Asian Journal of Information Technology 16 (10): 754-770, 2017
ISSN: 1682-3915


# Arabic Query Expansion: A Review

Jaffar Atwan and Masnizah Mohd
Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia,
43600 Bangi Selangor, Malaysia

**Abstract:** Finding a quick and efficient Query Expansion (QE) technique for all languages has become an urgent necessity because there is no query expansion technique that fully meets user's requirement. Each technique has its own advantages and disadvantages due to the differences in the languages spoken throughout the world and the vast developments occurred on World Wide Web. Focussing on Arabic, the main problems facing QE are term conflation, stemming and lemmatization and word sense disambiguation. Thus, the purpose of this study is to shed light on some of the techniques that employ Arabic QE available in literature and to encourage researchers to study these techniques. This study focuses on and summarizes the major techniques used in query expansion for Arabic information retrieval and discusses their strengths and weaknesses. It begins with some of the linguistic characteristics of the Arabic language the features of Arabic that are related to query expansion and the possible difficulties that they might present. Researchers concludes by offering suggestions for future research which of these approaches might be suitable for dealing with the specific features of the Arabic language.

**Key words:** Arabic query expansion, information retrieval, word sense disambiguation, stemming, lemmatization, difficulties

## INTRODUCTION

The term Query Expansion (QE) denotes when a search engine adds search terms to a user's query. When a search engine receives a user's initial query, Q = [$t_1$, $t_2$, ..., $t_i$], it generates an expanded query, Q' = [$t_1$, $t_2$, …, $t_i$,$et_1$, $et_2$, … , $et_i$] (i.e., et is the expanded term) from the initial user query and provides a new expanded query in the same language. Two key works on Information Retrieval (IR) are Croft *et al.* (2009) which provides a good overview of automatic QE and Carpineto and Romano (2012) which provides a comprehensive survey of automatic QE. Researchers are interested in developing a new query expansion technique. The most important reason is the vocabulary problem due to inaccurate description of user information needs. Large amount of information online has pushed users to rely heavily on search and filtering tools to find the information they need. Search engines offer an interface through which individuals can easily find information from a collection of text such as the web. Search engines gather and index information and use several methods to find relevant documents.

Another reason to continue in improving QE is that there is no information system that fully satisfies user's requirements in terms quality and relevancy of the retrieved data. Furthermore, the use of QE techniques can improve the effectiveness of information system throughput, giving immediate results while taking user needs into consideration. Finally, QE has become a major administrative activity in different IR systems for processing different natural languages. Thus, efficient techniques that work with special rules, for example rules that handle terms semantically, e.g., part of speech, semantic similarity measures should be available to create a useful IR system to improve the retrieval of natural language texts and to remove anomalies. Bhogal *et al.* (2007) suggest that in order to increase the number of relevant documents retrieved, queries need to be disambiguated by carefully examining the queries in their proper context. Query expansion techniques range from relevance feedback mechanisms to use-of-knowledge models such as ontologies that focus on resolving ambiguities.

Recently, for almost all natural language IR systems, the performance of QE has improved (Abdelali *et al.*, 2007). We found that many QE techniques are used in online search engines for example, www.google.com Google, www.bing.com Microsoft Bing and www.yahoo.com Yahoo which are all free online search engines. These sites apply QE techniques and support more than 40 different languages. Search engine QE techniques rely on searches based on synonyms identification, term reordering, misspelt query lookup or a

**Corresponding Author:** Jaffar Atwan, Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, 43600 Bangi, Selangor, Malaysia

related search of groups and categories. From the literature, most researchers of QE techniques currently focus on information integrationwhich works by making connections between data. Other researchers focus on utilizing knowledge bases such as Google with its new search technology knowledge graph that generates search results from semantic-search of information gathered from a wide variety of sources (Singhal, 2012).

Retrieved files sometimes include irrelevant documents or inadequate information for the user's needs. Therefore, an effective QE is a vital aspect of the retrieval process. A typical, brief internet query will undergo a process of improvement to enhance its retrieval efficiency. The majority of the current QE methods experience deterioration in retrieval performance as a result of the obscure option to add query terms during the QE process (Lopis *et al*., 2002). Therefore, this research not only reviews the literature but also the indicators identified by past research of the importance of the Arabic QE approaches.

The purpose of this study is to shed light on Arabic language features to characterize the main ideas of Arabic QE to provide a classification of the various approaches used in Arabic Information Retrieval (AIR) System and to investigate several existing QE approaches within literature related to AIR in terms of the strengths and weaknesses of approaches.

## MATERIALS AND METHODS

**The Arabic language:** Arabic is an international language that is spoken in more than 20 countries and one of the major languages spoken in the United Nations. Arabic is also the language of the Holy Quran, the holy book of the Islamic world and is read and spoken by hundreds of millions of Muslims across the globe (Diab and Habash, 2007; Farghaly and Shaalan, 2009). Most of the oral spoken Arabic is more divergent than the written Arabic due to dialectal interference. When subjected to morphological analysis, Arabic words are often ambiguous (Al-Sughaiyer and Al-Kharashi, 2004). Indeed, Arabic is one of the most morphologically complex languages in the world. Arabic has 28 letters which are written from the right to left. Further, a large number of words can be generated from a single root. In addition, Arabic characters can have diacritical marks on them called Damma, Fathah, Kasra, Shaddah which determine how a word should be pronounced as shown in Table 1. Arabic scripts do not have dedicated letters to represent the short vowels in the language. They are represented by diacritics above or below the letters.

Table 1: Example showing the effects of diacritical marks on the meaning of words

| Arabic words | English meaning |
|---|---|
| علم | He taught |
| علم | Science |
| علم | Flag |

تأهل منتخبا غانا وليبيا إلى الدور ربع النهائي من كأس أمم أفريقيا للاعبين المحليين المقامة في جنوب أفريقيا، وذلك بعد فوز الأول على إثيوبيا 1-صفر وتعادل الثاني مع الكونغو 2-2 يوم الثلاثاء في الجولة الثالثة الأخيرة من منافسات المجموعة الثالثة.

Fig. 1: Example of MSA in an extract frm news article nline www.aljazeera.net

Arabic speakers use classical Arabic classical language in their daily prayers and Modern Standard Arabic (MSA) i.e., Arabic without diacritics when reading or listening to news (Fig. 1). With family (at home) or with friends they use their own specific dialects (Farghaly and Shaalan, 2009).

In Arabic, there is often no special treatment of morphological variants. Arabic is rich and complex in morphological and syntactic structures. Therefore, it is possible for the size of its vocabulary to be in the tens or hundreds of thousands or even millions. Soudi *et al.* (2007) present a review of the salient issues in Arabic computational morphology, providing a broad coverage of the computational techniques for processing the Arabic morphology. They also present a detailed discussion of the linguistic approaches on which each computational treatment is based. As mentioned before, there are different types of Arabic language used in IR research which can be classified as the:

- Classic language with diacritics as is found in the Quran
- Slang language that used in social communication and varies from one country to another
- MSA which is common in media news, newspaper and other related fields.

Table 2 shows a brief description of some research works related to these classifications. The goal of using a QE technique to deal with an input query is to find the target documents that are most relevant to the corresponding user query. An example of a user query and its corresponding expansion terms is shown in Fig. 2.

We can see that, the user query word is expanded into several query words. Any system needs some kind of

Table 2: Types and description of research related to Arabic

| Types | Description | Researchers |
|---|---|---|
| Classical | Quran and hadith | Yunus *et al.* |
| | The holy book of | Hammo (2009) |
| | Muslims and sayings of | Bassam Hammo |
| | Prophet Muhammad | Noordin and Othman, |
| | Used in education and writing | Shoaib *et al.* |
| Slang | Social forum | Al-Gaphari and Al-Yadoumi |
| | Informal language | Al-Saidat and Al-Momani |
| | Non-standard words and phrases | Shatnawi *et al.*, |
| | Differ from one place to another | |
| Modern Standard Arabic (MSA) | Media news, newspaper; internet | Khafajeh and Yousef (2013) |
| | Arabic classical language | Khafajeh *et al.* (2010) |
| | without diacritics | Darwish (2002) |
| | | Abdelali *et al.* (2004) |
| | | Abdelali *et al.* (2004) |
| | | Abouenour *et al.* (2009) |
| | | Shaalan *et al.* (2012) |
| | | Otair *et al.* (2013) |

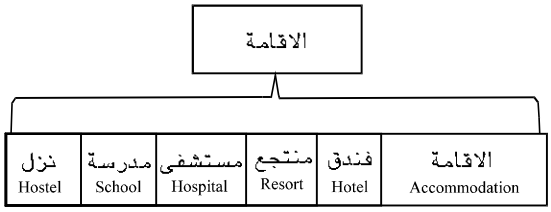


Fig. 2: Example of user query and its corresponding expended words

mechanism that can choose between the various possible options for each QE decision. The system also needs a mechanism to reorder words correctly because words with their equivalent meanings do not always appear in the same order in both the source and target queries. This reordering typically depends on the syntactic structure of the target language.

