Information Retrieval Systems: Between Morphological Analyzers and Systemming Algorithms

Afaf Abdel Rhman Mohamed[1], Chafika Ouni[1], Sarah Mustafa Eljack[1], and Fayez Alfayez[1]

[1] Department of Computer Science and Information, College of Science at Zulf, Majmaah University, Al-Majmaah 11952, Saudi Arabia

Abstract
The main objective of an Information Retrieval System (IRS) is to obtain suitable information within a reasonable time to satisfy a user need. To achieve this purpose, an IRS should have a good indexing system that is based on natural language processing.In this context, we focus on the available Arabic language processing techniques for an IRS with the goal of contributing to an improvement in the performance. Our contribution consists of integrating morphological analysis into an IRS in order to compare the impact of morphological analysis with that of stemming algorithms.

Key words:
Arabic language, Information retrieval systems, Morphological analyzer, Stemming algorithms.

1. Introduction

Today, usage of the Internet is widespread, and there has been a proliferation of heterogeneous resources (websites, e-books, e-newspapers, e-magazines, and online media) that has led to a high volume of information being generated. In this context, the importance of Information Retrieval (IR) has grown [1] [22]. IR refers to the process by which a user searches for certain information from a large amount of unstructured textual data using a query [4]. There is an imperative need for retrieval systems to match searchable documents with user needs (queries) in order to return relevant documents. An Information Retrieval System (IRS) is a library in which information is stored, processed, organized and retrieved to satisfy user needs [8]. The primary challenge for an IRS is to obtain the necessary information at the right time from a large amount of data.
A classical information system process contains three phases: indexing, query reformulation and matching. Indexing consists of selecting keywords from each document/query to build an index. The query is then reformulated to be suitable for the information retrieval model. Finally, the entered query is matched to the index to select the relevant documents. The selected documents are ranked in descending order and are displayed to the user [3,4,7].

We focus on the indexing phase, as this forms the core of the IRS. The main purpose of this work is to optimise the speed and performance of an IRS in finding relevant documents in response to search queries. Once an effective index has been built, the retrieval process is simplified [4, 5]. The rapid growth of Arabic content on the web offers motivation for the processing of an Arabic corpus. The Arabic language is classified the fourth most widely used language on the web, and in the last eighteen years (2000–2018), the number of Arabic-speaking internet users has grown by 8,616% (first language in user growth). The difficulty of indexing documents depends on the language processed [2] [4] [26].

This paper study the impact of morphological analysis on the performance of an IRS . It is organised as follows. Section 2 describes the morphology of the Arabic language. In Section 3, a literature review is carried out of morphological analysers and stemming algorithms. A new IRS based on morphological analyser techniques proposal is presented in Section 4. The experimental methodology and results are studied and analysed in Section 5. Finally, the conclusion and future works are given in Section 6.

2. Arabic morphology

The Arabic language has two forms:

Classical Arabic. This is the language of the holy Quran. It is vowelised (i.e. it uses Tashkil).

Modern Standard Arabic. This is used in official documents, newspapers and communication.

We focus on the second of these. The Arabic language has a very complex morphology compared to other languages;

Manuscript received March 5, 2022
Manuscript revised March 20, 2022
https://doi.org/10.22937/IJCSNS.2022.22.3.47

it is a highly inflectional language, and the sematics of words depend on various diacritics. On the other hand, unlike many other languages, Arabic does not have a standard rule for adding affixes to words. Another feature of the morphology is that pronouns are attached to a word as prefixes and suffixes [6-8].

The root is the basic unit of Arabic. The following example shown in fig 1 illustrates the complex structure of an Arabic word known as an agglutinated word: سيكتبونه means 'They will write it'

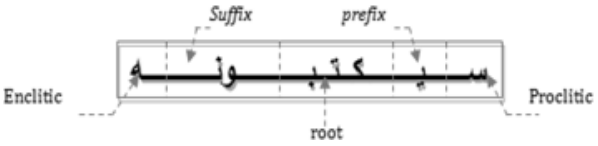

Fig. 1 Arabic morphology.

