JISTaP http://www.jistap.org
Journal of Information Science Theory and Practice

RESEARCH PAPER
J Inf Sci Theory Pract 9(2): 1-17, 2021

eISSN : 2287-4577 pISSN : 2287-9099

https://doi.org/10.1633/JISTaP.2021.9.2.1

Word Embeddings-Based Pseudo Relevance Feedback Using Deep Averaging Networks for Arabic Document Retrieval

Yasir Hadi Farhan*
Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, Selangor, Malaysia
E-mail: yasir.hadi87@yahoo.com

Shahrul Azman Mohd Noah
Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, Selangor, Malaysia
E-mail: shahrul@ukm.edu.my

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

Jaffar Atwan
Prince Abdullah Bin Ghazi, Faculty of Information Technology, Al Balqa Applied University, Salt, Jordan
E-mail: jaffaratwan@bau.edu.jo

ABSTRACT

Pseudo relevance feedback (PRF) is a powerful query expansion (QE) technique that prepares queries using the top *k* pseudo-relevant documents and choosing expansion elements. Traditional PRF frameworks have robustly handled vocabulary mismatch corresponding to user queries and pertinent documents; nevertheless, expansion elements are chosen, disregarding similarity to the original query's elements. Word embedding (WE) schemes comprise techniques of significant interest concerning QE, that falls within the information retrieval domain. Deep averaging networks (DANs) defines a framework relying on average word presence passed through multiple linear layers. The complete query is understandably represented using the average vector comprising the query terms. The vector may be employed for determining expansion elements pertinent to the entire query. In this study, we suggest a DANs-based technique that augments PRF frameworks by integrating WE similarities to facilitate Arabic information retrieval. The technique is based on the fundamental that the top pseudo-relevant document set is assessed to determine candidate element distribution and select expansion terms appropriately, considering their similarity to the average vector representing the initial query elements. The Word2Vec model is selected for executing the experiments on a standard Arabic TREC 2001/2002 set. The majority of the evaluations indicate that the PRF implementation in the present study offers a significant performance improvement compared to that of the baseline PRF frameworks.

Keywords: automatic query expansion, information retrieval, word embedding, deep averaging networks, pseudo relevance feedback, Arabic document retrieval on TREC collection

Received: March 16, 2021 **Revised:** May 5, 2021
Accepted: May 10, 2021 **Published:** June 15, 2021

***Corresponding Author:** Yasir Hadi Farhan
https://orcid.org/0000-0002-2378-9177
E-mail: yasir.hadi87@yahoo.com

1. INTRODUCTION

Search engines presently face the challenge of users' queries comprising two to three words, which are considered too little for a query (Berget & Sandnes, 2015). They are typically the consequence of users' challenges in expressing the requirement for information; these queries are not truly representative of users' information needs. The anomalous state of knowledge hypothesis by Belkin et al. (1982) suggests that query challenges are typically associated with users facing knowledge gaps. Automatic query expansion (AQE) is among the extensively-researched techniques to bridge that gap and facilitate better search queries by users (Azad & Deepak, 2019; Dalton et al., 2019).

AQE comprises global and local techniques, where the global techniques rely on a thesaurus to augment the original query instead of using retrieved results. Usually, WordNet is the most extensively used tool for identifying the new potential terms correlated with the terms of the initial query (Pal et al., 2014). On the other hand, local techniques rely on relevance-based information, where the first retrieved document are analysed and utilised to identify the most appropriate terms for augmenting the initial query (Miyanishi et al., 2013; Takeuchi et al., 2017). Pseudo relevance feedback (PRF) is a technique whereby it is presupposed that the *k* initial top documents yielded by the search are most relevant to identify new elements from the *k* documents; it is assumed to be a better expansion method since it facilitates automatic relevance feedback.

PRF is a potent query expansion (QE) technique to handle vocabulary mismatch between documents and user queries, which are then extended by identifying associated terms from the documents in the top results, also known as top pseudo-relevant documents (Farhan et al., 2020). Typically, PRF methods offer adequate performance (Carpineto et al., 2001; Clinchant & Gaussier, 2013); however, they rely on expansion term distribution pertaining to the pseudo-relevant document set, where the similarity shared by the original query elements and the expansion elements is not strongly considered. Fundamentally, in an ideal scenario, expansion elements should be identified considering their similarity to the query terms and their distribution pertaining to the pseudo-relevant document set. Several studies have proposed similar recommendations (Atwan et al., 2016; El Mahdaouy et al., 2019; Montazeralghaem et al., 2016).

Nevertheless, some of the discussed methodologies suffer from inherent challenges, such as the WordNet and other ontological databases returning keywords having several meanings; hence, QE should be performed only after word disambiguation (Farhan et al., 2020).

AQE research is of great interest to academicians who intend to utilise word embedding (WE) for semantic modelling to interpret text (ALMasri et al., 2016; Diaz et al., 2016; Roy et al., 2016) and handle the problems specified previously.

Under the natural language processing (NLP) domain, the umbrella term WE refers to modelling methods, where every term is denoted using a real-number vector, which means single-dimension mathematical WE is employed in a continuous vector space having substantially lower dimensions, where the embedding frameworks depend on proximity for corpus training. Word2Vec provides the skip-gram and continuous bag-of-words (CBOW) architectures for training the WE framework. CBOW uses proximal words for predicting the target, whereas skip-gram relies on the target to identify proximal words.

Nevertheless, considering the similar contexts shared by semantic and syntactically similar words, the objective of such training is to identify terms sharing syntactic and semantic similarity. Consequently, the *N* proximal terms may be combined, or neighbouring query-term techniques are employed for expansion in several AQE techniques (Roy et al., 2016).

1.1. The Need for Study

A majority of the AQE techniques that use WE rely on a single term to identify potentially identical terms, and fundamentally, the choice of candidate terms does not depend on the terms specific to the initial query. Nevertheless, we asserted that the need to enhance AQE effectiveness requires the query's semantic aspects to be designed entirely considering the potential vocabulary terms. Subsequently, the quality of the terms associated with queries can be enhanced; moreover, new terms having semantic correlation and high relevance can be added. Deep averaging networks (DANs) comprise a straightforward deep unordered framework that identifies the average WE in a piece of text, and subsequently, is classified using multiple linear stages.

Recently, WE has witnessed several developments for use in information retrieval (IR); however, representation concerning Arabic information retrieval (AIR) is being assessed. The extensive and complex structure of Arabic is among the extensively researched aspects concerning AIR frameworks, and despite its significance, Arabic lacks

 https://doi.org/10.1633/JISTaP.2021.9.2.1

sufficient NLP recognition (Alsmearat et al., 2014; Faqeeh et al., 2014). The fundamental challenge concerning Arabic is the scarcity of ontological or thesaurus knowledge-bases (Mohsen et al., 2018). Additionally, limited research has been conducted on the NLP side for diacritical Arabic compared to other languages such as English (Farghaly & Shaalan, 2009). To improve its efficacy, further studies are required due to weaknesses in conventional IR processes and techniques, particularly the Arabic system limitations.

Consequently, most AIR-specific research emphasises assessing or contrasting word-stemming methods (Abu El-Khair, 2007; Ben Guirat et al., 2016; Darwish & Mubarak, 2016; Larkey et al., 2002; Mustafa et al., 2008) that assess queries or documents to determine common-stemmed words used as the metric for ranking documents.

1.2. The Objective

Hence, we propose that the string-embedding method DANs be used to enhance AQE efficacy for Arabic, where the query vectors' average can be utilised with the PRF method to produce the potential expansion terms. Our objective is to enhance Arabic text retrieval by using the proposed DANs-based PRF technique, where the average vector of each query term's vectors has been used to find the candidate expansion vectors.

The present study uses the WE technique Word2Vec for training purposes, and the Okapi BM25 probabilistic framework has been used along with V2Q and EQE1 approaches for comparisons.

Furthermore, the proposed framework incorporates the BM25 probabilistic scheme along with EQE1 suggested by Zamani and Croft (2016) and V2Q formulated by Fernández-Reyes et al. (2018).

1.3. The Hypothesis

The present paper hypothesises that the DANs-based PRF technique, which utilises the similarity of the WE in generating the candidate expansion vectors, may be able to address the AIR term mismatch problem; consequently, AQE Arabic text retrieval performance can be enhanced. The *k* top documents identified during PRF are utilised to form the expansion vectors of the candidate terms.

2. RELATED WORKS

IR relates to storing, accessing, organising, and presenting information. IR systems are primarily used to offer users ease of access to information. Typically, users present their information requirements using queries that are processed by IR systems to provide the most relevant information.

