Univ Access Inf Soc (2018) 17:131–145
https://doi.org/10.1007/s10209-017-0522-3

LONG PAPER

Universal web accessibility and the challenge to integrate informal Arabic users: a case study

Aqil M. Azmi[1] · **Eman A. Aljafari**[1]

Published online: 3 February 2017

Abstract Most Arabs can read text written in Modern Standard Arabic (MSA). However, to easily express themselves, they may find it easier to switch to informal (colloquial) Arabic. The web is open for anyone to express him/herself freely, and people are expressing themselves through many social media platforms, such as blogs and forums increasingly in their native colloquies. Search engines are very good at handling queries in MSA, though not as good if the query is written in colloquial Arabic. Two issues will be addressed in this paper. First, many younger generation Arabs find it hard to write in MSA, which means that many results are missed due to improperly posted queries; and second, a query written in MSA will not retrieve documents written in colloquial Arabic. Thus, with the goal of universal accessibility of the web to all Arabic users, we need a successful mechanism that translates the query back and forth between MSA and the variety of colloquies spread throughout the Arab countries. As a case study, we investigate one of the local dialects in Saudi Arabia, a leading country in social media usage much of which is in colloquial language. We present a web information retrieval system for Arabic that addresses this concern. To test the proposed method, we compiled a corpus of over fourteen hundred documents and measured the performance of our system using 50 sample queries achieving an average recall and precision of 93.4 and 83.6%, respectively.

Keywords Arabic NLP · Colloquial Arabic · Web accessibility · Revised *n*-gram

1 Introduction

This paper addresses the subject of retrieving Arabic web contents based on a dialect, by studying different concepts in this area and the means to process an Arabic dialect. It is hoped that this work offers a simple yet comprehensive treatment method for parsing one of the regional Arabic dialects. We conclude by proposing a general framework for an Arabic information retrieval system. Throughout the paper, the words *colloquial* and *dialect* interchangeably are used. Indeed, Arabic is an old language that—to the surprise of many—precedes Islam. This is evident from the recent discoveries of pre-Islamic Arabic inscriptions from the second and the fourth centuries CE (see [3], pp. 123–129). What is more interesting is that even today, the majority, if not all Arabs, can read and understand the Holy Qur'an and the Hadith (Prophet Muhammad's sayings and tradition). Both are 1400-year-old texts. Bellamy [13] insists that the Arabic inscription at Jabal Ramm, believed to be from the fourth century CE, is closer to modern Arabic than Shakespeare's language to modern English. Arabic is the native language of over 300 million speakers [17] and over 1500 million worldwide Muslims who use it in their regular daily prayers.

Arabic is a Semitic language and can be classified as classical and modern. Classical Arabic represents the pure language used by the Arabs, the language the Qur'an was revealed in, while Modern Standard Arabic (MSA) is an

✉ Aqil M. Azmi
aqil@ksu.edu.sa

Eman A. Aljafari
eman_2242@hotmail.com

[1] Department of Computer Science, College of Computer and Information Sciences, King Saud University, Riyadh 11543, Saudi Arabia

evolving Arabic with constant borrowing and innovation to meet the modern challenges [26]. There are 28 basic letters in the Arabic alphabet. In addition, the Arabic orthographic system uses small diacritical markings to represent the three short vowels (a, i, u), shown in Table 1 (first three entries). The Arabic sound system consists of a total of thirteen different diacritics, the first seven of the basic diacritics (Table 1) and six compound diacritics. The latter are a combination of the syllabification mark *shaddah* along with either a short vowel or a nunation. Note that the diacritical mark *shaddah* does not appear in a standalone form. The markings are placed either above or below the letter to indicate the phonetic information associated with each letter to clarify the sense and meaning of the word. For example, the undiacritized word (عقد) has many different meanings some of which are: (عِقْد: *Eiqod*) *necklace*, (عُقَد: *Euqad*) *knots*, (عَقْد: *Eaqod*) *contract*, and (عَقَّد: *Eaq*∼*ad*) *complicated*. Wherever necessary, we use Buckwalter transliteration scheme (www.qamus.org/transliteration.htm) for those having difficulty following Arabic script. The Buckwalter transliteration has been used in different NLP publications, and its main advantage is that it is a strict one–one transliteration of Arabic using standard ASCII symbols (see [31], p. 21). The lack of diacritical markings often causes ambiguity. This is specially true for sentences, and even for natives adept at resolving, there are cases where they fail. Consider the undiacritized sentence, (سبقنا القطار*sbqnA AlqTAr*). It could either mean *we surpassed the train*, or *the train surpassed us*. With no context, only the proper diacritical marking would reveal the meaning. Another example where the lack of diacritical markings can only be resolved through world knowledge, (قررت الحركة الإسلامية وشخصيات معارضة الإنسحاب من حوار وطني*qrrt AlHrkp Al-AslAmyp w$xSyAt mEArDp AlAnsHAb mn HwAr wTny*) (see [12], p. 479). This sentence could either mean *The Islamic movement and (other) figures were opposed to withdrawal from the national dialogue*, or *The Islamic movement and opposition figures decided to withdraw from the national dialogue*. The meanings are completely opposite. To avoid such a scenario, this is one of the reasons why most of the religious texts are heavily vowelized.

Ferguson [27, 28] defined diglossia as a phenomenon whereby more than one variety of the same language exist side-by-side in the same speech community. Each variety is used for a specific purpose and in a distinct situation. In Arab countries, it is natural to see at least three varieties of Arabic coexist in a complex interaction [48, 26]. Classical Arabic is used in religious contexts and idiomatic expressions. Most written contexts use MSA, including television broadcasts that are read out loud. However, novels, personal letters, and increasingly Internet posts and texting are written colloquially [21]. Till recently, the society shunned those who wrote in colloquial; which was considered a domain for verbal communication. However, the web gave people equal opportunity to write in whatever language suites them. And so, the society's custody of the written language was relaxed. Presently, the printed media is typically written in MSA, yet in informal cases, e.g., personal communications, blogs, forums, tweets, people tend to communicate using a blend of MSA and colloquial Arabic. The colloquial or the dialectical Arabic differs from region to another, and the vocabulary of some dialects overlaps with MSA by as much as 90% (see [16], p. 254); however, differences include some very common words such as those meaning "see," "go," and "not," as well as phonology, syntax, and morphology rules [32]. Unfortunately, these rules are not written and there are no available dictionaries for their vocabulary. The range of dialects that involve Arabic is much more varied than the range of dialects that are typically considered to comprise European languages such as English and French. This makes the linguistic environment much more fluid and harder to translate using some automated techniques, e.g., machine translation [49]. It is noteworthy that automated tools, e.g., Google Translate, are able to translate MSA text

Table 1 The basic Arabic diacritics are grouped into three sets

Diacritic Set	Diacritic on the letter ب	Buckwalter transliteration	Name	Pronounciation
Short vowels	بَ	ba	fatha (فتحة)	/b//a/
	بُ	bu	damma (ضمة)	/b//u/
	بِ	bi	kasra (كسرة)	/b//i/
Nunation	بً	bF	tanween fath (تنوين فتح)	/b//an/
	بٌ	bN	tanween damm (تنوين ضم)	/b//un/
	بٍ	bK	tanween kasr (تنوين كسر)	/b//in/
Syllabification marks	بْ	bo	sukon (سكون)	/b/
	بّ	b∼	shaddah (الشدة)	/b//b/

The nunation can only be placed at the end of the word. The syllabification mark *shaddah* only occurs with either short vowel or nunation

reasonably well, though sadly fail for a text written in one of the many Arabic dialects. Cote [18] found prevalence of the Egyptian dialect throughout MENA (Middle East and North Africa) region, and it is not limited to the scope of "drama." We must also consider the popular culture of Egypt portrayed in media/print in the MENA region and the temporal dynamics of diffusion of the web in the region. Now, consider the tweet by +@alaa+ written in Egyptian colloquial:

في تقديري ده معناه ان العصار هو اللي مستقصد نوارة وبيبعت ناس تضربها، بيحكمنا عيل كذاب بيروج اشاعات واهبل كمان مش عارف يحبكها

Weyman [49] provided the following Google translate of the tweet, *In my opinion means that the de Assar is the Old Mstqsd Nawara Bebat and hit people, Bagmana Eil and rumors bouncing Barog ***hole violin not know Ihpkha*. At the time of writing, we did our own Google translate of the above tweet and found that Google has slightly improved its translation skills. For another example, consider the Nejdi Arabic colloquial expression (لايق عليتس الفوز: *lAyq Elyts Alfwz*) meaning *Winning befits you* (fem.). It was Google translated to, *The uncertainty Alits win*. Nonetheless, both are complete gibberish translations, and we expect a similar miserable handling by search engines for searches involving colloquial Arabic. A somewhat related story: Few years back, a friend asked Omar (son of the first author), who just started to learn Arabic in school, (شلونك: $*lwnk*) *how are you?* for the unfamiliar it very much sounds like (ايش لونك: *Ay*$ *lwnk*) *what is your color?*, and he innocently replied white.

