Combining Morphological and Ngram Evidence for Monolingual Document Retrieval

Jaap Kamps and **Christof Monz** and **Maarten de Rijke**

Language & Inference Technology Group, ILLC, U. of Amsterdam
Nieuwe Achtergracht 166, 1018 WV Amsterdam, The Netherlands
E-mail: {kamps,christof,mdr}@science.uva.nl

Abstract

We report on experiments in which we merged the results of linguistically informed and linguistically ignorant approaches to retrieval for European languages. We found that even high-quality base runs can be improved by means of fairly simple techniques for merging them with other runs, although the improvements no longer seem to be as dramatic as those reported on previous experiments on smaller collections than we used and with retrieval engines that are not as highly optimized as the one used in our experiments.

1 Introduction

It's a widely held belief that deep linguistic analysis does more harm than it helps in document retrieval (Lewis and Sparck Jones, 1996). Morphology seems to provide the level of analysis that is appropriate for document retrieval. Especially for non-English European languages there is evidence that linguistically informed morphological analyses helps improve effectiveness. For instance, in combination with lexical-based stemming compound splitting improves retrieval effectiveness for Dutch and German (Monz and de Rijke, 2002). Unfortunately, for many European languages other than English, lexical resources are hard to obtain or even non-existent. For this reason, various teams working in document retrieval for such languages have developed language independent morphological normalization tools, often based on ngrams (CLEF, 2002).

Rather than choosing for either linguistically motivated morphological approaches or linguistically ignorant ngram-based approaches for retrieval for European languages, our strategy is to merge the results of the two approaches. Assuming that high-quality morphological and ngram-based runs identify mostly the same relevant documents, but different non-relevant documents, such combinations should yield improvements in retrieval effectiveness over both base runs.

In this paper we report on experiments carried out in monolingual document retrieval for Dutch, French, German, Italian, and Spanish, using the collections and assessments made available in the CLEF evaluation campaign (CLEF, 2002). We found that even high-quality base runs can be improved by means of fairly simple techniques for merging them with other runs, although the improvements no longer seem to be as dramatic as those reported on previous experiments on smaller collections than we used and with retrieval engines that are not as highly optimized as the one used in our experiments. The parameters that we used to create the optimal combination of runs are collection dependent but they do seem to be fairly robust across topics.

2 Experimental Setup

All experiments were carried out with the **FlexIR** system developed at the University of Amsterdam (van Hage et al., 2002). **FlexIR** is a vector-space retrieval system; for the experiments for this paper it was used with the Lnu.ltc weighting scheme and with blind feedback turned on.

Table 1 lists the characteristics of the document collections that we used for the experiments in this paper. They are part of the document collections made available within the CLEF campaign.[1]

The topics used in the experiments were Topics 41–90 and 91–140; these were the topics used at CLEF-2001 and CLEF-2002, respectively. For evaluation purposes we used the qrels provided by the CLEF organizers.

3 Three Types of Runs

For Dutch, French, German, Italian, and Spanish we created three types of runs: morphological, ngram-based and merged.

3.1 Morphological Runs

The three main morphological phenomena, i.e., inflection, derivation, and compound words, all affect

[1] As of 2002, CLEF also includes Finnish and Swedish in the monolingual track. Unfortunately, we did not have access to (linguistically informed) morphological normalization tools for these languages.

Language	Collection	Year	Documents	Size (in MB)
Dutch	Algemeen Dagblad	1994/1995	106,483	241
	NRC Handelsblad	1994/1995	84,121	299
French	Le Monde	1994	44,013	157
	SDA French	1994	43,178	86
German	Der Spiegel	1994/1995	13,979	63
	Frankfurter Rundschau	1994	139,715	320
	SDA German	1994	71,677	144
Italian	La Stampa	1994	58,051	193
	SDA Italian	1994	50,527	85
Spanish	Agencia EFE	1994	215,738	509

Table 1: The document collections used.

the effectiveness of text retrieval. Documents are not retrieved if the search key does not occur in the index. For effective retrieval morphological processing is needed in most languages to handle inflected word forms. The morphological normalization may be stemming or lemmatization. In *stemming* affixes are removed from word forms (Porter, 1980); the output is a common root or stem of different forms, which is not necessarily a real word. In (lexicon-based) lemmatization word forms are turned into base forms which are real words. Morphological analysis also allows one to split compounds into their component words.

