Journal of Engineering and Applied Sciences 13 (23): 10092-10100, 2018
ISSN: 1816-949X

Using Semantic Similarity with Word Embeddings for Arabic Multi-Words Term Extraction

El-Khadir Lamrani, El Habib Ben Lahmer and Abdelaziz Marzak
Faculty of Science Ben M'sik, University Hassan II, Casablanca, Morocco

Abstract: Identifying and extract terms from textual source is an indispensable task in information retrival and question answering systems by experiments multi-word terms represent the best candidates to represent a specific domain in Arabic. In this research, we assumed that the Multi-Word Terms (MWTs) consist of words with similar contextual representations and we propose a hybrid method of extracting multi-word terms from Arabic texts combines between linguistic and semantic approach, based on word embeddings which we use a linguistic and morphosyntactic analysis of the Arabic language to find candidate terms and we use cosine similarity between distributed representation of words for ranking candidate terms. The proposed methodology has been tested in a case studies carried out in the environnemental domains with promising results.

Key words: Multiword terms extraction, features extraction, linguistic filtering, semantic similarity, word embedding, Arabic textes

INTRODUCTION

Terminology extraction is a central field of research for a number of text mining and knowledge engineering applications such as information retrieval, ontology building, text classification, sentiment analysis, question answering systems, etc. (Zhang *et al*., 2007; Boulaknadel *et al*., 2008a, b). Information extraction is the process of extracting specific (pre-specified) information from textual sources. Terms, the linguistic representation of concept, is an important element in texts. The extraction of the terminology used to determine and extract for a given text all relevant terms for that source. The term extraction process consists of two fundamental steps, identifying term candidates (either single or multiword terms) from text and filtering through the candidates to separate terms from non-terms. Multi-word term are specific terms that are more representative of the semantic content of texts (Korkontzelos *et al*., 2008). In fact these terms are syntactical constructions less ambiguous and less polysemical than the isolated single terms (Daille, 1994; Jacquemin *et al*., 1997; Jacquemin, 1999; Zhang *et al*., 2008). For example, the complex term '[illegible]' refers to oil but the term '[illegible]' may refer to the verb to going or the metal "gold" and the term may refer to the black color or plural of lion.

A MWT is a term that is composed of more than one word. The unambiguous semantics of a multi-word term depends on the knowledge area of the concept it describes and cannot be inferred directly from its parts (Frantzi *et al*., 2000; SanJuan *et al*., 2005). In this study, we present a new method based on semantic similarity with words distributions that form a MWT, our method belongs to the family of hybrid methods for the MWTs Extraction from a corpus of text documents that we combine with linguistic and semantic approach, after a segmentation processing and a Part-of-Speech (PoS) Tagging to label texts, morpho-syntactic structures are used, they are called syntactic patterns mining to extract MWT candidates, we define a thereafter phase of language filter based on grammatical rules of Arabic language to identify terms which consist of words that have a strong linguistic connection (Bounhas and Slimani, 2009), we applied semantic similarity to disambiguate and filter our terms.

Based on our linguistic method (Lamrani *et al*., 2014) to identify MWT candidate, we propose a new method to deal with ambiguities generated by the linguistic tools and to rank candidate terms according to their semantic similarity between words. We perform this task with a view to ontology engineering and information retrieval.

Literature review: In the first part of this study, we presents the existing Multi-Word Term (MWT) extraction works in different languages, mostly in English and in the second part focus in Arabic language works.

Multi-word terms extraction for English language: Much research has addressed the issueof MWT extraction for many languages. These researches have either used a

Corresponding Author: El-Khadir Lamrani, Faculty of Science Ben M'sik, University Hassan II, Casablanca, Morocco