In this digital age, the Internet constitutes the main source of information for many users. This presents not only a great opportunity, because the material is readily available under the fingertip, but also a major challenge, since the aspect of dialects hinders search effectiveness. Whereas major search engines support searching in MSA, the presence of dialect words in the query makes it harder to retrieve the desired result. Googling, for example, the query *what happens if the children met* expressed in a local Saudi dialects gets 115,000 hits, while its equivalent MSA query will result in 5,620,000 hits. There is a healthy amount of literature devoted to different aspects of MSA natural language processing, though it is a far cry from other more established languages, e.g., English. The processing studies for Arabic language dialects are relatively new and cannot be compared to the enormous work that is done in other languages such as English dialects [6]. One reason for being such a latecomer was the shortage of web content in dialectical Arabic. Initially, most of the Arabic content in the web were in MSA, and only recently did blogs, forums, and social media, etc., gain widespread acceptance among the Arabs. The social media is an area where colloquial Arabic is profusely used. A Wiki entry, under "Varieties of Arabic," lists the following as a major Arabic dialect group: Arabian Peninsula (includes Saudi and Arabian Gulf), Egyptian, Mesopotamian, Levantine, Yemeni, Sudanese, and Maghrebi dialects. This classification is rather based on geographical locations, and there are many instances whereby a certain region may have more than one dialect. Some authors do not recognize Saudi as a distinct dialect, rather placing it under the Gulf dialect (see [20], pp. 4–5). As a relatively large country, Saudi Arabia features different dialects. Only a fraction of the population (approximately 200,000) speaks the Gulf dialect (www.ethnologue.com/country/sa/languages).

Many of the younger generation Arabs have a hard time writing in MSA. However, even if we ignore this group of users, many of the relevant documents will not be looked at with the present search engines. The search engines themselves are unbiased. They use algorithms to match whatever they index against whatever they see as queries. Suppose a user who is incompetent in writing in MSA, queries the search engine using colloquial Arabic, the search engine will only retrieve certain matching documents. It will not retrieve any of the relevant documents in MSA since the keywords do not match. This argument is true both ways. An MSA query will overlook relevant documents written in colloquial Arabic. Therefore, there is a need to develop a transparent system that is able to map between MSA and its various dialects. This ensures that all related documents are retrieved regardless in what Arabic (standard, dialectal, or a combination) they are in, and irrespective of the Arabic used in the query itself. It might be a good idea to introduce a tag to the query telling the search engine what colloquial it is in. There is an explosive growth of Arabic web content, an increase of 2500% since the year 2000 [19]. Twitter is a particularly fast-growing domain, as the Arabic use on Twitter grew 22-fold in a one-year period between October 2010 and 2011, spurred by the uprisings and revolutions in the region, making Arabic the eighth most used language on Twitter [44]. These statistics suggest a dire need for a unified scheme to retrieve Arabic texts transcribed in MSA and Arabic dialects.

With so many Arabic dialects, it is difficult to include them all in a single study given that each can be treated as a separate language. After consideration, we decided to go for one of the local dialects used in Saudi Arabia. This is one of the fastest growing countries in social media use, where 97% of the users prefer to use Arabic for browsing [5]. According to Alexa (www.alexa.com) statistics for May 2013, Twitter was ranked the seventh most visited site in Saudi Arabia, a country that ranks second among the

world's fastest growing countries using Twitter [37]. Ninety percent of the tweets in Saudi Arabia are in Arabic [5]. The capital Riyadh alone accounts for 50 million tweets, that makes it grip the tenth position as a city worldwide in terms of tweets per month [37]. Due to its enormous geographic area, there are six dialectical groups within Saudi Arabia [47], we, however, will pick the dialect that is used in the capital of the kingdom, Nejdi Arabic. About 8 million of the kingdom's 30+ Million population uses this dialect (www.ethnologue.com/country/sa/languages). Since one-third of the Saudi population are expatriates, this translates to 40% of the native population using the Nejdi dialect.

The rest of the paper is organized as follows. In Sect. 2, we look into related works. Section 3 goes over the challenges and difficulties MSA and Arabic colloquies presents. We describe our proposed design in Sect. 4. In Sect. 5, we evaluate our proposed Algorithm. Finally, Sect. 6 concludes the paper with future work.

2 Related work

Arabic is one of the most widely used languages in the world [51]. The current Arabic language is an assortment of Modern Standard Arabic, which has a standard orthography, and dialectal Arabic, which does not have a writing standard and commonly used in everyday conversations and on the web discussion [34]. MSA is the language that the children are taught at school. The varieties of Arabic dialects are considered a lower form of expression; therefore, not granted the stature of MSA, which has a great impact in the lack of using Arabic dialects in daily writings. On the other hand, Arabic dialects have gained the stature of living languages in the web because they are the native tongue of millions of people. Consequently, a lot of serious efforts appeared in the last few years to study the syntax and morphology patterns in the varieties of Arabic dialects. These are not enough and there is a serious need for more effort to build robust tools and applications for processing these dialects [24].

As stated earlier, among the chief Arabic dialect groups are: Saudi, Khaliji, Egyptian, Iraqi, Levantine, Yemeni, Sudanese, and Maghrebi dialects. Looking over an MSA word form vs. its colloquial form, we see that some words in MSA are totally transformed in some of the dialects. For example, the MSA sentence: (أنا أكتب) which means *I am writing*, becomes: (أنا عمّال أكتب:*nA Em*~*Al* >*ktb*) in Egyptian; (أنا دأكتب:>*nA d*>*ktb*) in Iraqi; (أنا عم بكتب:>*nA Em bktb*) in Levantine; and (أنا كنكتب>*nA knktb*) in Moroccan. Also, different dialects have their own morphological rule. For example, in Levantine the present tense begins with the prefix (ب), often preceding by the morpheme (عم). To negate a sentence in Saudi colloquial, we have to use the morpheme (ما), so the negation of the sentence (أنا أعرف:>nA AErf) meaning *I know* is (أنا ما أعرف:>*nA mA*>*Erf*). While in Egyptian the suffix (ش) is appended to the word with (ما), so the negation of the sentence (أنا بعرف:>*nA bErf*) *I know* (Egyptian colloquial) is (أنا ما بعرفش:>*nA mA bErf*$).

There have been several attempts at trying to create Arabic resources through analyzing and processing Arabic dialects found online in blogs and social media. One such project was MAGEAD, a brainchild of Columbia University. MAGEAD (Morphological Analyzer and Generator for the Arabic Dialects) [33, 32] addresses the necessity for processing Arabic dialects morphology and Arabic dialects generation. It aims to define a unified processing architecture for all the Arabic dialects morphology besides MSA. To build an Arabic language morphological analyzer and generator for MSA and all of its dialects, the authors define the language words attribute-value for morphological features such as gender or number (single/dual/plural). MAGEAD represented the words in three levels: lexeme level, morpheme level, and surface level. The lexeme level represents the word in terms of stems and dialect-independent features. While the morpheme level the words are represented in terms of morphemes at the surface level, it gives the orthographic representation of the word. The authors devised "morphological behavior class" (MBC) which is used to map the features to their morphemes. MBC is useful in cases such as finding the feminine form of a word which is not always trivial, e.g., `[+FEM]`, for the morpheme (كاتب: *kAtb*) meaning *writer* is (كاتبة: *kAtbp*), while for (أبيض:>*byD*) meaning *white* it is (بيضاء: *byDA*'). The system was further enhanced to accept Levantine as a dialect alongside MSA by changing the linguistic knowledge representation for the work that is done on MSA. The MBC was expanded to include the Levantine postfix negation marker and aspectual particle, and the morphemes order in context-free grammar (CFG) was extended to handle these two situations. In the evaluation phase, MAGEAD analyzer was evaluated for both MSA and Levantine, and the results show that MAGEAD is a flexible analyzer for any Arabic dialects.

Al-Gaphari and Al-Yadoumi [8] designed a morphological rule-based method to convert the regional dialect of the capital of Yemen, known as San'ani dialect, to MSA. They used a simple MSA stemmer, and no root dictionary was involved in this step. The authors reported that many of the distorted words in the dialect depended on the immediate neighboring word. Based on this observation, they devised syntactic rules and a stemming process. Their method was able to handle around 77% of the words in the corpus.