For each of the languages we used a lexical-based stemmer, or lemmatizer, where available. For Dutch we used MBLEM, a memory-based lemmatizer developed at Tilburg University (van den Bosch and Daelemans, 1999); for French, German, Italian we used TreeTagger (Schmid, 1994), and for Spanish we used a Porter stemmer (CLEF-Neuchâtel, 2002).

For Dutch and German we complemented our lemmatizers with a compound splitter to analyze complex words such as *Autobahnraststätte* (English: highway restaurant) and *Vredesverdrag* (English: peace agreement). In addition to these noun-noun compounds there are several other forms of compounding, including verb-noun (e.g., German: *Tankstelle*, English: gas station), verb-verb (e.g., German: *spazierengehen*, English: taking a walk), noun-adjective (e.g., Dutch: *werkeloos*, English: unemployed), etc.. We used simple compound dictionaries, that consist of complex words and their parts, where each part is lemmatized; see (Monz and de Rijke, 2002) for further details.

For retrieval purposes, each document in the Dutch and German collections is analyzed and if a compound is identified, both the compound and all of its parts are added to the document. Compounds occurring in a query are analyzed in a similar way: the parts are simply added to the query, while keeping the compound.

3.2 Runs Based on Ngrams

In information retrieval ngrams have become a popular technique for identifying index terms; see, e.g., (Mayfield and McNamee, 1999; Savoy, 2001) for some recent examples of systems using ngrams. We fixed the ngram length to be the largest integer smaller than the average word length. For Dutch, German, Italian, and Spanish we used ngram length 5, and for French we used ngram length 4; see Table 2 for an overview. For each word we stored both the word itself and all possible ngrams that can be obtained from it without crossing word boundaries. For instance, the Dutch version of Topic 108 contains the phrase *maatschappelijke gevolgen* (English: societal consequences); using ngrams of length 5, this becomes:

> *maatschappelijke maats aatsc atsch tscha schap chapp happe appel ppeli pelij elijk lijke gevolgen gevol evolg volge olgen*

Some authors adopt ngram-based approaches in which ngrams are allowed to span word boundaries; see e.g., (McNamee and Mayfield, 2002). We did not find any consistent significant improvements in allowing ngrams to cross word boundaries, and stuck to our present set-up for reasons of space efficiency.

Stopword removal was done before ngrams were formed; we determined the 400 most frequent words, then removed from this list content words that we felt might be important despite their high frequency. We did not use a ‘stop ngram’ list. Diacritic characters were not replaced by the corresponding non-diacritic letters.

3.3 Merging Runs

We merged our morphological and ngram-based base runs in the following manner. First, we normalized the retrieval status values (RSVs), since different runs may have radically different RSVs. For each run we reranked these values in $[0.5, 1.0]$,

	Dutch	French	German	Italian	Spanish
Avg. word length	5.4	4.8	5.8	5.1	5.1
Ngram length	5	4	5	5	5

Table 2: Average word length and ngram length used for the ngram base runs.

using

$$RSV_i' = 0.5 + 0.5 \cdot \frac{RSV_i - min_i}{max_i - min_i},$$

and assigned the value 0.5 to documents not occurring in the top 1000; this is a variation of the Min_Max_Norm considered by Lee (Lee, 1997a). Next, we assigned new weights to the documents using a linear interpolation factor λ representing the relative weight of a run:

$$RSV_{new} = \lambda \cdot RSV_1 + (1 - \lambda) \cdot RSV_2.$$

For $\lambda = 0.5$ this is similar to the summation function used by (Fox and Shaw, 1994; Belkin et al., 1995; Lee, 1997a).

Table 3 lists our non-interpolated average precision scores for CLEF 2002, for the morphological and ngram-based base runs, and for the merged runs. The figures in brackets indicate the improvement of the merged run over the best underlying base run. For all languages, the merged run outperforms the underlying base runs. Moreover, these improvements occur at all recall levels, as illustrated by the P/R plots for German (CLEF 2002) in Figure 1. However, the relative improvements are far less dramatic than the 25% improvements reported in the literature (Lee, 1997b; Lee, 1997a), which were obtained using low-quality runs (by today's standards).

