Int J Speech Technol (2016) 19:191–201
DOI 10.1007/s10772-015-9284-6

Semantic similarity based approach for reducing Arabic texts dimensionality

Arafat Awajan[1]

Received: 19 February 2015 / Accepted: 28 May 2015 / Published online: 9 June 2015

Abstract An efficient method is introduced to represent large Arabic texts in comparatively smaller size without losing significant information. The proposed method uses the distributional semantics to build the word-context matrix representing the distribution of words across contexts and to transform the text into a vector space model (VSM) representation based on word semantic similarity. The linguistic features of the Arabic language, in addition to the semantic information extracted from different lexical-semantic resources such as Arabic WordNet and named entities' gazetteers are used to improve the text representation and to create word clusters of similar and related words. Distributional similarity measures have been used to capture the words' semantic similarity and to create clusters of similar words. The conducted experiments have shown that the proposed method significantly reduces the size of text representation by about 27 % compared with the stem-based VSM and by about 50 % compared with the traditional bag-of-words model. Their results have shown that the amount of dimension reduction depends on the size and shape of the windows of analysis as well as on the content of the text.

Keywords Text dimensionality reduction · Distributional semantics · Word-context matrix · Semantic vector space model · Arabic language processing · Word similarity

✉ Arafat Awajan
awajan@psut.edu.jo

[1] The King Hussein Faculty of Computing Sciences, Princess Sumaya University for Technology, Khalil Saket Street, Al-Jubaiha, P.O. Box 1438, Amman 11941, Jordan

1 Introduction

Texts are high-dimensional objects in which every word may be considered as an independent attribute. With the increasing size of available texts in electronic form, reducing text dimensionality becomes an important issue to be addressed in many natural language processing (NLP) tasks, as these tasks perform well with low-dimensional texts and poorly in high-dimensional texts. Therefore the techniques that reduce the text dimension while minimizing information loss are essential, as they positively impacts the time and space efficiency and improves the quality of different NLP tasks' results (Martins et al. 2003).

Different techniques have been proposed for reducing text dimensionality. They are always incorporated in the model used for text representation. They aim to identify a low-dimensional representation of original text through eliminating redundancy of terms and shortening the number of features representing the text in preserving, as much as possible, the original text content. The reduction of text dimensionality can be tackled based on knowledge specified manually by experts, derived automatically from corpus statistics, or computed from linguistic resources. The accumulated knowledge from these different levels of analysis leads to special feature selection mechanisms able to reduce the text dimension without losing significant information.

Generally, NLP applications receive textual documents as "bags of words". Two main representations of texts are used in literature: feature vector representation and graph representation. In the feature vector representation, a document is presented by a vector built according to the vector space model (VSM), where the components represent the different features of the text, principally its terms or words (Salton et al. 1975). The graph representation is an

alternative representation in which the terms are represented by nodes and the relations between them, such as their co-occurrence, are represented as edges (Mihalcea and Tarau 2004; Biemann 2006). However, these two different representations have limitations related to their high dimensionality. In addition, the VSM representation suffers from lack of semantics relations among its components.

The distributional semantics (DS) approach has been used as a paradigm to model languages and represent naturally occurring texts. It is a statistical-based model that uses the statistical distribution of words along with their contexts to determine the degree of semantic similarity between them. This model describes the words by context-vectors built on the distributional hypothesis, which states that similar words appear in similar contexts (Hagiwara 2008).

Our objective is to demonstrate that using the linguistic features of natural languages may yield to improve the text representation of texts and can be used to reduce the dimensionality by removing redundancies. We extend our previous work on Arabic text representation that aimed to obtain normalized and compact text representation by investigating the possibility of using the rich morphological structures of Arabic language for building a new, semantically enriched and reduced VSM of semantics of Arabic text (Awajan 2015). We have incorporated lexical relations and semantic information from word thesauri such as the Arabic WordNet and from named entities (NE) gazetteers to improve the text representation by adding semantic information and relations to this representation. The proposed method creates clusters of similar or related words extracted from the same root or stem and regrouped along with their synonyms. The distributional similarity is used for discovering words' semantic information that allows us to group similar words and eliminate word redundancy from the text (Hagiwara 2008).

This paper is organized as follows. After overviewing related works in Sect. 2, the features and characteristics of Arabic language are described in Sect. 3. We present the different NLP tools used for preprocessing the text and extracting its main linguistic features in Sect. 4. Section 5 describes the proposed model. The results and the evaluation of the proposed system are summarized in Sect. 6.

2 Related work

The VSM was proposed by Gerard Salton in the early 1970s in order to have a dense representation of textual document in which the terms are listed and weighted based on their frequencies in the text (Salton et al. 1975). Since its introduction, the VSM has had great success in the field of NLP, mainly in indexing, information retrieval, information extraction, document categorization, and keyword extraction.

The VSM was introduced to represent collections of documents. It has been used for measuring the similarities of natural language texts. In a collection of n documents that contains m unique terms, each document D_i is represented by a vector $D_i = < d_{i1}, d_{i2},\ldots, d_{im} >$ of dimension m that represents a vector in the vector space. The component d_{ij} represents the frequency or the weight of the jth term in the document D_i. Then a collection of documents can be represented as a term-by-document matrix of column vectors D_i such that the m rows represent terms and the n columns represent documents. In the vector D_i, the sequential order of the words, the structure of phrases is lost. However, these vectors mainly capture the similarity of documents that may be seen as an important aspect of semantics (Turney and Pantel 2010).

