Report on CLEF-2003 Monolingual Tracks: Fusion of Probabilistic Models for Effective Monolingual 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 third participation in the CLEF evaluation campaign, our first objective was to propose more effective and general stopword lists for the Swedish, Finnish and Russian languages along with an improved, more efficient and simpler stemming procedure for these three languages. Our second goal was to suggest a combined search approach based on a data fusion strategy that would work with various European languages. Included in this combined approach is a decompounding strategy for the German, Dutch, Swedish and Finnish languages.

Introduction

Based on our experiments of last year [Savoy 2002], we participate in French, Spanish, German, Italian, Dutch, Swedish, Finnish and Russian monolingual tasks without to rely on a dictionary. This paper presents the approaches we used in the monolingual tracks and is organized as follows: Section 1 contains an overview of our nine test-collections while Section 2 describes our general approach to building stopword lists and stemmers for use with languages other than English. In Section 3, we suggest a simple decompounding algorithm that could be used to decompound German, Dutch, Swedish and Finnish words. Section 4 evaluates two probabilistic models and nine vector-space schemes using the nine test-collections. Finally, Section 5 presents and evaluates various data fusion operators, together with our official runs.

1. Overview of the Test-Collections

The corpora used in our experiments included newspapers such as the *Los Angeles Times* (1994, English), *Glasgow Herald* (1995, English), *Le Monde* (1994, French), *La Stampa* (1994, Italian), *Der Spiegel* (1994/95, German) and *Frankfurter Rundschau* (1994, German), *NRC Handelsbald* (1994/95, Dutch), *Algemeen Dagblad* (1995/95, Dutch) *Tidningarnas Telegrambyrå* (1994/95, Swedish), *Aamulehti* (1994/95, Finnish), and *Izvestia* (1995, Russian). As an additional source of information, we included various articles edited by news agencies such as *EFE* (1994/95, Spanish), and the Swiss news agency (1994/95, available in French, German and Italian but without parallel translation).

As shown in Table 1a and 1b, these corpora are of various sizes, with the Spanish collection being the biggest and the German, English and Dutch collections second. Ranking third are the French, Italian and Swedish corpora, then somewhat smaller is the Finnish collection and finally the Russian collection is clearly the smallest. Across all the corpora the mean number of distinct indexing terms per document is relatively similar (around 112), but this number is a little bit larger for the English collection (156.9) and smaller for the Swedish corpus (79.25).

Tables 1a and 1b compare also 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 18.63 with the corresponding median being 7). These findings indicate that each collection contains numerous queries, yet only a rather small number of relevant items are found. For each collection, 60 queries have been created. However, relevant documents cannot be found for each request and each language. For the English collection, the Queries #149, #161, #166, #186, #191, and #195 do not have any relevant items; for the French corpus, these requests are #146, #160, #161, #166, #169, #172, #191, #194; for the German collection (Queries #144, #146, #170, #191); for the Spanish collection (Queries #169, #188, #195); for the Italian collection (Queries #144, #146, #158, #160, #169, #170, #172, #175, #191); for the Dutch collection (Queries #160, #166, #191, #194); for the Swedish collection (Queries #146, #160, #167, #191, #194, #197, #198); for the Finnish corpus (Queries #141, #144, #145, #146, #160, #167, #169, #175, #182, #186, #188, #189, #191, #194, #195). Appearing for the first time in a CLEF evaluation campaign is the Russian corpus, for which we have only 28 requests.

During the indexing process of our automatic runs, we retained only the following logical sections from the original documents: <TITLE>, <HEADLINE>, <TEXT>, <LEAD>, <LEAD1>, <TX>, <LD>, <TI> and <ST>. From the topic descriptions we automatically removed certain phrases such as "Relevant document report …", "Find documents …", "Trouver des documents qui parlent …", "Sono valide le discussioni e le decisioni …", "Relevante Dokumente berichten …" or "Los documentos relevantes proporcionan información …".

