Journal of King Saud University – Computer and Information Sciences (2015) 27, 104–112

King Saud University

Journal of King Saud University – Computer and Information Sciences

www.ksu.edu.sa
www.sciencedirect.com

A rule-based stemmer for Arabic Gulf dialect

Belal Abuata *, **Asma Al-Omari** [1]

Faculty of IT & Computer Sciences, Yarmouk University, Irbid 21163, Jordan

Received 22 November 2013; revised 12 March 2014; accepted 15 April 2014
Available online 24 March 2015

KEYWORDS

Arabic dialect stemmer;
Gulf dialect;
Rule base stemming;
Arabic NLP

Abstract Arabic dialects arewidely used from many years ago instead of Modern Standard Arabic language in many fields. The presence of dialects in any language is a big challenge. Dialects add a new set of variational dimensions in some fields like natural language processing, information retrieval and even in Arabic chatting between different Arab nationals. Spoken dialects have no standard morphological, phonological and lexical like Modern Standard Arabic. Hence, the objective of this paper is to describe a procedure or algorithm by which a stem for the Arabian Gulf dialect can be defined. The algorithm is rule based. Special rules are created to remove the suffixes and prefixes of the dialect words. Also, the algorithm applies rules related to the word size and the relation between adjacent letters. The algorithm was tested for a number of words and given a good correct stem ratio. The algorithm is also compared with two Modern Standard Arabic algorithms. The results showed that Modern Standard Arabic stemmers performed poorly with Arabic Gulf dialect and our algorithm performed poorly when applied for Modern Standard Arabic words.

1. Introduction

Nowadays, Arabs prefer to use their local dialect in daily conversation whenever it is not required for them to use the Modern Standard Arabic (MSA). Recently, dialect began to be used in both television and radio (Almeman and Lee, 2013). Arabs also use dialect instead of MSA in important fields such as online communication (chat rooms, SMS, Facebook, Twitters and others). Most of the research on Arabic is focused on MSA (Duwairi et al., 2007; Harrag et al., 2011; AI-Shalabi et al., 2003; Goweder et al., 2008). Currently, there are 12 different Arabic dialects spoken in 28 countries around the world. While most of these dialects are specific to a particular region (for example, "Sudanese Arabic" or "Iraqi Arabic"), the most commonly spoken Arabic dialect is Egyptian Arabic. This variety of Arabic dialects is largely due to the fact that, as Arabic spread and took hold in new regions, it often adopted traces of the language it replaced.

A limited number of Arabic dialect software have been developed and a limited number of research papers published (Al-Gaphari and Al-Yadoumi, 2010). Dialectal varieties have not received much attention due to the lack of dialectal tools and annotated texts. Hence working on dialect is difficult for many reasons (Al-Shareef and Hain, 2011). Firstly, dialect is not considered a written language, it is normally used in

* Corresponding author. Tel.: +962 2 7211111; fax: +962 2 7211128.
E-mail addresses: belalabuata@yu.edu.jo (B. Abuata), zadst@yahoo.com (A. Al-Omari).
[1] Tel.: +962 2 7211111; fax: +962 2 7211128.
Peer review under responsibility of King Saud University.

ELSEVIER Production and hosting by Elsevier

http://dx.doi.org/10.1016/j.jksuci.2014.04.003

spoken form and limited textual data exist in dialect form compared with MSA. Another reason is that dialect still inherits the complex morphological form of MSA. Moreover, additional affixes are introduced informally for each dialect, thereby increasing cross-dialectical differences. Finally no standard convention is agreed on how various dialects should be transcribed.

Modern Standard Arabic (MSA) is a form of Arabic language that is usually used in news media and formal speeches (Diab et al., 2004). There are no native speakers of MSA as stated in Mutahhar and Watson (2002). The importance of processing the dialects comes from here: "Almost no native speakers of Arabic sustain continuous spontaneous production of MSA. Dialects are the primary form of Arabic used in all unscripted spoken genres: "conversations, talk shows, interviews, etc." (Habash and Rambow, 2005). Dialects are becoming widely in use in new written media (newsgroups, weblogs, online chat etc). "Substantial Dialect-MSA differences impede direct application of MSA NLP tools" (Diab and Habash, 2006). Fields of researches such as Arabic NLP, Arabic Translations, Arabic and Cross language Retrieval and other related Arabic research fields suffer from lack of resources due to lack of standards for the dialects, as well as lack of written resources of dialects themselves as shown in Maamouri and Bies (2004). It is right to say that the dialect is very popular where the majority of people use it during their daily life, they use it for conversation and online chatting as well (Alghamdi et al., 2008). Unfortunately, not only is the dialect rarely used in writing, but it also has no written standard. It is realized as a language of heart and feeling where MSA is considered as a language of mind. It is a formal language that has a very good written standard (Al-Gaphari and Al-Yadoumi, 2010).

This paper is organized as follows: In Section 2, we give an overview of related work. In Section 3 we present our method/algorithm used to find the stem of dialect words used in Arabian Gulf Countries. In Section 4 we discuss the evaluation results and their analysis. In Section 5 we talk about conclusion and future work.

2. Background and related works

Arabic language belongs to Semitic group of languages unlike the English language which belongs to the Indo-European language group. Arabic is the fifth widely spoken language in the world used by 5% of people around the world (Kadri and Nie, 2006). It is the official language in 26 countries, located in the Arab world within the west Asia to North Africa. Arabic is the language of Islam, where hundreds of millions of Muslims use it for their religious daily uses.

