The Effectiveness of Combining Information Retrieval Strategies for European Languages

Jaap Kamps Maarten de Rijke
Language & Inference Technology Group
ILLC, University of Amsterdam
http://lit.science.uva.nl/

ABSTRACT

Building an effective Information Retrieval system requires various design choices, ranging from the weighting scheme to the type of morphological normalization. The combination of runs has become a standard technique to reap the benefits of different run types. Until now, systematic studies of the effectiveness of combination strategies have only been carried out for English. This paper provides an exploratory overview of the effectiveness of combination methods in nine European languages. We demonstrate that the combination of effective information retrieval strategies can lead to significant improvements of retrieval effectiveness. Furthermore, we analyze the relative impact of retrieving more relevant documents and of improved ranking of relevant documents. The experimental evidence is obtained using the 2003 test-suite of the cross-language evaluation forum (CLEF).

1. INTRODUCTION

When building an information retrieval (IR) system, one faces a variety of choices, ranging from the text representation used (whether to use morphological normalization or not, and which type of query formulation to use), to the choice of search strategy (which weighting scheme to use, and whether to use blind feedback). Unfortunately, there are few equivocal answers to these choices. Hoping to reap the benefits of more than one strategy, researchers have naturally resorted to combining different strategies.

The combination of retrieval runs is one of the recurring themes in IR. It goes back at least to Fox and Shaw [11], and it re-occurs at many retrieval evaluation exercises. A recent overview of combination approaches is given in [7], describing combinations of document representation, query formulations, ranking algorithms, and search systems. Here, we focus on the combination of runs made on different document representations, but using the same retrieval settings and weighting scheme. In particular, we build three indexes using different morphological normalization methods. Until now, systematic studies of combination methods have focused exclusively on English. We extend the domain of application to nine European languages (Dutch, English, Finnish, French, German, Italian, Russian, Spanish, and Swedish). There are notable differences between these languages, such as the complexity of inflectional and derivational morphology [12]. Because of such differences, results obtained for English need not carry over to monolingual IR in other European languages.

Below, we formulate a number of hypothesis we want to address in this paper. Our first hypothesis states that the differences between English and other European languages are negligible with respect to combination methods.

Hypothesis 1. The effectiveness of combining retrieval strategies does not differ between English and other European languages.

In the experiments of Lee [14], the standard combination methods lead to impressive improvements. However, there is evidence that the effectiveness of combination methods greatly diminishes with the increasing effectiveness of retrieval systems [3, 6]. So, can combination methods deliver a significant improvement over high performing retrieval strategies? This motivates our second hypothesis:

Hypothesis 2. The combination of high performing retrieval strategies does not lead to significant improvement over the best performing single strategy.

In general, high performing strategies will be more similar in the (relevant) documents they retrieve. This implies that very few relevant documents are retrieved by only one of the retrieval strategies. There are two principal ways in which a combination run can improve [4]. The first is increased recall by having more relevant documents than any single strategy. The second is increased precision by better ranking of relevant documents. The diminishing effectiveness of combination methods could be explained if fruitful combinations depend to a large extent on different strategies retrieving complementary sets of relevant documents.

Hypothesis 3. The effectiveness of combination methods is due to additional relevant documents being retrieved in the combined strategy.

We restrict ourselves to monolingual retrieval; an excellent overview of combination methods for bi- and multilingual retrieval (using a distributed index) is provided in [20].

The remainder of the paper is organized as follows. We first give an overview of combination methods and earlier results. In Section 3, we describe the experimental set-up

SAC'04, March 14-17, 2004, Nicosia, Cyprus

and the set of monolingual base runs for the nine European languages that are used in our experiments. Then, in Section 4 we apply the standard combination methods to those base runs, and evaluate their effectiveness. We also try to determine the causes of the improvements of run combinations. In Section 5, we discuss our findings in the light of combination methods results for English collections.