linguistic approach, a statistical approach or a combination of them (hybrid approach). Most recent MWT extraction methods rely on a hybrid approach to efficiently extract MWTs due to its higher accuracy compared to the two other approaches (Mahdaouy *et al*., 2014). In this regard, Frantzi *et al*. (2000) have proposed a hybrid approach to extract multi-word terms from English corpus combining linguistic and statistical. From linguistic point of view, their approach extracts the candidates of multi-word terminology based on some linguistic information such as POS tagging of the corpus which is then utilized in the linguistic filter. The linguistic filter includes all kinds of terminologies and generates beneficial outcome. The stop-list also prevents the extraction of candidates which are less likely to be terminology and enhances the precision of the output list. Furthermore, the C-value is utilized to ensure that, the extracted outcome is basically a multiword terminology. The C-value measure has been utilized for resolving the problem of nested terms. Generally, the terms used chemistry documents, automotive and biomedical articles, follow by a specific pattern of combined nouns and adjectives. In the syntactic point of view, generally they are compound nouns and associated constructions. For example, the terms found in biomedical abstracts of the Genia corpus are names of diseases and drugs, chemical elements, anatomy and other names. Consequently, methods for automatic identification of compound nouns when used on domain-specific data, might be effective in extracting multiword terminologies aimed at comparing MWTs interpretation and classification. SanJuan *et al*. (2005) have merged three linguistic resources and techniques such as: statistical associations, WordNet, and clustering, to construct a hierarchical model, as against manual annotation of terms of the corpus. They have compared their model with traditional clustering algorithms and with the manually created ontology of the corpus. Chen *et al*. (2006) have proposed a novel automatic statistical method for determining multiword terms depending on co-related text-segments present in a range of documents. The proposed method applied a novel and efficient statistical strategy for determining multi-word terms. Chen *et al*. (2006) have presented the multi-word term extraction system, a new automatic statistical approach to identify multiword terms based on co-related textsegments existing in a set of documents. The suggestion was used a new and effective statistical method for identifying multi-word terms. The above system consists of four components: text segment generator which utilizes a small predefined stop-list as an preliminary input, to classify a set of text documents into text-segments, text segment-weigher which computes the segment weight for each created text-segments, text segment-segmenter which segments all the textsegments depending on their segmentweights to create new text-segments-term candidates, the term candidates shall be re-input for additional segmentation or directly input to the subsequent component, term identifier which recognizes the resulting term candidates to become terms, according to a specific threshold. Other extraction method of complex terms, work conducted by Church and Hanks (1990) which is based on a statistical measure: mutual information, in this method the researchers considered that the words that often appear together in a statistically significant manner have chance to form complex terms and then compare the probability of occurrence of words together with the probability of occurrence of these words separately. Daille (1994) proposed the system ACABIT (Automatic Corpus Based Acquisition of Binary Terms), This system is dedicated to the French and English languages. Boulaknadel *et al*. (2008a, b) proposed a tool that is based on ACABIT and deals Arabic language, this tool consists of two steps: the first concerns linguistic identification of words using simple rules applied to tagged corpus, they used at this stage the system proposed by Diab (2007) to retrieve a list of candidate terms labeling, the list of these terms is contained in a second stage to several statistical measures. These measures are used to calculate the terminological status of the sequence encountered. Bounhas and Slimani (2009) followed the same approach that focuses on deep linguistic study based on syntactic models of complex names and specific linguistic rules for Arabic language, to propose a method for extracting knowledge from corpus labeled using again the tool (Diab, 2007) they evaluated their method using the same corpus used by Boulaknadel *et al*. (2008a, b) and they got results that exceed those obtained by them in terms of accuracy.

Arabic multi-word terms extraction: One of the Arabic MWT extraction works is the work which has been presented by Bounhas and Slimani (2009) who have proposed a hybrid method to extract multiword terminology from Arabic corpora. They have applied several tools to extract and identify the compound nouns. Their approach used the Arabic morphological analyser (AraMorph), proposed by Hajic *et al*. (2005). The AraMorph has been applied to compute the morphological features, required for the syntactic rules. Al Khatib and Badarneh (2010) has proposed a hybrid approach to extract multi-word terms from Arabic corpus. They concentrated on compound nouns as in important type of MWT and select bi-gram term. The approach relies on two main filters: linguistic filter where simple part (POS) tagger is use to extract candidate MWTs. This step contains, prepositions words classification, extraction of nouns sequence and nouns sequence that associated by prepositions, testing each extracted sequence based on MWTs syntactic patterns. Statistical method where log-likelihood ratio and C-value are used to rank bi-gram candidate MWTs.

MATERIALS AND METHODS

We propose an extraction method of multi-word terms from Arabic texts based on our previous research by Lamrani *et al*. (2014). Our method consists of four main steps will be explained in detail (Fig. 1). To ensure extraction of relevant information must first go through a deep linguistic analysis starts with segmentation and cutting text into words and labeling to assign to each word a labeled them which designates its kind and the group of words to which it belongs and the standardization of various non-standard words taking into account the different variations that a word can take. The second phase involves extracting complex terms according to well-defined patterns of extraction and the third phase will filter the results under the previous phase based on linguistic rules, we will define to extract candidate terms most relevant and finally in the last stage, we will use a statistical and semantic filter to extract key terms.

Pre-processing and Part-of-Speech (PoS) tagging: This phase includes two main steps:

Pre-processing: Pre-processing design tokenization and stop word elimination which tokenization aims to cut the text to lexical units. It's a hard task. In fact disambiguate borders of phrases and terms in Arabic is not a trivial thing as several ambiguities of different types can appear, the proportion of ambiguous words represent more than 90% in a non-vowelized Arabic text (Belguith *et al*., 2005), for example, the word 'وهم' may refer to a name illusion or a conjunction 'و' and' in English+pronoun 'هم': 'they' in English. And stop words elimination aims to remove words having no significant semantic relation to the context in which they exist.