The COLABA [24] is an ambitious project to produce loads of special resources and processing tools to serve Arabic dialects. It was initiated to process data from Arabic social media, blogs, forums, and chat rooms. Recognizing the fact that the language used in such forums is dialectical Arabic, the COLABA project focused on information retrieval (IR) as a way of processing dialectical Arabic. The IR system retrieves relevant dialectical Arabic data simultaneously with data under the standard MSA format, thus allowing users to retrieve as much relevant content as possible. In order to convert the query terms from MSA into the required dialect, an MSA to a dialect term lexicon was built to find the word's symmetry. In addition, the authors used MAGEAD [32] to find Arabic verbs and nouns varieties. The evaluation data were collected from the web covering different genres: politics, religion, and social issues in a variety of Egyptian, Iraqi, Levantine, and Moroccan dialects. The data were filtered in favor of those with more dialect contents. Then, the documents were ranked according to their degree of dialectness. This was determined using an MSA analyzer software that indicates the number of non-MSA words in documents. For each document, the dialect words were added to the lexicon. Finally, they manually annotated each word to determine its dialectal type. The COLABA IR system takes MSA query terms and expands them by generating their MSA-inflected forms along with their corresponding dialects forms. For example, the MSA word (أصبح: *>SbH*) is inflected to forms, e.g., (سيصبح: *sySbH*), (أصبحنا: *>SbHnA*); to their MSA-inflected form with the dialects affixes, e.g., (هيصبح: *hySbH*), (هيصبحوا: *hySbHwA*); and to their dialects forms with the dialects affixes, e.g., (بقى: *bqY*), (هيبقى: *hybqY*). This system used MAGEAD to analyze the MSA verbs and nominals by using MAGEAD's analysis system and then retrieve its dialect equivalents from the lexicon. After that, MAGEAD's generator is used to generate the MSA and dialects words different forms. Finally, all the generated words are returned in the original query context after removing the repeated words. Unlike COLABA which expects the input to be MSA word(s), our system does not impose any such restriction on the input allowing for a combination of MSA and dialectal Arabic words.

Shatnawi et al. [45] proposed a framework to improve the Arabic language IR by enabling users to write queries in Jordanian dialect. This system maps the user's dialect queries to their equivalent ones in MSA by using a CFG. The grammar was built to ensure that the query sentence conforms to the Jordanian dialect syntax. For CFG, the query terms must be type-known. The term type can be a verb, an adjective, or a noun. Depending on the structure of the sentence and the given set of affixes, it is possible to extract the term type. This is a major drawback of the system, since determining the type is a non-trivial task. The authors' simplistic scheme to determine the type by making use of the affixes associated with the three kinds of verbs: present, past, and imperative is problematic. The problem is there is an overlap between these verbs in the use of affixes. The system checks the queries convention with the proposed grammar, processing those that pass the convention. This is followed by preprocessing the dialectical query, e.g., stop-words removal, stemming. After that, the results are used to map the dialectical query terms affixes to their equivalent in MSA and the search is continued in the traditional way. The authors concluded that using dialectical queries yields slightly better results than pure MSA queries.

3 Challenges in using Arabic and its dialects for retrieving information

Arabic is a challenging language to work with in IR. Below, we list some features of the Arabic language and its dialects that show how significant the challenges are:

1. Orthographic variations (dialects only): Due to the absence of vocabulary dictionary of the dialectical Arabic, there is no standard orthography [40]. Often, the natives will spell the words/sentences phonetically, which means the possibility of multiple spelling of a single word within the same dialect, e.g., (بكره: *bkrh*) and (بكرى: *bkrY*) for *tomorrow*.
2. Complex morphology (MSA and dialects): There is a great complexity in morphological analysis, as Arabic is highly inflectional and derivational. Morphology deals with the internal structure of words and it is considered a base layer for other linguistic layers [11]. Arabic morphology is systematic though fairly complex. There are two properties that are used to build words: derivation and agglutination. The derivation process is a powerful word-generation mechanism that makes Arabic the richest vocabulary language compared to other languages [11]. Arabic words can be classified into: nouns, verbs, and particles. The words are generally based on a "root" which uses three consonants to define the underlying meaning of the word. The three consonants are represented by the letters ف، ع، ل which serves as generic letters to represent the first, second, and third letters of the Arabic triliteral roots. The derivation of a word from a given root and a pattern is done by replacing the generic letters of the root in the pattern with the given letters of the root (Fig. 1). This derivation process produces what we call "stem" [12]; and it justifies the reason for describing Arabic as a derivative language. The process of stem derivation yields a huge number of stems that gain their meaning from both roots and the patterns [11]. Classical Arabic has some 9000

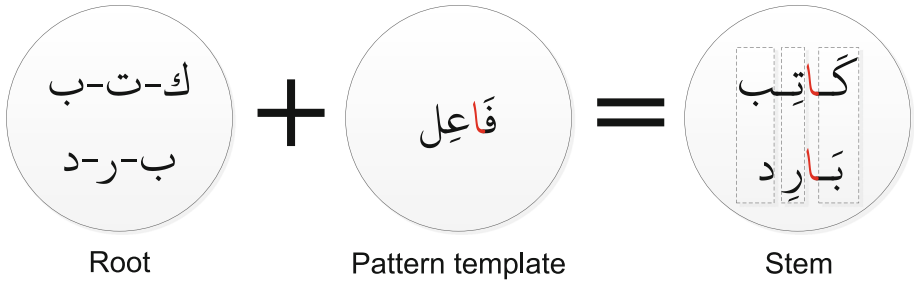

Fig. 1 Sample derivation process which produces a *stem*. The stem is generated by replacing the *letters* ف, ع, and ل in the pattern template with the first, second, and third letter (respectively) of the triliteral root

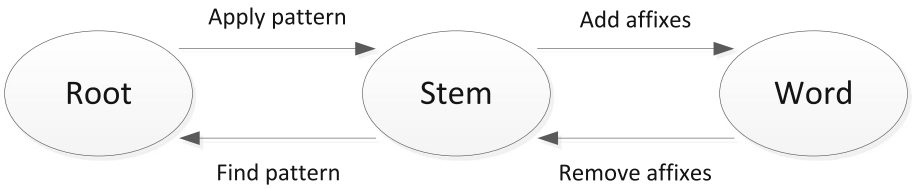

Fig. 2 General Arabic word construction system. An affix is a morpheme that is attached to a stem to form a new word. In Arabic, the affix could be any combination of prefix(es) and suffix(es)

roots, of which 1200 roots are in MSA [35]. The agglutination property of the Arabic language glues stems that were generated using the derivation property with affixes (prefixes and suffixes) to create a desired inflection of meaning. Figure 2 shows the general Arabic word construction system using the two properties of the language. For example, the word وكاتبتهم (= و + كاتب + ت + هم) is composed of a conjunction (و: *w*), a verb (كاتب: *kAtb*), a subject pronoun (ت: *t*), and an object pronoun (هم*hm*). The meaning of the word is, *and I wrote to them*. Its stem is (كاتبت: *kAtbt*) *I wrote to*, while the root is (كتب: *ktb*) *write*. The inflection introduces an additional challenge to Arabic IR, as the query keyword(s) may appear in a different inflected form in the document.

3. Ambiguity (MSA and dialects): In this regard, Arabic is a notorious language. Consider the word "read" in English, it could be read as a past tense or a present tense depending on its context. However, in Arabic, the ambiguity is more subtle. Words with similar spelling may have different pronunciations and meaning that can only be determined by the context and proper knowledge of the grammar [42]. A task where the natives are often successful at. Even then, there are times when ambiguity persists and the only way out is to use the diacritical markings علامات التشكيل (also known as حركات), which helps in clarifying the sense and meaning of the word[1]. Consider, for example, the sentence, (كتبا كتبت) which could mean (كَتَبْتُ كُتُبًا: *katabotu kutubF*) *I wrote books*, or (كَتَبْتَ كُتُبًا: *katabota kutubF*) *you wrote books* (masc.), or (كَتَبْتِ كُتُبًا: *kataboti kutubF*) *you wrote books* (fem.). Here, as always the marking have fully resolved the case. That is why classical Arabic texts and especially religious books have always used the diacritical marking. This practice has subsided, and MSA texts are seldom written with diacritics and are totally absent in the dialects. The ambiguity due to the absence of diacritics falls into several categories [12]. Of concern in the context of this work are syntactic ambiguity and internal word structure ambiguity. The previous example illustrates the case of syntactic ambiguity. The latter is due to the agglutination property of Arabic, e.g., (كتاب) may either be (كِتَاب: *kitaAb*) *book*, or (كُتَّاب: *kut* ∼ *aAb*) *writers*. Both cases can only be resolved through diacritics. When analyzing an undiacritized MSA text, Debili et al. [23] reported an average of 11.6 possible diacritizations for very non-diacritized word. What this means is for each undiacritized word it is possible to have about eleven different interpretations/meanings. Without the diacritical markings, many of the applications, such as text-to-speech, machine translation, and information retrieval, will suffer greatly [12]. On the other hand, we have a competing view which states that automatic diacritization is computationally expensive and is further compounded by the diacritization of previously unseen words which is generally intractable. Given that, we might as well remove all the diacritics before retrieval at the expense of

[1] In the Latin alphabet the diacritics are used to change the sound value of the letter to which they are added, while in Arabic they serve as a vowel pointing system. Distinct letters serve as long vowels, but for short vowels the diacritical markings are used. See Sect. 1 for more detail on the diacritical marking.

increased ambiguity (see [20], pp. 23–24). This view is based on the belief that retrieval is generally tolerant of ambiguity due to word sense [43], and that word sense disambiguation is akin to diacritization [20]. However, it is inconvenient when looking for a rare form of a word, or when all the outcomes are equally likely, e.g., querying (علم الدين). Without the disambiguating diacritical marking, it could either mean *science of religion* or *Alamuddin* (a name such as Amal Alamuddin, actor George Clooney's wife). Both are popular search queries, which mean there is only a 50% chance of retrieving the correct document.

