ISSN 1870-4069

A Framework for the Construction of a Historical Dictionary for Arabic

Rim Laatar, Chafik Aloulou, Lamia Hadrich Belguith

MIRACL Laboratory, ANLP Research Group,
Faculty of Economics and Management, University of Sfax, Sfax, Tunisia
rimlaatar@yahoo.fr
{chafik.aloulou,l.belguith}@fsegs.rnu.tn

Abstract. Arabic is one of the oldest Semitic languages in the world. But despite its rich historical heritage, Arabic is still bereft of a historical dictionary which traces the first use of its words and the evolution of their meanings and structures. Therefore, creating such a dictionary is of a great importance for the Arab world as it bridges the gap between its present and its past. This task should undergo several stages and requires a lot of effort. In this paper, we present our framework to help the linguists create a historical dictionary for Arabic. For this aim, we propose a platform which helps to trace the evolution of the meanings of a given word throughout time. The developed system allows the user to extract the meaning of an Arabic word according to the historical period in which it appeared. It also provides information about the oldest date of use of the word with a textual example in which it first appeared, and the first place where the word was used.

Keywords: Arabic historical dictionary, word sense disambiguation, word embedding, old Arabic, classical Arabic, modern standard Arabic.

1 Introduction

The human language is subject to a large number of factors and influences. The latter contribute to its development and to the evolution of its vocabulary and constructions. It may also lead to its erosion and fragmentation, or to its total extinction. It develops and renews itself when it finds the conditions that guarantee its development and may fade away when it is neglected by people (lack of use, forgetfulness, etc.).

Thus, the historical events and the political conditions that humanity has experienced have had a decisive impact on the subdivision of the languages of the ancient worls. In fact, each language can be divided into species and groups themselves, each giving rise to several languages linked together by historical and geographical bonds. In order to safeguard their languages, nations have resorted to the rooting of the language and the establishment of its history by means of historical dictionaries. According to Al-Said [2], a historical dictionary is a general dictionary of language which draws its importance from the human heritage gathered from sciences, arts and letters from the different ages and places. It studies the evolution of the construction of words and their meanings through the chronological stages the language has undergone. The historical dictionaries of a language are thus considered to be as the language body which helps to understand the entire human heritage.

However, despite its richness, the Arabic language does not yet have a historical dictionary which helps to monitor the semantic development of the Arabic language throughout history, and to understand its knowledge and scientific heritage correctly. Indeed, words in Arabic have gone through a historical process of growth marked by significance and expectation. Accordingly, certain words have changed in terms of vocations over time and others have completely disappeared from literature.

As a matter of fact, the evolution of the Arabic language from antiquity to the present day has given birth to several linguistic registers. According to Al-Said [3], the Arabic language can be divided into three periods: *(1)* Old Arabic, which is not used currently. It is found in ancient literary works (mainly poems).*(2)* Classical Arabic or literary Arabic, which is the language of the Quran. It is spread through Islamic conquests.*(3)* Modern Standard Arabic (MSA) is the official language of all the Arab countries [16].

In view of that, language evolves and changes by time in terms of its sounds, rules, and especially its meanings. Accordingly, the meanings of words is not fixed, but in a constant change and evolution from one age to another.

In this work, we propose to develop a framework for the construction of a historical dictionary for Arabic. In fact, creating such a dictionary has to go through several stages and requires a lot of effort. One of these stages is the extraction of the appropriate sense of a given word according to its appearance in the document. Recently, several studies have focused on disambiguating words in Modern Standard Arabic, but there seems to be no work concerned with disambiguating Arabic words according to their historical period in which they appeared. The principal objective is to disambiguate words appearing in Old and Classical Arabic in order to study the semantic evolution of each word of the language through its historical ages. Therefore, the main contributions of this paper are as follows:

- Propose a method which helps to extract automatically the meaning of a given word according to its historical period. This method aims not only to identify what a word means in a given context, but also to disambiguate it according to the historical period in which it appeared.
- Suggest a method that helps to give clear and precise information about the oldest use of a given word, its first date of appearance, its users, its first places of appearance, and the date of its sense evolution.
- Help the linguists build a historical dictionary for Arabic by proposing a tool which enables to automatically extract the meaning of a given word and describe its evolution historically and geographically.

The rest of this paper is divided as follows. Section 2 reviews some works on existing historical dictionaries in different languages as well as some previous attempts to build a historical dictionary for Arabic. It also gives an overview of the works which focus on Arabic word sense disambiguation. Section 3 explains our method to help the linguists to construct such a dictionary. Section 4 presents the developed system. Experimental results are tackled in section 5. In section 6 we draw a conclusion.

2 State of the Art

2.1 Historical Dictionaries Background

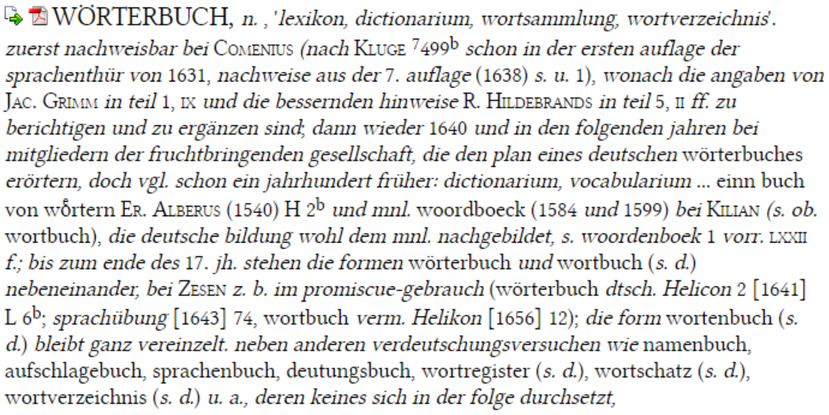
The idea of creating historical dictionaries appeared during the second half of the 19^{th} century following the appearance of the method of historical analysis [2]. The primordial objective of creating historical dictionaries was to gather information about the words of the language by studying their evolution over time in terms of phonetics, structure, form and meaning. Several international projects have been launched in different countries whose purpose was to develop a historical dictionary. The first attempt was with the German Historical dictionary in 1838. Then some other endeavors were with the Dutch Historical dictionary in 1849 and with the English Dictionary in 1849.

The German Historical Dictionary. The German historical dictionary *Deutsches WörterBuch* (DWB) is the most important German historical dictionary since the 16^{th} century. It is also called the Grimm dictionary, referring to the names of its creators the Grimm brothers (Jacob and Wilhelm Grimm), who began working on it in 1838 with more than 80 collaborators. It is a historical dictionary that traces the history of each word using many quotes. Indeed, the purpose of this dictionary is to analyze and explain exhaustively the origin and use of each German word [2]. Figure 1 shows an excerpt[1] from the dictionary for the word ' WÖRTERBUCH/Dictionary'.