Turney and Pantel (2010) surveyed three different classes of VSMs: term-document, word-context, and pair-pattern based representations yielding three different classes of applications. The term-document representation is used to measure the documents' similarity, the word-context measures the similarity of words, and the pair-pattern measures the similarity of relations. The word-context matrix, where the focus is on the word vectors, is used by different authors to extract semantic representation of words and to measure the words' similarity (Bullinaria and Levy 2012). In this representation, the context is given by words, phrases, sentences, or such patterns. The context is always very difficult to define, but in written text, the context of a word is often given by its neighbor words that occur in the same sentence; hence it can be measured by the co-occurrence frequency.

Reducing VSM dimensionality is one of the active research areas in the information retrieval and NLP research communities. There are two different approaches for dimensionality reduction: language-independent approach and language-dependent approach. Among the language-independent reduction methods, the singular value decomposition and independent component analysis are the most common. They reduce the dimensionality of the vector space by providing a reduced rank approximation in the column and row space of the document matrix (Baker 2013). This approach ignores the linguistic features of the text's language and considers the words as abstract orthogonal dimensions.

The language-dependent techniques investigate the use of linguistic features of the text language to reduce the representation of text. Van Rijsbergen (1979) suggested the usage of the language-dependent approach based on stemming techniques for reducing the size of index term and therefore achieving a high degree of relevancy in

information retrieval. He found that the text representation based on stems reduces the size of the document by 20–50 % compared with the full words representation.

Works on Arabic text representation and dimensionality reduction used several approaches. Most of these works are based on language-dependent techniques. They use stems of words for building the language model and representing the manipulated texts (Duwairi et al. 2009; Harrag et al. 2010; Froud et al. 2012). However, other alternatives may be used to achieve a more concise representation of Arabic text. Duwairi et al. (2009) compared three techniques for reducing Arabic texts: stemming, light stemming, and word clustering. Stemming techniques replace the words by their roots, light stemming removes suffixes and prefixes attached to the words, and word clustering groups synonyms. They found that the stemming technique gives the best results in term of size reduction. Harrag et al. (2010) compared and evaluated five dimension reduction techniques: stemming, light-stemming, document frequency (DF), term frequency–inverse document frequency (TF–IDF), and latent semantic indexing (LSI). The results showed that the DF, TF–IDF, and LSI techniques are more effective and efficient than the two other methods.

The analysis of the previous studies on Arabic texts shows that they ignore the semantics of words, which leads to regrouping terms that may have different meaning according to their contexts; hence they may produce errors in the results of the NLP applications. In addition, using light stemming techniques creates multiple entries in the VSM for different words that carry the same meaning or concepts, yielding a larger size of the representation. For example, the words (schools- المدارس), (teaching—الدراسة), and (teache—مدرسون) have different stems, while they are generated from the same root (teach, d r s—درس) and carry related key information. However, grouping words according to their root may generate incoherent results as different words referring to different concepts could in some cases share the same root. Therefore, there is a need to develop new techniques that take all these considerations into account and result in a more accurate and reduced representation of Arabic texts.

3 Linguistic features of Arabic words

The most important issue facing the representation of Arabic text and the reduction of its dimensionality is the abundance of unique word forms resulting from its rich morphology (Hmeidi et al. 1997). Figure 1 shows the growth in vocabulary of Arabic compared to the growth in English (Kirchhoff et al. 2006; Heintz 2010). To resolve this problem, we propose to study the features of Arabic language that negatively impact its VSM representation and to use its relevant properties to obtain more reduced representation of texts without losing information.

Arabic language is characterized by its complex inflectional and derivational morphology system able to generate a large number of words from well-defined basic forms called roots. The word formation rules in Arabic are characterized by the presence of templatic morphemes in addition to concatenative morphemes (Habash 2010). In the templatic morphology of Arabic, the word stems are derived from the roots according to a predefined set of patterns or templates, which results in the creation of many lexical variations. The root associated with a stem generally carries the main concept or abstract meaning of the word, and its pattern involves its possible part of speech and has a predictable semantic effect on the word. For example, the words (لعب،لاعب، لعبة، ملعب) (in English: play, player, game, playground) are generated from the same root (to play: لعب) according to four different patterns (Awajan 2007).

In its concatenative morphology, Arabic language allows adding two close classes of concatenative morphemes to stems in order to create the surface form of words: the affixes (prefixes and suffixes) and the clitics (proclitics and enclitics). The affixes determine the word's various attributes, such as person, gender, number, and tense, while the clitics, which in other languages such as English appear as separate words, are concatenated to the words to indicate definiteness, conjunction, various prepositions, and possessive forms (Shaalan 2014). These clitics can be attached to nouns "فمدرسينهم", to verbs "ويعاملونهم" as well as to each other "فيهم". The concatenative morphology of Arabic language allows the creation of an important number of variants of a same single word stem. Beesley (1998) estimated that the various combinations of prefixes and suffixes with stems generate more than 72,000,000 abstract words.

Based on the above, we can classify Arabic words into two categories: derivative and non-derivative words. The derivative words are generated according to the derivational and inflectional rules where a word pattern usually combines with a vast number of roots. The non-derivative words include fixed words and words borrowed from foreign languages, even though words borrowed from other languages can also be used to derive new words according to Arabic derivational rules. The fixed Arabic words are generally functional words or stop words such as pronouns, prepositions, conjunctions, question words, and the like.