4. Widespread use of synonyms (MSA and dialects) [51]: Some of the synonyms for the Arabic word for *lion* (أسد:*sd*) are: (أسامة:*sAmp*), (الحارث:*AlHArv*), (لبؤة: *lb&p*), (شبل: *$bl*), (ملك الغابة: *mlk AlgAbp*), (حمزة: *Hmzp*), etc. We counted 53 different synonyms for the *she camel* (الناقة: *AlnAqp*) [9]. Interestingly, some of the above listed synonyms for lion, e.g., Asad, Osama, Hamza, also happen to be common name for people. It will be a challenge to know when, for example, Asad refers to the animal or someone's name.
5. Prevalence of irregular plural (MSA and dialects): The identification of irregular (also known as broken) plural in MSA is a problematic issue for IR, e.g., tooth → teeth in English. An example in Arabic for broken plural is (قائد: *qA}d*) → (قواد: *qwAd*) *leaders*. About 41% of the Arabic plurals are broken [15], and these constitute approximately 10% of text in large Arabic corpora [29]. Detecting broken plurals is essential for the stemming procedure, which in turn is an important part of any information retrieval process. According to [30], the best scheme to identify the irregular plural is through a dictionary-based system. The authors continue using light-stemming with a scheme to recognize broken plural improves the performance of IR systems when compared to other schemes that are used in typical IR.
6. Gross misspelling (dialects) [14]: As part of a related work, we compiled a list of spelling errors and classified them into three groups. In the first group, we have errors due to the proximity in the sound of pair of letters: ض and ظ; ت and ة;; and س and ص. For example, (فريظة:*fryZp*) is correctly spelled (فريضة:*fryDp*). In the second group we have errors due to mixing up between the short (diacritical marking) and long vowels, e.g., (موظاعفة: *mwZAEfp*) whose correct spelling is (مضاعفة: *mDAEfp*). In the last group we have words with the letter hamza (ء). In Arabic, the letter hamza appears in one of the following forms: ء (standalone), أ (over the letter alif), إ (below the letter alif), ؤ (over the letter waw), or ئ (over the letter alif-maqsura). There is a complex set of rules which dictates how the letter hamza is written, which people often misspell, e.g., (هؤلاء:*h&lA'*) is misspelled as (هئولاء: *h}wlA'*); or just drop it altogether, e.g., (ياخذ: *yAx**), (ياكل: *yAkl*) instead of (يأخذ: *y>x**), (يأكل: *y>kl*), respectively. To simplify the matter, misspelling can be treated as orthographic variation.
7. Speech effects (dialects) [14]: This is a common phenomenon in social media messaging such as tweets, where one of the letters is repeated many times, e.g., (سلام: *slAm*) is rendered as (سلاااااام: *slAAAAAAm*).
8. Missing spaces between words (dialects) [14]: It is common to spot missing blank between words, e.g., (ماهوب) instead of (ما هوب: *mA hwb*), and (لويعطوني) instead of (لو يعطوني: *lw yETwny*). Alkanhal et al [7] devised a dictionary-centered stochastic scheme that is geared toward detecting and correcting such cases with a very high accuracy. The dictionary is based on a large Arabic corpus, mainly MSA. To handle the dialects, we need to include dialect corpus to the dictionary.
9. Out of vocabulary (OOV) words (MSA and dialects): These include words such as named entities, technical terms, and acronyms. The OOV words are a common source of error in any retrieval system. Davis and Ogden [22], and Al-Fedagi and Al-Anzi [4] report that around 50% of OOV words in Arabic are named entities. The proper way to handle these is by translation; however, when this is not possible, e.g., name of a person, the words are transliterated. Most people, unfortunately, do not follow a standard transliteration rule, resulting in different spellings for the same word. In [2] reported 15 different spellings for the name Condoleezza, with four different (كوندوليزا:*kwndwlyzA*), (كوندااليزا: *kwndAlyzA*), (كوندليزا: *kwndlyzA*) and (كوندوليسا: *kwndwlysA*) found in CNN-Arabic website alone. The same holds true when transliterating Arabic names into English. A study identified 32 different spelling in English for the name of the former Libyan leader, Muammar Gaddafi [50].
10. Foreign words (dialects): Though similar to OOV, we decided against including it since it affects dialects only. In their quest, the Arabs had contact with others; however, being a dominant culture, classical Arabic admitted few foreign words. In contrast, colloquial Arabic has always been open to borrowing from other languages and dialects, e.g., Levantine has a large number of loan words from languages such as Turkish, Persian, and French. Social media texts contain lots of words of foreign origin, particularly English, which are spelled in Arabic. For example, (جلاكسي:*jlAksy*) for the *Galaxy series of mobiles*, and (أوكي:*wky*) for *OK*.

Recently, a new phenomenon started showing up, particularly among the locals in Saudi Arabia and the Gulf region, a

hybrid language that is a combination of English and broken Arabic (colloquial). For example, (عفش حق كتشن: *Ef$ Hq kt$n*) meaning *kitchenware*; and (أنا فيه يروح الحين>*nA fyh yrwH AlHyn*) meaning *I will go now*. It is the consequence of the lack of locals interest in correcting the non-Arab expatriate workers language mistakes [10]. Following the boom years in late 1970s, there was a large influx of educated foreigners, e.g., management, technicians, and skilled workers, who mainly communicated in English. The prosperity saw also an influx of semi- or un-educated expatriate workers, e.g., domestic helpers. According to the UN, expatriates make up more than 30% of the total population, which is even higher in the capital Riyadh. The locals communicated verbally with the latter group using the hybrid language that is gender-free. Though this has not trickled down into the written form, most likely it will start showing up in the future, in particular the next generation of school kids who grew up with this language.

Some of the problems listed above, in particular numbers 7, 9, and 10, can be solved using the revised *n*-gram model. The plain *n*-gram model is used to compute the similarity coefficient of two words, which is defined as the ratio of the number of common *n*-grams in both words, divided by the number of unique *n*-grams in them. This definition, however, ignores the order of the *n*-grams in the target word. In other words, the possibility that a high matching score of two strings may not share the same concept [1]. For example, the bigram similarity coefficient between (التحالفات: *AltHAlfAt*) *the alliances* and (الفاتح:*AlfAtH*) *the conqueror* is $6/7 \approx 85.7\%$and is very high considering that both words are totally unrelated. Ahmad and Nürnberger [1] proposed a language-independent approach for conflation that does not require a prior knowledge of the language, or the predefined rules. The revised *n*-gram model insists that the order of the *n*-grams be maintained when comparing for similarities between the words. Let w_1 and w_2 be the words to be compared and assume without any loss of generality that the word w_1 is shorter of both words. We denote a substring of length k of the word w that starts at position i using w[i:k]. The substring will be empty if $k \leq 0$. Formally, the similarity score S for revised *n*-gram $(n \geq 2)$, and an odd-numbered window of size m is given by

$$S_{n,m}(w_1, w_2) = \frac{\sum_{i=2}^{|w_1|-n+1} \sum_{j=-(m-1)/2}^{(m-1)/2} \pi(w_1[i:n], w_2[i+j:n])}{\#\text{unique n-grams in union of } w_1 \text{ and } w_2}, \quad (1)$$

where $\pi(w, w') = 1$ if $w = w'$, and zero otherwise. The revised bigram similarity coefficient between earlier example words results in a score of $2/7 \approx 28.6\%$, a more reasonable value. Figure 3 features another example. This measure is very practical for Arabic nouns and verbs which are heavily affixed. One final example, the revised *n*-gram similarity coefficient for the words (سلام) and (سلاااام) is 75%. This shows the model is well capable of handling speech effects.

4 Proposed generalized framework for Arabic information retrieval

The wide array of dialects that were seen in many of the postings brought to light the significant differences in language, and therefore, the need to process these different dialects to be easily accessible. There is a need to develop a more refined Arabic text-based searching based on the utilization of Arabic slang and dialectical terms in the search queries. The suggestion is that a more unified framework ought to be in place to enable intended relevant documents be retrieved in both formats, with the classical format and the dialectical format. For any language, the effectiveness of the used query depends upon the system capacity to be compatible with the used language by means of understanding the language characteristics [25]. So we will start by going over the differences between Nejdi (a dialect of our choice for this study) and MSA.