There are several challenges during an interaction between IR systems and users; vocabulary mismatch is one of the significant ones (Carpineto & Romano, 2012; Farhan et al., 2020). Researchers have responded to this challenge by proposing several solutions including AQE. Schemes based on AQE augment the original query by automatically including new terms considering that the IR system's accuracy will be increased (Abbache et al., 2016). WE is among the leading AQE techniques that have gained widespread attention (ALMasri et al., 2016; Diaz et al., 2016; El Mahdaouy et al., 2019; Roy et al., 2016).

As specified previously, WE is an umbrella term from the NLP domain; it comprises feature learning methods and language modelling. Semantic parsing allows identifying the meaning of the text; consequently, the natural language can be interpreted. WE is one among the semantic vector space frameworks where real number vectors are employed to denote words in the corpus. The words are indicated using distributed or local representation. Local representation is also referred to as sparse representation. It uses the statistical presence of words while disregarding their potential associations. Distributed representation is based on the hypothesis that word representation in the WE vector space is similar for words sharing a context (Bengio, 2009; Kim et al., 2017; Turney & Pantel, 2010).

For instance, "ذكر" (male) and "انثى" (female) are words that share a similar context; consequently, there is a likelihood that the corresponding WE vector space distributions will be proximal. The works of researchers who have employed the WE vector space for word representation are discussed below.

Diaz et al. (2016) introduced an AQE technique by employing topic-guided local embeddings that were obtained from the documents retrieved after query submission. Document scoring requires using the Kullback-Leibler divergence (KLD) language framework. The extracted documents were re-ranked after computing the local embeddings corresponding to the queries. The results indicated that local embeddings provided superior results.

ALMasri et al. (2016) contrasted numerous expansion techniques using deep-learning derived vector representations. The resulting vectors comprise words such as "taxi" and "driver" that have a similar context (Mikolov et al., 2013b). Hence, the corpus comprises terms that are denoted using vectors of specific dimensions. Therefore, the similarity of the two terms is determined by normalising

the cosine similarity corresponding to the two vectors. The skip-gram technique was used to conduct assessments on four CLEF sets. It was established that deep-learning-based high-quality vector characterisation leads to a significant performance improvement compared to baseline language frameworks that lack expansion. Vector-based models were also superior to mutual information and PRF models (Manning et al., 2008).

Roy et al. (2016) suggested a relevance feedback framework that integrates semantic associations by utilising the terms compositionality and matching them with word vectors. Rather than implementing an ad-hoc method for determining the co-occurrence between the query terms and those present in the documents, kernel density estimation (KDE) is employed to methodically ascertain the co-occurrence. The objective is to formulate a systematic relevance feedback framework that offers a straightforward technique to integrate term-specific compositionality and semantic associations. The suggested technique considers query WE as data elements that are surrounded by kernel functions so that a density estimator can be formulated. This estimator attempts to compute the probability density expression that produces the detected query word vectors. Subsequently, KLD may be employed corresponding to the density function for document ranking determination. Further, this technique facilitates the use of term composition during QE. The proposed feedback scheme is assessed using several typical IR test sets, namely Web track test and TREC ad-hoc collections. The analysis indicates that the suggested KDE-based relevance feedback technique provides a statistically significant increase in relevance compared to model-based techniques that rely on quantitative co-occurrence of words and query elements in the top-ranked results.

Aklouche et al. (2018) proposed a QE technique that used NLP along with the Word2Vec toolkit (Mikolov et al., 2013a). In this technique, the TREC Washington Post Corpus is employed to train a CBOW and skip-gram to obtain fresh semantic elements associated with the initial query. Further, the reweighting technique and term selection are used to determine their effects on the retrieval performance. Vector similarity comparisons have been drawn using Euclidean distance. The vector selection is performed using the complete query or its constituent terms. Experimental observations indicated that optimal results were obtained using query reweighting, which was the best technique compared to the others. The retrieval performance is impacted when all the expanded query terms are set to the same weight.

El Mahdaouy et al. (2019) suggested a QE technique that uses the WE semantic similarities and integrates it with IR frameworks such as the Log-Logistic distribution Model, Language Model (LM), Okapi BM25, and smoothed power-law Model (SPL) to enhance Arabic text-retrieval efficacy. The present work proposes CBOW, skip-gram, and the Global Vector (GloVe) model, which are neural WE methods. The present scheme intends to handle term discrepancy by employing a low-dimensional vector space to position words. The suggested method integrates similarity and scoring expression. The initial query is reformulated using resembling terms identified during translation. These resembling terms comprise the most relevant words obtained using the vocabulary corpus. Fang and Zhai (2006) formulated semantic term matching constraints (STMCs) to assess the suggested augmentation. STMC comprises a scheme that facilitates integrating semantic comparability and regulating the weights corresponding to the initial query's words and the matching words. The present study proposes two schemes: The first comprises an assessment of the retrieved documents to identify potentially similar words. The second scheme identifies similar words for the entire corpus. The outcomes indicate that the suggested extension schemes offer a significant performance improvement over the baseline bag-of-words techniques. Further, the suggested method is superior to the three IR language frameworks founded on WE, namely the neural translation language model formulated by Zuccon et al. (2015), the LM-based WE, and the generalised language model (GLM) suggested by Ganguly et al. (2015). Hence, word vector indication is an appropriate beginning point for determining the semantic similarity shared by query terms and potential expansion candidates. Presently used QE methods proceed assuming that all query terms can identify the most useful expansion candidates (Esposito et al., 2020).

Nevertheless, considering that the query context is disregarded when an expansion is conducted, the critical factor is to create a context applicable for retrieving the relevant terms. Several studies such as that of Roy et al. (2016) have tried to address this question; however, the issue has not been fully resolved. Zamani and Croft (2016) and Fernández-Reyes et al. (2018) are two leading works that have attempted to address this challenge. Fernández-Reyes et al. (2018) suggested that terms be identified if they share a strong resemblance to the original query terms. It was hypothesised that there is a likelihood of all query terms not being adequate for expansion; several terms might impact expansion adversely. Hence, the au-

 https://doi.org/10.1633/JISTaP.2021.9.2.1

thors formulated the query-supplemented AQE technique based on the V2Q approach. The scheme comprises filters that disregard unnecessary words from the original query.

Zamani and Croft (2016) suggested that the PRF words which are embedded be based on embedding-specific relevance. This framework builds upon the relevance model technique. The model was evaluated using the TREC test set; the results highlighted a noteworthy performance increase compared to the baseline relevance framework. This technique evaluates the semantic similarity of the terms identified from WE vector similarity comparison. It is assumed that query terms are conditionally independent, thereby indicating that the terms comprising the expanded query should be related to the terms identified for augmenting the initial query. Considering that the present study uses both techniques in the baseline form, these techniques are discussed in detail.

El Mahdaouy et al. (2019) suggested building WE resemblance into the PRF frameworks for AIR. The concept is to choose expansion terms from PRF documents such that there is an alignment considering similarity and distribution with the initial query terms. AIR specific to the PRF model can use WE, because similar words can be included in a set at one side where they are proximal in the vector space. The primary objective is to enhance the weight and semantic-similarity to the terms of the original query. The present study assesses three WE frameworks: CBOW, skip-gram, and GloVe. Further, WE resemblance is integrated into four PRF frameworks, namely the KLD (Carpineto et al., 2001), Log-Logistic, Bo2 of the Family of Divergence from Randomness (Amati & Van Rijsbergen, 2002), and SPL of the information-based family of PRF models (Clinchant & Gaussier, 2013). Assessments were conducted on the TREC 2001/2002 Arabic test set using three neural WE frameworks. The study objective was to comprehend ways to integrate WE in the PRF methods for AIR. The outcomes indicate that the PRF extensions used in the present study offer a significant performance increase over baseline models. Further, an increase of 22% was observed for the elemental IR mean average precision (MAP) model; the robustness index was enhanced by 68%. Moreover, the performance differential of the three WE frameworks, namely skip-gram, CBOW, and GloVe was not statistically significant.

