INITIAL EXPERIMENTS WITH EXTRACTION OF STOPWORDS IN HEBREW

Yaakov HaCohen-Kerner and Shmuel Yishai Blitz
Department of Computer Science, Jerusalem College of Technology
21 Havaad Haleumi St., 91160 Jerusalem, Israel

Keywords: Hebrew, Information Retrieval, Stopwords, Zipf's law.

Abstract: Stopwords are regarded as meaningless in terms of information retrieval. Various stopword lists have been constructed for English and a few other languages. However, to the best of our knowledge, no stopword list has been constructed for Hebrew. In this ongoing work, we present an implementation of three baseline methods that attempt to extract stopwords for a data set containing Israeli daily news. Two of the methods are state-of-the-art methods previously applied to other languages and the third method is proposed by the authors. Comparison of the behavior of these three methods to the behavior of the Zipf's law shows that Zipf's succeeds to describe the distribution of the top occurring words according to these methods.

1 INTRODUCTION

Elimination of stopwords is an important issue in information retrieval (IR). Stopwords (stop-list words, common words, noise words or negative dictionary) are usually the most frequently occurring words. Common stopwords are for example: articles (e.g., 'a', 'an', 'the'), prepositions (e.g., 'in', 'on', 'of', 'with', 'to') and conjunctions (e.g., 'or', 'and', 'but').

A rather small number of stopwords make up a large fraction of the text of most documents. Francis (1982) found that the 10 most frequently occurring words in English typically account for 20 to 30 percent of the tokens in a document.

Stopwords are considered as non-useful as index terms since they carry low information content (Van Rijsbergen, 1979; Salton and McGill, 1983; Raghavan and Wong, 1986). Removal of stopwords reduces the index file and leads to improved performance (quality and time) of various IR tasks (Salton and Buckley, 1988; Yang, 1995).

In contrast to English and a few other languages, no commonly confirmed stopwords have been extracted for the Hebrew language. In order to improve the performance of natural language processing (NLP) tasks in general and IR in particular for the Hebrew language, there is a real need to produce a stopword list for Hebrew.

In this ongoing work, in addition to its uniqueness in handling Hebrew texts, specifically news documents, we implement three baseline methods that identify stopwords. We compare the behavior of these methods to the behaviour of the Zipf's law regarding the tested documents.

2 RELATED RESEARCH

Diverse systems have been developed in order to extract stopwords for a variety of languages, mostly for the English language. Examples of such systems are: English - Van Rijsbergen (1979), Fox (1990; 1992), Frakes and Baeza-Yates (1992), Sinka and Corne (2002; 2003), Lo et al. (2005), and Makrehchi and Kamel (2008); French - Savoy (1999); Greek - Lazarinis (2007); and Chinese - Zou et al. (2006a; 2006b). Nevertheless, there are still many languages (including Hebrew) that do not have standard stopword lists.

Fox (1990) generated a stop list for general text in English based on the Brown corpus (Francis, 1982) that contains 1,014,000 words drawn from a broad range of literature in English. The list includes 421 stopwords. Most of them are the top frequently occurring words. Frakes and Baeza-Yates (1992) created a similar list containing 425 words.

HaCohen-Kerner Y. and Yishai Blitz S..
INITIAL EXPERIMENTS WITH EXTRACTION OF STOPWORDS IN HEBREW .
DOI: 10.5220/0003093104490453
In *Proceedings of the International Conference on Knowledge Discovery and Information Retrieval* (KDIR-2010), pages 449-453
ISBN: 978-989-8425-28-7

Lazarinis (2007) presents the creation process of a stopword list for the Greek language. This process generated 99 stopwords using word lemmatization (i.e., normalization of words to their forms, which are used as the headwords in a dictionary). In addition, Lazarinis evaluated the effect of stopword elimination from web queries. The results show that removal of stopwords from user queries improved the average number of relevant documents within a shorter retrieval process.

Lo et al. (2005) proposed a new method for creating a stopword list for a given corpus. Their approach, called term-based random sampling approach, is based on how informative a given term is. Their approach achieves a comparable performance to four baseline approaches inspired by Zipf's law, using various standard collections. The proposed method has a lower computational overhead. They claim that the experimental results demonstrate that a more effective stopword list could be derived by merging Fox's classical stopword list with the stopword list produced by either the baselines or the proposed approach.

