CAMS-KG: a Classical Arabic Morpho-Semantic Knowledge Graph

Editor(s): Name Surname, University, Country
Solicited review(s): Name Surname, University, Country
Open review(s): Name Surname, University, Country

Ibrahim Bounhas [a,b,e,*], Nadia Soudani[a,c,e] Yahya Slimani[a,d,e]
[a] *LISI Laboratory of Computer Science for Industrial Systems, Carthage University, Tunisia*
[b] *Higher Institute of documentation, La Manouba University, Tunisia*
[c] *National school of Computer Sciences, La Manouba University, Tunisia*
[d] *Higher Institute of Multimedia Arts of Manouba, La Manouba University, Tunisia*
[e] *Jarir: Joint group for Artificial Reasoning and Information Retrieval (www.jarir.tn)*
{Bounhas.ibrahim, nadia.soudani, yahya.slimani}@gmail.com

Abstract. In this paper we propose to build a morpho-semantic knowledge graph from Arabic vocalized corpora. Our work focuses on classical Arabic as it has not deeply investigated in related works. We use a tool suite which allows analyzing and disambiguating Arabic texts, taking into account short diacritics to reduce ambiguities. At the morphological level, we combine Ghwanmeh stemmer and MADAMIRA which are adapted to extract a multi-level lexicon from Arabic vocalized corpora. At the semantic level, we infer semantic dependencies between tokens by exploiting contextual knowledge extracted by a concordancer. Both morphological and semantic links are represented through compressed graphs, which are accessed through lazy methods. These graphs are mined using BM25 measure to compute on-to-many similarity. Indeed, we propose to evaluate CAMS-KG in the context of Arabic Information Retrieval (IR). Several scenarios of document indexing and query expansion are assessed. That is, we vary indexing units for Arabic IR based on different levels of morphological knowledge, a challenging issue which is not yet resolved in related works. We also experiment several combinations of morpho-semantic query expansion. This permits to validate our resource and to study its impact on IR based on state-of-the art evaluation metrics.

Keywords: Morpho-semantic knowledge extraction, Classical Arabic text mining, Arabic information retrieval, graph-based knowledge representation.

* Corresponding author: E-mail: Bounhas.ibrahim@gmail.com.

1. Introduction

An emergent need sprouts to look for new techniques and tools to use the large mass of data accessible in the web in an effective and intelligent manner [62]. Hence, stored and processed data would be automatically transformed in useful information and effective knowledge. Search engines have to turn into conscious systems by integrating into their search process reasoning and inferring algorithms and models. These approaches aim to understand user queries and retrieve accurate information that suits and approximates the required user cognitive model. An important feature of semantic search engines is the quality of the lexico-semantic resource. The more the resource is semantically rich, formalized and accessible, the easier is the possibility to operate with it and to share its content. From this observation which is accorded to the semantic web view in handling and processing data on the web and how to cope better by integrating such resources with an overwhelming amount of information, we promote the fact to contribute in the construction of such resources [20].

We are trying to contribute in this field by exploiting existing electronic resources and transforming them in more structured and formal representations. Indeed, such resources are crucial to enhance IRS. Due to the shortness of queries and language ambiguities, many irrelevant documents may be evaluated as relevant. Besides, many relevant documents containing some morphologically or semantically related terms are not considered. To resolve such problems Information Retrieval Systems (IRS) in the era of semantic Web need to deeply analyze and interpret both queries and documents. Natural Language Processing (NLP) and semantic web techniques are combined to reach this goal. Knowledge Graphs (KG) constitute an emergent solution to support IRS in enhancing retrieval effectiveness [85, 86].

Several related works on the field of KG construction and mining focused on Latin origin languages like English [76]. For Arabic, growing efforts are substantially devoted to enhance NLP tools and IR applications [62]. Nevertheless, Arabic semantic IR is yet challenging given the complex morphology of this language [27]. Building Arabic knowledge graphs needs to consider and to model both morphological and semantic features. In this context, we suggest to build a morpho-semantic KG to support Arabic knowledge representation and IR. Nevertheless, our literature review (cf. section 4.1 and 4.2.1) shows that existent works mainly focused on modern texts which are actually produced and shared on the Web. Other important funds such as classical Arabic poesy and literature are not yet well studied, while they are available on the Web. For instance the Quranic corpus was studied in several works (e.g. [65]).