Fernández-Reyes et al. (2018) suggested a global AQE technique where the entire query is considered. This technique gathers pertinent information concerning potential expansion terms and uses it to reduce disambiguation issues and enhance precision and recall metrics. Word2Vec-based representation was initially suggested by Mikolov et al. (2013b); the present study uses this representation form for AQE (ALMasri et al., 2016) because of its computational efficiency. Fernández-Reyes et al. (2018) proposed two AQE techniques: a prospect-guided association scheme (V2Q) and a query-guided association scheme (Q2V). The Q2V technique identifies potential candidates using query terms. Conventional collection methods are used for this scheme. The top N closest candidates are identified and used as pre-selection elements; subsequently, they are subjected to refinement that produces the list of final expansion words. Consequently, this scheme produces $Rank_q$ for every query term q. This list comprises data pairs having words and their respective votes. The Q2V technique produces a candidate term list that is used for choosing expansion terms. After the ranking is completed, the list is sorted so that the expansion term selection methods may be used. Typically, $Rank_q$ is used to identify and select the top terms. Often, potential terms retrieved using the original query match the meaning of the original words; consequently, Q2V may lose the query context. In contrast, V2Q uses an opposite approach compared to the Q2V scheme. This scheme hypothesises that the potential expansion terms must provide votes for every query term; hence, candidates are likely to be selected. Several query terms might affect the expansion adversely; put differently, not all query elements can be expected to enhance expansion. Hence, the suggested V2Q scheme identifies words having a relatively higher semantic similarity to the query terms. Consequently, V2Q scores well over Q2V since the latter employs all query elements to identify candidate terms. There is a likelihood of disambiguation in the query. The suggested techniques enhance the precision and recall without relevance-specific feedback information. Experiments indicated that the outcomes were superior compared to classic IR frameworks such as Vector Space Model, LM, PRF model, and Okapi BM25.

Zamani and Croft (2016) assessed WE use to enhance the query language framework performance for ad-hoc retrieval tasks. An AQE-embedding relevance framework was suggested based on the work of Lavrenko and Croft (2017). The technique relied on semantic similarity at the WE vector level and suggested two different query language framework estimations. The first technique works under the presupposition that the query terms share conditional dependence. On the other hand, the second technique presupposes that two semantically similar words do not depend on the query. Words having weights close to

semantically synonymous query terms but not included in the query are assimilated to augment the original query. These techniques are used for the PRF relevance framework and termed EQE1 and EQE2. Three standard TREC sets, namely Robust (TRE C Robust Track 2004 collection), GOV2 (TREC Terabyte Track 2004-2006 collection), and AP (Associated Press 1988-1989), were used for the experiments. The experiments relied on the GloVe technique of Pennington et al. (2014) to extract WE from six billion tokens contained in Giga-words five and Wikipedia Dump 2014. The feedback document count was 10 for the PRF scenario. The assessment comprised four baselines: (1) standard maximum likelihood estimation for the query framework, (2) a WE (VEXP) heuristic QE technique (ALMasri et al., 2016), (3) document language model smoothing technique based on embedding (GLM) (Ganguly et al., 2015), and (4) a QE technique that used vocabulary term vectors' similarity along with the average embedding vector specific to all query terms. Cosine similarity specific to WE vectors was used for determining semantic similarity between words. The experiments indicated that embedding techniques have a significantly better performance than competitive baselines in most cases, considering MAP and average precision (AP) metrics. V2Q, EQE1, and previously-discussed techniques have a constraint considering the individual query elements.

Nevertheless, research indicates that individually-used query terms employed for QE might cause query drift by deflecting the query context that could be different from those individual query terms (Crimp & Trotman, 2018). Hence, the present study endeavours to assess the effects of AQE using WE on the full query instead of considering the query terms separately. Explicitly, this research focuses on AQE using DANs sentence embedding vectors.

3. METHOD

To augment AIR performance, we suggest using the DANs technique to formulate a WE-specific PRF scheme that integrates WE similarities into present PRF frameworks. The fundamental idea of the proposed technique is that performance enhancement can be obtained if expansion element similarity and distribution in pseudo-relevant documents are attached to the original query vectors using the average vector representation. Expansion information will be gathered using pseudo-relevant documents. Average vectors corresponding to DANs will be sorted, and the top results will be used to augment the original query.

To achieve the desired objective, we used DANs on the present bag-of-words IR framework. Probability-based Okapi BM25 framework is used along with two representative WE techniques, namely, EQE1 and V2Q for AQE. These techniques were formulated by Zamani and Croft (2016) and Fernández-Reyes et al. (2018), respectively.

We selected BM25 owing to its excellent performance in TREC retrieval tasks; furthermore, it has also influenced search engine ranking algorithms (Croft et al., 2010). BM25 is set as the baseline framework for non-AQE output. In contrast, V2Q and EQE1 schemes are used for comparing the output of DANs-based AQE with the presently used AQE schemes suggested by both techniques. Hence, we can evaluate whether DANs can output expansion terms having better relevance during PRF-based document retrieval.

3.1. Word2Vec and DANs

Word2Vec is a neural network-facilitated technique that outputs word vector representations; it was proposed by Mikolov et al. (2013b). It processes text input and provides the corresponding word vectors. The system is trained using training text, after which it learns how to represent words as vectors (Xue et al., 2014). The approach suggested in the present study relies on Word2Vec, that lists two WE model training frameworks: Skip-gram and CBOW. The CBOW framework estimates the middle word using distributed context-specific representations. On the other hand, the skip-gram framework estimates the context using distributed input word representation.

Most predictive frameworks for NLP activities suffer from performance degradation when unknown words that were not part of the training set due to low-frequency or other issues are used for experiments. Such words, referred to as Out-Of-Vocabulary (OOV), can affect NLP performance because of the inability of a representation framework to provide appropriate representation. To address OOV words, Arabic Wikipedia was used to augment the dataset. Hence, words not present previously were made available during training by dataset augmentation.

DANs is an average WE-based sentence embedding technique designed to convert text to vector format (Collobert et al., 2011). It is feasible to train DANs in little time with data (such as Arabic) with high syntactic variance. Embedding assumes a more abstract form as text length increases (Franco-Salvador et al., 2015). DANs combine the high speed and accuracy of unordered (such as Neural Bag-of-Word [NBOW]) and syntactic functions (such as Recursive Neural Networks [RecNN]), respectively.

 https://doi.org/10.1633/JISTaP.2021.9.2.1

Consequently, this approach provides superior results compared to the previously-used bag-of-words scheme. DANs implementation, specific to AQE, is described: the original query is processed using the WE framework, and the vectors are fed as input to DANs; consequently, the average vector is obtained as output. All inputs to DANs are passed across two linear layers. The average vector corresponding to every query term is computed to obtain a single vector or tensor (Iyyer et al., 2015). Fig. 1 depicts this process. Initially, a composition layer computes the average embedding. The intermediate data is fed through a hidden layer set comprising one or more layers; the network transforms the average. Lastly, a soft-max layer processes the intermediate data for prediction. In the case of using DANs for AQE, it follows three simple steps:

Step 1. The average of the embedding is taken, which is associated with the input sequence of tokens (in this case, the query terms).
Step 2. This average will pass through one or more feedforward layers.
Step 3. Implement a linear classification on the final representation of the layer.

The goal of this research is to investigate the impact of expanding a complete query sentence by using the DANs method based on the WE-based PRF technique on AQE. Thus, in the following sections, we illustrate how the DANs are being implemented in the three models, the basic BM25 IR model and the AQE approaches of EQE1 and V2Q. The BM25 is considered the basic approach of applying QE using DANs. We trained the Word2Vec CBOW model using the top-retrieved document of PRF technique on which search is performed. The dataset is the Arabic TREC 2001/2002 collections consisting of news about the Middle East and including articles from May 13, 1994, to December 20, 2000. Search is also performed on the same corpus. Initially, the dataset was divided into three groups: TREC 2001 contains 25 queries, TREC 2002 contains 50 queries, and TREC 2001/2002 contains 75 queries. The queries in TREC 2001/2002 are the combination of both TREC 2001 and TREC 2002 queries. Table 1 provides statistical information on the dataset.

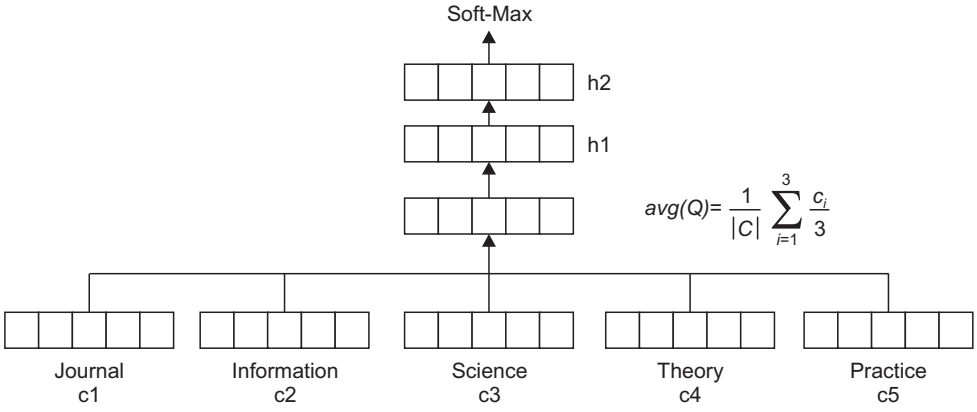

Fig. 1. Deep averaging networks model. Adapted from Iyyer et al., Association for Com putational Linguistics 2015.