According to this excerpt, we note that the search for a given word in the German electronic history dictionary makes it possible to present the various synonyms of the word in focus as well as its inflected forms. The articles which contain this word are classified by their date of appearance, their links, and their search links in other dictionaries. This work was manually and did not depend on NLP tools.

The Dutch Historical Dictionary WNT. The dictionary of the Dutch language (WNT) "Woordenboek der Nederlandsche Taal" was announced in 1849. The WNT contains about 95 000 main entries and about 1.7 million citations. This dictionary indicates for each word the grammatical characteristics, the origin, the meaning(s), their use in compounds, sentences and proverbs, and derivations [2].

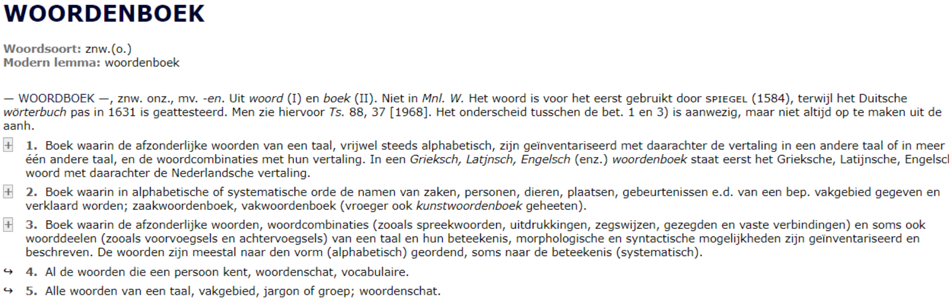
[1] `http://woerterbuchnetz.de/cgi-bin/WBNetz/wbgui_py?sigle=DWB&mode=Vernetzung&lemid`



WÖRTERBUCH, *n.* , *'lexikon, dictionarium, wortsammlung, wortverzeichnis'*. *zuerst nachweisbar bei* COMENIUS *(nach* KLUGE ^{7}499^b *schon in der ersten auflage der sprachenthür von* 1631, *nachweise aus der* 7. *auflage* (1638) *s. u.* 1), *wonach die angaben von* JAC. GRIMM *in teil* 1, IX *und die bessernden hinweise* R. HILDEBRANDS *in teil* 5, II *ff. zu berichtigen und zu ergänzen sind; dann wieder* 1640 *und in den folgenden jahren bei mitgliedern der fruchtbringenden gesellschaft, die den plan eines deutschen* wörterbuches *erörtern, doch vgl. schon ein jahrhundert früher: dictionarium, vocabularium* ... einn buch von wo̊rtern ER. ALBERUS (1540) H 2^b *und mnl.* woordboeck (1584 *und* 1599) *bei* KILIAN *(s. ob.* wortbuch), *die deutsche bildung wohl dem mnl. nachgebildet, s. woordenboek* 1 *vorr.* LXXII *f.; bis zum ende des* 17. *jh. stehen die formen* wörterbuch *und* wortbuch (*s. d.*) *nebeneinander, bei* ZESEN *z. b. im promiscue-gebrauch* (wörterbuch *dtsch. Helicon* 2 [1641] L 6^b; *sprachübung* [1643] 74, wortbuch *verm. Helikon* [1656] 12); *die form* wortenbuch (*s. d.*) *bleibt ganz vereinzelt. neben anderen verdeutschungsversuchen wie* namenbuch, aufschlagebuch, sprachenbuch, deutungsbuch, wortregister (*s. d.*), wortschatz (*s. d.*), wortverzeichnis (*s. d.*) *u. a., deren keines sich in der folge durchsetzt,*

Fig. 1. Excerpt from the DWB dictionary.

Figure 2 shows an excerpt[2] from the dictionary for the word ' WOORDENBOEK /Dictionary'. This figure indicates that the WNT dictionary provides for each word the part of speech, the lemma, the meaning, the date of appearance, the inflected forms, the quotations and the origin of this word. Once the word is composed, the dictionary displays the previous information for each word that composes it. This work was manually and did not depend on NLP tools.



WOORDENBOEK

Woordsoort: znw.(o.)
Modern lemma: woordenboek

— WOORDBOEK —, znw. onz., mv. *-en*. Uit *woord* (I) en *boek* (II). Niet in *Mnl. W.* Het woord is voor het eerst gebruikt door SPIEGEL (1584), terwijl het Duitsche *wörterbuch* pas in 1631 is geattesteerd. Men zie hiervoor *Ts.* 88, 37 [1968]. Het onderscheid tusschen de bet. 1 en 3) is aanwezig, maar niet altijd op te maken uit de aanh.

⊞ **1.** Boek waarin de afzonderlijke woorden van een taal, vrijwel steeds alphabetisch, zijn geïnventariseerd met daarachter de vertaling in een andere taal of in meer één andere taal, en de woordcombinaties met hun vertaling. In een *Grieksch, Latjnsch, Engelsch* (enz.) *woordenboek* staat eerst het Grieksche, Latijnsche, Engelscl woord met daarachter de Nederlandsche vertaling.

⊞ **2.** Boek waarin in alphabetische of systematische orde de namen van zaken, personen, dieren, plaatsen, gebeurtenissen e.d. van een bep. vakgebied gegeven en verklaard worden; zaakwoordenboek, vakwoordenboek (vroeger ook *kunstwoordenboek* geheeten).

⊞ **3.** Boek waarin de afzonderlijke woorden, woordcombinaties (zooals spreekwoorden, uitdrukkingen, zegswijzen, gezegden en vaste verbindingen) en soms ook woorddeelen (zooals voorvoegsels en achtervoegsels) van een taal en hun beteekenis, morphologische en syntactische mogelijkheden zijn geïnventariseerd en beschreven. De woorden zijn meestal naar den vorm (alphabetisch) geordend, soms naar de beteekenis (systematisch).

↪ **4.** Al de woorden die een persoon kent, woordenschat, vocabulaire.