The foremost challenge for Natural Language Processing (NLP) in Arabic is overcoming ambiguity (Albared *et al.*, 2009; Kamir *et al.*, 2002). It is not uncommon for the different possible translations of a word to have very different meanings and because of its rich and complex morphology, Arabic is notorious for its morphological ambiguity (Attia, 2006).

Alqudsi *et al.* (2012) discusse the complexities of the Arabic language such as the direction of Arabic writing, the absence of capital letters, the fact that some letters change shape depending on their location within a word and many other issues. In addition, these features are the common difficulties faced in Arabic translation, Arabic classical language research and MSA research.

In light of the foregoing it is clear that developers of AIR systems need to consider some major problems. One of the main problems faced by web search engines in the Middle East and North Africa is the lack of a large number of Arabic web pages with valuable information. This is clear even within the free encyclopaedia Wikipedia which

Table 3: Pseudo-relevance feedback: worst result

| Query | Candidate term | New query |
|---|---|---|
| السمكالاستوائي | الفواكه‘الغابات | الاسماك الاستوائيةاوالغابات |
| Tropical fish | Fruit, Forest | Tropical fish or fruit or forest |

enables internet users to create and edit different articles where the Arab contribution does not exceed 1% at best (Al-Kabi *et al.*, 2012). Further, from their evaluation of Google queries based on language preferences, Al-Eroud *et al.* (2011) concluded that, if an Arabic query is submitted in Arabic and if there are many relevant popular pages in English, it is not justifiable for Google to retrieve such popular pages, even if they are in English and the query is in Arabic. Arabic users prefer to use English terms instead of Arabic ones in their queries.

Due to the extremely inflective nature of Arabic, most IR systems suffer due to their inability to address morphological complexity which is compounded by issues such as lack of space between words and pronouns and ambiguity of symbols. For example, in the case of the alif ‘ا’, many people tend to not to write the hamza, so, the abstract alif could be either ‘ا’ or ‘أ’ or ‘آ’ or ‘إ’. Further, the prefixes and suffixes could be a combination of more than one grammatical symbol as in ‘يفهمونها’ or ‘سيفهمونكها’ ‘they will understand it’ ‘to have them to understand it to you’, respectively (Abdelali *et al.*, 2004). To solve the synonymy problem, researchers have developed methods to expand the original user query by adding synonyms of the query keywords. However, sometimes there are differences between the researchers keywords and the user’s keywords which often go beyond synonyms. Consider the short query ‘المواصلاتالعامةفيتونس’ ‘Tunisia’s public transport’ an actual query (#AR53) in the TREC-2 Adhoc test collection (Egozi *et al.*, 2011; Voorhees and Harman, 1999; Graff and Walker, 2001). In this example, a relevant document may discuss announcements by the transport minister in Tunisia without mentioning any direct synonym of any of the query keywords. To handle such problems, it has been suggested that the expansion of a user query should depend on corpus-based methods. For example, Xu and Croft (2000) suggest using expansion terms that are more related to the user query from the top-ranked documents terms, i.e., terms which co-occur with query keywords. The adoption of this type of approach has led to a significant improvement but it requires manual tuning to avoid negative effects on its performance. For example, as shown in Table 3 too few expansion terms may have no impact while too many may cause a query swerve (Mitra *et al.*, 1998). This example illustrates the worst scenario of a pseudo-relevance feedback problem.

Table 4: Morphological variation of Arabic word

| Word Form | Form 1 | Form 2 | Form 3 | Form 4 |
|---|---|---|---|---|
| Arabic words | قرأ | يقرأ | قارئ | مقروء |
| English meanings | Reads | Reading | Reader | Readable |

To tackle polysemy conducted a study that was motivated by the need to enhance monolingual Arabic searches. Arabic queries compound these difficulties as Arabic is a dynamic language that constantly acquires new words from other languages. Such words are problematic as they usually do not follow normal Arabic word structure with the said word being used differently by different users. Existing search engines look for the version submitted in the user query and do not attempt to find other variants in their text collection (Nwesri, 2008).

Term conflation, stemming and lemmatization and word sense disambiguation are the main problems related to Arabic language that must be handled by a retrieval system. These problems are discussed in the following sub-sections.

**Term conflation:** Term conflation denotes a situation where one term has different forms, known as linguistic variants. These variants conceptually share an original term in text occurrences. An information retrieval system can use conflation methods to retrieve a greater number of documents related to the user query (Galvez *et al*., 2005). The variants of a term may be morphological, semantic or graphical in origin. If an IR system ignores these variations this could lead to the retrieval of documents that have terms that are not related conceptually. Fundamentally, in the case of a term extraction system, single-word phrases might be polysemy in nature while, on the other hand, multi-word phrases have a term structurewhich is inclined to be subject to modifications (Arampatzis *et al*., 1998; Galvez *et al*., 2005; Savary and Jacquemin, 2003). In general, researchers classify variations into three main types as follows:

**Morphological variation:** Linked to the internal structure of words where a term can be represented in many forms, as shown in Table 4. This is assumed to be similar for all morphologically related terms.

**Lexico-semantic variation:** Linked to the semantic proximity of words where as shown in Table 5 different terms can represent the same meaning and many meanings can be represented by the same term.

**Syntactic variation:** Linked to the structure of multi-word terms where alternative syntactic structures are reduced

Table 5: Many meanings can be reduced to one term

| Arabic words | English meanings |
|---|---|
| اضطراب الحديد في الدم | Blood iron disorder |
| مشكلة الحديد في الدم | Blood iron problem |
| نقص الحديد في الدم | Blood iron deficiency |
| فقر الدم | Anemia |

Table 6: Syntactic variation reduced to a canonical syntactic structure

| Arabic words | English meanings |
|---|---|
| تبحث بعض المعلومات | Searching for certain information |
| تبحث هذه المعلومات | Searching for this information |
| بحث المعلومات | Search information |

Table 7: Many words are derived from the same root Ktb كتب

| Arabic words | English meanings | Arabic root(stem) | English root(stem) |
|---|---|---|---|
| مكتب | Office | كتب | Ktb |
| كاتب | Writer | كتب | Ktb |
| مكتبة | Library | كتب | Ktb |
| مكتوب | Written | كتب | Ktb |

to a canonical syntactic structure. Statements that are syntactically distinct but semantically equivalent such as those in Table 6 are conflated into a single syntactic structure, 'search information'. As stated earlier, the goal of using term conflation is to increase the effectiveness of an IR system. Term conflation aims to process information in order to decrease the variation in word forms in queries and documents by finding the terms that are conceptually related but do not match morphologically. In general, Morris categorizes term conflation into two approaches: feature reduction and statistical conflation. Typical feature reduction methods include approaches such as stemming and lemmatization while typical statistical conflation approaches include string similarity scoring approaches such as gram characterization, edit distance and Bayesian Models. The latest IR systems use a combination of the two types of approaches. Over the past decade, research in this area has focused on two main approaches: stemming and statistical conflation. The following subsection discusses the importance of stemming and lemmatizationwhich is of particular relevance to highly morphological languages such as Arabic (Croft *et al*., 2009)

**Stemming:** The term stemming refers to a conflation approach that attempts to find a common stem for a group of words that appear in a text as illustrated in Table 7. With this approach, one stem for a group of words that are relevant in term of form related can be found without the need to have a correct morphological root. Stemming approaches can be grouped into three general groups: lookup-based, rule-based and probabilistic (Ahmed and Nurnberger, 2009). Lookup-based approaches search for a stem of the word in the text in a lookup table that contains a list of words and their stems. If the search is

Table 8: Suffixes and prefixes that are removed by light10 (Arabic light stemmer)

| Prefixes | Suffixes |
|---|---|
| ال, وال, بال, كال, فل, و | ها, ان, ات, ون, ين, يه, ية, ه, ة, ي |

successful, the specific stem is returned. Even though, this approach yields highly accurate results, it has several obvious drawbacks including the need for linguistic expertise, a labour-intensive list formation process and the complexity of the system.

A rule-based stemmer is a 'light stemmer which employs a set of rules that is applied indiscriminately to remove suffixes or prefixes. However, this type of stemmer does not address the stemming of broken plurals at all. In contrast, morphological analysers use lexicons and morphological rules to remove the proper affixes prefix, infix and suffix as shown in Table 8. The analysers analyse all possible combinations of initial and final letters of a word and use the rules to validate the combination between these letters and the remaining stem for any given word. While such systems produce more accurate stems, they commonly return more than one possible stem for the same word, making it very difficult to determine the best stem that represents the word. Such systems are also less efficient than light stemmers, despite sometimes returning results that are very similar to those of the light stemmers (Larkey *et al*., 2007).

Stemmers are basic elements in query systems, indexing, Web search engines and IR systems. Stemming can be viewed as a recall-enhancing device or a precision-enhancing device. In the field of text mining, stemming is used to group semantically related words to reduce the size of the dictionary feature reduction. With Arabic stemming, words are reduced to their roots. Root-based indexing is aggressive in the sense that, it reduces words to their three-letter roots. This affects the semantics as several words with different meanings might have the same root.