Table 1. Statistics for Arabic TREC collections

Collections	TREC 2001	TREC 2002	TREC 2001/2002
Number of queries	25	50	75
The average number of words/queries	4.88	3.28	4.08
Number of documents		383,872	
Number of tokens		76 million	
Number of unique words		666,094	
Size (compressed)		209 MB	
Size (uncompressed)		869 MB	

3.2. Query Expansion Based-PRF using DANs (BM25+DANs-PRF)

The Okapi BM25 (Robertson et al., 1995) model is considered as one of the best-known term-weighting schemes derived from the probabilistic model. It considers three components, namely, the term frequency, inverse document frequency, and the length of the document (Robertson et al., 1995; Trotman et al., 2014; Vaidyanathan et al., 2015). The most common BM25 scoring function is shown in Equations 1 and 2:

$$\sum_{i\in Q} log \frac{\frac{(r_i+0.5)}{R-r_i+0.5}}{\frac{(n_i-r_i+0.5)}{(N-n_i-R+r_i+0.5)}} \cdot \frac{(k_1+1)\mathrm{f}_i}{K+\mathrm{f}_i} \cdot \frac{(k_2+1)q\mathrm{f}_i}{k_2+q\mathrm{f}_i} \quad (1)$$

$$K = k_1\left((1-b)+b.\frac{dl}{avdl}\right) \quad (2)$$

Where k_1, k_2, and K are parameters whose values set empirically. qf_i refers to how many times the same term repeated in the same query. R is the number of relevant documents in the query, and r_i is the number of relevant documents containing term i. The document length is dl. A typical value for k_1 is 1.2, and k_2 varies from 0 to 1,000 and b=0.75. $avdl$ is the average length of a document in the collection.

The application of QE using DANs in the BM25 model is a straightforward process, whereby the input is the entire query terms, and based on the WE model, vectors for each individual query terms will be extracted, and the average vector will be generated using DANs. The top k retrieved documents of PRF will be trained using CBOW model, where each term represented in the WE corpus as a vector of real number. Then, the most similar vectors in this corpus to the average vector of DANs are identified by using the cosine similarity measures, as illustrated in Equation 3. The most similar vectors are the potential candidate expansion vectors, and the candidate expansion terms related to the extracted vectors are identified using WE corpus. The top n candidates are used as expansion terms to the original query.

$$Cosine(A,B) = \frac{A\cdot B}{\|A\|\times\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\times\sqrt{\sum_{i=1}^{n} B_i^2}} \quad (3)$$

$$\frac{1}{|C|}\sum_{t=1}^{|C|}\log[\log[P(wt \setminus wt-c,\ldots,wt-1,wt+1,..,wt+c)]] \quad (4)$$

Equation 4 is the CBOW architecture of the Word2Vec model to learn the vector representation of the words, where it uses the surrounding words to predicts the target word, where $|C|$ is the number of words in the corpus, and c is the size of the dynamic context of wt.

$$avg(Q) = \frac{1}{|C|}\sum_{i=1}^{3}\frac{c_i}{3} \quad (5)$$

Equation 5 is used to find the average vector of the original query term vectors, where v_i is the vectors for each term i in query Q, $avg\ (Q)$ is the average vector of Q, and n the number terms in query Q.

The following steps illustrate the application of DANs-based PRF for QE on the BM25 model:

Step 1. Select a set of users' query Q consisting of n terms, whereby $Q = \{t_1, t_2, \ldots, t_n\}$;

Step 2. For each term t_i in Q, using the created WE model, we get the respective vector v_i from the WE corpus;

Step 3. For each vector v_i in query Q, we compute the average vector of this query Q using $s=avg(Q)=\frac{1}{|C|}\sum_{i=1}^{3}\frac{c_i}{3}$, where $avg\ (Q)$ is the average vector of Q, v_i is the vectors for each term i in query Q, and n the number terms in query Q;

Step 4. Select set $E = \{e_1, e_2, \ldots, e_n\}$, where E (consisting of terms e) is the top k document of PRF;

Step 5. For each term e in set E, using the created WE model, we get the respective vector ev_i from the WE corpus;

Step 6. For each vector ev_i, we compute the WE similarities to the $avg\ (Q)$ of DANs;

Step 7. Select the top k similar vectors to the $avg\ (Q)$ from set E (consists vectors ev_i) and save the selected vectors in set W, where set $W = \{w_1, w_2, \ldots w_k\}$ has the most similar vectors in set E to the $avg\ (Q)$ of DANs;

Step 8. Based on the created WE model, get the words $tw_1, tw_2, \ldots tw_k$ respective to the final candidate vectors $w_1, w_2, \ldots, w_k$ from the WE corpus;

Step 9. Add the new words $tw_1, tw_2, \ldots tw_k$ to the original query Q to get new query Q';

Step 10. Retrieve documents using the new query Q'.

The present study relied on the PRF method to identify potential expansion term vectors using top documents obtained during retrieval. QE implemented through DANs is easily integrated with the BM25 framework-

 https://doi.org/10.1633/JISTaP.2021.9.2.1

based PRF. The set of query terms is formed using the WE framework and forms the input. Vectors corresponding to every query element are computed; subsequently, DANs are employed to determine the average vector. Training is implemented using the CBOW scheme, where the top *k* documents comprising PRF are used. Real number vectors are used for denoting the elements of the WE corpus. Cosine similarity rules are employed to determine the vectors most similar to the DANs average vector; this process is specified using Equation 5. The vectors identified during the previous step are potential expansion element vectors whose corresponding word representations are potential query augmentation elements. The first *n* words are used for augmenting the query.

3.3. Embedding-Based Query Expansion Method using DANs (EQE1+DANs-PRF)

As indicated before, the EQE1 scheme suggested by Zamani and Croft (2016) suggested that query terms are presupposed to have conditional independence. This technique begins by identifying from the WE corpus a list of potential expansion candidates using similarity metrics. Subsequently, the terms identified in the previous step are compared with all query terms. Threshold candidates are those bearing high similarity to all query terms; these are chosen for QE. The EQE1 technique provides acceptable results when query terms have a substantial similarity. Nevertheless, most queries do not have a substantial similarity between all constituent words.

While implementing DANs using EQE1-based PRF, rather than identifying very similar candidates, we intended to identify the degree of similarity between the vectors corresponding to the top documents retrieved during PRF and the *average vector* corresponding to the initial query. Here, query-term vectors are identified using the WE corpus; these vectors are used to identify similar vectors from top PRF documents using the WE framework.

Once the vectors corresponding to the original query terms are obtained, cosine similarity is used to identify vectors from the top retrieved PRF documents resembling the initial query vectors. Furthermore, we used DANs on the initial query terms to determine the corresponding average vector. These two steps yielded the candidate and average vectors. Subsequently, the candidate and average DANs vectors were compared. Candidates having equal or greater than ≥ 0.7 similarity to average DANs vectors were selected.

Furthermore, the WE framework was used to reference the WE corpus to identify words using corresponding vectors. Lastly, new words are used for QE. Experiments indicated that 0.7 was an appropriate threshold for selecting similar words.

We proceeded under the assumption that the average query-term vector indicates a significant semantic similarity to the query. The top PRF documents retrieved during the process comprise valid words for QE. Furthermore, potential word vectors present in top PRF-retrieved documents having maximum similarity to the average vector are understood to be semantically analogous or associated with the original query's meaning because that vector denotes the query's centroid. The process of applying DANs in EQE1 can be summarized into the following steps:

Step 1. Select a set of users' query Q consisting of n terms, whereby $Q = \{t_1, t_2, ..., t_n\}$;
Step 2. For each term t_i in Q, using the created WE model, we get the respective vector v_i from the WE corpus;
Step 3. Compute the similarity (using cosine similarity) between v_i and all other vectors in the WE corpus;
Step 4. Select the top k similar vectors to v_i from the WE corpus and save the selected vectors in set E, where set $W = \{w_1, w_2, ..., w_k\}$ has the most similar vectors to v_i in the WE corpus;
Step 5. For each vector v_i in query Q, we compute the average vector of this query Q using $DANs = avg(Q) = \frac{1}{|C|}\sum_{i=1}^{3}\frac{c_i}{3}$, where $avg\ (Q)$ is the average vector of Q, v_i is the vectors for each term i in query Q, and n the number terms in query Q;
Step 6. Compute the similarity between $avg\ (Q)$ and set W;
Step 7. Select the vectors $\{vw_1, vw_2, ... vw_k\}$ from set W which have similarity ≥ *0.7* to $avg\ (Q)$;
Step 8. Based on the created WE model, get the words $tw_1, tw_2, ... tw_k$ respective to the final candidate vectors $vw_1, vw_2, ..., vw_k$ from the WE corpus;
Step 9. Add the new words $tw_1, tw_2, ... tw_k$ to the original query Q to get new query Q';
Step 10. Retrieve documents using the new query Q'.

