Journal of the Egyptian Mathematical Society

Volume (26) - Issue (3) - 2018

ISSN:1110-256X

DOI: 10.21608/joems.2018.2984.1053

AUTOMATIC BUILDING ARABIC DOMAIN MORPHOLOGICAL DICTIONARY USING PART OF SPEECH TAGGING

O.G El Barbary

Mathematics Department, Faculty of Science, Tanta University, Egypt
omniaelbarbary@yahoo.com

Received 24/2/2018 **Revised 5/3/2018** **Accepted 18/4/2018**

Abstract:

Arabic language is still facing some difficulties in automatic processing relating to the richness, morphology, phonetic and lexicon. This paper presents a new strategy for building a morphological field dictionary for Arabic language. Our strategy is divides into two parts. The first, extracts an efficient Field Association FA words for each domain specific. Second, generates the Part Of Speech (POS) tagging for these (FA) word and collect them in one frame. After that, the FA words with its frame collected in alphabetic order. The method of building the automatic morphological field dictionary using a main algorithm is discussed and studied. The advantage of our approach is to build an extended and updated automatic Arabic morphological field dictionary. The average of the accuracy measures (F measures) of the experimental results is up to 76 %.

Keywords: Arabic morphology; Field association terms; Arabic Field Association Terms; Morphological Dictionary; Information Retrieval.

Mathematics Subject Classification (MSC): 68T50; 68T35; 68U15; 68Q42; 68P20.

1. Introduction

Recently, the amount of information of all kinds available electronically has increased rapidly. So, there is a huge need to search and organize enormous amounts of information in text documents.

Text searching is one of the most essential operations in information retrieval systems. With the extensive use of the internet, with its different powerful searching capabilities and applications, this importance has gained a high impetus in the last few years. One of the main problems intrinsic in free-text searching is the variation encountered in word forms due to derivational and inflectional requirements. Hence, a simple matching process becomes irrelevant for efficient information retrieval purposes. This has led us to devise and develop other techniques for improving search performance.

A Field Association (FA) word is a new technique for selecting efficient words that can be related to specified field. The person can recognize a field like mathematics by finding any of these words quantity, structure, space, change, deduction, abstraction, counting, calculation and measurement. Readers generally identify the subject of a text when they notice specific terms, called field association terms [1, 2, 3].

Arabic is the most commonly spoken language after Chinese1. It is probable that with approximately 422 million native speakers, The rich morphology of Arabic and the more complex word formation all contribute to produce Arabic IR researches depending on Arabic morphology. It becomes an integral part of many Arabic information retrieval system. Arabic offers special challenges for data driven. An Arabic word consists of a stem with a consonantal root and pattern. Furthermore, it contains affixes and vowels; also sometimes the same root with different vowels stands for different meanings.

Most pervious work in AIR depends on stem [4]. Stemming is a tool used in IR to combat the vocabulary mismatch problem. This requires deleting the vowels and it is a big mistake because many words become the same although they differ in meaning.

The Arabic language has a special characteristic differs from other languages, most languages construct words out of morphemes which are just concatenated one after another, for example un+ fail + ing. In these languages like

[1] http://encarta.msn.com/encyclopedia 761570647 4/Language.html

English, the stemming technique is very effective. On the other hand, in Arabic language, it is misleading. The type of account of Arabic morphology that is generally accepted by linguists is that proposed by McCarthy [5].

The rest of this paper is organized as follows:

Research objectives of this work are given in Section 2. Section 3 gives more details about Arabic morphology. Previous work is described in section 4. FA words and how to extract efficient FA words from a document with building the FA words determination algorithm is the main purpose of Section 5. The main goal of Section 6 is to determine the Arabic morphology derivatives frame and designate the algorithm that extracts derivatives for PFA and SPFA words. The method of building the automatic morphological dictionary using a main algorithm is discussed in Section 7. The experimental results of this research have been appeared in Section 8. Section 9 explains the research conclusion and future works.

2. Research objectives

Because there are so many text documents available on the Internet and Intranets with a vast amount of potentially valuable knowledge buried within them. And because the number of these documents is usually very large spanning thousands or millions of documents. Hence there is an extreme need for building new automated techniques to efficiently organize, classify, summarize, label, and extract relevant information.

Therefore the objective of this work is to develop a new technique for building Arabic morphological dictionary using Field Association words Derivatives (FAD). To establish this dictionary we defined an algorithm that find all real derivatives included in Arabic derivation namely, active, passive and imperative for each verb. Hence we generated the active participle, the passive participle, the elative, and the noun of the instrument, the adverb and the intensive adjective for each noun. Finally, the Arabic morphological dictionary is constructed automatically.