Prior to the Text Retrieval Conference (TREC), the stemming of Arabic documents was not only performed manually but it was only applied on small corpora. Later, many researchers including both native and non-native Arabic speakers, created a considerable amount of Arabic stemming algorithms. Despite stemming errors, it has been empirically demonstrated that stemming improves retrieval in many languages including Arabic (Aljlayl and Frieder, 2002; Larkey and Connell, 2002). It is noteworthy that Arabic differs from the other Indo-European languages in terms of its syntax, morphology and semantics. Since, the morphological nature of Arabic is complex, there is a wide body of research in this area that particularly focuses on the impact of Arabic morphology on AIR. Based on the required level of analysis, Arabic stemmers are categorized as either root-based (Salton and Buckley, 1997) or stem-based (Baeza-Yates and Ribeiro-Neto, 1999; Lopis *et al*., 2002; Manning *et al*., 2008). In Arabic, the root is the original form of the word before any transformation process (Farghaly and Shaalan, 2009). However, a stem is a morpheme or a set of concatenated morphemes that can accept an affix (Alqudsi *et al*., 2012).

A superior root-based Arabic stemmer is Khoja's stemmer presented by Khoja (2001). The Khoja algorithm removes suffixes, infixes and prefixes and uses pattern matching to extract the roots. However, the algorithm suffers when dealing with names and nouns. There are several proposed Arabic stem-based light algorithms (Aljlayl and Frieder, 2002; Nwesri, 2008; Al-Ameed *et al*., 2006; Larkey *et al*., 2007; Larkey and Connell, 2002). The most widely used Arabic light stemmer is the light10 developed by Larkey *et al*. (2007) and Larkey and Connell (2002). Light stemming does not deal with patterns or infixes; it is simply a process of stripping off prefixes and/or suffixes. Unfortunately, an unguided removal of a fixed set of prefixes and suffixes causes many stemming errors especially where it is hard to distinguish between an extra letter and a root letter.

Although, light stemmers produce fewer errors than aggressive root-based stemmers, aggressive stemmers reduce the size of the corpus significantly. Both Arabic root-based and stem-based algorithms suffer from stemming errors. The main cause of this problem is the stemmer's lack of knowledge of the word's lexical category, e.g., noun, verb and preposition. Paice (1994, 1996) show that, light stemming reduces over-stemming errors but increases under-stemming errors. On the other hand, heavy stemmers reduce under-stemming errors while increasing over-stemming errors. Since, Arabic has more than 10,000 independent roots (Al-Fedaghi and Al-Anzi, 1989) it is timely expensive and not sufficient to use a dictionary to recover wrongly stemmed words.

The N-gram stemming technique is ineffective for Arabic text processing (Duwairi 2006; El-Kourdi *et al*., 2004). However, Khoja's root-extraction stemmer (Khoja, 2001) and Larkey's light stemmer (Larkey *et al*., 2007; Larkey and Connell, 2002) are the two most effective Arabic stemmers. Furthermore, Larkey *et al*. (2007) proposes multiple light stemmers that depend on heuristics which-along with statistical stemmers would be able to handle all instances of co-occurrence in Arabic text retrieval. However, for cross-language retrieval, light stemmers are more effective than morphological stemmerswhich deal with the root of each word. Additionally, Darwish (2002) investigates the impact of enhanced morphological analysis, particularly in terms of context-sensitive morphology on monolingual AIR. A

comparative analysis of context-sensitive morphology and non-context-sensitive morphology found that the former is more effective in AIR than the latter; moreover, Taghva *et al.* (2005) state that employing a root-extraction stemmer for the Arabic language yields the same result as the Khoja stemmer (without using root dictionary). Further, the root-extraction stemmer is equal to light stemmers in monolingual document retrieval tasks. Nevertheless, Khoja, (2001)'s method which initially removes the prefixes and suffixes of terms, later checks the root dictionary list. If a term is found in the list, it returns the root if not, it returns the original word without modifying it.

It is noteworthy that Arabic morphology in IR is aimed at finding words with identical or relevant meanings. Further, it has been identified that by means of indexing Arabic text, the efficiency of retrieving words or stems can be substantially increased by using roots (Abu-Salem *et al*., 1999).

According to Larkey *et al*. (2007) when dealing with Arabic, light stemming basically does not deal with patterns or infixes, rather it simply strips off prefixes and/or suffixes. However, even though light stemming is capable of accurately conflating several variations of words into huge classes of stems (Atwan *et al*., 2013) it is still not capable of conflating other forms. For instance, the broken plurals of nouns and adjectives do not get confounded with their singular forms and past-tense verbs will not conflate with their present tense due to the fact that they preserve some affixes and internal dissimilarities. Nevertheless, despite its ease-of-use and disadvantages, none of the other advanced approaches have been considered more efficient for IR. According to Aljlayl and Frieder (2002) it has been proven that stem-based IR is more effective than root-based IR.

Nwesri (2008) introduces new stemming techniques that minimize stemming errors in light stemmingwhich improved retrieval results in some cases. He uses a modified version of the light 10 stemmer to develop three new versions, the: light 11-13. The modified version of the light 10 uses the grammatical and morphological rules of Arabic words to validate affixes. All the three versions perform slightly better than the light 10 stemmer with the light 13 improving recall significantly when using relevance feedback over the TREC and Arabic GigaWord collections. He extends word normalization for improved retrieval effectiveness and also shows that automatic generation of stopword variants leads to a reduction in precision and recall. Overall, supporting light stemming with morphological rules aids the retrieval effectiveness.

As justification, the major challenge faced by researchers in root-based IR is the surface variants of words which do not convey similar meanings or interpretations. Despite the fact that each variant has its own meaning, it originates in the same root. Therefore, when using root-based IR, it is highly possible for researchers or document analysts to encounter increasing ambiguities and confusion with words. Thus, it is necessary to enhance the AIR System by integrating disambiguated word meanings.

These stemming limitations can cause problems in applications that have strict word matching requirements. The goal of any new stemmer is to address stemming accuracy by avoiding over-stemming, under-stemming and mis-stemming without adding too much complexity to the stemming algorithm and without using any type of morphological analysis.

Any new stemmer should tackle two main issues. First, it needs to consider the list of affixes that should be removed when applying its removing rule and second, it needs to perform a morphological analysis of the stemmed word. The definition of the stemmer is closely related to the definition of a lemma in linguistics or the dictionary form of a word (Manning *et al*., 2008). Traditionally, the process of reducing a word to its lemma is called lemmatization. However, lemmatizers rely more heavily on the linguistic features of a given text (Manning *et al*., 2008; Al-Shammari, 2013; El-Beltagy and Rafea, 2011). In this sense, the stemmer should be more accurate with less computational cost.

**Lemmatization:** Lemmatization is an advanced stemming process that involves the use of vocabulary and morphological analysis to reduce inflected or sometimes derived words to their stem, base or root, generally from a written word form. Recently, (Al-Shammari, 2013) a new Arabic lemmatizer has been developed that has a high range of accuracy. It uses syntactical knowledge to make stemming decisions. Al-Shammari (2013) proposes an Arabic advance stemmer called the Educated Text Stemmer (ETS) which uses the (Khoja, 2001; Larkey *et al*., 2007) root-based stemmers. The Khoja method uses a root-base stemmer that removes suffixes, infixes and prefixes as illustrated in Table 7 but also uses pattern matching to extract the roots whereas, Larkey's light 10, a well-known Arabic light stemmer that does not deal with patterns or infixes, simply strips off prefixes and/or suffixes. The ETS lemmatizer tackles the Arabic word lexicallywhich is a drawback compared to the other types of Arabic stemmers (Atwan and Mohd, 2012). It uses a new but long list of affixes and a short list of stopwords to distinguish nouns and verbs. ETS consist of two main algorithms: An initial algorithmwhich stems the word according to its previous stop-word and a second algorithmwhich stems the word by removing some

affixes and then uses pattern matching to compare the result word to a group of similar words in terms of a common threshold. In general, the ETS is computationally expensive its initial algorithm fails to improve IR due to the long list of affixes and the lack of stopwords lists that can distinguish between verbs and nouns.

El-Beltagy and Rafea (2011) introduce work similar to that of Al-Shammari and Lin (2008a, b) which is based on light 10 (Larkey *et al*., 2007) to handle the problem of the broken pluralswhich other stemmers are unable to do effectively. The researchers propose a set of rules for detecting broken plural patterns and transforming these plurals into their singular forms. Their method also uses a corpus to find out whether the word resulting from the proposed transformation exists in the corpus. Moreover, the same process is also applied to remove certain prefixes and suffixes. At the end of the process, if the resulting word appears in the corpus, then this word is considered to be the stem. Their stemmer is applied to a corpus consisting of about one million tokens and is able to achieve high accuracy. However, it has not been tested on an AIR System.

The lemmatizers proposed by Al-Shammari and Lin (2008a, b) and El-Beltagy and Rafea (2011) have neither been tested against standard benchmarks nor have they been compared with each other. Furthermore, they use different rules for normalization and different stopwords lists. However, the results of their work indicate that, as stated earlier, a stemmer should not only tackle the word morphologically it should also be careful with the set of affixes to be removed from that word.