3.4. Prospect-Guided Query Expansion Strategy Based on DANs (V2Q+DANs-PRF)

Fernández-Reyes et al. (2018) proposed V2Q and declared that not all query terms are necessarily practical for the expansion process; some terms might impact retrieval quality adversely. Hence, they suggested using a filtering stage to isolate the original terms before starting the ex-

pansion process. Potential expansion words are selected using filtered query words. Nevertheless, disregarding even a single word might lead to loss of query context and change in its meaning.

However, we have suggested using DANs to use the average vector corresponding to initial query terms rather than eliminating one or more query terms for candidate list creation. The vector is then used to prepare the potential candidate list. This technique allows keeping the query context unchanged. The candidate set will be created using the average DANs vector supposedly indicative of the original query context. This technique is understood as an augmented variant of V2Q.

Hence, we suggest that the average DANs vector be used instead of skimming the WE corpus for similar vectors. The top PRF documents will be searched for similar vectors. We propose to use V2Q+DANs-based PRF (V2Q+DANs-PRF). This technique comprises three primary steps: firstly, the initial candidate expansion set is created; this set is named W. It comprises vectors in the top PRF documents that closely resemble each query term vector. Consequently, D is the resulting candidate expansion vector list. The next step is to assess the W∪D set and calculate the cosine similarity of the vectors and compare it to the average vector *avg* (*Q*); all similarity values ≥ 0.7 are accepted, and the corresponding expansion vectors are selected. The terms corresponding to these vectors are used for QE. The following steps illustrate the process of applying DANs in the V2Q approach of QE.

Step 1. Select a set of user query Q consisting of n terms, whereby $Q = \{t_1, t_2, ..., t_n\}$;

Step 2. For each term t_i in Q, using the created WE model, we get the respective vector v_i from the WE corpus;

Step 3. Compute the similarity (using cosine similarity) between v_i and all other vectors in the WE corpus;

Step 4. Select the top k similar vectors to v_i from the WE corpus and save the selected vectors in set W, where set $W = \{w_1, w_2, ..., w_k\}$ has the most similar vectors in the WE corpus to v_i;

Step 5. For each vector v_i in query Q, we compute the average vector of this query Q using *DANs* = *avg* (*Q*) $= \frac{1}{|C|}\sum_{i=1}^{3}\frac{c_i}{3}$, where *avg* (*Q*) is the average vector of Q, v_i is the vectors for each term i in query Q, and n the number terms in query Q;

Step 6. Compute the similarity between *avg* (*Q*) and other vectors in the WE corpus;

Step 7. Select the top k similar vectors to *avg* (*Q*) from the WE corpus and save the selected vectors in set D, where set $D = \{d_1, d_2, ..., d_k\}$ has the most similar vectors in the WE corpus to *avg* (*Q*);

Step 8. Select the vectors from set W∪D which have similarity ≥ *0.7* to *avg* (*Q*) and save the selected vectors in set S, where set S $\{s_1, s_2, ..., s_n\}$ has the vectors from set W∪D that have similarity ≥ *0.7* to *avg* (*Q*);

Step 9. Based on the created WE model, get the words $ts_1, ts_2, ... ts_k$ respective to the final candidate vectors $s_1, s_2, ..., s_n$ from the WE corpus;

Step 10. Add the new words $ts_1, ts_2, ... ts_k$ to the original query Q to get new query Q';

Step 11. Retrieve documents using the new query Q'.

4. EXPERIMENTS

4.1. Experimental Setup

To assess the effects of the suggested sentence-embedding method DANs for PRF founded on QE, we evaluated three techniques, namely, basic BM25 framework, EQE1, and V2Q approaches. The previous section discusses the implementation of DANs as proposed. We contrast the outcomes of all DANs-based PRF techniques with the original techniques as formulated by researchers.

The TREC 2001/2002 Arabic newswire dataset was used for assessing the framework. Several researchers who worked on retrieval of Arabic text used the same dataset (Darwish & Ali, 2012; El Mahdaouy et al., 2019). This dataset comprises 2,117 Middle East-specific news documents dated between May 13, 1994, and December 20, 2000. The Word2Vec technique (Mikolov et al., 2013b) was used to build the WE corpus for experiments that used 383,872 Arabic news articles from Agence France Presse. The Linguistic Data Consortium disseminates these articles. The documents are encoded using UTF-8 and occupy about one gigabyte.

TREC 2001/2002 Arabic newswire contains three TREC sets: TREC 2001, TREC 2002, and TREC 2001/2002, comprising 25, 50, and 75 queries. TREC 2001/2002 comprises a combination of the queries of the constituent sets. Stemmer was used during experimentation because it has demonstrated a substantial impact on Arabic text retrieval performance. Farasa (Darwish & Mubarak, 2016) is considered the most efficacious stemmer for Arabic and was used in the present works of El Mahdaouy et al. (2019). Fig. 2 shows a comparison of precision and recall performance for 25 queries for TREC 2001, while Fig. 3 and Fig. 4 showing a comparison of

 https://doi.org/10.1633/JISTaP.2021.9.2.1

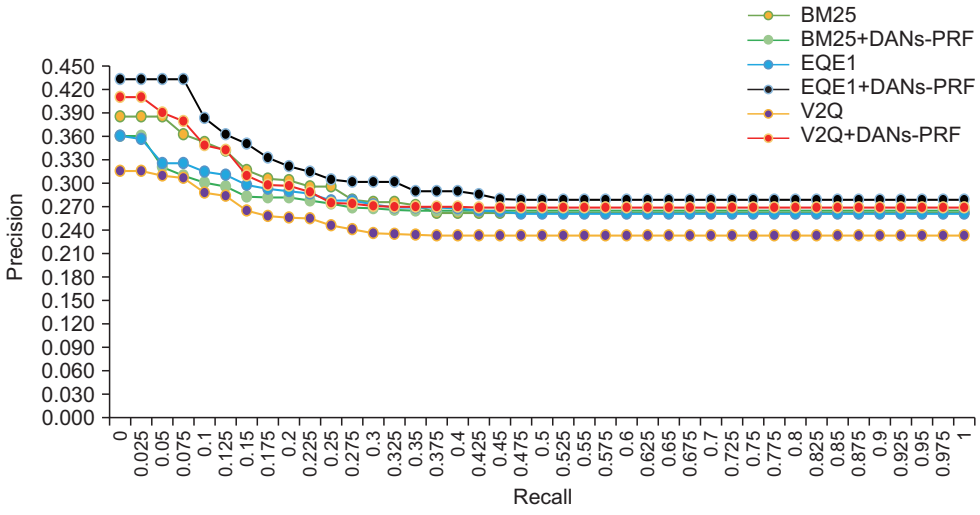

Fig. 2. Comparison of performance by precision and recall for 25 queries for TREC 2001 of all the models. DANs, deep averaging networks; PRF, pseudo relevance feedback.

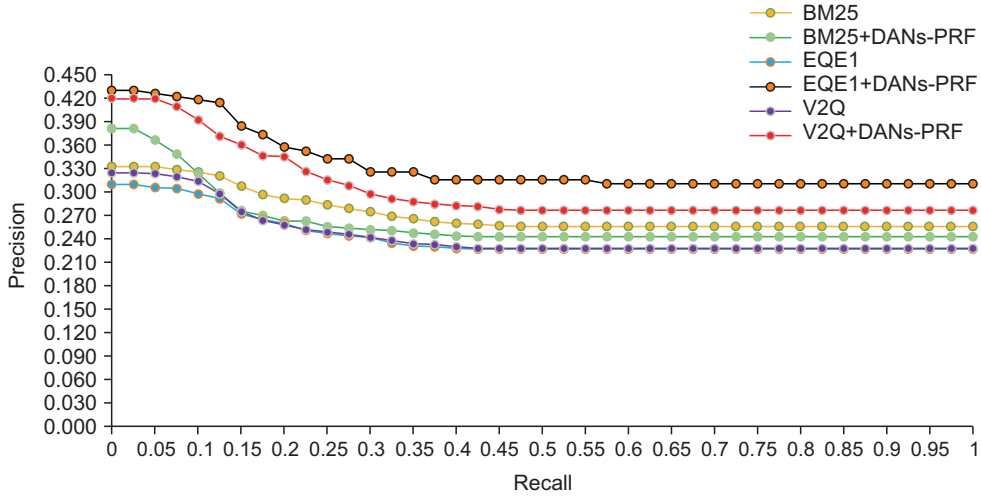

Fig. 3. Comparison of performance by precision and recall for 50 queries for TREC 2002 of all the models. DANs, deep averaging networks; PRF, pseudo relevance feedback.