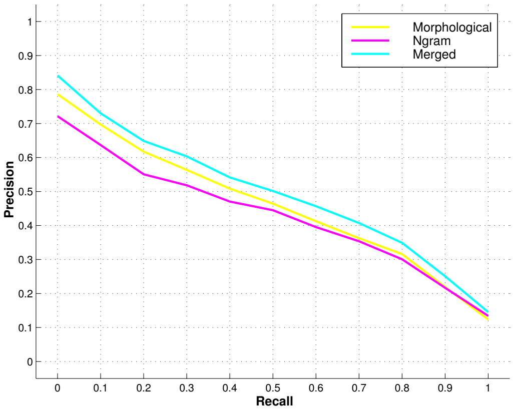

Figure 1: 11pt interpolated avg. precision for German, using the CLEF 2002 topics.

The optimal interpolation factors λ were obtained experimentally. Figure 2 suggests that the optimal interpolation is very stable across topic sets.[2]

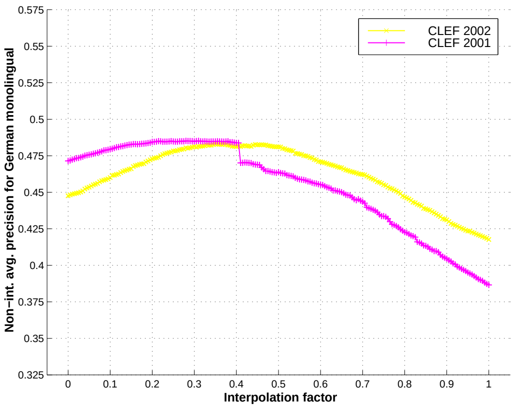

Figure 2: The interpolation factor λ. Effect on non-interpolated avg. precision scores for German, at CLEF 2001 and 2002, where $\lambda \in [0, 1]$.

Figure 3 shows that λ can be chosen from a broad interval of values without dramatic penalties in terms of non-interpolated avg. precision scores.

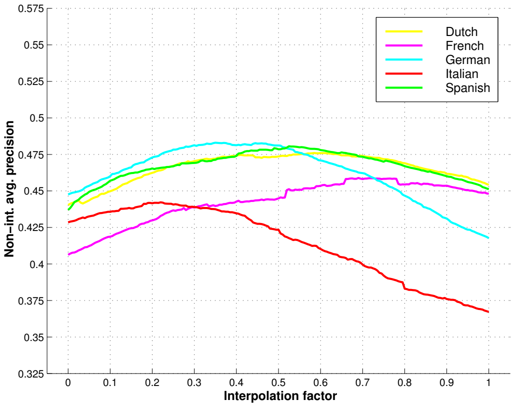

Figure 3: The interpolation factor λ. Effect on non-interpolated avg. precision scores for Dutch, French, German, Italian, and Spanish at CLEF 2002, with $\lambda \in [0, 1]$.

[2] Note that there is a marked discontinuity in the CLEF 2001 curve at 0.4; we have observed similar — but less dramatic — drops in curves for other merged runs, but have not found an unequivocal explanation yet.

	Dutch	French	German	Italian	Spanish
Morphological	0.4404	0.4063	0.4476	0.4285	0.4370
Ngram	0.4542	0.4481	0.4177	0.3672	0.4512
Merged	0.4760	0.4589	0.4830	0.4422	0.4806
	(+4.8%)	(+2.4%)	(+7.9%)	(+3.2%)	(+6.5%)

Table 3: Non-interpolated average precision scores for CLEF 2002.