Stop words are the terms that occur frequently in most of the documents in a given collection. Also, we create our own Arabic stop word list contains some of grammatical links such as the definite study (AL) (the), attached and separate prepositions, conjunctions, interrogative words, negative words, exclamations and calling letters, adverbs of time and place, also, they include all the pro-nouns, demonstratives, subject and object pronouns, the five distinctive nouns, some numbers, additions and verbs.

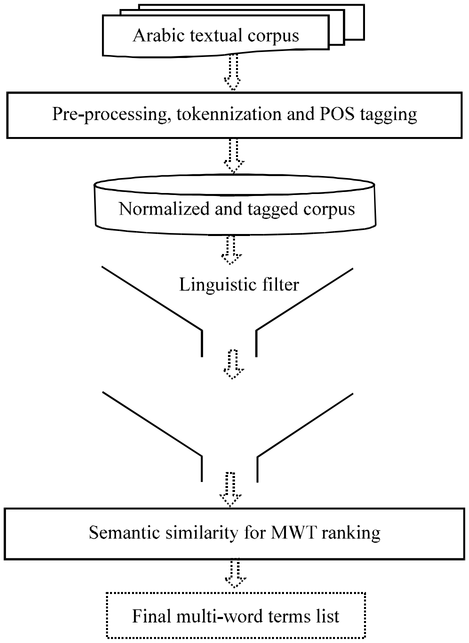

Arabic textual corpus
Pre-processing, tokennization and POS tagging
Normalized and tagged corpus
Linguistic filter
Semantic similarity for MWT ranking
Final multi-word terms list

Fig. 1: Architecutre of proposed method of MWT extraction

POS tagging and stemming: The POS tagging is used to assign each word a labeled them which designates its kind and the group of words to which it belongs plays an indispensable role in the extraction of information from texts, an Arabic word may take several grammatical labels, for example, the word ('كتب') has 5 distinct labels may designate a verb in the third person singular of achievement active (kataba: write), masculine plural noun (kutubun: books) singular masculine noun (katbun: written) verb of the third person masculine singular of passive accomplished (Kutiba: it was written or kuttiba correspondent factitive form) or imperative verb in the second person masculine singular (kattib: do write) (Fathi Debili labeling grammatical). Several studies have addressed the problem of labeling grammatical Arabic, our work is based on the research of Diab (2007) proposed a system for tokenization and grammatical labeling that achieves resolution thresholds exceed 96% labeling that achieves resolution thresholds exceed 96% for unvowelized Arabic texts. The underlying system uses Support Vector Machine (SVM).

Extraction of MWT candidates: This phase comprises two steps morph syntactical extraction structures based on linguistic rules, called extraction patterns and standardization to address the diversity of variants of the same structure. In this research, the most important structure that is treated the complex term, a complex term is a phrasme which consists of more than one word in our research is based on the research by Boulaknadel *et al*. (2008a,b) to define morph syntactical structures on which the terms appear by adding new structures for extracting and ternary terms and 4 naire as: "مرض فقدان الناعة الكتسبة" this term of 4 words noted by Term.. Belongs to the structure identified by Name. Term., a Term. may appear in the following structures: Name Term..: تقلص الجال الغابوي Name Prep Term..: أمراض في الأجهزة التنفسية

Thus, a Term. may appear in the following structures, the noun and the adjective may be definite or indefinite, in our research, we consider the determinant ‚ part of the word:

- Name Adj: التفاعل الكيميائي، تفاعل كيميائي
- Name. Name. prep:
- Name. Name.:

The second step is the normalization that solves the diversity of variations under which the same term may appear, e.g., morph syntactic structures "تلوث ألياه", "تلوث الاء", "التلوث الائي" and "تلوث للماء" are four variations of the same term. The result of this step is a list of all the complex terms that appear in the corpus of documents.

Linguistical filter: The language filter can submit multi word candidate terms resulting from the previous phase to a set of linguistic rules to determine enjoying the semantic richness of the Arabic language to measure the dependency between the words composing a complex term in Arabic a complex Term.. may be in the form Name. Term.., where Term is a noun phrase.

The language filter can submit complex candidate terms resulting from the previous phase to a set of linguistic rules to determine enjoying the semantic richness of the Arabic language to measure the dependency between the words composing a complex term in Arabic a complex Term.. may be in the form Name. Term.., where Term is a noun phrase (Fig. 2). Measuring dependence between the words composing a complex term noun phrase can be found in several grammatical categories: Name and adjective (النعت مالنعوت) Name and its annex (مضاف ومضاف إليه) and other (بدل، حال، إسناد، توكيد).

