JISTaP http://www.jistap.org
Journal of Information Science Theory and Practice
eISSN : 2287-4577 pISSN : 2287-9099

Research Paper
J Inf Sci Theory Pract 8(4): 67-86, 2020
https://doi.org/10.1633/JISTaP.2020.8.4.6

Survey of Automatic Query Expansion for Arabic Text Retrieval

Yasir Hadi Farhan*

Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, Bangi, Malaysia
E-mail: yasir.hadi87@yahoo.com

Masnizah Mohd

Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, Bangi, Malaysia
E-mail: masnizah.mohd@ukm.edu.my

Shahrul Azman Mohd Noah

Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, Bangi, Malaysia
E-mail: shahrul@ukm.edu.my

ABSTRACT

Information need has been one of the main motivations for a person using a search engine. Queries can represent very different information needs. Ironically, a query can be a poor representation of the information need because the user can find it difficult to express the information need. Query Expansion (QE) is being popularly used to address this limitation. While QE can be considered as a language-independent technique, recent findings have shown that in certain cases, language plays an important role. Arabic is a language with a particularly large vocabulary rich in words with synonymous shades of meaning and has high morphological complexity. This paper, therefore, provides a review on QE for Arabic information retrieval, the intention being to identify the recent state-of-the-art of this burgeoning area. In this review, we primarily discuss statistical QE approaches that include document analysis, search, browse log analyses, and web knowledge analyses, in addition to the semantic QE approaches, which use semantic knowledge structures to extract meaningful word relationships. Finally, our conclusion is that QE regarding the Arabic language is subjected to additional investigation and research due to the intricate nature of this language.

Keywords: information retrieval, Arabic text retrieval, query expansion

Open Access

Received date: October 6, 2020
Accepted date: December 8, 2020

***Corresponding Author:** Yasir Hadi Farhan
PhD Student
Universiti Kebangsaan Malaysia (UKM), Selangor, Bangi 43600, Malaysia
E-mail: yasir.hadi87@yahoo.com

1. INTRODUCTION

One outstanding and vital service of the Internet is web search, where, with extensive web documents available on the Internet, users discover and have access to valuable data and materials through search engines. For instance, if a person wishes to find out detailed information about في اللغة العربية الاخطاء اللغويه الشائعة (common linguistic mistakes in the Arabic language), he/she must just type الاخطاء اللغويه الشائعة في اللغة العربية in any of the search engines, and a large number of related web documents will be displayed, with the most suitable ones listed on top. Each search result comes with a title, URL address, and a snippet (a phrase or sentence taken by the search engines from the related web documents) (Cui, Wen, Nie, & Ma, 2002).

As the Internet is eminently usable, and online information and users' web sites have multiplied manifold, this information is usually not organized, posing difficulties to users in fulfilling their information needs. To get better results, users must formulate their queries properly, through which the correct term could retrieve the related documents. Although observers mandate that users' queries should be short, comprising just two or three words, formulating the user's queries is considered to be a tricky task (Berget & Sandnes, 2015). According to Belkin (1980), it is difficult for people to define precisely what their information need is, because that information is a gap in their knowledge, which he called the Anomalous State of Knowledge hypothesis.

Some methods have been proposed to fill the knowledge gaps, one of which is Query Expansion (QE) (Azad & Deepak, 2019; Dalton, Naseri, Dietz, & Allan, 2019). The hypothesis of QE is to suggest new related words to add to the original query in order to retrieve more related documents. Search engines such as Google and Yahoo offer suggestion-related queries to users, where the user has the option to retain his or her query or choose the suggested query (Wang, Lai, & Liu, 2009). QE is of four types: *Manual Query Expansion (MQE)*, *Automatic Query Expansion (AQE)*, *Interactive Query Expansion (IQE)*, and *Hybrid Query Expansion (HQE)*. In this review, the focus will be on AQE.

Over the years, AQE methods have been suggested as an effective tool for overcoming the term mismatch problem (Vechtomova, 2009; White & Horvitz, 2015). Vocabulary mismatch occurs when a user's query does not match the content of the stored documents. AQE aims to improve retrieval performance by inserting new words into the initial query, provided these are related to the original query terms. Researchers classify AQE techniques into local and global approaches; the latter use ontology or other semantic resources like WordNet to expand the original query independent of the query and the retrieved outcomes (Pal, Mitra, & Datta, 2014). On the other hand, local approaches use the documents retrieved from the initial query, envisaging the finding of useful terms in these documents, to add them to the original query for expansion. This relies on Relevance Feedback (RF) approaches.

For Information Retrieval (IR) purposes, the Arabic language has been less studied than other European (and to a certain extent Asian) languages (El Mahdaouy, El Alaoui, & Gaussier, 2018). However, in recent years, significant efforts have been devoted to improve the performance of IR and Natural Language Processing (NLP) tasks in the Arabic language (Batita & Zrigui, 2018), although, due to the complex morphology of Arabic, Arabic semantic IR is still an obstacle and challenge to researchers in this field (Belkredim, El Sebai, & Bouali, 2009; Bounhas, Soudani, & Slimani, 2020) .

Arabic is a right to left language, characterized by a typically huge vocabulary, with ambiguity in the meaning of numerous words (Abu-Errub, 2014). Four hundred and twenty-two million people use Arabic as the local language, and another 250 million use it as a second language. According to Abouenour (Abouenour, Bouzouba, & Rosso, 2010), the distinctive features of Arabic, viz., its lack of capital letters, complex morphology, and short vowels, pose significant challenges to the research community. These exceptional features also confound Arabic Information Retrieval (AIR).

Arabic consists of two major divisions: alphabets and diacritical marks. Alphabets represent the consonant sounds, whereas diacritical marks represent the short vowels causing variations in pronunciation. For instance, the word "كتب" can bring more than one valid interpretation, depending on the short vowels put on the letters of the word. "كَتَبَ" means (he wrote), but "كُتُب" means (books). Another word is "ذَهَبَ", which means (he went), whereas "ذَهَبْ" means (gold).

Many methods have been recommended for the enrichment of AIR (query correction, QE, as well as stemming and lemmatization), most of which encompass a vital task between NLP and IR. However, Arabic text comes with a new set of challenges, for the complexity of the language is as problematic as it is challenging. Farghaly and Shaalan (2009) consider developing the Arabic language in the IR field a novel study. The queries and research are quite inadequate compared to the work in other languages like English.

The main aim of this paper is to review works related to QE for AIR, the intention being to identify the recent state-of-the-art of this burgeoning area. Earlier review papers (Atwan & Mohd, 2017) regarding Arabic QE techniques focus merely on reviewing and classifying the traditional QE techniques without highlighting current techniques like Word Embedding (WE), considered one of the most important techniques for improving retrieval performance. Therefore, our review endeavours to cover all the Arabic QE techniques and highlight their advantages and disadvantages.

2. QUERY EXPANSION

QE occurs when a new word is added to the original query, with the aim to enhance retrieval performance. As indicated previously, there are four different ways to expand the query:

- *MQE* is based on the skilful decision of the user, who selects the candidate terms and reformulates the initial query manually. However, earlier studies found that manual choice of candidate terms for the expansion may retrieve only twenty-five percent from the relevant documents in the collection (Sharma, Pamula, & Chauhan, 2019).
- *IQE* is also called semi-automatic, where the system gives suggestion terms, and the user selects the proper expansion terms which can retrieve more relevant documents in the next iteration.
- *AQE*: The weight of each candidate term is computed, and the highest weighted terms are selected to add to the initial query (the system executes the whole process without involving the user).
- *HQE* refers to combining two or more methods for the expansion process, as in Han and Chen (2009), who proposed a hybrid method combining two techniques: ontology-based and neural networks.

2.1. Automatic Query Expansion

In our review, we focus on AQE, which is the process of automatically enhancing retrieval effectiveness, without involving users in the assortment of terms. AQE selects the words which have the highest weight, which will contribute to the reformulation of the original query in improving retrieval performance. For the results to be improved, suitable weight terms are required (Ooi, Ma, Qin, & Liew, 2015). Fig. 1 shows the main steps of AQE.

2.2. Automatic Query Expansion Approaches

In this type, the whole process of the expansion, consisting of getting the expansion terms set, choosing the proper terms from this set, assigning a weight to each term, and then formulating the query, is implemented by the system without involving the user. AQE can be classified into 1) statistical approaches, 2) semantic approaches, and 3) hybrid approaches (Raza, Mokhtar, & Ahmad, 2018).

2.2.1. Statistical Query Expansion Approaches

The idea of statistical QE approaches is to apply a statistical or probabilistic method to the content of documents to mine term × term association or term × query association, which is utilized to identify the candidate expansion terms for the query. The statistical QE approaches focus on the analysis of the document corpus, which can be either a collection of web documents, search and browse history documents, or text documents. During the expansion process, non-relevant terms may be added to the query as new expansion terms, which may cause a query drift, i.e., straying to topics other than the main query. The statistical QE approaches reduce the risk of such drifting.

2.2.2. Semantic Query Expansion Approaches

The semantic QE approaches use the original search query to extract meaningful word relationships using semantic knowledge structures, where the query can be expanded semantically based on these relationships. The knowledge structure, also known as thesaurus, lexicon, or ontology, consists of a set of concepts and a set of relationships among the concepts (Gruber, 1995). The knowledge structure can be generated manually (e.g., WordNet (Miller, Beckwith, Fellbaum, Gross, & Miller, 1990)), or automatically (e.g., similarity thesaurus (Qiu & Frei, 1993)), and may be dependent on a specific domain, or be independent, where the relationships between concepts can be either synonym, antonym, part-of, hyponym, or located-at.

2.2.3. Hybrid Query Expansion Approaches

The hybrid QE approaches combine two or more QE approaches such as statistical or semantic, or a combination of statistical and semantic approaches, to improve results by

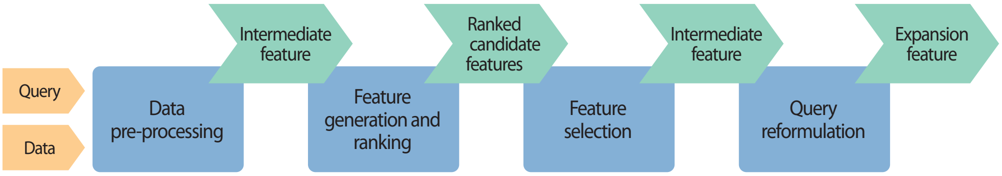

Fig. 1. Automatic query expansion.

obtaining the positive characteristics of each approach. The main idea of hybrid QE approaches is to obtain an advantage greater than that from applying only one approach, either statistical or semantic. Recently, to overcome the limitations of individual QE techniques, many studies have combined various QE methods for achieving better performance.