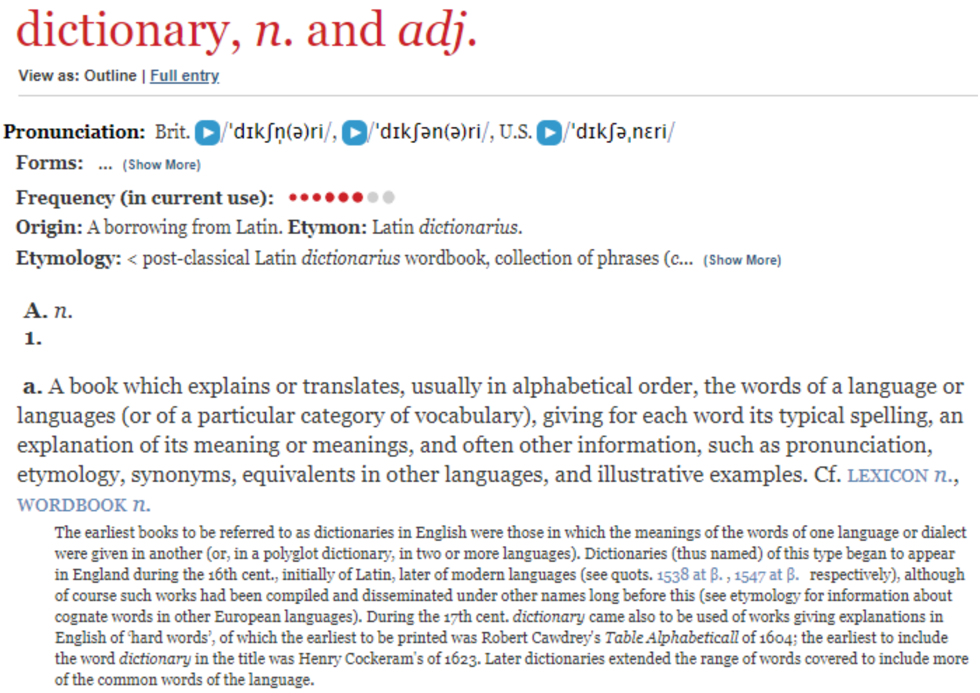
↪ **5.** Alle woorden van een taal, vakgebied, jargon of groep; woordenschat.

Fig. 2. Excerpt from the WNT dictionary.

The English Historical Dictionary OED. The Oxford English Dictionary (OED) is a reference dictionary for the English language. It is published by The Oxford University Press. The essential task of a dictionary would be to trace the history and the trajectory of each word, illustrating with quotations the nuances of meaning and uses that have emerged over time [2]. Figure 3 shows an excerpt[3] from the dictionary for the word 'Dictionary'. This excerpt shows

[2] http://gtb.inl.nl/iWDB/search?actie=article&wdb=WNT&id=M087164
[3] http://www.oed.com/view/Entry/52325

how the search for a word in the OED online dictionary gives information on the different pronunciation of this word, its etymology and its different meanings.



dictionary, *n.* and *adj.*

View as: Outline | Full entry

Pronunciation: Brit. ▶ /ˈdɪkʃn̩(ə)ri/, ▶ /ˈdɪkʃən(ə)ri/, U.S. ▶ /ˈdɪkʃəˌnɛri/

Forms: ... (Show More)

Frequency (in current use): ●●●●●●○○

Origin: A borrowing from Latin. **Etymon:** Latin *dictionarius.*

Etymology: < post-classical Latin *dictionarius* wordbook, collection of phrases (c... (Show More)

A. *n.*

1.

a. A book which explains or translates, usually in alphabetical order, the words of a language or languages (or of a particular category of vocabulary), giving for each word its typical spelling, an explanation of its meaning or meanings, and often other information, such as pronunciation, etymology, synonyms, equivalents in other languages, and illustrative examples. Cf. LEXICON *n.*, WORDBOOK *n.*

The earliest books to be referred to as dictionaries in English were those in which the meanings of the words of one language or dialect were given in another (or, in a polyglot dictionary, in two or more languages). Dictionaries (thus named) of this type began to appear in England during the 16th cent., initially of Latin, later of modern languages (see quots. 1538 at β. , 1547 at β. respectively), although of course such works had been compiled and disseminated under other names long before this (see etymology for information about cognate words in other European languages). During the 17th cent. *dictionary* came also to be used of works giving explanations in English of 'hard words', of which the earliest to be printed was Robert Cawdrey's *Table Alphabeticall* of 1604; the earliest to include the word *dictionary* in the title was Henry Cockeram's of 1623. Later dictionaries extended the range of words covered to include more of the common words of the language.

Fig. 3. Excerpt from the OED dictionary.

According to the study of the state of the art on the creation of historical dictionaries of other languages, we can see that the objectives of these dictionaries revolve around the study of the evolution of the meanings of the words since their first appearance.

Arabic Historical Dictionary. As for the historical dictionary for Arabic language, a first attempt of the historical dictionary was launched and directed by August Fischer in Egypt in 1935.

In April 2004, the Historical Dictionary of the Arabic Language Committee was founded by a decision of the Arab Language and Science Counselling Union in Cairo (Egypt). This project aims to create a historical dictionary of the Arabic words and their uses in order to indicate the change of their meanings through time and space [1]. It is still under the studying stage of the corpus.

A third attempt of the project was by the Arab Center for Research in Doha in 2013. The initial steps have been to prepare a reference bibliography of the sources of the linguistic corpus of the dictionary.

As a best of our knowledge, the creation of historical dictionaries for other languages does not use the natural language processing (NLP) tools, whereas for the attempts that have been made to create the historical dictionary for Arabic, they were not successful.

As for the attempt at the Arab Center for Research in Doha, the team pointed out the need to rely on NLP tools in the different steps of building the desired dictionary. In this context, Khalfallah et al.[12], proposed a platform of Automatic Natural Language Processing (ANLP) tools which permit the automatic indexing and research from Arabic texts corpus. They developed a system which allows to extract contexts from the entered corpus and to assign meaning by the user [13]. Another primordial step in the creation of a historical dictionary is by determining the correct meaning of a word in a given context, also known as WSD.

2.2 Word Sense Disambiguation: State of the Art

WSD Approaches. WSD is a natural language processing task of identifying the particular word senses of polysemous words used in a sentence [28]. WSD has become a prominent research area in the field of NLP. There are three main methods to WSD: knowledge-based method, supervised and unsupervised method [24].

Knowledge based methods rely on dictionaries, thesaurus, and knowledge to extract the definition of the ambiguous word. Unsupervised methods are based on training sets and do not use any structured resource while supervised methods are based on manually sense-annotated data sets [22]. Recently, the great advance in distributed semantics paved the way for the appearance of word embedding. Word embedding enables the computation of semantically related words. It can also be used to represent other linguistic units, such as phrases and short texts. Yet the problem of WSD has been approached from various perspectives in the context of word embeddings [25].

Word embeddings are of a major importance as they exhibit certain algebraic relations and can, therefore, be used for meaningful semantic operations. The latter include computing word similarity and capturing lexical relationships [23].

Recently some works have focused on word representation in vector space for the Arabic language such as[29,27,9].

However, most of the works used word embedding or any other techniques related to WSD were applied to Latin languages like English and French. But in the last decade, some attempts were applied to the Arabic language.

Arabic WSD Approaches. To the best of our knowledge, there seems to be no work concerned with disambiguating Arabic words according to their historical period in which they appeared. Hence, the idea of disambiguating words appearing in Old and Classical Arabic in order to create a historical dictionary is original.

