Monolingual Document Retrieval: English versus other European Languages

Jaap Kamps Christof Monz Maarten de Rijke Börkur Sigurbjörnsson

Language & Inference Technology Group
ILLC, University of Amsterdam
{kamps,christof,mdr,borkur}@science.uva.nl
http://lit.science.uva.nl/

ABSTRACT

The vast majority of research in information retrieval is done using English collections and topics. This raises questions about the effectiveness of retrieval strategies for other languages. To examine this issue, we focus on document retrieval in nine European languages. In particular, we investigate the effectiveness of language-dependent approaches to document retrieval, such as stemming and decompounding; of language-independent approaches, such as character n-gramming; and of the combination of the two types of approaches. The experimental evidence is obtained using the 2003 test-suite of the cross-language evaluation forum (CLEF).

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; H.3.4 Systems and Software; H.3.7 Digital Libraries

General Terms

Measurement, Performance, Experimentation

Keywords

Information Retrieval, Cross-Language Information Retrieval

1. INTRODUCTION

Researchers in Information Retrieval (IR) have experimented with a wide variety of approaches to document retrieval for European languages. Differences between these approaches range from the text representation being used (whether to apply 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). A recent overview of monolingual document retrieval can be found in [10]. In this paper, we focus on approaches that differ in the type of document representations that they employ, but all use the same retrieval settings and weighting scheme. In particular, we focus on different approaches to morphological normalization or tokenization. We conduct experiments on nine European languages (Dutch, English, Finnish, French, German, Italian, Russian, Spanish, and Swedish). There are notable differences between these languages that may have an impact on retrieval performance; these differences include the complexity of inflectional and derivational morphology [15].

We formulate a number of research questions that we want to address in this paper. Our main focus is on monolingual retrieval in European languages, and our overall aim is to investigate differences between languages with respect to approaches to document retrieval.

Aim 1 Investigate different document representations of English and other European languages for monolingual document retrieval.

We investigate the effectiveness of language-dependent approaches to document retrieval, which require a detailed knowledge of the particular language at hand. An well-known example of a language-dependent approach is the use of stemming algorithms. There is no consistent evidence on the effectiveness of stemming in English [9, 11].

Aim 2 Evaluate the effectiveness of stemming for European languages.

Another example of a language-dependent approach is the use of decompounding strategies for compound-rich European languages, such as Dutch and German. Compounds formed by the concatenation of words are rare in English, although exceptions like *database* exist.

Aim 3 Evaluate the effectiveness of decompounding for compound-rich European languages.

Furthermore, we investigate the effectiveness of language-independent approaches to document retrieval, which do not

DIR 2003 4th Dutch-Belgian Information Retrieval Workshop, Amsterdam, The Netherlands

depend on knowledge of the language at hand. The best example of language-independent approaches is the use of character n-gramming techniques.

Aim 4 Evaluate the effectiveness of character n-gramming for European languages.

Finally, building on the outcomes of some of the previous aims, we investigate whether both approaches to document retrieval can be fruitfully combined.

Aim 5 Evaluate the effectiveness of combining language-dependent and language-independent approaches for European languages.

The remainder of the paper is organized as follows. In Section 2, we describe the set-up of our experiments, and give a detailed overview of the used approaches to monolingual document retrieval. Then, in Section 3 we apply these approaches to monolingual document retrieval in nine European languages. Finally, in Section 4, we discuss our findings and draw some conclusions.

2. EXPERIMENTAL SET-UP

Experimental evaluation is done on the test-suite of the Cross-Language Evaluation Forum [CLEF, 2], using the 2003 documents, topics and assessments for monolingual Dutch; English; Finnish; French; German; Italian; Russian; Spanish; and Swedish. In total, there are 60 topics, collection sizes range from 16,716 documents (Russian) to 454,045 documents (Spanish).