The Arabic alphabet is used in several languages such as Persian, Malay, and Urdu (Al-Fedaghi and Al-Sadoun, 1990). Arabic Alphabets consist of letters, numbers, punctuation marks, space and special symbols (e.g. mathematical notations). It is different from English in its vowels and diacritic marks. Diacritics are used in the form of over and underscores with Arabic letter. However, most recent written Arabic texts are unvocalized.

The Arabic language is considered a member of a highly sophisticated category of natural languages which has a very rich morphology. Generally speaking, its richness is attributed to the fact that one root can generate several hundreds of words having different meanings. Arabic language orientation is right-to-left which is the opposite in English. There are 30 letters used in the Arabic language. The difficulty in dealing with Arabic language is due not only to its orientation, but also to the language diacritization of scripts, vowels which may or may not be included in, and its complex morphological analysis. All of these and other factors such as its sensitivity to gender, number, case, degree, and tense, make it very difficult to deal with the Arabic language (Abu-Salem et al., 1999).

Arabic is classified into three variants: Classical Arabic, Modern Standard Arabic (MSA) and Colloquial or Dialectal Arabic. The Classical Arabic is the language of the Holly Qur'an and used from the Pre-Islamic Arabia to that of the Abbasid Caliphate. However modern authors never used Classical Arabic. They instead use a literary Language known as MSA. MSA uses the classical vocabulary that is not presented in the spoken varieties. Dialectal Arabic refers to many national varieties which formalize the daily spoken language in the Arab world. It is unwritten and often used in informal spoken media. It differs from MSA on all levels of linguistic representation; the most extreme differences are on phonological and morphological levels.

In the following two subsections, an overview of MSA and Arabic dialect is presented. Also, some of the related works to Arabic dialect are mentioned.

2.1. Modern Standard Arabic

Modern Standard Arabic (MSA) is a form of Arabic language that is used widely in news media and formal speeches (Diab et al., 2004). There are no native speakers of MSA as stated in Mutahhar and Watson (2002). The grammatical system of the Arabic language is based on a root-and-pattern structure and considered as a root-based language with not more than 10,000 roots (Ali, 1988). A Root in Arabic is the bare verb form which can be trilateral, which is the overwhelming majority of Arabic words, and to a lesser extent, quadrilateral, pentaliteral, or hexaliteral, each of which generates increased verb forms and noun forms by the addition of derivational affixes (Saliba and Al-Dannan, 1990).

A stem is a combination of a root and derivational morphemes to which an affix (or more) can be added (Gleason, 1970). However, when applying this definition to Arabic, the verb roots and their verb and noun derivatives are considered as stems. Affixes in Arabic are: prefixes, suffixes (or postfixes) and infixes (morphemes). Prefixes are attached at beginning of the words, where suffixes are attached at the end, and infixes are found in the middle of the words. For example, the Arabic word الطالبات (altalibat) which means "women students", consists of the elements shown in Table 1:

El-Sadany and Hashish (1988), described the formation of Arabic affixes in the Arabic verbs as a derivation of the following rule:

Prefixe1 + Prefix2 + Stem + Suffix1 + Suffixe2 + Suffix3

Table 1 Example of Arabic Affixes.

Word	Prefix	Suffix	Infix	Root
الطالبات	ال	ات	ا	طلب

where the prefixes and suffixes in the above rule are lists of finite length. The properties of the above rule are given as follows:

Prefix1:	The elements of which serve like conjunctions, examples for such an element are ف,ب,, and و (“faa’, baa’ and waw”)
Prefix2:	The attributes associated with its elements help in the partial determination of the tense and the features of the subject pronoun. An example of an element from Prefix2 is (ي ‘yaa’). For this case the tense is present (complete determination of tense), subject pronoun is for the third person (the number and gender of the subject are not determined). In case the Prefix2 is not present (nil) the tense can’t be determined (may be past or imperative)
Suffixe1:	Its elements partially determine the tense and completely determine the features from the subject pronoun (person, gender, number). An example of the Suffixe1 list is (ون “woon”) as in the Arabic word يدرسون (yadrusoon) which means “they study now”
Suffixe2 and Suffixe3:	The attributes associated with their elements which determine the features of the first and second objects pronouns respectively. An example from these lists is (هم “hom”)
Stem:	It is formed by substituting the characters of the root in certain forms called measures or templates (أوزان “awzan”). An example for the trilateral measure is: تفاعل measure ع ن ق root تعانق stem

2.2. Arabic dialect

Arabic dialect is a collective term of the daily spoken language through the Arab world. It is radically different from the literary language (MSA). Arabic dialect is spoken by more than 400 million persons in nearly two dozen countries and holds the dual distinction of being the fifth most widely spoken as well as one of the fastest growing languages in the world (Cote, 2009). Arabic dialects are primarily oral languages; written material is almost invariably in MSA. As a result, there is a serious lack of Language Model (LM) training material for dialectal speech (Alotaibi et al., 2009). Each Country has its main dialect and this main dialect can be divided into a group of sub-dialect e.g. the Saudi dialect includes Najdi (Central) dialect, Hejazi (Western) dialect, Southern dialect (Almeman and Lee, 2013). One factor that differentiates dialects is the influence from the languages previously spoken in the areas. This influence have typically provided a significant number of new words, and also influenced pronunciation and word order.

All the dialects derive ultimately from the same ancient Arabic source, but have undergone changes in grammatical structure, vocabulary and so on. A dialect is different from an accent, which is a version of a standard language with different pronunciations. Arabic dialects can be classified either historically or popularly (Zina Saadi, 2013).