2.3. Automatic Query Expansion Stages

According to Ooi et al. (2015), AQE can be divided into four phases, viz., pre-processing of the data, ranking and generating the candidate terms, selection of the new terms, and reformulating the query.

2.3.1. Pre-processing of Data

A new suitable style of the data depends on this step, for the query to be expanded. In this phase, the data source will be transformed into another format to be more effectively processed by subsequent phases (Carpineto & Romano, 2012). In this phase, the following steps are to be followed:

- Extracting the text from the source corpus, like Word Doc., PDF, etc.
- Tokenization (only the individual words are selected)
- Stop-word Removal (prepositions and common words will be deleted)
- Stemming of the words (returning the deriving words to their roots)
- Word Weighting (each term will be assigned a weight)

After these steps are applied, it is easier to deal with this text, because each term has a weight, and each term is already returned to its root.

2.3.2. Ranking and Generating the Candidate Terms

Ranking and generating the candidate terms are considered the most crucial stage, because the selected terms will be a small number that is added to the original query. The data source and the users' queries are the input in the current step, and a collection of new terms with their weights is the output. Generating and ranking of the new words is a task of the system.

2.3.3. Selection of the New Terms

The top candidate terms are selected after the query terms are ranked. An individual basis is adopted to select the top terms, and the mutual dependencies of the candidate terms are ignored (Lin & Murray, 2005).

2.3.4. Reformulating the Query

Reformulating the query is the last stage of AQE, also well known as the described expanded query that will be run into the search box to get a better result. The process is to reconstruct the original query and add the new candidate terms which have the highest weight. Numerous methods, such as the Rocchio algorithm, have been proposed to find the new candidate terms correlated with the query terms, using diverse techniques that depend on various data sources.

3. AUTOMATIC QUERY EXPANSION METHODS

Carpineto and Romano (2012) categorize AQE methods into five groups based on the conceptual paradigm applied to obtain enlargement features, viz., Linguistic Methods, Corpus-Specific Statistical Methods, Query-Specific Statistical Methods, Search Log Analysis Method, and Web Data. The Linguistic Methods, Corpus-Specific Statistical Methods, and Query-Specific Statistical Methods are related to our review and are described below.

Linguistic analysis refers to the scientific analysis of a language sample, involving at least one of the five main branches of linguistics, viz., phonology, morphology, syntax, semantics, and pragmatics. Linguistic analysis can be used to describe the unconscious rules and processes that speakers of languages use to create spoken or written language and can be useful to those desiring to learn a language or translate from one language to another. It is usually found in thesauri, dictionaries, or other related knowledge sources like WordNet. As the development of features is generally made independent of the complete query and the database's content searched, it is usually more sensitive to Word-Sense Disambiguation (WSD) (Carpineto & Romano, 2012).

Query-Specific Local Methods are more effective than corpus-specific techniques because they exploit the local context provided by the query. In contrast, corpus-specific techniques might be based on documents that are found frequently in the corpus which may be irrelevant for the query—additionally, *Query-Specific Local Methods* utilize the top-ranked retrieved documents (Carpineto & Romano, 2012).

4. ARABIC QUERY EXPANSION APPROACHES

As mentioned earlier, Arabic has a large vocabulary. The language used for official writing is Standard Arabic, which is an antique Arabic language including borrowed words originating from other languages. Different borrowed words and synonyms tend to be applied in diverse writings. The Arabic structure style also tends to differ across Arab countries (Abdelali, 2004).

Another prominent feature of Arabic is its great structure of inflection morphology, its high incidence of word polysemy and its shape, and root ambiguity like “Pattern Polysemy” and “Root-Homonymy” (Haddad, 2014; Hattab, Haddad, Yaseen, Duraidi, & Shmais, 2009). The human-system relations can adversely affect each other, with cases of incomplete and unreliable representation mostly due to little vague words, which may also be diffused with cognitive phonetic spelling errors (Haddad, 2014; Hattab et al., 2009), inducing inappropriate or even appropriate information to appear, for example, reducing precision or recall in the IR system and confounding the inter-cognitive message. Such inadequacy in communication can be considered a type of misapprehension and a worrying inter-cognitive communication between machines and humans (Hattab et al., 2009).

However, due to the vital position of the Arabic language, the research to develop IR for it is still not adequate compared to other languages (Abbache, Meziane, Belalem, & Belkredim, 2016a; Abu-Errub, 2014). Additionally, emphasis needs to be placed on this language, so that assistance can be provided to its users in their exploration and quests.

One benefit brought by QE is that there are more chances of retrieving a relevant document that did not appear with the original query words, with a clear increase in recall. For example, if the query *Al-Qaeda* is expanded to *Al-Qaeda*, *al-Qaida*, *al-Qa’ida*, *Osama bin Laden*, *terrorist Sunni organization*, or September 11, 2001, it can yield a better result. In this case, this initial query not only retrieves or obtains the documents with the original or initial term (Al-Qaeda), but also other spellings in the other documents. Experiments and research efforts investigating the active role of QE in the retrieval of Arabic text are still few and far between, compared to studies carried out in other languages like English (Farghaly & Shaalan, 2009).

4.1. Arabic Query Expansion Based on WordNet

One of the linguistic QE techniques is to find the related words of the given query, by using the thesaurus to pick the synonyms for that word. One of the most common methods is WordNet (Jiang & Conrath, 1996; Mandala, Takenobu, & Hozumi, 1998), a lexical database that can group words into a set of synonyms called synsets. This technique expands the original query by analyzing the expansion features such as lexical, morphological, semantic, and syntactic term relationships.

Trad, Mustafa, Koroni, and Almaghrabi (2012) presented the AQE technique based on Arabic WordNet (AWN) Ontology for the Arabic language; this work is compared and evaluated with English WordNet ontology. A software system was built to evaluate the two different datasets with different languages, constructed with ASP.NET technology via C# language. The proposed method deals with two different datasets—the Arabic corpus and the English corpus. The content of each corpus is a medical document. However, the system deals with queries in different languages (Arabic and English). The proposed system was able to scan and index the documents through the multi-threaded procedure, where the output documents were stored in the Structured Query Language (SQL) database, which represents the inverted index. The proposed system had three steps: to find the proper synonyms of each query term, to select the best synonyms for the expansion, and the expansion process based on the words chosen. Boolean Model, Vector Space Model, and Latent Semantic Indexing model were employed for conducting the experiments. The subsystem of the proposed method expanded the original query based on WordNet ontology in three steps:

1. For each term in the original query, the set of synonyms was found (synset).
2. The most useful synonyms were selected to add to the original query.
3. Finally, the original query was expanded by adding useful synonyms based on the retrieval model used.

In the same domain, Mahgoub, Rashwan, Raafat, Zahran, and Fayek (2014) introduced a new AQE technique for semantic QE using a domain-independent semantic ontology, which depends on three Arabic resources: (1) Arabic Wikipedia, (2) Al Raed Dictionary, and (3) Google WordNet Dictionary. Two AQE strategies were explored, the first being to produce a single expanded query wherever this query contains all the expanded terms. The second strategy was to create multiple queries by expanding each word at a time, thus getting a single result list by combining the results of these queries. In the second strategy, for each expanded query, the noise source was the only one which was the expanded term, unlike the other words left without expansion, making this strategy less sensitive to noise. The expansion terms of every single query were set to be a maximum of 50 expansion terms, to increase recall measure. Light-10 stemmer was implemented due to its high performance compared to other stemming methods. Instead of having the entire quantity stemmed before indexing, the group of the word was clustered with a similar stem and placed in a similar document as a wordlist, which was then employed in extension. This method reduced the possibility of similarity among dual words conveying the same stem, but with different senses, which must originate from a similar document in mass to employ in extension. Al-Raed and Google WordNet can be used for further development. The open-source domain Lucene utilizes the proposed method. Lucene is an open-source IR library authorized under the Apache Software License (Białecki, Muir, & Ingersoll, 2012). Initially inscribed in Java, it has now been extended to additional software design languages. To identify how significant a document is to a user's query, Lucene hangs on the Vector Space Model (VSM) of IR and the Boolean Model. The results showed both strategies outperforming the baseline, with the second strategy performing better than the first.

Further, Al-Chalabi, Ray, and Shaalan (2015) presented the AQE method, which added a semantic term to the original query using the semantic resources AWN. Using the ontology resource AWN, each word in the query extracted the top ten synonyms. The Boolean operators (and & or) were applied to reformulate the initial query; hence, the new query was tested using the Google search engine. For example, the initial query was: العالم ما هي أكبر دولة في (Which is the largest country in the world). This query was expanded by employing AQE using AWN, changing into the new query: أو أوسع دولة أو بلد أو وطن في العالم ما هي أكبر أو أعظم (Which is the largest or biggest or widest country or land or homeland in the world). Fifty Arabic queries of the TREC & CLEF dataset were selected to test the proposed method.