We use the FlexIR system developed at the University of Amsterdam [17]. FlexIR is implemented in Perl and supports many types of preprocessing, scoring, indexing, and retrieval tools. FlexIR implements a number of retrieval models; for the runs in this paper we make use of the standard vector space model. All our base runs use the Lnu.ltc weighting scheme [1] 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.

Both topics and documents were stopped using the stopword lists from the Snowball stemming algorithms [21], for Finnish we used the Neuchâtel-stoplist [3]. The Russian collection and topics are encoded using the UTF-8 or Unicode character encoding, which we converted into a 1-byte per character encoding KOI8 or KOI8-R. We did all our processing, such as lower-casing, stopping, stemming, and n-gramming, in this KOI8 encoding. Finally, we converted the resulting documents and queries into the Latin alphabet using the Volapuk transliteration (using the Perl package `Convert::Cyrillic`).

Blind feedback was applied to expand the original query with related terms. Term weights were recomputed by using the standard Rocchio method [19], 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.

Finally, to determine whether the observed differences between two retrieval approaches are statistically significant, we used the bootstrap method, a non-parametric inference test [5, 6]. The method has previously been applied to retrieval evaluation by, e.g., Wilbur [22] and Savoy [20]. We take 100,000 resamples, and look for significant improvements (one-tailed) at significance levels of 0.95 (*), 0.99 (**), and 0.999 (***).

2.1 APPROACHES

A wide variety of approaches has been applied to monolingual document retrieval in non-English [10]. One can divide the approaches in two categories. The first category are language-dependent approaches, such as stemming and lemmatizing. The second category are language-independent approaches like (character) n-grams of various lengths that sometimes span word boundaries.

We decided to focus on the following types of runs:

2.1.1 Baseline

We consider as a baseline the straightforward indexing of the words as encountered in the collection. We do some limited sanitizing: diacritics are mapped to the unmarked character, and all characters are put in lower-case. Thus a string like `Information Retrieval` is indexed as the two tokens `information retrieval` and a string like the German `Raststätte` (English: motorway restaurant) is indexed as `raststatte`.

2.1.2 Stemming

The stemming or lemmatization of words is a widely used language-dependent approach to document retrieval. An overview of stemming algorithms can be found in [8]. We use the set of stemmers implemented in the Snowball language [21]. The string processing language Snowball is specifically designed for creating stemming algorithms for use in Information Retrieval. It is partly based on the familiar Porter stemmer for English [18], and provides rule-based stemming algorithms for all the nine European languages that we consider in this paper. We perform the same sanitizing operations as for the word-based run. Thus a string like `Information Retrieval` is indexed as the stems `inform retriev`.

2.1.3 Decompounding

For the compound rich languages, Dutch, German, Finnish, and Swedish, we apply a decompounding algorithm. We treat all the words occurring in the respective CLEF corpora as potential base words for decompounding, and also use their associated collection frequencies. We ignore words of length less than four as potential compound parts, thus a compound must have at least length eight. As a safeguard against oversplitting, we only regard compound parts that have a higher collection frequency than the compound itself. We consider linking elements `-s-`, `-e-`, and `-en-` for Dutch; `-s-`, `-n-`, `-e-`, and `-en-` for German; `-s-`, `-e-`, `-u-`, and `-o-` for Swedish; and none for Finnish. We prefer a split with no linking element over a split with a linking element, and a split with a single character linker over a two character linker.

Each document in the collection is analyzed and if a compound is identified, all of its parts are added to the document (while the original compound is also retained). Thus a string like the Dutch `boekenkast` (English: bookshelf) is indexed as the three tokens `boekenkast boek kast`. Compounds occurring in a query are analyzed in a similar way: the parts are simply added to the query. Since we expand both the documents and the queries with compound parts, there is no need for compound formation [13].

2.1.4 n-Gramming

Character n-gramming is a widely used language-independent approach to document retrieval. Character n-grams are an old technique for improving retrieval effectiveness, dating back at least to [4]. An excellent overview of n-gramming techniques is given in [16]. We apply character 4-grams not spanning word-boundaries, and add the n-grams to the documents, while also retaining the original words. Again, we perform the same sanitizing operations as for the word-based run. This means that the string `Information Retrieval` is indexed as the 16 tokens `information info nfor form orma rmat mati atio tion retrieval retr etri trie riev ieva eval`.

