The Influence of Basic Tokenization on Biomedical Document Retrieval

Dolf Trieschnigg
Human Media Interaction
University of Twente
Enschede, The Netherlands
trieschn@ewi.utwente.nl

Wessel Kraaij
TNO ICT
Delft, The Netherlands
kraaijw@acm.org

Franciska de Jong
Human Media Interaction
University of Twente
Enschede, The Netherlands
fdejong@ewi.utwente.nl

ABSTRACT

Tokenization is a fundamental preprocessing step in Information Retrieval systems in which text is turned into index terms. This paper quantifies and compares the influence of various simple tokenization techniques on document retrieval effectiveness in two domains: biomedicine and news. As expected, biomedical retrieval is more sensitive to small changes in the tokenization method. The tokenization strategy can make the difference between a mediocre and well performing IR system, especially in the biomedical domain.

Categories and Subject Descriptors

H.3.1 [**Content Analysis and Indexing**]: Indexing methods

General Terms

Algorithms, Measurement, Performance, Experimentation

Keywords

Biomedical document retrieval, tokenization, lexical analysis

1. INTRODUCTION

Tokenization is the process of converting a stream of characters into a stream of words or tokens. In the context of Information Retrieval the process is also known as lexical analysis and its goal is to identify candidate index terms.

For general IR purposes, a simple tokenizer extracting sequences of letters will suffice. Additionally, the words can be conflated to a root form using a stemmer (e.g. 'activation' and 'activate' can be stemmed to 'activ') and non-informative words can be discarded using a list of stopwords (e.g. 'the').

Handling biomedical literature is characterized by its problems related to terminology [5]. Its terminology is inconsistently spelled, abbrevations are frequently used, words and abbreviations can have different meanings (homonymy) and concepts are described in more than one way (synonymy). Tokenization plays an important role in handling inconsistently spelled terminology, especially for the purpose of keyword based searching. Biomedical terms contain digits, capitalized letters within words, greek letters, roman digits, hyphens and other special characters.

Many different tokenization approaches have been proposed and used, but few or limited comparative studies have been carried out on its actual influence on search performance. In this work we compare several simple tokenization approaches in the context of biomedical document retrieval. We focus on: case sensitivity, the use of digits, the treatment of special characters, removal of stopwords, stemming and the expansion of compound terms.

SIGIR'07, July 23–27, 2007, Amsterdam, The Netherlands.

2. RELATED RESEARCH

Most related research comes from the field of biomedical text mining in which named entity recognition is an important problem. Term variation is one of the most frequent causes of gene name recognition failures [2, 4, 5].

Many tokenization approaches have been proposed during The TExt Retrieval Conference Genomics tracks [3]. Zhou et al [9] applied conditional Porter stemming to prevent biomedical terminology from being stemmed incorrectly. Wan et al [7] separate sequences of either letters or digits as index terms. Urbain et al [6] normalize gene/protein terms with variants. Many systems use domain specific stop word lists and expand queries with lexical variants. Unfortunately often no comparisons are made to a baseline, making it difficult to assess the added value of these approaches. Wang et al [8] studied the influence of handling hyphenation and Greek letters for preprocessing queries.

3. EXPERIMENTS

We use two IR models in our experiments: (1) probabilistic Language Modeling (LM) using Jelinek-Mercer smoothing and (2) the out-of-the-box TFIDF model from Lucene [1].

A biomedical and news corpus are used for the experiments. Firstly, the TREC 2006 Genomics collection and topics, which consists of around 160.000 full-text biomedical articles from Highwire Press. The topics are used unaltered in the experiments, that is maintaining phrases such as "What is the role of". Secondly, the TREC ad hoc collection consisting of around 520.000 newswire documents (FBIS, FR, FT and LA) is used as a reference corpus, using topics 351 to 450 of the TREC 6, 7, 8 ad hoc evaluations. We only use the title section of the topic descriptions for our queries. Table 1 gives an overview of the 14 tested to-

Name	Description	Example tokenization "NF-κ B/CD28-responsive"	Genomics LM		Genomics TFIDF		News LM		News TFIDF	
LL	lowercase (lc), letters	nf κ b cd responsive	0.3173		0.2832		0.1713		0.1616	
L	letters	NF κ B CD responsive	0.2967	-6%	0.2789	-2%	0.1633	-5%	0.1550	-4%
LLD	lc, letters or digits	nf κ b cd28 responsive	0.3304	4%	0.3201	† 13%	0.1723	‡ 1%	0.1634	† 1%
LD	letters or digits	NF κ B CD28 responsive	0.2919	-8%	0.2852	1%	0.1641	-4%	0.1560	-3%
SLLD	lc, either letters or digits	nf κ b cd 28 responsive	0.3414	8%	0.2938	4%	0.1724	‡ 1%	0.1635	‡ 1%
TGLL	replace greek letters, LL	nf kappa b cd responsive	0.3191	1%	0.2817	‡ -1%	0.1713	0%	0.1616	0%
LLP	LL, porter stemmer	nf κ b cd respons	0.3306	4%	0.2804	-1%	0.2030	19%	0.1951	21%
LLSW	LL, general stopword list	nf κ b cd responsive	0.3511	† 11%	0.2904	3%	0.1742	‡ 2%	0.1659	‡ 3%
W	non whitespace	NF-κ B/CD28-responsive	0.1701	‡ -46%	0.1850	† -35%	0.1245	‡ -27%	0.1126	‡ -30%
LW	lc non whitespace	nf-κ b/cd28-responsive	0.1795	‡ -43%	0.2135	† -25%	0.1318	‡ -23%	0.1195	‡ -26%
COM	compound tokenizer	nf κ nfκ b cd 28 responsive bcd28responsive	0.3651	15%	0.3305	17%	0.1767	‡ 3%	0.1679	‡ 4%
CUS	custom tokenizer	nf kappa nfkappa b cd 28 respons bcd28respons	0.4019	† 27%	0.3537	† 25%	0.2099	‡ 23%	0.2001	‡ 24%