Abbache, Meziane, Belalem, and Belkredim (2016b) presented the AQE technique exclusively focusing on short queries using the lexical database AWN and association rules to generate the new query terms. The users usually send their information needs as an initial query. However, short queries are a poor representation of users' information needs, and they do not express the actual user's intention; therefore, reformulating these queries becomes necessary. The proposed technique started in the pre-processing phase, where the stop words were removed to make the query ready for the next step. The next step was the extraction of the expansion candidate terms from the term selection source AWN. The Analyser (Query Analysis), which uses the light-stemming algorithm, was utilized to extract the useful expansion terms. Two strategies were used to expand the query: IQE and its application in sequence. In the first strategy, IQE, the system provides the terms of a suggestion, and the user must decide which term is to be selected to add to the initial query. The user's query is sent to the Index Search after analysis by the Analyser to get the results. And for each term in the initial query, the Synonyms Extractor supplies different parts of speech to assist the user in the selection. The user chooses the appropriate synonyms to add to the initial query, after which the search will commence with the new query. Though AQE employs the same steps as IQE, in the former, the system chooses the useful synonyms instead of the user. There must be a relation between the terms and their synonyms to ensure the choice of useful synonyms. The choice of the QE terms is based on the assumption that words occurring often in the documents most probably contain related or connected meanings and, as such, can be associated. The association can occur between the synonym and one of the query terms or their synonyms. The experiments were conducted with four different strategies:

(1) Simple search, called (R1), performed using the Lucene Index without expansion
(2) AQE based on synonyms, called (R2), involving expansion of the original query by synonym using Lucene
(3) AQE based on WordNet, called (RS), which concerns using WordNet and the association rules to expand the original query
(4) IQE, called (RI), where expansion depends on the user's selection of the new expansion terms

Based on the experiment results, all the proposed strategies (R1, R2, RS, and RI) showed significant improvement in terms of recall measure, while the strategies R1, R2, and RS indicated a decrease in the precision measure, especially in the first strategy, R1. The fourth strategy, RI, outperformed the others in terms of mean average precision (MAP). The reason behind the decreasing precision values was that the selection and addition of all the valid synonyms could insert noise, which will hurt the expansion, as justified by Abbache et al. (2016b). Additionally, AWN is a global database and does not contain synonyms for all Arabic words. Another challenge is the polysemous nature of Arabic words, where a term could indicate many meanings, making it ambiguous.

He and Wang (2009) used disambiguation on extended terms in an English-Arabic Cross-Language IR (CLIR), which improved the search performance. Corpus-based disambiguation was selected to retain the highest appropriate extended terms in the extended query. Words received utilizing QE were examined and collated with the meanings of the initial query terms specified in WordNet. The exact words employed in the meanings of a query team were reflected in the last extended query. In this research, an English-Arabic CLIR method employed QE, where English queries were used to search for Arabic documents. When this program received the input of the English query, it was interpreted with an English-to-Arabic dictionary. In any case, where one of the query words was not able to be construed or translated, transliteration was used.

As soon as the Arabic query was prepared, QE was adopted by Indril apparatuses. The extended terms were converted back to English using an Arabic-to-English dictionary, and every word was reflected by one translation. The word with the maximum chance was selected using a thesaurus-based disambiguation approach, to improve the efficiency of that method. The procedure of interpreting a query from one language to another frequently presents ambiguity and uncertainty. Query terms generally go beyond a single translation, and those translations could hold alternative words. Ascertaining the accurate translations to employ in the target query can be quite tricky. The sole method employed to improve the query translation is QE, used to improve the translated query with relevant terms extracted from the collected document. The simple method of QE comes in two phases; in the first phase, the assumed group of related documents is ascertained. For the second phase, the relevant terms are employed for query enhancement. This procedure of accumulating the proper words to the translated query helps improve the accuracy and recollection, as more related documents could be recognized and retrieved (Ballesteros & Croft, 1997). The experimental outcomes revealed that QE improved by disambiguation provided better efficiency.

Beirade, Azzoune, and Zegour (2019) proposed the QE method, which used the semantic relations between the words of the Holy Qur'an in creating the Quranic ontology, exhibiting the meaning of each term with their relations. The proposed method has been developed to recognize semantic terms in the Holy Qur'an. Apache Lucene's search engine, which uses an indexed dictionary of words in the Qur'an, was employed in this study. The content of the indexed dictionary is the Quranic exegesis of Ibn Kathir and Hadith of the Prophet Mohammed. To enhance the performance of Arabic QE, the term-to-term similarity and association techniques were utilized to create the Arabic thesauri.

The proposed system had three main stages: preparing documents, building a traditional IR system, and building the thesauri. Besides this, the procedure of the QE had three phases: the first was to send the query terms to the thesaurus, and the second was to get the related terms, followed by reformulation of the query. The experiment results demonstrated that the proposed system was useful for using semantic treatments with ontology.

4.2. Arabic Query Expansion Based on Pseudo Relevance Feedback

RF is considered the most effective technique for expanding query utilizing the extracted terms from retrieved documents to reformulate the original query. The method most similar to RF is Pseudo Relevance Feedback (PRF), also known as blind feedback. PRF is a local and vital QE technique (Singh & Sharan, 2017). PRF, proposed by ALMasri, Berrut, and Chevallet (2016) and Croft and Harper (1979), assumes that the top retrieved documents are relevant, and selects from these documents the useful terms to add to the initial query for the expansion purpose. The selection of the proper words is done automatically, without involving the user. Each word of the top-retrieved documents will get a score based on its co-occurrence of the query terms. Hence, the word with the highest score will be selected to formulate the initial query. The findings of some of the researchers who rely on this category are discussed below.

Atwan, Mohd, Rashaideh, and Kanaan (2016) proposed the AQE method, which relies on the semantic information to PRF to improve conventional IR processes for enhancing the AIR framework. Besides this, several Arabic stemmers were investigated, and a list of Arabic stop-words created. This study utilized the Arabic newswire TREC 2001 dataset to create an enhanced Arabic framework to improve the retrieval performance of AIR. The experiments were done using three lists of stop-words: the Khoja stop-words list, the Abu El-Khair general stop word list, and the combined stop words list. The third list combined the Abu El-Khair stop-words list with Khoja stop-words to make a new, combined list. The combined list was able to improve retrieval performance in terms of precision and reducing the corpus size. To find the semantic similarity between the query terms and the documents, AWN was utilised. Experimental results demonstrated that the AIR technique-based AWN was successful in picking up the useful synonym and reducing ambiguity.

The process of combining different Arabic word formats to their stem or root is called stemming. Indexing a text collection using stems or roots is superior to using original word formats. However, Arabic text stemming has adverse effects on words. It conflates words with a different meaning under one index term. It frequently occurs in Arabic when using stems, and it becomes more frequent when choosing roots to index the collection. Stemming is not always perfect, and stemming errors usually occur.

El Mahdaouy, El Alaoui, and Gaussier (2019) proposed incorporating WE similarity into PRF models for AIR, where WE similarities were the similarity between two vectors obtained via WE. The main idea was to select the expansion terms using their distribution in the set of top-relevant documents of PRF, along with their similarity to the original query terms. This study hypothesized that WE can be exploited in the PRF framework for AIR to deal with term mismatch, since similar words and words that should be grouped to the same stem would be close to each other in the vector space. The main goal was to boost the weights of semantically related terms to the original query terms. This work investigated three neural WE models, viz., Skip-gram, Continuous Bag of Words (CBOW), and the Global Vector (GloVe) models.

The WE similarities were incorporated into four PRF models, viz., the Kullback-Leibler divergence (KLD) (Carpineto, De Mori, Romano, & Bigi, 2001), the Bose-Einstein 2 (Bo2) of the Family of Divergence from Randomness models (Amati & Van Rijsbergen, 2002), and the Log-Logistic (LL), as well as the Smoothed Power-Law (SPL) of the information-based family of PRF models (Clinchant & Gaussier, 2013). Evaluations were performed on the standard Arabic TREC 2001/2002 test collection using the three neural WE models mentioned above. This study aimed to learn how WE might be exploited in PRF techniques for AIR. The results showed that the proposed PRF extensions significantly outperformed the baseline PRF models. Moreover, they enhanced the baseline IR model by 22% for MAP and the robustness index by 68%. The difference in performance among the three WE models (GloVe, CBOW, and Skip-gram models) was not statistically significant.

4.3. Arabic Query Expansion Based on Stemming

Another linguistic-based approach which expands the query globally is word stemming. Stemming is the process of reducing inflected words to their morphological root or word stem. The purpose of stemming is to combine the words with one stem into one index term, assuming that these words have the same meaning. The stemming technique can be either simple, removing pluralization suffixes from words, or more complicated, attempting to maintain meanings and incorporate dictionaries (Farrar & Hayes, 2019). Stemming is considered to be one of the earliest AQE techniques. The findings of researchers who used this approach to enhance retrieval performance are discussed below.

Khafajeh, Yousef, and Kanaan (2010) constructed an Arabic thesaurus based on the Term-Term Similarity and Association technique, which could be used to enhance the retrieval performance in any unique domain, to retrieve more documents relevant to the user's query. Two hundred forty-two Arabic abstract documents and fifty-nine Arabic queries were chosen from abstracts related to information about computer science and information system, presented at the Saudi Arabian National Computer Conference. An automatic IR system was devised and constructed from scratch to handle Arabic data. The association thesaurus helped improve precision and recall, compared to the similarity thesaurus.

However, the association thesaurus also contains numerous restrictions beyond the traditional IR system, particularly in precision and recall levels. Automatic stemmed words and full word index were constructed using an inverted file technique. Subject to these cataloguing words, scholars have created three IR systems. The first IR system employs a Traditional IR system using frequency-inverse document frequency (TF-IDF) to catalogue term weights. In the second, scholars employed Similarity Thesaurus, using the VSM with four similarity sizes (Dice, Cosine, Jaccard, and Inner product). TF-IDF was again used to catalogue term weights, and the similarity dimensions were compared for identifying the excellent one to be employed in developing the Similarity Thesaurus. For the third IR system, scholars employed an Association Thesaurus, using the Fuzzy Model for terms' weights to be catalogued.

The research dictates that the catalogued terms recur about two to seven times in the text. Aljlayl and Frieder (2002), after examining certain documents and recognizing the average number of incidences for some words, found the superlative catalogue terms recurring within the average would be the ones used. The test revealed that the Jaccard (which measures the degree of similarity between the retrieved documents) and Dice (used as a threshold and weight in ranking the retrieved documents) and the VSM model relationship dimensions were identical, while the Cosine and Inner also showed similar relationship dimensions. However, they were better than the method of Dice and Jaccard measures, and all the queries provided a closely similar position or standing. The researcher felt it to be better to employ stemmed words with the corresponding thesaurus in the retrieving system of the Arabic language, than to employ complete terms short of the thesaurus. Applying stemmed words with Association thesaurus in the Arabic language retrieving system is far better than employing complete words using Association thesaurus. All other studies in different languages approve of it. Association thesaurus is considered a better option than the traditional retrieval system. Its retrieval performance is even more enhanced than employing the corresponding thesaurus for the Arabic language retrieval system. When the stemming and Association thesaurus were jointly applied, excellent results were accomplished.