	English	French	German	Spanish
Size (in MB)	579 MB	331 MB	668 MB	1,086 MB
# of documents	169,477	129,806	294,809	454,045
# of distinct terms	426,757	355,691	1,666,538	774,263
Number of distinct indexing terms / document				
Mean	156.9	118.5	111.9	112.9
Standard deviation	118.77	95.72	100.06	55.75
Median	129	89	84	100
Maximum	1,881	1,621	2,424	642
Minimum	2	3	1	5
Number of queries	54	52	56	57
Number rel. items	1,006	946	1,825	2,368
Mean rel. / request	18.63	18.19	32.59	41.54
Standard deviation	28.61	33.16	36.95	57.37
Median	7	8	24	22
Maximum	139 (#Q:157)	193 (#Q:181)	226 (#Q:181)	303 (#Q:181)
Minimum	1 (#Q:141)	1 (#Q:141)	1 (#Q:160)	1 (#Q:175)

Table 1a: Test-collection statistics

	Italian	Dutch	Swedish	Finnish	Russian
Size (in MB)	363 MB	540 MB	352 MB	137 MB	68 MB
# of documents	157,558	190,604	142,819	55,344	16,716
# of distinct terms	560,087	883,953	767,504	1,444,232	345,728
Number of distinct indexing terms / document					
Mean	116.4	110	79.25	114	124.5
Standard deviation	88.24	107.03	64.00	91.35	124.53
Median	84	77	62	87	41
Maximum	1,395	2,297	1,547	1,946	1
Minimum	1	1	1	1	1,769
Number of queries	51	56	53	45	28
Number rel. items	809	1,577	889	483	151
Mean rel./ request	15.86	28.16	16.77	10.73	5.39
Standard deviation	20.32	43.10	25.09	15.78	7.11
Median	8	14.5	11	5	3
Maximum	110 (#Q:197)	226 (#Q:181)	170 (#Q:181)	82 (#Q:181)	31 (#Q:192)
Minimum	1 (#Q:145)	1 (#Q:195)	1 (#Q:141)	1 (#Q:149)	1 (#Q:147)

Table 1b: Test-collection statistics

2. Stopword Lists and Stemming Procedures

In order to define general stopword lists, we first accounted for the top 200 most frequent words found in the various languages, together with articles, pronouns, prepositions, conjunctions or very frequently occurring verb forms (e.g., to be, is, has, etc.). As compared to last year's stopword lists [Savoy 2002], we only modified those for the Swedish and Finnish languages, and we created a new one for the Russian language (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 430 words for Italian, 463 for French, 603 for German, 351 for Spanish, 1,315 for Dutch, 747 for Finnish, 386 for Swedish and 420 for Russian.

Once it removes high-frequency words, an indexing procedure generally applies a stemming algorithm in an attempt to conflate word variants into the same stem or root. In developing this procedure for various European languages, 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 try to

reduce various word declensions into the same stem, such as those used in the German, Finnish and Russian languages.

More sophisticated schemes have already been proposed for the removal of derivational suffixes (e.g., "-ize", "-ably", "-ship" in the English language), 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, our stemming approach tried to remove some derivational suffixes (e.g., "communicateur" -> "communiquer", "faiblesse" -> "faible"). For the Dutch language we used the Kraaij & Pohlmann's stemmer [Kraaij 1996]. Our various stemming procedures can be found at www.unine.ch/info/clef/. Currently, it is not clear whether a stemming procedure such ours removes only inflectional suffixes from nouns and adjectives, and better retrieval effectiveness may be achieved by a stemming approach that also accounts for verbs or that removes both inflectional and derivational suffixes.

Finally, diacritic characters are usually not present in English collections (with some exceptions, such as "résumé"); and as with the Italian, Dutch, Finnish, Swedish, German, Spanish and Russian languages, these characters are replaced by their corresponding non-accentuated letter. For this latter language, we convert and normalize the Cyrillic Unicode characters into Latin alphabet (perl script available at www.unine.ch/clef/).

3. Decompounding Words

Most European languages manifest other morphological characteristics with compound word constructions being just one example (e.g., handgun, worldwide). In German for example, compound words are widely used and they may cause more difficulties than do those in English. For example, an insurance company would be "Versicherungsgesellschaft" ("Versicherung" + "S" + "Gesellschaft"). However the morphological marker ("S") is not always present (e.g., "Atomtests" built as "Atom" + "Tests"), and sometimes the letter "S" belongs to the decompounded word (e.g., "Wintersports" for "Winter" + "Sports"). In Finnish, we also encounter similar constructions as such as "rakkauskirje" ("rakkaus" + "kirje" for love & letter) or "työviikko" ("työ" + "viikko" for work & week). Recently, Braschler [2003] shows that decompounding German words may significantly improve retrieval performance.