In fact, all the works that focus on Arabic Word Sense Disambiguation are concerned with identifying the meaning of words in MSA.

In this Section, we are going to review the existing works related to the disambiguation of words in MSA. The work proposed by Bouhriz et al.[8] takes into consideration the local context and the global context defined by the full text during the disambiguation process. They have represented local context, global context and each sense of the ambiguous word with the help of vectors. Then, the appropriate sense of the target word is the sense that has the closest semantic proximity to its local and global context.

Alian et al.[5], relied on Arabic Wikipedia to extract the different meanings of the ambiguous word. They have applied Vector Space Model as a mathematical representation for documents. Vector Space Model serves to represent each retrieved texts from Wikipedia as a vector. Then, each text represented with the help of vectors is compared with the context of the word using cosine distance. The appropriate sense of the ambiguous word is therefore the concept of having the most cosine similarity.

Another method was proposed by Menai[17]. They have used genetic algorithms to solve word sense disambiguation problem. They tested their approach using a sample text in Arabic then they compared with naïve Bayes classifier [6].

Zouaghi et al.[30] have proposed an approach based on information retrieval measures. They have generated the contexts of use for each sense of the ambiguous word using its glosses. Then, the most probable sense is chosen by measuring the similarity between the different contexts generated and the current context of the ambiguous word.

The method proposed by Zouaghi et al.[31] is a hybrid method that combines unsupervised and knowledge based methods. They have used a context Matching algorithm that measures the similarity between the contexts of use corresponding to the glosses of the target word and the original sentence [6].

The most recent work proposed by Alkhatlan et al.[7] aims to disambiguate Arabic words using Arabic Wordnet and word embeddings. The main idea of this work is to represent each sense of the ambiguous word by a vector based on word2vec and Glove. The system proposed by Alkhatlan et al.[7] lists all the synsets which represent the ambiguous word along with their similarity to the context. It also chooses the synset that has the maximum similarity among synsets.

Table 1 presents a comparative study in the field of word sense disambiguation for Arabic. This comparison is performed using these criteria:

- The used method for WSD,
- The resources used to WSD,
- The testing data (number of ambiguous word used),
- The rate of precision.

Thanks to this study, we note that most of the works used a knowledge based approach for AWSD because these approaches provide a higher precision than the unsupervised approach.

[4] https://sites.google.com/site/mouradabbas9/corpora

Table 1. Comparative study of some AWSD approaches.

Author	WSD method	Used resources	Testing data	Precision
Bouhriz et al. [8]	Knowledge based AWSD	- Arabic Wordnet	A sample text in Arabic	74%
Alian et al. [5]	Knowledge based AWSD	- Arabic Wikipedia - Arabic Wordnet	7 ambiguous words	-
Menai [17]	Knowledge based approach	- Arabic WordNet - A sense annotated corpus	A sample text in Arabic	79%
Zouaghi et al. [30]	Knowledge based approach	- Arabic dictionary Alwassit - A collected corpus of 1500 Arabic texts	- 50 ambiguous words - 130 contexts of use for every word	73%
Zouaghi et al. [31]	Hybrid AWSD	- Arabic dictionary Alwassit - A collected corpus of 1500 Arabic texts	- 10 ambiguous words - 130 contexts of use for every word	79%
Alkhatlan et al. [7]	Knowledge based approach	- Arabic WordNet - Watan and Khaleej corpora[4]	- 10 ambiguous words - A collected corpus of 240 training samples	79%

3 Proposed Method

Our research consists in developing a framework to help linguists create a historical dictionary for Arabic. Thus, we have been inspired from the different outputs of the historical dictionaries already done and based on a description submitted by Doha site[5] of their aims behind the draft of the historical lexicon of Doha for the Arabic language. We have noticed that building the desired dictionary requires fundamental steps. One of them is by tracing the evolution of the meanings of a given word throughout time in addition to its historical information such as the date and the location of its first appearance. So we began by presenting the methodology of determining the meaning of a word in context, also known as Word Sense Disambiguation and then we detail the process of extraction of its historical information.

3.1 Methodology of Word Sense Disambiguation in Arabic

We propose here a method which permits to determine the meaning of an ambiguous word. This method aims not only to identify what a word means in

[5] https://www.dohadictionary.org/AR/Lexical_Services/Pages/Bibliography.aspx

a given context, but also to disambiguate it according to the historical period in which it appeared. Recently, word embedding has become a mainstay of natural language processing thanks to their ability to solve many NLP problems such as machine translation, sentiment analysis and even word sense disambiguation(WSD).

Word embeddings consist in building word representations in vector space based on the distributional hypothesis [11]: words that occur in the same contexts tend to have similar meanings [10].

In the spirit of representing words as vectors in a highly dimensional space, our method also benefits from word embedding to disambiguate Arabic words. This method is made up of two stages. The first one is about building an Arabic word embedding model using the skip gram technique [19]. The second one consists in calculating the similarities between the context of the ambiguous word and its definitions after representing, with the help of vectors, the context of the word to be disambiguated and its different glosses.

Arabic Word Embedding Model. The Word2vec tool [18] remains a popular choice benefited from its fast training and good results. In this work, we explore skip gram architecture to build neural word embeddings for Arabic. In fact, to build the word embedding model, we have used the Historical Arabic Dictionary Corpus (HADC) [4], which is originally designed to build a historical dictionary. It contains texts in Old Arabic, Classical Arabic and Modern Standard Arabic. A preprocessing step has been done before training word2vec model. First of all, we removed punctuation and non Arabic words. Then, we removed the stop words from the corpus based on a predifined list of stop words.

Similarity Calculation. To attribute for each ambiguous word its appropriate sense, we choose the sense with the closest semantic similarity to its local context. To measure the similarity between the context of use and each sense definition, three methods have been used. In what follows, we explain how to compute the semantic similarity among the context of use of the ambiguous word and its sense definition.

No weighting method. The simplest strategy to compare the context and the sense definition of the ambiguous word is by computing the sum of their words vectors. The similarity is subsequently measured for each meaning of the ambiguous word by using a cosine distance metric.

IDF weighting method. The core principle of this methodology is to assign a weight to each word in the context of the ambiguous word and its sense definition. These weights are based on the Inverse Document Frequency. The idea behind this is the word needed to determine most of the sentence's semantics usually have higher idf values [20]. The context vector (respectively sense vector) is represented by the sum of each word vector multiplied with its idf score.

To create a historical dictionary, we should take into consideration that certain words disappeared from the language and some new words appeared, and therefore the idf is calculated according to the periods in which the words appeared. Indeed, the corpus can be divided into three main periods. The first period contains texts in classical Arabic, the second period with texts in middle-age Arabic and the third period with texts in modern Arabic. Then, for each period, idf is calculated using the following formula [21]:

$$idf(w) = \log \frac{S}{WS}, \qquad (1)$$

where w is a word that appeared in a specephic period, S is the total number of sentences in this period and WS is the number of sentences containing the word w. The similarity is subsequently measured by using a cosine distance metric.