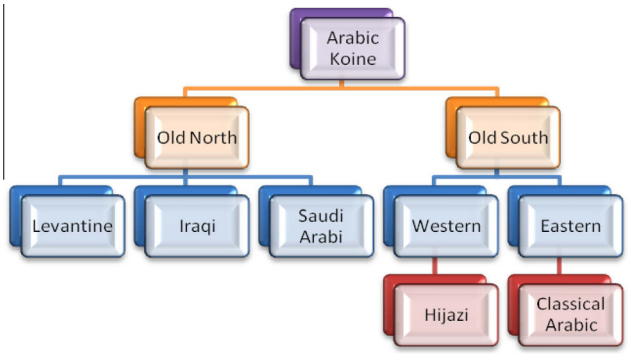

Figure 1 Pre-Islamic or Pre-Arabic Expansion.

For Historical Classification; dialects can be either:

- Pre-Islamic or Pre-Arabic Expansion (6th century BC to 6th century AD) as in Fig. 1.
- Post-Islamic or Post-Arabic Expansion (since 6th century AD) as in Fig. 2.

As for Popular Classification, the Arabic dialects can be divided into various as shown in Fig. 3. Some of these groups are:

Sudanese Arabic – Mostly spoken in Sudan.
Levantine Arabic – This dialect is often heard in Syria, Lebanon, Palestine, and western Jordan.
Gulf Arabic – Mostly heard throughout the Gulf Coast from Kuwait to Oman.
Najdi Arabic – This dialect is most often heard in the desert and oasis areas of central Saudi Arabia.
Yemeni Arabic – This dialect is most common in Yemen.
Iraqi Arabic – The dialect most commonly spoken in Iraq.
Hijazi Arabic – This dialect is spoken in the west area of present-day Saudi Arabia, which is referred to as the Hejaz region.
Egyptian Arabic – This is considered the most widely spoken and understood “second dialect.” It’s mostly heard in Egypt.
Moroccan Arabic – Spoken mostly in Morocco.
Tunisian Arabic - Spoken mostly in Tunisia.
Hassaniiya Arabic – Most often spoken in Mauritania.
Andalusi Arabic – This dialect of the Arabic language is now extinct, but it still holds an important place in literary history.
Maltese Arabic – This form of Arabic dialect is most often found in Malta.

There are few works carried out for stemming of Arabic dialect. Most of these works studied a specific dialect used in one country. Examples of these are the works by Al-Gaphari and Al-Yadoumi (2010) and Alamlahi and Ahmed (2007).

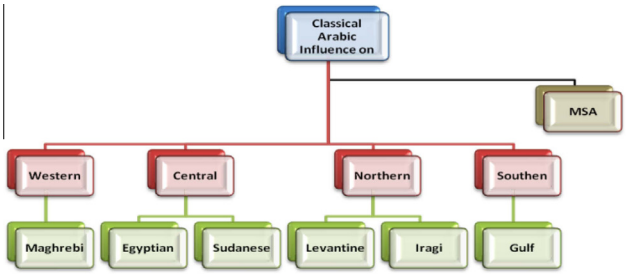

Figure 2 Post-Islamic Post-Arabic Expansion.

Their work is based on morphological rules related to Sana'ani dialect as well as MSA. These rules assist in the conversion of dialect to its corresponding MSA. Their method tokenizes the input dialect text and divides each token into two parts: stem and its affixes. Such affixes can be categorized into two categories: dialect affixes and/or MSA affixes. At the same time, the stem could be dialect stem or MSA stem. Therefore, their method implemented by using a simple MSA stemmer; must pay attention to such situations. Then their dialect stemmer is applied to strip the resulting token and extract dialect affixes (Al-Gaphari and Al-Yadoumi, 2010). Their work uses the dialect stemming as a system that translates the Sana'ani dialect to MSA.

There are a large number of linguistic differences between MSA and the regional dialects. Some of those differences are not found in written form if they are on the level of short vowels, which are deleted in Arabic text anyway.Some of the main differences are (Zaidan and Callison-Burch, 2013):

- MSA's morphology is richer than dialects' along some dimensions such as case and mood. For instance, MSA has a dual form in addition to the singular and plural forms, whereas the dialects mostly lack the dual form. Also, MSA has two plural forms, one masculine and one feminine, whereas many (though not all) dialects often make no such gendered distinction.5 On the other hand, dialects have a more complex cliticization system than MSA, allowing for circumfix negation, and for attached pronouns to act as indirect objects.
- Dialects lack grammatical case, while MSA has a complex case system. In MSA, most cases are expressed with diacritics that are rarely explicitly written, with the accusative case being a notable exception, as it is expressed using a suffix (+A) in addition to a diacritic (e.g. on objects and adverbs).
- There are lexical choice differences in the vocabulary itself. Table 2 gives several examples. Note that these differences go beyond a lack of orthography standardization.
- Differences in verb conjugation, even when the trilateral root is preserved. See the lower part of Table 2 for some conjugations of the root š-r-b (to drink).

Table 2 shows a few examples illustrating similarities and differences across MSA and two Arabic dialects: Levantine and Gulf. Even when a word is spelled the same across two or more varieties, the pronunciation might differ due to differences in short vowels (which are not spelled out). Also, due to the lack of orthography standardization, and variance in pronunciation even within a single dialect, some dialectal words could have more than one spelling (e.g. Levantine "He drinks" could be byšrb). (Table 2 uses the Habash–Soudi–Buckwalter transliteration scheme to represent Arabic orthography, which maps each Arabic letter to a single, distinct character (Habash et al., 2007).

