 can be seen from the "Lnu-ltc" model, with a greater improvement however. 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, reached a maximum point and then subsequently fell slowly (with a greater variability for the Prosit model however). When a few terms were added to the original query however, the Prosit model usually performed at lower levels than did the Okapi. When this number of additional terms

was increased however, the Prosit model tended to result in better mean average precision than did the Okapi scheme. However, when more than 100 terms are added, the Okapi model produced a better retrieval effectiveness than the Prosit model.

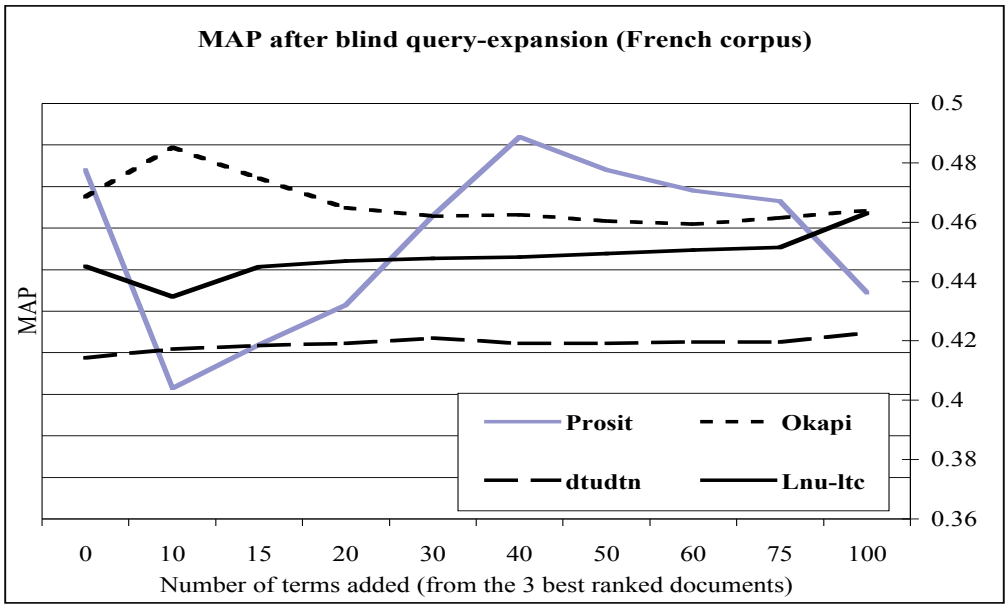

Figure 1: Mean average precision using blind-query expansion within different retrieval models

4. Data Fusion

For the each language, we may assume that different indexing and search models would retrieve different pertinent and non-relevant items and that combining different search models should improve retrieval effectiveness. More precisely, when combining different indexing schemes we would expect to improve recall due to the fact that different document representations may retrieve different pertinent items (Vogt & Cottrell 1999). On the other hand, when combining different search schemes, we would 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 (that can be 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 order to combine two or more indexing schemes, we evaluated various fusion operators, and their precise descriptions are listed in Table 10. For example, the Sum RSV operator indicates that the combined document score (or the final retrieval status value) is simply the sum of the retrieval status value (RSV_k) of the corresponding document D_k computed by each single indexing scheme (Fox & Shaw 1994). We can thus see from Table 10 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 γ_i (usually equal to 1) in order to favor the *i*th more efficient retrieval scheme.

In addition to using these data fusion operators, we also considered the round-robin approach, whereby in turn we take one document from all individual lists and remove 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 *i*th result list, we needed to compute the average of the RSV_k (denoted $Mean^i$) and the standard deviation (denoted $Stdev^i$). Based on these values, we would 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 10, Min^i (Max^i) denotes the minimal (maximal) RSV value in the *i*th result list.

Sum RSV	SUM ($\gamma_i \cdot RSV_k$)
Norm Max	SUM (γ_i • (RSV$_k$ / Max^i))
Norm RSV	SUM [γ_i • ((RSV$_k$ − Min^i) / (Max^i − Min^i))]
Z-Score	γ_i • [((RSV$_k$ − Mean^i) / Stdev^i) + δ^i] with δ^i = [(Mean^i − Min^i) / Stdev^i]

Table 10: Data fusion combination operators used in this study

Table 11 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 sometimes improve retrieval effectiveness (for the French or Russian corpora however, no improvement can be found). When combining two retrieval models, the Z-score scheme tended to produce the best, or at least, a good performance. In Table 11, under the heading "Z-scoreW", we attached a weight of 2 to the Prosit model, and 1.5 to the Okapi model.