4.1 An in-depth look at the Nejdi dialect

As there is no known corpus for the Nejdi dialect, we compiled our own. We started with the set of comments written by the online readers of the electronic edition of Alriyadh (www.alriyadh.com), one of the most widely circulated printed newspapers in the capital, Riyadh. This turned out to be a good source for a text that is a rich combination of MSA and dialect. For larger samples of dialectal writing, we turned to another resource. Given that Saudi Arabia is among the world's fastest growing countries with Twitter [37], a prolific resource for dialectal writing, we actively looked into tweets. Going over a large collection of tweets, we manually compiled a small corpus of 240 tweets. Combining both resources (online comments and the tweets), we compiled a large list of Nejdi dialect words (verbs and nouns), along with their stem, and the corresponding MSA equivalent word and dialect stop-words list. The list of 255 dialectal words and their corresponding MSA words were divided into nine categories. With the exception of the first category, verbs and nouns, the rest were treated as stop-words. This list is necessary to do a successful back and forth conversion between MSA and dialect, as well as a rich resource to study the properties of the Nejdi dialectal writing. The full compiled list is available upon e-mail request to the corresponding author. Following a careful analysis, we did not observe any syntactic differences between the Nejdi dialect and MSA, though there were numerous morphological differences. Very late into the project, we became aware of two resources for Nejdi dialect words [36, 41]. These resources, albeit old, were pointed out by one of the anonymous reviewers for which we are thankful. We were happy to note that these resources

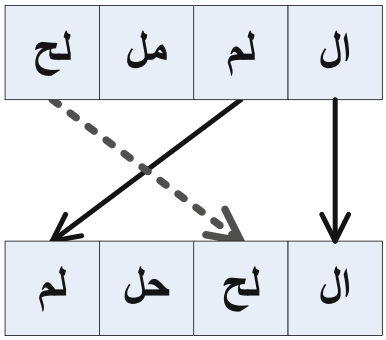

Fig. 3 The bigram similarity between the Arabic word (الملح) *the salt* and (اللحم) *the meat*. The plain bigram similarity measure (all three *arrows*) is $3/5 = 60\%$, and for the revised bigram (*solid arrows* only) it is $2/5 = 40\%$. As both words are unrelated, the second measure is more realistic

second our findings, however, there were some minor differences. Both works go into great length in covering subregional differences in the Nejdi dialect. For instance, [36] divided the Nejdi dialects into four subgroups: Central Nejdi, Northern Nejdi, Mixed Northern-Central, and Southern. The first three groups differ from each other in various ways that are easily characterizable phonologically and morphologically, while the southern group is marked by syntactic and lexical features which link it to the dialects of the south, in particular the Yemen. In this work, we, however, have focused on the dialect used in the capital Riyadh. Originally, the capital city sported a single dialect; however, this is no more true. With a steady growth of population from about half a million, in the early 1970s, to over 7 million inhabitants as of 2015, the dialect used in Riyadh is not as clear cut as it was when these reference books were researched. Many traditional villages and towns in the area around traditional Riyadh which the urban sprawl reached and currently encompasses, and with the population migration from different Nejdi villages into the capital, rendered it having a mixture of various Nejdi dialects. Summarizing the differences:

- There are no orthographic rules for the dialects, so it is very likely words will be spelled differently by different individuals, e.g., (بكرة: *bkrp*) and (بكرى: *bkry*) for *tomorrow*.
- Diminutive form in the dialect has an additional pattern (افعيل) that is not in MSA, e.g., (ابنية: *Abnyp*), (اوليد: *Awlyd*), and (ارجيل: *Arjyl*) in place of MSA words (بنت: *bnt*), (ولد: *wld*), and (رجل *rjl*), respectively, meaning *girl*, *boy*, and *man*, respectively.
- The words, (حقين: *Hqyn*), and (حق: *Hq*) appear often in the dialect. These dialect words have emerged from the MSA root (حقّ) which means *right*, as in human right. In dialect, however, it stands for "belonging" or "that of," a meaning which is unrelated to the root (http://ar.mo3jam.com/term/%D8%AD%D9%82). For example, (حقين سدير) means *those belonging to Sudayr*, or *the people of Sudayr*; and (حق الهيئة) means *an employee of Haia*. The Haia is a short for "The General Presidency of the Promotion of Virtue and the Prevention of Vices."
- The words hundreds, thousands, and months names may be preceded by the letter (ت: *t*) when they are headed by a number, e.g., (خمس تشهر: *xms t$hr*) and (ست تالاف:*st tAlAf*) which, respectively, means *five months* and *six thousand*.
- In the dialect, we have a single relative pronoun (اللي: *Ally*) vs. several in MSA, e.g., (الذي: *Al*y*), (التي: *Alty*), etc.
- The speaker pronoun (ي: *y*) is omitted in verbs. For example, (عطن: *ETn*), (يوجعن: *ywjEn*), and (عرفن: *Erfn*), instead of the correct form (اعطني: *AETny*), (يوجعني: *ywjEny*), and (عرفني: *Erfny*), respectively. These, respectively, mean *give it to me*, *it hurts me*, and *he recognized me*.
- Some MSA words are combined to form a single word in the dialect, such as (منهو: *mnhw*) instead of (من هو: *mn hw*) *who is he?*, and (قلنالكم: *qlnAlkm*) instead of (قلنا لكم:*qlnA lkm*).
- The letter (ك: *k*) in MSA verbs and nouns is converted to (تس: *ts*) in the dialect. For example, *to lie* (كذب: *k*b*) becomes (تسذب: *ts*b*).
- The dialect nouns, circumstances, and adjectives may be preceded by the prefixes (هالـ: *hAl*), (فال: *fAl*), and (عال: *EAl*), e.g., (هالحتسي: *hAlHtsy*) *this talk* and (عالماشي: *EAlmA$y*) *just passing through*.
- The prefix (سي: *sy*) in the present tense verbs in MSA is replaced by (بت: *bt*) or (بي: *by*) in the dialect. The word for *he will sign* in MSA (سيوقع: *sywqE*) becomes (بيوقع: *bywqE*) in the dialect.
- The letter (ئ:}) in MSA is replaced with the letter (ي: *y*) in the dialect, if it is preceded by the letter (ا: *A*), e.g., (جايزة: *jAyzp*) instead of (جائزة: *jA}zp*).

Since there are no available stemmers that can process the dialect, we have to define some conversion rules. The rules will be used to convert a dialect word into MSA, after which it can be stemmed using a standard MSA stemmer. The rules can easily be deducted from above. In addition, we do need Arabic root lexicon to check whether a word can be inflected from a root.

4.2 The IR system

Considering the data posted online, the Arabic retrieval system has to process texts at different levels: pure MSA texts, a text that is a mixture of MSA and colloquial, and pure dialectal texts. It is not uncommon to find online posts written in MSA especially if the writer is mindful of the fact that a wider audience will be reading his/her post. In the pure dialect texts, all of the text is written in a colloquial language, and some features of another different dialect may be found erroneously. A proper system must transparently retrieve information expressed in any of the three levels, equally well, regardless of

whether the query was written in MSA, dialect, or a combination of both. Algorithm 1 is a general framework for our proposed Arabic IR system that caters to all users. One of the standard practices in Arabic NLP is letter normalization. There are some letters in Arabic that are often misspelled using variants, and researchers find it more useful to completely make these variants ambiguous (normalized) (see [31], pp. 22–23). For instance, in Egypt, but not necessarily in other Arab countries, a final Ya (ي: *y*) is often written dotless, *i.e.,* as an Alif-Maqsura (ى: *Y*). So the name Ali (علي: *Ely*) is written (على: *ElY*). As we have no idea of the user's writing style, we normalize Alif-Maqsura to Ya, and the Ya to Alif-Maqsura. This is an added flexibility without imposing any restriction onto the user. Lines 12–14 in Algorithm 1 fall under this category. We need a simple affix removal stemming scheme along with a small lexicon for the dialect-MSA terms and some conversion rules. Table 2 summarizes the set of conversion rules between MSA and dialect (both ways). For affix removal, we decided to adapt the stemmer in Taghva et al. [46] for our problem. The stemmer is meant to handle MSA words, so the changes are either done prior to applying the stemmer or to its output. This same stemmer was used to handle the San'ani dialect in [8].

Algorithm 1: A framework of the proposed algorithm for an MSA/dialectal Arabic information retrieval system.

Input: Search string S

1 **begin**
2 Remove the diacritical marking.
3 Remove special symbols, *e.g.* ~, !, @, #, $, %, etc.
4 Apply conversion rules.
5 Remove the stop-words.
6 Apply stemming algorithm to obtain words stems.
7 Do lexical mapping between dialects and MSA terms.
`// generate all possible related terms`
8 **begin**
9 Look for the stem matching with any affix.
10 If the stem begins with (ا) as an unessential letter, look for a matching stem without its first letter and with all possible affixes.
11 If the stem ends with any of (ة | ت | ه), look for a matching of the stem after replacing (ة → ت and ه), (ت → ة), (ه → ة) respectively.
12 If the stem ends with any form of the letter Ya, *i.e.* (ي | ى), look for a matching stem after replacing it with the other form of Ya.
13 If the stem contains any of (ئ | ء | ؤ), then look for a matching of the stem after replacing (ئ → ء and ؤ), (ء → ا, ؤ, and ئ), (ؤ → ء and ئ) respectively.
14 If the stem contains any form of the letter Alif, *i.e.* (آ | إ | أ | ا), look for a matching stem after replacing it with the other form of Alif.
15 Do the search process.