The growth of Arabic information on the Web presents many challenges to researchers in the field of QE as they have to find methods to deal with issues such as short vowels, absence of capital letters and complex morphology (Abouenour *et al*., 2010). In addition, stemmer algorithms are not designed to take into account the semantics of the stemmed wordwhich is very important for languages such as Arabic (Croft *et al*., 2009). Another big problem that needs to be addressed in developing an AIR system is the ongoing growth in the number of foreign words within the Arabic text. Current AIR systems are not able to handle the problem of retrieving different versions of the same Foreign word (Abdelali *et al*., 2004). In this context, Foreign words are words that are borrowed from other languages and transliterated into Arabic as they are pronounced differently by Arabic speakers with some segmental and vowel changes. The application of stemming is not helpful for such words as they have no clear affixes. In fact, stemming would be detrimental to accuracy because the core letters that match Arabic affixes would be removed, resulting in the word being mapped to another index term (Nwesri, 2008). Therefore, any new stemmer should be developed to take into consideration not only the Arabic words but also any incorporated Foreign words. The next subsection defines the word sense disambiguation problem and sheds light on its importance in the context of Arabic QE.

Table 9: Synonymous Arabic phrases

| Arabic phrase | English meanings |
|---|---|
| كاتب المقالة | Writer of essay |
| مؤلف المقالة | Author of essay |
| محرر المقالة | Clerk of essay |

**Word sense disambiguation:** Basically, the quality of user search outcome is enhanced by the capability of a search engine's QE. Normally, users do not often create search queries using ideal terms. However, it is essential to use ideal terms because the database might not contain the terms entered by the user (Pinto and Perez-Sanjulian, 2008).

Word Sense Disambiguation (WSD) is the task of selecting the correct sense for a word (Jurafsky and Martin, 2008). It is considered as "AI-complete problem, that is, a task whose solution is at least as difficult as the most difficult problems in artificial intelligence" (Navigli, 2009). Unfortunately, the identification of the specific meaning that a word assumes in context is not that simple. An example is shown in Table 9. While most of the time humans do not, even think about the ambiguities of language, machines need to process unstructured textual information and transform it into data structures which must be analysed in order to determine the underlying meaning.

The rule of ambiguity in IR systems causes poor performance. The inherent difficulty of WSD is also attested to by the lack of applications that can be used for real-world tasks. The exponential growth of the Internet community, together with the fast-paced development of information technology has led to the production of a vast amount of unstructured data such as document warehouses, web pages, collections of scientific articles and blog corpora. As a result, there is an increasing urge to treat this mass of information by using automatic methods but researchers continue to struggle to develop high-quality, error-free methods.

WSD is typically configured as an intermediate task, either as a standalone module or as a properly integrated part of an application thus, performing disambiguation implicitly. However, the success of WSD in real-world applications is still to be determined. The application-oriented evaluation of WSD remains an open research area, even though various works and proposals have been published on the topic.

The results of recent comparative evaluations of WSD systems which mostly concern standalone WSD, show that most disambiguation methods, among other issues, have inherent limitations in terms of performance and generalization capability when fine-grained sense distinctions are employed. Conversely, the increasing availability of wide-coverage, rich lexical knowledge resources as well as the construction of large-scale coarse-grained sense inventories, could pave the way for new disambiguation approaches, especially semantically enabling applications in the area of human-language technology (Navigli, 2009).

## RESULTS

**Arabic query expansion approaches**: There are many different approaches in carrying out Arabic QE. This section provides a brief explanation of the main approaches that have been used in previous research. Over the years, many techniques have been used to enhance the performance of Arabic QE. We categorize QE approaches into four types:

- Feedbackwhich uses terms from documents retrieved from the initial query
- Data extracted from a corpuswhich use expansion terms from a collection
- Using external resources
- Hybrid approachwhich uses more than one technique together

Table 10 shows QE categorization and some related works. In the following subsections, these four types of approaches are further classified according to how expansion terms are used. This section concludes with a discussion of the more recent hybrid approaches.

**Query expansion using feedback:** In the field of IR, the use of data from relevant or non-relevant documents was the first approach used by researchers. This approach can be accomplished automatically pseudo-relevance feedback or through user interaction relevance feedback. This category of QE approach using feedback has evolved over time through the addition of features that handle the user query language characteristics morphology, syntax and semantics resulting in several new approacheswhich are discussed in the following subsections.

**Relevance feedback:** Relevance feedback is provided by humans according to their linguistic knowledge and information needs. The strength of this type of approach is that, it can deeply analyse both syntax and semantics. A key component of any relevance feedback QE system is its lexical resources. In practice, a relevance feedback QE system tries to overcome natural language problems like morphological variations, polysemy and synonymy by eliciting user feedback on the relevance of the ranked documents obtained in response to the initial query and then uses this feedback to refine the query automatically (Chinnakotla *et al*., 2010). The system assumes that the top *n* number of documents is the most relevant to the query and then takes the terms from these documents to reweight the query within the weighting algorithm. The modified query is then used to retrieve a new set of documents for presentation to the user (Attar and Fraenkel, 1977).

By Salton and Buckley (1997), the relevance feedback approach primarily involves selecting significant terms or expressionswhich are connected to specific documents that were previously retrieved which were recognized as appropriate by the users and then improving the significance of these terms in a fresh formulation of the query.

Abusalah *et al.* (2009) use relevance feedback from Arabic native speakers to evaluate a Cross-Language Information Retrieval (CLIR) System which uses ontology consisting of 200 Arabic concepts in the travel domain. This ontology-based approach to improve query translation significantly outperformed the Machine Readable Dictionary (MRD) ranslation baseline using mean average precision as a metric in a user-centred experiment. Abouenour *et al.* (2009) use user feedback and text segmentation to extract knowledge from a database of Prophetic traditions or 'Hadiths'; the database contains texts of 340 Hadiths. The overall result improved in terms of precision and recall. Kanaan *et al.*

Table 10: QE categorization and some related research

| Types | Description | Researchers/Years |
|---|---|---|
| QE using feedback | Relevance feedback | Larkey and Cornell (2002) |
| | Pseudo-relevance Feedback | Robertson and Gao Darwish (2002) |
| QE using data extracted from a corpus | Expansion based on stem/root | Hammo *et al.* (2002) Rachidi *et al.* (2003) El-Emary and Atwan (2005) |
| | Expansion based on co-occurrence | Bassam Hammo Khafajeh and Yousef (2013) |
| | Expansion based on attribute | Khafajeh *et al.* (2010) Bellare *et al.* (2007) |
| QE using external resources | Arabic wordnet | Black and Elkateb Abouenour *et al.* (2009) Abouenour *et al.* (2010) Al-Ameed *et al.* (2006) |
| | Ontology | Abd-El-Jaber and Sembok, Abusalah *et al.* (2009) Zaidi *et al.* (2005) |
| Hybrid QE technique | | Otair *et al.* (2013) Menai and Alsaeedan (2012) Liu (2006) |


Fig. 3: Process of PRF AQE

(2007) compare the performance of Interactive and Automatic QE IQE, AQE on a collection of 242 documents by using Term Frequency Inverse Document Frequency (TFIDF) weighting and average precision in their evaluation. Using IQE and AQE techniques, the top 15 expansion terms were added from the top 10 documents to the original query. The best results in terms of average precision were obtained by IQE because users can truly identify good expansion terms. Yet, AQE gives good results when compared with the baseline system, without expansion. However, the weakness of the relevance feedback approach is that it is time-consuming in terms of the time required to wait for user feedback because the users need to have as well as to apply a high level of linguistic knowledge.

**Pseudo-relevance feedback:** This QE feedback technique which is also known as Pseudo-Relevance Feedback (PRF) or blind feedback is done automatically. Essentially, the same work is done using relevance feedback but without user interaction. The PRF system assumes that the top n number of documents is the most relevant to the query and then takes terms from these documents to reweight the query within the weighting algorithm.

Pseudo-relevance feedback is used by most IR systems but usually, the source language text is not structurally analysed beyond the syntactic or semantic levels as the expansion is based on the top n ranked documents. The process of PRF consists of the following steps as illustrated in Fig. 3:

- Select topped ranked document
- Choose top weighted terms, i.e., no syntactic or semantic analysis
- Expand the user query by adding new terms
- After the terms are added, apply simple reweighting equation
- Generate new expanded Query Q'

Carpineto and Romano (2012) and Carpineto *et al.* (2002) not only used but also proved that the AQE approach can effectively improve IR. Indeed, its effectiveness is attested to by the fact that versions of the AQE approach can be found in most IR systems. The AQE approach has been of interest for some time but it is only recently that it has reached a level of scientific and experimental maturity, especially in laboratory settings such as TREC (Carpineto and Romano, 2012).

However, AQE does have one limitation, it lacks analysis of the source languagewhich may cause several problems as words are retrieved without disambiguation of their syntactic or semantic role.

