Received: May 6, 2019

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

Query Expansion Based on Wikipedia Word Embedding and BabelNet Method for Searching Arabic Documents

Maryamah Maryamah[1*] Agus Zainal Arifin[1] Riyanarto Sarno[1] Yasuhiko Morimoto[2]

[1] *Department of Informatics, Faculty of Information and Communication Technology, Institut Teknologi Sepuluh Nopember, Indonesia*
[2] *Department of Information Engineering Graduate School of Engineering, Hiroshima University, Higashihiroshima, Japan*
* Corresponding author's Email: maryamahfaisol02@gmail.com

Abstract: Document searching using queries that can understand the context can affect the intent and purpose of the user's desire when searching documents. Many studies have been conducted on understanding the context of the query, but differences in terms of language can lead to different methods of context understanding; therefore, methods implemented in the previous studies need to be improved. In this paper, we proposed a query expansion method based on BabelNet search and Word Embedding (BabelNet Embedding). Query expansion method focuses on developing queries based on semantic relationships on queries to understand the context of the query. Candidate queries were developed by finding synonyms, measuring similarity using WordNet, Word Embedding on all articles of Wikipedia, and BableNet Embedding on articles Wikipedia Online. We compared our proposed method with the existing semantic query expansion. Our result provided better result in retrieving relevant document with accuracy of 89% in searching Arabic documents.

Keywords: Semantic query expansion, Word embedding, BabelNet embedding, Arabic document.

1. Introduction

The Arabic language document has grown rapidly over time. This is evidenced by more than 100 million Arabic web page content [1]. The field of Arabic Natural Language Processing (ANLP) has poor attention compared to English that make Arabic has scarce resource, such as corpora and semantic [2]. Arabic has more challenge because there are many unique characteristics, complex nature (derivational, inflectional, etc) to understand Arabic text source [3]. Implementing Arabic in terms of language and literacy have more undeniable challenge in the field of document search.

Searching for documents performed by a user requires keywords in order to help users obtain the document. Keyword represents the input query from the user in searching the document. However, the user has difficulty in expressing the question into a relevant query. Ineffectiveness of information retrieval systems is often caused by the query inaccuracy. Hundreds of thousands of irrelevant documents will be returned if the selected keywords are too general. Retrieve information from the internet using an information retrieval system often requires precise keywords from multiple field to achieve the best result [4]. The user needs additional query from the system to help user obtain the relevant documents. One method that can be done is query expansion. Query expansion is a method of extending the query term by adding multiple query candidates to improve performance in document search.

In general, automatic query expansion is divided into three: statistical, semantics and hybrid [5]. Statistical method is a query expansion that explores queries based on the structure of a given word user in searching. Semantics is a method of query expansion by analysing and

understanding the meaning of query given by user to obtain the relevant document.

Hybrid is combination method of statistical and semantics method. Hybrid is a method that is often implemented because it generates better results in information retrieval [5]. Hybrid method combines two or three query expansion methods. The application of hybrid query expansion methods is needed to help understand the context of the query [6]. Understanding contexts on document search is also needed to help users obtain relevant document results.

Compared to statistical-based, semantic method is more effective in terms of information retrieval, especially if the retrieval list is very large and has many irrelevant documents. In addition, for a small retrieval list, semantic method is also more effective, since query candidates are based on, for example thesaurus or ontology independent corpus [7].

Query expansion based on semantic divided into three, namely linguistic, based ontology and mix mode (hybrid) [7]. The linguistic method uses word sense from the thesaurus and other word sense relations based on user's initial query. Ontology-based method is an expansion method that utilizes the ontology concept to obtain additional candidate queries. The mix mode method is a combined method of linguistic and ontology method in query expansion.

The disadvantage of the semantic method is that the candidate query result is general, so a mechanism to overcome the issue is needed, one of which is by applying statistical method to obtain relevant documents and the document will be used as the material in searching word sense (semantic method) for the query. A combination of statistical and semantic method is called Hybrid method. Hybrid is a method that is often implemented because it generates better results in information retrieval [5]. Hybrid method combines two or three query expansion methods. The application of hybrid query expansion method is needed to help understand the context of the query [6]. Understanding contexts on document search is also needed to help users obtain relevant document results.

There are several other methods that can be used to obtain documents that will be candidate queries, such as classification and clustering. Semantic search document is needed to obtain relevant document for candidate queries. One of the semantic searches that can be used is BabelNet method. BabelNet conducting semantic search documents that combine knowledge of Wikipedia articles and lexicographic from Wordnet [8]. Searching for BabelNet on Wikipedia is based on links that are linked between articles (see also, categories and external links in Wikipedia). The application of the BabelNet method can shorten the time compared to the pseudo-relevance feedback method because the method of finding articles that match the query and other related articles does not process all articles. BabelNet can be used to obtain articles that are semantic related to queries that do not cover the document extraction process. Document extraction is used to obtain a candidate term for searching documents. Document extraction is an important step because improper extraction will result in irrelevant candidate terms. Document extraction carried out in previous studies is only based on term frequency, the similarity with queries. The query expansion used only focuses on searching document or document extraction.

In this paper, we are proposed query expansion based on BabelNet search and Word Embedding. BabelNet search is used to search Wikipedia articles coherent with queries. Word Embedding is used for document extraction in selected articles using Word2Vec (skip-gram and continuous bag of word). The result of term extraction would be a candidate term for query expansion. Candidate term for query expansion is also obtained based on synonym, Wordnet, Word embedding (Word2Vec and GloVe) with train all Wikipedia Indonesia. Prioritizing a candidate term based on a semantic relationship with a query is expected to help retrieve relevant document.