We cann't talk about the dependence between words with a complex term than in the reference case (إسناد) or Annex nonverbal by cons in other grammatical categories there are no dependencies between words even if they are found linked several times. We propose in this study an algorithm that allows us to recognize the multi-word terms (complex term) relevant. Reference (إسناد) is always identified by the existence of the reference letter (ي) but there are cases where the letter (ي) is the original word for Annex, we can recognize it by analyzing the canonical form of the first word that composes the complex term to see if it is a derivative or a proper name and the pattern of the word is extracted then tested if it satisfies the canonical forms data, the name n can be an adjective, a proper noun or a named entity.

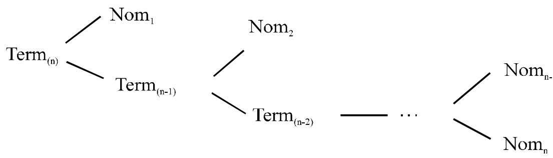


Fig. 2: Structure of multi-word terms in Arabic

Linguistic filtering algorithm: Given a list of n complex words L = WT, WT., ..., WT, ...; Wt (Fig. 3). For each i belongs to {1, ..., n} WT. = WW'. The function filter () eliminates the word W. and replace WW. by 'W.' if W. is a proper name, the function extract () adds WW. to the list of complex candidate terms, the extraction schemes is as follows: the root word is determined based on the algorithm khoja [Shereen Khoja] then, we subtracted this racine of full word and replace the voids, respectively by (فعل) if the root is a triple consonant (فعلل) if it is a root of four consonant, e.g., word (msalm مسلم), the root (slm سلم) subtraction, we gives (م...), we replace the '.', respectively (فعل), the scheme is obtained (مفعل).

Statistical filter: The statistical filter classifies the key term candidates in order of importance, based on the linguistic filter returns a list containing the complex terms. Statistical measures are applied to classify the list of complex terms extracted by the linguistic filtering.

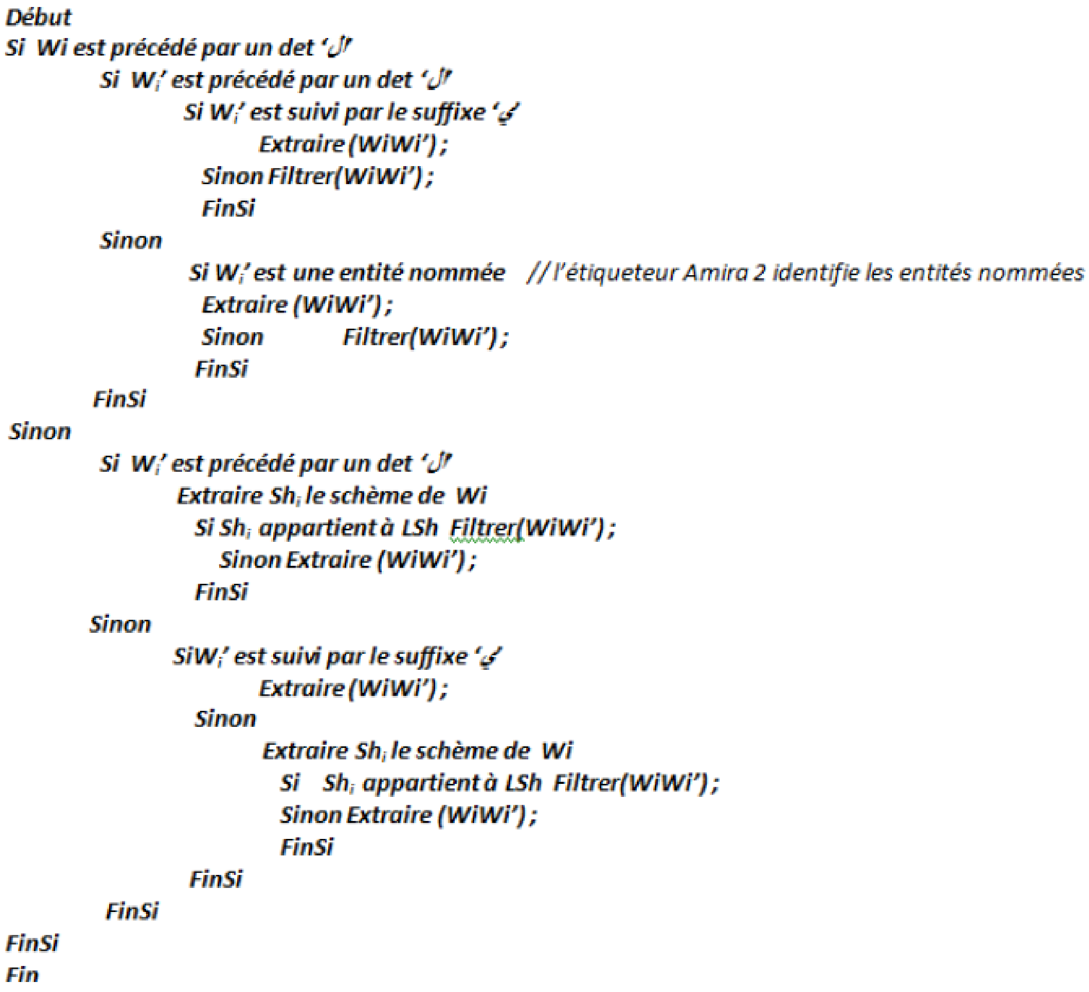
The C-value: The C-value measures the termhood of a candidate string on the basis of several characteristics: number of occurrences, term nesting, and term length. It is defined as:

$$\mathrm{C-Value}(a) = \begin{cases} \log(|a|).f(a) \text{ if } a \text{ is not nested} \\ \log(|a|).f(a) - g(a) \text{ otherwise} \end{cases} \quad (1)$$

Where:
|a| = The length in words of candidate term a
f(a) = The number of occurrences of a

$$g(a) = \frac{1}{|T|(a)} \sum_{b \in T} f(b) \quad (2)$$

Where:
T(a) = The set of longer candidate terms into which a appears
(|T(a)| = The cardinality of this set



```
Début
Si  Wi est précédé par un det ‘ال’
            Si  $W_i$’ est précédé par un det ‘ال’
                      Si $W_i$’ est suivi par le suffixe ‘يّ’
                                Extraire (WiWi’) ;
                        Sinon Filtrer(WiWi’) ;
                        FinSi
            Sinon
                       Si $W_i$’ est une entité nommée    // l’étiqueteur Amira 2 identifie les entités nommées
                        Extraire (WiWi’) ;
                        Sinon          Filtrer(WiWi’) ;
                       FinSi
           FinSi
Sinon
            Si  $W_i$’ est précédé par un det ‘ال’
                      Extraire $Sh_i$ le schème de  Wi
                        Si $Sh_i$ appartient à LSh  Filtrer(WiWi’) ;
                          Sinon Extraire (WiWi’) ;
                        FinSi
          Sinon
                     Si$W_i$’ est suivi par le suffixe ‘يّ’
                                Extraire (WiWi’) ;
                        Sinon
                                Extraire $Sh_i$ le schème de  Wi
                                  Si    $Sh_i$ appartient à LSh  Filtrer(WiWi’) ;
                                  Sinon Extraire (WiWi’) ;
                                  FinSi
                       FinSi
            FinSi
FinSi
Fin
```

Fig. 3: Arabi linguistic filtering pseudocode

As one can note, if the candidate term is not nested, its score is solely based on its number of occurrences and length. If it is nested, then its number of occurrences is corrected by the number of occurrences of the terms into which it appears.

The NC-value: The NC-value combines the contextual information of a term together with the C-value. The contextual information is calculated based on the N-value which provides a measure of the terminological status of the context of a given candidate term. It is defined as:

$$\text{N}-\text{value(a)} = \sum_{a \in C_a} f_a(b).\frac{|T(b)|}{n} \qquad (3)$$

Where:

C. = The set of distinct context words of a
f(b) = The number of times b occurs in the context of a
n = The total number of terms considered

This measure is then simply combined with the C-value value to provide the overall NC-value measure:

$$\text{NC-value(a)} = 0.8\times\text{C-value(a)}+0.2\times\text{N-value(a)} \qquad (4)$$

We use statistical measurement NC-value which exceeds other measures in terms of accuracy for bi-grams and trigrams (Mahdaouy *et al*., 2014).

Semantic similarity between words in terms candidates: Most of the commonly used methods represent words in a corpus using values, thus ignore the context a word is used in. In 2013, three papers on the topic of distributed word embeddings to catch semantic similarities between words published by Mikolov *et al*. (2013a-c) which resulted in Googles Word2Vec Software. It’s a novel way to create vector representations of words in a way that preserves their meaning, i.e., words vectors with similar meaning tend to be located in similar positions when represented in the vector space. We can illustrate the model as a two-layer neural net that processes text. We'll start describing the word embedding then the Skip-Gram Model that Word2Vec uses for learning.

Word embedding: Word embedding is the collective name for a set of language modelling and feature learning techniques in NLP where words or phrases from the vocabulary are mapped to vectors of real numbers in a low-dimensional space relative to the vocabulary size. So, the basic idea is to store the same contextual information in a low-dimensional vector, each word is now represented by a D-Dimensional vector where D is a relatively small number (typically between 50 and 1000).