**Query expansion using data extracted from a corpus:** Expansion based on data extracted or captured from a collection varies, based on the stem or root of terms that is shared between the original query words and the collection. Furthermore, expansion use terms in the case of term-based co-occurrence synonym or concept relationship and sometimes by extracting attributes surrounding the word of the query in the documents within the collection.

**Expansion based on stem or root:** There are many Arabic words that share one stem or root. One of the QE approaches that can solve this ambiguity does so by adding new words to the original query that share the same stem or root. El-Emary and Atwan (2005) compare to full-word indexing with root indexing using a traditional modelling technique. They employed a Vector Space Model (VSM) with a corpus of 242 Arabic documents using cosine similarity. They found that root indexing enhances the average accuracy when compared with full-word indexing.

Hammo *et al.* (2002) developed a system that uses techniques from IR stemming and NLP part of speech tagging by extending the initial query with terms that have a similar stem or root to the initial query terms. In this way, the query is extended to involve all the terms verbs and nouns, extracted from verbs that appear in the index file and comprise the same roots that were generated from the initial query terms. This approach is used in a question-answering system that provides short answers to questions expressed in the Arabic language. The result showed that, recall is increased but this is sometimes at the expense of precision. In another research, Harrag *et al.* (2009) expanded the query with terms that share the same stem for retrieving the verses of the Quran. The expanded query is efficient and retrieves more verses than the original query.

Rachidi *et al.* (2003) employ word root extraction and thesauri, i.e., group of words shared similar meaning constructed from automatic Arabic document classifications from a database consisting of 1000 documents in their QE method. They state that their QE approach shows significant improvement in recall. However, it is noteworthy that the overall success of their

system is limited to the amount of available tools developed for the Arabic language. As mentioned before, Arabic stemmers and lemmatizers still suffer from many problems. Using this type of QE approach requires a good stemmer to get the best stem or root for multiple words but such a stemmer is still unavailable.

**Expansion based on co-occurrence:** In this type of research, a general thesaurus is built that uses the relationship between synonyms or concepts. A thesaurus is a reference work that details words that are arranged together based on the likeness of meaning including synonyms and occasionally antonyms as against a dictionary that includes descriptions and diction. Furthermore, its aim is to guide both an indexer and a searcher to choose identical terms or a combination of preferred terms to signify a specific topic (Dextre and Zeng, 2012).

In the research by Hammo (2009) queries are automatically enhanced with relevant terms which are generated from a vowelized index by utilizing a stemmer and a thesaurus of semantic synonym classes. The methods presented in these works are applied to a collection of verses from the Quran. The researchers state that the QE for searching Arabic text in this way is encouraging and that the effectiveness could probably be enhanced further. Khafajeh and Yousef (2013) uses a similarity thesaurus for QE with full words. Other methods retrieve search results by using local context analysis which breaks documents into passages based on an original query and retrieves the top-ranked concepts from the top-ranked passages. In conducting a full-word search of 242 documents, they claim to have successfully utilized QE approaches for the retrieval process and improved IR performance through the use of stemming. As a result, these techniques improve the average precision of a search.

Khafajeh *et al.* (2010) used AQE to develop an AIR System with a corpus-based thesauruswhich uses similarity term-term similarity and association fuzzy set theory thesauri featuring full and stemmed words from over 242 Arabic documents for QE. The best results are achieved by using stemmed words with an association the saurus as opposed to using stemmed words with a similarity thesaurus. Both of these methods improve the precision and recall of AIR systems.

Shaalan *et al.* (2012) expand queries, based on similarity of terms to improve AIR. They suggest a method for QE in AIR that employs the co-occurrence algorithm to select relevant terms in order to expand the query and remove the non-related terms. The algorithm was tested on an INFILE test collection of CLEF 2009 and the experiments showed that QE that considers the similarity of terms improves precision and retrieves more relevant documents. Furthermore, by using this method, recall can be increased while maintaining precision at the same level.

Expanding terms during query or document indexing time from a thesaurus in any IR system appear to result in improved performance. However, AIR suffers from a lack of tools that fully support IR. In addition, AIR still needs to develop tools that support IR by lexicon or thesaurus that structure Arabic language with lexical-semantic relations. Hence, researchers that are developing QE for AIR using the co-occurrence approach need to concentrate on finding the best way to build a thesaurus or finding the best terms co-occurrence relationship.

**Expansion based on attribute:** Bellare *et al.* (2007) expands the query based on the extracted attributes from the corpus of the entity using the Gale corpus a mix of English, Chinese and Arabic newswire, blogs and broadcast news which improves document retrieval performance. However, there is an inadequate number of instances for other entities in the corpus from which attributes were extracted where some other queries have not improved.

**Query expansion using external resources:** The use of external resources is a new QE technique that currently is unable to handle language morphology. However, tools such as WordNet and some specific, manually built domain ontologies have helped improve IR performance using this approach.

**Arabic WordNet:** Arabic WordNet (AWN), described by Elkateb *et al.* (2006) is a lexical resource that depends on the design and topics of the globally acknowledged Princeton WordNet (PWN). Furthermore, it might be directly mapped with PWN 2.0 and EuroWordNet (EWN) which will enable translation on the lexical level into English and a multitude of other languages.

Arabic WordNet is a linguistic resource which constitutes a profound conventional semantic basis. Apart from the standard WordNet manifestation of senses (Fig. 1), the meaning of words is described with semantics which could be understood by machines in first order logic. The Suggested Upper Merged Ontology (SUMO) is the foundation for these semantics which is affiliated with domain ontologies (Elkateb *et al.*, 2006). Ontology as a concept and some related works are described in the next subsection. A number of studies use WordNet to identify the word senses in query terms.

Table 11: Example Arabic word and its synonyms

| Queries | Words | Synonyms |
|---|---|---|
| Arabic words | ملعب | استاد ,ميدان , مدرج ,منتجع ,ساحة |
| English meanings | Playground | Stadium, field, Runway, park, square, play, recreation ground |

When the perception of a query term is identified, its synonyms, hyponyms, words from its meaning and its compound words are considered for probable inclusion in the query. Table 8 shows some synonymous Arabic phrases and Table 11 shows an example of an Arabic word and its synonyms.

Al-Ameed *et al.* (2006) created a WordNet prototype that uses a word-sense based on synonym search approach with user feedback to select the senses that match the query for AIR systems. Their prototype consists of 4,000 words and their overall results indicate that word sense when it is based on synonyms will improve AIR system performance. However, they only tested a number of related senses in terms of delivered outcomes without testing real IR performance in terms of precision or recall. By Abouenour *et al.* (2008, 2009, 2010) and Abouenour (2011) AWN is used in a question-answering retrieval system to semantically expand a query based on a synonym and its relationships synonym, definition, supertypes and subtypes. Researchers successfully tested word-based and QE evaluations over a set of CLEF and TREC questions. The best performance achieved by QE was in retrieving answers. However, this research has some limitations due to low coverage of the considered questions in AWN because the questions which were translated from other languages do not have an answer key.

**Ontology:** Ontology includes a range of forms, components and connection types. Furthermore, it basically signifies knowledge as a group of aspects of a sector and the associations among those aspects. It could be employed to explain the components within a particular field and could possibly be applied to explain that field (Petrov, 2011; Zaidi *et al*., 2005) has expanded the original term by adding synonyms (relative derivatives as well as generic or specific concepts) in legal domain. Using their own manually built ontology, the researchers utilized the first 50 top ranked documents to determine the relevance of their proposed algorithm based on the frequency of the initial key wordswhich will have priority compared to the words released by the ontology. The results revealed a significant improvement in both recall and precision using ontology in QE for Arabic. Furthermore, (Abusalah *et al.* (2009) developed a cross-language retrieval system by manually developing a bilingual Arabic, English ontology in the travel domain that consists of 200 concepts. Many studies regard ontology as the generic selection of informationwhich might be employed to search engines by extending the query. They compared their ontological QEwhich uses a corpus consisting of 8,000 documents collected from AL-Nahar www.annahar.com newspaper from 1996-1999 and documents collected from the Palestinian Ministry of Tourism http://www.mota.ps with the results from the MRD and found that their method outperformed MRD in terms of precision.

By Abouenour *et al.* (2008), the researchers use semantic QE to answer a set of 82 questions from the CLEF that were translated into Arabic and which were classified into different question-answering system domains by looking for the correct answer in the first five search results returned by Google. Then, they expanded each question semantically, based on their built ontology, the Amine AWN (AAWN) ontology and its semantic relations (synonym, definition, supertypes and subtypes) and ran it through Google again. This program uses the existing mappings between English synsets WordNet and SUMO concepts and adds Arabic synonyms to those types based on the equivalence relation between English synsets and AWN synsets. Semantic QE using AAWN was found to improve the accuracy of the returned expected answers.

In light of the foregoing, the use of ontologies in IR looks promising. They seem to improve performance in term of accuracy by disambiguating and reformulating the user query. However, the successful use of ontologies in QE depends on many factors; the quality of the ontology must be accurate, stable, comprehensive and up to date. Also, a user needs to be familiar with such ontologies if a user can navigate them with ease, this increases their effectiveness (Bhogal *et al.*, 2007).