Our paper focuses on Gulf Arabic dialect group which includes Kuwait, Bahrain, Qatar, United Arab of Emirates and some parts of East of Saudi Arabia and some parts of South Iraq.

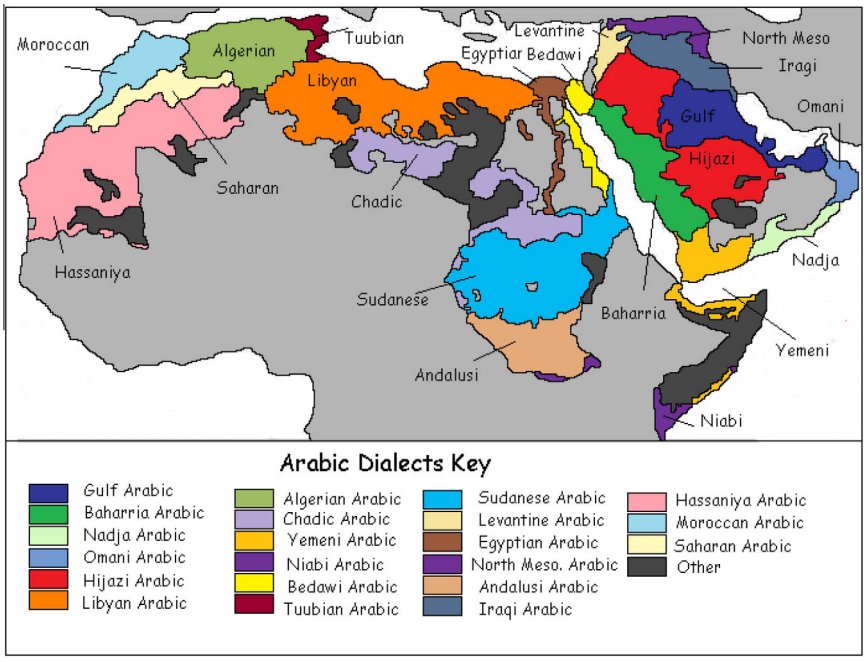

Figure 3 Arabic dialect groups (http://www.importanceoflanguages.com/LearnArabic/tag/arabic-dialects-map).

Table 2 Examples of similarities and differences across MSA and two Arabic dialects.

English	MSA	LEV	GLF
Book	ktAb	ktAb	ktAb
Year	snħ	snħ	snħ
Money	nqwd	mSAry	flws
Come on!	hyA!	ylA!	ylA!
I want	Aryd	bdy	AbŸý
Now	AlĀn	hlq	AlHyn
When?	mtý?	Aymtý?	mtý?
What?	mAäA?	Ayš?	wš?
I drink	Âšrb	bšrb	Ašrb
He drinks	yšrb	bšrb	yšrb
We drink	nšrb	bnšrb	nšrb

Table 3 Examples of non-Arabic dialect words.

Dialect words	Origen	Meaning	Affixes removed	Result
Beshtek بشتك	Fares	Cloak	Kaf ك	Besht بشت
Altegoree التجوري	India	Storage	Alef-Lam الـ	Tegoree تجوري
Dlagaat دلاغات	Turkey	Socks	Alef taa ات	Dlag دلاغ
Abajorat ابجورات	France	Rolling shutters	Alef taa ات	Abajor ابجور
Drawel دريول	England	Driver	–	Drawel دريول

The important thing that must be acknowledged is that Arabic language influenced and gets influenced by other languages. It lent some words to other language like Persian, Turkish, Hindi and Malay. Arabic literary affects Europeans culture especially, in mathematics, science and philosophy. Also it has borrowed words from other languages such as Hebrew, Greek and Persian in early centuries. In modern times it borrowed from English, French and Turkish.

3. Method

The primary goal of this study is to derive an efficient algorithm to extract the stem of dialect words used in Arabian Gulf countries (Kuwait, Bahrain, Qatar, UAE, Saudi Eastern Area, and South of Iraq). These words are collected from the chatting and the Internet forums. The list does not include the old words which are not used these days (e.g. *sahed 'fever'* صاهد, *karfaia "bed"* كرفايه) and includes some new words added to the Gulf dialect (e.g. *majase 'stubbed'* مجسي, *ja'as 'mean'* جعص).

Gulf dialect words include some Non-Arabic words which come from different countries. Some come from India and Iran (Fares) because before the discovery of oil the Arabian Gulf traders used to go to these countries. Other words come from England, France and Turkey because after the discovery of oil European traders from these countries came to Arabian Gulf countries. The proposed algorithm has to remove the affixes added to these words first before comparing it to a Non-Arabic word list since these words must not be analyzed. Table 3 shows example of these words.

Also a Gulf dialect contains many stop words and these words must not be analyzed so the proposed algorithm has to compare the word with a stop word list. Some examples of stop words are *Meno 'Who'* منو, Sheno What' شنو, Shloon How' شلون. Sometimes stop words include affixes; these affixes must be deleted before comparing it to a stop word list, e.g. *Shloonkum 'How are you'* شلونكم. This stop word contains a suffix *kum* كم which has to be deleted first so it became *Shloon 'How'* شلون.