This paper is organized as follows. Section 2 present the related word of this research contain existing method in query expansion. Section 3 describes proposed method. The result and experiment of this paper are explained in Section 4. The conclusion and future work are presented in Section 5.

2. Related work

There are several stages carried out in this paper as follows.

In survey paper [5], it is explained that in general the query expansion method is divided into two:

2.1 Query expansion manual

Manual-based expansion query is done by expanding additional queries using the strategy of

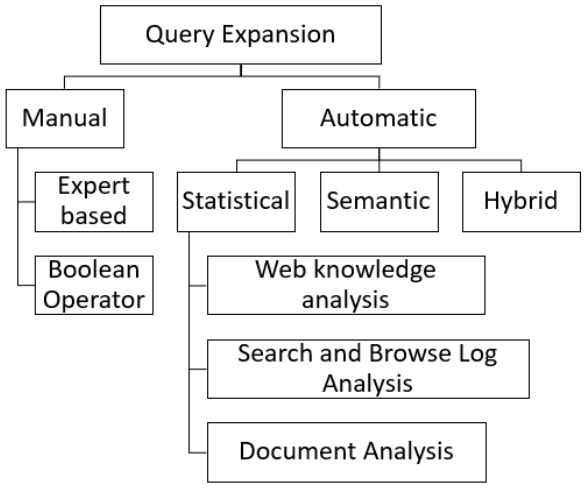

Figure 1. Survey query expansion

the researcher. The manual expansion query is divided into two:

2.1.1. Expert based

Expert based method is a query expansion that expands queries based on the knowledge of the in-field experts, since they have different strategies in expanding additional queries. The advantages of this method include additional queries relevant document results.

2.1.2. Boolean operator

The Boolean operator method is the addition of a query that is done by determining the words to be added by expanding a Boolean operator. The Boolean operator used in this paper is OR and AND. Implementing manual query expansion method with the help of expert is highly recommended due to the highly relevant additional query results. Expert help is needed to determine additional queries that are implicit in the query. The Boolean operator is very appropriate to use because the researcher knows which query is appropriate in order to obtain relevant documents. The disadvantage of the manual method is the old method of executing the process because it requires the help of additional queries from both experts and researchers, so a query expansion method that can provide additional automatic query is proposed.

2.2 Query expansion automatic

Expansion automatic is a query addition method that is done by searching for words related to the query. The relation used can be in the form of word similarity, word weights, and so on. Automatic-based expansion query is usually used in the field of health due to the limited knowledge in terms of health vocabularies. The user additional inputs are not relevant. A number of studies on health have been conducted by replacing the query with the Unified Medical Language Systems (UMLS) concept by calculating relevant documents that have the highest similarity using fuzzy [9], integrating MeSH (Medical Subject Headings) in tree form [10], retrieving information from MEDLINE vocabulary about information from various diseases, extracting and predicting based on semantic information, and turning information into UMLS concepts and predicting which treatment can be done [11]. Automatic query expansion is divided into three.

2.2.1. Statistical

Statistical query expansion is a method that focuses on analysing query words from corpora documents, web documents, browser history and text documents. Query expansion is divided into three based on the source of corpus:

a. Web Knowledge analysis
   Expansion query using web knowledge is done by using corpus web knowledge to obtain candidate queries. Expansion query using web knowledge is divided into three: Online Knowledge based search, text analysis links, social data search.
b. Search and Browse Log Analysis
   This method is divided into two: search log analysis and personalize search analysis. Search log analysis consists of click through data based and past log-based queries. Personalize search analysis is divided into four: implicit user profile, search behaviour based, explicit user profile, context aware analysis.
c. Document Analysis
   Expansion query using document corpus is done by analysing the corpus collection document to determine candidate queries. Expansion query using document analysis is divided into four: clustering based, local context analysis, global analysis, and local analysis. Local analysis is divided into two: relevance feedback and pseudo relevance feedback [12].

2.2.2. Semantics

Semantics is a query expansion method by extracting word relations, analyzing and understanding of the original word query to obtain the relevant documents. By determining the relation between the original word query and

other words, the semantic candidate query can be obtained. Query expansion based on semantic divide into three: linguistic, ontology based dan mix mode (hybrid) [7].

a. Linguistic method

The linguistic method uses word sense from the thesaurus such as synonym, hyponym, hypernym, meronym and other word sense relations based on the initial query. The word sense results will be the relevant candidate queries. The linguistic method is divided into two, namely morphological expansion and related term expansion.

1. Morphological expansion is a query expansion method that adds words derived from stem words, part of speech, adjective and others. Morphological expansion is one of the several very effective methods in the field of information retrieval [13]. A number of studies on query expansion used morphological methods in terms of stemming words [14], part of speech for phrase extraction using WordNet [15], part of speech for synonym extraction using WordNet [16]. The stemming word can be used for adding candidate queries and this has proven to be efficient for information retrieval [14]. Tag post is also proven to increase precision and recall significantly [7].
2. Related term expansion is a query expansion method that adds words based on semantically-related words such as synonyms, hypernyms, etc. A number of studies that used the related term expansion method are Wordnet [17-19], synonym [20], synonym for finding relevant tweet [21].

b. Ontology based method

Ontology based method is an expansion method that utilizes the ontology concept to obtain additional candidate queries. Ontology based method is divided into two based on the nature of the domain used.