As shown in this example, an Arabic word can express an entire sentence. The prefixes or suffixes used may also change the semantics of the word; for example, مكتبة means 'library'. Furthermore, the absence of vowels (diacritics) causes an ambiguity problem. Several words of different meaning will have the same root; for example, the word (حسب) can have several meanings, including 'calculate', 'think that', or 'accordingt o' [8]. This ambiguity poses a problem for IR applications, and the richness of this morphology makes it difficult to develop natural language processing applications for Arabic IR[26].

The most important challenge facing developers of an IRS is that of how to extract a suitable root without changing the semantics of the word. The indexing process is based on root extraction (stemming), and there are several algorithms in the literature that can handle the stemming problem.

3. Literature review

Stemming is a linguistic process in which the various morphological variants of the words are mapped to their base forms [8,10]. In Arabic morphology, finding a suitable stem is very difficult [8]. Stemming algorithms can be classified into two basic groups: root extraction stemmers (morphological analysers), and light stemmers [6,7,9,11,12,16].

3.1 Stemming Algorithms

Light stemming algorithms remove prefixes and suffixes from words. The most popular algorithms used for IRSs are the Larkey stemmers (light 1, light 2…….light 10) [7,13]. In this context, the authors of [7,8,15] have developed a new stemmer based on light 10, and have proved that the effectiveness of their algorithms is better than light 10. In [7,12,14,15], the authors built stemming algorithms based on various methods, including statistical methods and big data. To evaluate their approaches, they compared them with a Khoja stemmer (a morphological analyser). The results obtained in all cases confirmed that Arabic stemming was better than Khoja stemming.

Discussion:

Although Arabic stemming methods of this type have achieved good results, they face many problems. Stemmers do not use morphological rules to determine the correct affixes [10], and may obtain roots that are very different from the original word in their meaning. Additionally, light stemming encounters problems such as over-stemming, mis-stemming and under-stemming [8]. These approaches also do not remove the infixes of the words, which decreases their effectiveness [11]. These issues increase the ambiguity, thus reducing the performance of the IRS.

3.2. Root Extraction Stemmers: Morphological Analysers

A root extraction stemmer recognises the composition of a word, and can provide specific morphological information about words [13]. This approach consists of reducing morphological variants to linguistically correct root morphemes, using trilateral Arabic verb patterns [11,15]. The most successful morphological analyser was developed by Khoja [11]. In this section, we are interested in the most widely used morphological analysers in the literature.

3.2.1. ARAMORPH.

Aramorph is a free, open source morphological analyser distributed under a GPL license. The original version was developed in Perl by Backwater, although it is now available in Java Aramorph defines three lexicons: one for prefixes, one for suffixes and a third for lexemes. The analysis proceeds as follows: the Arabic words are first transliterated (for example, كتاب is transliterated to 'ktb'). Aramorph then attributes morph syntactic traits to each word (e.g. gender, number).

3.2.2. MORPH2.

MOPPH2 [13] is a morphological analyser for Arabic texts, and can analyse vocalised Arabic words. It was developed using the Java programming language, based on the version proposed in [16]. The new version developed in the current work can increase the precision and recall from 69.77% to 89. 77% and from 68.51% to 82.51%, respectively.

3.2.3. Qutuf قُطُوْف [27]

This uses finite state automata and rules for agreement developed for cliticalisation parsing. It is based on the AlKhalil Morpho Sys 2 database

3.2.4. AlKhalil (AlKhalilMorpho Sys).

AlKalil is a morphosyntactic analyser for standard Arabic words taken out of context. Version 1.0 was distributed at the ICCA 2010 conference in the form of a free Java source code. AlKhalil analyses partially or totally vowelised words. The authors of [17] have developed an improved version of AlKhalil Morpho Sys 2, and have introduced several improvements, such as patterns and lemma tags to increase text coverage (99%). According to these authors, AlKhalil can index every word in the text by specifying the frequency of its occurrence and its locations in the text.

In our work, we are interested only in the open source

morphological analysers AlKhalil and Aramorph. The following section aims to study these two morphological analysers and to evaluate them using a text file encoded in CP-1256 from the EASC corpus.