The Proposed algorithm for the Gulf dialect Arabic in order to analyze the words to extract the stem will remove all the Affixes usually appeared in MSA form, e.g. *Alef-Lam* الـ, *Waw-Noon* ون, *Vowels (Alef* ا, *Yaa* ي, *Waw* و*)*. The proposed algorithm supposes that the smallest length of a stem is three letters since more than Three-quarters of the MSA words have a root or stem of size three. The main steps of the proposed algorithm are as follows:

- Read the word
- Check the size of the word ($< = 3$). This control is performed every time when a letter is removed from the word.
- Remove suffixes and prefixes found in suffixes and prefixes set
  - Delete the following prefixes if found (ال ، لل ، وال، بال، ولل).
  - Delete the following prefixes if found (اش، وش).
  - Delete the following suffixes (ات، ون، ك، كم، وكم، تني، ونه، ينه، هـ، ته، هم،ها، ونهم، ين، ت، تي، ني،تك، تكم، الكم،الك).
- Check if the word belongs to Non-Arabic word or to Stop word:
  - If true then stop.
  - If false, delete the following prefix (با ، وا، ا، ت، ي، م، مت، ما ، ان، من، يت).
- If the first letter is (ا، ن، ت، م) and the third letter is (ت) then delete both.
- If the first letter is ا، ت، ي، م then delete it
- Check each letter in the word if it is a vowel then delete it:
  - If we have a vowel with non-vowels neighbored then deletes it.
  - If we have two consecutive vowel letters then we have to delete one of them according to the following order (ا) then (و) and then (ي).
  - If we find three consecutive vowels keep the one in the middle and delete the two neighbored letters.
- After deleting the vowel letters, if we get two letters similar to the neighbored then delete one of them. E.g. رجاجيل delete the (ا ، ي) we get رججل delete the ج we get رجل.

In this pseudo code a file of Non-Arabic word and a file of the stop words have to be created first and before starting the program:

```
Begin
Define WLF a file of the word list to be stemmed, FNAW a file of
Non-Arabic words and another file FSW for stop-words
  WHILE WLF is NOT Empty
    W = Read Single Word ()
    SET wordLength to Length of W
    IF wordLength < = 3 THEN
      Stop and Exit While
    ENDIF
    WHILE wordLength > 3
      IF W contains Prefix (ال ، لل ، وال، بال، ولل) THEN
        Delete Prefix from W
        IF wordLength < = 3 THEN
          Print W as the stem AND Exit While
        ENDIF
      ENDIF
    ENDWHILE
    WHILE wordLength > 3
      IF W contains Prefix (اش، وش) THEN
        Delete Prefix from W
        If wordLength < = 3 THEN
          Print W as the stem AND Exit While
        ENDIF
      ENDIF
    ENDWHILE
    WHILE wordLength > 3
      IF W contains Suffix (ات،ون،ك، كم،وك، تني،ونه،ه، ته،هم،ونهم،ي
ك،ت،تي،ني،الكم،الك) THEN
        Delete Suffix from W
        IF wordLength < = 3 THEN
          Print W as the stem AND Exit While
        ENDIF
      ENDIF
    ENDWHILE
    IF W found in FNAW Then
      Print W as the stem AND Exit While
    ENDIF
    IF W is found in FSW THEN
      Print W as the stem AND Exit While
    ENDIF
    WHILE wordLength > 3 do
    IF W contains Prefix (با، وا، ا، ت، ي، م، مت، ان، من) THEN
        Delete Prefix
        IF wordLength < = 3 THEN
          Print W as the stem AND Exit While
        ENDIF
      ENDIF
    ENDWHILE
    WHILE wordLength > 3
    IF 1st Letter of W is (ا، ن، ت، م) AND 3rd letter of W is (ت)
THEN
          Delete Both Letters from W
        IF wordLength < = 3 THEN
          Print W as the stem AND Exit While
        ENDIF
      ENDIF
    ENDWHILE
//Check each letter in the word if it is a vowel or not
    WHILE not all letters scanned and wordLength >3
      Set L to Letter of W
//Consecutive vowels must be deleted according to the priority (alef
ا, waw , Yaa ي , Yaa ي)
      IF neighbored letters of L not vowels AND L is Vowel
THEN
        Delete L
      ENDIF
      IF one of L neighbored is vowel THEN
        Delete the vowel neighbored with less priority
      ENDIF
      IF two neighbored of L are vowels THEN
        Delete both neighbors with less priority
      ENDIF
//Scan if there are two consecutive letters that are the same then
delete one of them
      IF neighbor of L = L THEN
        Delete neighbor of L OR L
        IF wordLength < =3 THEN
          Print W as the stem AND Exit While
        ENDIF
      ENDIF
    ENDWHILE
  Print W as the stem AND Exit While
End While,
End
```

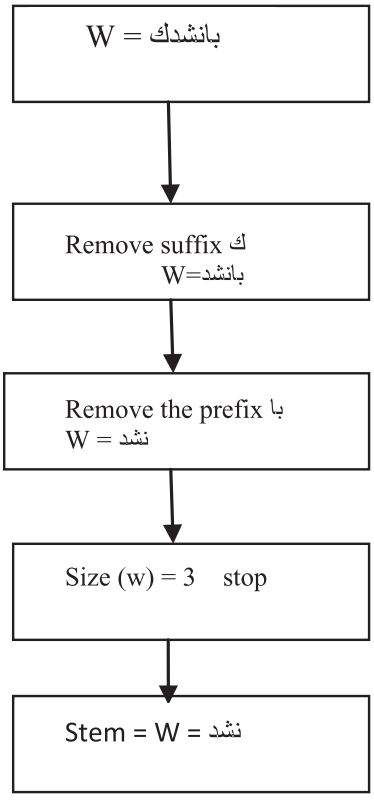

Figure 4 Stemming the word بانشدك (I urge you).