There is a lack of resources and tools that support AIR and only a little effort has been made in this area in comparison with work on QE in English. Belkredim and Meziane (2008), Belkredim *et al.* (2009), Moawad *et al.* (2010), Beseiso *et al.* (2010), Hoseini (2011) and Jarrar (2011) have used ontology in Arabic text search. Al-Rajebah and Al-Khalifa (2014) and Al-Rajebah *et al.* (2010) proposed ontology built and its semantic relations extracted from Arabic Wikipedia www.wikipedia.com for the Arabic language using semantic field linguistic theory. The evaluation of their method showed that, it extracts concepts and their semantic relations with a degree of 65% in term of average precision but their work not tested in the field of AIR.

Most of the work done in this area uses a thesaurus based on a corpus, manually built specific domain ontologies or WordNet and Wikipedia as an

ontological resource all of which have the potential to improve AIR. Despite the dearth of these tools and the promise of success from their application in AIR all these tools still not only have limitations but also lack some linguistic aspects for Arabic. Also, there is no standard Arabic ontology available. Therefore, further research needs to be done to build a standard Arabic ontology and adapt it to the field of AIR.

**Hybrid query expansion technique:** Basically, hybrid techniques combine two or more techniques together or to construct a technique that is based on the outcomes of the new techniques or tools used in the previous phase. Also, as one of the techniques used to overcome this apparent weakness of the AQE approach, a hybrid approach has been developed.

Liu (2006) introduce a retrieval model that expands the user query in two levels. The first level represents the query in terms and phrases and expands the initial user query by finding its senses using WordNet and Wikipedia. The second level uses the top-ranked documents from first process for local PRF. New terms and phrases from the feedback form the new query and are submitted as the final result. The model was tested using different datasets and the overall result showed that it improves precision. Menai and Alsaeedan (2012) use a genetic algorithm to find the most appropriate word from senses retrieved from AWN which was also designed to improve precision. However, the dataset on which they tested their method was too small to draw definite conclusions about the method's efficacy.

More recently, Otair *et al.* (2013) combined a thesaurus with feedback techniques using a corpus consisting of 242 documents and 60 queries. The best result in terms of recall and precision was achieved by combining interactive QE. Also, the next best result was obtained by AQE using standard Rocchio's (Joachims, 1996) to reweight the query terms and build the new query, i.e., the expanded query.

## DISCUSSION

In light of the foregoing, it is clear that QE can be and in fact has been addressed using a variety of different approaches. Here, the focus has been on the application of these approaches to AIR problems, taking special note of the approaches that might be suitable for dealing with the specific features of the Arabic language. Although, there are many common elements in the systems that have been proposed thus far, there is also a growing diversity within the systems. The automatic expansion of the semantic knowledge base approach has made rapid progress and there is great optimism about its potential for future success. This automatic expansion approach promises a high level of improvement in performance. However and it is more suited to certain types of Arabic IR challenges due to the advantages associated with the automatic approach, i.e., it is applicable and attainable without user intervention.

An automatically derived knowledge base expansion approach requires less time and effort to develop than an interactively derived knowledge base expansion approach. This is why most new systems apply the automatic approach. One noticeable trait identified in this review is the reality that most approaches that have so far been proposed for AIR may have only been tested on limited domains and not on standard corpora. This situation is understandable from a practical perspective. However, from a research point of view, this lack of wider testing often makes it very difficult to assess a system's capability or to compare different systems. Indeed, there seems to be some confusion in the field about how different automatic expansion approaches should be formally tested and evaluated. Different performance measures will often be chosen and a different number of trials will be performed in the analysis and testing of these systems.

Improving the techniques without addressing the issue of lack of resources, e.g., insufficiency in standard corpuses will not lead to greater advancement. At the moment, the TREC corpus is the most popular in terms of Arabic text IR and is adequate enough to be considered for system evaluation. However, researchers face problems with stopwords lists due to many researchers have created their own lists. Normalization is another problem that affects AIR because there are no standard steps for researchers to follow, therefore, researchers create and follow their own normalization steps. As discussed above, the stemming process is very important for highly morphological languages such as Arabic because it has an impact on AIR in terms of removing inappropriate affixes. Here too, there is no standardization as researchers use different set of affixes with different lengths.

As we have seen in this review, different approaches and different challenges lead researchers to seek to achieve different sets of objectives, making it difficult to perform comparisons in many cases. However, difficulties in comparisons are also real in many of the studies. Most researcher's evaluations focus on the expansion of their own collected texts. There is very little work on the expansion of standard corpora, particularly those that describe much of the information found on the internet where expansion is in demand. Usually,

researchers test their expansion techniques by using recall, precision, average recall, average precision, mean average precision and the f-measure which are evaluation measures commonly used in the research and development of IR technology.

## CONCLUSION

Arabic has some commonalities with other languages such as English. Nevertheless, it also has its own unique characteristics with regards to its heritage, declassing characteristics, internal construction, strong association with Islam and the Arabic traditions and individuality. Hence, an Arabic NLP system that fails to take into account these traits is highly unlikely to succeed. The challenges faced by Arabic language researchers are not restricted to the ethnic factors of the language. They also extend to its normal linguistic formwhich is explained in brief below.

Recently, research on AIR has gained momentum and is being expanded in many fields. Most of the researchers working on AIR are well aware that it is essential to use different techniques to retrieve the most highly related data on behalf of the user. However, they are hindered in their efforts not only by a lack of techniques that can be applied in AIR but also by a lack of resources, especially those which handle Arabic morphological structures though this has been somewhat offset by WordNet, a new open-source program that handles Arabic semantics. It is therefore necessary to develop more techniques that can improve the performance of AIR and retrieve more data related to user queries. Furthermore, Automated Intelligent WSD for AIR is a crucial but a difficult task to complete due to the complexity of the Arabic lexical structure. Ultimately, the employment of QE should be able to semantically improve the performance of AIR systems and help the user to locate the required information.

## RECOMMENDATIONS

From this review, it is clear that, there is a lack of techniques to verify or present exceptional solutions to ensure that, standardized AIR systems can be developed. Also, from this review it seems that, the most encouraging course for research in this field is to pursue methods that employ hybrid approaches, i.e., methods that use approaches that incorporate automatic expansion of a semantic knowledge base.

This study has reviewed a number of QE techniques to expand Arabic text queries and discussed the main Arabic language features and problems related to IR that present challenges for researchers. However, since, the variants of the techniques differ from each other here, we offer some suggestions for future research. We suggest that, there is a need to find standard stopwords and standard normalization processes for MSA and that the pre-processing in AIR systems should be enhanced.

Most IR systems focus on the expansion of news and official texts while not many focus on open domain retrieval. AIR focuses mainly on informal genres which take much of their information from the internet for which expansion is in great demand. Knowledge base expansion has grown quickly and AQE will surely follow. Nonetheless, these IR systems still do not meet human requirements.

In the future, we plan to develop a new AIR system that takes the reordering of language challenges into consideration. We will attempt to develop a hybrid technique by combining the semantic relationship between the query words and the corpus to retrieve documents that are semantically related to the user query which will adequately address the WSD problem. It is envisaged that the technique will be able to retrieve documents that are relevant to the user query and improve the performance of AIR systems.

## ACKNOWLEDGEMENT

This research project was supported by the Malaysian Government under Fundamental Research Grant Scheme (FRGS/2/2013/ICT02/UKM/02/1).

## REFERENCES

Abdelali, A., J. Cowie and H.S. Soliman, 2004. Arabic information retrieval perspectives. Proceedings of the 11th Conference on Natural Language Processing, Journes d'Etude sur la Parole-Traitement Automatique des Langues Naturelles (JEP-TALN'04), April 19-22, 2004, JEP-TALN, France, pp: 391-400.

Abdelali, A., J. Cowie and H.S. Soliman, 2007. Improving query precision using semantic expansion. Inf. Process. Manage., 43: 705-716.

Abouenour, L. and K. Bouzouba and P. Rosso, 2010. An evaluated semantic query expansion and structure-based approach for enhancing Arabic question/answering. Int. J. Inform. Commun. Technol., 3: 37-51.

Abouenour, L. and K. Bouzoubaa and P. Rosso, 2009. Structure-based evaluation of an arabic semantic query expansion using the JIRS passage retrieval system. Proceedings of the EACL 2009 Workshop on Computational Approaches to Semitic Languages, March 31, 2009, Athens, Greece, pp: 62-66.

Abouenour, L., 2011. On the Improvement of Passage Retrieval in Arabic Question/Answering (Q/A) Systems. In: Natural Language Processing and Information Systems, Munoz, R., A. Montoyo and E. Metais (Eds.). Springer, Berlin, Germany, ISBN:978-3-642-22326-6, pp: 336.

Abouenour, L., K. Bouzoubaa and P. Rosso, 2008. Improving Q/A using Arabic wordnet. Proceedings of the 2008 International Arab Conference on Information Technology (ACIT'08), December 16-18, 2008, Sfax University Tunisia, North Africa, pp: 1-8.

