When Stopword Lists Make the Difference

Ljiljana Dolamic and Jacques Savoy

Computer Science Department, University of Neuchâtel, 2009 Neuchâtel, Switzerland.
E-mail: {Ljiljana.Dolamic, Jacques.Savoy}@unine.ch

In this brief communication, we evaluate the use of two stopword lists for the English language (one comprising 571 words and another with 9) and compare them with a search approach accounting for all word forms. We show that through implementing the original Okapi form or certain ones derived from the *Divergence from Randomness* (DFR) paradigm, significantly lower performance levels may result when using short or no stopword lists. For other DFR models and a revised Okapi implementation, performance differences between approaches using short or long stopword lists or no list at all are usually not statistically significant. Similar conclusions can be drawn when using other natural languages such as French, Hindi, or Persian.

Introduction

During automatic indexing, frequently occurring word forms having no real purpose in describing document contents are removed for two main reasons (Manning, Raghavan, & Schütze, 2008). First, each match between a query and a document should be based on pertinent terms rather than retrieving document simply because they contain words such as "the," "your," or "of." This would not constitute an intelligent search strategy since these nonsignificant words in fact represent noise and may actually damage retrieval performance because they do not discriminate between relevant and nonrelevant items. Second, the size of the inverted file would hopefully be reduced (by ∼30–50%).

Although these objectives seem clear, some arbitrariness is required because no clear methodology has been suggested for developing stopword lists (Fox, 1990). The SMART system, for example, suggests 571 English words while Fox (1990) proposed only 421 words (also called *function* or *grammatical* words). These two solutions include mainly determinants, prepositions, conjunctions, pronouns, and certain very frequent verbs forms ("is," "can," etc.). Commercial services, on the other hand, tend to be more conservative and limit the size of their stopword list. For example, the list used by the DIALOG information service (Harter, 1986) includes nine items ("an," "and," "by," "for," "from," "of," "the," "to," "with") while the list contains only a single item ("the") in other cases (Moulinier, 2004).

Removing very frequent word forms may, however, reduce retrieval effectiveness. In English for example, queries might encounter problems with terms such as "language c," "vitamin a," "IT engineer," or "US citizen" where the forms "c," "a," "it," or "us" are usually removed during indexing. These examples thus explain why commercial information-retrieval (IR) systems may index the documents under all available forms, and apply only a very short stopword list when analyzing the request (Moulinier, 2004).

The main objective of this brief communication is to analyze and evaluate various stopword lists using a relatively large number of queries. The rest of the communication is organized as follows: We briefly describe the IR methods applied during our experiments, then depict the main characteristics of the test collections. Next, the performance of various IR models combined with various stopword lists, IR models, and natural languages is evaluated. The main findings are presented in the conclusion.

IR Models

To evaluate the impact of the various stopword lists with respect to different IR models, we first applied the classical *tf idf* model (Manning et al., 2008) wherein the weight attached to each indexing term was the product of its term occurrence frequency (tf_{ij} for indexing term t_j in document d_i) and the logarithm of its inverse document frequency [$idf_j = \log(n/df_j)$]. If a word occurs in all documents, its *idf* weight will be $\log(n/n) = 0$. To measure similarities between documents and requests, we compute the inner product after normalizing (cosine) the indexing weights.

To complement this vector-space model, we implemented certain probabilistic models such as the Okapi (or BM25) approach (Robertson, Walker, & Beaulieu, 2000) together with a modified version named *Okapi**. In the Okapi model, the *idf* formula plays a crucial role. Using the original *idf* formula $idf = \log[(n - df_j + 0.5)/(df_j + 0.5)]$, we have noticed that when the underlying term t_j occurs in more than half of the documents ($df_j > n/2$), the resulting *idf* value would be negative, and the final document score also could be negative. As a means of estimating *idf*, we therefore suggest a new variant defined as $idf = \log\{1 + [(n - df_j + 0.5)/(df_j + 0.5)]\}$.