2.1.5 Combining

To combine runs, we use a weighted combination. First, we normalize the retrieval status values (RSVs), since different runs may have radically different RSVs. Following Lee [14], both scores are normalized by mapping them to the interval $[0, 1]$ using $RSV_i' = \frac{RSV_i - min_i}{max_i - min_i}$ with min_i (max_i) the minimal (maximal) RSV score over all topics in the run. Next, we assign 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 simple (but effective) combSUM function used by Fox and Shaw [7]. The interpolation factors λ were obtained from experiments on the CLEF 2002 data sets (whenever available).

3. RESULTS

3.1 BASELINE

The mean-average-precision (MAP) scores for our baseline runs are shown in Table 1. For most languages, the baseline run performs fairly well. This is certainly the case for Dutch, where the baseline runs has a MAP of 0.4800.

3.2 STEMMING

Next, we consider runs in which stemming was applied. The results are shown in Table 2. The results are mixed. On the one hand, we see a decrease in retrieval effectiveness for Dutch, English, and Russian. On the other hand, we see an increase in retrieval effectiveness for Finnish, French, German, Italian, Spanish, and Swedish. The improvements for Finnish, German, and Spanish are statistically significant.

3.3 DECOMPOUNDING AND STEMMING

Next, we consider decompounding documents and queries for the four compound-rich languages: Dutch, Finnish, German, and Swedish. After decompounding, we apply the same stemming procedure as in Section 3.2 above. The results are shown in Table 3. The results for decompounding are positive overall. For Dutch, we now see an improvement over the baseline run (unlike the case in which only stemming is applied). We also see improvements for Finnish, German, and Swedish. For all the four compound-rich languages, the score in Table 3 exceeds that of Table 2.

3.4 N-GRAMMING

Recall that our n-gram runs use character n-grams of length 4, and that we retain the original words in the index. The results are shown in Table 4. We see a decrease in performance for English and Italian, and an improvement for the other seven languages: Dutch, Finnish, French, German, Russian, Spanish, and Swedish. The improvement is significant for five of the languages, namely Finnish, German, Russian, Spanish, and Swedish. However, the decrease in performance for English is also significant.

3.5 COMBINATION

As is clear from the results above, there is no equivocal best strategy for monolingual document retrieval. For English, our baseline run scores best. For Italian, the stemmed run scores best. For the other seven languages, Dutch, Finnish, French, German, Russian, Spanish, and Swedish, 4-gramming scores best. So what is a good uniform retrieval strategy that, we hope, will do well for all languages? Since both language-dependent approaches and language-independent approaches to document retrieval have their respective merits, we consider the combination of both types of runs. In particular, we combine the decompounded (whenever available) and stemmed run with the 4-gram run. We apply a weighted combination method, also refered to as linear fusion, to combine the language-dependent and language-independent approaches to document retrieval. The results are shown in Table 5. We find only positive results: all languages improve over the baseline, even English! Even though both English runs scored lower than the baseline (one of them even signicantly lower), the combination improves over the baseline. The improvement for six of the languages, Finnish, French, German, Russian, Spanish, and Swedish, is significant.

4. DISCUSSION AND CONCLUSIONS

In this paper we have discussed a variety of approaches to monolingual document retrieval for European languages. In Section 1, we formulated a number of research questions that provided the rationale for the experiments reported in this paper.

Our second aim was to evaluate the effectiveness of stemming for European languages. Our results in Table 2 show mixed results, stemming does help for six of the languages, but hurts performance for three languages (Dutch, English, Russian). Our result for English is in line with earlier experiments [9]. Although the Russian stemmed run failed to improve for monolingual retrieval, other experiments not reported in the paper showed improvement for the English to Russian biligual runs [12]. For three languages, Finnish, German, and Spanish, the improvements of retrieve effectiveness are statistically significant.