Just to give a working idea on the proposed algorithm, we will hypothetically apply the algorithm on the sample dialect query (ايش يصير إذا اجتمعوا البزارين: Ay$ ySyr A*A AjtmEwA AlbzAryn), the Nejdi dialect for *what happens if the children met*. For convenience, we will follow the Arabic word with its transliteration (Buckwalter). Steps 2–4 are not applicable on this example. In step 5, we remove the two dialect stop-words, (ايش) and (إذا). Step 6 applies the stemming algorithm, resulting in (يصير: *ySyr*) → (صير: *Syr*), etc. Next, we do lexical mapping and get, (صير: *Syr*) will be mapped into (حصل: *HSl*); (اجتمع: *AjtmE*) will not be mapped; and (بزار: *bzAr*) will be mapped into (أطفال: *>TfAl*). In steps 8–11, for each of the five words: (صير: *Syr*), (حصل: *HSl*), (اجتمع: *AjtmE*), (بزار: *bzAr*) and (أطفال: *>TfAl*), we generate all possible word patterns will all possible affixes, e.g., (بزار: *bzAr*) → (بزارين: *bzAryn*), (وللبزارين: *wllbzAryn*), (البزارين: *Al-bzAryn*), (بزارينهم: *bzArynhm*), etc). Finally, we do a normal search for all the generated word patterns.

For the searching process, we can have two possible options: search for the original query or search for the expanded query. In the former, the system retrieves only the relevant documents of the query terms, while in the latter option it retrieves the relevant documents of the query terms along with the results of conversion rules and the dictionary correspondence terms.

5 Performance evaluation of the proposed algorithm

The main objective of this paper is to integrate the dialectic Arabic users into the realm of standard Arabic users. In the previous section, we proposed Algorithm 1 that processes the queries written in Arabic regardless of what form it was in. The system was implemented using ASP.NET in C# running under Windows. Figure 4 is a screen shot of retrieved results.

To assess the design, we manually compiled a corpus of 1429 documents. The documents in the corpus are composed of 377 MSA documents covering different genres from two local newspapers, Al-Khaleej and Al-Watan, and 1052 manually filtered tweets mostly in colloquial Arabic including tweets expressed in a combination of MSA and colloquial as well. The objective of filtering is to avoid short tweets with profanity. This corpus is different from the one mentioned in Sect. 4.1 which was used to compile the features of the Nejdi dialect. Below are sample tweets from the corpus, followed by their translation.

مواليد السعودية يرتفعون إلى 600 ألف بزر في السنة

The number of birth in Saudi has risen to 600 thousand a year.

Table 2 The set of conversion rules between dialect and MSA

Rule	Dialect → MSA	MSA → Dialect
Numbers	If we have a number expressed in words that is followed by any of the words (تشهر | تالاف | تميات | تمية | تيام), then append (ة) at the end of the number and convert the following words: (تشهر | تالاف | تميات | تمية | تيام) → (أشهر | آلاف | مئات | مئة | أيام) respectively.	A number (ثلاثة | أربعة | خمسة | ستة | سبعة | ثمانية | تسعة) if followed by any of (أيام | مئة | مئات | آلاف | أشهر), then delete the letter (ة) from the number and the following words are rewritten as (تشهر | تالاف | تميات | تمية | تيام) respectively.
Clitics	If any of the clitics (ما | يا | مو | لو | من) appears as a prefix of a word, check the remainder of the word and if it is a pronoun or if it can be inflected from a root, separate them.	
تس	If the word can be inflected from a root then leave it, otherwise replace (تس → ك).	Replace (ك → تس) and force stemmer to treat (تس) as a single letter.
ي	A word containing (اي) is replaced with (ائ) if it cannot be inflected from a root.	Replace (ائ → اي).
اللي	Convert the word (اللي) if followed by (singular masc. | plural masc. | singular fem. | plural fem.) to (الذي | الذين | التي | اللاتي) respectively.	Replace (الذي | الذين | التي | اللاتي) → (اللي).

Rules that calls for checking whether the word can be inflected from a root requires consulting Arabic root lexicon

أعرف واحد مدمن مباريات لدرجة انه مره شاف اخوه يلعب بليستيشن "فيفا" قال : أصبر لا تلعب لين اجي ، وراح يصلح له شاهي وجاء يتفرج

I know one game addict, to the point where he once saw his brother playing Playstation "FIFA", and said, wait, do not play till I return, where he went and prepared a tea and came watching.

A standard measure to evaluate information retrieval with binary classification is precision and recall. Precision (P) is a fraction of the retrieved instances that are relevant, while recall (R) is the fraction of relevant instances that are retrieved. Precision can be considered as a measure of exactness or quality, while recall is a measure of completeness or quantity. Both measures are expressed as numbers ranging between 0 and 1 inclusive. More precisely, they are defined [38] as:

$$P = \frac{\#\text{retrieveditemswhicharerelevant}}{\text{total\#retrieveditems}} \quad (2)$$

$$R = \frac{\#\text{retrieveditemswhicharerelevant}}{\#\text{relevantitemsinthecorpus}} \quad (3)$$

Fig. 4 Screen shot showing the retrieved results of a query

To evaluate the system, we measured its performance on a total of 50 different queries using both options, original and expanded (see Sect. 4.2). To get a good picture, we included queries in MSA and in colloquial Arabic. There were no specific criteria for picking a query other than it should be either in MSA or colloquial. All retrieved results were manually verified. Table 3 summarizes the results. For the original query, the average performance measures for all the queries was 86.64 and 65.04% for precision and recall, respectively. If we go for expanded query, the precision slightly drops to 83.609% though the recall goes up significantly to 93.42%. The slight drop in the precision for the expanded query is normal as the number of retrieved instances has increased and some of them may not be correct. From the results, it can be argued that the expanded query provides a better performance than that we got from the original query.

To give a flavor of the system suppose we issue the query, (مولم عزيمة: *mwlm Ezymp*) which means *prepared a feast*. The word (مولم: *mwlm*) is a colloquial, and using the

Table 3 Performance measure for 50 queries (16 MSA and 34 colloquial) using the search options "original query" and "expanded query"

Queries	Original query		Expanded query	
	P	R	P	R
MSA queries (الأكل حار: *Al>kl HAr*), (الجوال الذكي الأفضل: *AljwAl Al*ky Al>fDl*), (الشاي والقهوة: *Al$Ay wAlqhwp*), (أفضل وظيفة: *>fDl wZyfp*), (جالس عند تويتر: *jAls End twytr*), (سرق: *srq*), (سعر كيلو الطماطم: *sEr kylw AlTmATm*), (عقوبة السرقة: *Eqwbp Alsrqp*), (غسيل الأواني: *gsyl Al>wAny*), (غلط: *glT*), (نحن جاهزون: *nHn jAhzwn*), (حماس: *HmAs*), (مسجون: *msjwn*), (كذبة أبريل: *k*bp >bryl*), (نوم الأطفال: *nwm Al>TfAl*), (جامعة حائل: *jAmEp HA}l*)	0.93	0.74	0.93	0.96
Colloquial queries (كيف نجود الفلوس: *kyf njwd Alflws*), (أخبار دحدرة البترول: *>xbAr dHdrp Albtrwl*), (أسعار التتن: *>sEAr Alttn*), (ابوي يعصب: *Abwy yESb*), (اخوي هرب: *Axwy hrb*), (العزايم في عايلتنا: *AlEzAym fy EAyltnA*), (الغذاء والدواء تعذرب في مياه القصيم: *Alg*A' wAldwA' tE*rb fy AlqSym*), (الكهرب: *Alkhrb*), (الكهرب يطفي: *Alkhrb yTfy*), (الولد زعلان: *Alwld zElAn*), (اليمن تتشره على السعودية: *Alymn tt$rh Ely AlsEwdyp*), (اوباما يجدع من راتبه: *AwbAmA yjdE mn rAtbh*), (أربع تالاف: *ArbE tAlAf*), (أشرب موية: *A$rb mwyp*), (أوقات دخول العوايل في الجنادرية: *>wqAt dxwl AlEwAyl fy AljnAdryp*), (دايم أتحلطم: *dAym >tHlTm*), (سواليفه سامجة: *swAlyfh sAmjp*), (شرهان: *$rhAn*), (عيالنا ما يذاكرون: *EyAlnA mA y*Akrwn*), (فلم سامج: *flm sAmj*), (قايلتن لكم اخمدوا: *qAyltn lkm AxmdwA*), (لايق عليتس الفوز: *lAyq Elyts Alfwz*), (مبسوط: *mbswT*), (مبطي ما صورت: *mbTy mA Swrt*), (مخلوق بشر: *mxlwq bvr*), (مولم عزيمة: *mwlm Ezymp*), (مية: *myp*), (ميتين: *mytyn*), (يا ويل اللي يزعل بناتنا: *yA wyl Ally yzEl bnAtnA*), (يروح يرجم الجمرات: *yrwH yrjm AljmrAt*), (البنات ما يصبرن عن النوتيلا: *AlbnAt mA ySbrn En AlnwtylA*), (مين اللي مبتحل بالمساجين:)myn Ally mbtHl bAlmsAjyn, (انكب العشاء: *Ankb AlE$A'*), (بيدبرون بترول لمصر: *bydbrwn btrwl lmSr*)	0.84	0.61	0.79	0.92
Overall average	0.87	0.65	0.84	0.93