We also have implemented three models derived from *Divergence from Randomness* (DFR) paradigm (Amati & van Rijsbergen, 2002), which combine the two information measures formulated next:

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

in which $\mathrm{Prob^1_{ij}}$ is the probability of finding by pure chance the tf_{ij} occurrences of the term t_j in a document. On the other hand, $\mathrm{Prob^2_{ij}}$ is the probability of encountering a new occurrence of term t_j in the document, given that tf_{ij} occurrences of this term already had been found. To estimate the first probability, we could use a Poison approximation of a binomial distribution (denoted by P) or the inverse expected document frequency model [$\mathrm{I(n_e)}$]. For the second probability, the estimates could be based on Laplace's law (L) or on the ratio between two binomial distributions (B). In the following experiments, the DFR-PL2, DFR-PB2, and DFR-$\mathrm{I(n_e)}$B2 models were used.

From these different probabilistic IR models, we can expect the presence of very common words to have little impact on document retrieval, and a marginal and nonsignificant effect on their rankings. Is this really the case? And when considering natural languages other than English, can we come to similar conclusions? To answer these questions, we conducted the set of experiments described in the following two sections.

Test Collections

In our evaluations, we used an English corpus built during the CLEF 2001 through CLEF 2006 evaluation campaigns (Peters et al., 2008). This corpus consists of newspaper articles published in the *Los Angeles Times* (1994) and the *Glasgow Herald* (1995). The collection contains a total of 169,477 documents, and each article contains about 250 content-bearing terms, on average.

This collection contains 284 topics, each subdivided into a brief title (denoted as T), a full statement of the information need (called description or D), plus any background information that might help assess the topic (narrative or N). These topics cover various subjects (e.g., "El Niño and the Weather," "Chinese Currency Devaluation," "Eurofighter," "Victories of Alberto Tomba," "Marriage Jackson-Presley," or "Computer Animation"), including both regional ("Films Set in Scotland," "Area of Kaliningrad") and international coverage ("Oil Prices," "Sex in Advertisements"). In our evaluations, we built the queries based on the title (T) and descriptive (D) parts of the topic formulation, corresponding to the official query format in the CLEF evaluation campaigns.

To compare our conclusions with another test collection, we selected a corpus written in French and having a more complex morphology than if in English (Sproat, 1992). This second collection also was built during the CLEF 2001 through CLEF 2006 evaluation campaigns. It is composed of newspaper articles published in *Le Monde* (France, 1994 & 1995) as well as articles extracted from the *ATS* news agency (Switzerland, 1994 & 1995). This collection contains a total of 177,452 documents, and each article contains about 180 content-bearing terms together with 299 topics, on average.

For the Persian language, we used a test collection built during the CLEF 2008 evaluation campaign (Peters et al., 2008), comprising newspaper articles (*Hamshahri*, during 1996–2002). This corpus contains 166,774 documents and 100 topics descriptions, and each article contains about 202 indexing terms (after stopword removal), on average. Finally, the Hindi corpus was developed during the FIRE campaign (see www.isical.ac.in/~fire/) and comprises 45 queries and 95,215 articles, in mean, containing 356 indexing terms per document. From a morphological point of view, we can clearly consider Hindi to be the most complex language used in our evaluation while Persian is simpler, with its morphology complexity being comparable to French.

To automatically index these documents, we applied a light stemmer (removing the plural in all cases, and grammatical cases attached to nouns and adjectives for the Persian and Hindi languages).

For the English language, we used both a long list (SMART system) comprising 571 entries and a short one containing nine words (DIALOG system). As a light stemmer, we used the S-stemmer suggested by Harman (1991), which applies three rules to remove the plural suffix "-s." For French, we used both a long list (464 entries) and a shorter one (20 words: "de," "la," "le," "l," "a," "les," "et," "des," "d," "en," "du," "un," "une," "est," "dans," "il," "pour," "au," "que," "qui"). The light stemmer used for this language was described in Savoy (1999), and both the stemmer and the stopword list are freely available at www.unine.ch/info/clef/ For Persian, we built a stopword list containing 881 terms. Documents written in this language separate various suffixes from the stem by inserting a small space. In our implementation, we replaced this small space with a regular space so that the various suffixes included in the stopword list would be removed, and thus the stopword list for this language is longer. For Hindi, the stopword list contains 165 words, and a light stemmer was suggested.