Abu-Salem, H., M. Al-Omari and M.W. Evens, 1999. Stemming methodologies over individual query words for an Arabic information retrieval system. J. Assoc. Inf. Sci. Technol., 50: 524-529.

Abusalah, M., J. Tait and M. Oakes, 2009. Cross language information retrieval using multilingual ontology as translation and query expansion base. Polibits, 40: 13-16.

Ahmed, F. and A. Nurnberger, 2009. Evaluation of N-gram conflation approaches for Arabic text retrieval. J. Assoc. Inf. Sci. Technol., 60: 1448-1465.

Al-Ameed, H.K., S.O. Al-Ketbi, A.A. Al-Kaabi, K.S. Al-Shebli and N.F. Al-Shamsi *et al*., 2006. Arabic search engines improvement: A new approach using search key expansion derived from arabic synonyms structure. Proceedings of the IEEE International Conference on Computer Systems and Applications, March 8, 2006, IEEE, Dubai, UAE., ISBN:1-4244-0211-5, pp: 944-951.

Al-Eroud, A.F., M.A. Al-Ramahi, M.N. Al-Kabi, I.M. Alsmadi and E.M. Al-Shawakfa, 2011. Evaluating Google queries based on language preferences. J. Inf. Sci., 37: 282-292.

Al-Fedaghi, S. and F. Al-Anzi, 1989. A new algorithm to generate Arabic root-pattern forms. Proceedings of the 11th National Computer Conference and Exhibition, NCCE'1989, Dhahran, Saudi Arabia, pp: 4-7.

Al-Kabi, M., H. Wahsheh, I. Alsmadi, E. Al-Shawakfa and A. Wahbeh *et al*., 2012. Content-based analysis to detect Arabic web spam. J. Inf. Sci., 38: 284-296.

Al-Rajebah, N.I. and H.S. Al-Khalifa, 2014. Extracting ontologies from arabic wikipedia: A linguistic approach. Arabian J. Sci. Eng., 39: 2749-2771.

Al-Rajebah, N.I., H.S. Al-Khalifa and A.S. Al-Salman, 2010. Building ontological models from Arabic Wikipedia: A proposed hybrid approach. Proceedings of the 12th International Conference on Information Integration and Web-based Applications and Services, November 8-10, 2010, ACM, Paris, France, pp: 899-902.

Al-Shammari, E. and J. Lin, 2008b. A novel Arabic lemmatization algorithm. Proceedings of the 2nd Workshop on Analytics for Noisy Unstructured Text Data, July 24-24, Singapore, pp: 113-118.

Al-Shammari, E.T. and J. Lin, 2008a. Towards an error-free Arabic stemming. Proceedings of the 2nd ACM Workshop on Improving non English Web Searching, October 30, 2008, ACM, California, USA., pp: 9-16.

Al-Shammari, E.T., 2013. Lemmatizing, stemming and query expansion method and system. US Patent No. 8,473,279, United States Patent and Trademark Office, Washington, DC., USA. https://www.google.com/patents/US8473279.

Albared, M., N. Omar and M.J.A. Aziz, 2009. Classifiers combination to Arabic morphosyntactic disambiguation. Proceeding of the International Conference on Electrical Engineering and Informatics, August 5-7, 2009, Selangor, Malaysia, pp: 163-171.

Aljlayl, M. and O. Frieder, 2002. On Arabic search: Improving the retrieval effectiveness via a light stemming approach. Proceedings of the 11th International Conference on Information and Knowledge Management, November 04-09, 2002, ACM, McLean, Virginia, ISBN:1-58113-492-4, pp: 340-347.

Alqudsi, A., N. Omar and K. Shaker, 2012. Arabic machine translation: A survey. Artif. Intell. Rev., 42: 549-572.

Al-Sughaiyer, I.A. and I.A. Al-Kharashi, 2004. Arabic morphological analysis techniques: A comprehensive survey. J. Assoc. Inf. Sci. Technol., 55: 189-213.

Arampatzis, A.T., T. Tsoris, C.H.A. Koster and T.P.V.D. Weide, 1998. Phase-based information retrieval. Inf. Process. Manage., 34: 693-707.

Attar, R. and A.S. Fraenkel, 1977. Local feedback in full-text retrieval systems. J. ACM., 24: 397-417.

Attia, M., 2006. An ambiguity-controlled morphological analyzer for modern standard Arabic modelling finite state networks. Proceedings of the Conference on Challenges of Arabic for NLP/MT Vol. 200610, October 23, 2006, British Computer Society, London, UK., pp: 48-67.

Atwan, J. and M. Mohd, 2012. Arabic information retrieval: A semantic query expansion technique. Proceedings of the 2nd National Doctoral Seminar on Artificial Intelligence Technology, November 19-20, 2012, UNITEN Residence Hotel, Selangor, Malaysia, pp: 19-25.

Atwan, J., M. Mohd and G. Kanaan, 2013. Enhanced Arabic Information Retrieval: Light Stemming and Stop Words. In: Soft Computing Applications and Intelligent Systems, Noah, S.A., A. Abdullah, H. Arshad, A.A. Bakar and Z.A. Othman *et al*. (Eds.). Springer, Berlin, Germany, ISBN:978-3-642-40566-2, pp: 219-228.

Baeza-Yates, R. and B. Ribeiro-Neto, 1999. Modern Information Retrieval. Pearson, London, UK., ISBN:978-81-317-0977-1, Pages: 517.

Belkredim, F.Z. and F. Meziane, 2008. Dear-onto: A derivational Arabic ontology based on verbs. Intl. J. Comput. Process. Lang., 21: 279-291.

Belkredim, F.Z., A. El-Sebai and U.H.B. Bouali, 2009. An ontology based formalism for the Arabic language using verbs and their derivatives. Commun. IBIMA., 11: 44-52.

Bellare, K., P.P. Talukdar, G. Kumaran, F. Pereira and M. Liberman *et al*., 2007. Lightly-supervised attribute extraction. Proceedings of the NIPS 2007 Workshop on Machine Learning for Web Search Vol. 3, December 7, 2007, National Institute of Population Studies, Whistler, British Columbia, Canada, pp: 44-53.

Beseiso, M., A.R. Ahmad and R. Ismail, 2010. A survey of Arabic language support in semantic web. Intl. J. Comput. Appl., 9: 35-40.

Bhogal, J., A. Macfarlane and P. Smith, 2007. A review of ontology based query expansion. Inform. Process. Manage., 43: 866-886.

Carpineto, C. and G. Romano, 2012. A survey of automatic query expansion in information retrieval. ACM. Comput. Surv., 44: 1-1-1-50.

Carpineto, C., G. Romano and V. Giannini, 2002. Improving retrieval feedback with multiple term-ranking function combination. ACM. Trans. Inf. Syst., 20: 259-290.

Chinnakotla, M.K., K. Raman and P. Bhattacharyya, 2010. Multilingual pseudo-relevance feedback: Performance study of assisting languages. Proceedings of the 48th Annual Meeting on Association for Computational Linguistics, July 11-16, 2010, ACM, Uppsala, Sweden, pp: 1346-1356.

Croft, B., D. Metzler and T. Strohman, 2009. Search Engines: Information Retrieval in Practice. 1st Edn., Addison Wesley, London, UK., ISBN: 978-0136072249.

Darwish, K., 2002. Building a shallow Arabic morphological analyzer in one day. Proceedings of the ACL-02 workshop on Computational Approaches to Semitic Languages, (WCASL'2002), Philadelphia, Pennsylvania, pp: 1-8.

Dextre, C.S.G. and M.L. Zeng, 2012. From ISO 2788 to ISO 25964: The evolution of thesaurus standards towards interoperability and data modelling. Inf. Stand. Q., 24: 20-24.

Diab, M. and N. Habash, 2007. Arabic dialect processing tutorial. Proceedings of the Conference on Human Language Technology NAACL, Companion Volume: Tutorial Abstracts, April 22-27, 2007, Association for Computational Linguistics, Vancouver, Canada, pp: 5-6.

Duwairi, R.M., 2006. Machine learning for Arabic text categorization. J. Am. Soc. Inform. Sci. Technol., 57: 1005-1010.

Egozi, O., S. Markovitch and E. Gabrilovich, 2011. Concept-based information retrieval using explicit semantic analysis. ACM. Trans. Inf. Syst., 29: 8-1-8-34.

El-Beltagy, S.R. and A. Rafea, 2011. An accuracy-enhanced light stemmer for Arabic text. ACM. Trans. Speech Lang. Process., 7: 2-1-2-22.

El-Emary, I.M.M. and J. Atwan, 2005. Designing and building an automatic information retrieval system for handling the arabic data. Am. J. Applied Sci., 2: 1520-1525.

El-Kourdi, M., A. Bensaid and T.E. Rachidi, 2004. Automatic Arabic document categorization based on the Naive Bayes algorithm. Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages, August 28, 2004, ACM, Geneva, Switzerland, pp: 51-58.

Elkateb, S., W. Black, H. Rodriguez, M. Alkhalifa, P. Vossen, A. Pease and C. Fellbaum, 2006. Building a WordNet for Arabic. Proceedings of The 5th International Conference on Language Resources and Evaluation, May 22-28, 2006, Genoa-Italy, pp: 29-34.