Examples of applying this algorithm on some words are shown in Figs. 4–6.

Table 4 contains some examples that show how the algorithm handles some of the special rules found in some words (these rules are found in the pseudo code).

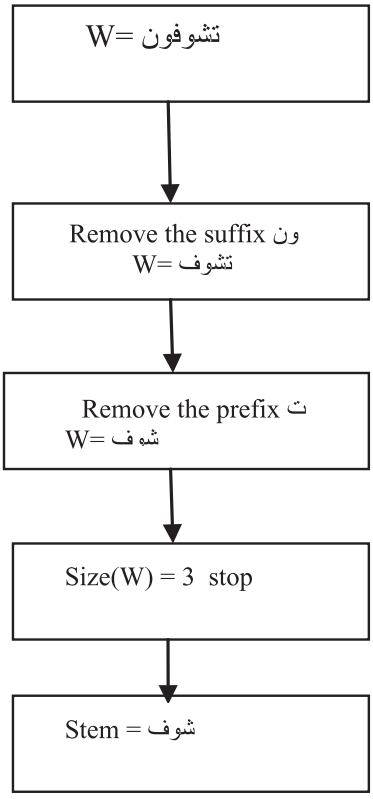

Figure 5 Stemming of تشوفون (You will see).

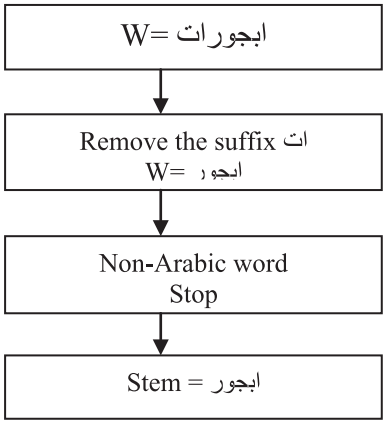

Figure 6 Stemming of ابجورات.

4. Results and analysis

The data test corpus used to test the algorithm is obtained from many places related to Arabic Gulf dialects.[2,3,4,5] We browsed different sites in order to obtain as many as possible Gulf word varieties. The initial data corpus is 15486. The corpus was then analyzed and preprocessed to remove duplicates and MSA words. The resulted corpus consists of 5436 distinct Gulf dialect words. Table 5 and Table 6 show some characteristics of the test corpus.

After applying this algorithm on the Arabic dialect test corpus we get the results reported in Table 7.

Table 4 Examples of some special rules of the algorithm.

No	Rule	Examples
(1)	Delete the prefix (ال، لل،وال،...) The example here also good for the last rule in the pseudo code about taking scan letters and removing vowels	العيال = عيال Then the size is 4 so we take the letter at position 2 (ي) vowel which is followed by other vowel (ا) so keep the (ي) and remove (ا) so the root will be (ع ي ل). This example shows also the rule if we have two vowels in the pseudo code
(2)	If the first letter is (ا، ن، ت، ي، م وليس و) and the third is (ت) we have to delete	ابتلش ، نبتلش، يبتلش، تبتلش، مبتلش = بلش This is the same as MSA e.g.: ارتحل، نرتحل، يرتحل، ترتحل، مرتحل = رحل
(3)	For some prefixes: • If the word starts with (وش) or (اش) we have to delete The example here is also good for the last rule in the pseudo code about taking scan letters and removing vowels	وشعلومك = علومك Then the size is five so we have to take the character at position 3 which is a vowel (و) so delete and shift to the right to get (ل) with the two neighbors (ع ، م) so the root is (ع ل م)
	• If it starts with (ان) we have to remove	انحاش = حاش انخشَ = خشَ انترس = ترش
	• If we have (م), (مت) or (من)	مشخال = شخال Then remove the (ا) which is a vowel to get the root (ش خ ل) and the same for (مرزام = رزم) متحدر = حدر ، متبرز = برز منحاش = حاش
	• If we have (ي) or (يت)	يرطن = رطن ، يجندس = جندس يتغلي = غلي ، يتعلث = علث This can be the same as in MSA. e.g.: يكتب = كتب ، يرسم = رسم يتفكر = فكر ، يتأمل = أمل
(4)	For some suffix: • If it ends with (ك) or (كم)	يغربلك = غربل وهقكم = وهق This also the same as in MSA e.g.: كتابك = كتب ، كتابكم = كتبكتابك = كتب ، كتابكم = كتب
	• If we have (ون), (ونه) or (ونهم)	يخشَونه = خشَ يخشَونهم = خشَ يخشَون = خشَ And this is the same as in MSA e.g.: يظهرون = ظهر يظهرونه = ظهر يظهرونهم = ظهر
	• If we have (ات)	ماصخات = مصخ قفشات = قفش And this is the same as in MSA e.g.: جلسات = جلس ومضات = ومض

[2] http://www.alamuae.com/uaedic/index.html.
[3] http://ar.mo3jam.com/.
[4] http://www.7bna.com/vb/showthread.php?t=93153.
[5] http://www.majma.org.jo/majma/index.php/2009-02-10-09-36-00/648-mag80-5.html.

Table 5 Arabic dialect corpus word frequencies based on word length.

Word length	Word frequency	Word ratio
7	85	1.56%
6	305	5.61%
5	913	16.80%
4	1872	34.44%
3	2213	40.71
2	48	0.88
Totals	5436	100%

Table 6 Examples of Arabic dialect corpus set of different word lengths.