Other challenges imposed by the language should be faced. First, certain letters can be written in different ways, leading to sparser representation. For example, the letter ALIF may appear with or without HAMZA or MADDA. Second, special marks called diacritics are used as short vowels and may appear optionally in texts. Although these

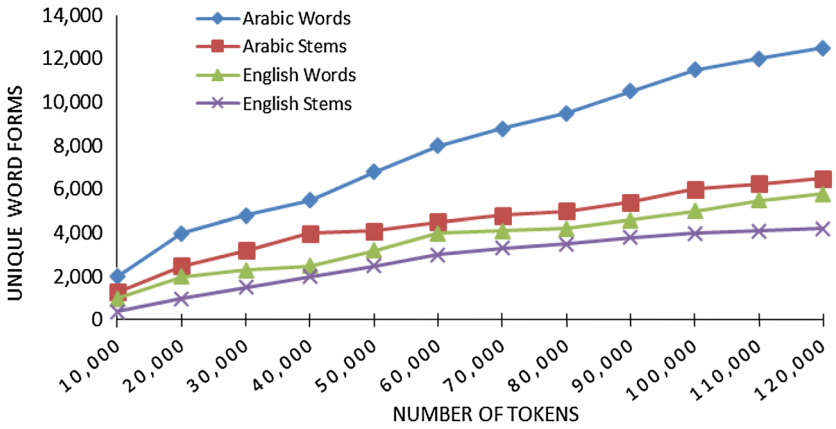

Fig. 1 Growth of unique words in Arabic and English texts (Heintz 2010)

marks are omitted in most Modern Standard Arabic (MSA) texts, we can still find words in some texts with these marks (Xu et al. 2002). Third, Arabic language is rich in synonyms because they are in general appreciated in written Arabic texts as they are considered to be an element of good writing style (Hasnah and Al-Ja'am 2002). However, manipulating synonyms or grouping them is challenging, as perfect synonyms are rare. We need always to check the context of the words to decipher their meaning and determine if they are synonyms carrying the same information.

Furthermore, the recognition of NE, which is one of the most important task in NLP applications, is another important challenge facing the representation and semantic analysis of Arabic texts. This difficulty is due mainly to the lack of capitalization, the lack of uniformity of writing styles and the shortage of available and annotated resources. Unlike in European languages, capitalization is not a distinguishing orthographic feature of Arabic script for recognizing NEs such as proper names, acronyms, and abbreviations, which leads to ambiguities in the interpretation and recognition of NEs (Shaalan 2014). For example, the word "البحرين" may be interpreted as a named entity representing the name of the Arabic country "Bahrain," or it may be interpreted as "the two seas" in the sentence "مرج البحرين يلتقيان". Furthermore, most NE are difficult to differentiate from common nouns and adjectives. For example, the company name "شركة مذيب حداد: Mudiab Haddad Company" may be considered as three words: two common nouns, "حداد: blacksmith" and "شركة: company," and an adjective, "مذيب: solvent." Thus, the morphological analysis of the texts should include a component able to detect and extract the NE before performing the analysis at the word level.

4 Text preprocessing

The proposed method proceeds in four phases, namely: text preprocessing, word-context matrix construction, similarity measurement, and semantic VSM representation. Different NLP tools are used in the preprocessing stage in order to tokenize the text, normalize letters, recognize NE, extract stems and roots, and remove stop words. It is applied and tested using undiacritized MSA texts. MSA texts are used today in written media, official speeches, and lectures, and they represent most of the available electronic texts throughout the Arabic world.

The first phase preprocesses the text and transforms it into a sequence of tokens in which each one is labeled to identify its category: derived, non-derived, and stop words, and, for the derived words, their root and pattern. The result of the text processing phase is a new presentation in which each word except the NE is described according to the following general structure where the characters [] delimit the optional components:

$$[\text{Proclitic(s)}] + [\text{Prefix(es)}] + \textbf{Stem}[\text{Root} + \text{Patten}] + [\text{Suffix(es)}] + [\text{Enclitic}].$$

4.1 Tokenization and normalization

The preprocessing phase starts with the text tokenization. The tokenizer breaks the text into sentences and tokens (words) by detecting the sentences' boundaries and isolating the individual words. As the clitics represent tokens that are attached to the word (preposition, conjunction, definite article, or object pronoun), they have to be taken out from the word and isolated as independent tokens. The tokenizer works in two passes; the first pass isolates the sentences and the words; the second pass removes all the clitics (proclitics and enclitics) from each detected word.

Normalization of letters is also needed to solve the problem we face when some Arabic characters are written in several ways which create different surface forms for the same word. This is the case we face with the letters ALF "أ" and YA "ي". For example, in modern writing style, the word أحمد is usually written as احمد. In this work, by normalization, we mean transforming all these different shapes of the same letter into a normalized one.

4.2 Named entities identification

The extraction of NEs is important for both the reduction of text dimensionality and the semantic analysis of text. As previously described, NE recognition for Arabic text is challenging due to the rich morphology of Arabic language and its script. Our approach in dealing with NE recognition and extraction has been limited to satisfy our objective in obtaining reduced representation of texts. We integrated a manually constructed gazetteer of NEs in the preprocessing phase of our proposed system. The text is scanned to recognize and extract the NEs and to replace each one by a code representing its entry in the gazetteer. This code conveys additional semantic features of the NE, such as its category (names of persons, organizations, locations, date, etc.).

4.3 Morphological analysis

Morphological analysis is used to reduce the words to their basic form and to obtain information on the morphological structure of each word. All the texts' words except the previously detected NEs are analyzed to extract their stems. The extracted words' features are then used to determine the category of each word (stop words, derivative words, or non-derivative words) as well as its morphological structure.