2. COMBINATION METHODS

2.1 Standard Combination Methods

We focus on the following standard combination methods as introduced by Fox and Shaw [11, p.245/246]: *combMAX* (take the maximal similarity score of the individual runs); *combMIN* (take the minimal similarity score of the individual runs); *combSUM* (take the sum of the similarity scores of the individual runs); *combANZ* (take the sum of the similarity scores of the individual runs, and divide by the number of non-zero entries); *combMNZ* (take the sum of the similarity scores of the individual runs, and multiply by the number of non-zero entries); and *combMED* (take the median similarity score of the individual runs).

Similarity score distributions may differ radically across runs. We apply the six combination methods to *normalized* similarity scores. That is, instead of directly applying the methods to the retrieval status values (RSVs), we follow [13, p.185] and normalize them into $[0, 1]$ using the minimal and maximal similarity scores. We calculate

$$\mathsf{sim}_{\text{norm}} = \frac{\mathsf{sim}_{\text{original}} - \mathsf{sim}_{\text{min}}}{\mathsf{sim}_{\text{max}} - \mathsf{sim}_{\text{min}}}$$

with $\mathsf{sim}_{\text{min}}$ ($\mathsf{sim}_{\text{max}}$) the minimal (maximal) RSV score over all topics in the run.

Fox and Shaw [11] found combSUM to be the best performing combination method. Lee [13, 14] conducted extensive experiments with the Fox and Shaw combination rules. Lee [13] introduced the normalization method for RSVs which allows for the combination of runs using different weighting schemes. In Lee [14], combMNZ emerges as the best combination rule. The work of Aslam and Montague [1, 2] confirms Lee's findings.

2.2 Linear Combination Methods

Not all of the base runs per language exhibit the same level of performance. Instead of the unweighted combination methods discussed above, we could also assign different weights to the different base runs. This refinement of Fox and Shaw [11]'s combSUM rule is known as *linear combination* in Vogt and Cottrell [23]. We calculate new similarity scores for documents in the following way. To combine n runs run_1 through run_n, we set $\mathsf{sim}_{\text{new}} = \Sigma_n^{i=1} \mathsf{w}_i \cdot \mathsf{sim}_i$, where w_i is the relative weight of run_i. To simplify our set-up, we apply the linear combination only to pair-wise combinations of base runs. For this case, the linear combination function can be simplified by using a single combination factor $\lambda \in [0, 1]$ representing the relative weight of the first mentioned run:[1] $\mathsf{sim}_{\text{new}} = \lambda \cdot \mathsf{sim}_1 + (1 - \lambda) \cdot \mathsf{sim}_2$.

2.3 Rationale of Combination Methods

After observing the small overlap between sets of retrieved documents, Saracevic and Kantor [19] found that the greater the number of runs in which a document was retrieved, the greater the odds of a particular document being relevant. Turtle and Croft [22, p.218] observe that the combination of two runs could be effective even if one base run retrieved a subset of the relevant documents of the other. Thus, in this case, the observed improvement is entirely due to a better ranking of the relevant documents. Two theoretical rationales for the effectiveness of combination methods have been formulated by Belkin et al. [4, p.339]: "The first derives from the observation that different representations ... retrieve different sets of documents (both relevant and non-relevant). ... The other rationale ... suggests that the more sources of evidence are available ... the more accurate judgment of the probability of relevance ... will be." These two rationales have distinct and testable consequences. If the first holds, the runs improves due to the retrieval of a larger set of relevant documents than the underlying base runs. If the second is valid, this boils down to a better ranking of relevant documents in the combined run in comparison with the underlying base runs. Note that these two effects are the only ways in which the score of a run can improve.

[1] For $\lambda = 0.5$ this is similar to the combSUM function used by Fox and Shaw [11], and discussed in Section 4.1 above.