Our proposed decompounding approach shares some similarity with Chen's algorithm [2002]. Before using it, we create a word list composed of all words appearing in the given collection (without stemming). Associated with each word, we also store the number of its occurrences in the collection (some examples are given in Table 2).

computer	2452	port	1091
computers	79	ports	2
sicherheit	6583	sport	1483
sicher	4522	sports	199
heit	4		
		winter	1643
bank	9657	winters	148
bund	7032	wintersport	44
bundes	2884	wintersports	2
bundesbank	1453		
präsident	24041		

Table 2: Examples of German words included in our words list

In order to present an overview of our decompounding approach, we will take as an example the German word "Computersicherheit," composed of "Computer" + "Sicherheit" (security). This compound word does not appear in our German word list as depicted in Table 2, so our algorithm starts the decompounding process by attempting to split a word following the k = 4 last letters (given the two strings "computersicher" and "heit"). During the entire procedure, we only consider words having a length greater than a given threshold (fixed at 3 for all languages in our experiments). If both components appear in the word list, then we have a candidate for decompounding; otherwise the k limit is increased by one. Since, in our case, the string "computersiche" does not appear in the German word list, splitting is rejected. When k = 9, our algorithm will find the word "computers" in the word list, but will fail to find the word "icherheit". With k = 10, our algorithm will find both the word "computer" and "sicherheit" in the German word list (see Table 2) and this solution becomes the top level decompounding suggestion. Recursively, the system now tries to decompound the two parts, namely the words "computer" and "sicherheit". During this recursive process, the system is allowed to ignore some short sequences of letters at the end of a word (such as "-s" or "-es" in German, or "-s" for the Swedish language)

because such morphological markers may indicate the genitive form (such as "'s" in the noun phrase "John's book").

After this generative part, the system responds a tree of possible formats in which the compound construction can be broken down, and with each component, we find the number of its occurrences in the corpus. In our example, the answer will be (computer 2452, sicherheit 6583 (sicher 4522, heit 4)). Thus, from this result, we know that the word "Sicherheit" appears 6583 times in the corpus, and we may consider decompounding this term into the words "sicher" and "heit". From this we can add (or replace) the compound word in the document (or in the request) by all decompound candidates ("computer" + "sicherheit", and "computer" + "sicher" + "heit" in our case) or only by decompounding only the minimum number of terms ("computer" + "sicherheit" in our case).

However, when faced with multiple candidates, our algorithm will try to select the single "best" one. To achieve this, our system will consider the total number of occurrences for the component words and if this value is greater than the number of occurrences for the compound construction, the decompounded candidate will be selected. In our example, the system will not decompound the word "Sicherheit" because the number of occurrences of the words "sicher" (4522) and "heit" (4) will not produce a total (4526) greater than the number of occurrences of the word "sicherheit" (6583).

If we consider the German word "Bundesbankpräsident" (president of the (German) federal bank), the generative part of our algorithm would return (bundesbank 1453 (bund 7032, bank 9657), präsident 24041) and the final decompounding approach would return (bund 7032, bank 9657, präsident 24041). In this case, the number of occurrences of "bundesbank" (1453) is smaller than the sum of the occurrences of the words "bund" and "bank". However, our approach does not always generate the appropriate components of a compounded term. For example, based on the compound construction "wintersports", the system answers with (winter 1643, port 1091) instead of (winter 1643, sport 1483). This problem is due to the fact that the first part of our approach ignores backtracking and will stop when it encounters the first splitting of the compound into two parts.

4. Indexing and Searching Strategy

In order to obtain a broader view of the relative merit of various retrieval models, we first adopted a binary indexing scheme within which each document (or request) is 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 could account for the term occurrence frequency (retrieval model notation: "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, a cosine normalization could prove beneficial and each indexing weight could vary within the range of 0 to 1 (retrieval model notation: "ntc-ntc", Table 3 depicts the exact weighting formulation).

Other variants might also be created. For example, the tf component may be computed as $0.5 + 0.5 \cdot [\text{tf} / \max \text{tf in a document}]$ (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 1996], "doc=dtu" [Singhal 1999].

Besides the previous models based on the vector-space approach, we also considered probabilistic models. In this vein, we used the Okapi probabilistic model [Robertson 2000] within with:

$$K = k_1 \cdot [(1 - b) + b \cdot (l_i / avdl)]$$

represents the ratio between the length of D_i measured by l_i (sum of tf_{ij}) and the collection mean noted by avdl. In Table 3, the value of nt_i indicates the number of distinct indexing terms including in the representation of D_i.