The text words are analyzed using both the Alkhalil Morph-Syntactic System (AMSS) (Boudlal et al. 2010) and the Stanford Arabic part-of-speech (POS) tagger (Green and Manning 2010). These two analyzers are freely available at their official websites. The AMSS identifies all the possible morphological and syntactic features, specifically proclitics, prefixes, stem, word type, word pattern, word root, POS, suffixes, and enclitics. Meanwhile, the Stanford Arabic parser provides the POS tag associated with the words, given that they are already tokenized. The POS tags provided by the Stanford Arabic parser are compared with those provided by AMSS, and only compatible solutions provided by AMSS will be retained as final features of the word. A simple greedy regular expression expression-based stemmer is developed to extract the stems of non-derivative words that AMSS fails to analyze. This stemmer is repeatedly applied until the word stops changing, producing a new representation of each word as a sequence of clitics, suffixes, and stems. The preprocessing phase assigns to each input word a category (derivative, non-derivative, or stop word), its stem, its POS tag, and the root and pattern for the derivative words. Table 1 illustrates the results of the preprocessing phase applied to the sentence "وسيعالجونها بالحاسبات" which means in English "and they will process it by the computers".

4.4 Stop word removal

Although stop words are limited in number, their total number would be approximately one-third of the total number of words in normal text. Their lexical meaning is not clearly separable from their surrounded words, and they are generally considered uninformative terms with little semantic discrimination power.

A table of stop words is used in the removal process. This table includes the most frequent words in the language that are considered as uninformative terms in addition to the list of clitics. This process reduces the size of the text by about 35 %. There are two basic advantages to removing these words. First, it reduces the size of the text representation. Second, it allows computing similarity between sentences to be more accurate and easier to depict (Salton and McGill 1986).

The text is transformed at the end of the preprocessing phase into a basic vector space S, where each entry represents a named entity or a stem associated with the number of words in the text generated from this stem. The main results of the morphological analyzer such as the stem's category (derived or non-derived), its morphological structure (pattern, root), and part of speech are also provided with each entry of S. The list of items in S represents on average <60 % of the original size of the text bag of words. The delimiters of sentences are saved at this level because the sentence information is needed in the semantic analysis of words, as this analysis is done only at the level of sentences.

5 Semantic similarity analysis

Our approach to reducing the dimensionality of Arabic text is based mainly on the analysis of semantic similarity. Similar words are regrouped in clusters and represented by a single index. The DS principle is used to identify the similar words based on analysis of the contexts where the words appear in the text. The analysis of word context is conducted at the local level analysis of the word occurrence and used to compute the word-context matrix. The semantic of a word is defined by the set of contexts in which it occurs in texts. A window of analysis that represents a sequence of words surrounding the current word is defined to represent the local context of the word.

5.1 Word-context matrix

The word-context matrix is the result of extracting information about semantic properties of words using text-based statistics, which is extremely common in semantic-based NLP applications. This matrix can be used to induce the

Table 1 Results of the tokenizer passes

Text	وسيعالجونها بالحاسبات								
Pass 1 (Tokenization)	بالحاسبات				وسيعالجونها				
Pass 2 (Tokenization)	حاسبات		ال	ب	ها	سيعالجون			و
Stemming	ات	حاسب	ال	ب	ها	ون	يعالج	س	و
Morpheme type	Suffix Plural feminine	Stem	Clitic definite article	Clitic preposition	Clitic object pronoun	Suffix plural	Stem	Prefix Future tense	Clitic conjunction (and)
Retained stems	Computer: حاسب				Process: يعالج				
Root extraction	حسب				عالج				

semantic similarity and the aspect of the meaning of words represented by their stems from their contexts in text.

The word-context matrix is derived from basic counts of words appearing in the predefined contexts. It is a co-occurrence matrix D representing a document where the row i represents an entry S_i from the vector S produced at the end of the preprocessing phase, and the column j represents a context C_j. Herein, a context C_j of a stem S_i is given by another stem of the vector S that occurs with S_i in the same window. Consequently, the element $D_{i,j}$ of this matrix represents the number of times that the stem S_i occurs with the context C_j (or the stem S_j). The raw $D_{i:}$ corresponds to one of the vector S entries, and the column $D_{:j}$ represents a context. Table 2 represents a set of stems and a set of contexts, their meaning in English and their frequency in a selected text.

Table 3 shows the word-context matrix of a sample text calculated for a window with variable size determined by the boundaries of the current sentence. The working window is passed over the text being analyzed, and the words and contexts within this window are recorded. The element (i,j) of the matrix represents the number of times that the stem S_i appears in the context C_j.

The word-context matrix is used to measure the words' similarity and to build a new representation of the text. In this presentation, the similar words under certain conditions are clustered and considered as one entry in the VSM. Our approach is based on the distributional hypothesis, declaring that words occurring in similar contexts tend to have similar meanings (Harris 1954). Two words have the same context if they always occur with the same words, which can be translated by having similar vectors in the word-context matrix.