3. EXPERIMENTAL SET-UP

Experimental evaluation is done on the test-suite of the Cross-Language Evaluation Forum (CLEF), using the 2003 documents, topics and assessments for monolingual Dutch; English; Finnish; French; German; Italian; Russian; Spanish; and Swedish. We use the FlexIR system developed at the University of Amsterdam [16]. FlexIR is implemented in Perl and supports many types of preprocessing, scoring, indexing, and retrieval tools. All our base runs use the Lnu.ltc weighting scheme [5] to compute the similarity between a query and a document. For the experiments on which we report in this paper, we fixed *slope* at 0.2; the pivot was set to the average number of unique words per document.

Blind feedback was applied to expand the original query with related terms. Term weights were recomputed by using the standard Rocchio method [18], where we considered the top 10 documents to be relevant and the bottom 500 documents to be non-relevant. We allowed at most 20 terms to be added to the original query.

To determine whether the observed differences between two retrieval approaches are statistically significant, we used the bootstrap method [9, 10]. We take 100,000 resamples, and look for significant improvements (one-tailed) at significance levels of 0.95 (*); 0.99 (**); and 0.999 (***).

3.1 Monolingual Base Runs

A variety of approaches has been applied to monolingual retrieval in non-English [12]. These can be divided in two categories: 1. language-dependent approaches, such as stemming and lemmatizing; and 2. language-independent approaches like (character) n-grams of various lengths that sometimes span word boundaries (see [15] for an overview).

We decided to focus on three types of runs:

Words. This is a vanilla base run that indexes the words as encountered in the collection. We do some sanitizing: diacritics are mapped to the unmarked character, and all characters are case-folded. E.g., the German '`Raststätte`' (English: motorway restaurant) is indexed as '`raststatte`'.

Stems. We use the set of stemmers implemented in the string processing language Snowball [21], which is specifi-

Table 1: MAP of CLEF 2003 topics (rows 2–4) and of combinations for CLEF 2003 topics (rows 5–10).

	Dutch	*English*	*Finnish*	*French*	*German*	*Italian*	*Russian*	*Spanish*	*Swedish*
Words	**0.4800**	**0.4483**	0.3175	0.4313	0.3785	0.4631	0.2551	0.4405	0.3485
Stems	0.4652	0.4273	0.3998	**0.4511**	0.4504	**0.4726**	0.2536	**0.4678**	0.3707
4-Grams	0.4488	0.3731	**0.4676**	0.4142	**0.4639**	0.3883	**0.2871**	0.4545	**0.3751**
combMAX	0.4642	0.4146	0.4336	0.4503	0.4686	0.4434	0.2793	0.4674	0.3877
combMIN	0.5062	**0.4519**	0.4279	0.4557	0.4235	0.4354	0.2614	0.4611	0.3990
combSUM	**0.5190**	0.4509	**0.4850**	**0.4821**	**0.4941**	**0.4816**	0.3000	**0.4887**	**0.4392**
combANZ	0.4765	0.4360	0.4285	0.4544	0.4412	0.4658	**0.3046**	0.4624	0.3855
combMNZ	0.5181	0.4401	0.4778	0.4749	0.4817	0.4727	0.2915	0.4849	0.4290
combMED	0.4652	0.4273	0.3998	0.4511	0.4504	0.4726	0.2536	0.4678	0.3707
%Change	+8.1%***	+0.8%	+3.7%	+6.9%	+6.5%*	+1.9%	+6.1%	+4.5%*	+17.1%**

cally designed for creating stemming algorithms for use in IR. It is partly based on the familiar Porter stemmer for English [17], and provides stemming algorithms for all languages considered here. We perform the same sanitizing operations as for the word-based run.

4-Grams. We apply character 4-grams not spanning word-boundaries. E.g., '`Information Retrieval`' is indexed as '`info nfor form orma rmat mati atio tion retr etri trie riev ieva eval`'. Character n-grams are an old technique for improving retrieval effectiveness, dating back at least to [8]. Again, we perform the same sanitizing operations as for the word-based run.