Nwesri and Alyagoubi (2015) revealed how to apply stemming from getting the same outcomes devoid of indexing the stemmed text. This method was based on indexing the original words and extracting different words using the stemmer, thus adding them to the initial query. This study dealt with the Arabic dataset, where the words in the language are constructed based on pattern systems, as in other Semitic languages. The authors constructed a table for each term in the corpus using a stemmer, instead of applying stemming to all the terms in the corpus. The main goal of this study was to overcome stemming issues such as the conflating of the words under one index term. The use of stemming was to generate stem-word clusters for expanding the query, but the stemmed text would not be indexed. Four stemmers were utilized: *light10*, *Khoja*, *Buckwalter morphological analyzer*, and *Alstem*. Lemur Toolkit was used to run experiments, where TREC 2001/2002 queries with the Agence France-Presse (AFP) text collection used in this study. Lemur Toolkit is an open-source software framework used to construct language modelling and IR software. It supports sundry retrieval models such as the VSM, Okapi BM25 model, and In-Query Retrieval model. In this study, the In-Query Retrieval model was utilised to test the experiments. The procedures applied to investigate the effects of using the stemmers on AQE formed the first step of this study: removing the diacritics, normalizing the text, removing stop-words, and then indexing the collection. The stemmer creates a list of stem-word pairs from the collection, where the created pairs are sorted based on their frequency in the collection from high to low. Hence, each query term is expanded by stemming it, and the most repeated terms in that list would be retrieved. These retrieved terms would be added to the original query as synonyms to that query terms. Based on the experiment results, the expansion using any of the stemmers could achieve significant improvement if the number of the terms added to the original query exceeded 40 words. Below ten terms, the results were lower than the baseline; if the expansion terms were 10, the results were at the baseline, and when it was 40 words, the results outperformed the baseline. Besides this, recall measure showed a good result when using light10, Buckwalter morphological analyzer, and Alstem stemmers, while using the Khoja root stemmer would hurt recall.

Hassan and Hadi (2017) developed an AQE technique to improve the AIR system, adopting two stages. At the first stage, the most relevant candidate terms were carefully chosen by using the initial query and the best synonyms detected at the same time via the corpus. At the second stage, extraction of the most relevant candidate terms was done from the first document set, and the query expanded to produce the ranked list. WSD was utilized in this study, and three different traditional semantic measurers: LCH, WUP, and PATH, were applied to avoid any mistake arising from word sense. Based on the experiment results, the PATH measure was found to be the best to solve ambiguity among the other measures. The authors claimed that their system outperformed the traditional method of precision, recall, and latency.

Hammo, Sleit, and El-Haj (2007) presented a new searching method without diacritics, where the Quran-related verses matched a user's query retrieved through AQE methods. The recommended method used an interpersonal database search engine accessible and mobile through Relational Database Management System (RDBMS) platforms and offered sophisticated and fast retrieval. The assimilation of the IR system notion with RDBMS was improved to preserve and operate on the scripts of the Quran. A search engine QE was mechanically created to identify all the verses containing words interrelated to the roots of the query's bag-of-words (BOW), where stemmer on the BOW was used. The root was obtained from the query, the Root Index was searched, and all the words related to the particular root were fetched from the Vowelized Words Index. The next step involved the initial query to be submitted for the second time for the Posting Table to search for the incidence of all verses containing particular words. Simultaneously, intensifying a quest query to search for the alternative words that the user inserts can help to similarly improve the recall, sometimes even accurately. When it is employed for IR, terms and their synonyms are made better in the vector, allowing a new search to start. Synonyms or words similar to the Quran's words are gathered and semantically clustered. A set of tests carried out on word search of the Quran words showed that QE was favourable to Arabic text search, with the bright possibility of its effectiveness being enhanced more. Rule-based stemmer was employed in this research to ascertain the efficiency of Arabic passage retrieval and question answering. The results from the study showed the efficiency of this method, particularly for a question answering system (Hammo et al., 2007).

4.4. Arabic Query Expansion Based on Co-occurrence of Words

The primary way to quantify the semantic relations between words is by computing the co-occurrence of these words. A hypothesis by Harris (1968) and Lindén and Piitulainen (2004) is that semantically similar words occur in the same contexts, where they co-occur with the same other words. Therefore, the co-occurrence of the term was used in IR systems for a while (Pôssas, Ziviani, Meira Jr, & Ribeiro-Neto, 2005; Zaïane & Antonie, 2002).

Shaalan, Al-Sheikh, and Oroumchian (2012) recommended a method for AQE on AIR through Expectation-Maximization (EM). EM is a statistical method used for finding the maximum likelihood of parameters and is typically used to compute the maximum likelihood estimates, given incomplete samples. It is guaranteed to find a local optimum of the data log-likelihood. In this research, EM is used to indicate the similarity between two words based on their co-occurrence in a set of documents. EM distance between word A and word B is calculated by dividing the total number of documents in which both the words appeared together, by the sum of the total number of documents that each word appeared separately. In this situation, the EM distance indicates the degree to which word A and word B are bonded, in terms of their concurrence in similar documents. The hypothesis is that the less the EM distance between two terms, the more bonded they are. EM algorithm is used when related terms are selected, so that the query can increase, and unrelated terms can be tidied out. INFILE assessment gathering of CLEF2009 further verified the algorithm. The recommended method revolved around the simultaneous occurrence of words when queries were expanded. In a paragraph about كأس العالم (world cup), familiar related words like كرة (ball), كرة القدم (football), أهداف (goals), حماسة- (excitement), and كأس البطولة (championship cup) are commonly found. Such terms are also frequently found in the documents that do not contain the correct connection to the keywords كأس العالم- (world cup). However, it still possesses the word موندیال, (Mondial) whose meaning is still related to the (world cup). The recommended method started with the examination of documents that possess the correct wording of the query, so that a list of co-occurring contextual words could be ascertained. This word list was utilized to expand the existing query. From then on, the expanded query was employed to expand the current query. A new set of documents do not necessarily have the exact words as the original query. This approach is likely to enlarge a query built on the similarity of terms to refine AIR. The test results showed that expanding queries regained more appropriate documents for queries, compared to the baseline.

Further, expanding the query using the proposed method improves the overall recall precision for the final list of retrieved documents. The readings from the baseline revealed significant data concerning the user information required in query titles and explanations, both of which supplement each another. The test runs showed that the recommended method enhanced the retrieval process in Arabic, while retaining a similar precision.

Due to the complexity of Arabic morphology, Bounhas et al. (2020) proposed the creation of a new morpho-semantic resource from classical Arabic vocalized corpus, based on a text mining process. The Classical Arabic Morpho-Semantic Knowledge Graph (CAMS-KG) resource was generated to analyse complexity in the Arabic text, with the proposed CAMS-KG resource rich in semantic knowledge. Arabic words have inflections or Harakat, also called Tashkeel, which are short vowels that can be positioned either at the top of the letter in the word or below it. The meaning of the words depends on these short vowels, with their position causing change in the meaning of the words. To reduces the ambiguities of Arabic words by identifying their actual meaning, the short vowels were considered while creating this resource. The proposed resource merges morphological and semantic knowledge. To compute the similarities of the Arabic tokens (terms), the BM25 model was employed.

The terms in the proposed resource most similar to the original query terms were selected and added to the initial query for the purpose of expansion, taking into account the contextual relationships between tokens and their whole meaning. Based on the BM25 model, the selected terms were weighted according to a normalized similarity score. Tashkeela (64) and ZAD (65) were two corpora utilized in this study for the experiments.

4.5. Arabic Query Expansion Based on Word Embedding

Traditional IR models are based on the BOW paradigm, where relevance scores are computed based on the exact matching of keywords between documents and queries, without allowing distinct semantically-related terms to match each other and contribute to the retrieval score. Although these models have already achieved excellent performance, it has been shown that most of the unsatisfactory cases in relevance are due to the term mismatch between queries and documents (El Mahdaouy, El Alaoui, & Gaussier, 2016). It is hence necessary to adopt a method that can consider all semantically-related terms in each document for a given query term.

Estimating distributed representation of each word is based on term proximity in the large collection, such as co-occurrence of the terms in the windows. Recently, the WE vector has gained significant attention, and has been successfully employed in various NLP and IR tasks (Kusner, Sun, Kolkin, & Weinberger, 2015; Vulić & Moens, 2015; Zhou, He, Zhao, & Hu, 2015). Further, using the WE was found to be a promising technique, sufficient for QE (Kuzi, Shtok, & Kurland, 2016; Sordoni, Bengio, & Nie, 2014).

Though the WE has proved its success as a promising technique for most of the NLP researchers in recent years, a limited number of studies focus on this technique (Diaz, Mitra, & Craswell, 2016; Zamani & Croft, 2016; Zuccon, Koopman, Bruza, & Azzopardi, 2015).

The undermentioned studies focus on exploiting the WE techniques by incorporating their semantic similarities into the existing IR models.

El Mahdaouy et al. (2016) introduced a new Semantically Enhanced Term Frequency (SMTF) based on the distributed representation of word vectors for AIR. To compute the semantic similarities between the query terms and terms of a given document weighted by their within-document frequency, three WE models were employed: CBOW model, Skip-Gram model (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013), and the GloVe model (Pennington, Socher, & Manning, 2014). The similarity between two-word vectors (trained using WE models) was computed using the cosine distance in the vector space given by $\cos(w_1, w_2) = \frac{W_1 \cdot W_2}{\| W_1 \| \| W_2 \|}$. The resulting term frequency (TF) was incorporated into the document growth functions of the probabilistic family of Information-Based models Log-Logistic Distribution (LGD), SPL Model , and the Okapi BM25 model. The method tried to address the mismatch between the query terms and the documents by calculating the TF based on the occurrence of the term in the query with semantically similar terms in the document. The experiments were performed on the standard Arabic newswire dataset TREC 2001/2002. Based on experiment results, the proposed method of incorporating SMTF into the probabilistic IR models outperformed the baseline BOW models. Moreover, the three suggested WE models showed the same performance on most test collections, with the CBOW and GloVe models found to be more suitable and flexible in incorporating similarity into the traditional IR models SPL and BM25.

