International Journal of Religion
2025
Volume: 6| Number 1 | pp. 121 – 127
ISSN: 2633-352X (Print) | ISSN: 2633-3538 (Online)
ijor.co.uk
DOI: https://doi.org/10.61707/k7rnm813

Arabic Stop Words for Information Retrieval Systems

Tengku Mohd Tengku Sembok[1], Belal Mustafa Abuata[2]

Abstract

This paper explores the classification, analysis, and implications of Arabic stop words in the context of natural language processing (NLP) and information retrieval systems. Arabic, with its complex morphological and syntactic structures, poses unique challenges for automated data processing. Stop words, such as prepositions, adverbs, conjunctions, and interjections, often comprise a significant portion of text but carry minimal semantic value. Their removal is essential for improving the efficiency and accuracy of text analysis tools. The study employs both manual selection and statistical analysis to develop a comprehensive stop word list, focusing on Quranic text as a reference corpus. By integrating linguistic insights and computational techniques, the paper highlights the importance of stop words in stemming algorithms and indexing processes. The findings emphasize the potential for improved retrieval performance and reduced computational overhead through effective stop word handling. This work contributes to advancing Arabic text processing and serves as a foundation for future research in this field.

Keywords: *Stop Words, Information Retrieval, Natural Language Processing, Arabic*

INTRODUCTION

The rapid growth of digital information and the increasing reliance on automated systems for data processing have highlighted the need for effective natural language processing (NLP) tools. In the context of Arabic, a rich and complex language with unique morphological and syntactic structures, one of the fundamental challenges lies in handling stop words, non-content words that carry little semantic value but are abundant in textual data (Abdul-Maseh, 1981; Dahdah, 1985; Hilal, 1990; Saliba & Al-Dannan, 1990).

Stop words, which include prepositions, adverbs, conjunctions, and interjections, account for a significant portion of Arabic text. While these words are essential for grammatical correctness, they often hinder the performance of information retrieval systems by inflating the size of indexing structures and reducing retrieval efficiency (Salton et al., 1975; Salton & McGill, 1993). Identifying and removing these stop words is, therefore, a critical step in the preprocessing phase of text analysis (Al-Kharashi & Evens, 1994).

This paper delves into the classification, analysis, and implications of Arabic stop words, emphasizing their role in information retrieval and stemming processes. By leveraging linguistic expertise and computational techniques, this study aims to refine stop word lists and optimize Arabic text processing. Through a combination of manual selection and statistical analysis, a comprehensive stop word list has been developed and evaluated, with a particular focus on Quranic text as a reference corpus.

Arabic Non-Content Words

Arab grammarians have classified Arabic articles based on their usage and meanings in the Arabic speech (Abdul-Maseh 1981; Dahdah 1985). These are prepositions, adverbs, conjunctions, and interjections. These articles are widely used in Arabic as connectors and complimentary to sentences. Arabic articles can be attached to pronouns, words, and to each other.

Having knowledge in these articles will help in the understanding of the Arabic morphological structure. Consequently, this knowledge is important in the development of Arabic stemming and information retrieval systems. Therefore, in the next sections we will briefly describe each of these four classes of Arabic articles.

[1] INTERNATIONAL ISLAMIC UNIVERSITY MALAYSIA E-mail: TMTS@IIUM.EDU.MY

[2] YARMOUK UNIVERSITY, JORDAN

Prepositions

This type of articles is called by grammarians as the particles of attraction (حروف الجر), the particles which govern the genitive (Saliba & Al-Dannan, 1990). They are divided into two groups: the separable prepositions and inseparable prepositions in accordance to their state of being attached or not to other words. Table-1 shows inseparable prepositions and Table-2 shows those of the separable type.