Rows 2–4 in Table 1 contain the results of our base runs; the best run per language is in bold-face. There is no unique best indexing approach: words are best for Dutch and English; stems are best for French, Italian, and Spanish; and 4-grams are best for the remaining languages.

4. RESULTS AND ANALYSIS

4.1 Standard Combination Methods

For all nine languages, we apply the six standard combination methods to the three base runs. The resulting scores are listed in rows 5–10 in Table 1; the best base run score and the best combined run score are in bold-face. The percentage difference between the best base run and the best combined run is indicated in the bottom row of the table.

Let us analyze the outcomes. First, the best combination always improves over the best base run, ranging from 0.8% (English) to 17.1% (Swedish). The improvement over the best underlying base run is significant for four languages (Dutch, German, Spanish, and Swedish). Which combination method is the most effective? When looking at the relative performance of the six combination methods, combMIN scores best for English, combANZ for Russian, and combSUM for the other languages. Generalizing over languages, combSUM performs the best (for example when considering the mean reciprocal rank of the combination methods). CombSUM is the only combination method that improves over the best performing base run for all languages. Moreover, there is no language for which one of the other combination methods scores significantly better than combSUM.

Our results concur with the original experiments of Fox and Shaw [11] for English mentioned before. It contradicts studies of combination method effectiveness in English that claim the superiority of combMNZ. In our experiments the combSUM rule outperforms combMNZ for all nine languages. Lee [14, p.269] attributes the difference with the prevalence of combSUM in Fox and Shaw's experiments to the lack of score normalization. Interestingly, we did perform score normalization, and still combSUM prevails.

4.2 Linear Combination Methods

We looked at all combinations of pairs of base runs, while varying the combination factor with steps of 0.05. For all 27 combinations of base runs, Figure 1 plots the mean average precision (MAP) scores against the used combination factor. The best pairwise combined runs per language are:

Language	Combination	MAP	%Change
Dutch	0.40 4-Grams, 0.60 Words	0.5323	+10.9%***
English	0.50 Words, 0.50 Stems	0.4855	+8.3%***
Finnish	0.45 4-Grams, 0.55 Stems	0.4993	+6.8%
French	0.50 4-Grams, 0.50 Stems	0.4824	+6.9%*
German	0.45 4-Grams, 0.55 Stems	0.5025	+8.3%**
Italian	0.65 Words, 0.35 Stems	0.4919	+4.1%*
Russian	0.50 Words, 0.50 Stems	0.2937	+2.3% over
		4-Grams;	15.1%* over Words)
Spanish	0.45 4-Grams, 0.55 Stems	0.4904	+4.8%**
Swedish	0.60 4-Grams, 0.40 Words	0.4371	+16.5%***

The best linear combination is better than the best base run for all nine languages; for seven languages, the improvement is significant. Moreover, for Russian, the combination of words and stems improves significantly over the words base run (but not over the 4-grams base run not used in the combination). Observe that all curves in Figure 1 are convex. If we were to draw a straight line between the scores of the base runs, then the linear combination curve is above this line. This indicates the tendency of the combination to improve over the weighted mean of the base run scores. In case both base runs have a comparable performance level, this can result in a considerable improvement (e.g., the Dutch and Swedish combinations). In case there is a sizable difference in the performance of both base runs, a similar gain over the weighted mean of the base runs will result in a much lower improvement over the best underlying base run (e.g., the Finnish 4-grams combination).

How does one find the optimal value of the combination factor? Fortunately, the factors seem to be stable across topic sets, so if earlier topic sets are available, the close-to-optimal values can be obtained experimentally. If earlier topic sets are not available, there are some rules of thumb. If one expects that one base run will be superior to the other ones, one may assign a somewhat higher weight to the superior run. Finally, one may resort to assigning equal weights to the base runs, and thus effectively be using combSUM.