Word	Arabic meaning	English meaning	Derivations	Stem
زحلاقية	لعبة التزحلق	Sliding game	تزحلق, يتزحلقون,	زحلق
ديوانية	مكان لتجمع الرجال	'Men hall	الديوانيه, الدواوين,	دون
تكمكمت	تغطت	Covered herself	يتكمكمون - تكمكموا	كمكم
بوشلاخ	كاذب	Lie	يشلخون, تشليخ	شلخ
اشلون	كيف	How	اشلونكم - اشلونك -	شلون
ابريج	اناء الماء - إبريق	Water jug	اباريج -- ابريجكم	بريج
انخش	اختبأ	Hide	اتخشون - خشيت	خش
اركد	اهدأ	Calm	ركدوا - تركدون	ركد
يبي	يريد	He want	تبون. تبين.	ابي
يبوق	سرق	He steal	- تبوقين - مبيوق	باق
حصه	الؤلؤ الابيض	White pearl	حصايص, يحصص	حص
صج	صدق	Truly	صجك - الصج	صج

The reasons which cause the wrong stems are:

- Some Arabian Gulf country convert *Jeem* ج to Yaa ي, e.g. *Mayhood 'Effort'* ميهود instead of Majhood مجهود. In this case the algorithm has to remove *Yaa* ي because it is a vowel
- Try to stem nouns – name of things – e.g. *Kamee 'Dry milk'* كامي, Krothat 'Nuts' كروظات .
- Some plurals for Non-Arabic words are not in a standard form, i.e. ending with *Alef Taa*"ات "، Waw noon "ون " or Yaa noon "ين ". E.g. The plural of the Non-Arabic word -Persian- *Ebreej 'Jug'* ابريج is *Abareej* اباريج. The problem here is that the algorithm searches for suffix and prefix in the word only before it checks whether the word is Non-Arabic or not.

Since there is no stemmer for the Gulf dialect words we used two Arabic stemmers to compare our stemmer with them. Here we want to prove that MSA stemmers are not applicable to Arabic dialect and their performance is low. These stemmers

Table 7 Results of the three stemmers for Arabic gulf dialect corpus.

Stemmer name	Total number of words	Accuracy (%)	Right root	Not stem	Wrong root
Khoja's stemmer	5436	39	2121	1684	1631
Darwish's stemmer	5436	28	1524	2132	1780
New stemmer	5436	88	4784	0	652

Table 8 Results of the three stemmers for MSA corpus.

Stemmer name	Total number of words	Accuracy (%)	Right root	Not stem	Wrong root
Khoja's stemmer	5436	92	5016	65	355
Darwish's stemmer	5436	76	4140	314	982
New stemmer	5436	52	2825	982	1629

are: Khoja's stemmer (Khoja and Garside, 1999), and Darwish's stemmer (Kareem Darwish, 2002). We selected these two stemmers as they are considered among the Arabic stemmers in terms of accuracy levels when applied on MSA words. Khoja's stemmer is a heavy stemmer whereas Darwish's stemmer is a light stemmer. The test corpus is the same as the one used for testing the new algorithm which consists of 5436 words. Table 7 shows the results of applying the three stemmers on the Dialect list:

Another test is carried out for the same three stemmers where the corpus is changed to MSA list of words collected form various MSA Arabic text. Table 8 shows the results of this test.

Depending on the previous results we found that:

- MSA stemming is different from Dialect stemming. This is shown from the results obtained by the MSA stemmers (Khoja's and Darwish's stemmers) and the new stemmer on Gulf dialect corpus and the MSA corpus as shown in Tables 7 and 8. Hence, dialect Arabic requires special stemmers.
- The proposed stemmer has to stem all words even the Non-Arabic words because sometimes these words have some affixes e.g. the word *Abajorat 'Rolling shutters'* ابجورات has a suffix *Alef and Taa* ات which has to be removed first before comparing the word with the Non-Arabic word list. These types of words are not stemmed both of Khoja and Darwish stemmers.
- Wrong roots came from stemming Non-Arabic words e.g. *Estekanah 'cup of tea'* استكانه and *Asansoor 'Elevator'* اسنسور. Also it try to stem a stop words since they are not the same as in MSA language e.g. *Meno 'Who'* منو, Sheno 'What' شنو
- The right root came from that for dialect words they have the same affixes used in MSA language which Khoja stemmer can recognize and delete e.g. *Rekdo 'be quite'* ركدوا has a suffix *Waw and Alef* وا which has to be removed, other example is *Alathwal 'Stubbed'* الاثول which has a prefix *Alef Lam Alef* الا .

5. Conclusion and future work

MSA stemming algorithms are not applicable to Arabic Dialect. In this paper we showed that MSA stemmers perform poorly when applied to Arabic Gulf dialect and the Arabic Gulf dialect cannot be applied for MSA words. There is no stemming algorithm that handles Arabic Gulf dialect words. Only few algorithms are available that handle only a singular Arabic dialect. In this paper we presented a new rule based Dialect stemmer built for the Gulf dialects. The algorithm is built of a set of predefined rules for Gulf dialects. This new

stemmer accuracy is acceptable and it gave superior results compared to other stemming algorithms. The algorithm can handle many dialects.

This new algorithm can handle all known Arabic dialects by defining new rules and integrating these rules with the current used rules. Improvement to our stemmer can also be added to handle all the non-Arabic words used in Arabic dialects. More tests are also required for the use of the algorithm in Arabic translation and Arabic Sentiment analysis.

References

Abu-Salem, H., Al-Omari, M., Evens, M.W., 1999. Stemming methodologies over individual query words for an Arabic information retrieval system. J. Am. Soc. Inf. Sci. 50 (6), 524–529.