As a second probabilistic approach, we implemented the Prosit (PRObabilistic Sift of Information Terms) approach [Amati 2002a, 2002b] which is based on the following indexing formula:

$$w_{ij} = Inf1_{ij} \cdot Inf2_{ij} = (1 - Prob1_{ij}) \cdot Inf2_{ij} \quad \text{with}$$

$$Prob1_{ij} = tfn_{ij} / (tfn_{ij} + 1) \quad \text{with } tfn_{ij} = tf_{ij} \cdot \log_2[1 + ((C \cdot \text{mean dl}) / l_i)]$$

$$Inf2_{ij} = -\log_2[1 / (1+l_j)] - tfn_{ij} \cdot \log_2 [l_j / (1+l_j)] \quad \text{with } l_j = tc_j / n$$

in which tc_j indicates 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 are fixed according to values listed in Table 4.

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	$w_{ij} = \left((k_1 + 1) \cdot tf_{ij}\right) \Big/ \left(K + tf_{ij}\right)$	Lnu	$w_{ij} = \dfrac{\left(1 + \ln(tf_{ij})\right) \Big/ \left(\ln(\text{mean tf}) + 1\right)}{(1 - \text{slope}) \cdot \text{pivot} + \text{slope} \cdot nt_i}$
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	$w_{ij} = \left(\ln(tf_{ij}) + 1\right) \cdot idf_j \Big/ \sqrt{\sum_{k=1}^{t} \left(\left(\ln(tf_{ik}) + 1\right) \cdot idf_k\right)^2}$		
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

Language	Index	b	k_1	avdl	C	mean dl
English	word	0.8	2	800	1.5	167
French	word	0.75	3	900	1.25	182
Spanish	word	0.4	1.2	400	1.75	157
German	word	0.5	1.5	600	3	152
German	5-gram	0.3	1	500	2.5	475
Italian	word	0.55	1.5	800	1.25	165
Dutch	word	0.8	3	600	2.25	110
Dutch	5-gram	0.6	1.2	600	1.75	362
Finnish	word	0.75	2	900	1.25	114
Finnish	5-gram	0.6	1.2	800	2	539
Swedish	word	0.7	2	500	3	79
Swedish	4-gram	0.75	2	900	1.75	292
Russian	word	0.7	2	800	1.5	124
Russian	5-gram	0.75	1.2	750	1.75	451
Russian	4-gram	0.75	1.2	750	1.75	468

Table 4: Parameter setting for the various test-collections

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, Spanish and Italian collections using words as indexing units. The evaluation of our two probabilistic models and nine vector-space schemes are given in Table 5a.

In order to represent German, Dutch, Swedish, Finnish and Russian documents and queries, we considered the n-gram, decompounded and word-based indexing schemes. The resulting mean average precision for these various indexing approaches is shown in Table 5b (German and Dutch corpora), in Table 5c (Swedish and Finnish languages) and in Table 5d (Russian collection).

It was observed that pseudo-relevance feedback (blind-query expansion) seems to be a useful technique for enhancing retrieval effectiveness. In this study, we adopted Rocchio's approach [Buckley 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 we enlarged the query by the 10 to 175 terms provided by the 3 or 10 best-retrieved articles.

	Mean average precision			
Query TD Model	English 54 queries	French 52 queries	Spanish 57 queries	Italian 51 queries
Prosit	48.19	**52.01**	47.23	47.17
doc=Okapi, query=npn	**48.83**	51.64	**48.85**	**48.80**
doc=Lnu, query=ltc	44.51	48.26	45.79	45.32
doc=dtu, query=dtn	43.17	46.58	45.03	45.71
doc=atn, query=ntc	45.55	45.48	44.04	45.77
doc=ltn, query=ntc	34.68	39.01	42.40	42.56
doc=ntc, query=ntc	27.12	32.74	27.08	28.90
doc=ltc, query=ltc	28.14	34.41	29.74	28.63
doc=lnc, query=ltc	33.89	37.98	33.52	32.68
doc=bnn, query=bnn	15.97	24.01	26.48	25.33
doc=nnn, query=nnn	6.50	12.27	19.84	22.36

Table 5a: Mean average precision of various single searching strategies (monolingual)