Our approach is based on a deep study of Arabic morphology. We study the impact of diacritics and morphological disambiguation on KG construction and IR. Indeed, short diacritics affect the morpho-syntactic features of words and their semantics (cf. section 2). Moreover, we model the different levels of the Arabic lexicon to study all word forms. To fulfill these goals, we build a new morpho-semantic resource from Arabic vocalized corpora based on a text mining process. This resource gathers rich morphological and semantic knowledge. It is stored by means of compressed graphs, using the Webgraph[1] framework, which exploits lazy methods to access and mine large-scale graphs [73]. This allows us to compute the similarities of Arabic tokens taking advantage of the richness of the Arabic lexicon and a contextual knowledge base extracted from corpora. CAMS-KG is evaluated on the context of Arabic IR by exploiting morphological and semantic knowledge for query expansion. Furthermore, as we model several levels of the Arabic lexicon, we could perform a comparative experimental examination by varying indexing units in Arabic IR.

In the following section, we study morphological and semantic ambiguities in Arabic texts and focus on their complexities and relationships. Then, we detail the process adopted in this paper to build and asses CAMS-KG. We start by specifying the components of our resource in Section 3. Section 4 details the steps allowing to build CAMS-KG, which is evaluated in Section 5. In all these sections, we study existent approaches and tools to justify our choices. Finally, we conclude this paper and provide some perspectives for future research.

[1] http://webgraph.di.unimi.it/

2. Arabic morphology

2.1. Arabic Language features

Arabic is a right to left Semitic language. It is cursive, agglutinative, highly inflectional and derivational [3, 23, 29]. Words are obtained by adding affixes to the base of the word which is derived from a root. Moreover, Arabic uses specific signs written above or below letters to indicate the proper pronunciation of words. Called diacritics or short vowels, they help induce the correct meaning of a word.

The morphology of Arabic language defines several rules which allow studying the structure of words [25, 57]. It permits to analyze word forms and morphological relations based on derivation and inflection. Automatic morphological analysis aims to recognize the different lexical units (Nouns, Verbs and Particles) and their morpho-syntactic attributes [57]. The Arabic lexicon is built based on three principal processes; namely derivation, inflection and word construction [67].

i) Derivation: it consists in deriving lemmas from roots by applying patterns/templates. For example, from the root (ع،ل، م), we may derive the verb "عَلِمَ" (Ealima/to know) and the noun "عِلْم" (Eilom/science), by applying the patterns "فَعِلَ" and "فِعْل" respectively. It is also possible to derive "augmented verbs" from which nouns can be derived. For example, from the same root, we may derive the augmented verb"عَلَّمَ" (Eal~ama/ to teach), by applying the pattern "فَعَّل". From this verb, we may generate the noun "مُعَلِّم" (muEal~im/a teacher) by applying the pattern "مُفَعِّل".

ii) Inflection: consists in conjugating singular forms of nouns and infinitive forms of verbs, which results in changing the values of their morpho-syntactic attributes, such as the gender for nouns and the aspect for verbs[2]. Often different affixes (prefixes, infixes and suffixes) are added to the original word. For example, the verb "عَلَّمَ" (Eal~ama/ to teach) may be conjugated with the first singular person to obtain "عَلَّمْتُ" (Eal~amotu/I teached) and "أُعَلِّمُ" (>uEal~imu/I teach) in the perfect and imperfect forms respectively. The plural form of the word "مُعَلِّم" (muEal~im/a teacher) is "مُعَلِّمُون" (muEal~imuwn/teachers), etc.

iii) Word construction: proclitics and/or enclitics are agglutinated to stems to get more complex forms of words called maximal written words. For example, when we add a determiner ("ال"; Al; the) and a conjunction ("و"; w; and) to the previous stem, we obtain "والمُعَلِّمُون" (wAlmuEal~imuwn /and the teachers).