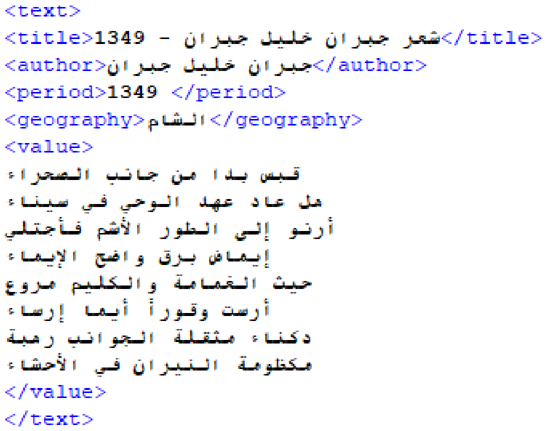
Word mover distance (WMD) method. WMD was introduced by [14]. WMD is a method that allows us to measure the distance between two documents (two sentences in our case). It takes into account the word's similarities in word embedding space. Indeed, we have used WMD to calculate the similarity between the context of use of the ambiguous word and its senses definition.

3.2 Methodology for Extraction of Historical Information for a Word

After the extraction of the meaning of the given word, we must determine the historical period when this word was first used. More precisely, our aim is to determine, for a given Arabic word, the date of its first appearance and the date of its sense's transformation. Then, we will store the history of Arabic words in an XML format [15].

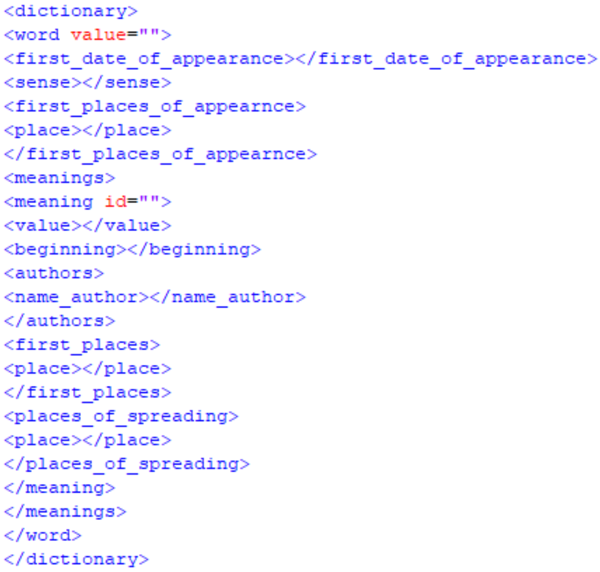
The first step is to represent the texts of the corpus in an XML format. The XML format allows us to capture the historical information of each document. In fact, the title of each document in the corpus is saved under the headings of: Author's name - Date of death. Therefore, the XML format of each document is automatically created by using the title of the document, the author and the period representing the date of death of the author. In fact, the date of death describes the historical period when that specific meaning was used. This could be explained with the fact that the date of the author's death gives a precise idea about the person's details.

However, the geography of appearance is extracted from Arabic Wikipedia. This step involves an XML description of all the text figures in the HADC corpus. The texts of the corpus are thus represented in the form of two files under different extension (TXT and XML). The TXT extension contains the value of the text and the XML extension contains the title, the author, the period and the geography of appearance of the text as well as its value (see figure 4).



```
<text>
<title>1349 - شعر جبران خليل جبران</title>
<author>جبران خليل جبران</author>
<period>1349 </period>
<geography>الشام</geography>
<value>
قبس بدا من جانب الصحراء
هل عاد عهد الوحي في سيناء
أرنو إلي الطور الأشم فأجتلي
إيماض برق واضح الإيماء
حيث الغمامة والكليم مروع
أرست وقوراً أيما إرساء
دكناء مثقلة الجوانب رهبة
مكظومة النيران في الأحشاء
</value>
</text>
```

Fig. 4. Structure of a XML file of a text extracted from the corpus.



```
<dictionary>
<word value="">
<first_date_of_appearance></first_date_of_appearance>
<sense></sense>
<first_places_of_appearnce>
<place></place>
</first_places_of_appearnce>
<meanings>
<meaning id="">
<value></value>
<beginning></beginning>
<authors>
<name_author></name_author>
</authors>
<first_places>
<place></place>
</first_places>
<places_of_spreading>
<place></place>
</places_of_spreading>
</meaning>
</meanings>
</word>
</dictionary>
```

Fig. 5. XML description of the dictionary of meaning.

The second step is to study the variation of the meanings of the word through time. Hence, the principle is as follows: for a given word, we first extract its meaning by applying our disambiguation method presented in section 3.1. Second, we stoke the historical information (period, place,user) related to the meaning of the word. This process will be repeated recursively for each document of the corpus containing the word being analyzed. Then, the meaning and the historical period in which this meaning was used is stoked in the XML format. It is worth pointing out that the XML model is automatically updated once the meaning of the word is found in an older document.

As for the word's first date of appearance, first place of appearance and origin, we will extract that automatically from the corpus. In fact, suppose that the texts of the corpus are historically classified from the oldest to the most recent ones, the first document containing the word in focus will reflect its first date of appearance, its origin and its user. We present an extract of the structure of our XML model in Fig. 5. It is necessary here to note that the texts of the corpus are lemmatized in order to identify the word's first occurrence. Finally, the XML model stores the following information:

- The oldest use of the word, together with its users, first date of appearance and first places of appearance,
- The meanings of the word historically classified according to its appearance in the corpus, specifying for each meaning the history of its oldest use, its oldest place of appearance and its users.

4 An Overview of the Developed System

In this section, we present the implementation details of our tool. The architecture of our system is shown in Fig 6.