Further, El Mahdaouy et al. (2018) proposed an AQE technique by incorporating the WE semantic similarities into the existing IR models. Language Model (LM), LGD Model , SPL Model, and Okapi BM25 were employed to incorporate the similarities of the WE techniques into these models, to improve the retrieval performance for Arabic text. Three neural WE techniques were proposed: Skip-gram, CBOW, and the GloVe models. The present work distributes the words in a vector space in a low-dimensional vector to address the term mismatch problem. The method proposed by these authors depends on the merger of the word similarity into their scoring function, where the suggested extensions contain substantial similarity to the translation models presented by Baeza-Yates et al. (2012). The set of translations for a given query was replaced by its set of similar terms, where the letter set is a set created from the top relevant words from the vocabulary corpus. The Semantic Term Matching Constraints (SMTCs) proposed by Fang and Zhai (2006) were utilized to investigate the proposed extension. The SMTCs are a framework that are capable of merging the semantic similarity and adjusting the term weights of the initial query and their similar words. Two strategies were proposed in this work, the first strategy is to select similar candidate words from the retrieved documents, while the second strategy is to selects the similar candidate terms from the whole corpus.

The experiments were conducted on TREC 2001/2002 Arabic newswire dataset. The results revealed that the proposed extension techniques significantly outperformed the baseline BOW models. The proposed method was better than three WE-based IR language models, and exceeded the semantic indexing approach proposed by Abderrahim (Abderrahim, Dib, Abderrahim, & Chikh, 2016). Additionally, there was no significant difference between the two strategies.

Since PRF has proved successful in dealing with the vocabulary mismatch problem, El Mahdaouy et al. (2019) presented a new method that incorporates the WE similarities into the current PRF models to enhance the AIR. The hypothesis was that the semantic terms which are relevant to the query terms are essential; therefore, the work aimed to boost the weights of these vital terms. The WE models Skip-gram, CBOW, and GloVe were utilised in this study, representing each word in the vector space by a single vector in a low-dimensional vector. Four PRF models, including the *Bo2* model (Amati & Van Rijsbergen, 2002), *LL* model, *KLD* (Carpineto & Romano, 2012), and the SPL (Clinchant & Gaussier, 2013), were proposed to incorporate the semantic similarity of the WE into these models. This work aimed to study how the WE semantic similarity could be employed in PRF techniques. The idea of the proposed method was to combine the distribution of the expansion terms into the set of PRF in the integrated PRF framework. The experiments and evaluations were performed on the standard Arabic newswire TREC 2001/2002 dataset. The proposed technique was better than the baseline IR model by 22% for MAP. Additionally, there was no significant difference among the three WE models (Skip-gram, CBOW, and GloVe models) in terms of performance.

Maryamah, Arifin, Sarno, and Morimoto (2019) proposed an AQE method based on *BabelNet* using the WE technique Word2Vec. BabelNet is a semantic search dictionary that combines the knowledge of Wikipedia articles and lexicography from Wordnet (Navigli & Ponzetto, 2012). The use of WordNet is to obtain the synset relationship between the words based on lexical and semantic relationships, while Wikipedia uses the relationship between entities on the Wikipedia page. BabelNet can shorten the time compared to PRF, because this method does not process all the articles in finding the relevant documents of the query. GloVe, skip-gram, and CBOW architectures were used for the representation of the word vectors. The proposed method was focused on the semantic AQE method to reduce the possibility of disambiguation term problems. Synonym and WordNet were used as an AQE method based on the WE technique. The purpose of using BabelNet was to investigate how it works in semantic searching to determine the documents highly similar to the original query based on Wikipedia, and not to use the knowledge of BabelNet. Based on the experiment results that used 40 queries, the average accuracy of these queries was 90%, which makes this a promising method for AQE in searching Arabic documents.

Alshalan, Alshalan, Al-Khalifa, Suwaileh, and Elsayed (2020) introduced an embedding-based QE (EQE) technique that uses the semantic representations obtained by WE for QE. Two QE methods were introduced: first, embedding-based PRF (EPRF)

Table 1. Summary of the studies which use QE approaches for the Arabic language

Author	Problem specification	QE approach	Dataset	Evaluation metrics
Hammo et al. (2007)	The ambiguity caused by the short vowels in Arabic	QE based synonymy, QE based stemming, QE through word sensing, & QE through paraphrasing	The Holy Qur'an	Precision and recall
He and Wang (2009)	The ambiguity in the user's queries	CLIR and Cross-Language Latent Semantic Indexing (CL-LSI)	Bilingual corpora	MAP
Khafajeh et al. (2010)	The vocabulary problem	VSM, Jaccard, Dice & cosine similarity	242 Arabic abstract documents	Precision and recall
Trad et al. (2012)	The ambiguity in the user's queries	Boolean Model, Vector Model, and LSI Model	Arabic & English medical corpus	Precision and recall
Shaalan et al. (2012)	User's queries are short	Expectation-Maximization (EM), Cut-off Run1, Run2, & Run3	CLEF2009	Precision and recall
Mahgoub et al. (2014)	The ambiguity in the user's queries	VSM single expanded queries & multiple expanded queries	Zad Al Ma'ad book written by the Islamic scholar Ibn Al-Qyyim	Precision, recall, F-Score, and NDCG
Nwesri and Alyagoubi (2015)	The negative effects of the Arabic text stemming the on words	BM25 & InQuery retrieval model with four different stemmers used: light10, Khoja, Buckwalter morphological analyzer, and Alstem	Arabic TREC 2001/2002	MAP, precision @10, and recall
Al-Chalabi et al. (2015)	Term mismatch problem	Manual QE & AQE using Boolean operators "AND" and "OR"	TREC & CLEF Arabic questions	Mean Reciprocal Ratio (MRR)
Abbache et al. (2016b)	Term mismatch problem	Interactive QE & AQE	The Xnh-45004 Arabic corpus	Precision, recall, F-Measure, & MAP
El Mahdaouy et al. (2019)	Term mismatch problem	LGD, SPL, LM, BM25, & PRF	Arabic TREC 2001/2002	MAP & P@10
Atwan et al. (2016)	The weakness of Arabic stop-word list and stemming algorithms	PRF and Arabic WordNet	Arabic TREC 2001	MAP, precision, and recall
Hassan and Hadi (2017)	The ambiguity in the user's queries	WSD based *LCH* , *WUP*, and *PATH* semantic measures	Zad Al Ma'ad book written by the Islamic scholar Ibn Al-Qyyim	Precision, recall, and MAP
El Mahdaouy et al. (2016)	Term mismatch problem	SMTF, LGD, & SPL models	Arabic TREC 2001/2002	MAP & P@10
Maryamah et al. (2019)	Vocabulary mismatch problem	Synonym & WordNet	Wikipedia and BabelNet	Precision & recall
El Mahdaouy et al. (2018)	Term mismatch problem	LM, LGD, SPL, and BM25	Arabic TREC 2001/2002	MAP & P@10
Beirade et al. (2019)	Vocabulary mismatch problem	Al-fanous system & semantic search engine	Holy Al Qur'an	Precision & recall
Bounhas et al. (2020)	Term mismatch problem	BM25 based different stemming algorithm	ZAD Test Collection	MAP Recall F-Score P@5 P@10 P@15 P@20
Rahman et al. (2019)	Term mismatch problem	BM25 based DBpedia and Hypernym	SemEval-2016 Task 3 CQA dataset	MAP
ALMarwi et al. (2020)	Term mismatch problem	Hybrid semantic query expansion based WE, WordNet, and TF	News collected from Al-Alam, BBC, CNN, and Al-Jazeera channels	Precision and recall
Alshalan (2020)	Term mismatch problem	Embedding-Based Query Expansion (EQE), Embedding-Based PRF (EPRF), & PRF with Embedding-Based Reranking (PRF+ERerank).	EveTAR test collection	MAP

QE, Query Expansion; Query Expansion; CLIR, Cross-Language Information Retrieval; MAP, mean average precision; VSM, Vector Space Model; LSI, Latent Semantic Indexing; NDCG, Normalized Discounted Cumulative Gain; AQE, Automatic Query Expansion; LGD, Log-Logistic Distribution; SPL, Smoothed Power-Law; LM, Language Model; PRF, Pseudo Relevance Feedback; WSD, Word-Sense Disambiguation; SMTF, Semantically Enhanced Term Frequency; CQA, community question-answering; WE, Word Embedding; TF, term frequency.

technique, which incorporates the WE similarity scores to weight the additional query terms, and the second, embedding-based re-ranking to re-rank the top documents retrieved by PRF. Three different test collections of EveTAR (Hasanain, Suwaileh, Elsayed, Kutlu, & Almerekhi, 2018) were employed for the evaluations. Besides this, to incorporate the distributed WE in the ad-hoc IR task, three methods were proposed: 1) EQE, where the aim is, through the retrieval stage, to add semantic-related words to the original query online; 2) EPRF, the hypothesis of which is to get candidate expansion words from the relevant tweets terms; and 3) (PRF+ERerank), PRF with Embedding-Based Reranking. This method aims to re-rank the PR documents by leveraging the pre-trained embedding model. Based on the experiment results, the proposed EQE technique under-performed the baseline PRF technique. The second technique, based on PRF, significantly exceeded the baseline and other approaches.

Further, a hybrid semantic QE method for AIR was offered by ALMarwi, Ghurab, and Al-Baltah (2020). The proposed method had the advantage of the statistical and semantic approaches. Hence, the common limitation of the statistical approaches was addressed by authorizing similar semantic terms to contribute to the scoring function. The candidate expansion terms related to the meaning of the whole query were generated. To calculate the term weights and know the degree of importance of each term, three pieces of evidence were considered: the WE, the TF, and WordNet. In order to avoid query drift, the semantic filtering method, particle swarm optimization, was utilized to remove the noise from the candidate expansion terms. The experimental results demonstrated that the proposed technique improved the effectiveness of the IR system in terms of precision values.