	AlKhalil	Aramorph
Agglutinated words	+	+
Vocalised words	+	±
Text cover	+	±
Availability	Open source Java code	Open source Java code
Arabic language type	Classic and modern Arabic	Classical Arabic

3.3. Evaluation of Arabic Morphological Analysers

Table 1: Evaluation metrics

3.3.1. Aramorph

This approach is able to:
- Provide information about the word in question: glossary in English, possible combinations of analysis, voyellations;
- Assign the appropriate morphological categories; and
- Decrease ambiguity with respect to XEROX [20].

However, it has several shortcomings [19]:
- It is not based on rules i.e. the word forms are inserted manually, meaning that the cost of updating is high;
- The components of agglutinated words must be presented in the lexicon;
- The treatment of question morphemes is poorly specified, meaning that the limitations of ambiguity and the coverage decrease; and
- It does not analyse abbreviations; if the word is decomposed, it loses its semantics.

3.3.2. AlKhalil

AlKhalil has two main interfaces: one that allows the text to be entered for analysis and another to display the results of analysis. Users can choose the desired type of output, and the scan result can be saved as an HTML file.
AlKhalil is characterised by a wide coverage of textual words, a high speed of analysis, and accurate indexing of textual words in terms of their location, frequency and identification number. However, also it has some disadvantages, in that it does not analyse the proper names, and has difficulty updating the database.

3.3.3. Comparative Study Between AlKhalil and Aramorph.

To enrich our theoretical study, we analysed a text file encoded in CP-1256 taken from the EASC corpus. The results are presented in Table 1.

Discussion.

The first two metrics (agglutinated and vocalised words) were selected from [18]. We find that both Aramorph and AlKhalil are able to analysevowelised Arabic words, regardless of the type of vowel assigned. Unlike Aramorph, AlKhalil covers all the agglutinated words presented in the file.

However, AlKhalil covers the text by providing all possible analysis solutions, in contrast to Aramorph. Hence, AlKhalil is the most useful morphological analyser. In addition, many authors have used AlKhalil in their research, including the present authors [21,22].
The main goal of developing a stemmer or morphological analyser is to increase and support the search effectiveness of an IRS. Hence, to evaluate the impact of a stemming algorithm or morphological analyser, the best approach is to use it with an IRS [5].

4. An IRS Based on Morphological Analyser Techniques

To carry out a comparison between a stemmer and a morphological analyser, we used an open source IRS called Lucene, based on a light stemmer from the Arabic Computational Linguistics project. The indexing process of Lucene is illustrated in the figure2. It consists of three phases:

- The encapsulation phase, in which parsers transform the file into an object called a document;
- An analysis of the document is carried out using a suitable analyser; and
- Creation of the index using an IndexWriter module, in a location determined by the directory

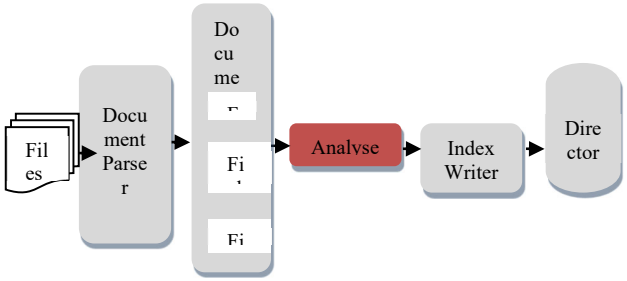

Fig. 2 the indexing process in Lucene.

Our contribution is made at the level of the analysis phase of a document, and consists of developing a new

Java class based on the AlKhalil morphological analyser.

5. Experimentation and validation

5.1. Test collection

Our corpus is taken from the Islamic website www.islamweb.net . We chose the rubric of consecutive notice الفتاوى in order to build the collection test, and considered each question posed as a query. The answers to this questionrepresent the relevant documents. Our corpus contains 586 documents, and an example of one of these documents is shown in Figure 3.