The pointwise mutual information (PMI) is commonly used instead of the direct count of the frequency of co-occurrence of pairs of words S_i and C_j (word and context). PMI measures how often a word S_i occurs in a context represented by C_j, compared with what is expected if they were independent. The PMI is given by:

$$PMI(S_i, C_j) = \log_2 \frac{P(S_i, C_j)}{P(S_i)P(C_j)}$$

Table 2 The words and contexts of the selected window and their count

Selected stems							
Symbol	S1	S2	S3	S4	S5	S6	S7
Stem	مدرسة	شركة	منشأة	معهد	حراسة	عناية	اهتمام
English meaning	School	Company, corporation	Plant, facility	Institute	Guarding	Care, attention	Interest
Count	6	5	4	6	3	2	2
Selected context							
Symbol	C1	C2	C3	C4	C5	C6	C7
Context	مدرس	مشروع	مهندس	تدريب	مدير	رأسمال	طالب
English meaning	Teacher, tutor	Project	Engineer	Training	Director	Capital	Student
Count	9	5	4	6	4	5	9

Table 3 The word-context matrix of counts computed for Window matching sentences' boundaries

Stems	Contexts						
	C1	C2	C3	C4	C5	C6	C7
S1	4	0	0	2	1	0	5
S2	0	2	2	0	1	2	0
S3	0	2	1	1	1	2	0
S4	3	0	1	3	1	0	4
S5	0	3	0	0	1	1	0
S6	2	0	0	0	0	0	1
S7	2	0	0	1	0	0	1

where $P(S_i, C_j) = \frac{D_{i,j}}{\sum_{k=1}^{N}\sum_{l=1}^{N} D_{l,k}}$, $P(S_i) = \frac{\sum_{j=1}^{N} D_{i,j}}{\sum_{k=1}^{N}\sum_{l=1}^{N} D_{l,k}}$, and $P(C_j) = \frac{\sum_{i=1}^{N} D_{i,j}}{\sum_{k=1}^{N}\sum_{l=1}^{N} D_{l,k}}$.

However, the word-matrix is sparse, with most of the entries equal to zero, representing unseen combinations of stem-context. As the PMI is biased toward these infrequent situations, we use the Laplace smoothing algorithm before applying the PMI. This algorithm adds 1 to all the entries. Finally, we use the positive pointwise mutual information (PPMI) to represent the word-context matrix, as the PPMI replaces the negative values of PMI with zeros.

To reduce the dimension of the texts, we propose to regroup the words using two different types of knowledge. The first type corresponds to the knowledge that can be extracted from the word-context matrix, which mainly measures the distributional similarity of words based on their contextual appearance in the text. The second is based on knowledge obtained from linguistic resources, mainly the morphological structure of words, the list of potential synonyms in the language, and the co-referent NEs. For example, the stems generated from the same root and their synonyms are considered as candidates to be regrouped if the first level of knowledge is satisfied, as such words with the same contextual appearance tend to represent the same concept.

5.2 Similarity measures

The similarity between pairs of words is a measure of the degree of correspondence between their contexts and can be seen as the distance between their two vectors in the context space represented by the word-context matrix. The more correspondence there is, the smaller the distance between them and the greater their similarity. Let S_i and S_j are two stems of the text represented by the two row vectors $D_{i:}$ and $D_{j:}$ where $D_{i:}$ is the vector $< D_{i,1}, D_{i,2},\ldots, D_{i,N} >$ and $D_{j:}$ is the vector $< D_{j,1}, D_{j,2},\ldots, D_{j,N} >$. Two words are similar in meaning if their vectors are similar, whereas they have similar neighbors.

Different distances are always used to capture the similarity of words including the Euclidean distance, cosine distance, Jaccard distance, and Dice distance. The distance between two word-vectors $D_{i:}$ and $D_{j:}$ is used to measure their semantic similarity where the points that are close together are semantically similar and points that are far apart are semantically distant (Turney and Pantel 2010). In this work, we consider the cosine distance as a measure of the words' similarity. Cosine similarity encodes the similarity between two words by giving the cosine of the angle between their corresponding vectors. The similarity of the two stems S_i and S_j represented in the word-context matrix by the vectors $D_{i:}$ and $D_{j:}$ representing the PPMI values for the stems S_i and S_j is given by:

$$\mathrm{CosSim}(S_i, S_j) = \frac{D_{i:} * D_{j:}}{|D_{i:}||D_{j:}|} = \frac{\sum_{k=1}^{N} D_{i,k} * D_{j,k}}{\sqrt{\sum_{k=1}^{N} D_{i,k}^2}\sqrt{\sum_{k=1}^{N} D_{j,k}^2}}$$

Distributional similarity has been widely used to detect similar words that occur in the same context. However, these captured words are not necessarily carriers of the same meaning or concept. Therefore, the results of the similarity measurement should be considered only for the words that may have similar meaning: the potential synonyms and the stems generated from the same root.

5.3 Grouping similar name entities

An NE may appear in different forms, and several co-reference expressions referring to the same entity may be found in a text. For dealing with the problem of matching co-referent NEs in text, we manually reshape the manually constructed gazetteer of NEs by grouping the names entities referring to the same entities in clusters. In the manipulated text, all the NEs belonging to the same cluster and referring to the same entity will be replaced by the shortest NE considered the cluster representative. Table 4 represents some entries of the clustered NE.