	Mean average precision					
Query TD Model	German words 56 queries	German decompound 56 queries	German 5-gram 56 queries	Dutch words 56 queries	Dutch decompound 56 queries	Dutch 5-gram 56 queries
Prosit	42.14	45.53	42.88	**47.15**	48.36	39.41
doc=Okapi, query=npn	**44.54**	**46.93**	**44.27**	46.86	**48.73**	**40.23**
doc=Lnu, query=ltc	40.64	45.44	39.63	43.38	45.08	33.63
doc=dtu, query=dtn	42.60	43.95	39.08	42.69	43.78	33.82
doc=atn, query=ntc	40.98	43.67	40.36	41.92	43.52	36.43
doc=ltn, query=ntc	39.07	39.32	38.57	38.45	39.51	32.47
doc=ntc, query=ntc	27.40	32.64	31.59	29.27	30.36	29.42
doc=ltc, query=ltc	28.85	36.02	32.76	30.97	32.41	28.24
doc=lnc, query=ltc	30.16	35.93	32.10	31.39	33.15	28.53
doc=bnn, query=bnn	23.63	23.31	21.07	26.14	26.80	21.16
doc=nnn, query=nnn	15.97	10.85	9.78	11.35	10.64	9.82

Table 5b: Mean average precision of various single searching strategies (German & Dutch collections)

	Mean average precision					
Query TD Model	Swedish words 53 queries	Swedish decompound 53 queries	Swedish 4-gram 53 queries	Finnish words 45 queries	Finnish decompound 45 queries	Finnish 5-gram 45 queries
Prosit	39.26	40.86	**40.23**	46.35	46.96	**49.03**
doc=Okapi, query=npn	**39.98**	**41.43**	40.05	46.54	46.61	48.97
doc=Lnu, query=ltc	38.03	39.82	37.87	**48.73**	**47.31**	46.03
doc=dtu, query=dtn	38.14	40.32	36.40	44.44	44.78	43.54
doc=atn, query=ntc	36.56	37.85	39.95	42.91	43.99	48.56
doc=ltn, query=ntc	33.81	35.49	36.11	42.47	43.11	42.94
doc=ntc, query=ntc	25.08	26.82	26.13	32.73	33.46	35.64
doc=ltc, query=ltc	26.57	28.65	25.46	37.27	38.34	37.72
doc=lnc, query=ltc	26.91	29.17	29.03	36.93	39.18	37.21
doc=bnn, query=bnn	19.75	21.89	25.67	17.95	15.17	20.06
doc=nnn, query=nnn	11.55	11.75	12.47	13.85	13.21	14.83

Table 5c: Mean average precision of various single searching strategies (Swedish & Finnish collections)

The results depicted in Tables 6 (depicting our best results) indicate that the optimal parameter setting seems to be collection-dependant. Moreover, performance improvement also seems to be collection dependant (or language dependant), with no improvement for the English corpus yet an increase of 8.55% for the Spanish corpus (from a mean average precision of 51.71 to 56.13), 9.85% for the French corpus (from 48.41 to 53.18), 12.91% for the Italian language (41.05 to 46.35) and 13.26% for the German collection (from 41.25 to 46.72, combined model, Table 6b).

Query TD Model	Mean average precision Russian words extended stemmer 28 queries	Russian words light stemmer 28 queries	Russian 5-gram 28 queries	Russian 4-gram 28 queries
Prosit	36.69	34.89	30.44	**34.43**
doc=Okapi, query=npn	34.26	34.58	30.31	32.51
doc=Lnu, query=ltc	36.34	**36.30**	27.36	29.75
doc=dtu, query=dtn	32.67	32.95	28.49	30.55
doc=atn, query=ntc	**37.06**	33.22	**31.29**	31.41
doc=ltn, query=ntc	29.55	30.89	23.83	22.05
doc=ntc, query=ntc	33.47	30.14	28.69	27.39
doc=ltc, query=ltc	32.34	28.74	26.40	27.52
doc=lnc, query=ltc	32.58	24.47	20.65	21.88
doc=bnn, query=bnn	14.84	15.23	13.13	9.05
doc=nnn, query=nnn	12.27	11.41	7.95	5.83

Table 5d: Mean average precision of various single searching strategies (Russian collection)