TABLE-1: Inseparable Prepositions

Preposition	Meaning
و	and, in swearing
ك	same as, like
ل	to, for
ب	in, at, by
ت	by, in swearing

TABLE-2: Separable Prepositions

Preposition	Meaning
إلى	to
حتى	until, up to
حاشى	excluding, except
خلا	except, apart from
ربٍ	many, may
عدا	except, exclude
على	on, over
عن	from, after
في	in, inside
كي	in order
لولا	because, if not
بين	between
وسط	middle
وراء	behind
تحت	under
دون	without
بعد	after
مع	with
مذ	recently, since lately
من	from, on account of
منذ	recently, since lately

Adverbs

Arabic adverbs are divided into three types: adverbs of various origins, indeclinable nouns, and nouns in the accusative. The adverbs of various origins can be partly inseparable or separable. Table-3 shows examples of these types.

TABLE-3: Examples of Arabic Particle Adverbs

Adverb	Meaning	Type
أ	is used for questioning	*various origins (inseparable)*
س	is used to express futurity	*various origins (inseparable)*
ل	surely	*various origins (inseparable)*
أجل	certainly	*various origins (separable)*
إذا	if	*various origins (separable)*
إذن	so	*various origins (separable)*
إن	truly	*various origins (separable)*
فقط	only	*various origins (separable)*
هو	he	*indeclinable nouns*
الذي	whom	*indeclinable nouns*
بينما	while	*indeclinable nouns*
كيف	how	*indeclinable nouns*
هذا	this	*indeclinable nouns*
جدا	very much	*accusative*
جميعا	all	*accusative*
حيثما	wherever	*accusative*
قليلا	a little	*accusative*
كثيرا	a lot	*accusative*
غدا	tomorrow	*accusative*

Conjunctions

The conjunctions particles in Arabic consist of different groups with reference to their implication. Some of these groups are: حروف العطف (*connective particles*), حروف الشرط (*conditional particles*), حروف النسخ (paste particles), حروف الجزم (clipping particles), حروف النصب (accusative/subjunctive particles), etc. These particles can be separable or inseparable. In Table-4 are examples of these particles with their meanings.

TABLE-4: Examples of Arabic Particle Conjunctions

Adverb	Meaning	Type
أم	or	*connective*
أو	or	*connective*
بل	rather	*connective*
ف	then	*connective*
حتى	until	*connective*
إذ	and then	*conditional*
إذا	whether	*conditional*
أما	as for	*conditional*
إن	if	*paste*
أن	that	*paste*
كأن	like	*paste*
لكن	but	*paste*
لعل	perhaps	*paste*

ليت	would	*paste*
إذن	therefore	*accusative/subjunctive*
كي	so that	*accusative/subjunctive*
لما	whereas	*clipping*
لم	not	*clipping*

Interjections

This type of particles is called أصوات (sounds or tones), which are mostly used in situations like being in pain, danger, as a complimentary, and surprise. Some of these particles are shown in Table-5 together with their respective meaning.

TABLE-5: Examples of Arabic Particle Interjections

Adverb	Meaning
حي	*come!*
آ	*O!*
أف	*pish!*
بس	*then*
آه	*Oh!*
هش	*to call a horse*

Development of an Arabic Stop Word List

Words that are frequently occurred in documents and do not give any hint value to the content of their documents are labeled as stop words. Thus, they are eliminated from the set of index terms so that they will not interfere with retrieval process. Salton & McGill (1993) reported that such words comprise around 40-50% of a collection of documents text words. Eliminating the stop words from consideration early in automatic indexing will speed up processing, save a huge amounts of space in indexes, and will enhance retrieval effectiveness (Frakes & Baeza-Yates, 1992).

There are many approaches used to determine stop words list which having the same aim of finding those words of no content values (van Rijsbergen, 1979; Popovic, 1991; Fatimah, 1995). These approaches range from manual selection of such words to statistical methods of identifying occurrences of words with very high or very low frequency.

The approach used to find the stop words list in our experiment is the combination of manual selection method and statistical counting in determining highly frequency words. The sources used as references are as follows:

1. Al-Omari set of 117 stop words (Al-Omari, 1994);
2. Words obtained with reference to Abdul-Maseh (1981);
3. Words obtained with reference to Al-Mawrid Arabic dictionary (Al-Baalbaki, 1996);
4. Words obtained with reference to Al-Dawody (1997);
5. Words obtained with reference to Al-Baki (1981);
6. Words obtained from the chapters in Al-Quran (Ali, A.Y. 1983).