4.3 Analysis of Run Combinations

We will now give a detailed analysis of two combined strategies that show considerable improvement. We choose

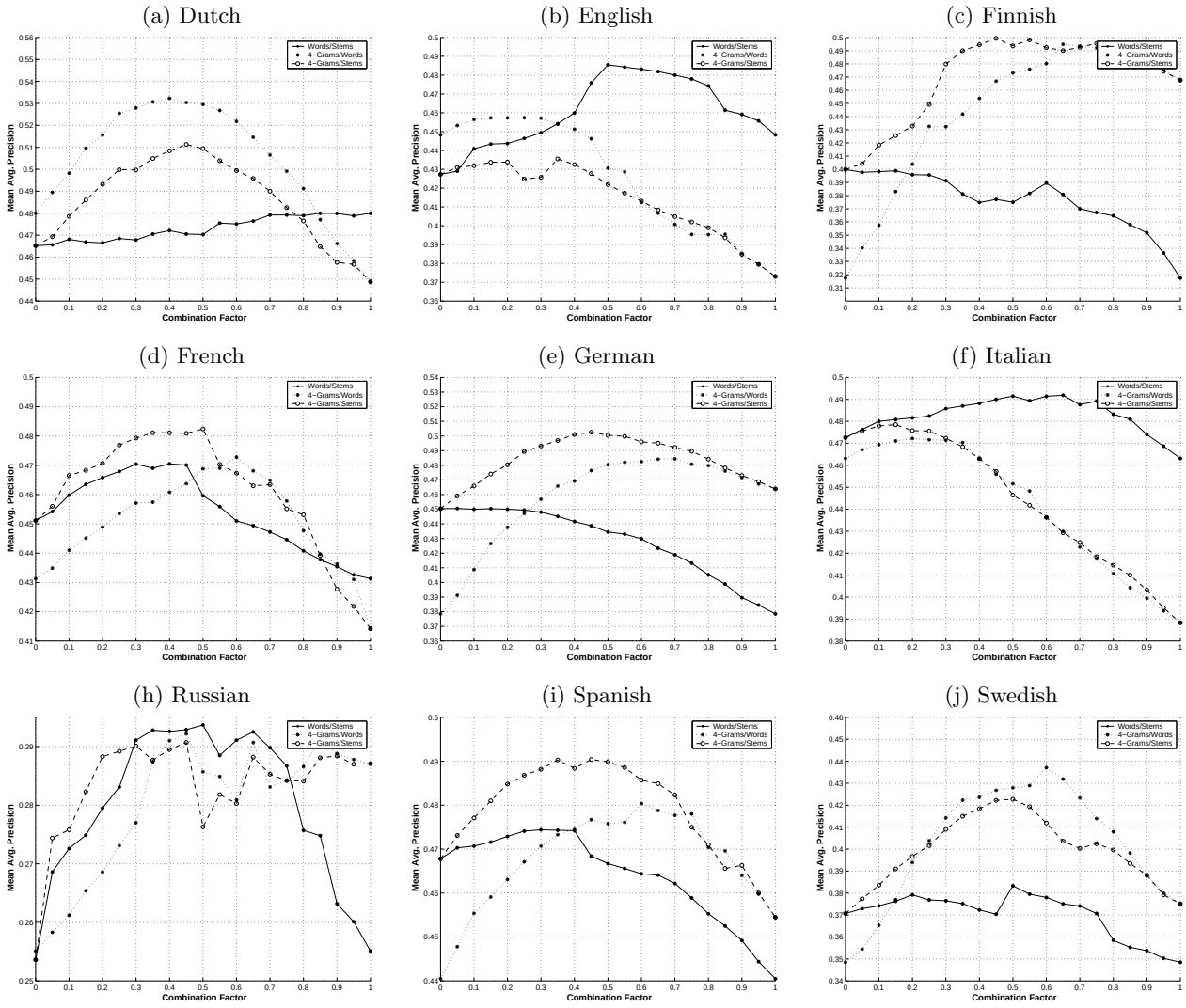

Figure 1: Pairwise combined runs using linear combination.