Query TD Model	Mean average precision English 54 queries	French 52 queries	Spanish 57 queries	Italian 51 queries
doc=Okapi, query=npn	**48.83**	51.64	48.85	48.80
5 docs / 10 best terms	48.79	51.33	52.74	52.97
5 docs / 15 best terms	48.15	**51.91**	52.87	**53.39**
5 docs / 20 best terms	47.37	51.30	**53.02**	52.35
10 docs / 10 best terms	45.70	49.81	52.51	51.33
10 docs / 15 best terms	44.10	48.59	52.55	51.17
10 docs / 20 best terms	45.62	49.68	52.79	51.94

Table 6a: Mean average precision using blind-query expansion

Query TD Model	Mean average precision German words 56 queries	German decompound 56 queries	German 5-gram 56 queries	Dutch words 56 queries	Dutch decompound 56 queries	Dutch 5-gram 56 queries
Okapi	44.54	46.93	44.27	46.86	48.73	40.23
k doc. / m terms	5/10 46.46	5/10 50.32	5/50 **47.26**	5/10 52.32	5/10 54.60	5/100 43.12
	5/20 47.83	5/20 51.40	5/100 46.96	5/30 53.39	5/30 54.79	5/150 43.32
	5/40 **48.39**	5/50 **51.64**	5/125 46.88	5/50 **54.14**	5/40 **55.56**	5/200 **43.90**
	10/10 45.98	10/15 50.32	10/40 46.46	10/15 51.26	10/15 53.07	10/100 42.34
	10/15 46.31	10/30 50.20	10/100 46.50	10/20 51.14	10/20 52.81	10/150 42.67
	10/20 46.08	10/40 50.33	10/125 46.59	10/40 51.72	10/30 53.77	10/200 42.54

Table 6b: Mean average precision using blind-query expansion (German & Dutch collections)

Query TD Model	Mean average precision Swedish words 53 queries	Swedish decompound 53 queries	Swedish 4-gram 53 queries	Finnish words 45 queries	Finnish decompound 45 queries	Finnish 5-gram 45 queries
Prosit	39.26	40.86	40.23	46.35	46.96	49.03
k doc. / m terms	3/20 **45.93**	3/10 **48.01**	3/30 42.13	3/20 52.50	3/10 52.03	3/15 50.98
	3/30 44.50	3/15 46.23	3/40 42.16	3/30 **52.71**	3/20 **53.37**	3/50 49.44
	3/60 42.59	3/40 43.58	3/50 **42.57**	3/40 50.04	3/30 52.93	3/125 49.06
	5/20 43.29	5/30 47.15	5/30 39.44	5/20 49.69	5/10 48.82	5/30 52.45
	5/30 43.86	5/40 46.66	5/40 41.10	5/30 47.90	5/15 47.85	5/60 **52.92**
	5/40 43.40	5/50 46.29	5/50 41.37	5/50 49.77	5/20 48.85	5/75 52.67

Table 6c: Mean average precision using blind-query expansion (Swedish & Finnish collections)

Query TD Model	Mean average precision Russian words extended stemmer 28 queries	Russian words light stemmer 28 queries	Russian 5-gram 28 queries	Russian 4-gram 28 queries
doc=Okapi, query=npn	34.26	34.58	30.31	32.51
5 docs / 20 best terms	**34.81**	32.68	29.27	30.76
5 docs / 30 best terms	32.46	34.69	29.10	30.45
5 docs / 40 best terms	31.87	**34.81**	29.64	30.62
10 docs / 20 best terms	30.84	31.30	30.25	29.92
10 docs / 30 best terms	29.24	33.00	30.07	30.17
10 docs / 40 best terms	29.28	30.24	30.03	29.84
10 docs / 50 best terms	27.99	28.88	29.32	29.46

Table 6d: Mean average precision using blind-query expansion (Russian collection)

5. Data Fusion

For the English, French, Spanish, Italian and Russian languages, we assumed that the n-gram indexing and word-based document representation approaches are distinct and independent sources of evidence regarding the content of documents. For the German, Dutch, Swedish and Finnish languages, we added the decompounding indexing approach in our documents (and queries) representation scheme.