النـدم - وإن كان الركن الأساسي والشرط الأهم في التوبة - فإنه لا يكفي فيها وحده، بل لابد معه من استكمال شروط التوبة كلها حتى تكون توبة صحيحة صادقة، قال الأخضري: وشروط التوبة الندم على ما فات، والنية أن لا يعود إلى ذنب فيما بقي عليه من عمره، وأن يترك المعصية في ساعتها إن كان متلبسا بها.

Means

' Although the main pillar and the most important condition in repentance - it is not enough in them alone, but must complete with the conditions of repentance in order to be all true repentance sincere, Al-Akhdari said: the conditions of repentance is lost, and the intention not to return to the guilt of what remained Of his age, and to leave disobedience at the time if he was in flagrante delicto'.

Fig. 3 the indexing process in Lucene

Table 2: query types

Long query	ما حكم من صدق الكاهن، رغم علمه أنه لا يعلم الغيب إلا الله، ثم قال الشهادتين في صلاته؟ **Means** **' What is the ruling on the sincerity of the priest, despite knowing that he knows only the unseen but Allah, then he said the two testimonies in his prayer? '**	**Doc 88, Doc 12, Doc 24**
Short query	ما حكم من يسرف في صرف المال ؟ **Means** **' What is the ruling of wasting money? '**	**Doc 20, Doc 50, Doc 55**

5.2. Query

Twenty-five queries were collected for our experiments. Table 2 presents the relevant documents for two queries of different lengths.

5.3. Evaluation Methodology

In our experiments, we use the program trec_eval[25]. In particular, we are interested in the following measures:

- The main average precision (MAP) for a comparison of the overall performance of the SRI;
- The precision/recall report for a behaviour evaluation of the SRI;
- The precision calculated at different points (after the first 5, 10, 15 and 100 documents returned) to measure the quality of the results returned by the IRS after the first n documents.

Our assessment is a comparative study between Lucene based on stemming algorithm light 10 and Lucene based on morphological analyser Alkhalil (our proposal).
This study is carried out according to one main parameter: the variation in the query length. We therefore classified the submitted queries into two types, based on their length: short and long queries. The results are presented in the following section.

6. Results

6.1. Precision for n Documents Retrieved.

lmprovement of +22.22% in the precision values P @ 5 and P @ 10 using our proposal, as compared to Lucene.

An increase in the query length results in a decrease in the precision values P @ 5 and P @ 10; however, this variation is lower in our proposal than in Lucene.

6. 2. Main Average Precision.

As our second result, we obtained the MAP.

These results show that we obtain an improvement in the MAP of +11.62% in our proposal compared to Lucene. However, the MAP shows a loss of −18.61% when submitting long queries.

Table3. P@5 and P@10 documents retrieved

P@5			ΔMAP
	Lucene	Our_proposal	
10 short queries	**0.1800**	0.2200	+22.22%
15 long queries	**0.1133**	0.1067	−5.08%
P@10			
10 short queries	**0.0800**	0.11	+37.5%
15 long queries	**0.1133**	0.1067	−5.8%

Table 4. MAP results

MAP			ΔMAP
	Lucene	**Our_proposal**	
10 short queries	**0.5233**	**0.6753**	**+29.04%**
15 long queries	**0.8286**	**0.6729**	**−18.61%**

6.3. Precision/Recall Curve.

The impact of the variation in query length on the first 11 points of the recall is illustrated in Figures 4 and 5 for short and long queries, respectively.

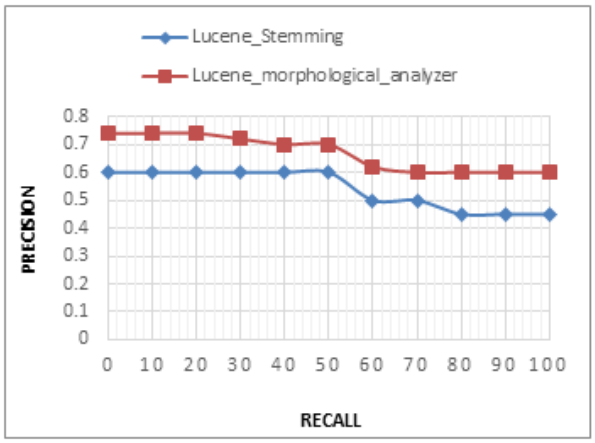

Figure 4. Short queries.