Our third aim was to evaluate the effectiveness of decompounding for the compound-rich European languages Dutch,

Table 1: Word-based run.

	Dutch	English	Finnish	French	German	Italian	Russian	Spanish	Swedish
Word-based (baseline)	0.4800	0.4483	0.3175	0.4313	0.3785	0.4631	0.2551	0.4405	0.3485

Table 2: Snowball stemming algorithm.

	Dutch	English	Finnish	French	German	Italian	Russian	Spanish	Swedish
Stemmed	0.4652	0.4273	0.3998	0.4511	0.4504	0.4726	0.2536	0.4678	0.3707
%Change over baseline	-3.1	-4.7	+25.9	+4.6	+19.0	+2.1	-0.6	+6.2	+6.4
Stat.Significance	-	-	*	-	***	-	-	*	-

Table 3: Compound splitting and stemming algorithms.

	Dutch	Finnish	German	Swedish
Comp.Split+Stemmed	0.4984	0.4453	0.4840	0.3957
%Change over baseline	+3.8	+40.3	+27.9	+13.5
Stat.Significance	-	***	***	-

Table 4: 4-Gramming.

	Dutch	English	Finnish	French	German	Italian	Russian	Spanish	Swedish
4-Grammed	0.4996	0.4119	0.4905	0.4616	0.5005	0.4227	0.3030	0.4733	0.4187
%Change over baseline	+4.1	-8.1	+54.5	+7.0	+32.2	-8.7	+18.8	+7.4	+20.1
Stat.Significance	-	*(!)	***	-	***	-	*	*	*

Table 5: Combination of (Compound-splitting and) Stemming and 4-Gramming.

	Dutch	English	Finnish	French	German	Italian	Russian	Spanish	Swedish
Combination	0.5072	0.4575	0.5236	0.4888	0.5091	0.4781	0.2988	0.4841	0.4371
%Change over baseline	+5.7	+2.1	+64.9	+13.3	+34.5	+3.2	+17.1	+9.9	+25.4
Stat.Significance	-	-	***	**	***	-	*	***	**

German, Finnish, and Swedish. The decompounding improves scores for all four languages (see our results in Table 3). It is of interest to observe that the results for English are radically different from the results for the other languages. For English, the language-dependent techniques turn out to hurt performance. For the other European languages, we may conclude that these language-dependent approaches can help retrieval effectiveness.

Our fourth aim was to evaluate the effectiveness of character n-gramming for European languages. Our results in Table 4 show a decrease in performance for English and Italian, and an improvement for the other seven languages. The improvement is significant for Finnish, German, Russian, Spanish, and Swedish; as is the decrease for English. Again, we see a clear difference between the results for English, where n-gramming even leads to a significant drop in performance, and the other European languages. For the other European languages, we may conclude that the language-independent approach of n-gramming can help retrieval effectiveness.

In order to find out whether language-dependent and language-independent approaches have complementary retrieval enhancing effects, our fifth aim was to evaluate the effectiveness of combining the two types of approaches. Our results in Table 5 show improvement for all nine languages. The improvement for six of the languages, Finnish, French, German, Russian, Spanish, and Swedish, is significant. For eight of the languages, the combination results in the best overall score, only for Russian the 4-gram run is better than the combination. The English combination improves even though both underlying runs scored lower than the baseline (one of them even signicantly lower).

Our overall aim was to investigate differences between document representations in English and other European languages for monolingual document retrieval. Our results on the effectiveness of both language-dependent and language-independent approaches show considerable difference between English and the other languages. These differences can be explained in part by the differences in inflectional and derivational morhology between English and the other European languages [15]. This result supports an observation made in much recent work on non-English language and information processing: be careful when carrying over results from English to other languages. Because of the differences between the English language and other Indo-European languages, information retrieval results need not carry over to the latter languages.