	Mean average precision				
Query TD Model	English word queries	French word 49 queries	Finnish 4-gram 45 queries	Russian word 34 queries	Portuguese word 46 queries
Okapi expand doc/term	3/15 0.5581	3/10 **0.4851**	0/0 0.5389	0/0 **0.3800**	10/20 0.4731
Prosit expand doc/term	3/10 0.5427	10/30 0.4484	3/40 0.5684	0/0 0.3448	10/50 0.5030
Round-robin	**0.5699**	0.4693	0.5647	0.3545	0.4847
Sum RSV	0.5461	0.4665	0.5597	0.3695	0.5154
Norm Max	0.5592	0.4777	0.5718	0.3580	0.5157
Norm RSV	0.5575	0.4838	0.5703	0.3580	0.5188
Z-Score	0.5580	0.4839	**0.5731**	0.3577	0.5175
Z-ScoreW	0.5582	0.4796	0.5716	0.3572	**0.5231**

Table 11: Mean average precision using different combination operators (γ_i = 1, with blind-query expansion)

Run name	Language	Query	Index	Model	Query expansion	Combined	MAP
UniNEfr1	French	TD TD	word word	dtu-dtn Prosit	5 best docs / 40 terms 10 best docs / 30 terms	Round-robin	0.4437
UniNEfr2	French	TD TD	word word	Prosit Okapi	10 best docs / 30 terms 3 best docs / 10 terms	Z-Score	**0.4849**
UniNEfr3	French	TDN TDN	word word	Prosit dtu-dtn	5 best docs / 20 terms 10 best docs / 30 terms	Z-ScoreW	0.4785
UniNEfi1	Finnish	TD TD	4-gram word	Prosit Prosit	3 best docs / 40 terms 3 best docs / 20 terms	Z-ScoreW	0.4967
UniNEfi2	Finnish	TD TD TD	4-gram word 4-gram	Prosit Prosit Okapi	3 best docs / 40 terms 3 best docs / 20 terms 3 best docs / 20 terms	Sum RSV	**0.5453**
UniNEfi3	Finnish	TDN TDN	4-gram word	Prosit Prosit	3 best docs / 30 terms 3 best docs / 20 terms	Z-ScoreW	0.5454
UniNEru1	Russian	TD TD	word word	Prosit Lnu-ltc	3 best docs / 20 terms	Round-robin	**0.3546**
UniNEru2	Russian	TD TD	word word	Prosit Okapi		Z-score	0.3545
UniNEru3	Russian	TDN TDN	word word	Prosit Okapi	10 best docs / 15 terms 5 best docs / 15 terms	Round-robin	0.4070
UniNEpt1	Portuguese	TD TD	word word	Okapi Prosit	5 best docs / 15 terms 10 best docs / 10 terms	Norm RSV	0.5004
UniNEpt2	Portuguese	TD TD	word word	Prosit Lnu-ltc	5 best docs / 30 terms 10 best docs / 15 terms	Z-score	0.5105
UniNEpt3	Portuguese	TD TD	word word	Okapi Prosit	10 best docs / 20 terms 10 best docs / 50 terms	Norm RSV	**0.5188**

Table 12: Description and mean average precision (MAP) of our official runs

Finally, in Table 12 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 to be found per language. In the "UniNEfi1" run, we removed all documents appearing in the year 1994 (in order to search all newspaper articles that described events occurring in the year 1995. However, 66 (over 413) relevant items have been published in year 1994).

Conclusion

In this fifth CLEF evaluation campaign, we proposed a general stopword list and stemming procedure for the Portuguese language. Currently it is not clear if a stemming procedure, such as the one we suggested whereby only inflectional suffixes were removed from nouns and adjectives, could result in better retrieval effectiveness than a stemming approach that takes both inflectional and derivational suffixes into account. In order to achieve better retrieval results, we used a data fusion approach based on the Z-score, where it was required that document (and query) representation be based on two or three indexing schemes.

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

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

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

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

Chen, A. (2003). Cross-language retrieval experiments at CLEF 2002. In C. Peters, M. Braschler, J. Gonzalo, & M. Kluck, (Eds), *Advances in Cross-Language Information Retrieval*, (pp. 28-48), Springer-Verlag, Berlin, LNCS #2785.

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

Hedlund, T., Airio, E., Keskustalo, H., Lehtokangas, R., Pirkola, A. & Järvelin, K. (2004). Dictionary-based cross-language information retrieval: Learning experiences from CLEF 2000–2002. *IR Journal*, 7(1-2), 99-119.

Lovins, J.B. (1968). Development of a stemming algorithm. *Mechanical Translation and Computational Linguistics*, 11(1), 22-31.

Porter, M.F. (1980). An algorithm for suffix stripping. *Program*, 14, 130-137.

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. (2003). Report on CLEF-2003 monolingual tracks Fusion of probabilistic models for effective monolingual retrieval. In *Proceedings CLEF-2003*, (pp. 179-188). Trondheim.

Savoy, J. (2004a). Combining multiple strategies for effective monolingual and cross-lingual retrieval. *IR Journal*, 7(1-2), 121-148.

Savoy, J. (2004b). Report on CLIR task for the NTCIR-4 evaluation campaign. In *Proceedings NTCIR-4*, (pp 178-185). Tokyo: NII.

Singhal, A., Choi, J., Hindle, D., Lewis, D.D. & Pereira, F. (1999). AT&T at TREC-7. In *Proceedings TREC-7*, (pp. 239-251). Gaithersburg: NIST Publication #500-242.

Sproat, R. (1992). *Morphology and Computation*. Cambridge, MA: The MIT Press.

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