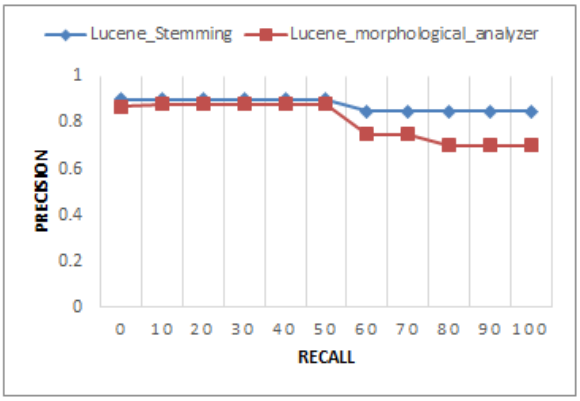

Figure 5. Long queries.

7. Conclusion

In this work, we describe the impact of the integration of a morphological analyzer for the Arabic language in an open source IRS. Our experimental results show that morphological analysis improves the performance of the IRS terms of precision/recall, MAP and precision for n documents returned.

So that, Morphological analyser find the suitable root of the word by matching it with different patterns. This analysis contribute to the creation of a good index of the document (query). Hence, an IRS can select more relevant document for user query.

This improvement is shown only when short queries are used; for long queries, a stemming algorithm is better than a morphological analyzer. This regression in IRS performance may be related to the ambiguity in the solution provided by the morphological analyzer. To address this problem, a new stemming algorithm should be developed based on morphological analyser techniques.

Acknowledgements:

The authors would like to thank the Deanship of Scientifc Research at Majmaah University for supporting this work under Project Number R-2022-18.

References

.[1] Zulaini Y, MuhamadTaufik A, Azreen A, Rabiah, A.Query translation using concepts similarity based on Quran ontology for cross-language information retrieval. Journal of Computer Science. 2013 June,9(7),pp 889-897.

[2] Ali A, Mosa E , Abdullah B. An intelligent use of stemmer and morphology analysis for Arabic information retrieval. Egyptian Informatics Journal.2020 March, 209–217

[3] Essam H, HayelK. Arabic studies' progress in information retrieval. International Journal of Advanced Computer Science and Applications (IJACSA). 2016November,Vol 7, pp 234-238.

[4] Sangita K, Soumen S. New concept-based indexing technique for search engine. Indian Journal of Science and Technology.2017 May, 10(18),pp 1-10.

[5] Maher, A., Mohammed, A.L.: The effectiveness of classification on information retrieval system (case study). (2018). arXiv: 1804.00566 cs.IR.

[6] Ahmed K, Zakir K, Mirza A. Arabic stemmer for search engines information retrieval. InternationalJournal of Advanced Computer Science and Applications (IJACSA). 2016November, 7(1), pp 407-411.

[7] Ahmad A, WafaaA: Arabic stemming techniques: Comparisons and new vision. Proceedings of the 8th IEEE GCC Conference and Exhibition, Muscat Oman, 2015, pp 1–4.

[8] Kheireddine A, SihamHalim. S. A novel robust Arabic light stemmer. Journal of Experimental & Theoretical Artificial Intelligence.2016 July,29(3),pp 1-17

[9] Yaser A, Khawlah M, Mohammad H. Conditional Arabic light stemmer: CondLight. The International Arab Journal of Information Technology.2018 April, 15(3A),pp.559-564.

[10] Jasmeet, S., Vishal, G. Text stemming: Approaches, applications, and challenges. ACM Computing Surveys. 2016 September, 49(3), pp 1-46.

[11] Mohamad A, Riyad A, Ghassan K, AlaaA :Building an effective rule-based light stemmer for Arabic language to improve search effectiveness. The International Arab Journal of Information Technology2012July, 9(4), pp 368-372.

[12] Youness M, Mohammed E, Jamaa, B.Arabic stemmer based big data. Journal of Electronic Commerce in Organizations. 2018January,16(1), pp 17-28.