5. ACKNOWLEDGMENTS

We thank Valentin Jijkoun for his help with the Russian collection. Jaap Kamps was supported by the Netherlands Organization for Scientific Research (NWO) under project numbers 400-20-036 and 612.066.302. Christof Monz was supported by NWO under project numbers 612-13-001 and 220-80-001. Maarten de Rijke was supported by NWO under project numbers 612-13-001, 365-20-005, 612.069.006, 612.000.106, 220-80-001, 612.000.207, and 612.066.302.

REFERENCES

C. Buckley, A. Singhal, and M. Mitra. New retrieval approaches using SMART: TREC 4. In D. K. Harman, editor, *The Fourth Text REtrieval Conference (TREC-4)*, pages 25–48. National Institute for Standards and Technology. NIST Special Publication 500-236, 1996.

CLEF. Cross language evaluation forum, 2003. `http://www.clef-campaign.org/`.

CLEF-Neuchâtel. CLEF resources at the University of Neuchâtel, 2003. `http://www.unine.ch/info/clef`.

T. de Heer. The application of the concept of homeosemy to natural language information retrieval. *Information Processing & Management*, 18:229–236, 1982.

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

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

E. A. Fox and J. A. Shaw. Combination of multiple searches. In D. K. Harman, editor, *The Second Text REtrieval Conference (TREC-2)*, pages 243–252. National Institute for Standards and Technology. NIST Special Publication 500-215, 1994.

W. Frakes. Stemming algorithms. In W. Frakes and R. Baeza-Yates, editors, *Information Retrieval: Data Structures & Algorithms*, pages 131–160. Prentice Hall, 1992.

D. Harman. How effective is suffixing? *Journal of the American Society for Information Science*, 42:7–15, 1991.

V. Hollink, J. Kamps, C. Monz, and M. de Rijke. Monolingual document retrieval for European languages. *Information Retrieval*, 7:31–50, 2004.

D. Hull. Stemming algorithms – a case study for detailed evaluation. *Journal of the American Society for Information Science*, 47:70–84, 1996.

J. Kamps, C. Monz, M. de Rijke, and B. Sigurbjörnsson. Language-dependent and language-independent approaches to cross-lingual text retrieval. In C. Peters, M. Braschler, J. Gonzalo, and M. Kluck, editors, *Cross-Language Information Retrieval, CLEF 2003*, Lecture Notes in Computer Science. Springer, 2004.

W. Kraaij and R. Pohlmann. Viewing stemming as recall enhancement. In H.-P. Frei, D. Harman, P. Schabie, and R. Wilkinson, editors, *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 40–48. ACM Press, New York NY, USA, 1996.

J. H. Lee. Combining multiple evidence from different properties of weighting schemes. In E. A. Fox, P. Ingwersen, and R. Fidel, editors, *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 180–188. ACM Press, New York NY, USA, 1995.

P. H. Matthews. *Morphology*. Cambridge University Press, 1991.

P. McNamee and J. Mayfield. Character n-gram tokenization for European language text retrieval. *Information Retrieval*, 6, 2003.

C. Monz and M. de Rijke. Shallow morphological analysis in monolingual information retrieval for Dutch, German and Italian. In C. Peters, M. Braschler, J. Gonzalo, and M. Kluck, editors, *Evaluation of Cross-Language Information Retrieval Systems, CLEF 2001*, volume 2406 of *Lecture Notes in Computer Science*, pages 262–277. Springer, 2002.

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

J. J. Rocchio, Jr. Relevance feedback in information retrieval. In G. Salton, editor, *The SMART Retrieval System: Experiments in Automatic Document Processing*, Prentice-Hall Series in Automatic Computation, chapter 14, pages 313–323. Prentice-Hall, Englewood Cliffs NJ, 1971.

J. Savoy. Statistical inference in retrieval effectiveness evaluation. *Information Processing and Management*, 33:495–512, 1997.

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

J. Wilbur. Non-parametric significance tests of retrieval performance comparisons. *Journal of Information Science*, 20:270–284, 1994.