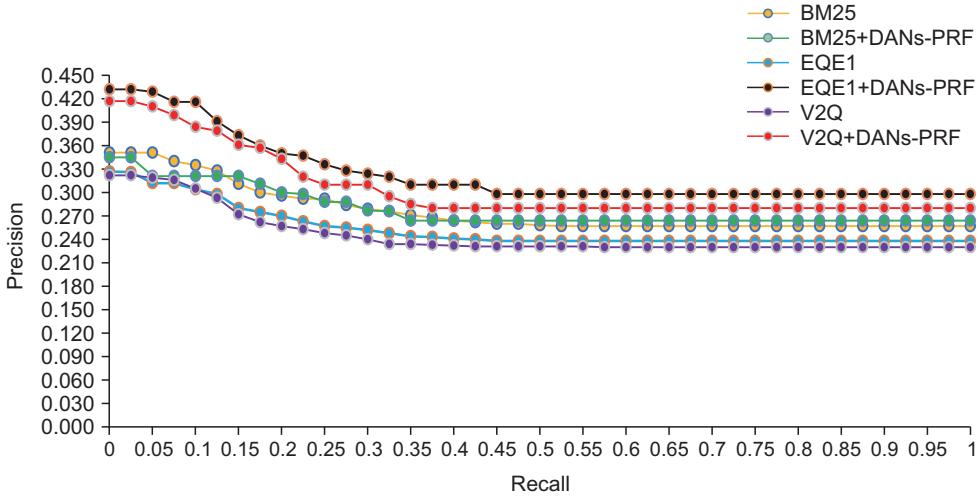

Fig. 4. Comparison of performance by precision and recall for 75 queries for TREC 2001/2002 of all the models. DANs, deep averaging networks; PRF, pseudo relevance feedback.

precision and recall performance for 50 queries for TREC 2002, and 75 queries for TREC 2001/2002 respectively, for all the baseline approaches and the proposed methods.

The system was set to retrieve 100 documents. The assessment comprised three baselines: (1) Okapi BM25 probabilistic framework used without expansion, (2) Embedding-based Query Expansion (EQE1), and (3) Prospect-Guided QE strategy (V2Q). Whoosh is a Python-based search engine library (Mukherjee & Kumar, 2019), which was used for all experiments. WE vector cosine similarity measure was used to calculate the semantic similarity between the terms.

4.2. Evaluation Metrics

The DANs QE-based PRF is benchmarked against six baseline techniques using MAP, corresponding to the top 100 documents for determining retrieval efficacy. Furthermore, precision for the top ten documents (P@10) was also used for performance measurement. The expressions for the performance metrics are specified below. Equations 6 and 7, respectively, express the recall and precision metrics, while Equation 8 is for the MAP, and the AP metric is presented in Equation 9.

$$Recall = \frac{|Rel| \cap |Ret|}{|Rel|} \tag{6}$$

$$Precision = \frac{|Rel| \cap |Ret|}{|Rel|} \tag{7}$$

Where:

Ret is the total number of retrieved documents, and *Rel* is the total number of relevant documents in the dataset.

$$MAP = \frac{1}{N}\sum_{n} AP_n \tag{8}$$

Where:

AP is the average precision value for a given query from the evaluation set of *n* queries, defined as:

$$AP = \frac{\sum_n P@r}{R} \tag{9}$$

Where:

r being the rank of each relevant document, *R* is the total number of the relevant documents, and *P@r* is the precision of the top-r retrieved documents.

5. RESULTS

The results of the DANs-based PRF used in this study are presented in this section. Conventional DANs-based QE is applied to the BM25 framework (BM25+DANs). DANs-based EQE1 and V2Q approaches were assessed for identification and selection performance pertaining to potential expansion terms. This study did not use the entire WE corpus, unlike the traditional AQE techniques. Instead, we used the top PRF documents for QE. These augmented techniques were designated BM25+DANs-PRF, EQE1+DANs-PRF, and V2Q+DANs-PRF, respectively. It is hypothesised that DANs can provide better expansion terms using top-retrieved PRF documents instead of the entire WE corpus. Consequently, AIR system performance can be enhanced.

Table 2 presents experimental results of all schemes, considering MAP and top-10 precision level (P@10). Ta-

Table 2. MAP and P@10 for all the models

Techniques	Collections					
	TREC 2001		TREC 2002		TREC 2001/2002	
	MAP	P@10	MAP	P@10	MAP	P@10
BM25	31.30	42.10	28.70	35.80	29.50	37.90
BM25+DANs-PRF	25.60	22.90	29.30[a)]	37.60[a)]	31.10[a)]	34.50
EQE1	30.70	42.90	25.70	33.50	26.90	30.40
EQE1+DANs-PRF	31.30[a)]	43.00[a)]	30.20[a)]	34.30[a)]	32.30[a)]	34.20[a)]
V2Q	27.00	35.20	26.20	33.30	26.50	33.90
V2Q+DANs-PRF	28.30[a)]	35.70[a)]	31.20[a)]	34.60[a)]	30.70[a)]	34.10[a)]

MAP, mean average precision; DANs, deep averaging networks; PRF, pseudo relevance feedback.
[a)]These values indicate that the corresponding scheme produced superior results than the baseline DANs-based PRF.

 https://doi.org/10.1633/JISTaP.2021.9.2.1

Table 3. MAP values for topic-by-topic analysis for all the models

Topics	BM25	BM25+DANs-PRF	EQE1	EQE1+DANs-PRF	V2Q	V2Q+DANs-PRF
1. Politics	0.169	0.295[a)]	0.217	0.216	0.244	0.261[a)]
2. Arts	0.235	0.330[a)]	0.277	0.303[a)]	0.273	0.283[a)]
3. Technology	0.093	0.148[a)]	0.098	0.118[a)]	0.127	0.185[a)]
4. Diseases	0.190	0.371[a)]	0.327	0.337[a)]	0.346	0.358[a)]
5. Waters	0.095	0.144[a)]	0.125	0.155[a)]	0.117	0.119[a)]
6. Environment	0.106	0.168[a)]	0.130	0.204[a)]	0.119	0.174[a)]
7. Tourism	0.101	0.161	0.078	0.082[a)]	0.112	0.220[a)]

MAP, mean average precision; DANs, deep averaging networks; PRF, pseudo relevance feedback.
[a)]These values indicate that the corresponding scheme produced superior results than the baseline DANs-based PRF.

ble 3 presents the MAP values for topic-by-topic analysis of all the proposed techniques in addition to the baseline approaches, respectively.

6. TOPIC-BY-TOPIC ANALYSIS

As mentioned earlier, the dataset used in this study is Arabic newswire TREC 2001/2002, which has two collections, namely TREC 2001 and TREC 2002. The first collection consists of 25 queries and the second 50 queries, and the total is 75 queries from various topics. However, to carry out a topic-by-topic analysis, these queries have been grouped into seven domains: politics (34 queries), arts (11 queries), technology (11 queries), diseases (seven queries), waters (six queries), environment (four queries), and tourism (two queries).

7. DISCUSSION

Empirical data presented in Table 2 suggests that the augmented BM25+DANs-PRF technique is superior to the BM25 probabilistic framework for TREC 2002 and TREC 2001/2002 datasets when evaluated using MAP. Considering P@10, the technique proposed in this present article yielded better results than the traditional BM25 executed only for the TREC 2002 dataset. For TREC 2001, the proposed BM25+DANs-PRF technique failed to pass the baseline model BM25 in terms of MAP and P@10.

Moreover, the proposed EQE1+DANs-PRF approach was superior to the traditional EQE1 technique when measured using MAP and P@10 for all TREC datasets. At the same time, the proposed V2Q+DANs-PRF approach used in the present study is superior to the traditional V2Q approach when evaluated using P@10 and MAP for all TREC datasets.

Furthermore, based on the results of topic-by-topic which were presented in Table 3, we can note that, in terms of MAP, the proposed technique BM25+DANs-PRF has outperformed their baseline BM25 (without expansion) for all the topics, where it significantly outperformed their baseline model for the topics "Politics" and "Diseases." The proposed EQE1+DANs-PRF technique has outperformed its baseline EQE1 approach in terms of MAP for all the topics, except for the "Politics" topic, where it has almost similar values. In addition, the proposed technique V2Q+DANs-PRF has outperformed its baseline V2Q approach for all the topics in terms of MAP.