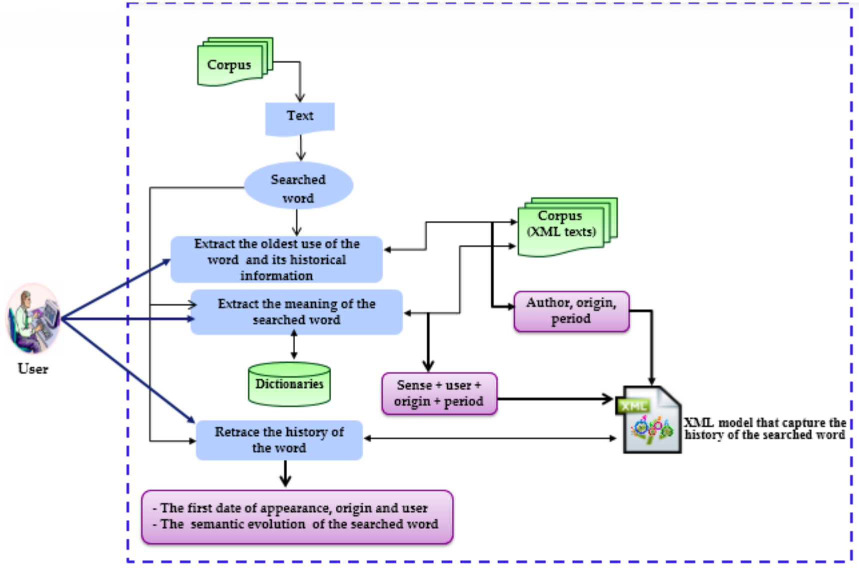

Corpus
Text
Searched word
Extract the oldest use of the word and its historical information
Corpus (XML texts)
Extract the meaning of the searched word
Author, origin, period
User
Dictionaries
Sense + user + origin + period
XML
XML model that capture the history of the searched word
Retrace the history of the word
- The first date of appearance, origin and user
- The semantic evolution of the searched word

Fig. 6. Architecture of the developed system.

To develop our tool, we used python programming language. We also used Gensim[6] toolkit to explore the pre-trained model, and PyQt4[7] toolkit to build our interface tool.

More precisely, the first version of our tool includes three main zones. The first one is called "The current text" in which the text will be uploaded to

[6] https://pypi.org/project/gensim/

[7] https://pypi.org/project/PyQt4/

display. The second one is called “Research operation” which represents the research possibilities (with criteria by entering a word to search, etc.). The last one contains three parts: “Historical information”, “Research Sense”, and “The Semantic evolution of a word”.

Therefore, our tool allows users to execute the following tasks:

– Extract the first date of appearance of a given word, as well as its users and its first places of appearance. The historical information extraction interface is presented in Fig. 7.
– Extract the meaning of the given word according to the historical period in which it appeared. The word sense disambiguation interface is presented in Fig. 8.
– Generate the meanings of the word historically classified according to its appearance in the corpus, specifying for each meaning the history of its oldest use, its oldest place of appearance and its users.

5 Evaluation

To evaluate our work, we conduct a series of experiments regarding the ability of word embeddings to solve WSD problem.

– Our corpus of test consists of 183 texts that appear in different historical periods.
– These texts have been selected from the Historical Arab Corpus HADC and Open Source Arabic Corpora (OSAC) corpus [26].
– We have tested about 100 ambiguous words. For each ambiguous word, we have used AntConc[8] to extract its contexts of use from the test corpus.
– As we have previously mentioned, our test corpus contains documents that appear in different periods from Classical Arabic to Modern Arabic. Then, we have randomly extracted, for each period, 100 contexts of use for each ambiguous word.
– We have used 150-dimensional Skip gram word embeddings.
– Moreover, to extract the different meanings of the ambiguous word taking into account the historical period in which the word appeared in the document, we have used four Arabic dictionaries that describe the different historical periods of the Arabic language.
  - Tahdhib Allougha Dictionary[9], for Old Arabic by Abou Mansour Azhari,
  - Tej Alarous Dictionary[10] by Murtadha Zbidi, For Intermediate Arabic Dictionaries,

[8] http://www.laurenceanthony.net/software/antconc/

[9] AlAzhari, Abu Mansour, Refining the Language. Dar Alamaarif, Cairo, 1976.

[10] Zabidi, Sayed Mortadha, Tej Alarous, Kuwait Government Press and the National Council for Culture and Arts, Kuwait from 1965 to 2002.

- Alwassit dictionary[11] and dictionary of contemporary Arabic language[12] for modern Arabic.

Table 2. The average precision obtained with stop words removal from the corpus when training words vector.

Pre-processing step	Precision		
	Old Arabic	Classical Arabic	MSA
With stop words removal	42,5%	43,9%	49%
Without stop words removal	48.54%	47.48%	59.42%

Table 3. Results of disambiguation words according to its appearance in the document.

Method	Precision		
	Old Arabic	Classical Arabic	MSA
No weighting method	48.54%	47.48%	59.42%
MDA distance method	46,94%	44,10%	50,26%
IDF weighting method	48,89%	48,57%	63,44%

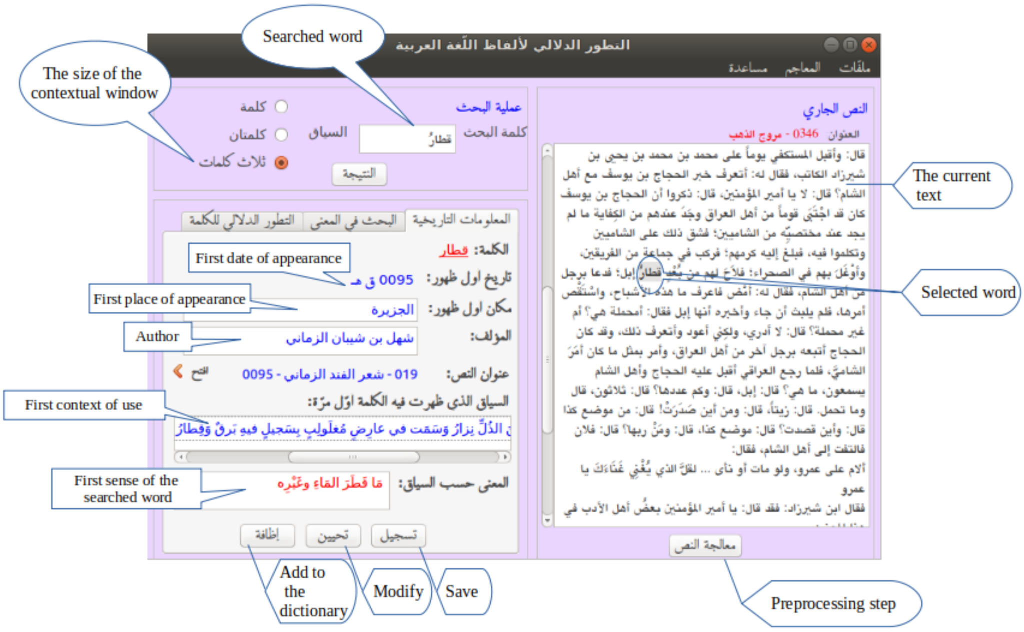
We have semi-automatically developed a structured electronic dictionary with an XML format containing the glosses of 100 ambiguous words in the Old Arabic. Similarly, we have developed a dictionary that contains the glosses of 100 ambiguous words extracted from Tej Alarouss. Thus, the last two dictionaries Tahdhib Alougha and Tej Alarous are manually structured because they have complex structures, which varies from one entry to another and are characterized by a quasi-absence of marker. For words in Modern Standard Arabic, the two dictionaries Alwaseet and Almouasera are used. Indeed, we have an HTML version of these two dictionaries.