5.4 Synonym grouping

Synonym is defined by Merriam-Webster On-Line as "one of two or more words of the same language that have the same or nearly the same meaning in some or all contexts". Two words are synonyms if they are syntactically identical, and the substitution of Y for X in a declarative sentence does not change its meaning. Synonyms cover all kind of words (noun, adjectives, verbs, and adverbs), but identifying them is a challenging task, as perfect synonyms are rare. Two words may often be considered to be synonyms

Table 4 Named entities clusters

Named entity cluster entry	English meaning	Examples of clustered NE members
الأردن	Hashemite Kingdom of Jordan	المملكة الاردنية الهاشمية، الأردن، المملكة الاردنية، الدولة الاردنية، ...
الملك السعودي	The King of Saudi Arabia	الملك السعودي، خادم الحرمين، العاهل السعودي، عاهل المملكة العربية السعودية، ...
بن باديس	Abdelhamid Ben Badis	بن باديس، ألامام بن باديس، الشيخ بن باديس، الامام عبد الحميد بن باديس، عبد الحميد بن باديس، ....
يناير	January	كانون أول، جانفيه، يناير ...
الملكية الاردنية	Royal Jordanian	الملكية الاردنية، الخطوط الجوية الملكية الاردنية، شركة الطيران الاردنية، ...
الواشنطن بوست	The Washington Post	صحيفة الواشنطن بوست، جريدة الواشنطن بوست، الواشنطن بوست، صحيفة واشنطن بوست، جريدة الواشنتن بوست، ...

in some contexts but not in other contexts. Therefore, the process of synonym grouping must address this issue by considering the contextual information measured by the cosine similarity. For the purpose of this work, two stems are synonyms of each other if they satisfy the following conditions:

1. They have the same POS.
2. They are linguistically considered as potential synonyms.
3. They appear in the text in similar context.

To address this problem we combine the distributional similarity with available language resources that can provide sets words' synonyms. A table of synonyms is built as a prototype for testing the proposed method. It includes synonyms from different linguistic resources: the Arabic WordNet (AWN), Almaany (2014) and Parkinson (2005). Table 5 shows some entries of the synonyms tables. AWN is composed of groups of near-synonyms, instantiates a sense or concept, called synsets (synonym sets). It is constructed according to the development process of Princeton WordNet and Euro WordNet as an open-source electronic lexical database. It provides a list of synsets related to a given term and the relationships with other concepts, as well as information about the corresponding English/Arabic synsets (Elkateb et al. 2006). These resources provide lists of potential synonyms, but they don't themselves provide a word-pair similarity metric.

The following algorithm represents the detection of synonyms task:

1. For each word, we extract all the possible synonyms from the table of synonyms.
2. For each stem S_i, compare its synonyms list against all the other stems S_j ($j = 1,\ldots, N$).

Table 5 Examples from the table of synonyms

Word types	Arabic word	English meaning	Synonyms
Nouns	أصل	Origin, source	أرومة، جذع، جذر، مرجع، مَحْتِد، أساس، قاعدة، مبدأ، قاعدة، مصدر، منشأ، عِرْق، مَنْبِت، نَبْعَة
	حجاب	Veil	برقع، خمار، ستر، ستار، قناع، لثام، نقاب، غطاء، ستارة، حاجِز، ساتِر، فاصِل، مانِع، مُغَطٍّ، سَدّ، فاصِل
	جهاز	Apparatus	أداة، آلة، عدة، لوازم، ماكينة، هيئة، عضو، مجلس
Verbs	ترك	To leave	برح، بارح، خرج، تخلى، انسل، انصرف، غادر، كف، توقف، تخلى، أقلع، خلف، استسلم أَبْقَى ، إِعْتَزَلَ، إِفْتَرَقَ، إِنْفَصَلَ تَجَنَّبَ، تَحَاشَى، فارَقَ، هَجَرَ، زَهِدَ
	بدأ	To begin	ابتدأ، دشن، شرع، طفق، استهل، أحدث، أنشأ، أسس، أوجد، إِسْتَهَلَّ، إِفْتَتَحَ ، باشَرَ، إِنْبَلَجَ
	جاء	To come	ورد، أتى، حضر، قدم، ظهر، أَقْبَلَ، زار، أَقْبَلَ، قَدِمَ
Adjective	ثابت	Firm	محكم، راسخ، متين، وطيد، وثيق، مستقر، صلب، محقق، مؤكد، أكِيد، باقٍ، دائِم، مَكِين

2.1. If S_j is a synonym of S_i and they have the same POS

2.1.1. Calculate the similarity measure CosSim(S_i,S_j).

2.1.2. If CosSim(S_i,S_j) > Threshold, merge the two vectors $D_{i:}$ and $D_{j:}$ by accumulating their statistics, and remove the row $D_{j:}$ from the word-context matrix.

3. End.

5.5 Stems grouping

The derivational morphology system in Arabic allows the production of different stems from the same root according to different patterns or templates. The different stems generated from the same root may or may not carry the same meaning according to the meaning added by the pattern. For example, the stems (دراسة مدرس، دارس،) are generated from the same root (درس) according to different patterns, and they refer to the same concept. On the contrary, the words (كتاب، اكتتاب) respectively (subscription, book) carry different meanings, although they are generated from the same root (كتب). Therefore, we have to consider the context, as two different stems generated from the same root may have different meanings if they appear in different contexts.

The different stems generated from the same roots are first tested to calculate their similarity. They are grouped together, and their statistics are accumulated if they are found to be similar. This operation considerably reduces the number of rows in the word-context matrix without changing the number of contexts in order to keep the contexts as varied as possible. The following algorithm represents the clustering stems generated from the same root:

1. For each stem S_i, compare its root against all the other stems roots S_j (j = 1,…, N and i ≠ j).

1.1. If S_j is derived from the same root as S_i,

Calculate the similarity measure CosSim (S_i,S_j).
If CosSim(S_i,S_j) > Threshold, merge the two vectors $D_{i:}$ and $D_{j:}$ by accumulating their statistics, and remove the row $D_{j:}$ from the word-context matrix.