4 Discussion

The following rationale has been put forward for combining (high quality) runs: try to maximize the overlap of relevant documents between the base runs, while minimizing the overlap of non-relevant documents (Lee, 1997a); this way, the RSVs of relevant documents should get a boost, but those of non-relevant documents not. The following coefficients $R_{overlap}$ and $N_{overlap}$ have been proposed for determining the overlap between two runs run_1 and run_2:

$$R_{overlap} = \frac{R_C \times 2}{R_1 + R_2} \qquad N_{overlap} = \frac{N_C \times 2}{N_1 + N_2},$$

where R_C (N_C) is the number of common relevant (non-relevant) documents, and R_i (N_i) is the number of relevant (non-relevant) documents in run_i ($i = 1, 2$). (A document is relevant if its relevance score in the qrels provided by CLEF is equal to 1.)

Table 4 shows the overlap coefficients for the base runs used to produce merged runs; the coefficients are computed over all topics.

Contrary to Lee (Lee, 1997a)'s rationale, for our high quality base runs there does not seem to be an obvious correlation between the overlap coefficients and the improvements obtained by combining them.

5 Conclusions

We reported on experiments in which we merged the results of linguistically informed and linguistically ignorant approaches to retrieval for European languages. We found that even high-quality base runs can be improved by means of fairly simple techniques for merging them with other runs, although the improvements no longer seem to be as dramatic as those reported in the literature.

Acknowledgments

We would like to thank the anonymous reviewers for their valuable feedback, and Vera Hollink for her technical support. Jaap Kamps was supported by the Netherlands Organization for Scientific Research (NWO), grant # 400-20-036. Christof Monz was supported by the Physical Sciences Council with financial support from NWO, project 612-13-001. Maarten de Rijke was supported by grants from NWO, under project numbers 612-13-001, 365-20-005, 612.069.006, 612.000.106, 220-80-001, and 612.000.207.

References

N.J. Belkin, P. Kantor, E.A. Fox, and J.A. Shaw. 1995. Combining evidence of multiple query representations for information retrieval. *Information Processing & Management*, 31(3):431–448.

CLEF-Neuchâtel. 2002. CLEF resources at the University of Neuchâtel. `http://www.unine.ch/info/clef`.

CLEF. 2002. Cross-Language Evaluation Forum. `http://www.clef-campaign.org`.

E.A. Fox and J.A. Shaw. 1994. Combination of multiple searches. In *Proceedings TREC-2*, pages 243–252.

J.H. Lee. 1997a. Analyses of multiple evidence combination. In *Proceedings SIGIR'97*, pages 267–276.

J.H. Lee. 1997b. Combining multiple evidence from different relevant feedback networks. In *Database Systems for Advanced Applications*, pages 421–430.

D.D. Lewis and K. Sparck Jones. 1996. Natural language processing for information retrieval. *Communications of the ACM*, 39(1):92–101.

J. Mayfield and P. McNamee. 1999. Indexing using both n-grams and words. In E.M. Voorhees and D.K. Harman, editors, *The Seventh Text REtrieval Conference (TREC-7)*, pages 419–423. NIST Special Publication 500-242.

P. McNamee and J. Mayfield. 2002. Scalable multilingual information access. In C. Peters, editor, *Working Notes for the the CLEF 2002 Workshop*, pages 133–140.

C. Monz and M. de Rijke. 2002. Shallow morphological analysis in monolingual information retrieval for Dutch, German and Italian. In C. Peters, M. Braschler, J. Gonzalo, and M. Kluck, editors, *Proceedings CLEF 2001*, LNCS 2406, pages 262–277. Springer Verlag.

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

J. Savoy. 2001. Report on CLEF-2001 experiments. In C. Peters, editor, *Working Notes for the CLEF 2001 Workshop*, pages 11–20. ERCIM-01-W04.

H. Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing*.

A. van den Bosch and W. Daelemans. 1999.

	Dutch	French	German	Italian	Spanish
$R_{overlap}$	0.9443	0.9606	0.9207	0.9021	0.9172
$N_{overlap}$	0.3790	0.5187	0.4180	0.4510	0.5264

Table 4: Degree of overlap among relevant and non-relevant documents for the base runs used to form the merged runs. The coefficients are computed over all CLEF 2002 topics.

Memory-based morphological analysis. In *Proceedings ACL'99*, pages 285–292.

W. van Hage, V. Hollink, J. Kamps, C. Monz, and M. de Rijke. 2002. The **FlexIR** information retrieval system. Manual, Language & Inference Technology Group, ILLC, U. of Amsterdam.