Experiments

To measure retrieval performance (Buckley & Voorhees 2005), we adopted the mean average precision (MAP) computed by TREC_EVAL, based on a maximum of 1,000 retrieved items. To statistically determine whether a given search

TABLE 1. Mean average precision (MAP) for various stopword lists using a light stemmer (Harman, 1991) (284 TD queries).

	MAP		
Model	SMART	Short	None
Okapi	0.4516	0.4402†	0.3839†
Okapi*	0.4520	0.4589†	0.4595†
DFR-I(n_e)B2	**0.4702**	**0.4743**	**0.4737**
DFR-PL2	0.4468	0.4463	0.3159†
DFR-PB2	0.4390	0.3258†	0.0287†
tf idf	0.2742	0.2535†	0.2293†

†statistically significant performance difference.

TABLE 2. Mean average precision (MAP) for various stopword lists using a light stemmer (Savoy, 1999) (299 TD queries).

	MAP		
Model	Long	Short	None
Okapi	0.4321	0.4286†	0.2457†
Okapi*	0.4332	0.4311	0.4302
DFR-I(n_e)B2	**0.4499**	**0.4490**	**0.4467**
DFR-PL2	0.4247	0.4216	0.3080†
DFR-PB2	0.4167	0.4172	0.0469†
tf idf	0.2867	0.2758†	0.2436†

†statistically significant performance difference.

strategy is statistically better than another, we applied the *t* test whereby the null hypothesis H_0 states that both retrieval schemes result in similar performance levels (computation done with the R system; Crawley, 2007). In the experiments presented in this brief communication, statistically significant differences were detected by applying a two-sided test (significance level $\alpha = 5\%$). This null hypothesis thus would be accepted if two retrieval schemes returned statistically similar means; otherwise, it would be rejected. If the underlying normality assumption is not always respected (e.g., using the Shapiro–Wilk test; Royston, 1982), we found that the *t* test returns the same conclusion as the bootstrap test (Savoy, 1997) that does not impose such assumption. The strong correlation between these two tests was previously found in Abdou and Savoy (2006), and tends to confirm the robustness of the *t* test.

For the English language, the MAP achieved using three stopword lists is depicted in Table 1. Table 2 indicates the MAP obtained with the French test collection while Table 3 shows the MAP using the Persian and Hindi languages. In these tables, the best performance for a given stopword list is always depicted in boldface.

When using the SMART stopword list as baseline (column 2 in Table 1), statistically significant performance differences are denoted by the symbol †. For the Okapi or the last three models, performance usually decreased significantly when considering either a short stopword list (column 3) or when we ignored this list (last column). For the Okapi* or DFR-I(n_e)B2 models, however, performance improved when

TABLE 3. Mean average precision (MAP) for various stopword lists using a light stemmer (TD query).

	MAP			
	Persian (100 queries)		Hindi (45 queries)	
Model	Stoplist	None	Stoplist	None
Okapi	0.4559	0.4110†	0.3036	0.2111†
Okapi*	0.4612	0.4564	0.3087	0.2823†
DFR-I(n_e)B2	0.4476	0.4428	0.3301	**0.2912**†
DFR-PL2	**0.4785**	**0.4576**†	0.3271	0.2151†
DFR-PB2	0.4617	0.1882†	**0.3436**	0.0180†
tf id	0.2744	0.2176†	0.2060	0.1371†

†statistically significant performance difference.

applying either a short stopword list or none at all (Yet, with Okapi*, the differences were significant.)

Table 2 lists the MAP for French using the two stopword lists of different sizes as well as one experiment in which very frequent word forms were not removed. With a long stopword list, retrieval effectiveness usually was higher. Moreover, when comparing a long stopword list with none at all, the differences were usually statistically significant (indicated in Table 2 by a † after the corresponding MAP values) when considering the Okapi, DFR-PL2, DFR-PB2, and the classical *tf idf* IR model. For the Okapi* and DFR-I(n_e)B2 models, retrieval performance was similar across the three search contexts.