8. CONCLUSION AND FUTURE WORKS

This research has evaluated the effects of DANs sentence-embedding methods on PRF-based AQE used for AIR. The intent is to identify the most appropriate terms from the top PRF documents. The vectors representing the terms are compared to the average vector of the initial query. The most-similar vectors are used to identify terms for AQE.

The present study is based on the Word2Vec WE model. Three baselines, namely, Okapi BM25, EQE1, and V2Q were used for performance benchmarking. The present study contributes by using a different approach: Rather than identifying the similarity between average DANs vectors and WE corpus, we attempted to determine the degree of similarity between the average DANs vector and top PRF documents.

The candidate vectors were generated based on the average vector of DANs in the enhanced BM25+DANs-PRF technique from the top retrieved documents of PRF. Meanwhile, in the other enhanced techniques EQE1+DANs-PRF and V2Q+DANs-PRF, the candidate

vectors are generated based on the original query term vectors in addition to the average vector of DANs, with the help of the top-retrieved documents of PRF.

The augmented BM25+DANs-PRF technique relies on the top-retrieved PRF documents and identifies potential vectors using DANs average vector. At the same time, other augmented methods like V2Q+DANs-PRF and EQE1+DANs-PRF rely on the vectors of the initial query terms and DANs average vectors to identify terms using top PRF documents.

Overall, the experiment results illustrated in Table 2 showed that the proposed DANs-based PRF, when incorporated into the probabilistic model BM25, are unable to pass their baseline standard BM25 (without expansion) in most cases in terms of MAP and P@10. On the other hand, the proposed DANs-based PRF when incorporated into the EQE1 and V2Q approaches has proven its success, where it outperformed the standard baseline approaches EQE1 and V2Q; however, in most of the cases, there is no significant difference between the results of the baselines approaches and the proposed techniques.

In addition, based on the topic-by-topic analysis, we can notice that the proposed techniques BM25+DANs-PRF, V2Q+DANs-PRF, and EQE1+DANs-PRF have significantly outperformed their baselines in most cases in terms of MAP. Finally, we can conclude that the proposed DANs technique is performing better with the specific topics than with the global topics.

This study's objective was to address the vocabulary mismatch challenge using the top-retrieved PRF documents that have strong similarity to the average vector of the query terms. We hypothesise that the top-retrieved PRF documents might have the required terms that demonstrate a high degree of similarity to the average vector of the initial query terms. Consequently, Arabic text retrieval performance can be enhanced.

ACKNOWLEDGMENTS

We want to thank the LDC for providing us with the LDC2001T55 Arabic Newswire Part 1 at no cost, and for awarding us with the Fall 2012 LDC Data Scholarship. This study is partially supported by the Universiti Kebangsaan Malaysia, grant: DCP-2017-007/4.

CONFLICTS OF INTEREST

No potential conflict of interest relevant to this article was reported.

REFERENCES

Abbache, A., Meziane, F., Belalem, G., & Belkredim, F. Z. (2016). Arabic query expansion using WordNet and association rules. *International Journal of Intelligent Information Technologies*, 12(3), 51-64. http://doi.org/10.4018/IJIIT.2016070104.

Abu El-Khair, I. (2007). Arabic information retrieval. *Annual Review of Information Science and Technology*, 41(1), 505-533. https://doi.org/10.1002/aris.2007.1440410118.

Aklouche, B., Bounhas, I., & Slimani, Y. (2018, November 14-16). *Query expansion based on NLP and word embeddings*. Paper presented at the TREC 2018, Gaithersburg, MD, USA.

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. In N. Ferro, F. Crestani, M.-F. Moens, J. Mothe, F. Silvestri, G. M. Di Nunzio, C. Hauff, & G. Silvello (Eds.), *Proceedings of the 38th European Conference on IR Research* (pp. 709-715). Springer. https://doi.org/10.1007/978-3-319-30671-1_57.

Alsmearat, K., Al-Ayyoub, M., & Al-Shalabi, R. (2014, November 10-13). An extensive study of the Bag-of-Words approach for gender identification of Arabic articles. In A. Bouras, Z. Tari, A. Erradi, & S. Abdelwahed (Eds.), *Proceedings of the 2014 IEEE/ACS 11th International Conference on Computer Systems and Applications* (pp. 601-608). IEEE. https://doi.org/10.1109/AICCSA.2014.7073254.

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. https://doi.org/10.1145/582415.582416.

Atwan, J., Mohd, M., Rashaideh, H., & Kanaan, G. (2016). Semantically enhanced pseudo relevance feedback for Arabic information retrieval. *Journal of Information Science*, 42(2), 246-260. https://doi.org/10.1177%2F0165551515594722.

Azad, H. K., & Deepak, A. (2019). Query expansion techniques for information retrieval: A survey. *Information Processing & Management*, 56(5), 1698-1735. https://doi.org/10.1016/j.ipm.2019.05.009.

Belkin, N. J., Oddy, R. N., & Brooks, H. M. (1982). Ask for information retrieval: Part II. Results of a design study. *Journal of Documentation*, 38(3), 145-164. https://doi.org/10.1108/eb026726.

Ben Guirat, S., Bounhas, I., & Slimani, Y. (2016). Combining indexing units for Arabic information retrieval. *International Journal of Software Innovation*, 4(4), 1-14. https://

doi.org/10.4018/IJSI.2016100101.

Bengio, Y. (2009). Learning deep architectures for AI. *Foundations and Trends® in Machine Learning*, 2(1), 1-127. https://doi.org/10.1561/2200000006.

Berget, G., & Sandnes, F. E. (2015). Searching databases without query-building aids: Implications for dyslexic users. *Information Research: An International Electronic Journal*, 20(4), 689.

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. https://doi.org/10.1145/366836.366860.

Carpineto, C., & Romano, G. (2012). A survey of automatic query expansion in information retrieval. *ACM Computing Surveys*, 44(1), 1. https://doi.org/10.1145/2071389.2071390.

Clinchant, S., & Gaussier, E. (2013, September 29-October 2). A theoretical analysis of pseudo-relevance feedback models. In O. Kurland, D. Metzler, C. Lioma, B. Larsen, & P. Ingwersen (Eds.), *Proceedings of the ICTIR '13: International Conference on the Theory of Information Retrieval* (pp. 6-13). Association for Computing Machinery. https://doi.org/10.1145/2499178.2499179.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., & Kuksa, P. (2011). Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12(76), 2493-2537.

Crimp, R., & Trotman, A. (2018, December 11-12). Refining query expansion terms using query context. In B. Koopman, A. Trotman, & P. Thomas (Eds.), *Proceedings of the ADCS '18: 23rd Australasian Document Computing Symposium* (article no.: 12). Association for Computing Machinery. https://doi.org/10.1145/3291992.3292000.

Croft, W. B., Metzler, D., & Strohman, T. (2010). *Search engines: Information retrieval in practice*. Addison-Wesley.

Dalton, J., Naseri, S., Dietz, L., & Allan, J. (2019, April 14-18). Local and global query expansion for hierarchical complex topics. In L. Azzopardi, B. Stein, N. Fuhr, P. Mayr, C. Hauff, & D. Hiemstra (Eds.), *Proceedings of the 41st European Conference on IR Research*, *ECIR 2019* (pp. 290-303). Springer. https://doi.org/10.1007/978-3-030-15712-8_19.

Darwish, K., & Ali, A. (2012, July 8-14). Arabic retrieval revisited: Morphological hole filling. In H. Li, C.-Y. Lin, M. Osborne, G. G. Lee, & J. C. Park (Eds.), *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics* (pp. 218-222). ACL.

Darwish, K., & Mubarak, H. (2016, May 23-28). Farasa: A new fast and accurate Arabic word segmenter. In N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, & S. Piperidis (Eds.) *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC'16)* (pp. 1070-1074). European Language Resources Association.

Diaz, F., Mitra, B., & Craswell, N. (2016, August 7-12). Query expansion with locally-trained word embeddings. In K. Erk, & N. A. Smith (Eds.), *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* (pp. 367-377). Association for Computational Linguistics. https://doi.org/10.18653/v1/P16-1035.

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. https://doi.org/10.1177%2F0165551518792210.

Esposito, M., Damiano, E., Minutolo, A., De Pietro, G., & Fujita, H. (2020). Hybrid query expansion using lexical resources and word embeddings for sentence retrieval in question answering. *Information Sciences*, 514, 88-105. https://doi.org/10.1016/j.ins.2019.12.002.

Fang, H., & Zhai, C. (2006, August 6-11). Semantic term matching in axiomatic approaches to information retrieval. In S. Dumais, E. N. Efthimiadis, D. Hawking, & K. Järvelin (Eds.), *Proceedings of the SIGIR '06: 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 115-122). Association for Computing Machinery. https://doi.org/10.1145/1148170.1148193.