Yahya Alamlahi, Fateh Ahmed, 2007. Sana'ani dialect to modern standard Arabic: rule-based direct machine translation, Computer Science Dep., Sana'a University, Sana'a, Yemen.

Al-Fedaghi, S.S., Al-Sadoun, H.B., 1990. Morphological compression of Arabic text. Inf. Process. Manage. 26 (2), 303–316.

Al-Gaphari, G.H., Al-Yadoumi, M., 2010. A method to convert Sana'ani accent to modern standard Arabic, 2010. Int. J. Inf. Sci. Manage. 8 (1), 39–49.

Alghamdi, M., Alhargan, F., Alkanhal, M., Alkhairy, A., Eldesouki, M., Alenazi, A., 2008. Saudi accented Arabic voice bank. J. King Saud Univ. Comp. Inf. Sci. 20 (1), 45–62 (Riyadh).

Ali, N., 1988. Computers and Arabic Language. Al-Khat Publishing Press, Ta'reep, Egypt.

Almeman, K., Lee, M., 2013. Automatic building of Arabic multi dialect text corpora by bootstrapping dialect words. Communications, Signal Processing and their Applications (ICCSPA), 1st International Conference on 12–14 Feb, pp. 1–6.

Alotaibi, Y.A., Selouani, S., Cichocki, W., 2009. Investigating emphatic consonants in foreign accented Arabic. J. King Saud Univ. Comp. Inf. Sci. 21 (1), 13–25 (Riyadh).

AL-Shalabi, R., Kannan, G., AI-Serhan, H., 2003. New approach for extracting Arabic roots. Proc. of 2003 International Arab conference on Information Technology (ACIT'2003), Alexandria, pp. 42–59.

Al-Shareef, Sarah, Hain, Thomas, 2011. An investigation in speech recognition for colloquial Arabic. In proceeding of: INTERSPEECH 2011, 12th Annual Conference of the International Speech Communication Association, Florence, Italy, August 27–31, pp. 2869–2872.

Cote, Robert A., 2009, Choosing one dialect for the Arabic speaking world: a status planning dilemma, 75 Arizona Working Papers in SLA & Teaching 16, 75–97.

Diab, M., Habash, N., 2006. Arabic Dialect Processing. AMTA, Bostan.

Diab, M., Hacioglu, K., Jurafsky, D. 2004. Automatic tagging of Arabic text: from row text to base phrase chunks. In 5th Meeting of the North American Chapter of the Association for Computational Linguistics/Human Language Technologies Conference (HLTNAACL04), Boston, MA.

Duwairi, R., Al-Refai, M., Khasawneh, N., 2007. Stemming versus light stemming as feature selection techniques for Arabic text categorization. Innovations in Information Technology, IIT '07. 4th International Conference on, 18–20 Nov, pp. 446–450.

El-Sadany, T.A., Hashish, M.A., 1988. Semi-automatic vowelization of Arabic verbs. Proceedings of 10th National Computer Conference, pp. 45–56.

Gleason, H.A., 1970. An Introduction to Descriptive Linguistics, third ed. Holt. Rinehart and Winston, New York.

Goweder, A., Alhami, H., Rashed, T., Al-Musrati, A., 2008. A hybrid method for stemming Arabic text. In Proceedings of the 9th International Arab Conference on Information Technology (ACIT 2008) (Tunis, 2008).

Habash, N., Rambow, O., 2005. Tokenization, morphological analysis, and part-of-speech tagging for Arabic in one fell swoop. In Proceeding of the Association for Computational Linguistic (ACL).

Habash, N., Soudi, A., Buckwalter, T., 2007. On Arabic transliteration. In: van den Bosch, A., Soudi, A. (Eds.), Arabic Computational Morphology: Knowledge-Based and Empirical Methods. Springer.

Harrag, F., El-Qawasmah, E., Al-Salman, A.M.S., 2011. Stemming as a feature reduction technique for Arabic text categorization. Programming and Systems (ISPS), 10th International Symposium on, 25–27 April, pp. 128–133.

Kadri, Y., Nie, J.Y., 2006. Effective stemming for Arabic information retrieval. In Proceedings of the Challenge of Arabic for NLP/MT Conference. The British Computer Society. London, UK.

Kareem Darwish, 2002. Al-stem: a light Arabic stemmer. [Online]. Available: http://www.glue.umd.edu/~kareem/research.

Khoja, S., Garside, R., 1999. Stemming Arabic text. Computing Department, Lancaster University, Lancaster, 15 (April 2012). Doi: http://zeus.cs.pacificu.edu/shereen/research.htm.

Maamouri, M., Bies, A., 2004. Developing an Arabic treebank: Methods, guidelines, procedures, and tools. Linguistic Data Consortium (LDC).

Mutahhar, A.R., Watson, J., 2002. Social issues in popular Yemeni culture. Yemeni–British project supported by the British Embassy, Social Fund for Development and Leigh Douglas Memorial Fund, Sana'a, Yemen.

Saliba, B., Al-Dannan, A., 1990. Automatic morphological analysis of Arabic: a study of content word analysis. Proc. First Kuwait Comput. Conf., 231–243

Zaidan, O.F., Callison-Burch, C., 2013. Arabic dialect identification. Comput. Linguist. 1 (1), 1–36.

Zina Saadi, One language, many dialects: an analysis of Arabic dialects. Computational Linguist. Middle Eastern Languages Specialist. Basis Technology corp. Found online (accessed 03.2013).