two runs of the linear combination from Section 4.2: the Dutch combination of 4-grams (with weight 0.40) and words (0.60); and the Swedish combination of 4-grams (0.60) and words (0.40). The Dutch base runs retrieve 1428 (words) and 1428 (4-grams) documents, the combined run 1476; the Swedish base runs retrieve 806 (words) and 906 (4-grams) documents, the combined run 920. Thus, the Dutch combination retrieves 48 additional relevant documents (3.4%), and the Swedish combination retrieves 14 additional relevant documents (1.5%). So there is a clear recall-enhancing effect, but does this explain the performance improvement?

To answer this question we modified our evaluation script so that the run results can be manipulated before the MAP scores are calculated. Table 2 shows the analysis for the Dutch and Swedish combined runs. We compare the combined run against both underlying base runs. First, we can treat documents in the combined run as if occurring at their base run's rank. This neutralizes the effect of documents receiving a better ranking due to the combined evidence, which allows us to pin-point the contribution of additional relevant documents in the combined run. Second, we can ignore the ranking of documents that did not occur in the underlying base run. This neutralizes the effect of additionally retrieved relevant documents, and allows us to isolate the contribution of documents receiving a better ranking.

The outcome is interesting. Only for the Swedish combination (920 relevant documents) against the words base run (806 relevant documents), we see that both effects are clearly present. However, it is more fair to compare the combined run against the base run retrieving the largest set of relevant documents. Here, the situation is quite different: the additionally retrieved documents play a very limited role, and the gain in performance can almost exclusively be attributed to documents receiving a better ranking. We see the same for Dutch, the improved ranking of documents accounts for almost all of the improvement. This finding corroborates the observations of [22].

5. CONCLUSIONS

In this paper we investigated the effectiveness of combin-

Table 2: Analysis of effects causing combination improvement.

	Dutch			Swedish		
	Avg.Prec.	% Diff.	Rel.docs	Avg.Prec.	% Diff.	Rel.docs
Words	**0.4800**		1428	0.3485		806
4-Grams	0.4488		1428	**0.3751**		906
4-Grams/Words	0.5323	+10.9%	1476	0.4371	+16.5%	920
Against the words base run						
Docs at base run rank	0.4830	+0.6%		0.3638	+4.3%	
Restrict to base run docs	0.5262	+9.6%		0.4176	+19.8%	
Against the 4-grams base run						
Docs at base run rank	0.4542	+1.2%		0.3764	+0.3%	
Restrict to base run docs	0.5260	+17.2%		0.4357	+16.2%	

ing IR strategies for nine European languages. We explored the combination of runs made on indexes using different morphological normalization methods, but using the same retrieval settings and weighting scheme. Our results in Section 4.1 show considerable differences between English and the other European languages, thereby refuting our first hypothesis. The respective improvements range from 0.8% for English to 17.1% for Swedish. The small improvement for English using combSUM, and the fact that combMNZ does not improve over the best base run, helps to explain the earlier negative results on English in the literature.

Our second hypothesis stated that combining effective retrieval strategies does not lead to significant improvement of retrieval effectiveness. Our results refute this hypothesis; we found improvements for all nine languages, and significant improvements for four languages. The combSUM combination method improves for all nine languages, and improves significantly for four languages. We obtain even better results for linear combination using optimal settings. Again, we find improvements for all nine languages, with significant improvements for seven languages.

We also analyzed in detail the relative contribution of additional relevant documents, and of better ranking of relevant documents. Although there is, at the outset, a clear recall enhancing effect, its contribution to the score is very limited. Almost all of the improvement can be accounted for by the improved ranking of relevant documents already retrieved by the best underlying base run. This refutes our third hypothesis and sheds light on the conditions required for the successful application of combination methods.