These two dictionaries are distinguished by a set of markers facilitating the transformation of their raw content to a structured version in XML. Then we have automatically transformed them into a structured electronic XML format.

The first part of this evaluation is to test the impact of removing stop words from the corpus when training word vectors. Results are shown in table 2. For the word sense disambiguation task, we have noticed that without removal stop words from the corpus, our trained model exhibits stronger performances compared to the model obtained with trained corpus without stop words.

[11] The 5th Edition of the Alwasseet Dictionary published by the Arabic Language Complex, in Cairo in 2011.

[12] Mokhtar, Omar Ahmed, Modern Arabic Language, The Universe of Books, Cairo, 2008.



كلمة البحث: قطارٌ

السياق: ● ثلاث كلمات

الكلمة: قطار

تاريخ اول ظهور: 0095 ق هـ

مكان اول ظهور: الجزيرة

المؤلف: شهل بن شيبان الزماني

عنوان النص: 019 - شعر الفند الزماني - 0095

السياق الذي ظهرت فيه الكلمة اوّل مرّة:

الذُّلِّ نِزارٌ وسَمَت في عارِضٍ مُغلَولِبٍ يسَحِيلٍ فيهِ بَرقٌ وقِطارٌ

المعنى حسب السياق: مَا قطَرَ القاءِ وغَبْرِه

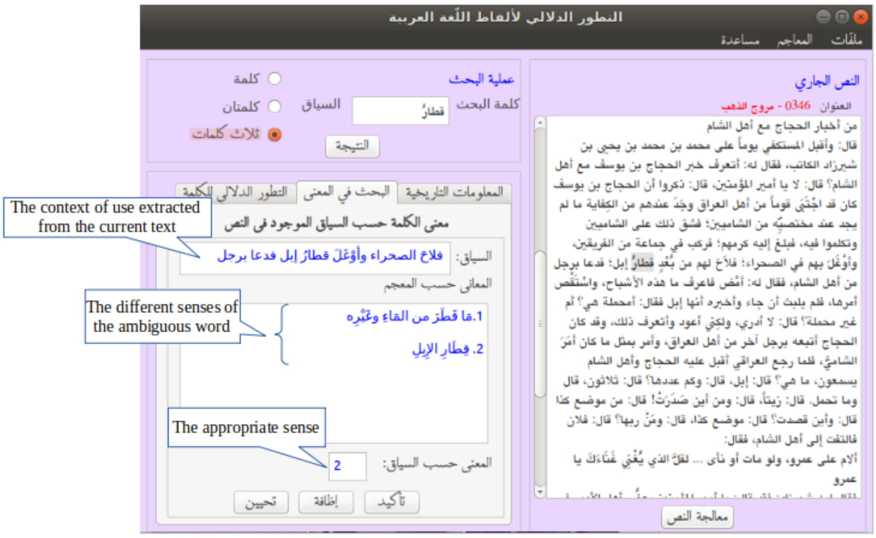
العنوان 0346 - مروج الذهب

قال: وأقبل المستكفي يوماً على محمد بن محمد بن يحيى بن شيرزاد الكاتب، فقال له: أتعرف خبر الحجاج بن يوسف مع أهل الشام؟ قال: لا يا أمير المؤمنين، قال: ذكروا أن الحجاج بن يوسف كان قد اجْتَبَى قوماً من أهل العراق وجَدَ عندهم من الكِفاية ما لم يجد عند مختصيِّه من الشاميين؛ فشق ذلك على الشاميين وتكلموا فيه، فبلغ إليه كرمهم؛ فركب في جماعة من الفريقين، وأوْغَلَ بهم في الصحراء؛ فلاَحَ لهم من بُعْدٍ قطارٌ إبل؛ فدعا برجل من أهل الشام، فقال له: أمْضِ فاعرف ما هذه الأشباح، واسْتَقْصِ أمرها، فلم يلبث أن جاء وأخبره أنها إبل فقال: أمحملة هي؟ أم غير محملة؟ قال: لا أدري، ولكِنِي أعود وأتعرف ذلك، وقد كان الحجاج أتبعه برجل آخر من أهل العراق، وأمر بمثل ما كان أمَرَ الشاميَّ، فلما رجع العراقي أقبل عليه الحجاج وأهل الشام يسمعون، ما هي؟ قال: إبل، قال: وكم عددها؟ قال: ثلاثون، قال وما تحمل. قال: زيتاً، قال: ومن أين صَدَرَتْ! قال: من موضع كذا قال: وأين قصدت؟ قال: موضع كذا، قال: ومَنْ ربها؟ قال: فلان فالتفت إلى أهل الشام، فقال:

ألام على عمرو، ولو مات أو نأى ... لقلّ الذي يُغْنِي غَنَاءَكَ يا عمرو

فقال ابن شيرزاد: فقد قال: يا أمير المؤمنين بعضُ أهل الأدب في

Fig. 7. Historical information extraction interface.



كلمة البحث: قطارٌ

السياق: ● ثلاث كلمات

معنى الكلمة حسب السياق الموجود فى النص

السياق: فلاحَ الصحراء وأوْغَلَ قطارٌ إبل فدعا برجل

المعانى حسب المعجم

1.مَا قطَرَ من القاءِ وغَبْرِه

2. قِطَارِ الإِبِلِ

المعنى حسب السياق: 2

العنوان 0346 - مروج الذهب

من أخبار الحجاج مع أهل الشام

قال: وأقبل المستكفي يوماً على محمد بن محمد بن يحيى بن شيرزاد الكاتب، فقال له: أتعرف خبر الحجاج بن يوسف مع أهل الشام؟ قال: لا يا أمير المؤمنين، قال: ذكروا أن الحجاج بن يوسف كان قد اجْتَبَى قوماً من أهل العراق وجَدَ عندهم من الكِفاية ما لم يجد عند مختصيِّه من الشاميين؛ فشق ذلك على الشاميين وتكلموا فيه، فبلغ إليه كرمهم؛ فركب في جماعة من الفريقين، وأوْغَلَ بهم في الصحراء؛ فلاَحَ لهم من بُعْدٍ قطارٌ إبل؛ فدعا برجل من أهل الشام، فقال له: أمْضِ فاعرف ما هذه الأشباح، واسْتَقْصِ أمرها، فلم يلبث أن جاء وأخبره أنها إبل فقال: أمحملة هي؟ أم غير محملة؟ قال: لا أدري، ولكِنِي أعود وأتعرف ذلك، وقد كان الحجاج أتبعه برجل آخر من أهل العراق، وأمر بمثل ما كان أمَرَ الشاميَّ، فلما رجع العراقي أقبل عليه الحجاج وأهل الشام يسمعون، ما هي؟ قال: إبل، قال: وكم عددها؟ قال: ثلاثون، قال وما تحمل. قال: زيتاً، قال: ومن أين صَدَرَتْ! قال: من موضع كذا قال: وأين قصدت؟ قال: موضع كذا، قال: ومَنْ ربها؟ قال: فلان فالتفت إلى أهل الشام، فقال:

ألام على عمرو، ولو مات أو نأى ... لقلّ الذي يُغْنِي غَنَاءَكَ يا عمرو

Fig. 8. Word sense disambiguation interface.