Savoy (1999) presented a stopword list for French corpora that contains 215 words, based based on removal of plural suffixes. Savoy concludes that using both stopwording and stemming significantly improves retrieval effectiveness.

Zou et al. (2006b) proposed an algorithm that constructs a stopword list in Chinese. Comparison between the produced Chinese stopword list and a standard stopword list in English shows that the percentage of stop words intersection is very high. Experiments conducted on Chinese segmentation using their stopwords showed that the accuracy of Chinese segmentation was improved significantly.

There is continuous need for automatic extraction of stopword lists (Sinka and Corne, 2003; Lo et al., 2005) due to language updates and cultural, technological and web influences.

There are two main categories of stopwords: general and domain-specific. Domain-specific stopwords are words, which have no discriminate value within a specific domain. Such stopwords varied from one domain to another. For instance, the term 'government' might be a stopword in the domain of politics, but an important term in the domain of sports.

Automatic extraction of domain-specific stopwords for the English language has been performed by Makrehchi and Kamel (2008).

3 BASELINE METHODS

Three baseline methods were implemented by us to automatically extract a stopword list for a tested data set: Term Frequency Normalized (*TFN*), Term Frequency Double Normalized (*TFDN*) and Inverse Document Frequency Normalized (*IDFN*).

These methods are based on the Zipf's law (1949). According to Zipf’s law, in a corpus of natural language text, the usage frequency of any word is considered to be inversely proportional to its frequency rank. Therefore, the term's rank-frequency distribution can be fitted very closely by formula (1), where *r* is the term's rank in the frequency table, *F(r)* is the term's frequency, $\alpha \approx 1$ and $c \approx 0.1$.

$$F(r) = C/r^{\alpha} \qquad (1)$$

Zipf's law is an empirical law formulated using mathematical statistics. It claims that given a corpus of natural language utterances (for most languages) the frequency of any word is inversely proportional to its rank. Therefore, the most frequent word will occur approximately twice as often as the second most frequent word, which occurs twice as often as the fourth most frequent word, etc. For instance, in the Brown Corpus "the" is the most frequently occurring word, and by itself accounts for nearly 7% of all word occurrences. The second-place word "of" accounts for slightly over 3.5% of words.

TFN is a normalized version of Term frequency (*TF*). TF_k is defined for a certain term *k* as the number of times that *k* appears in all the documents (Doc_i) included in a specific data set, as shown in formula (2).

$$TF_k = \sum TF_k\ (Doc_i) \qquad (2)$$

TFN is a normalized version of *TF* by the total number of tokens in the data set. TFN_k is calculated using formula (3),

$$TFN_k = -log(TF_k/v) \qquad (3)$$

where TF_k is the term frequency of *k* and *v* is the total number of tokens in the data set.

TFDN is a double normalized version of *TF*. It is the term frequency of *k* normalized by V_i the number of tokens in Doc_i summed over all the documents in the data set and then re-normalized by *NDoc* the number of documents in the data set.

$$TFDN_k = \sum (TF_k\ (Doc_i)/V_i)/NDoc \qquad (4)$$

To the best of our knowledge, $TFDN_k$ is a novel method for measuring the frequency of words. In contrast to previous *TF* measures, it takes into account also the relative importance of *k* in each document.

IDFN is a normalized version of *IDF*. IDF_k is computed for a certain term *k* as shown in formula (5), where *NDoc* is the total number of documents in the data set and D_k is the number of documents containing term *k*. *IDF* assumes that infrequently occurring terms have a greater probability of occurring in relevant documents and should be considered as more informative and therefore of more importance in these documents.

$$IDF_k = log(NDoc/D_k) \quad (5)$$

IDFN is perhaps the most common variant of the *IDF* measure. *IDFN* was defined by Robertson and Sparck-Jones (1976). *IDFN* normalizes *IDF* with respect to the number of documents not containing the term *k* and adds a constant of 0.5 to both numerator and denominator to moderate extreme values. $IDFN_k$ is calculated using formula (6), where *NDoc* is the total number of documents in the data set and D_k is the number of documents containing term *k*.

$$IDFN_k = log\ (((NDoc - D_k) + 0.5)/(D_k + 0.5)) \quad (6)$$

4 EXPERIMENTS

The examined dataset includes texts in Hebrew that were published in Arutz 7[1] – Israel National News. All documents belong to the same domain: Free Daily Israel Reports in Hebrew from the year of 2008. The entire dataset includes 13,342 news documents. They include 3,463,871 tokens where 171,814 of them are unique. Each document includes in average 259.6 tokens, while 191.5 occur only once.