1. Domain dependent uses specific properties of a particular domain. Some studies using dependent domains are [22, 23].
2. Domain independent uses a more general properties and can consist of several domains in one ontology. Some knowledge structures that are built using an independent domain are Freebase [24], DBpedia [25], [26], [27], UNIpedia [28].

c. Mix mode method

The mix mode method is a combined method of linguistic and ontology method in query expansion. The mix mode method utilizes the advantages of both methods, so that it is expected to obtain more relevant results. An example of a mix mode method is semantic query expansion in sport domains [29].

Some previous studies using semantic-based query expansion are conducted by adding candidate queries using Harman calculations, croft, okapi with algorithm lesk [30]; using co-occurrence from top feedback document; using WordNet to understand semantic queries; combining the document co-occurrence calculation process and its semantics using WordNet; using local word embedding (processing relevant feedback based on document dataset); and calculating its semantic relation.

Word embedding uses i.e. Word2Vec and GloVe to measure similarity between queries and documents [31]. Process the dataset document using Word2Vec to search for term candidates using semantic similarity term. Calculate proximity of term candidates by query using cosine similarity and is described as bipartite graph. Calculate the proximity of queries to candidate terms using the maximum cosine similarity. Calculate the proximity of candidate term with query using average cosine similarity [32].

2.2.3. Hybrid

Hybrid expansion query is a combination of statistic and semantic-based query expansion method. Some researchers previously combined this method because the application of the previous method is not enough to obtain relevant documents, so that they require a combination of both methods. Previous studies also found more relevant results using the hybrid method compared to the application of statistic or semantic-based method only.

3. The proposed semantic query expansion methods

In this paper, Fig. 2 show that proposed method contains several methods that are applied in the semantic based on synonyms, WordNet, and Word Embedding. The method applied from input query obtained based on forum questions in the field of religious social. This paper focuses on the semantic query expansion method because semantically related additional query has the possibility of additional relevant queries and optimal results. Additional query search using the semantic method also reduces the possibility of disambiguation term problems.

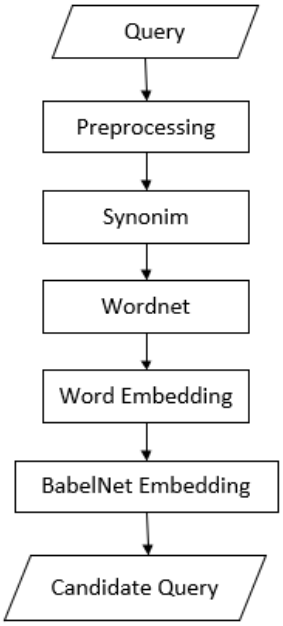

Figure 2. Proposed method

Semantic-based expansion query is a query expansion method that is carried out by analyzing and understanding the query meaning given by the user to obtain relevant documents. Semantic method implementation offers the ultimate precision rate and recall rate, but it demands a comparatively complex query language [19]. There are several query expansion methods applied in this study:

3.1 Synonym

The expansion query technique that is performed is thesaurus or finding word similarity from the query. The use of the synonym method is due to the existence of references containing words similar to the query, so that the reference is expected to be successfully obtained. Reference authors usually use similar words in composing documents. Finding word similarity is based on the query that has been done by preprocessing. Word expansion using synonym can improve efficiency of retrieved document. The result is similar and help overcome vocabulary mismatch.

3.2 WordNet

WordNet is a database created by Ford University, in which the dictionary contains of definition and relation of word called synsets [33]. Several synsets in Wordnet is synonym, hypernim, hyponym, and meronym. Synonym is a relationship based on the similarity of words or same meaning [34]. Hyponym and hypernim have a special relationship where hypernim is a common word from hyponym or hyponym is a special relationship of hypernim. Semantics of words to form WordNet synonym sets to attain better recall and precision rate.

3.3 Word embedding

Word Embedding is a method that is done looking for the similarity of words based on the words that accompany it. The more words mentioned together the more the word looks like. There are two Word Embedding method used in this study is Global Vectors (GloVe) and Word2Vec.

3.2.1. GloVe

GloVe is an unsupervised learning algorithm to obtain a vector representation of words (word embeddings). The training process of the GloVe model is done by involving all statistical information from the corpus by forming word-of-occurrence matrix. The GloVe training process is more efficient than other techniques because GloVe only involves training on matrix elements of which value are non-zero.

GloVe generates good Word Embedding evidenced by the success rate up to 75% on the word analogy test. GloVe Word Embedding also outperformed other Word Embedding models on word-to-word similarity tests and object name recognition [35]. Words generated by GloVe are projected to contribute greatly in handling the diversity of the thesaurus. This is due to the vector representation of Word Embedding generating good results on the nearest neigbors trial.