We resume the different types of Arabic lexical entries as follows [57, 67]:

- **Root**: an abstract unit representing the origin of a word (e.g. "ع،ل، م").
- V**erbed pattern**: refers to pattern/template of the simple or the augmented verb derived from a root (e.g. "عَلَّمَ" (Eal~ama/ to teach)) in the infinitive form (third singular masculine person).
- **Lemma**: it is obtained after derivation (e.g. "مُعَلِّم"(muEal~im/a teacher)). The lemma constitutes the minimal lexical unit having a sense without any inflection. Most of lemmas are singular masculine nouns or verbs conjugated with the third singular masculine person.
- **Stem**: the inflected form of verbal and nominal lemmas. It is also the form of the word obtained after adding inflectional marks (e.g. "مُعَلِّمُون" (muEal~imuwn/teachers)).
- **Word**: written forms obtained by eventually adding enclitics and/or proclitics to stems.

For a more detailed description of Arabic word construction process, we refer the readers to Habash's book [67] and SARF[3].

2.2. Arabic morpho-semantic ambiguities

The absence of short diacritics may lead to morphological ambiguities [33, 34], which can affect the semantic interpretation of unique words and whole sentences. For example, the sentence "المَلِك عَادِل يُحِبُّ رَعِيَّتَه" (Almalik EaAdil yuHib~u raEiy~tah) can have two different possible interpretations namely, i) "The king is fair and he loves his people"; and, ii) "The king Adel loves his people". In fact, the word "عادل/ EaAdil " can be an adjective that means "fair" or a proper noun where EaAdil is the first name of the king. Despite the

[2] Table 1 provides the complete list for Arabic morpho-syntactic attributes.

[3] https://sourceforge.net/projects/sarf

word is vocalized, its Part of Speech (POS) (i.e. grammatical category) could not be correctly determined. Indeed, words derived from the same root may share some semantic features. For example, the words الإجارة (Al<jArp), الأجرة(Al>jrp) and الأجارة(Al>jArp) have in common a shared meaning of "a reward for a work".

Morphological features as gender may influence meaning. For example, the masculine word "مَكتَب" (maktab) means "office", while its feminine ("المكتبة"; Almaktabap) is a synonym of "library". Heteronyms that have the same spelling (homographs), but not the same vocalization (heterophones) differ in meaning and they are a source of ambiguity especially if they are not vocalized as حمل (Hml) which can be a verb or a noun. It can be حَمَلَ (Hamala /to carry) or حُمِل(passive voice of hamala:Humil / to be carried) or حَمّلَ (Ham~la /to charge) or حُمّل (Hum~l /to be placed on). It can also be حَمْل (Hamol /the fetus or the pregnancy or the action of carrying) and so on. Enhancing Arabic semantic retrieval needs to consider and model these features in KG construction.

3. CAMS-KG specification

3.1. Existent resources

This section studies existent Arabic semantic and morpho-semantic resources and their limits tune some choices in CAMS-KG specification .

3.1.1. Semantic resources

In the semantic axis, Arabic WordNet (AWN) constitutes an important resource exploited for different purposes including Query Expansion (QE) [36, 48, 66]. For example, Atwan et al. (2016) [36] proposed a QE approach based on the test collection combined with AWN. Abbache et al. (2016) [2] performed short query expansion by combining AWN and a sub-corpus from the Xinhua collection. They used association rules between terms to select the appropriate synonyms from AWN. In their experiments, they exhibit that this approach can improve the effectiveness of an Arabic IRS. Authors in [54, 55, 60] proposed semantic indexing, which allows to recognize for each term in a query or in a document a concept from AWN. Mahgoub et al. (2006) [13] presented a QE approach using three Arabic resources; namely Arabic Wikipedia, "Al Raed" dictionary and "Google_WordNet", which is the Google translation of English Wordnet into Arabic. For Wikipedia, authors grouped terms by studying the redirect links and article glosses.

3.1.2. Morpho-Semantic resources

Due to the characteristics of Arabic Language and the so deep and direct dependence between morphology and semantics, some researchers tried to model the relationships between semantics and morphology. For example, Belkredim et al. (2008) developed a derivational Arabic ontology based on verbs [28]. Other researchers tried to exploit the morphological relations between terms when building terminologies by grouping similar terms having a common root [78]. ElKateb et al. (2006) [78] proposed to extend AWN by considering that words having the same root are semantically similar. Also, Hattab et al. (2009) [63] exploited the morpho-syntactic attributes of words as the POS, the root and the scheme to determine the similarity degree between two words.