Accordingly, we considered the word embeddings obtained without removing stop words from the corpus when trained words vectors. The second part is to evaluate the capacity of word embeddings to represent the sentence containing the ambiguous word and its sense definitions. Specifically, we consider three method: no weighting method, IDF weighting method and WMD distance method.

As illustrated in table 3, IDF weighting method achieves better results on WSD tasks. Unexpectedly, the cosine distance between average word vectors (No weighting method) is more powerful than WMD metric as it captures the meaning similarities between the context of use of the ambiguous word and its sense definitions.

6 Conclusion

In this paper, we presented our tool that aims to help linguists to create a historical dictionary of Arabic. The implemented method consists of two steps: in the first one we extract the meaning of an ambiguous Arabic word according to the historical period in which it appeared together with its first date of appearance, its users and its first place of appearance. Second, we generate the different meanings of a given word historically classified according to its appearance in the corpus. 100 ambiguous words have been chosen for the test. Each context of use has been extracted according to to three specific periods. Experiments have shown an accuracy of 48,89% for the Old Arabic, 48,57% for the Classical Arabic and 63,44% for the Modern Standard Arabic.

References

1. Abdel-Aziz, M.H.: A Historical Dictionary for Arabic Language: Documents and examples (2008)
2. Al-Said, A.B.: A Corpus-based Historical Arabic Dictionary: Linguistic and Computational processing. Ph.D. thesis, Cairo University (2011)
3. Al-Said, A.B.: The historical arabic dictionary resources. Journal Of the Arab languages (2015)
4. Al-Said, A.B., Medea-García, L.: The historical arabic dictionary corpus and its suitability for a grammaticalization approach. In: 5^{th} international conference in linguistics (2014)
5. Alian, M., Awajan, A., Al-Kouz, A.: Arabic word sense disambiguation using wikipedia. International Journal of Computing and Information Sciences (2016)
6. Alian, M., Awajan, A., Al-Kouz, A.: Arabic word sense disambiguation - survey. In: International Conference on New Trends in Computing Sciences (2017)
7. Alkhatlan, A., Kalita, J., Alhaddad, A.: Word sense disambiguation for arabic exploiting arabic wordnet and word embedding and word embedding. In: The 4th International Conference On Arabic Compitational Linguistics (ACLing) (2018)
8. Bouhriz, N., Benabbou, F., Lahmar, E.H.B.: Word sense disambiguation approach for arabic text. International Journal of Advanced Computer Science and Applications (2016)
9. Dahou, A., Xiong, S., Zhou, J., Haddoud, M.H., Duan, P.: Word embeddings and convolutional neural network for arabicsentiment classification. In: Proceedings of COLING (2016)

10. Frej, J., Chevallet, J.P., Schwab, D.: Enhancing translation language models with word embedding for information retrieval. The Computing Research Repository (CoRR) (2018)
11. Harris, Z.S.: Distributional structure. Word (1954)
12. Khalfallah, F., Aloulou, C., Belguith, L.H.: Had, a platform to create a historical dictionary. In: AICCSA (2016)
13. Khalfallah, F., Aloulou, C., Belguith, L.H.: A platform based anlp tools for the construction of an arabic historical dictionary. In: • (ed.) NLDB (2016)
14. Kusner, M.J., Sun, Y., Kolkin, N.I., Weinberger, K.Q.: From word embeddings to document distances. In: Proceedings of the 32^{nd} International Conference on Machine Learning (2015)
15. Laatar, R., Aloulou, C., Hadrich-Belguith, L.: An xml model for an arabic historical dictionary. In: LPKM (2018)
16. Masmoudi, A., Bougares, F., Ellouze, M., Estève, Y., Belgui, L.: Automatic speech recognition system for tunisian dialect. Lang Resources and Evaluation (2017)
17. Menai, M.E.B.: Word sense disambiguation using evolutionary algorithms – application to arabic language. Computers in Human Behavior (2014)
18. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. In: Proceedings of the International Conference on Learning Representations (ICLR) (2013)
19. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.A.: Distributed representations of words and phrases and their compositionality. In: The 26th International Conference on Neural Information Processing Systems (2013)
20. Nagoudi, E.M.B., Schwab, D.: Semantic similarity of arabic sentences with word embeddings. In: WANLP-EACL (2017)
21. Nagoudi, E.M.B., Schwab, D.: Semantic similarity of arabic sentences with word embeddings. In: WANLP@EACL (2017)
22. Navigli, R.: Word sense disambiguation : A survey. ACM Computing Surveys (2009)
23. Oele, D., van Noord, G.: Distributional lesk: Effective knowledge-based word sense disambiguation. In: International Conference on Computational Semantics (2017)
24. Pal, A.R., Saha, D.: Word sense disambiguation: A survey. International Journal of Control Theory and Computer Modeling (IJCTCM) (2015)
25. Pelevina, M., Arefiev, N., Biemann, C., Panchenko, A.: Making sense of word embeddings. In: Proceedings of the 1st Workshop on Representation Learning for NLP (2016)
26. Saad, M., Ashour, W.: Osac: Open source arabic corpora. In: International Conference on Electrical and Computer Systems (2010)
27. Soliman, A.B., Eissa, K., El-Beltagy, S.R.: Aravec: A set of arabic word embedding models for use in arabic nlp. In: 3rd International Conference on Arabic Computational Linguistics, ACLing (2017)
28. Ustalov, D., Teslenko, D., Panchenko, A., Chernoskutov, M., Biemann, C., Ponzetto, S.P.: An unsupervised word sense disambiguation system for under-resourced languages. In: In Proceedings of the 11^th Conference on Language Resources and Evaluation (2018)
29. Zahran, M.A., Magooda, A., Mahgoub, A.Y., Raafat, H.: Word representations in vector space and their applications for arabic. In: CICLing (2015)
30. Zouaghi, A., Merhbene, L., Zrigui, M.: Combination of information retrieval methods with lesk algorithm for arabic word sense disambiguation. Artificial Intelligence Review (2012)

31. Zouaghi, A., Merhbene, L., Zrigui, M.: A hybrid approach for arabic word sense disambiguation. International Journal of Computer Processing Of Languages (2012)