3. Arabic morphology

All previous studies are based on FA words in English and Japanese, and the extension of FA words to another language such Arabic could be definitely strengthened further researches. Motivated by the need to enhance Arabic searching, we investigate techniques that improve AIR effectiveness. We test supporting FA words with morphological and grammatical rules to Arabic information retrieval, such that building Arabic morphological field dictionary.

Morphology is the field of linguistics which study word structure and formation. It consists of inflectional morphology and derivational morphology. Inflectional morphology is defined as the use of morphological methods to form, inflected word forms from a lexeme. Inflection word forms indicate grammatical relations between words. On the other hand, derivational morphology is concerned with the derivation of new words from other words using derivational affixes. Arabic offers a special challenge for derivational morphology.

An Arabic word may be composed of a stem consisting of a consonantal root and a pattern. Furthermore, it contains affixes and vowels. There are 15 trilateral forms, of which at least 9 are common. Within each conjugation pattern, an entire paradigm is found. Arabic contains two voices (active and passive), two tenses (perfect and passive) and five moods (indicative, subjective, jussive, imperative and energetic).

4. Previous work

There are some necessary and important steps to build an automatic dictionary of any language. Gina-Anne Levow et. al. In [6] defining the key issues in dictionary-based CLIR. They also developed unified frameworks for term selection and term translation. Their developments help researchers to explain the relationships among accessible techniques, and illustrate the effect of those techniques. El-Sayed Atlam and others in [7] presented a strategy for building a morphological matching dictionary of the English language that infers meaning of derivations by considering morphological affixes and their semantic classification. Their strategy depended on grouped derivations into a frame that is accessible to semantic stem and knowledge base. They also proposed in [8] an efficient method for selecting compound Field Association (FA) terms from a large pool of single FA terms for any specialized fields.

In 2013 many researchers have been modified Arabic information retrieval using FA words. El-Monsef, M E Abd and others in [9] investigated three different methods of vector space models using FA words. They are developed K- nearst neighbor classification algorithm, Rocchio document classification algorithm and centroid based algorithm. In [10] O. G. El-Barbary and El-Sayed Atlam have presented a new technique for Arabic document summarization using a fuzzy ontology. This approach depends essentially on fuzzy linguistic variable ontology and FA words. They have predefined the domain ontology with various events Arabic language. The document preprocessing mechanism generated the meaningful terms based on Arabic corpus and Arabic language dictionary defined by the domain expert. They also proved that the meaningful terms have been classified according to an FA term classifier algorithm. Moreover, they addressed some process based on the fuzzy ontology is also developed for Arabic document summarization. Such as every fuzzy concept has a set of membership degrees associated with various events of the domain ontology. In addition, in [11] she developed another method that makes use of FA words to

classify the Arabic news. Another development from her in [12] for using FA words with Arabic morphology and apply them for Arabic document classification.

The construction of an Arabic field dictionary using field association words and its morphological derivatives is an important step in Arabic information retrieval areas. Some research has touched on the extraction of Arabic words of the documents, but is still far from the use of Arabic morphology. El-Sayed Atlam and others in [13] have been presented a new method to extract Arabic FA terms from domain-specific corpora using Part-Of-Speech (POS) pattern rules and corpora comparison. Another method for automatically building new FA words have been developed in [14] by El-Sayed Atlam and others. They developed the www search engine to extract FA word candidates from document corpora. New FA word candidates in each field are automatically compared with previously determined FA words. Then new FA words are appended to an FA word dictionary. As an application to Arabic information retrieval, Meshrif.

5. Field association terms

A single FA word indicates a minimum unit (word) with semantic meaning that identifies a particular field e.g., The words "Protocol", "WAN" are single FA words. These words identify the field network. A compound FA word consists of two or more single FA words. e.g " Application Server" is compound FA words of the field network. There are five groups of FA words based on how well they indicate specific fields [1, 2]. Some FA Words can uniquely identify a certain field, while some FA Words may belong to two or more fields. Thus, each FA Term has a different scope to associate with a field. We can conclude it as the following. Perfect-FA words (PFA) associate with one terminal field. Semi-perfect FA words (SPFA) associate with more than one terminal field in one medium field. Medium-FA words (MeFA) associate with one medium field only. Multiple-FA words (MuFA) associate with more than one terminal field and more than one medium field. Non-Specific FA words (NSFA) do not specify terminal fields or medium fields. Non-Specific FA words include stop words (e.g. articles, prepositions, pronouns).
,From the previous classification of FA words, we can decide that group PFA and SPFA are the most efficient group for identifying fields. For this reason we will build our dictionary on this efficient group.