2. End.

5.6 Text representation

The new representation of the text is built from the final word-context matrix after merging related or similar rows. The document is transformed at the end of this phase into a semantic vector space model where each entry represents an equivalence class that clusters terms representing equivalent meaning and concept. The new representation has two types of entries: NE and stems. Each cluster of stems is represented by the most frequent of its members. On the other hand, the shortest named entity is selected to represent the list of co-reference NE found in the text. A weight is provided with each entry representing the accumulated frequencies of all its cluster members. The entry weight is used to reflect the importance of the concept or meaning associated with the cluster in the document.

6 Experiments

We have tested the proposed approach on a set of documents from the BBC Arabic news collection in order to evaluate its performance in term of dimensionality reduction (Saad and Ashour 2010). We select texts belonging to six different categories: Middle East news, World news, Business, Sport, Sciences, and Arts. Texts from the same category were then merged together to produce six text files of different sizes.

We have used the publicly available Alkhalil Morph-Syntactic System (AMSS) and Stanford Arabic part-of-speech tagger in addition to home-built modules that combine the results of the two systems to come up with more accurate morphological analysis of texts. A built-in list of stop words, a manually constructed table of NE and a table of synonyms are used in the implementation of the proposed method. The performances of the different text representations are described in terms of the dimension reduction ratio (DRR) defined by:

$$DRR = \frac{Size\ of\ the\ Reduced\ Representation\ of\ Text}{Size\ of\ the\ Original\ Text\ without\ Stop\ Words}$$

The results of two text representations i and j are compared using the improvement gain realized by the presentation model i compared to the presentation model j defined by:

$$Gain(i,j) = 1 - \frac{DRR(i)}{DRR(j)}$$

Three experiments were conducted to evaluate the performances of the proposed approach in terms of reduction ratios and the impacts of the text size, the windows of analysis size, and the text category.

6.1 Experiment 1

The proposed new representation is implemented and tested on the dataset files. The results were compared with both the

Table 6 Dimension reduction ratio and gain

Size of the original text	DRR				Gain	
	Text without stop words	Bag-of-words (Unique words)	Stem-based reduction	Proposed approach	Proposed approach compared to stem-based approach	Proposed approach compared to bag-of-words approach
1 KB	0.67	0.95	0.77	0.63	0.18	0.34
3 KB	0.71	0.56	0.50	0.40	0.20	0.29
30 KB	0.63	0.29	0.19	0.13	0.32	0.55
50 KB	0.65	0.24	0.15	0.10	0.33	0.58
80 KB	0.60	0.20	0.11	0.08	0.27	0.60
100 KB	0.64	0.17	0.09	0.06	0.33	0.65

Table 7 Dimension reduction ratio in different windows of analysis

Size of the original text	DRR		
	Bigram	Trigram	Variable size of the window
1 KB	0.74	0.72	0.63
3 KB	0.47	0.46	0.40
30 KB	0.17	0.15	0.13
50 KB	0.13	0.11	0.10
80 KB	0.10	0.09	0.08
100 KB	0.09	0.07	0.06

Table 8 Dimension reduction ratio for different text categories

Text category	Middle east news	World news	Business	Sports	Science and technology	Arts and culture
DRR	0.10	0.12	0.13	0.10	0.08	0.09

traditional text representation based on the bag-of-words representation and the traditional VSM representation. In the bag-of-words representation, the text is represented as a vector of unique words. The traditional VSM using the stems of words is the general representation used by almost all of the published works in the Arabic language.

Table 6 shows the amount of reduction for each one of the three representations. The results show that the proposed approach has considerably reduced the dimension of text. The reduction is improved on average by a factor of 18–33 % compared with the methods that consider only stems, and by a factor of 30–65 % compared with the bag of words (unique words). The results show the impact of the size of the text on the performance of the different tested methods. The performance of all the approaches is better on larger texts; performance increases with the increase of the total number of words in the text.

6.2 Experiment 2

This experiment aims at analyzing the impact of the size of the window of analysis on the performance of the proposed text representation. Table 7 compares the reduction ratios obtained with bigram, trigram, and variable sized windows. Our best results were achieved for the variable size of window that matches the sentence containing the analyzed word.

6.3 Experiment 3

In the third experiment, we tested the system using six files of about 50 KB each. These files represent six different categories of texts: Middle East news, World news, Business, Sport, Science and Technology, and Arts and Culture. Table 8 shows how the amount of dimensionality reduction varies with the text category. It shows that the best performances are obtained with scientific texts. The worst results are found when the texts are related to business issues.

7 Conclusion

This work has presented a new method by which high-dimensional Arabic texts can be represented by a reduced-size semantic vector space. It uses semantic similarity analysis to capture similar words and regroup them in clusters. The proposed approach creates a semantic vector space model that extends the standard VSM by embedding linguistic information extracted from lexical resources such as Arabic WordNet and gazetteer of NE. The dimension of the text representation is reduced dramatically by regrouping synonyms and similar words generated from the same roots. The associated weight of each retained word or feature is computed by accumulating the weights of its synonyms and class terms. Experiments on different datasets have shown that using semantic similarity for grouping

text's features significantly reduces the size of text representation by about 27 % compared with the stem-based vector space model and by about 50 % compared with the traditional bag-of-words model. The results have shown that the amount of dimension reduction depends on the size and shape of the windows of analysis as well as on the nature of the text and its category. Future work could focus on the analysis of the impacts of the proposed approach on the performance of different NLP applications, especially text classifications, and text summarization.

References

Almaany. (2014). Dictionary and glossary. http://www.almaany.com/.

Awajan, A. (2007). Arabic text preprocessing for the natural language processing applications. *Arab Gulf Journal of Scientific Research, 25*(4), 179–189.