Farghaly, A. and K. Shaalan, 2009. Arabic natural language processing: challenges and solutions. ACM Trans. Asian Language Inform. Process. Assoc. Comput. Mach., 8: 1-22.

Galvez, C., F.D. Moya-Anegon and V.H. Solana, 2005. Term conflation methods in information retrieval: Non-linguistic and linguistic approaches. J. Doc., 61: 520-547.

Graff, D. and K. Walker, 2001. Arabic newswire part 1. Linguistic Data Consortium, Philadelphia, Pennsylvania.

Hammo, B., H. Abu-Salem, S. Lytinen and M. Evens, 2002. QARAB: A question answering system to support the Arabic language. Proceedings of the 40th Association for Computational Linguistics on Computational Approaches to Semetic Languages, (ACLCASL'2002), Pennsylvania, USA., pp: 55-65.

Hammo, B.H., 2009. Towards enhancing retrieval effectiveness of search engines for diacritisized Arabic documents. Inf. Retrieval, 12: 300-323.

Harrag, F., A. Hamdi-Cherif, A.M.S. Al-Salman and E. El-Qawasmeh, 2009. Experiments in improvement of Arabic information retrieval. Proceedings of the 3rd International Conference on Arabic Language Processing (CITALA'09), May 4-5, 2009, IEEE, Rabat, Morocco, pp: 71-81.

Hoseini, M.A.S., 2011. Modeling the Arabic language through verb based ontology. Intl. J. Acad. Res., 3: 800-804.

Jarrar, M., 2011. Building a formal Arabic ontology (invited paper). Proceedings of the Experts Meeting on Arabic Ontologies and Semantic Networks, July 26-28, 2011, Alecso, Tunis, Tunisia, pp: 1-11.

Joachims, T., 1996. A probabilistic analysis of the rocchio algorithm with TFIDF for text categorization. MCs Thesis, Defense Technical Information Center, Virginia, USA.

Jurafsky, D. and J.H. Martin, 2008. Speech and Language Processing: An Introduction to Speech Recognition, Computational Linguistics and Natural Language Processing. 2nd Edn., Prentice Hall, New York, pp: 1024.

Kamir, D., N. Soreq and Y. Neeman, 2002. A comprehensive NLP system for modern standard Arabic and modern Hebrew. Proceedings of the Workshop on Computational Approaches to Semitic Languages (ACL-02), July 11, 2002, ACM, Philadelphia, Pennsylvania, pp: 1-9.

Kanaan, G., R. Al-Shalabi, S. Ghwanmeh and B. Bani-Ismail, 2007. A comparison between interactive and automatic query expansion applied on arabic language. Proceedings of the 4th International Conference on Innovations in Information Technology, November 18-20, 2007, Dubai, pp: 466-470.

Khafajeh, H. and N. Yousef, 2013. Evaluation of different query expansion techniques by using different similarity measures in Arabic documents. Intl. J. Comput. Sci., 10: 160-166.

Khafajeh, H., N. Yousef and G. Kanaan, 2010. Automatic query expansion for Arabic text retrieval based on association and similarity thesaurus. Proceedings of the European, Mediterranean and Middle Eastern Conference on Information Systems (EMCIS'10), April 12, 2010, EMCIS, Abu Dhabi, UAE., pp: 1-17.

Khoja, S., 2001. APT: Arabic part-of-speech tagger. Proceedings of the Student Workshop at the Second Meeting of the North American Chapter of the Association for Computational Linguistics. Carnegie Mellon University, Pittsburgh, Pennsylvania. June 2001.

Larkey, L., L. Ballesteros and M. Connell, 2007. Light stemming for Arabic information retrieval. Arabic Comput. Morphol., 38: 221-243.

Larkey, L.S. and M.E. Connell, 2002. Arabic information retrieval at UMass in TREC-10. Master Thesis, National Institute of Standards and Technology, Gaithersburg, Maryland.

Liu, S., 2006. Improve text retrieval effectiveness and robustness. Ph.D Thesis, University of Illinois at Chicago, Chicago, Illinois.

Lopis, F., J.L. Vicedo and A. Ferrandez, 2002. Passage selection to improve question answering. Proceedings of the 2002 Conference on Multilingual Summarization and Question Answering, August 31, 2002, Association for Computational Linguistics Stroudsburg, Pennsylvania, USA., pp: 1-6.

Manning, C.D., P. Raghavan and H. Schutze, 2008. Introduction to Information Retrieval. Cambridge University Press, Cambridge, UK.,.

Menai, M.E.B. and W. Alsaeedan, 2012. Genetic algorithm for Arabic word sense disambiguation. Proceedings of the 13th ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel and Distributed Computing (SNPD'12), August 8-10, 2012, IEEE, Kyoto, Japan, ISBN:978-1-4673-2120-4, pp: 195-200.

Mitra, M., A. Singhal and C. Buckley, 1998. Improving automatic query expansion. Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 24-28, 1998, ACM, Melbourne, Australia, ISBN:1-58113-015-5, pp: 206-214.

Moawad, I.F., M. Abdeen and M.M. Aref, 2010. Ontology-based architecture for an Arabic semantic search engine. Proceedings of the 10th Conference on Language Engineering, December 15-16, 2010, Egyptian Society of Language Engineering (ESOLE), Cairo, Egypt, pp: 67-73.

Navigli, R., 2009. Word sense disambiguation: A survey. ACM Comput. Surv., Vol. 41, No. 2. 10.1145/1459352.1459355

Nwesri, A., 2008. Effective retrieval techniques for Arabic text. Ph.D Thesis, RMIT University, Melbourne, Victoria.

Otair, M.A., G. Kanaan and R. Kanaan, 2013. Optimizing an Arabic query using comprehensive query expansion techniques. Int. J. Comput. Applic., 71: 42-49.

Paice, C.D., 1994. An evaluation method for stemming algorithms. Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Jul. 3-6, Dublin, Ireland, pp: 42-50.

Paice, C.D., 1996. Method for evaluation of stemming algorithms based on error counting. J. Assoc. Inf. Sci. Technol., 47: 632-649.

Petrov, V., 2011. Ontological Landscapes: Recent Thought on Conceptual Interfaces Between Science and Philosophy. Walter de Gruyter, Berlin, Germany,.

Pinto, F.J. and C.F. Perez-Sanjulian, 2008. Automatic query expansion and word sense disambiguation with long and short queries using WordNet under vector model. Conf. Software Eng. Databases, 2: 17-23.

Rachidi, T., M. Bouzoubaa, L. El-Mortaji, B. Boussouab and A. Bensaid, 2003. Arabic user search query correction and expansion. Proc. Copstic, 3: 11-13.

Salton, G. and C. Buckley, 1997. Improving Retrieval Performance by Relevance Feedback. In: Readings in Information Retrieval, Jones, K.S. and P. Willett (Eds.). Morgan Kaufmann Publishers, Burlington, Massachusetts, pp: 355-363.

Savary, A. and C. Jacquemin, 2003. Reducing Information Variation in Text. In: Text- and Speech-Triggered Information Access, Renals, S. and G. Grefenstette (Eds.). Springer, Berlin, Germany, ISBN:978-3-540-40635-8, pp: 145-181.

Shaalan, K., S. Al-Sheikh and F. Oroumchian, 2012. Query Expansion Based-on Similarity of Terms for Improving Arabic Information Retrieval. In: Intelligent Information Processing, Shi, Z., D. Leake and S. Vadera (Eds.). Springer, Berlin, Germany, ISBN:978-3-642-32890-9, pp: 167-176.

Singhal, A., 2012. Introducing the knowledge graph: Things, not strings. Official Google Blog, New York, USA.

Soudi, A., G. Neumann and A. Bosch, 2007. Arabic Computational Morphology: Knowledge-Based and Empirical Methods. In: Arabic Computational Morphology, Soudi, A., A. van den Bosch and G. Neumann (Eds.). Text, Speech and Language Technology Volume 38, Springer, The Netherlands, ISBN: 978-1-4020-6045-8, pp: 3-14.

Taghva, K., R. Elkhoury and J. Coombs, 2005. Arabic stemming without a root dictionary. Proceedings of the International Conference on Information Technology: Coding and Computing (ITCC'05) Vol. 1, April 4-6, 2005, IEEE, Las Vegas, Nevada, pp: 152-157.

Voorhees, E.M. and D. Harman, 1999. The seventh text retrieval conference (TREC-7). Master Thesis, National Institute of Standards and Technology, Gaithersburg, Maryland.

Xu, J. and W.B. Croft, 2000. Improving the effectiveness of information retrieval with local context analysis. ACM. Trans. Inf. Syst., 18: 79-112.

Zaidi, S., M.T. Laskri and K. Bechkoum, 2005. A cross-language information retrieval based on an Arabic ontology in the legal domain. Proceedings of the International Conference on Signal-Image Technology and Internet-Based Systems (SITIS'05), November 27-December 1, 2005, Hotel Hilton Suites, Lahore Pakistan, pp: 86-91.