Hammo et al. (2008) [18] developed an IRS that was run on scripts of the holy Quran and a collection of forty non-vocalized words obtained from 10 college students. Each student has been asked to provide 4 words that he memorizes from the Quran. Three types of indexes are provided: Vocalized-Word Index, Non-Vocalized-Word Index and Root Index by using Khoja stemmer. A thesaurus of semantic classes is used to expand user queries.

Anizi and Dichy (2009) [56] proposed dictionary-based search. They used DIINAR.1 to add words which have the same lemmas as the query terms. Another approach that consists in adding words appearing in the contexts of query terms in ArabiCorpus[4] is also carried out.

Al-zoghby and Shaalan (2015) [9, 10] proposed a semantic IRS based on morpho-semantic indexing, but no morphological or semantic disambiguation task was performed. The experiments are performed on three types of indexing spaces; namely (i) MTS: Morphological Term Space; (ii) STS: Semantic Term Space; and, (iii) CS: Conceptual Space. MTS is a documents index where terms are enriched with morphological expansions by adding derivations with use of the RDI Mor-

[4] http://arabicorpus.byu.edu/

phological Analyzer and the synonymous terms. STS enhances MTS by computing similarities based on Universal WordNet (UWN). Whereas, CS is a Concept-Space by which semantically related terms are grouped into concepts based on semantic relations as synonymy, generalization properties (Super-Classes, Instance-Of) and specialization properties (Sub-classes, Has-Instances).

3.1.3. Synthesis

Existent works build or exploit several types of resources such as dictionaries [69], ontologies [24, 61] conceptual spaces including semantic relations [31, 4] and AWN [1, 8, 36, 48]. Although some existing resources have a good coverage and are rich with semantic relations, we note the lack of a resource which combines both morphological and semantic knowledge. For example, DIINAR [56] contains morphological relations only, while AWN is a semantic network. A typical resource should contain morphological relations representing the derivational and flectional process, morpho-syntactic attributes (e.g. gender, number, POS, etc.) and distributional knowledge allowing to compute non-boolean similarity measures (e.g. co-occurrence links). For example, in AWN the relations between tokens are not weighted and thus have the same degree of similarity.

Furthermore, short diacritics have a crucial role in understanding the interpretation of words in the morphological and semantic levels. These signs may be used to reduce ambiguities in building knowledge graphs. Unfortunately, most Arabic NLP tools (cf. section 4.1.1), do not take them into account. Besides, only few works tried to exploit them [33, 34]. Finally most related works focus on standard modern Arabic and classical Arabic is not well investigated.

3.2. CAMS-KG components

Based on the limits of existent resources, we propose to build new resources which combine morphological and semantic knowledge. We focus on classical Arabic and consider short vowels while mining our corpora to reduce ambiguities.

We generate from a corpus a graph representing the morpho-semantic relations between the words where different types of nodes are modeled. The process of knowledge extraction is based on a hybrid approach combining: i) linguistic knowledge consisting of lexical units and their morpho-syntactic attributes; and, ii) distributional knowledge representing co-occurrence links. CAMS-KG covers the different levels of the Arabic lexicon; namely roots, verbed patterns, lemmas, stems and words in their vocalized and non vocalized forms.

3.2.1. Nodes and edges

In this section, we specify the types of nodes which should be included in our KG i.e. the types of lexical entries detailed in section 2.1. Related works in the fields of Arabic text mining and IR discussed this issue to select the best indexing unit which represents the meaning of words. In this context, Elayeb and Bounhas [15] reveal that indexing documents by surface words leads to a high precision, but a low recall in IR. They justify this by the fact that other morphological variants of the same word may exist in documents. Other researchers [5, 51, 52] carried out document indexing by lightly stemmed words. Light stemming-based indexing allows better precision and optimizes the search process in terms of storage size and processing time, compared to surface word-based indexing [52]. As reported in [43], some studies compared roots, stems and surface words for indexing. However, these observations cannot be generalized [15].