Using Persian or Hindi, the retrieval effectiveness was higher with a stopword list; yet, differences were usually statistically significant (indicated in Table 3 by a † after the corresponding MAP values).

The poor performance levels for certain DFR implantations can be explained as follows. In some cases, very frequent words do not follow the expected random distribution. In such cases, the IR system detects a divergence from randomness, and thus the corresponding document score is increased. The determinant "the," for example, would be very frequent in any document, yet it occurs only one or twice in some articles. If the term "the" also occurs in the query, these documents will be ranked close to the top of the list. In our English corpus, we found a document comprising 3,409 terms, but with only one occurrence of "the" (This article was on sport results.) Because the expected number of occurrences of the word "the" in an article of this length would be much greater than the observed number (1 in this case), the corresponding document would be listed as the first retrieved item (for all queries having a "the"). For other function words such as "in" or "of," we found that a similar pattern does exist.

Finally, the results depicted in the tables clearly demonstrate better overall performance of the Okapi* model over the classical Okapi model.

Conclusion

Using the MAP as a retrieval-effectiveness measure, we demonstrated that a stopword list may significantly change performance levels, usually by improving them when compared to a search without stopword removal. For the English language, a short stopword list (9 words) usually results in performance levels similar to a longer one (571 words). From French, a similar conclusion can be reached. With Hindi and Persian, we compared a long stopword list with an indexing strategy that accounts for all word forms. For both languages, the removal of very frequent word forms usually leads to a significant improvement in MAP.

Moreover, implementing the traditional Okapi IR model as well as certain implementations of the DFR paradigm might result in low retrieval-performance levels when accounting for all word forms. Such a low retrieval performance can be explained either by the implementation of the *idf* computation (Okapi) or by observing a very low frequency for word forms usually having a high occurrence frequency in all documents written in a given language (e.g., "the," "of," and "or" in English).

Acknowledgments

This research was supported in part by Swiss NSF Grant 200021-113273.

References

Abdou, S., & Savoy, J. (2006). Statistical and comparative evaluation of various indexing and search models. In Proceedings of the AIRS (pp. 362–373). Lecture Notes in Computer Science, 4182. Berlin, Germany: Springer-Verlag.

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

Buckley, C., & Voorhees, E.M. (2005). Retrieval system evaluation. In E.M. Voorhees & D.K. Harman (Eds.), TREC: Experiment and evaluation in information retrieval (pp. 53–75). Cambridge, MA: MIT Press.

Crawley, M.J. (2007). The R book. Chichester, England: Wiley.

Fox, C. (1990). A stop list for general text. ACM-SIGIR Forum, 24, 19–35.

Harman, D. (1991). How effective is suffixing? Journal of the American Society for Information Science, 42(1), 7–15.

Harter, S.P. (1986). Online information retrieval. Concepts, principles, and techniques. San Diego, CA: Academic Press.

Manning, C.D., Raghavan, P., & Schütze, H. (2008). Introduction to information retrieval. Cambridge, England: Cambridge University Press.

Moulinier, I. (2004). Thomson Legal and Regulatory at NTCIR-4: Monolingual and pivot-language retrieval experiments. In Proceedings of the NTCIR-4 (pp. 158–165), Tokyo.

Peters, C., Jijkoun, V., Mandl, T., Müller, H., Oard, D.W., Peñas, A., & Santos, D. (Eds.). (2008). Advances in multilingual and multimodal information retrieval. Lecture Notes in Computer Science, 5152. Berlin, Germany: Springer-Verlag.

Robertson, S.E., Walker, S., & Beaulieu, M. (2000). Experimentation as a way of life: Okapi at TREC. Information Processing & Management, 36(1), 95–108.

Royston, P. (1982). Algorithm AS 181: The W test for normality. Applied Statistics, 31(2), 176–180.

Savoy, J. (1997). Statistical inference in retrieval effectiveness evaluation. Information Processing & Management, 33(4), 495–512.

Savoy, J. (1999). A stemming procedure and stopword list for general French corpora. Journal of the American Society for Information Science, 50(10), 944–952.

Sproat, R. (1992). Morphology and computation. Cambridge, MA: MIT Press.