This is what these algorithms basically do: they start from a random vector for each word in the vocabulary. Then they go over a large corpus and at every step, observe a target word and its context (neighbours within a window). The target word's vector and the context words vectors are then updated to bring them close together in the vector space (and therefore increase their similarity score). Other vectors (all of them or a sample of them) are updated to become less close to the target word. After observing many such windows, the vectors become meaningful, yielding similar vectors to similar words. Using this words representation we can express similarity of words and even analogies between them:

Most similar words: we can easily find the most similar words to a certain word by finding the most similar vectors.

Analogies: word embeddings exhibit some semantic and syntactic patterns. For example, we can observe relationships or analogies such as:

$$u_{\overrightarrow{king}} - u_{\overrightarrow{man}} \approx u_{qeen} - u_{\overrightarrow{woman}}$$

This definitely helps to solve analogy questions: a is to b as c is to d given a-c, find the missing word d. To conclude this part, we have to remember that word embedding affect for each word in our vocabulary a random vector. Then while processing the training data, the words vectors are optimized by making vectors of similar words (words in the same context or neighbours) more closer. Word2Vec uses a specific learning model to optimize the vectors, it's the Skip-Gram Model which we explain in the next part.

Word2Vec; CBOW and Skip-Fram Models: Skip-Gram or CBOW (Continuous Bag of Words) are architectures that describe how the neural network learns the vector representation for each word. For CBOW, we try to predict the word given its context while skip-gram try to predict the context given a word. In simpler words, CBOW tends to find the probability of a word occurring in a neighbourhood (context). So, it generalises over all the different contexts in which a word can be used. Whereas skip-gram tends to learn the different contexts separately. So, Skip-Gram needs enough data with reference to each context. Hence Skip-Gram requires more data to train, also, skip-gram (given enough data) contains more knowledge about the context. We use the Skip-Gram Model because it produces more accurate results on large datasets. The training objective of the Skip-Gram Model is to find word representations that are useful for predicting the surrounding words in a sentence or a document. More formally, given a sequence of training words w., w., ..., w., the objective of the skip-gram model is to maximize the average log probability:

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{j\in[-c,\ c],\ j\neq 0}\log(p(w_{t+j} \mid w_t))$$

where c is the size of the training context. The projection learned from skip-gram model could be used to compute the vector representation of a single word. The basic Skip-gram formulation defines p(w.|w.) using the softmax function:

$$p(w_O \mid w_I) = \frac{\exp(u'_{wO}\ .u_{wI})}{\sum_{w=1}^{W}\exp(u'_{wO}\ .u_{wI})} \qquad (5)$$

Where:

u. and u'. = The input and output vector representations of w

W = The number of words in the vocabulary

This formulation is impractical because the cost of computing .log p(w.|w.) is proportional to W which is often large (10-10 terms). The training objective is to learn word vector representations that are good at predicting the nearby words (Fig. 4).

Semantic similarity with word embeddings:

$$\cos(\theta) = \frac{W_1.W_2}{\|W_1\|.\|W_2\|} = \frac{\sum_{i=1}^{n} W_1.W_2}{\sqrt{\sum_{i=1}^{n} W_{1i}^2}\sqrt{\sum_{i=1}^{n} W_{2i}^2}} \qquad (6)$$

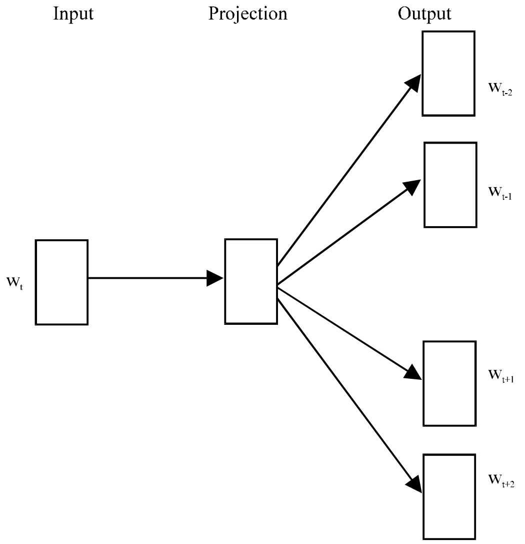


Fig. 4: The Skip-Gram Model architecture

RESULTS AND DISCUSSION

We have used the Skip-Gram word representations for Arabic model using a large data collected by Zahran *et al*. (2015) (//sites.google.com/site/mohazahran/data) from different sources counting more than 5.8 billion words.

Table 1: The skip-gram model parameters

Parameters	Values
Number of features	300
Window size	5
Hierarchical softmax	Yes

Table 2: Precision of MWT methods

Methods	Precision (%)
Boulaknadel method	85.0
Linguistic and NC-value	91.0
Linguistic and semantic	92.5
Linguistic and semantic and NC-value	92.0

Training the Arabic Skip-Gram Model require choice of some parameters affecting the resulting vectors. All the parameters are shown in Table 1.