Rahman, Hisamoto, and Duh (2019) explored various QE methods to overcome the issues of question re-ranking in a CLIR setting for community question-answering platforms. The proposed methods supposed that the mistranslations caused by CLIR were often irrelevant to the original query concepts and that using similar words for the expansion might retrieve the needed related terms. The QE method introduced was based on WE, DBpedia concepts linking, and Hypernym to expand the query in question-question re-ranking tasks. The query translation followed by QE was adopted to provide the initial query. Moreover, using external resources to expand each word, a search engine such as ElasticSearch was employed to match those terms against the existing questions. The experiment results revealed the proposed QE technique outperforming the current methods on Cross-Language (CL) question re-ranking. Table 1 summarises the studies in the literature review, based on the problem of the research, QE approach, dataset used, and the evaluation metrics.

Fig. 2 illustrates statistics of the reviewed papers based on their publication year.

Fig. 3 shows the reliable resources (such as IEEE, Springer, and ScienceDirect) of the published articles, where the top resource in our review is Springer. Our survey covered studies in the last ten years, except for two pieces of research in 2007 and 2009; they have been selected because they are too related in this review.

Fig. 4 illustrates statistically the percentage of the Arabic datasets used in the reviewed studies. In contrast, Fig. 5 showed the percentage of statistical problem identification in the studies of that review, where most of the researchers focused on addressing the term mismatch problem between the query terms and the documents.

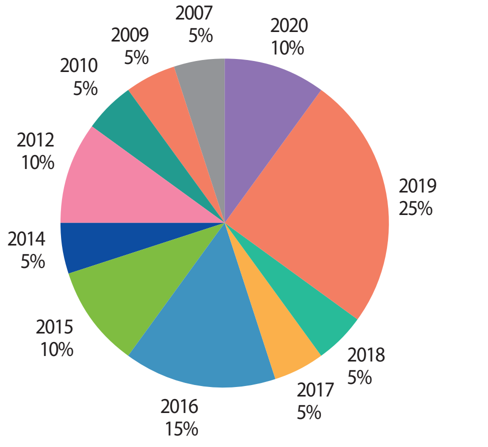

Fig. 2. Percentage of the surveyed papers by publication year.

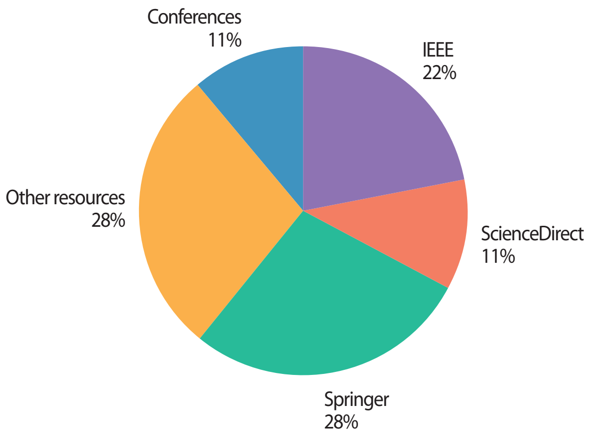

Fig. 3. Statistics of the surveyed papers, according to publishers.

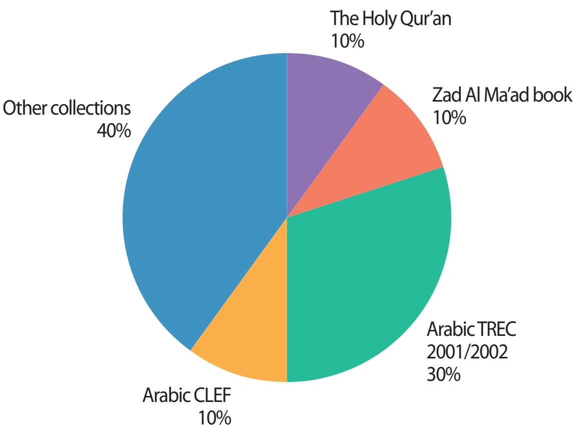

Fig. 4. Percentage of the data collections used in the literature review.

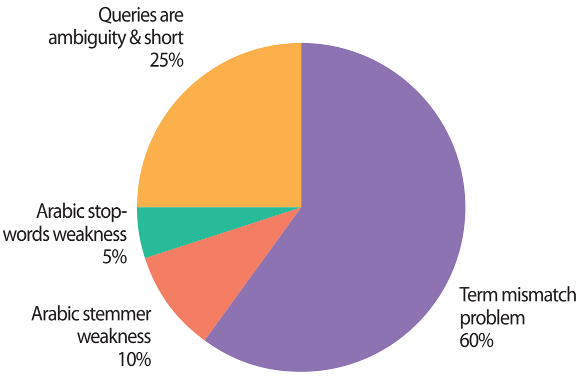

Fig. 5. Percentage of problem identification based on the literature review.

5. DISCUSSION

Based on our review presented above, QE methods may enhance the retrieval performance by a certain percentage. However, some of these methods may not consider the relationship between query terms and their actual meaning and cannot fully show the user's intention; or, even if it is able to improve retrieval performance, it still has its disadvantages and drawbacks. Further, it indicates that QE for Arabic text is incapable even of accomplishing perfect, significant results due to the complexity and high morphology of Arabic. Therefore, works on QE for Arabic are still subject to additional investigation and research. Among the unused approaches is the usage of large scale (big data) for backing up QE. The following paragraphs discuss each QE technique, based on our review.

- **PRF** is a vital and common method for improving retrieval accuracy. The importance of the PRF concept is its presumption that a restricted number of top-ranked documents in the initial retrieval results are significant, from which it selects connected terms to improve the query representation via QE. Although conventional PRF methods generally improve retrieval performance on average, they are not entirely adequate to assist some queries and upset other queries, restricting their value in real retrieval functions (Collins-Thompson, 2008); and this occurs due to continual dependence on the top-ranked documents, which may be irrelevant and can introduce noise into the feedback process. Lastly, in this case, the QE process will be upset (Lee, Croft, & Allan, 2008).
- **A stemming algorithm** is an additional method employed to expand queries. This method has been proven to overtake the conventional method for QE. Further, stemming reduces the words to their roots, where *light stemmer* employs some set of rules to remove the suffixes or prefixes. Arabic is different from other Indo-European languages in terms of its syntax, morphology, and semantics, and has broken plurals (inflections). However, this kind of stemming algorithm fails with broken plurals completely and conflates the words with a different meaning under one index term. This happens regularly in Arabic in applying stems, and it develops more while selecting roots to index the collection (Nwesri & Alyagoubi, 2015).

  A superior root-based Arabic stemmer, *Khoja's stemmer*, is another stemming algorithm proposed by (Khoja, 2001). The rule of this algorithm is to remove suffixes, infixes, and prefixes and use pattern matching to extract the roots. However, this algorithm suffers a practical issue, relating mostly to nouns and foreign names like those of countries and cities (Larkey, Ballesteros, & Connell, 2007). Finally, the central issue of stemming Arabic text is the lack of knowledge of the stemmer in the word's lexical category (e.g., noun, verb, and preposition).
- **CLIR** is another method which can be adopted to enhance retrieval performance. This procedure of supplementing connected terms to the interpreted query helps improve precision and recall, as more important documents can be recognized and retrieved. However, the process of interpreting or translating a query from one language to another often presents ambiguity. Another concern is to translate the query from Arabic to English to get the results and then retranslate it to Arabic, which is time-consuming.

Trad et al. (2012) compared AWN Ontology with English WordNet Ontology, finding no actual benefit concerning Arabic development in any of the assessment measures adopted, notwithstanding the employment of specialized corpus. In other words, using A WN Ontology is weak, compared to its English version. On the other hand, the main issue of the pre-processing task for Arabic QE is that there is no general standard stop-word list. Most of the researchers, such as Larkey et al. (2007) and Larkey and Connell (2001), used the list of stop-word Lemur Toolkit, created by Khoja (2001) and containing only 168 words.

- **Ontology and Thesaurus**

Using ontology or other similar resources such as WordNet seems useful to improve retrieval effectiveness in some cases. However, based on various studies, using ontology does not enhance retrieval effectiveness continually. Further, expanding the query based on ontology may cause two main issues: first, the global/general ontology often introduces ambiguity, with the possibility of insertion of noise terms into the queries. Second, global ontology cannot pick the specific properties of a domain (Bhogal, MacFarlane, & Smith, 2007; Raza, Mokhtar, Ahmad, Pasha, & Pasha, 2019).

Fig. 6 shows each reference with its number of citations. In light of the aforesaid review, one can conclude that the most cited papers are for Hammo et al. (2007), Shaalan et al. (2012), Mahgoub et al. (2014), and Abbache et al. (2016b), where all of these references have more than twenty-five citations.

6. CONCLUSION

As mentioned earlier, users' queries are usually short, reflecting weakness of information knowledge and hence incapable of expressing their information needs. Further, the vocabulary term mismatch between the query terms and documents in the corpus is the most common issue in the IR system; hence the researchers' attempt to fill this gap by expanding the original query by adding some useful terms to it.

One of the most useful and vital ways to address the above issues is AQE. The main aim of QE methods is to improve retrieval performance by adding related terms to the initial query. As already mentioned, AQE approaches can be classified into statistical approaches, semantical approaches, and hybrid approaches (Raza et al., 2018). Statistical approaches such as TF and Term Co-Occurrence focus on computing the number of terms occurring in the documents and selecting the most frequent word. This, however, does not work well with short text (Bhogal et al., 2007).