In previous studies, the technique of Word Embedding was divided into two: the matrix factorization method and context-window based method. The matrix factorization method obtains Word Embedding by making a matrix based on the appearance of the word on the corpus and converting it into a vector of certain dimensions. The context-window based method obtains the meaning of word by learning the words that appear around it (which is in the long window range). The matrix factorization method has advantages over the context-window method in obtaining statistical information that represents the entire corpus because the context-window method only scans the context-window locally. In fact, the value of the word repetition statistics on the corpus is important information. On the other hand, learning the meaning of words from the words around them is a good idea.

$$X_{ij,t+1} = X_{ij,t} + \frac{1}{\mu};\ \mu = \{1,2,\dots,s\}. \quad (1)$$

$X_{ij,t}$ is co-occurrence term j in context word i in current time. $X_{ij,(t+1)}$ is co-occurrence term j in context word i in next time t $(t+1)$. s is length *window* matrix and the value of each context co-occurrence word is not always one and $\mu = \{1,2,\dots,s\}$ is word context distance in *window* s.

$$P_{ij} = P(j|i) = X_{ij}/X_i \tag{2}$$

P_{ij} is used to calculate probability of co-occurrence word in j in context word i. P_{ij} can be obtain with calculate mean of co-occurrence term in context word. X_{ij} is co-occurrence term j in context word i. X_i is co-occurrence context word i.

$$F\left(w_i, w_j, \widetilde{w}_k\right) = \frac{P_{ik}}{P_{jk}}. \tag{3}$$

F in Eq (3) is function consists of variables w_i, w_j, and $\widetilde{w}_k$ whose value must be close to the ratio value P_{ik}/P_{jk}. $w \in \mathbb{R}^d$ is word vector and $\widetilde{w} \in \mathbb{R}^d$ is context vector kata. F can be determined by changing the ratio P_{ik}/P_{jk} in vector to simplify calculation of distance.

$$F\left(w_i - w_j, \widetilde{w}_k\right) = \frac{P_{ik}}{P_{jk}}. \tag{4}$$

Eq (4) is modification of F function to vector. Vector can be determined to scalar with dot product operation.

$$F\left((w_i - w_j)^T \widetilde{w}_k\right) = \frac{P_{ik}}{P_{jk}}. \tag{5}$$

F function vulnerable to occur mixing dimensions because word co-occurrence matrices is arbitrary and can be exchange the two roles. F must be symmetry to avoid this issue. F must be homomorphic between ($\mathbb{R}$,+) and ($\mathbb{R}_{>0}$,x).

$$F\left((w_i - w_j)^T \widetilde{w}_k\right) = \frac{F(w_i^T \widetilde{w}_k)}{F(w_j^T \widetilde{w}_k)}, \tag{6}$$

Eq (4) is solve by

$$F\left(w_i^T \widetilde{w}_k\right) = P_{ik} = \frac{X_{ik}}{X_i}, \tag{7}$$

Eq (7) can be simplify

$$w_i^T \widetilde{w}_k = \log(P_{ik}) = \log(X_{ik}) - \log(X_i)\,. \tag{8}$$

$\log(X_i)$ not dependent on variable k, then it can be transformed into bias b_i for w_i. Equation can return to symmetry by adding $\tilde{b}_k$ for $\widetilde{w}_k$.

$$w_i^T \widetilde{w}_k + b_i + \tilde{b}_k = \log(X_{ik})\,. \tag{9}$$

w_i^T is vector word and $\widetilde{w}_k$ is vector context. b_i is bias for w_i and $\tilde{b}_k$is bias for w_k. However, there is a fallback in this equation when the element in the X matrix is zero, whereas the zero value in the X matrix can amount to 75-95% of the total element in the matrix (depending on the size of the corpus). Therefore, a zero-value element is not processed. For correction, the weighting technique is to provide a weighting function $f(X_{ij})$ in cost function obtain cost function model

$$J = \sum_{i,j=1}^{V} f(X_{ij})(w_i^T \widetilde{w}_j + b_i + \tilde{b}_j - logX_{ij})^2\,, \tag{10}$$

$f(x)$ as a weighting function formulated with

$$f(x) = \begin{cases} (x/x_{\max})^{\alpha}, & x < x_{max} \\ 1, & otherwise \end{cases}. \tag{11}$$

$\alpha = 3/4$ and $x_{max} = 100$ for all the experiments [35]. x_{max} is determined as 100 because each pair of main words and context words that appear 100 times are considered too often and change weight to 1.

3.2.2. Word2Vec

Word2Vec is used for learning word vectors which is divided into two called continuous skip-gram and continuous bag of words (CBOW) [36]. CBOW predicts the current word based in the context. Skip-gram uses maximize classification of word based on another word in same sentence. The difference between the two methods is that skip-gram does not use multiplications dense matrix, so the training process is more efficient and optimal [37]. In this paper, skip-gram is used to generate relevant candidate query. Equation of skip-gram is as follows.

$$Q = C \times \left(D + D \times log_2(V)\right) \tag{12}$$

where C is maximum distance of the word, D is the word representation and V is size of the vocabulary.

3.3 BabelNet embedding

BabelNet is a dictionary that can be built with extensive and multilingual semantic networks. BabelNet uses knowledge from WordNet and Wikipedia. WordNet is used to obtain synset relationship between words based on lexical and semantic relationship; while, Wikipedia uses relationship between entities on the Wikipedia page. The two relationships between the knowledge is combined, so that each Wikipedia page entity is interconnected using the WordNet relationship [8]. Wikipedia documents can avoid the final results of documents that are not relevant to the query compared to using dataset documents [38].