Where:

- Number of features: dimensionality of the word vectors
- Window size the limit on the number of words in each context

We used Skip-Gram Model in order to identify the near matches between two words W_i and W_j (e.g., W_i and W_j are obtained by comparing their vector representations V_i and V_j, respectively. We use the cosine similarity to evaluate the similarity between V_i and V_j, e.g:

$$Sim(\text{التلوث البيئي}) = \cos\left(V_{skip-gram}(\text{'التلوث'}), V_{skip-gram}(\text{'البيئي'}\right)$$

$$Sim(\text{'فقدان' ، 'الناعة' ، 'الكتسبة'}) = \frac{1}{3}\Big[\cos\left(V(\text{'فقدان'}), V(\text{'الناعة'})\right) + \cos\left(V(\text{'فقدان'}), V(\text{'الكتسبة'})\right) + \cos\left(V(\text{'الكتسبة'}), V(\text{'الناعة'})\right)\Big]$$

Testing collection and evaluation: Since, there is no standard domain-specific Arabic corpus, we have built in order to evaluate our approach, a new corpus specialized on the environmental domain with similar properties as the ones described. The corpus built contain 1666 files comprising 53569 different tokens (without stop words) extracted from the web site Al-Khat Alakhdar 1. It covers various environmental topics such as pollution, noise effects, water purification, soil degradation, forest preservation, climate change and natural disasters.

The evaluation of automatic MWTs extraction is a complex process and is usually performed by comparing each MWT candidate extracted to a domain-specific reference list. When there is no reference list available in the language retained, one can first translate the MWT candidates (using a machine translation system or a bilingual dictionary) and use a reference list available in another language. For the evaluation purpose, we have constituted automatically a reference list of all Arabic MWTs available in the latest version of AGROVOC2 thesaurus and then use the stemming algorithm to remove prefixes and suffixes for each MWT in the reference list and the extracted MWT list (Table 2).

Our results using our lingisitic filtering method and semantic similarity between words embedding are better in terms of precision then those achieved by Boulaknadel *et al.* (2008a, b) and Lamrani *et al.* (2014).

CONCLUSION

We describe in this study an extraction method of MWTs from a corpus of Arabic texts based firstly on the use of syntactic extraction patterns, after a deep linguistic analysis and morph syntactic and secondly on a linguistic filter enjoying grammatical richness of the Arabic language and a statistical/semantic measure to evaluate the link between the words composing the complex term. To make our method more complete, we will use algorithms for learning to rank candidates MWTs for improving accuracy.

REFERENCES

Al Khatib, K. and A. Badarneh, 2010. Automatic extraction of Arabic multi-word terms. Proceedings of the 2010 International Multiconference on Computer Science and Information Technology (IMCSIT), October 18-20, 2010, IEEE, Wisla, Poland, ISBN:978-83-60810-27-9, pp: 411-418.

Belguith, L., L. Baccour and G. Mourad, 2005. [Segmentation of Arabic texts based on the contextual analysis of punctuation and certain particles]. Proceedings of the 12th Annual International Conference on Automatic Processing of Natural Languages, June 6-10, 2005, TALN, Dourdan, France, pp: 451-456 (In French).

Boulaknadel, S., B. Daille and A. Driss, 2008a. [Acabit: A tool for extracting complex terms]. Proceedings of the 2008 Symposium on Act of the New Information Technologies: Opportunities for Lamazighe, November 24-25, 2008, IRCAM, Paris, France, pp: 75-82.

Boulaknadel, S., B. Daille and D. Aboutajdine, 2008b. Multi-word term indexing for Arabic document retrieval. Proceedings of the IEEE International Symposium on Computers and Communications (ISCC'08), July 6-9, 2008, IEEE, Marrakech, Morocco, ISBN:978-1-4244-2702-4, pp: 869-873.

Bounhas, I. and Y. Slimani, 2009. A hybrid approach for Arabic multi-word term extraction. Proceedings of the 2009 International Conference on Natural Language Processing and Knowledge Engineering NLP-KE, September 24-27, 2009, IEEE, Dalian, China, ISBN:978-1-4244-4538-7, pp: 1-8.

Chen, J., C.H. Yeh and R. Chau, 2006. Identifying multi-word terms by text-segments. Proceedings of the 7th International Conference on Web-Age Information Management Workshops, June 17-19, 2006, IEEE, Hong Kong, China, pp: 19-19.

Church, K. and P. Hanks, 1990. Word association norms, mutual information and lexicography. Comput. Linguist., 16: 22-29.