The values for precision and recall are averaged over all the queries

Table 4 Summary of the results for the query, (مولم عزيمة: *mwlm Ezymp*) using both search options

	Original query	Expanded query
Number of relevant instances in the corpus	7	7
Number of retrieved instances	6	8
Number of retrieved instances which are relevant	5	7
List of keywords found in the retrieved instances	(عزيمة، عزيمه، مولمين)	(عزيمة، عزيمه، مولمين، مجهزين، مجهزة)
List of irrelevant keywords found in the retrieved instances	retrieved a document with the phrase (عزيمة رونالدو)	
List of keywords in the relevant instances that were not retrieved	(مجهزين، مجهزة)	–
(Precision P, Recall R)	(0.83, 0.71)	(0.87, 1.0)

dictionary, the corresponding MSA term is (مجهز: *mjhz*). The second word is stemmed into (عزم: *Ezm*). Table 4 details the result for this query with both search options. An erroneously retrieved document has the Arabic phrase (عزيمة رونالدو: *Ezymp rwnAldw*) which means *Ronaldo's resolve*. The MSA trilateral root (عزم) has several meanings, and this is a case of a homonym word[2]. To exclude cases as such, we need to devise a sophisticated post-processing which is outside the scope of this work. The original query retrieved 6 instances with 5 of them being relevant. So, the precision in this case is $5/6 = 83\%$, and the recall is $5/7 = 71\%$ since there are 7 relevant instances in the corpus. For the expanded query, it retrieves 8 instances, 7 of which are relevant, and so the precision is $7/8 = 87\%$, and the recall is 100%.

As another example, consider the colloquial query (شرهان: *$rhAn*) meaning *angry*. Table 5 summarizes the result for this query with both search options. A possible erroneously retrieved document has the name, (الغضبان: *mrym AlgDbAn*) which is a local actress' name literally meaning *Maryam the Angry*.

For the final example consider another colloquial query (لايق عليتس الفوز: *lAyq Elyts Alfwz*) which means *Winning*

[2] Two or more words having the same spelling but different meanings and origins, e.g., lie (untrue) and lie (recline).

Table 5 Summary of the results for the dialectal query (شرهان: $*rhAn*) using both search options

	Original query	Expanded query
Number of relevant instances in the corpus	8	8
Number of retrieved instances	4	9
Number of retrieved instances which are relevant	4	8
List of keywords found in the retrieved instances	((شرهان، شرهانه))	((شرهان، شرهانه، زعلان))
List of irrelevant keywords found in the retrieved instances	an erroneous document having the name (مريم الغضبان)	
List of keywords in the relevant instances that were not retrieved	(زعلان)	–
(Precision *P*, Recall *R*)	(1.0, 0.5)	(0.88, 1.0)

Table 6 Summary of the results for the colloquial query (لايق عليتس الفوز: *lAyq Elyts Alfwz*) using both search options

	Original query	Expanded query
Number of relevant documents in the corpus	69	69
Number of retrieved documents	46	61
Number of retrieved documents which are relevant	44	59
List of keywords found in the retrieved documents	(لايق، عليتس، فوز، الفوز، بفوزه، وفوز، فوزا، بالفوز، للفوز، يفوز، وفوزا، فوزها، للفوز، بفوز)	(لايق، عليتس، فوز، الفوز، بفوزه، وفوز، فوزا، بالفوز، للفوز، يفوز، وفوزا، فوزها، للفوز، بفوز، اللانق، اللانقة، عليك)
List of irrelevant keywords found in the retrieved documents	two documents w/names (فوزي ناس) and ((فوزي مبارك))	
List of keywords in the relevant documents that were not retrieved	(اللانق، اللانقة، عليك، عليكم، عليكام)	((عليكم، عليكام))
(Precision *P*, Recall *R*)	(0.95, 0.63)	(0.96, 0.85)

Table 7 Performance measure for selected 17 queries using the 'expanded query' option based on the top 10, 20 and 30 returned results

Queries	*P*@10	*P*@20	*P*@30
مبطي ما صورت، الكهرب يطفي، مية، فلم سامج، أشرب موية، دايم أتحلطم، مخلوق بثر، جالس عند تويتر، العزايم في عايلتنا، يروح يرجم الجمرات، قايلتن لكم اخمدوا، أوقات دخول العوايل في الجنادرية، لايق عليتس الفوز، مولم عزيمة، يا ويل اللي يزعل بناتنا، سواليفه سامجة، الغذاء والدواء تعذرب في مياه القصيم	0.768	0.782	0.793

Table 8 Performance comparison between our system (expanded query) and [45]

	Average *P*	Average *R*
Our system	0.84	0.93
Shatnawi et al. [45]	0.54	0.66

The latter is intended for the Jordanian dialect. The performance measure for the other system is as reported in the corresponding literature

befits you (fem.). The search results are summarized in Table 6. We note two erroneous documents being retrieved.

For many applications, particularly web search, what is more important is how many good results are there on the first page or the first three pages. For this, we measure precision at fixed number of retrieved results, say 10 documents. This is referred to as "Precision @ 10," or *P*@10 (see [39], p. 148). Table 7 lists the precision for the top 10, 20 and 30 retrieved results for 17 queries using the expanded query option. These queries are a subset of the 50 queries featured in Table 3.

To see how the system contrasts with other comparable systems, we decided to compare the performance with [45], a system that handles Jordanian colloquial. Though both systems handle different dialects of Arabic, nevertheless it will provide a rough idea on respective performances. The results are summarized in Table 8. It should be noted that each system was assessed on a different corpus using a different set of queries. There could be many reasons for

the difference in the performance: one possibility is the underlying system and another possibility is that certain dialects of Arabic are much harder to handle. In Sect. 2, we mentioned that [45] mapped the dialectal queries to MSA through CFG grammar using a complex task that involved determining the type of the query terms. We can argue that the proposed system is a more feasible solution.

6 Conclusion and future work

Arabic dialects, the spoken form of the language, have moved into the realm of the written domain. Now, the dialect is present in online discussions, emails, social media, blogs, etc. Arabic dialects face many challenges in natural language processing techniques because they are less controlled and more speech like. We have plenty of tools at our disposal to process Modern Standard Arabic (MSA). These tools, when applied to dialects, yield significantly lower performance. This suggests the need to develop dedicated tools for dialect processing. In this work, we looked into the web information retrieval problem. A good information retrieval system must successfully handle queries expressed in either MSA, dialect, or both. The system should be transparent to the user, retrieving all related documents regardless of the Arabic expressed in. The colloquial or the dialectical Arabic differs from region to another; each has its own vocabulary, phonology, syntax, and morphology rules. Complicating the matter is the lack of dialectal vocabulary dictionary. In a way, each dialect can be considered as a separate language. One of the biggest problems facing a researcher is the lack of properly prepared resources covering each dialect. The proposed system addresses many of the challenges presented in MSA and the dialects. We presented a model system that should efficiently handle queries in MSA as well as dialectal, using as a case study one of the local Arabic dialects in Saudi Arabia. We offer two search options, original and expanded query. In the original query, we retrieve the relevant documents of the query terms, while in the expanded query, we additionally retrieve the results of the conversion rules and dictionary equivalence terms. Testing on a manually compiled corpus of over 1400 documents confirms the improved performance we get through the expanded query. The average precision for 50 queries was 83.6%, and the average recall was 93.4%.

With regard to future work, we intend to include other local dialects in Saudi Arabia and compile a comprehensive dictionary of MSA to/from all Saudi dialects. A longer-term goal is to cover other Arabic dialects. In the longer term, we intend to build a system that will handle all the Arabic dialects transparently.

Acknowledgements We would like to thank all the anonymous reviewers for their helpful comments. This work was supported by a special fund in the Research Center of the College of Computer and Information Sciences (CCIS) at King Saud University.