In this paper, we proposed BabelNet Embedding for query expansion. This method doesn't use knowledge on BabelNet but idea of how BabelNet work in Semantic Search to determine additional relevant documents highly similar to the query using Wikipedia document. The idea of BabelNet used in business process field to mapping Wikipedia Page to WordNet to build lexical database [39]. Fig. 3 explains that the first query will be carried out by the N-Gram process with the maximum number of queries. N-Gram will generate a word list that is used to search for documents that have the same title on Wikipedia. Based on the list word, a document search will be carried out. The results of the document search will be saved for the document extraction process. The process of document extraction is done using the Word Embedding (Word2Vec CBOW and Skip-gram). Result of document extraction is the candidate query used to help determine relevant documents.

3.4 Cross language

In this study, there is a process of changing languages from Indonesian to Arabic. This is done because the input is in Bahasa Indonesia and the output is in Arabic. The cross-language process in this study was conducted using google translate. The language change process is done after using the expansion query process. The query expansion result is changed into Arabic and searches for the Arabic e-books.

3.5 Search document

Document searches carried out in this study used the result of the extended query from the previous results. The returned document will be checked by reference based on the expert. The results of relevant documents can be evaluated using evaluation accuracy.

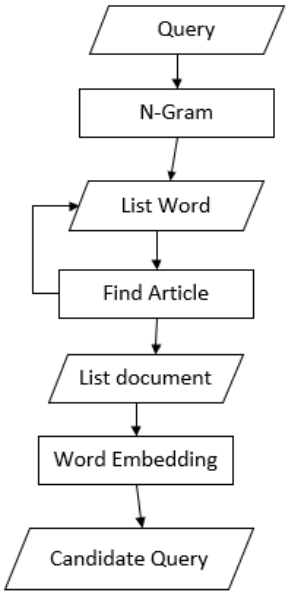

Figure 3. BabelNet embedding

4. Result and experiment

The document used in this paper is the literation and language Arabic. Arabic e-book documents were downloaded at www.shamela.ws. Indonesian knowledge data in Wikipedia were downloaded at https://dumps.wikimedia.org/idwiki/latest/ and trained in the Word Embedding and BabelNet Embedding method. The experiment data using the question posed at the previous religious forum. The results of the forum can be seen in www.piss-ktb.com.

4.1 Experiment

The experiment carried out in the study was used to search questions from the religious forums that discussed and do not have a clear legal basis. The example of the query used for experiment is "Hukum membayar zakat dengan uang (the law of paying zakat fitrah using money)". The steps taken to overcome these problems are as follows.

4.1.1. Pre-processing

The problems discussed in the forum are pre-processing and the results obtained are "bayar (pay), zakat, fitrah, uang (money)". Result of pre-processing can be seen in Table 1 and will be query tokens for document searches.

Table 1. Result of syntax query expansion

Indonesian Query	Arabic Query
Bayar (pay)	دفع
Zakat	زَكَاة
Uang (money)	نقود
Fitrah	فطر
zakat fitrah	زكاة الفطرة
zakat fitrah	زكاة الفطر
zakat fitrah	زكاة فطرة
zakat fitrah	زكاة فطر

Table 2. Result of synonym query expansion

Indonesian Query	Arabic Query
Sedekah (alms)	صدقة
Derma (charity)	مؤسسة خيرية

Table 3. Result of WordNet query expansion

Indonesian Query	Arabic Query
Sedekah (alms)	صدقة

The sample query expansion results used in this study are as follows

4.1.2. Synonym search from the query

The expansion query technique that is performed is thesauri or finding synonyms of the query. The search result of the book using the word synonym query is in Table 2. In terms of concept, it is even more different with the intention of input query questions. But searching for synonyms from words can provide an additional 1 relevant book. Because the 5 other relevant books had not been retrieved, the query is further expanded.

4.1.3. Term query search on WordNet

Searching for semantic relationship on WordNet is done by searching for synonym, hypernym, hyponym, meronym. Result of WordNet in Table 3 is the same as searching for synonym, so that the search result is the same as the previous search. It happens because most of the cases of experiment do not have the hypernym, hyponym and others.

4.1.4. Word embedding related to the query

Word embedding is done to expand the result of the expansion in order to help in finding relevant books according to references. The word embedding method that is done is the GloVe method. The GloVe method is an improved method of Word2Vec made by Stanford in the word embedding search. The GloVe method is applied using Indonesian-language Wikipedia resources to help analogy words related to search queries. Word embedding is done using only the main query: zakat. The result of word embedding "zakat" use the GloVe is in Table 4 and Word2Vec in Table 5.

Table 4. Result of GloVe query expansion

Indonesian Query	Similarity score
Infaq	0,562
Shodaqoh (alms)	0,516
Fitrah	0,505
Amil	0,451
Wakaf (waqf)	0,436
Sadaqah (alms)	0,429
Menunaikan (fulfill)	0,410
Tunaikan (fulfill)	0,387
sadaqah	0,385
Disyariatkan (required)	0,384

Table 5. Result of Word2Vec query expansion

Indonesian Query	Similarity score
infak (infaq)	0,775
wakaf	0,681
shadaqah	0,671
infak	0,659
amil	0,640
halal	0,606
fitrah	0,596
fithrah (fitrah)	0,593
kewajiban (obligation)	0,576
wadiah	0,570

The result of word embedding also adds 1 relevant book according to reference. Searching using word embedding can improve the search for relevant books if documents in the Arabic are included as resource in searching for word analogy. This allows the analogy of the resulting word to obtain relevant words according to the available books.