Table 1: The accumulative frequencies' rates of top occurring words.

# of top occurring words	Accum. freq. rate	# of top occurring words	Accum. freq. rate
10	9.27%	100	20.11%
20	11.45%	200	25.53%
30	13.12%	300	29.36%
40	14.5%	400	32.35%
50	15.7%	500	34.79%

Table 1 presents the accumulative frequencies' rates of the top 500 occurring words. It shows rather surprising findings: The top 10, 20, 100 and 500 words occur only 9.27%, 11.45%, 20.11% and 34.79% of the tokens of the data set, respectively.

These results are in contrast to findings in other datasets in other languages, e.g., (1) The 10 most frequently occurring words in English typically account for 20 to 30 percent of the tokens in a document (Francis, 1982), (2) The top 20 and 99 words occur 30.26% and 44.16% of the total lemmas, respectively (Lazarinis, 2007), and (3) Only 135 vocabulary items are needed to account for half the Brown Corpus[2].

The explanation might be that the examined documents are not uniformly distributed across the categories.

Table 2 presents an overlapping comparison of top occurring Hebrew and English words (the English list is presented in the Appendix of (Fox, 1990) based on the Brown corpus (Francis, 1982).

Table 2: Overlapping comparison of top occurring Hebrew and English words.

# of top words in both lists	Rate of overlapping of Hebrew and English top words
10	20%
20	30%
30	33%
40	50%
50	50%
100	60%

In contrast to quite high overlapping rate (above 80%) between top 100 English[3] and Chinese stopwords (Zou et al., 2006b), there is about medium overlapping rate (60%) between top occurring Hebrew and English words.

Only two words are included in the top 10 words in the English and Hebrew lists: 'of' and 'he'. Four additional overlapped words are contained in the 20 top word lists: 'not', 'on', 'that', and 'it'. The word 'the' that is placed on the first place in the English list appears only at the 87th place in the Hebrew list.

These findings might be due to: (1) The Hebrew corpus analyzed in this research which is a news corpus is not general as the English corpus (many top occurring Hebrew words are related to Israel and its politics, e.g., Israel, security, IDF, government, parliament) and (2) The complex morphology of Hebrew (Choueka et al., 2000) is one of the sources for major differences between the Hebrew and English stopword results. For instance, many English articles, prepositions, and conjunctions (e.g.,

[1] http://www.inn.co.il

[2] http://en.wikipedia.org/wiki/Zipf's_law

[3] based also on the Brown corpus (Francis, 1982).

'a', 'an', 'the', 'in', 'and') are usually not presented as single words in Hebrew, but rather as prefixes of the words that come immediately after.

As can be seen from figure 1, Zipf's law (with the values of c=0.017 & α = 0.7 in formula (1) succeeds to describe the distribution of the top 100 words according to the three implemented methods. Another important finding is the fact that TFN presents the smoothest curve. Indeed, this fact is quite trivial since the graph deals with the frequencies of top occurring words according to their place and TFN is the only method that actually expresses this relation.

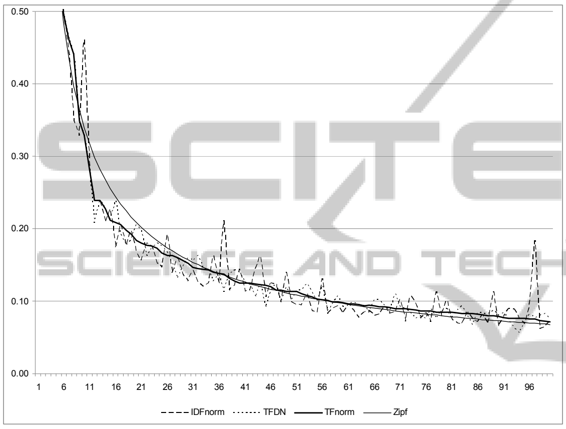

Figure 1: Frequencies' rates according to Zipf's law and the baseline methods for the top 100 words.