Algorithm 1: $PFA, SPFA$ **Words Determination Algorithm**

$Set\ PFA = \{\ \}, SPFA = \{\ \}$

$set\ root =< S >, set\ child < s/c >.$

$for\ the\ root < S > and\ any\ child < S/C >, w\ is\ a\ set\ of\ candidates\ FA$

$calculate\ conc(w, < S >) = ((\ Normalization\ (w, < C >))/(\ Normalization\ (w, < S >))),$

$Normalization\ (w, < T >) = ((Frequancy(w, < T >))/(Total - Frequancy(< T >)))$

$if\ (conc\ (w, < S >) \wedge conc(w, < S/c >)) \geq \propto, \propto$ *a threshold to judge FA words ranks*

$Then\ set\ w\ in\ class\ PFA$

$Else$

$if\ (conc\ (w, < S >) \geq \propto \wedge conc(w, < S/c >)) < \propto$

$Then\ set\ w\ in\ class\ SPFA$

End

6. The Arabic morphology derivatives frame.

In order to determine the derivatives frame, some useful derivatives must be located. These derivatives are present, past and imperative verbs for masculine, feminine, dual and plural. Also, you must determine actor, object, source, adjective, and exaggeration formula. Also, the name of preference, the place name, the time name and instrument name must be known.

6.1 Morphological derivatives mechanisms

We apply an inference mechanism to build this frame. The mechanism will generate the FA words at first. Every word has its derivatives in the language associated with various POS tags. In the following we describe the mechanism in detail; this process consists of two steps.

Step 1: This step called input linguistic; in this first step we apply the FA term extraction algorithm. The input vectors are the term set of concepts that retrieved from FA term algorithm these concepts are classified into two categories PFA, SPFA. The nodes in first step transmit input values to next step.

Step 2: This step performs POS tags for all FA terms derived. The Term Part-of- Speech (POS) is a piece of an algorithm that reads text in some language and apportions POS to each word (and other token), such as verb, noun, adjective, etc. This step run by the following algorithm.

Algorithm 2: morphological derivatives extraction algorithm

1- *Let* $PFA, SPFA$ *set of words*

2- *The output is set of PFAD and SPFAD words*

3- $PFAD = \{ w_{1PFAD}, w_{2PFAD}, \dots, w_{nPFAD} \}$. *where,* w_{iPFAD} *is a set of derivatives for the extracted PFAW* $w_i, 1 \le i \le n.$: $w_{iPFAD} = \{ w_{iPFAD_1}, w_{iPFAD_2}, \dots, w_{iPFAD_m} \}$,*and*

4- $SPFAD = \{ w_{1SPFAD}, w_{2SPFAD}, \dots, w_{nSPFAD} \}$. *where,* w_{iSPFAD} *is a set of derivatives for the extracted SPFAW* $w_i, 1 \le i \le n.$: $w_{iSPFAD} = \{ w_{iSPFAD_1}, w_{iSPFAD_2}, \dots, w_{iSPFAD_m} \}$.

5- Let $PFAD = \{ w_{1PFAD}, w_{2PFAD}, \dots, w_{nPFAD} \}$,
$SPFAD = \{ w_{1SPFAD}, w_{2SPFAD}, \dots, w_{nSPFAD} \}$.

6- $\forall\, w_{iPFAD} \in PFAD,\ 1 \le i \le n \ \wedge\ \forall w_{iSPFAD} \in \{ w_{1SPFAD}, w_{2SPFAD}, \dots, w_{nSPFAD} \}$.

7- *If* $L_{w_{iPFAD}} \vee L_{w_{iSPFAD}} = 3$, $L_{w_{iPFAD}}$ *is the length of PFAW word,* $L_{w_{iSPFAD}}$ *is the length of SPFAW word* $\forall W \in (w_{iPFAD} \vee w_{iSPFAD}) = l_1 l_2 l_3$, *where l's is an actual letter or constant.*
do
$PrSM = yl_1\ l_2\ l_3$, $PrSF = Tl_1\ l_2\ l_3$, $PrDM = yl_1\ l_2\ l_3 a$, $PrDF = Tl_1\ l_2\ l_3 a$, $PrPM = yl_1\ l_2\ l_3 wo$, $PrPF = Tl_1\ l_2\ l_3$

End

7. The presented algorithm for building the dictionary

Automatic building of morphological dictionary:

Outline of the presented method:

Figure 1 shows the outline of the presented method for building an Arabic morphological dictionary. To perform perfect morphological Arabic FA term dictionary our methods require the following:

- A set of reference keywords < w> for help to find perfect FA.
- A set of document data collection <D> from a large collection of documents by using the www search engine.
- In the new approach, PFA term is extracted from a large collection of data set. After that, we find all logical derivatives for each PFA term

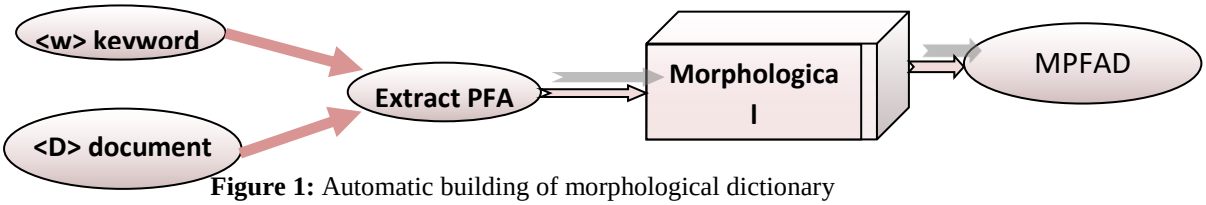

Figure 1: Automatic building of morphological dictionary

Algorithm 3: building FAD dictionary

Input: *(a) A set of documents <D>*
(b) A set of keywords <w> in a given <D>.

Output: *The MPFAD dictionary.*

Method: *1)* $\forall$ *d in <D> determine PFA term using an algorithm(1).*
2) $\forall$ *PFA term, apply the algorithm (2), find set of DPFA.*
3) Append DPFA to MPFAD.
end

8. Experimental evaluation

Arabic morphology, though considerably more difficult than the morphology found in the commonly studied European languages is fully susceptible to derivative analysis techniques. In addition, each word in the Arabic language has its own derivatives according to its length.For evaluating the performance of our approach; we adopt the performance measures, Precision (P) and Recall (R). In our method the performance of our technique evaluated on a variety of documents in different topics likes economy, health, sports and others. Our experiments, trained the method using Arabic documents collected from the internet. It mainly collected from Al-Jazeera Arabic news channel which is the largest Arabic site, Al-Ahram newspaper, Al-Watan newspaper, Al-Akhbar, Al-Arabiya, Al-hayaha and Wikipedia the free encyclopedia.

8.1 Experimental Evaluation

Precision or positive predictive value is the portion of relevant items among the retrieved instances, while recall or sensitivity is the portion of relevant items that have been retrieved over the whole amount of relevant instances

The formulas of precision and recall measures utilized in this paper are as follows

$$Precision = P = \frac{\text{Re}\mathit{levant}\ \mathit{E}\text{fficient Arabic morphological}\ \mathit{Selected\ Automatically\ append\ to\ DIC}}{\mathit{Automatically\ Retreived\ Efficient\ Arabic\ morphological}}$$

$$Recall = R = \frac{Relavant\ Efficient\ Arabic\ morphological\ Selected\ Automatically\ append\ to\ DIC}{Total\ Relavant\ Efficient\ Arabic\ morphological}$$

$$F - measure = \frac{2 \times P \times R}{P + R}$$

Data	P	R	F
التغذية (al taghzya- which means feeding in English)	0.95	0.97	0.96
الرياضة (al ryadah- which means sports in English)	0.89	0.90	0.895
الصحة(al sahah- which means health in English)	**0.54**	**1**	**0.7013**
الطب (al tab- which means medicine in English)	0.91	0.92	0.915
الطفل(al tafel- which means child in English)	0.74	0.80	0.77
الاستنساخ (al estensakh- which means the cloning in English)	0.98	0.99	0.985
الأمراض(al amrad- which means diseases in English)	0.79	0.86	0.83
البيئة (al beaa- which means the environment in English)	0.93	0.95	0.94
التكنولوجيا (al tecnologia- which means technology in English)	**0.47**	**1**	**0.64**

Table 1: The accuracy achieved by the experiment

From Table 1 we can deduce that our method for building a morphological dictionary using FA words is efficient in experiment except two fields. For the field الصحة (al sahah- which means health in English) the reason refers to the existing of the abnormal words. Any natural language contains abnormal words and verbs, in Arabic language the abnormal words has its own morphological derivatives and do not support in our algorithm. In addition, in the field التكنولوجيا (al tecnologia- which means technology in English) there are some words extraneous language like كمبيوتر (computer which means computer in English) this word is not an Arabic word and the Arabic word is الحاسوب (al hasob which means computer in English). These extraneous words don't have any morphological derivatives. So, we can treat this defect by exclusion this words before start algorithm 2.