[13] JNouha K, Lamia B ,Abdelmajid B. The MORPH2 new version: A robust morphological analyzer for Arabic texts.Proceedings of 10th International Conference on Statistical Analysis of Textual Data,Sapienza University of Rome,2010,pp 1034-1044.

[14] Mohammed A, Saif K, BelalA.Novel root based Arabic stemmer. Journal of King Saud University – Computer and Information Sciences. 2015Marh, 27,pp 94–103 .

[15] Mohammed A.Towards improving Khoja rule-based Arabic stemmer. Proceedings of IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT), Amman-Jordan, 2013,pp 1-6.

[16] Osama M E. An improved Arabic light stemmer. Proceedings of the 3rd International Conference on Research and Innovation in Information Systems – ICRIIS'13,Kuala Lumpur-Malaysia,2013,pp33-38.

[17] BelguithHadrich L, Chaâben N. Analyse et désambiguïsation morphologiques des textes arabes non voyellés. In Actes de la 13ème édition de la conférence sur le Traitement Automatique des Langues Naturelles (TALN), Belgique, 2006,pp 493-501.

[18] Mohamed B, Azzeddine M, Mohamed O, Abdelhak L, Abderrahim B: AlKhalilMorpho Sys 2: A robust Arabic morpho-syntactic analyzer. Journal of King SaudUniversity – Computer and Information Sciences. 2016 June, 29, pp 141–146.

[19] Mesfar, S. Analyse morpho-syntaxique automatique et reconnaissance des entités nommées en arabe standard. Doctoral dissertation, Besançon, University of FrancheComté, 2008.

[20] AttiaM.: An ambiguity-controlled morphological analyzer for modern standard Arabic modelling finite state networks. Proceedings of the Challenge of Arabic for NLP/MT Conference, The British Computer Society Conference, London, 2006, pp 4–67 .

[21] Ababou, N., Mazroui, A.: A hybrid Arabic POS tagging for simple andcompoundmorphosyntactic tags. International Journal of Speech Technology. 2016 June, 19(2), pp 289–302.

[22] Chennoufi, A., Mazroui, A.: Impact of morphological analysis and a large training corpus on the performances of Arabic diacritization. International Journal of Speech Technology. 2016 June,19(2), pp 269–280.

[23] Adnen M, Mounir Zrigui: Semantic Similarity Analysis for Corpus development and Paraphrase Detection in Arabic. The International Arab Journal of Information Technology. January 2021, Vol. 18, No. 1.

[24] Hassanin A., kamal J, Sherif A , Mohsen R. Arabic Documents Information Retrieval for Printed, Handwritten, and Calligraphy Image.IEEE access.2021 March, volume 9

[25] TREC, http ://trec.nist.gov/trec_eval, Date accessed : 19/01/2021.

[26] Internet world users by language, Top 10 Languages, https://www.internetworldstats.com/stats7.htm, Date accessed : 12/10/2021

[27] Morphological Analyzer & Part-Of-Speech tagger, http://qutuf.com/, Date accesed: 30 /06/2021

.Authors :

Afaf Abdel Rahman Mohamed An assistant Professor in the Department of computer science and information in the collage of science at The University of Majmaah. Her research focuses on AI and data mining. She is particularly interested in understanding new technologies in the field of machine learning; her work focuses on evolution of the machine learning applications

Chafika Ouni received the master degree in intellegent information system from Kairouan University in 2012 Tunisia. She is a lecturer in Computer Science and Information department at Majmaah University. Her research interests are Information Retrieval System, Language processing, data mining, deep learning.

Sarah Mustafa Eljack An assistant Professor in the Department of computer science and information in the collage of science at The University of Majmaah. Her research focuses on wireless communication CR , WSN , Machine learning . She is particularly interested in understanding new techniques in the field of machine learning; her work is focused on extrication the origin and evolution of the machine learning applications

Fayez AlFayez received the Ph.D. degree in computer science from Manchester Metropolitan University in 2016. He is an Assistant Professor His research interests are in wireless sensor network systems and their applications. His main expertise is in the area of optimizing and designing cross layer communication protocols for large scale distributed systems