As mentioned above, one of the methods used to find the list of stop words is from the chapters of Al-Quran. Al-Quran is the Muslim's holy book consisting of 114 chapters. After extracting all the words of Al-Quran and compute their frequency, we rank these words in descending order. Table 4.6 shows the list of the top 40 most frequent words. Among them are *prepositions* (و, من, على, أو,عن , عليهم ,إلى), *adverbs* (إذا, ,إن), *conjunctions* (إذ), and words such as (ذلك,الذي). These words are the candidates for stop words.

TABLE-6: A List of the 40 Most Frequent Words in the Al-Quran Collection

Rank	Term	Frequency	Percentage
1	و	10131	11.51%
2	من	3113	3.54%
3	الله	2393	2.72%
4	ما	1727	1.96%
5	لا	1474	1.67%
6	في	1242	1.41%
7	إن	1219	1.38%
8	الذين	975	1.11%
9	على	709	0.81%
10	أن	707	0.80%
11	إلا	666	0.76%
12	قال	501	0.57%
13	الأرض	444	0.50%
14	هو	437	0.50%
15	إلى	432	0.49%
16	لهم	416	0.47%
17	كان	407	0.46%
18	هم	398	0.45%
19	يا	361	0.41%
20	لكم	356	0.40%
21	إذا	355	0.40%
22	ثم	340	0.39%
23	به	329	0.37%
24	أو	327	0.37%
25	قل	315	0.36%
26	قالوا	311	0.35%
27	بما	302	0.34%
28	له	298	0.34%
29	ذلك	293	0.33%
30	الذي	283	0.32%
31	عن	272	0.31%
32	آمنوا	270	0.31%
33	كانوا	264	0.30%
34	يوم	261	0.30%
35	لم	261	0.30%
36	كل	257	0.29%
37	فيها	244	0.28%
38	إذ	237	0.27%
39	ربك	231	0.26%

40	عليهم	215	0.24%
	Total Number of Words in Al-Quran	88052	

Some words in Table-1 such as الله(God's name), ربك (your god), آمنوا (they believed) are content bearing words that are also frequently occurred in Al-Quran. These common words contain in every chapter and many sentences of Al-Quran. Therefore, their inclusion in the list of stop words does comply with the fact that they are not good indicators of the content of the collection. But, on the other hand, the user might want to retrieve those sentences with these words.

In our analysis, we found that there are few words such as إلهنا, كذلكم, مهما and others which has low frequency of occurrences. These words, theoretically, should be included in the stop word list. This complies with the approach taken by Fatimah (1995) in the development of the Malay stop words using the Malay translation of Al-Quran.

As a result of the detailed study, analysis and with references to some Arabic morphological books as mentioned at the beginning of the paper, we came out with a final list of around 753 Arabic stop words (see *appendix*).

Evaluation of the New Arabic Stop Words List

The evaluation of the new Arabic stop word list was performed on Al-Quran chapters. The total number of words in Al-Quran is about 88052 words. The evaluation of this list is considered as an introductory stage towards the aim of developing the new stemming algorithm. Both the new stop words list and the new stemming algorithm will be incorporated together and used to develop the experimental Arabic document retrieval.

In our evaluation, we compared all the words from Al-Quran to the new stop words list. The total number of words that matched the stop words list is 40620. This has resulted in 47432 words remaining in Al-Quran. This means 44% of the words in Al-Quran are stop words and if these words are not being considered as index terms then the compression ratio of 44% will be achieved. The evaluation performed using the stop words list developed by Al-Omari (1994) resulted in 35% compression rate. However, the above compression rate achieved for Arabic words is less than what has been reported by Fatimah (1995) for the Malay translation of the Al-Quran, where a compression of 53% was obtained using her Malay stop words list.

One of the interesting results we found in our analysis of the Arabic stop words is that, the 10 most frequently occurred words in the collection (shown in Table-2) constitutes around 26.90% of the words, whereas the same ten stop words constitutes around 58% of the total stop words.