The statistical QE approach of document analysis provides useful expansion terms but depends on the quality of the document corpus. One of the major limitations of the semantical QE approaches is that it may have more relevance to the document collection than to the searcher's query. Moreover, the terms in the semantical QE approaches are ranked based

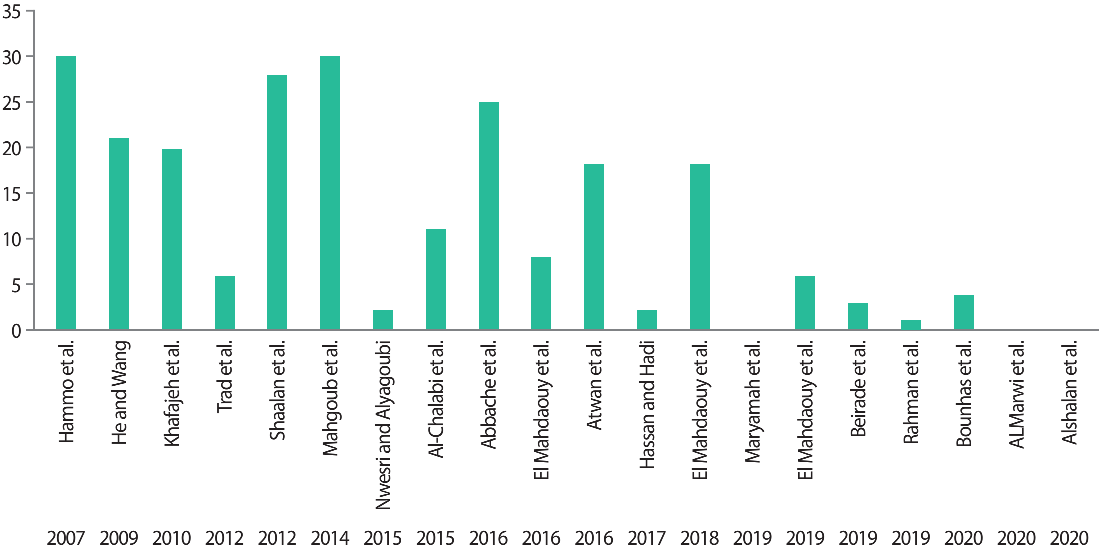

Fig. 6. Number of citations for each reference. Adapted from Al-Ghuribi and Noah. IEEE Access 2019;7:169446-169468.

on a documents database, not taking into account the intent of the users, as proved by recent studies (Carpineto et al., 2001; Carpineto & Romano, 2012).

On the other hand, the semantic approaches which use a thesaurus or ontology, or those based on the WE, address the ambiguity issue by relying on the meaning of the query in context for the expansion purpose.

Semantic QE approaches can be more effective in disambiguation of the query terms, compared to the statistical QE approaches (Pinto, Martinez, & Perez-Sanjulian, 2008). Although the semantic QE approaches have proved successful in improving retrieval performance, they have some drawbacks. The main issue is the difficulties in building a knowledge structure with accurate concepts and relationships, which requires an expert or a sophisticated knowledge system (Bhogal et al., 2007). Another issue is the drifting of the query from the real topic during the expansion process, due to the addition of non-relevant terms.

However, all the previous approaches depend on the individual terms in the expansion process. Expanding highly morphological texts like Arabic based on individual query terms may hurt the expansion process or change the meaning of the query.

Moreover, until now, no technique has been able to solve the vocabulary mismatch problem in IR fully. Some of the QE methods may ignore the relationship between query terms and their actual meaning and cannot fully express the user's intention; thus, using such a term in the query does not reflect the user's intention. Further, it confirms the inability of the semantic and statistical QE approaches for Arabic text to find the proper expansion terms without shortcomings, resulting in persisting issues relating to the expansion process. Therefore, works on QE for Arabic should be subjected to additional investigation and research.

REFERENCES

Abbache, A., Meziane, F., Belalem, G., & Belkredim, F. Z. (2016a). Arabic query expansion using WordNet and association rules. *International Journal of Intelligent Information Technologies*, 3(12), 4.

Abbache, A., Meziane, F., Belalem, G., & Belkredim, F. Z. (2016b). Arabic query expansion using WordNet and association rules. *Information retrieval and management: Concepts, methodologies, tools, and applications* (pp. 1239-1254). IGI Global.

Abdelali, A. (2004). Localization in modern standard Arabic. *Journal of the American Society for Information Science and Technology*, 55(1), 23-28.

Abderrahim, M. A., Dib, M., Abderrahim, M. E. -A., & Chikh, M. A. (2016). Semantic indexing of Arabic texts for information retrieval system. *International Journal of Speech Technology*, 19(2), 229-236.

Abouenour, L., Bouzouba, K., & Rosso, P. (2010). An evaluated semantic query expansion and structure-based approach for enhancing Arabic question/answering. *International Journal on Information and Communication Technologies*, 3(3), 37-51.

Abu-Errub, A. (2014). Arabic text classification algorithm using TFIDF and Chi Square measurements. *International Journal of Computer Applications*, 93(6), 40-45.

Al-Chalabi, H., Ray, S., & Shaalan, K. (2015, April 17-20). Semantic based query expansion for Arabic question answering systems. *2015 First International Conference on Arabic Computational Linguistics (ACLing)*. IEEE.

Al-Ghuribi, S. M., & Noah, S. A. M. (2019). Multi-criteria review-based recommender system-the state of the art. *IEEE Access*, 7, 169446-169468.

Aljlayl, M., & Frieder, O. (2002, November 6-8). On Arabic search: Improving the retrieval effectiveness via a light stemming approach. *Proceedings of the Eleventh International Conference on Information and Knowledge Management* (pp. 340-347). Association for Computing Machinery.

ALMarwi, H., Ghurab, M., & Al-Baltah, I. (2020). A hybrid semantic query expansion approach for Arabic information retrieval. *Journal of Big Data*, 7(1), 39.

ALMasri, M., Berrut, C., & Chevallet, J. -P. (2016, March 20-23). A comparison of deep learning based query expansion with pseudo-relevance feedback and mutual information. *ECIR 2016: Advances in Information Retrieval* (pp. 709-715). Springer.

Alshalan, S., Alshalan, R., Al-Khalifa, H., Suwaileh, R., & Elsayed, T. (2020, November 7-9). Improving Arabic microblog retrieval with distributed representations. *The Information Retrieval Technology: 15th Asia Information Retrieval Societies Conference, AIRS 2019* (pp. 185-194). Springer.

Amati, G., & Van Rijsbergen, C. J. (2002). Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Transactions on Information Systems*, 20(4), 357-389.

Atwan, J., & Mohd, M. (2017). Arabic query expansion: A review. *Asian Journal of Information Technology*, 16(10), 754-770.

Azad, H. K., & Deepak, A. (2019). Query expansion techniques for information retrieval: A survey. *Information Processing*

& Management, 56(5), 1698-1735.

Baeza-Yates, R., de Vries, A. P., Zaragoza, H., Cambazoglu, B. B., Murdock, V., Lempel, R., & Silvestri, F. (2012, April 1-5). Advances in information retrieval. *34th European Conference on IR Research, ECIR 2012*. Springer-Verlag Berlin Heidelberg.

Ballesteros, L., & Croft, W. B. (1997, July 10). Phrasal translation and query expansion techniques for cross-language information retrieval. *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. Association for Computing Machinery.

Batita, M. A., & Zrigui, M. (2018, January 8-12). Derivational relations in Arabic Wordnet. *The 9th Global WordNet Conference GWC*. Nanyang Technological University.

Beirade, F., Azzoune, H., & Zegour, D. E. (2019). Semantic query for Quranic ontology. *Journal of King Saud University-Computer and Information Sciences*, 31(2), 135-274.

Belkin, N. J. (1980). Anomalous states of knowledge as a basis for information retrieval. *Canadian Journal of Information Science*, 5(1), 133-143.

Belkredim, F. Z., & El Sebai, A. (2009). An ontology based formalism for the Arabic language using verbs and their derivatives. *Communications of the IBIMA*, 11(5), 44-52.

Berget, G., & Sandnes, F. E. (2015). Searching databases without query-building aids: Implications for dyslexic users. *Information Research: An International Electronic Journal*, 20(4), n4.

Bhogal, J., MacFarlane, A., & Smith, P. (2007). A review of ontology based query expansion. *Information Processing & Management*, 43(4), 866-886.

Białecki, A., Muir, R., & Ingersoll, G. (2012, August 20). Apache Lucene 4. *Proceedings of the SIGIR 2012 Workshop on Open Source Information Retrieval* (pp. 17-24). Department of Computer Science, University of Otago.

Bounhas, I., Soudani, N., & Slimani, Y. (2020). Building a morpho-semantic knowledge graph for Arabic information retrieval. *Information Processing & Management*, 57(6), 102124.

Carpineto, C., de Mori, R., Romano, G., & Bigi, B. (2001). An information-theoretic approach to automatic query expansion. *ACM Transactions on Information Systems*, 19(1), 1-27.

Carpineto, C., & Romano, G. (2012). A survey of automatic query expansion in information retrieval. *ACM Computing Surveys*, 44(1), 1.

Clinchant, S., & Gaussier, E. (2013, September 29-October 2). A theoretical analysis of pseudo-relevance feedback models. *Proceedings of the 2013 Conference on the Theory of Information Retrieval* (pp. 6-13). Association for Computing Machinery.

Collins-Thompson, K. (2008, December 15-17). Estimating robust query models with convex optimization. *Neural Information Processing Systems 21* (NIPS 2008). Curran Associates Inc.

Croft, W. B., & Harper, D. J. (1979). Using probabilistic models of document retrieval without relevance information. *Journal of Documentation*, 35(4), 285-295.

Cui, H., Wen, J. -R., Nie, J. -Y., & Ma, W. -Y. (2002, May 7). Probabilistic query expansion using query logs. *Proceedings of the 11th international conference on World Wide Web* (pp. 325-332). Association for Computing Machinery.

Dalton, J., Naseri, S., Dietz, L., & Allan, J. (2019, April 14-18). Local and global query expansion for hierarchical complex topics. *European Conference on Information Retrieval* (ECIR 2019) (pp. 290-303). Springer.

Diaz, F., Mitra, B., & Craswell, N. (2016, August 7-12). Query expansion with locally-trained word embeddings. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* (pp. 367-377). Association for Computational Linguistics.

El Mahdaouy, A., El Alaoui, S. O., & Gaussier, E. (2016, October 24-26). Semantically enhanced term frequency based on word embeddings for Arabic information retrieval. *2016 4th IEEE International Colloquium on Information Science and Technology* (CiSt). IEEE.

El Mahdaouy, A., El Alaoui, S. O., & Gaussier, E. (2018). Improving Arabic information retrieval using word embedding similarities. *International Journal of Speech Technology*, 21, 121-136.

El Mahdaouy, A., El Alaoui, S. O., & Gaussier, E. (2019). Word-embedding-based pseudo-relevance feedback for Arabic information retrieval. *Journal of Information Science*, 45(4), 429-442.

Fang, H., & Zhai, C. (2006, August 2). Semantic term matching in axiomatic approaches to information retrieval. *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 115-122). Association for Computing Machinery.