† ($p < 0.05$), ‡ ($p < 0.005$): significant differences with LL (sign test)

Table 1: Mean average precision of document retrieval for different tokenization approaches

kenization methods. The methods and their performance measured in terms of (document based) mean average precision will be discussed in the next section.

4. DISCUSSION & RESULTS

The baseline system (LL) converts the characters to lowercase and treats sequences of letters as terms. Both LM and TFIDF models perform around the median of TREC Genomics 2006 submissions (31% MAP). Case folding shows to be beneficial: its case sensitive variant (L) performs slightly worse (-6% and -2% for LM and TFIDF respectively).

Adding digits to the index (LLD, LD and SLLD) also improves performance. What the best treatment is of digits depends on the IR model. TFIDF performs better using terms consisting of both letters and terms, where LM gives the best performance when sequences of either letters or digits (SLLD) are used as tokens.

Rewriting Greek letters to their full name (TGLL) did not show large differences, probably because only few topics contained references sensitive to this approach.

Porter stemming (LLSP) shows strong improvements on the news collection (19% and 21% respectively). Despite the inccorect stemming of some gene names, the performance on the Genomics collection increases slightly for LM (4%) and decreases for TFIDF (-1%).

Applying a general stopword (LLSW) list shows a strong improvement on the genomics collection, probably because the topic descriptions contain many stopwords. The results on the news collection show a modest improvement.

As expected, naively tokenizing (lowercased) sequences of non-whitespace characters (W and LW) performs poor.

The compound tokenizer (COM) tokenizes both compounds and its components. The sequence of lowercased characters first is split on whitespace. This may result in strings containing special characters such as punctuation and hyphens (e.g. "cd28-responsive"). The special characters are stripped to form the first token (e.g. cd28responsive). Finally, the string is tokenized using the SLLD tokenizer; this results in additional tokens (e.g. cd 28 responsive). The rationale behind this tokenizer is that one token represents a normalized form of the original compound in the text. The additional tokens will match lexical variations containing the same components. The compound tokenizer shows a strong improvement over the baseline (15% and 17%).

The custom tokenizer (CUS) combines the tested tokenizers: it uses the compound tokenizer, removes stopwords and performs Porter stemming. The resulting indices show a very strong improvement over the baseline (27% for LM and 25% for TFIDF). The same tokenizer also performs best on the news collection.

The results show that a small change in tokenization strategy can improve a mediocre 2006 TREC genomics submission (MAP average: 29%) to the top quarter of the submissions (36%-54%). Normalization and splitting compounds to multiple terms shows to be very beneficial for the tested IR models which assume term independence in both queries and documents. We expect that incorporation of proximities of related terms in the retrieval model will even further improve retrieval performance.

5. ACKNOWLEDGEMENTS

This work was part of the BioRange programme of the Netherlands Bioinformatics Centre (NBIC), which is supported by a BSIK grant through the Netherlands Genomics Initiative (NGI).

6. REFERENCES

[1] Apache Lucene project. http://lucene.apache.org/.

[2] S. Ananiadou, D. B. Kell, and J.-i. Tsujii. Text mining and its potential applications in systems biology. *Trends in Biotechnology*, 24(12):571–579, Dec 2006.

[3] W. Hersh, A. Cohen, P. Roberts, and H. Rekapalli. TREC 2006 Genomics Track Overview. In *The Fifteenth Text REtrieval Conference (TREC 2006) Proceedings*, 2006.

[4] L. Hirschman, A. Yeh, C. Blaschke, and A. Valencia. Overview of BioCreAtIvE: critical assessment of information extraction for biology. *BMC Bioinformatics*, 6 Suppl 1:S1, 2005.

[5] M. Krauthammer and G. Nenadic. Term identification in the biomedical literature. *Journal of Biomedical Informatics*, 37(6):512–526, Dec 2004.

[6] J. Urbain, N. Goharian, and O. Frieder. Iit trec 2006: Genomics track. In *The Fifteenth Text REtrieval Conference (TREC 2006) Proceedings*, 2006.

[7] R. Wan, I. Takigawa, H. Mamitsuka, and V. N. Anh. Combining vector-space and word-based aspect models for passage retrieval. In *The Fifteenth Text REtrieval Conference (TREC 2006) Proceedings*, 2006.

[8] Y. Wang, H. Hung, C. Wu, R. T.-H. Tsai, C. Liu, and W. Hsu. An empirical study of lexical variation methods for biomedical information retrieval. In *NCS-2005*, Taiwan, 2005.

[9] W. Zhou, C. Yu, V. Torvik, and N. Smalheiser. A concept-based framework for passage retrieval at genomics. In *The Fifteenth Text REtrieval Conference (TREC 2006) Proceedings*, 2006.