5 CONCLUSIONS AND FUTURE WORK

In this ongoing work, we present an implementation of three baseline methods that attempt to extract stopwords in Hebrew for a data set containing Israeli daily news. Two of the methods are state-of-the-art methods previously applied to other languages and the third method is proposed by us.

A comparison of the behavior of these methods to the behaviour of the Zipf's law shows that Zipf's law succeeds to describe the distribution of the top occurring Hebrew words.

Future directions for research are: (1) Applying this research into larger and more diverse corpora to produce a general stopword list and domain-specific stopword lists, (2) Performing additional and extended experiments to evaluate the various stopword lists through retrieval of web queries in Hebrew, (3) Investigating whether other methods can be discovered to achieve more effective stopword lists for IR tasks, and (4) Defining new stopword lists using word lemmatization.

REFERENCES

Choueka, Y., Conley, E. S., Dagan, I. 2000. A comprehensive bilingual word alignment system: application to disparate languages - Hebrew, English. in *Veronis J. (Ed.), Parallel Text Processing*, Kluwer Academic Publishers, 69-96

Fox, C., 1990. A Stop List for General Text. *ACM-SIGIR Forum*, 24, 19–35.

Fox, C., 1992. Lexical analysis and stoplists. In *Information Retrieval - Data Structures & Algorithms*, 102-130, Prentice-Hall.

Francis, W., 1982. *Frequency Analysis of English Usage: Lexicon and Grammar*. Houghton Mifflin.

Frakes, W., Baeza-Yates, R. 1992. *Information retrieval: Data Structures and Algorithms*. Englewood Cliffs, NJ: Prentice Hall.

Lazarinis, F., 2007. Engineering and utilizing a stopword list in Greek Web retrieval. *JASIST,* 58(11): 1645-1652.

Lo, R. T.-W., He, B., Ounis, I., 2005. Automatically Building a Stopword List for an Information Retrieval System. *Journal Of Digital Information Management: Special Issue On The 5th Dutch-belgian Information Retrieval Workshop (dir'05)*, 31(3), 3-8.

Makrehchi, M., Kamel M. S., 2008. Automatic Extraction of Domain-Specific Stopwords from Labeled Documents. In *proc. of ECIR-08*, 222-233.

Raghavan, V. V., Wong S. K. M., 1986. A critical analysis of vector space model for information retrieval, *Journal of the American Society for Information Science*, 37(5), 279-287.

Robertson S. E., Sparck-Jones. K., 1976. Relevance weighting of search terms. *Journal of the American Society for Information Science*, 27(3): 129-146.

Salton, G., McGill, M. J., 1983. *Introduction to Modern Information Retrieval*. New York: McGraw-Hill.

Salton, G., Buckley, C., 1988. Term-Weighting Approaches in Automatic Text Retrieval. *Information Processing and Management*, 24, 513-523.

Savoy, J. A., 1999. Stemming Procedure and Stopword List for General French Corpora. *Journal of the American Society for Information Science*, 50(10), 944-952.

Sinka, M. P., Corne, D. W., 2002. A large benchmark dataset for web document clustering, in *Soft Computing Systems: Design, Management and Applications, Volume 87 of Frontiers in Artificial Intelligence and Applications*, 881-890.

Sinka, M. P., Corne, D. W., 2003. Evolving better stoplists for document clustering and web intelligence. *Design and application of hybrid intelligent systems*, 1015-1023.

Van Rijsbergen. C. J., 1979. *Information Retrieval,* 2nd edition. Dept. of Computer Science, University of Glasgow.

Yang, Y. M., 1995. Noise Reduction in a Statistical Approach to Text Categorization, In *Proc. of SIGIR-95, 18th ACM Int. Conference on Research and Development in Information Retrieval*. 256-263.

Zipf, G. K., 1949. *Human Behaviours and the Principle of Least Effort*, Addison-Wesley, Cambridge, MA.

Zou, F., Wang, P. F. L., Deng, X., Han, S., 2006a. Evaluation of Stop Word Lists in Chinese Language, In *Proc. of the 5th Int. Conference on Language Resources and Evaluation (LREC 2006)*, 2504-2207.

Zou, F., Wang, P. F. L., Deng, X., Han, S., 2006b. Automatic Identification of Chinese Stop Words, *Research on Computing Science*, 18, 151-162.