Experiments with Extraction of Stopwords in Hebrew

Yaakov HaCohen-Kerner

and

Shmuel Yishai Blitz

בית הספר הגבוה לטכנולוגיה בירושלים

Jerusalem College of Technology

Stopwords

- Stopwords (also called 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', 'to') and conjunctions (e.g., 'or', 'and', 'but').
- In information retrieval (IR), an important issue is the task of eliminating stopwords.
- No commonly confirmed stopwords for Hebrew

Baseline Methods for Extraction of Stopwords

The basis: the Zipf's Law

- The frequency of any word is inversely proportional to its rank.
- The most frequent word will occur approximately twice as often as the second most frequent word, etc.

Baseline Methods

(1) Term Frequency Normalized (*TFN*)
(2) Term Frequency Double Normalized (*TFDN*) and
(3) Inverse Document Frequency Normalized (*IDFN*)

Term Frequency Normalized (*TFN*)

TFN is a normalized version of Term frequency (***TF***).

TFk is defined for a certain term *k* as the # of times that *k* appears in all the documents (***Doci***) included in a specific corpus

$$TFk = \sum TFk\ (Doci)$$

TFN is a normalized version of *TF* by the total number of tokens in the data set.

$$TFNk = -log(TFk/v)$$

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

Term Frequency Double Normalized (*TFDN*)

TFDN is a double normalized version of ***TF***. It is the term frequency of ***k*** in ***Doc**$_i$* 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.

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

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.

Inverse Document Frequency Normalized (*IDFN*)

IDFN is a normalized version of ***IDF*.**

IDFk is computed for a certain term ***k*** where ***NDoc*** is the total number of documents in the data set and ***Dk*** is the number of documents containing term ***k*.**

$$\boldsymbol{IDFk = \log(NDoc/Dk)}$$

IDFN is a common variant of the *IDF* measure. *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.

$$\boldsymbol{IDFNk = \log\left(((NDoc - Dk) + 0.5)/(Dk + 0.5)\right)}$$

Data Set

The examined dataset includes Free Daily Israel Reports in Hebrew that were published in 2008 in Arutz 7 – Israel National News (http://www.inn.co.il).

- 13,342 news documents
- 3,463,871 tokens.
- 171,814 of them are unique.
- Each document includes in average 259.6 tokens.

The accumulative frequencies' rates of top occurring words

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

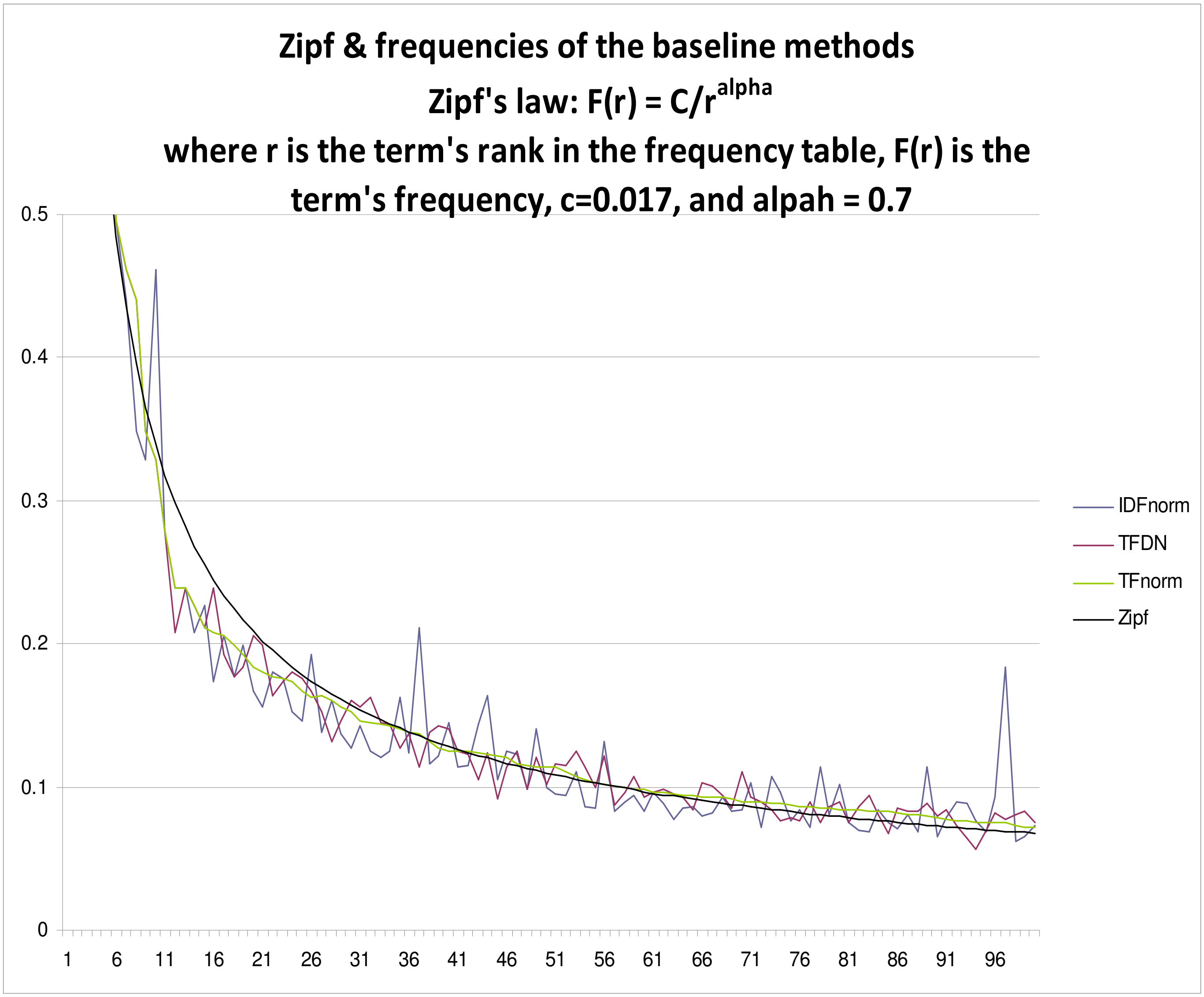
Overlapping comparison of top occurring Hebrew and English words