The experiment result with 40 queries shows that the average accuracy in this study was 90% as shown in Table 6. Experiment was carried out by calculating the documents returned and compared with relevant documents from experts. The relevant document in Table 6 is a document that originates from an expert according to the input query. Whereas, the retrieval document is the result document from the same system with the relevant document from the expert. In some cases, the query

Table 6. Accuracy of document query

Number Query	Proposed Method	Token query	Synonym	WordNet	Word2Vec	GloVe
1	**88%**	25%	50%	50%	50%	75%
2	**31%**	0%	0%	0%	0%	0%
3	**60%**	0%	0%	0%	0%	0%
4	**80%**	0%	0%	0%	0%	0%
5	50%	50%	50%	50%	50%	50%
6	100%	0%	100%	100%	100%	100%
7	**100%**	40%	**100%**	100%	40%	40%
8	**100%**	29%	43%	86%	29%	29%
9	100%	100%	100%	100%	100%	100%
10	50%	50%	50%	50%	50%	50%
11	**100%**	75%	**100%**	100%	75%	75%
12	**100%**	0%	0%	100%	0%	0%
13	**100%**	67%	**100%**	100%	67%	67%
14	**100%**	50%	50%	50%	50%	50%
15	100%	0%	100%	100%	100%	100%
16	**100%**	71%	71%	71%	**100%**	**100%**
17	**86%**	14%	57%	57%	86%	86%
18	100%	0%	100%	100%	100%	100%
19	50%	50%	50%	50%	50%	50%
20	100%	100%	100%	100%	100%	100%
21	**100%**	67%	**100%**	100%	67%	67%
22	100%	100%	100%	100%	100%	100%
23	100%	0%	100%	100%	100%	100%
24	100%	100%	100%	100%	100%	100%
25	**89%**	67%	78%	78%	78%	78%
26	**100%**	50%	**100%**	100%	50%	50%
27	100%	100%	100%	100%	100%	100%
28	67%	67%	67%	67%	67%	67%
29	100%	100%	100%	100%	100%	100%
30	**100%**	50%	**100%**	50%	50%	50%
31	**100%**	67%	67%	67%	67%	67%
32	67%	67%	67%	67%	67%	67%
33	100%	100%	100%	100%	100%	100%
34	**100%**	86%	86%	100%	100%	100%
35	**100%**	67%	**100%**	67%	67%	67%
36	75%	75%	75%	75%	75%	75%
37	100%	100%	100%	100%	100%	100%
38	**100%**	50%	50%	50%	**100%**	**100%**
39	**67%**	33%	33%	33%	33%	33%
40	100%	100%	100%	100%	100%	100%
Average	**89%**	54%	74%	75%	69%	70%

accuracy result shows a 100% result which indicates that with extended query based on semantic query expansion it can retrieve the entire document. However, there are some cases that fail to return relevant documents.

Some reference reasons are given that cannot be replaced. First, experts in conducting searches use analogies or figures of speech based on their knowledge. Knowledge of experts is needed in document search. Second, there are some of reference and the e-books in this study is the same as the experts in finding references. This can be done if experts also use the same e-book references.

Table 6 shows a comparison of evaluations using the accuracy of the proposed method with other methods: syntax, synonyms, WordNet, and Word Embedding. The result shows that the average accuracy of the proposed method has a higher value compared to the other methods. It indicates that the proposed method can obtain more relevant documents than the other methods. There are queries where using syntax is enough to obtain relevant documents; it indicates that the input query includes the necessary information. The input query also uses terms that match the terms in the document, but there is a condition that the user does not know the

term used in the document, so that the syntax is not enough to return the document.

Synonym and WordNet have the same results because WordNet also contains the word synonym too. The WordNet database also has the word hypernym, hyponym, and meronym. The experiment of adding synonyms can return relevant documents, so that WordNet and synonym have small different results. This result only contains a few relevant candidate queries, such as hypernym, hyponym, and meronym, on problem 8 and other problems only using the word synonym. Query expansion using WordNet and Paraphrase lexicon have been done to detect plagiarism candidates [40]. The paper use synonym synset to get an additional term. Paraphrase lexicon is also used to increase the results of WordNet which cannot detect phrases. In this paper, the synonym is not enough to retrieve relevant Arabic documents so that you need to add another method.

Expansion query using Word Embedding is proven to increase result compared to synonym and WordNet. Word Embedding can return relevant documents in addition to training-based queries on Wikipedia data. Training using Wikipedia completes the word semantic relationship, but the disadvantage of using all Wikipedia data is that the semantic word relationship with the query remains too general and not specific according to the context of the problem needed to be overcome. It causes a lot of noise term in the Word Embedding result. The proposed method uses Wikipedia data that has a link with the main query, so that additional queries are more in accordance with the desired context.

The proposed method has better result especially at number queries 2,3,4 where the other existing method cannot provide relevant documents. This is because the knowledge used in the proposed method has the same domain with the context of the problem. The proposed method also contain relation. In other methods, the knowledge used is relatively common and WordNet knowledge there are no words that match the problems discussed. In this study, understanding the context needs to be improved. This can be seen in Table 6 number queries 5, 10, 19 and other cases containing technical words in the religious field.

5. Conclusion

In this paper, we proposed query expansion based on BabelNet search and Word Embedding. BabelNet search is used to search Wikipedia articles coherent with queries. Word Embedding is used for Extracting terms in selected articles using Word2Vec (skip-gram and continuous bag of word). The result of term extraction is a candidate term for query expansion. Besides the proposed method, candidate term for query expansion is also obtained based on synonym, wordnet, Word embedding (Word2Vec and GloVe) with training all Wikipedia Indonesia. Candidate terms will be searched for the relevant book according to the expert according to the problem discussed. The proposed query expansion successfully retrieves relevance document in searching Arabic documents and obtains an average accuracy of 89%.