Faqeeh, M., Abdulla, N., Al-Ayyoub, M., Jararweh, Y., & Quwaider, M. (2014, August 27-29). Cross-lingual short-text document classification for Facebook comments. In M. Younas, I. Awan, & A. Pescape (Eds.), *Proceedings of the FiCloud 2014: 2nd International Conference on Future Internet of Things and Cloud* (pp. 573-578). IEEE. https://doi.org/10.1109/FiCloud.2014.99.

Farghaly, A., & Shaalan, K. (2009). Arabic natural language processing: Challenges and solutions. *ACM Transactions on Asian Language Information Processing*, 8(4), 14. https://doi.org/10.1145/1644879.1644881.

Farhan, Y. H., Noah, S. A. M., & Mohd, M. (2020). Survey of automatic query expansion for arabic text retrieval. *Journal of Information Science Theory and Practice*, 8(4), 67-86. https://doi.org/10.1633/JISTaP.2020.8.4.6.

Fernández-Reyes, F. C., Hermosillo-Valadez, J., & Montes-y-Gómez, M. (2018). A prospect-guided global query expansion strategy using word embeddings. *Information Processing & Management*, 54(1), 1-13. https://doi.org/10.1016/j.ipm.2017.09.001.

Franco-Salvador, M., Rangel, F., Rosso, P., Taulé, M., & Martít, M. A. (2015, September 8-11). Language variety identification using distributed representations of words and docu-

ments. In J. Mothe, J. Savoy, J. Kamps, K. Pinel-Sauvagnat, G. Jones, E. San Juan, L. Capellato, & N. Ferro (Eds.), *Proceedings of the 6th International Conference of the CLEF Association*, *CLEF'15* (pp. 28-40). Springer. https://doi.org/10.1007/978-3-319-24027-5_3.

Ganguly, D., Roy, D., Mitra, M., & Jones, G. J. F. (2015, August 9-13). Word embedding based generalized language model for information retrieval. In R. González-Ibáñez, & N. Hidalgo (Eds.), *Proceedings of the SIGIR '15: 38th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 795-798). Association for Computing Machinery. https://doi.org/10.1145/2766462.2767780.

Iyyer, M., Manjunatha, V., & Daumé, H., III. (2015, July 26-31). Deep unordered composition rivals syntactic methods for text classification. In C. Zong, & M. Strube (Eds.), *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing* (pp. 1681-1691). Association for Computational Linguistics. https://doi.org/10.3115/v1/P15-1162.

Kim, H. K., Kim, H., & Cho, S. (2017). Bag-of-concepts: Comprehending document representation through clustering words in distributed representation. *Neurocomputing*, 266, 336-352. https://doi.org/10.1016/j.neucom.2017.05.046.

Larkey, L. S., Ballesteros, L., & Connell, M. E. (2002, August 11-15). Improving stemming for Arabic information retrieval: Light stemming and co-occurrence analysis. In K. Järvelin, R. Baeza-Yates, & S. H. Myaeng (Eds.), *Proceedings of the SIGIR '02: 25th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 275-282). Association for Computing Machinery. https://doi.org/10.1145/564376.564425.

Lavrenko, V., & Croft, W. B. (2017). Relevance-based language models. *ACM SIGIR Forum*, 51(2), 260-267. https://doi.org/10.1145/3130348.3130376.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. Cambridge University Press.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013a). *Efficient estimation of word representations in vector space*. https://arxiv.org/abs/1301.3781v3.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013b, December 5-10). Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, & K. Q. Weinberger (Eds.), *Proceedings of the NIPS'13: 26th International Conference on Neural Information Processing Systems* (pp. 3111-3119). Curran Associates.

Miyanishi, T., Seki, K., & Uehara, K. (2013, October 27-November 1). Improving pseudo-relevance feedback via tweet selection. In Q. He, A. Iyengar, W. Nejdl, J. Pei, & R. Rastogi (Eds.), *Proceedings of the CIKM '13: 22nd ACM international conference on Information & Knowledge Management* (pp. 439-448). Association for Computing Machinery. https://doi.org/10.1145/2505515.2505701.

Mohsen, G., Al-Ayyoub, M., Hmeidi, I., & Al-Aiad, A. (2018, April 3-5). On the automatic construction of an Arabic thesaurus. In M. Quwaider (Ed.), *Proceedings of the 2018 9th International Conference on Information and Communication Systems* (pp. 243-247). IEEE. https://doi.org/10.1109/IACS.2018.8355431.

Montazeralghaem, A., Zamani, H., & Shakery, A. (2016, July 17-21). Axiomatic analysis for improving the log-logistic feedback model. In R. Perego, F. Sebastiani, J. Aslam, I. Ruthven, & J. Zobel (Eds.), *Proceedings of the SIGIR '16: 39th International ACM SIGIR conference on Research and Development in Information Retrieval* (pp. 765-768). Association for Computing Machinery. https://doi.org/10.1145/2911451.2914768.

Mukherjee, S., & Kumar, N. S. (2019, December 9-11). Duplicate question management and answer verification system. In M. Chang, R. Rajendran, Kinshuk, S. Murthy, & V. Kamat (Eds.), *Proceedings of the 2019 IEEE Tenth International Conference on Technology for Education* (pp. 266-267). IEEE. https://doi.org/10.1109/T4E.2019.00067.

Mustafa, M., AbdAlla, H., & Suleman, H. (2008, December 2-5). Current approaches in Arabic IR: A survey. In G. Buchanan, M. Masoodian, & S. J. Cunningham (Eds.), *Proceedings of the 11th International Conference on Asian Digital Libraries*, *ICADL 2008* (pp. 406-407). Springer. https://doi.org/10.1007/978-3-540-89533-6_57.

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. https://doi.org/10.1002/asi.23143.

Pennington, J., Socher, R., & Manning, C. (2014, October 25-29). GloVe: global vectors for word representation. In Y. Marton (Ed.), *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing* (pp. 1532-1543). Association for Computational Linguistics. https://doi.org/10.3115/v1/D14-1162.

Robertson, S. E., Walker, S., Jones, S., Hancock-Beaulieu, M. M., & Gatford, M. (1995). *Okapi at TREC-3*. Paper presented at the 3rd Text REtrieval Conference (TREC-3), Gaithersburg, MD, USA.

Roy, D., Paul, D., Mitra M., & Garain, U. (2016). *Using word embeddings for automatic query expansion*. Paper presented at the Neu-IR '16 SIGIR Workshop on Neural Informa-

tion Retrieval, Pisa, Italy.

Takeuchi, S., Sugiura, K., Akahoshi, Y., & Zettsu, K. (2017). Spatio-temporal pseudo relevance feedback for scientific data retrieval. *IEEJ Transactions on Electrical and Electronic Engineering*, 12(1), 124-131. https://doi.org/10.1002/tee.22352.

Trotman, A., Puurula, A., & Burgess, B. (2014, November 27-28). Improvements to BM25 and language models examined. In J. Culpepper, L. Park, & G. Zuccon (Eds.), *Proceedings of the ADCS '14: 2014 Australasian Document Computing Symposium* (pp. 58-65). Association for Computing Machinery. https://doi.org/10.1145/2682862.2682863.

Turney, P. D., & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37(1), 141-188.

Vaidyanathan, R., Das, S., & Srivastava, N. (2015). A study on retrieval models and query expansion using PRF. *International Journal of Scientific & Engineering Research*, 6(2), 13-18.

Xue, B., Fu, C., & Shaobin, Z. (2014, June 27-July 2). A study on sentiment computing and classification of Sina Weibo with Word2vec. In P. Chen, & H. Jain (Eds.), *Proceedings of the 2014 IEEE International Congress on Big Data* (pp. 358-363). IEEE. https://doi.org/10.1109/BigData.Congress.2014.59.

Zamani, H., & Croft, W. B. (2016, September 12-16). Embedding-based query language models. In B. Carterette, & H. Fang (Eds.), *Proceedings of the ICTIR '16: 2016 ACM International Conference on the Theory of Information Retrieval* (pp. 147-156). Association for Computing Machinery. https://doi.org/10.1145/2970398.2970405.

Zuccon, G., Koopman, B., Bruza, P., & Azzopardi, L. (2015, December 8-9). Integrating and evaluating neural word embeddings in information retrieval. In L. A. F. Park, & S. Karimi (Eds.), *Proceedings of the ADCS '15: 20th Australasian Document Computing Symposium* (article no.: 12). Association for Computing Machinery. https://doi.org/10.1145/2838931.2838936.