6. ACKNOWLEDGMENTS

Jaap Kamps was supported by the Netherlands Organization for Scientific Research (NWO) under project numbers 400-20-036 and 612.066.302. Maarten de Rijke was supported by grants from NWO, under project numbers 612-13-001, 220-80-001, 365-20-005, 612.069.006, 612.000.106, 612.000.207, and 612.066.302.

REFERENCES

[1] J. A. Aslam and M. Montague. Bayes optimal metasearch: A probablistic model for combining the results of multiple retrieval systems. In *SIGIR'00*, pp. 379–381, 2000.

[2] J. A. Aslam and M. Montague. Models for metasearch. In *SIGIR'01*, pp. 276–284, 2001.

[3] S. M. Beitzel et al. Disproving the fusion hypothesis: An analysis of data fusion via effective information retrieval strategies. In *ACM SAC'03*, pp. 823–827, 2003.

[4] N. J. Belkin, C. Cool, W. B. Croft, and J. P. Callan. The effect of multiple query representations on information retrieval system performance. In *SIGIR'93*, pp. 339–346, 1993.

[5] C. Buckley, A. Singhal, and M. Mitra. New retrieval approaches using SMART: TREC 4. In *TREC-4*, pp. 25–48, 1996.

[6] A. Chowdhurry, O. Frieder, D. Grossman, and C. McCabe. Analyses of multiple-evidence combinations for retrieval strategies. In *SIGIR'01*, pp. 394–395, 2001.

[7] W. B. Croft. Combining approaches to information retrieval. In *Advances in Information Retrieval*, pp. 1–36. Kluwer, 2000.

[8] T. de Heer. The application of the concept of homeosemy to natural language information retrieval. *IP&M*, 18:229–236, 1982.

[9] B. Efron. Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, 7:1–26, 1979.

[10] B. Efron and R. J. Tibshirani. *An Introduction to the Bootstrap*. Chapman and Hall, New York, 1993.

[11] E. A. Fox and J. A. Shaw. Combination of multiple searches. In *TREC-2*, pp. 243–252, 1994.

[12] V. Hollink, J. Kamps, C. Monz, and M. de Rijke. Monolingual document retrieval for European languages. *IR*, 6, 2003.

[13] J. H. Lee. Combining multiple evidence from different properties of weighting schemes. In *SIGIR'95*, pp. 180–188, 1995.

[14] J. H. Lee. Analyses of multiple evidence combination. In *SIGIR'97*, pp. 267–276, 1997.

[15] P. McNamee and J. Mayfield. Character n-gram tokenization for European language text retrieval. *IR*, 6, 2003.

[16] C. Monz and M. de Rijke. Shallow morphological analysis in monolingual information retrieval for Dutch, German and Italian. In *CLEF-2001*, pp. 262–277, 2002.

[17] M. Porter. An algorithm for suffix stripping. *Program*, 14 (3):130–137, 1980.

[18] J. J. Rocchio, Jr. Relevance feedback in information retrieval. In *The SMART Retrieval System: Experiments in Automatic Document Processing*, chapter 14, pp. 313–323. Prentice-Hall, 1971.

[19] T. Saracevic and P. B. Kantor. A study of information seeking and retrieving. III. searchers, searches, overlap. *JASIST*, 39:197–216, 1988.

[20] J. Savoy. Combining multiple strategies for effective monolingual and cross-language retrieval. *IR*, 6, 2003.

[21] Snowball. Snowball stemmers, 2003. `http://snowball.tartarus.org/`.

[22] H. Turtle and W. B. Croft. Evaluation of an inference network-based retrieval model. *ACM TOIS*, 9:187–222, 1991.

[23] C. C. Vogt and G. W. Cottrell. Predicting the performance of linearly combined IR systems. In *SIGIR'98*, pp. 190–196, 1998.