TABLE-2: The List of The Top10 Most Frequently Occurred Words in the Quranic Collection and their respective percentage

Rank	Word	Occurrence	Percentage (/total # words)	Percentage (/total # stop words)
1	و	10131	11.51%	25%
2	من	3113	3.54%	8%
3	الله	2393	2.72%	6%
4	ما	1727	1.96%	4%
5	لا	1474	1.67%	4%
6	في	1242	1.41%	3%
7	إن	1219	1.38%	3%
8	الذين	975	1.11%	2%
9	على	709	0.81%	2%
10	أن	707	0.80%	2%
	Total	**23690**	**26.90%**	**58%**
total # words			**total # stop words**	
88052			40620	

CONCLUSION

The study and classification of Arabic stop words are critical for the development of effective information retrieval systems and morphological analysis tools. By categorizing and understanding the functions of prepositions, adverbs, conjunctions, and interjections, we can better identify non-content words that do not contribute to the semantic meaning of texts. This understanding facilitates the development of accurate stemming algorithms and improves the overall efficiency of Arabic text processing (Frakes & Baeza-Yates, 1992; Gheith & El-Sadany, 1987).

Our findings emphasize the importance of integrating linguistic insights with computational approaches to create robust stop word lists that enhance indexing and retrieval processes. Future work may focus on refining these lists further through advanced statistical methods and evaluating their applicability across diverse Arabic corpora. Such endeavours will significantly contribute to the advancement of Arabic language processing and its applications in modern information systems.

REFERENCES

Abdul-Maseh, M.,1981. Arabic Language Structure. Beirut: Lebanon Library.

Al-Kharashi, I.A. & Evens, M.W. 1994. Comparing words, Stems, and Roots as Index Terms in an Arabic Information Retrieval System. Journal of the American Society for Information Science. 45(8): 548-560.

Al-Baalbaki, R. 1996. Al-Mawrid: A Modern Arabic-English Dictionary. Beirut, Lebanon: Dar El-Ilm Lilmalayin.

Al-Baki, M.F. 1981. The Indexed Terminology of Al-Quran Al-Kareem Words. Beirut, Lebanon: Dar Al-Fikir Publication.

Ali, A.Y. 1983. The Holy Quran: Text, Translation and Commentary. Maryland: Amana Corp.

Al-Omari, H. 1994. ALMAS: An Arabic Language Morphological Analyzer System. National University of Malaysia. Bangi, Selangor.

Al-Dawody, SA, A. (ed). 1997. Al-Quran Pronounced Terms. Al-Asafahani. Syria: Dar Al-Kalam.

Dahdah, A. 1985. Arabic Language Grammar Dictionary.

Ahmad, F. D. (1995). A Malay language document retrieval system: An experimental approach and analysis. Faculty of Information Science and Technology. UKM.

Frakes, W. B., & Baeza-Yates, R., 1992. Information Retrieval: Data Structures & Algorithms. Englewood Cliffs, NJ: Prentice Hall.

Gheith, M., & El-Sadany, T., 1987. Arabic Morphological Analyzer on a Personal Computer. Arabic Morphology Workshop, Stanford University.

Hilal, Y. 1990. Automatic Processing of Arabic Language and Applications. Proceedings of the Arabic Language Processing Using Computer Conference: 213-219.

Popovic. M. 1991. Implementations of A Slovene Language-Based Free-Text Retrieval System. University of Sheffield. UK.

Saliba, B., & Al-Dannan, A.,1990. Automatic Morphological Analysis of Arabic: A Study of Content Word Analysis. Proceedings of the First Kuwait Computer Conference, 231-243.

Salton, G., & McGill, M. J., 1993. Introduction to Modern Information Retrieval. McGraw-Hill. Beirut: Lebanon. Lebanon Library.

Salton, G. Yang. & Yu. 1975. A Theory of Term Importance in Automatic Text Analysis. Journal of American Society for Information Science. 26:33-44.

van Rijsbergen, C. J.,1979. Information Retrieval. London: Butterworths.