In order to combine these two and three indexing schemes respectively, we evaluated various fusion operators, as suggested by Fox and Shaw [Fox 1994]. Table 7 shows their precise description. For example, the combSUM 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. CombNBZ specifies that we multiply the sum of the document scores by the number of retrieval schemes that are able to retrieve the corresponding document. In Table 7, we can see that both the combRSV% and combRSVnorm 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. In addition to use 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.

combMAX	MAX ($\alpha_i \cdot RSV_k$)
combMIN	MIN ($\alpha_i \cdot RSV_k$)
combSUM	SUM ($\alpha_i \cdot RSV_k$)
combANZ	SUM ($\alpha_i \cdot RSV_k$) / # of nonzero (RSV_k)
combNBZ	SUM ($\alpha_i \cdot RSV_k$) * (# of nonzero (RSV_k))
combRSV%	SUM ($\alpha_i \cdot (RSV_k / MAX_{RSV})$)
combRSVnorm	SUM [$\alpha_i \cdot ((RSV_k\text{-}MIN_{RSV}) / (MAX_{RSV}\text{-}MIN_{RSV}))$]

Table 7: Data fusion combination operators

Query TD Model	Mean average precision English 54 queries	French 52 queries	Spanish 57 queries	Italian 51 queries	Russian 28 queries
Okapi expand doc/term	0/0 48.83	10/10 49.81	10/10 52.51	10/20 51.94	10/20 31.30
Prosit expand doc/term	3/15 50.99	5/30 52.30	10/10 50.19	10/50 50.82	5/30 35.41
combMAX	48.83	52.27	50.19	50.82	35.41
combMIN	2.88	42.77	8.21	18.62	24.96
combSUM	51.13	53.58	51.89	51.87	**35.68**
combANZ	37.95	53.25	43.97	50.05	35.60
combNBZ	51.11	53.66	51.89	51.86	35.65
combRSV%	**53.60**	54.50	53.30	53.58	34.43
combRSVnorm	53.25	**54.69**	**53.49**	54.37	34.30
round-robin	50.24	52.61	53.16	**54.47**	34.11

Table 8a: Mean average precision using different combination operators ($\alpha_i = 1$, with blind-query expansion)

Run name	Language	Query	Index	Model	Query expansion	combined	MAP
UniNEfr	French	TD	word	Okapi	10 best docs / 10 terms		
		TD	word	Prosit	5 best docs / 30 terms	round-robin	52.61
UniNEfr2	French	TD	word	Okapi	10 best docs / 10 terms		
		TD	word	Prosit	5 best docs / 30 terms	RSV%	**54.50**
UniNEsp	Spanish	TD	word	Okapi	10 best docs / 10 terms		
		TD	word	Prosit	10 best docs / 10 terms	RSVnorm	**53.80**
UniNEsp2	Spanish	TD	word	Okapi	5 best docs / 10 terms		
		TD	word	Prosit	10 best docs / 10 terms	RSVnorm	53.69
UniNEde	German	TD	word	Prosit	5 best docs / 20 terms		
		TD	decomp.	Prosit	10 best docs / 40 terms		
		TD	5-gram	Prosit	5 best docs / 175 terms	RSVnorm	54.58
UniNEde2	German	TD	word	Pro+Oka	5 best docs / 20 terms		
		TD	decomp.	Pro+Oka	10 best docs / 40 terms		
		TD	5-gram	Pro+Oka	5 best docs / 175 terms	sumRSV	**56.03**
UniNEit	Italian	TD	word	Okapi	10 best docs / 20 terms		
		TD	word	Prosit	10 best docs / 50 terms	RSV%	**52.23**
UniNEit2	Italian	TD	word	Okapi	10 best docs / 20 terms		
		TD	word	Prosit	10 best docs / 50 terms	sumRSV	51.56
UniNEnl	Dutch	TD	word	Okapi	10 best docs / 20 terms		
		TD	decomp.	Okapi	10 best docs / 20 terms		
		TD	5-gram	Prosit	10 best docs / 150 terms	round-robin	**50.65**
UniNEnl2	Dutch	TD	word	Okapi	10 best docs / 20 terms		
		TD	decomp.	Okapi	10 best docs / 20 terms		
		TD	5-gram	Prosit	10 best docs / 150 terms	sumRSV	50.24
UniNEsv	Swedish	TD	word	Pro+Oka	3 best docs / 15 terms		
		TD	decomp.	Pro+Oka	3 best docs / 15 terms		
		TD	4-gram	Pro+Oka	3 best docs / 40 terms	RSV%	48.19
UniNEsv2	Swedish	TD	word	Pro+Oka	5 best docs / 30 terms		
		TD	decomp.	Pro+Oka	5 best docs / 50 terms		
		TD	4-gram	Pro+Oka	5 best docs / 30 terms	RSVnorm	**48.69**
UniNEfi	Finnish	TD	word	Prosit	5 best docs / 30 terms		
		TD	decomp.	Prosit	5 best docs / 15 terms		
		TD	5-gram	Prosit	3 best docs / 125 terms	sumRSV	**54.51**
UniNEfi2	Finnish	TD	word	Prosit	5 best docs / 30 terms		
		TD	decomp.	Prosit	5 best docs / 15 terms		
		TD	5-gram	Prosit	3 best docs / 125 terms	sumRSV	53.55
UniNEru	Russian	TDN	word	Okapi	10 best docs / 20 terms	ext. stemmer	
		TDN	word	Prosit	5 best docs / 30 terms	sumRSV	35.32
UniNEru1	Russian	TD	word	Okapi	10 best docs / 20 terms	ext. stemmer	
		TD	word	Prosit	5 best docs / 30 terms	sumRSV	31.83
UniNEru2	Russian	TD	5-gram	Okapi	10 best docs / 50 terms		
		TD	5-gram	Prosit	5 best docs / 40 terms		
		TD	4-gram	Okapi	10 best docs / 50 terms		
		TD	4-gram	Prosit	5 best docs / 40 terms	sumRSV	**32.77**
UniNEru3	Russian	TDN	word	Okapi	10 best docs / 10 terms	ext. stemmer	
		TDN	word	Prosit	5 best docs / 20 terms	sumRSV	**42.24**