Farghaly, A., & Shaalan, K. (2009). Arabic natural language processing: Challenges and solutions. *ACM Transactions on Asian Language Information Processing*, 8(4), 14.

Farrar, D., & Hayes, J. H. (2019, May 27-27). A comparison of stemming techniques in tracing. *2019 IEEE/ACM 10th International Symposium on Software and Systems Traceability* (SST). IEEE.

Gruber, T. R. (1995). Toward principles for the design of ontologies used for knowledge sharing? *International Journal of Human-Computer Studies*, 43(5-6), 907-928.

Haddad, B. (2013). Cognitive aspects of a statistical language model for Arabic based on associative probabilistic Root-PATtern relations: A-APRoPAT. Infocommunications Journal, 2013(3).

Hammo, B., Sleit, A., & El-Haj, M. (2007, May 7-9). Effectiveness of query expansion in searching the Holy Quran. *The Second International Conference on Arabic Language Processing CITALA'07* (pp. 7-10). UNSPECIFIED.

Han, L., & Chen, G. (2009). HQE: A hybrid method for query expansion. *Expert Systems with Applications*, 36(4), 7985-7991.

Harris, Z. (1968). *Mathematical structures of language*. New York: John Wiley and Sons.

Hasanain, M., Suwaileh, R., Elsayed, T., Kutlu, M., & Almerekhi, H. (2018). EveTAR: Building a large-scale multi-task test collection over Arabic tweets. *Information Retrieval Journal*, 21(4), 307-336.

Hassan, A. K. A., & Hadi, M. J. (2017). Automatic query expansion for Arabic text retrieval. *Iraqi Journal of Science*, 58(4C), 2447-2457.

Hattab, M., Haddad, B., Yaseen, M., Duraidi, A., & Shmais, A. A. (2009, April 21-23). Addaall Arabic search engine: Improving search based on combination of morphological analysis and generation considering semantic patterns. *2nd International Conference on Arabic Language Resources and Tools* (pp. 159-162). MEDAR consortium.

He, D., & Wang, J. (2009). Cross-language information retrieval. In A. Goker, & J. Davies (Eds.), *Information retrieval: Searching in the 21st century* (pp. 233-254). Wiley Telecom.

Jiang, J. J., & Conrath, D. W. (1996). A concept-based approach to retrieval from an electronic industrial directory. *International Journal of Electronic Commerce*, 1(1), 51-72.

Khafajeh, H., Yousef, N., & Kanaan, G. (2010, April 12-13). Automatic query expansion for Arabic text retrieval based on association and similarity thesaurus. *European, Mediterranean & Middle East Conference on Information Systems* (EMCIS). Brunel University.

Khoja, S. (2001, June 2-7). 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* (NAACL2001). Carnegie Mellon University.

Kusner, M., Sun, Y., Kolkin, N., & Weinberger, K. (2015, July 7-9). From word embeddings to document distances. *Proceedings of the 32nd International Conference on International Conference on Machine Learning* (pp. 957-966). JMLR.org.

Kuzi, S., Shtok, A., & Kurland, O. (2016, October 24-28). Query expansion using word embeddings. *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management* (pp. 1929-1932). Association for Computing Machinery.

Larkey, L. S., Ballesteros, L., & Connell, M. E. (2007). Light stemming for Arabic information retrieval. In A. Soudi, A. van den Bosch, & G. Neumann (Eds.), *Arabic computational morphology* (pp. 221-243). Springer.

Larkey, L. S., & Connell, M. E. (2001, November 13-16). *Arabic Information Retrieval at UMass in TREC-10*. Paper presented at the Tenth Text REtrieval Conference (TREC 2001), Gaithersburg, MA, USA.

Lee, K. S., Croft, W. B., & Allan, J. (2008, July 27-28). A cluster-based resampling method for pseudo-relevance feedback. *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 235-242). Association for Computing Machinery.

Lin, J., & Murray, G. C. (2005, August 9-11). Assessing the term independence assumption in blind relevance feedback. *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 635-636). Association for Computing Machinery.

Lindén, K., & Piitulainen, J. (2004, August 29). Discovering synonyms and other related words. *Proceedings of CompuTerm 2004: 3rd International Workshop on Computational Terminology* (pp. 63-70). COLING.

Mahgoub, A., Rashwan, M., Raafat, H., Zahran, M., & Fayek, M. (2014, October 25). Semantic query expansion for Arabic information retrieval. *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)* (pp. 87-92). Association for Computational Linguistics.

Mandala, R., Takenobu, T., & Hozumi, T. (1998, August 16). The use of WordNet in information retrieval. *Usage of WordNet in Natural Language Processing Systems* (pp. 31-37). COLING.

Maryamah, M., Arifin, A. Z., Sarno, R., & Morimoto, Y. (2019). Query expansion based on Wikipedia word embedding and BabelNet method for searching Arabic documents. *International Journal of Intelligent Engineering & System*, 12(5), 202-213.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013, December 5-10). Distributed representations of words and phrases and their compositionality. *Proceedings of the 26th International Conference on Neural Information Processing Systems* (pp. 3111-3119). Curran Associates Inc.

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., & Miller, K. J. (1990). Introduction to WordNet: An on-line lexical database. *International Journal of Lexicography*, 3(4), 235-244.

Navigli, R., & Ponzetto, S. P. (2012). BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193, 217-250.

Nwesri, A. F. A., & Alyagoubi, H. A. (2015, September 1-4). Applying Arabic stemming using query expansion. *2015 26th International Workshop on Database and Expert Systems Applications (DEXA)*. IEEE.

Ooi, J., Ma, X., Qin, H., & Liew, S. (2015, August 19-21). A survey of query expansion, query suggestion and query refinement techniques. *2015 4th International Conference on Software Engineering and Computer Systems (ICSECS)*. IEEE.

Pal, D., Mitra, M., & Datta, K. (2014). Improving query expansion using WordNet. *Journal of the Association for Information Science and Technology*, 65(12), 2469-2478.

Pennington, J., Socher, R., & Manning, C. (2014, October 25-29). GloVe: Global vectors for word representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1532-1543). Association for Computational Linguistics.

Pinto, F. J., Martinez, A. F., & Perez-Sanjulian, C. F. (2008). Joining automatic query expansion based on thesaurus and word sense disambiguation using WordNet. *International Journal of Computer Applications in Technology*, 33(4), 271-279.

Pôssas, B., Ziviani, N., Meira Jr, W., & Ribeiro-Neto, B. (2005). Set-based vector model: An efficient approach for correlation-based ranking. *ACM Transactions on Information Systems*, 23(4), 397-429.

Qiu, Y., & Frei, H. -P. (1993, July 20-22). Concept based query expansion. *Proceedings of the 16th annual international ACM SIGIR conference on Research and Development in Information Retrieval* (pp. 160-169). Association for Computing Machinery.

Rahman, M. M., Hisamoto, S., & Duh, K. (2019). Query Expansion for Cross-Language Question Re-Ranking. *CoRR,* abs/1904.07982.

Raza, M. A., Mokhtar, R., & Ahmad, N. (2018). A survey of statistical approaches for query expansion. *Knowledge and Information Systems*, 61, 1-25.

Raza, M. A., Mokhtar, R., Ahmad, N., Pasha, M., & Pasha, U. (2019). A taxonomy and survey of semantic approaches for query expansion. *IEEE Access*, 7, 17823-17833.

Shaalan, K., Al-Sheikh, S., & Oroumchian, F. (2012, October 12-15). Query expansion based-on similarity of terms for improving Arabic information retrieval. *7th International Conference on Intelligent Information Processing (IIP)* (pp. 167-176). Springer.

Sharma, D., Pamula, R., & Chauhan, D. S. (2019). A hybrid evolutionary algorithm based automatic query expansion for enhancing document retrieval system. *Journal of Ambient Intelligence and Humanized Computing*, 1-20.

Singh, J., & Sharan, A. (2017). A new fuzzy logic-based query expansion model for efficient information retrieval using relevance feedback approach. *Neural Computing and Applications*, 28(9), 2557-2580.

Sordoni, A., Bengio, Y., & Nie, J. -Y. (2014, July 27-31). Learning concept embeddings for query expansion by quantum entropy minimization. *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence* (pp. 1586-1592). AAAI Press.

Trad, R., Koroni, R., Mustafa, H., & Almaghrabi, A. (2012, November 21-23). Evaluating Arabic WordNet Ontology by expansion of Arabic queries using various retrieval models. *2012 10th International Conference on ICT and Knowledge Engineering*. IEEE.

Vechtomova, O. (2009). Query expansion for information retrieval. In L. Liu, M. T. Özsu (Eds.), *Encyclopedia of database systems* (pp. 2254-2257). Boston: Springer.

Vulić, I., & Moens, M. -F. (2015, August 9-13). Monolingual and cross-lingual information retrieval models based on (bilingual) word embeddings. *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 363-372). Association for Computing Machinery.

Wang, X., Lai, G., & Liu, C. (2009). Recovering relationships between documentation and source code based on the characteristics of software engineering. *Electronic Notes in Theoretical Computer Science*, 243, 121-137.

White, R. W., & Horvitz, E. (2015). Belief dynamics and biases in web search. *ACM Transactions on Information Systems*, 33(4), 18.

Zaïane, O. R., & Antonie, M. -L. (2002, January 13). Classifying text documents by associating terms with text categories. *Proceedings of the 13th Australasian Computer Science Communications* (pp. 215-222). Australian Computer Society.

Zamani, H., & Croft, W. B. (2016, September 12-16).

Embedding-based query language models. *Proceedings of the 2016 ACM International Conference on the Theory of Information Retrieval* (pp. 147-156). Association for Computing Machinery.

Zhou, G., He, T., Zhao, J., & Hu, P. (2015, July 26-31). Learning continuous word embedding with metadata for question retrieval in community question answering. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing* (pp. 250-259). Association for Computational Linguistics.

Zuccon, G., Koopman, B., Bruza, P., & Azzopardi, L. (2015, December 8-9). Integrating and evaluating neural word embeddings in information retrieval. *Proceedings of the 20th Australasian Document Computing Symposium* (pp. 1-8). Association for Computing Machinery.