The higher performance of the proposed method compared to other method shows that it is promising for query expansion in searching Arabic document. To better understand the context of demand, future work of the research is implemented word sense disambiguation and use other knowledge especially in depend on the domain. Paraphrase lexicon is also needed to detect the phrase in the document. There are several previous studies that applied word sense disambiguation and were proven to improve results and more understanding context [41, 42].

Acknowledgments

We would like to express our gratitude to Ministry of Research, Technology, and Higher Education, Indonesia with grant number 5/E1/KP.PTNBH/2019.

References

[1] I. Moawad, W. Alromima, and R. Elgohary, “Bi-Gram Term Collocations-based Query Expansion Approach for Improving Arabic Information Retrieval”, *Arab. J. Sci. Eng.*, Vol. 43, No. 12, pp. 7705–7718, 2018.

[2] G. Mohsen, M. Al-ayyoub, I. Hmeidi, and A. Al-aiad, “On the Automatic Construction of an Arabic Thesaurus”, In: *Proc. of 2018 9th Int. Conf. Inf. Commun. Syst.*, No. April, 2018.

[3] M. Al-ayyoub, A. Nuseir, K. Alsmearat, Y. Jararweh, and B. Gupta, “Deep learning for Arabic NLP : A survey”, *J. Comput. Sci.*, Vol. 26, pp. 522–531, 2018.

[4] J. Ooi, X. Ma, H. Qin, and S. C. Liew, “A survey of query expansion, query suggestion and query refinement techniques”, In: *Proc. of 2015 4th Int. Conf. Softw. Eng. Comput. Syst. ICSECS 2015 Virtuous Softw. Solut. Big Data*, pp. 112–117, 2015.

[5] M. A. Raza, R. Mokhtar, and N. Ahmad, “A survey of statistical approaches for query expansion”, *Knowl. Inf. Syst.*, 2018.

[6] B. El Ghali and A. El Qadi, "Context-aware query expansion method using Language Models and Latent Semantic Analyses", *Knowl. Inf. Syst.*, Vol. 50, No. 3, pp. 751–762, 2017.

[7] M. A. Raza, R. Mokhtar, and N. Ahmad, "A Taxonomy and Survey of Semantic Approaches for Query Expansion", *IEEE Access*, Vol. 7, pp. 17823–17833, 2019.

[8] R. Navigli and S. P. Ponzetto, "BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network", *Artif. Intell.*, Vol. 193, pp. 217–250, 2012.

[9] Q. T. Zeng, J. Crowell, R. M. Plovnick, E. Kim, L. Ngo, and E. Dibble, "Assisting consumer health information retrieval with query recommendations", *J. Am. Med. Informatics Assoc.*, Vol. 13, No. 1, pp. 80–90, 2006.

[10] X. Mu, K. Lu, and H. Ryu, "Explicitly integrating MeSH thesaurus help into health information retrieval systems: An empirical user study", *Inf. Process. Manag.*, Vol. 50, No. 1, pp. 24–40, 2014.

[11] L. Wang, G. Del Fiol, B. E. Bray, and P. J. Haug, "Generating disease-pertinent treatment vocabularies from MEDLINE citations", *J. Biomed. Inform.*, Vol. 65, pp. 46–57, 2017.

[12] J. Singh, M. Prasad, O. K. Prasad, E. Meng Joo, A. K. Saxena, and C. T. Lin, "A Novel Fuzzy Logic Model for Pseudo-Relevance Feedback-Based Query Expansion", *Int. J. Fuzzy Syst.*, Vol. 18, No. 6, pp. 980–989, 2016.

[13] F. Moreau, V. Claveau, and S. Pascale, "Automatic morphological query expansion using analogy-based machine learning", *Eur. Conf. Inf. Retrieval. Springer*, Vol. 4425, pp. 222–233, 2007.

[14] C. Moral, A. De Antonio, and R. Imbert, "A survey of stemming algorithms in information retrieval", *Inf. Res. An Int. Electron. J.*, Vol. 19, No. 1, 2014.

[15] J. Jayanthi, "Personalized Query Expansion based on Phrases Semantic Similarity", In: *Proc. of the 3rd Int. Conf. Electron. Comput. Technol.*, Vol. 4, pp. 273–277, 2011.

[16] M. Lu, X. Sun, S. Wang, D. Lo, and Y. Duan, "Query Expansion via Wordnet for Effective Code Search", In: *Proc. of the 22nd Int. Conf. Softw. Anal. Evol. Reengineering*, pp. 545–549, 2015.

[17] S. Kapidakis, A. Mastora, and M. Peponakis, "Query Expansion of Zero-Hit Subject Searches : Using a Thesaurus in Conjunction with NLP Techniques", In: *Proc. of Int. Conf. Theory Pract. Digit. Libr.*, pp. 433–438, 2012.

[18] C. H. C. Leung, Y. Li, A. Milani, and V. Franzoni, "Collective Evolutionary Concept Distance Based Query Expansion for Effective Web Document Retrieval", *Comput. Sci. Its Appl. – ICCSA*, Vol. 7974, pp. 657–672, 2013.