Awajan, A. (2015). Semantic vector space model for reducing arabic text dimensionality. In *Proceedings of the 5th international conference on digital information and communication technology and its applications, Lebanon*, (pp. 129–135). April 29–May 1, 2015.

Baker, K. (2013). Singular value decomposition tutorial. Note for NLP Seminar. 1–24. Accessed December 2013, from www.ling.ohio-state.edu/∼kbaker/pubs/Singular_Value_Decomposition_Tutorial.pdf.

Beesley, K. R. (1998). Consonant spreading in Arabic stems. In *COLING-ACL'98*, vol *1*, pp 117–123, Montreal, Quebec, Canada, August 10–14.

Biemann, C. (2006). Chinese whispers—An efficient graph clustering algorithm and its application to natural language processing problems. Workshop on TextGraphs, at HLT-NAACL 2006, pp. 73–80

Boudlal, A., Lakhouaja, A., Mazroui, A., Meziane, A., Ould Abdallahi, O. B. M., & Shoul, M. (2010). Alkhalil Morpho Sys: A morphosyntactic analysis system for Arabic texts. In *International Arab conference on information technology*. http://www.itpapers.info/acit10/Papers/f653.

Bullinaria, J. A., & Levy, J. P. (2012). Extracting semantic representations from word co-occurrence statistics: Stop-lists, stemming and SVD. *Behavior Research Methods, 44*, 890–907.

Duwairi, R., Al-Refai, M. N., & Khasawneh, N. (2009). Feature reduction techniques for Arabic text categorization. *Journal of the American Society for Information Science and Technology, 60* (11), 2347–2352.

Elkateb, S., Black, W., Rodríguez, H., Alkhalifa, M., Vossen, P., Pease, A., & Fellbaum, C. (2006). Building a WordNet for Arabic. In *Proceedings of the fifth international conference on language resources and evaluation (LREC 2006)*. Genoa, Italy, May 22–28, 2006.

Froud, H., Lachkar, A., & Ouatik, S. A. (2012). A comparative study of root-based and stem-based approaches for measuring similarity between Arabic words for Arabic text mining applications. *Advanced Computing: An International Journal (ACIJ)*, *3*(6).

Green, S., & Manning, C. D. (2010). Better Arabic parsing: Baselines, evaluations, and analysis. In *COLING*, Beijing (pp. 394–402).

Habash, N. (2010). *Introduction to Arabic natural language processing*. San Rafael: Morgan & Claypool Publishers.

Hagiwara, M. (2008). A supervised learning approach to automatic synonym identification based on distributional features. In *Proceedings of the ACL-08, Columbus*, June 2008 (pp. 1–6).

Harrag, F., El-Qawasmah, E., & Al-Salman, A. M. (2010). Comparing dimension reduction techniques for Arabic text classification using BPNN algorithm. In *IEEE first international conference on integrated intelligent computing*, pp. 6–11.

Harris, Z. (1954). Distributional structure. *Word, 10*(23), 146–162.

Hasnah, A. M., & Al-Ja'am, J. M. (2002). Thesaurus-based query disambiguation method for cross-language information retrieval. *International Journal Intelligent Computing and Information Sciences, 2*(2), 58–68.

Heintz, I. (2010). Arabic language modeling with stem-derived morphemes for automatic speech recognition. *Ph.D. thesis*, Graduate School of The Ohio State University.

Hmeidi, I., Kanaan, G., & Evens, M. (1997). Design and implementation of automatic indexing for information retrieval with arabic documents. *Journal of the American Society for Information Science, 48*(10), 867–881.

Kirchhoff, K., Vergyri, D., Duh, K., Bilmes, J., & Stolcke, A. (2006). Morphology-based language modeling for conversational Arabic speech recognition. *Computer Speech & Language, 20*(4), 589–608.

Martins, C. A., Monard, M. C., & Matsubara, E. T. (2003). Reducing the dimensionality of bag-of-words text representation used by learning algorithms. In *Proceedings of 3rd IASTED international conference on artificial intelligence and applications (AIA2003), Benalmádena, Espanha* (pp. 228–233). Calgary: Acta Press.

Mihalcea, R., & Tarau, P. (2004). TextRank: Brining order into texts. In *Proceedings of EMNLP 2004*. Association for Computational Linguistics, Barcelona, Spain (pp. 404–411).

Parkinson, D. B. (2005). *Using Arabic synonyms*. Cambridge: Cambridge University Press.

Saad, M. K., & Ashour, W. (2010). OSAC: Open Source Arabic Corpus, the 6th International Symposium on Electrical and Electronics Engineering and Computer Science, European University of Lefke, Cyprus, from http://sourceforge.net/projects/ar-text-mining/files/ArabicCorpora.

Salton, G., & McGill, M. J. (1986). *Introduction to modern information retrieval*. New York, NY: McGraw-Hill. Inc.

Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. *Communication of the ACM, 18*(11), 613–620.

Shaalan, K. (2014). A survey of Arabic named entity recognition and classification. *Computational Linguistics, 40*(2), 469–510. doi:10.1162/COLIa00178.

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

Van Rijsbergen, C. J. (1979). *Information retrieval* (2nd ed.). Cambridge: Computer Laboratory, University of Cambridge.

Xu, J., Fraser, A., & Weischedel, R. (2002). Empirical studies in strategies for Arabic retrieval. In *SIGIR'02, Proceedings of the 25th annual international ACMSIGIR conference on Research and development in information retrieval, Tampere, Finland* (pp. 269–274). August 11–15, 2002.