9. Conclusion and future work

This paper presented an adaptation of existing Arabic morphological analysis techniques to make them suitable for the requirements of AIR applications. In this paper, derivation frames based on POS tagging and knowledge bases of verb lexicons can be related to produce a detailed representation of texts. The Arabic morphological dictionary in our approach that uses FA words and its derivatives is accurate with respect to the results shown. This experimental evaluation is carried out for 9 different fields using 187 MB of domain specific corpora obtained from Al-Jazeera Arabic news channel which is the largest Arabic site, Al-Ahram newspaper, Al-Watan newspaper, Al-Akhbar, Al-Arabiya, Al-hayaha and Wikipedia the free encyclopedia. The results show that the proposed methodology is effective for building an Arabic morphological dictionary using FA terms of accuracy up to 76%. Future studies will further improve the proposed methodology by adding a document classification module so that documents can be classified automatically and FA Term candidates extracted from them. In the future, we hope to reach accuracy up to 99% of this dictionary by using more new techniques.

References

[1] El-Sayed Atlam, Ghada Elmarhomy, Kazuhiro Morita, Masao Fuketa, Jun-ichi Aoe, Automatic building of new Field Association word candidates using search engine, INFORM PROCESS MANAG, **Volume 42(4)**, (July 2006), Pages 951-962.

[2] El-Sayed Atlam, K Morita, M Fuketa, Jun-ichi Aoe A new method for selecting English field association terms of compound words and its knowledge representation, INFORM PROCESS MANAG, **Volume 38(6)**, (November2002), Pages 807-821.

[3] El-Sayed Atlam, M. Fuketa, K. Morita, Jun-ichi Aoe, Documents similarity measurement using field association terms, INFORM PROCESS MANAG, **Volume 39(6)**, (November 2003) Pages 809-824.

[4] Sami Boudelaa, Friedemann Pulvermüller, Olaf Hauk, Yury Shtyrov, William Marslen-Wilson, Arabic Morphology in the Neural Language System, J COG NEUROSC, **Volume 22 (5)**,(2010), Pages 998-1010: 998-1010. .

[5] John McCarthy, A prosodic theory of non-concatenative morphology. LINGUIST INQ, **Volume 12(3)**, (1981), Pages 373-418.

[6]Gina-Anne Levow, W. Oard Douglas, Philip Resnik , Dictionary-based techniques for cross-language information retrieval, INFORM PROCESS MANAG, **Volume 41(3)**,(May 2005) Pages 523-547.

[7] El-Sayed Atlam, K Morita, M Fuketa, Jun-ichi Aoe A new method for selecting English field association terms of compound words and its knowledge representation, INFORM PROCESS MANAG, Volume 38(6), (November2002), Pages 807-821.
[8] M. Fuketa, S. Lee, T. Tsuji, M. Okada and J. Aoe , "A document classification method by using field association words", INFORM SCIENCES , **(Volume 126)**, (2000), Pages 57-70,
[9] El-Monsef, M E Abd; Atlam, El-Sayed; El-Barbary, O. G. , Combining FA words with vector space models for Arabic text categorization , INFORM TOKOYU. **Volume 16(6**) , (Jun 2013), Pages 3517-3528.
[10]O. G. El-Barbary and El-Sayed Atlam, Arabic document summarization using FA fuzzy ontology, INT J INNOV COMPUT I, **Volume 10(4)**, (August 2014), Pages 1351-1367.
[11]O. G. El-Barbary, Arabic News Classification Using Field Association Words, ADV RES, **Volume 6(1)**, (2016), Pages 1-9.
[12]O. G. El-Barbary, Using Arabic skeleton morphology and Maximum Entropy for Arabic Document classification, Article no.BJMCS.23055, BRIT J MATH COMPUT SC, **Volume 14(3)**, (2016), Pages 1-9.
[13] El-Sayed Atlam, Masao Fuketa, Kazuhiro Morita, and Jun-ichi Aoe, Automatic Building an Extensive Arabic FA Terms Dictionary, INT J COMP ELECTRIC AUTO CON INFORM ENG **Volume 4**(8) ,(2010), pages 1290- 1296 .
[14]El-Sayed Atlam, Ghada Elmarhomy, Kazuhiro Morita, Masao Fuketa, Jun-ichi Aoe, Automatic building of new Field Association word candidates using search engine, INFORM PROCESS MANAG, **Volume 42(4)**, (July 2006), Pages 951-962.