[19] S. S. Laddha, A. R. Laddha, and P. M. Jawandhiya, "New paradigm to keyword search: A survey", In: *Proc. of 2015 Int. Conf. Green Comput. Internet Things*, Vol. 431001, pp. 920–923, 2016.

[20] A. Babu, "An Information Retrieval System for Malayalam Using Query Expansion Technique", In: Proc. of *2015 Int. Conf. Adv. Comput. Commun. Informatics*, pp. 1559–1564, 2015.

[21] V. Nakade, M. Aibek, and A. Travis, "Preliminary Research on Thesaurus-Based Twitter Query Expansion for Data Extraction", In: *Proc. of ACMSE 2018 Conf.*, No. 14, 2018.

[22] H. Wang, Q. Zhang, and J. Yuan, "Semantically Enhanced Medical Information Retrieval System : A Tensor Factorization Based Approach", *IEEE Access*, Vol. 5, pp. 7584–7593, 2017.

[23] L. Guo, X. Su, L. Zhang, G. Huang, and X. Gao, "Query Expansion Based on Semantic Related", *Pacific Rim Int. Conf. Artif. Intell.*, Vol. 1, No. 1, pp. 19–28, 2018.

[24] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor, "Freebase : A Collaboratively Created Graph Database For Structuring Human Knowledge", In: *Proc. of 2008 ACM SIGMOD Int. Conf. Manag. data*, pp. 1247–1249, 2008.

[25] C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives, "DBpedia : A Nucleus for a Web of Open Data", *Semant. Web*, pp. 722–735, 2007.

[26] C. Bizer, J. Lehmann, G. Kobilarov, C. Becker, and R. Cyganiak, "DBpedia - A Crystallization Point for the Web of Data", In: *Proc. of IEEE Fourth Int. Conf. Semant. Comput.*, 2010.

[27] J. Lehmann, A. Jentzsch, and D. Kontokostas, "DBpedia – A Large-scale , Multilingual Knowledge Base Extracted from Wikipedia", *Semant. Web*, No. January, pp. 167–195, 2015.

[28] M. Kalender, J. Dang, and S. Uskudarli, "UNIpedia : A Unified Ontological Knowledge Platform for Semantic Content Tagging and Search", In: *Proc. of 2010 IEEE Fourth Int. Conf. Semant. Comput.*, pp. 293–298, 2010.

[29] M. U. Devi and G. M. Gandhi, "Wordnet and Ontology Based Query Expansion for Semantic Information Retrieval in Sports Domain", *J.*

Comput. Sci., Vol. 11, No. 2, pp. 361–371, 2015.
[30] A. Zouaghi, L. Merhbene, and M. Zrigui, "Combination of information retrieval methods with LESK algorithm for Arabic word sense disambiguation", *Artif. Intell. Rev.*, Vol. 38, No. 4, pp. 257–269, 2012.
[31] F. Diaz, B. Mitra, and N. Craswell, "Query Expansion with Locally-Trained Word Embeddings", *arXiv Prepr. arXiv1605.07891*, 2016.
[32] F. C. Fernández-Reyes, J. Hermosillo-Valadez, and M. Montes-y-Gómez, "A Prospect-Guided global query expansion strategy using word embeddings", *Inf. Process. Manag.*, Vol. 54, No. 1, pp. 1–13, 2018.
[33] P. Pambudi, R. Sarno, and E. Faisal, "Searching Word Definitions in WordNet Based on ANEW Emotion Labels", *Int. Semin. Appl. Technol. Inf. Commun.*, pp. 253–256, 2018.
[34] T. S. Utomo and R. Sarno, "Emotion Label from ANEW dataset for Searching Best Definition from WordNet", *Int. Semin. Appl. Technol. Inf. Commun.*, pp. 249–252, 2018.
[35] J. Pennington, R. Socher, and C. D. Manning, "GloVe : Global Vectors for Word Representation", In: *Proc. of 2014 Conf. Empir. methods Nat. Lang. Process.*, pp. 1532–1543, 2014.
[36] T. Mikolov, G. Corrado, K. Chen, and J. Dean, "Efficient Estimation of Word Representations in Vector Space", *arXiv Prepr. arXiv1301.3781*, pp. 1–12, 2013.
[37] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Distributed Representations of Words and Phrases and their Compositionality", *Adv. Neural Inf. Process. Syst.*, pp. 3111–3119, 2013.
[38] A. Keikha, F. Ensan, and E. Bagheri, "Query expansion using pseudo relevance feedback on wikipedia", *J. Intell. Inform. Syst.*, No. 50, pp. 455–478, 2018.
[39] E. W. Pamungkas, R. Sarno, and A. Munif, "B-BabelNet : Business-Specific Lexical Database for Improving Semantic Analysis of Business Process Models", *Telkomnika*, Vol. 15, No. 1, pp. 407–414, 2017.
[40] R. Muhammad, A. Nawab, M. Stevenson, and P. Clough, "Retrieving Candidate Plagiarised Documents Using Query Expansion", *Adv. Inf. Retr.*, pp. 207–218, 2012.
[41] S. Bagus and R. Sarno, "Adapted Weighted Graph for Word Sense Disambiguation", *Int. Conf. Inf. Commun. Technol.*, Vol. 4, 2016.
[42] F. Nurifan, R. Sarno, and C. S. Wahyuni, "Developing Corpora using Word2vec and Wikipedia for Word Sense Disambiguation", *Indones. J. Electr. Eng. Comput. Sci.*, Vol. 12, No. 3, pp. 1239–1246, 2018.