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

Tables 8a and 8b depict an evaluation of various data fusion operators, comparing them to the single approach using the Okapi and the Prosit probabilistic models. As shown in these tables, the combRSVnorm or combRSV% fusion strategies usually improve the retrieval effectiveness over the best single retrieval model.

	Mean average precision			
Query TD Model	German 56 queries	Dutch 56 queries	Swedish 53 queries	Finnish 45 queries
Prosit word doc/term	5/20 48.40	10/20 51.14 (Okapi)	3/60 42.59	5/30 47.90
Prosit decomp doc/term	10/40 51.40	10/20 51.81 (Okapi)	3/40 43.58	5/15 47.85
Prosit n-gram doc/term	5/175 49.46	10/150 44.23	3/40 42.16	3/125 49.06
combMAX	49.97	44.23	42.94	50.22
combMIN	35.54	6.30	33.91	33.36
combSUM	53.71	50.24	47.58	54.51
combANZ	47.85	31.90	41.14	49.25
combNBZ	53.70	50.81	47.29	**55.60**
combRSV%	54.46	53.99	47.95	54.49
combRSVnorm	**54.58**	**54.30**	**48.12**	54.16
round-robin	50.83	50.65	44.14	48.73

Table 8b: Mean average precision using different combination operators ($\gamma_i = 1$, with blind-query expansion)

Conclusion

In this fourth CLEF evaluation campaign, we proposed a general stopword list and stemming procedure for eight European languages (excluding English). Currently it is not clear if a stemming procedure such as that suggested and that only removes inflectional suffixes from nouns and adjectives, could produce better retrieval effectiveness than a stemming approach that takes both inflectional and derivational suffixes into account. We also suggested a simple decompounding approach for the German, Dutch, Swedish and Finnish language. In order to achieve better retrieval performance, we used a data fusion approach, one requiring that document (and query) representation be based on two or three indexing schemes.

Acknowledgments

The author would 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 2002a] Amati, G., Carpineto, C. & Romano, G. (2002). Italian monolingual information retrieval with PROSIT. In Proceedings of CLEF-2002, (pp. 145-151). Roma.

[Amati 2002b] 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 2003] Braschler, M. & Ripplinger, B. (2003). Stemming and decompounding for German text retrieval. In Proceedings 25th European Conference in IR (pp. 177-192). Berlin: Springer.

[Buckley 1996] 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 2002] Chen, A. (2002). Cross-language retrieval experiments at CLEF-2002. In Proceedings of CLEF-2002, (pp. 5-20). Roma.

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

[Kraaij 1996] Kraaij, W. & Pohlmann, R. (1996). Viewing stemming as recall enhancement. In Proceedings of the ACM-SIGIR'96, (pp. 40-48). New York: The ACM Press.

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

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

[Robertson 2000] 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 2002] Savoy J. (2002). Report on CLEF-2002 experiments: Combining multiple sources of evidence. In Proceedings of CLEF-2002, (pp. 31-46). Roma.

[Singhal 1999] 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.