References

1. Ahmad, F., Nürnberger, A.: N-gram conflation approach for Arabic text processing. In: Proceeding of the International Workshop on Improving Non English Web Searching (iNEWS '07), Amsterdam, The Netherlands, pp. 39–46 (2007)
2. Ahmad, F., Nürnberger, A.: Evaluation of N-gram conflation approaches for Arabic text retrieval. J. Am. Soc. Inform. Sci. Technol. **60**(7), 1448–1465 (2009)
3. Al-Azami, M.: The History of the Qur'anic Text: From Revelation to Compilation, 2nd edn. Al-Qalam Publishing, Sherwood Park (2011)
4. Al-Fedagi, S., Al-Anzi, F.: A new algorithm to generate Arabic root-pattern forms. In: Proceedings of the 11th National Computer Conference, Dhahran, Saudi Arabia, pp. 4–7 (1989)
5. Al-Khotani, S.: Kingdom leads growth in Arabic digital content. *Saudi Gazette*, 10 Sep 2013. http://saudigazette.com.sa/index.cfm?method=home.regcon&contentid=20130910179928 (2013)
6. Alamlahi, Y., Ahmed, F.: Sana'ani dialect to modern standard Arabic: rule-based direct machine translation. In: Proceedings of the 2011 International Conference on Artificial Intelligence (ICAI'11) (2011)
7. Alkanhal, M., Al-Badrashiny, M., Alghamdi, M., Al-Qabbany, A.: Automatic stochastic Arabic spelling correction with emphasis on space insertions and deletions. IEEE Trans. Audio Speech Lang. Process. **20**(7), 2111–2122 (2012)
8. Al-Gaphari, G.H., Al-Yadoumi, M.: A method to convert Sana'ani accent to modern standard Arabic. Int. J. Inf. Sci. Manag. **8**(1), 39–49 (2010)
9. Almaktebah AlShamela: http://shamela.ws/browse.php/book-7057/page-69 (2013)
10. Al-Qanair, H.: The effect of migrant workers on the Arabic language in the Gulf region (in Arabic). Alriyadh, 30 Jun 2013. http://www.alriyadh.com/848196 (2013)
11. Attia, M.: Large scale computational processor of the Arabic morphology, and applications. Master's thesis, Cairo, Egypt (2000)
12. Azmi, A., Almajed, R.: A survey of automatic Arabic diacritization techniques. Nat. Lang. Eng. **21**(3), 477–496 (2015)
13. Bellamy, J.: Two pre-islamic arabic inscriptions revised: Jabal Ramm and Umm AlJimal. J. Am. Orient. Soc. **108**(3), 369–372 (1988)
14. Benajiba, Y., Diab, M.: A web application for dialectal Arabic text annotation. In: Proceedings of the Workshop on Semitic Language Processing (LREC-2010), Malta (2010)
15. Boudel, A., Gaskell, M.: A re-examination of the default system for Arabic plurals. Lang. Cognit. Process **17**(3), 321–343 (2002)
16. Cadora, F.: Lexical relationships among Arabic dialects and the Swadesh list. Anthropol. Linguist. **18**(16), 237–260 (1976)
17. CIA: Central Intelligence Agency: World Factbook. Washington, DC (2008)
18. Cote, R.: Choosing one dialect for the Arabic speaking world: a status planning dilemma. In: Arizona Working Papers in SLA & Teaching, vol. 16, pp. 75–97 (2009)
19. Curley, N.: The rise of Arabic on the web. http://wamda.com/2012/04/the-rise-of-arabic-on-the-web-infographic (2012)
20. Darwish, K., Magdy, W.: Arabic information retrieval. Found. Trends Inf. Retr. **7**(4), 239–342 (2013)

21. Daoudi, A.: Globalisation and e-Arabic: the emergence of a new language at the literal and figurative levels. In: Hasselblatt, C., Houtzagers, P., Pareren, R.V. (eds.) Language Contact in Times of Globalization, pp. 61–76. Rodopi, Amsterdam (2011)
22. Davis, M.W., Ogden, W.C.: Free resources and advanced alignment for cross-language text retrieval. In: Proceedings of the 6th Text Retrieval Conference (TREC-6), Gaithersburg, MD, pp. 385–395 (1997)
23. Debili, F., Achour, H., Souissi, E.: De l'etiquetage grammatical a la voyellation automatique de l'arabe. Technical Report. Correspondances de l'Institut de Recherche sur le Maghreb Contemporain 17 (2002)
24. Diab, M., Habash, N., Rambow, O., Altantawy, M., Benajiba, Y.: COLABA: Arabic dialect annotation and processing. In: Proceedings of the Workshop on Semitic Language Processing (LREC-2010), pp. 66–74 (2010)
25. El-Khair, I.: Arabic information retrieval. Annu. Rev. Inf. Sci. Technol. **41**, 505–533 (2008)
26. Farghaly, A., Shaalan, K.: Arabic natural language processing: challenges and solutions. ACM Trans. Asian Lang. Inf. Process. (TALIP) **8**(4), 14 (2009)
27. Ferguson, C.: Diglossia. Word **15**(2), 325–340 (1959)
28. Ferguson, C.: Epilogue: diglossia revisited. In: In Contemporary Arabic Linguistics in Honor of El-Said Badawi, The American University in Cairo (1996)
29. Goweder, A., De Roeck, A.: Assessment of a significant Arabic corpus. In: Arabic Language Processing: Status and Prospects at ACL/EACL: Workshop, pp. 73–79. Toulouse, France (2001)
30. Goweder, A., Poesio, M., De Roeck, A., Reynolds, J.: Identifying broken plurals in unvowalised Arabic text. In: Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, Special Interest Group of the ACL (EMNLP), Barcelona, Spain, pp. 246–253 (2004)
31. Habash, N.: Introduction to Arabic Natural Language Processing. Morgan & Claypool Publishers, San Rafael (2010)
32. Habash, N., Rambow, O.: MAGEAD: a morphological analyzer and generator for the Arabic dialects. In: ACL '06: Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the ACL, pp. 681–688. Association for Computational Linguistics, Sydney, Australia (2006)
33. Habash, N., Rambow, O., Kiraz, G.: Morphological analysis and generation for Arabic dialects. In: Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages, Association for Computational Linguistics, pp. 17–24 (2005)
34. Habash, N., Eskander, R., Hawwari, A.: A morphological analyzer for Egyptian Arabic. In: Proceedings of the Twelfth Meeting of the Special Interest Group on Computational Morphology and Phonology, Montréal, Canada, pp. 1–9 (2012)
35. Habib, M.B.: An intelligent system for automated Arabic text categorization. Master's thesis, Cairo, Egypt (2008)
36. Ingham, B.: Najdi Arabic: Central Arabian. John Benjamins Pub. Co., Amsterdam/Philadelphia (1994)
37. Jiffry, F.: Saudi Arabia world's 2nd most Twitter-happy nation. The Arab News, 20 May 2013. http://arabnews.com/news/452204 (2013)
38. Kent, A., Berry, M.M., Luehrs Jr., F.U., Perry, J.W.: Machine literature searching VIII. Operational criteria for designing information retrieval systems. Am. Doc. **6**(2), 93–101 (1955)
39. Manning, C.D., Raghavan, P., Schütze, H.: Introduction to Information Retrieval. Cambridge University Press, Cambridge (2008)
40. Mustafa, M., AbdAlla, H., Suleman, H.: Current approaches in Arabic IR: a survey. In: The 11th International Conference on Asia-Pacific Digital Libraries (ICADL 2008), Bali, Indonesia (2008)
41. Prochazka Jr., T.: Saudi Arabian Dialects. Kegan Paul Int./ Routledge, London (1988)
42. Rashwan, M., Al-Badrashiny, M., Attia, M., Abdou, S., Rafea, A.: A stochastic Arabic diacritizer based on a hybrid of factorized and unfactorized textual features. IEEE Trans. Audio Speech Lang. Process. **19**(1), 166–175 (2011)
43. Sanderson, M.: Word sense disambiguation and information retrieval. In: Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 142–151, (1994)
44. Semiocast Corporation: Arabic highest growth on Twitter, English expression stabilizes below 40%. http://semiocast.com/publications/2011_11_24_Arabic_highest_growth_on_Twitter (2011)
45. Shatnawi, M., Yassein, M., Mahafza, R.: A framework for retrieving Arabic documents based on queries written in Arabic slang language. J. Inf. Sci. **38**(4), 350–365 (2012)
46. Taghva, K., Elkhoury, R., Coombs, J.: Arabic stemming without a root dictionary. In: ITCC '05: International Conference on Information Technology: Coding and Computing, pp. 152–157 (2005)
47. Versteegh, K.: The Arabic Language. Edinburgh University Press, Edinburgh (2001)
48. Wahba, K.: Arabic language use and the educated language user. In: Wahba, K., Taha, Z., Englands, L. (eds.) Handbook for Arabic Language Teaching Professionals in the 21st Century, pp. 125–138. Routledge, New York (2006)
49. Weyman, G.: Translating tweets from the Arabic spring: towards a translation workbench for twitter. http://meedan.org/2012/03/translation-twitter-middle-east-arabic/ (2012)
50. Whitaker, B.: Arabic words and the Roman alphabet. Tech. rep. www.al-bab.com/arab/language/roman1.htm (2002)
51. Xu, J., Fraser, A., Weischedel, R.M.: Empirical studies in strategies for Arabic retrieval. In: SIGIR '02: Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 269–274. Tampere, Finland (2002)