Daille, B., 1994. [Mixed Approach for Terminology Extraction: Lexical Statistics and Linguistic Filters Voorkant]. Paris Diderot University, Paris, France, Pages: 228 (In French).

Diab, M.T., 2007. Improved Arabic base phrase chunking with a new enriched POS tag set. Proceedings of the 2007 International Workshop on Computational Approaches to Semitic Languages: Common Issues and Resources, June 28, 2007, Association for Computational Linguistics, Stroudsburg, Pennsylvania, USA., pp: 89-96.

Frantzi, K., S. Ananiadou and H. Mima, 2000. Automatic recognition of multi-word terms: The C-value/NC-value method. Int. J. Digit. Lib., 3: 115-130.

Hajic, J., O. Smrz, T. Buckwalter and H. Jin, 2005. Feature-based tagger of approximations of functional Arabic morphology. Proceedings of the 4th International Workshop on Treebanks and Linguistic Theories (TLT), December 10, 2005, Universitat de Barcelona, Barcelona, Spain, pp: 1-54.

Jacquemin, C., 1999. Syntagmatic and paradigmatic representations of term variation. Proceedings of the 37th Annual International Meeting on Association for Computational Linguistics on Computational Linguistics, June 20-26, 1999, Association for Computational Linguistics, Stroudsburg, Pennsylvania, USA., ISBN:1-55860-609-3, pp: 341-348.

Jacquemin, C., J.L. Klavans and E. Tzoukermann, 1997. Expansion of multi-word terms for indexing and retrieval using morphology and syntax. Proceedings of the 35th Annual Meeting and 8th International Conference on the Association for Computational Linguistics and European Chapter of the Association for Computational Linguistics, July 7-12, 1997, Association for Computational Linguistics, Stroudsburg, Pennsylvania, USA., pp: 24-31.

Korkontzelos, I., I.P. Klapaftis and S. Manandhar, 2008. Reviewing and evaluating automatic term recognition techniques. Proceedings of the 6th International Conference on Advances in Natural Language Processing, August 25-27, 2008, Springer, Gothenburg, Sweden, ISBN:978-3-540-85286-5, pp: 248-259.

Lamrani, E.K., A. Marzak and H. Ballaoui, 2014. Mixed method for extraction of domain terminology from text: Linguistic and statistical filtering. Proceedings of the IEEE 2014 3rd International Conference on Colloquium in Information Science and Technology (CIST), October 20-22, 2014, IEEE, Tetouan, Morocco, ISBN:978-1-4799-5979-2, pp: 291-295.

Mahdaouy, A.E., S.E. Ouatik and E. Gaussier, 2014. A study of association measures and their combination for Arabic MWT extraction. Proceedings of the 10th International Conference on Terminology and Artificial Intelligence, September 10, 2014, Cornell University, Ithaca, New York, USA., pp: 1-8.

Mikolov, T., I. Sutskever, K. Chen, G.S. Corrado and J. Dean, 2013a. Distributed Representations of Words and Phrases and their Compositionality. In: Advances in Neural Information Processing Systems, Burges, C.J.C., L. Bottou, M. Welling, Z. Ghahramani and K.Q. Weinberger (Eds.). Curran Associates Inc., New York, USA., pp: 3111-3119.

Mikolov, T., K. Chen, G. Corrado and J. Dean, 2013b. Efficient estimation of word representations in vector space. J. English Lit., 1: 1-12.

Mikolov, T., W.T. Yih and G. Zweig, 2013. Linguistic regularities in continuous space word representations. Proc. NAACLHLT., 13: 746-751.

SanJuan, E., J. Dowdall, F. Ibekwe-SanJuan and F. Rinaldi, 2005. A symbolic approach to automatic multiword term structuring. Comput. Speech Lang., 19: 524-542.

Zahran, M.A., A. Magooda, A.Y. Mahgoub, H. Raafat and M. Rashwan *et al*., 2015. Word representations in vector space and their applications for Arabic. Proceedings of the 16th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing'15), April 14-20, 2015, Springer, Cairo, Egypt, ISBN:978-3-319-18110-3, pp: 430-443.

Zhang, W., T. Yoshida and X. Tang, 2007. Text classification using multi-word features. Proceedings of the IEEE International Conference on Systems, Man and Cybernetics (ISIC'07), October 7-10, 2007, IEEE, Montreal, Quebec, Canada, ISBN:978-1-4244-0990-7, pp: 3519-3524.

Zhang, W., T. Yoshida and X. Tang, 2008. A study on multi-word extraction from Chinese documents. Proceedings of the International Conference on Asia-Pacific Web, April 26-28, 2008, Springer, Berlin, Germany, ISBN978-3-540-89375-2, pp: 42-53.