# 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%

Top 18 occurring words in the 3 baseline methods

#	TFN	TFDN	IDFN
1	של	של	של
2	את	את	את
3	על	על	על
4	כי	כי	כי
5	לא	לא	לא
6	עם	עם	עם
7	ישראל	ישראל	הוא
8	הוא	הוא	גם
9	גם	גם	כל
10	כל	כל	ישראל
11	זה	זה	זה
12	היא	היום	בין
13	בין	בין	היא
14	אמר	אמר	היום
15	הממשלה	הממשלה	אמר
16	היום	היא	כך
17	יש	ראש	יש
18	או	לאחר	לאחר

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



Identification of important topics using top occurring words

	Top occurring content words related to the discussed topics for the last 4 months in 08			
Month	SEP	OCT	NOV	DEC/08
Topic				
Gilad Shalit		Gilad (362)	Gilad (221) Shalit (228)	Gilad (424)
Jewish holidays		Rosh (459) Hashanah (91) Kippurim (307) Sukkot (333) chag ("festival") (409)		
Hamas and Gaza	Hamas (464) Gaza (295)	Gaza (429)	Gaza (181)	Hamas (214) Gaza (59) rockets (461)
Disorderly conduct in Acco	Jews (331)	Acco (341) Jews (136) Arabs (270)	Jews (314)	Jews (266)
Shas		Shas (476) Yosef (478)		

Identification of important topics using top occurring words (2)

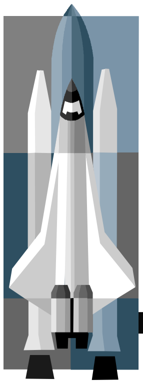
	Top occurring content words related to discussed topics **in the five weeks of AUG 08**				
Month	**Week 1**	**Week 2**	**Week 3**	**Week 4**	**Week 5**
Topic					
Jewish holidays	Rosh (15) Hashanah (39) Sukkot (333) chag (274)	Rosh (27) Hashanah (63) Kippurim (85) Sukkot (464)	Rosh (40) Hashanah (91) Kippurim (170) Sukkot (109) chag (82)	Hashanah (185)	Hashanah (109)
Disor-derly conduct In Acco	Jews (159) Arabs (357)	Acco (126) Jews (136) Arabs (270) Disorderly conduct (489)	Acco (341) Jews (314) Arabs (270) Disorderly conduct (461)	Jews (132) Arabs (384)	Jews (298)
Shas			Yosef (346)	Shas (266) Yosef (419)	Shas (357)

Web-Queries to evaluate the quality of the produced stopword lists

method	average precision results		rate of improvement when omitting unique stopwords from queries
	with unique stopword	without unique stopword	
TFN	15%	15%	0%
TFDN	14%	7%	-7%
IDFN	12%	15%	3%

Conclusions

- We present an application of three baseline methods (one of them us novel) that attempt to extract stopwords for a data set containing Israeli daily news.
- In similar to many other languages, the Zipf's law succeeds to describe the distribution of the top occurring words.
- Other experiments identify important topics as a function of time (months annd weeks) using content words that appear in the top occurring words.



Future Work

(1) Applying this research into larger and/or other data sets,

(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 in order to achieve more effective stopword lists for IR tasks, and

(4) Defining new stopword lists using word lemmatization.