That is, we can conclude that the indexing unit affects remarkably the IR results. For instance, the study done by Darwish et al. (2005) [43] revolves that returning all the possible solutions of a given word by the Arabic Morphological Analysis (AMA) tools then considering them in the retrieval process "*complicates retrieval, because it introduces ambiguity in the indexing phase as well as the search phase*". Moreover, they unfold that the main limits for AMA-based corpus processing are related to "*issues of coverage and correctness*". In addressing the issue of correctness, several statistical techniques are used to sort out inaccurate and irrelevant morphological solutions. For example, Sebawi morphological analyzer [41] evaluates any segmentation of a given word relies on the product of three probabilities computed respectively based on the frequencies of the prefix, the suffix and the template of the stem of the word. Darwish and Oard (2007) [44] compared words, n-grams, light stems, *agressive* stems and top-ranked roots returned by Sebawi. The experiments testify to the fact that 3-grams or 4-grams, lightly

or aggressively stemmed are better for indexing Arabic documents. In another work [91] in which the IBM-LM tool is used, authors indexed documents with 3-gram morpheme language model trained on LDC's Arabic Treebank: Part 1 v 2.0 to filter all the possible segmentations of a word. A Prefix-Suffix filter is also applied to measure how likely a prefix may occur with a suffix.

As far as coverage is concerned, recent analyzers have a better coverage of the language. This is shown through results stemmed by AMIRA in IR [40]. Moreover, MADA was used in diverse types of applications as Cross-Language Information Retrieval [26], question answering [64, 74] and profile-based IR [32]. In a recent work [70], authors proposed a semantic retrieval approach varying NLP tools, resources and indexing units. Approaches based respectively on roots, stems and lemmas are compared. Authors showed that the lemma-based semantic retrieval outperforms all the other indexing units.

From this study, we can conclude that the problem of indexing unit is not yet resolved [15, 70]. That's why we integrate in our knowledge graphs several levels of the Arabic lexicon. CAMS-KG includes morphological knowledge modeling the derivational and flectional process of Arabic words. It contains six types of nodes corresponding to the different layers of the Arabic lexicon (root, verbed pattern, lemma, stem, vocalized word, non-vocalized word). On the one hand, we define morphological links (e.g. the flectional relation between a lemma and a stem). In the semantic layer, we represent co-occurrence links between all the types of tokens which will allow computing semantic similarities. Furthermore, we capture the morpho-syntactic attributes of each token as detailed bellow.

3.2.2. Attributes

Table 1 summarizes the morpho-syntactic attributes used to describe Arabic lexical units. Some of them are common between all word categories. Others apply only for nouns, verbs or particles. An example is unfolded in table 2..

Table 1

Morpho-syntactic attributes of Arabic morphemes and words

Attribute	Values/remarks	Verb	Noun	Particle
	Common attributes			
Aspect	Imperative; Perfect; Imperfect	X		
Case	Accusative; Genitive; Nominative		X	X (Rule 1)
State	Defined; Not Defined; Defined by annexation, NOUN_PROP, DET		X	
Gender	Masculine; Feminine	X	X	X (Rule 2)
Lemma	The lemma	X	X	X
Mood	Only for imperfect verbs: Aprocoped; Indicative; Subjunctive	X		
Number	Singular; Plural; Dual	X	X	X (Rule 2)
Person	1; 2; 3.	X		X (Rule 3)
POS	The grammatical category.	X	X	X
Stem	The stem	X	X	X
Translation	The translation	X	X	X
Voice	Active; Passive	X		
	Word attributes			
CONJ	Does the word contain a conjunction (Y/N)	X	X	X
PREP	Does the word contain a preposition (Y/N)	X	X	X
PART	Does the word contain a particle (Y/N)	X	X	X
PRON	Does the word contain a pronoun (Y/N)	X	X	X

Rules 1: Applys only for demonstrative pronouns, relative pronouns and some particles related to proper nouns like "أبو" (Father of)
Rules 2: Applys only for pronouns and the particles "كلا" (masculine) and "كلتا" (feminine) meaning "both"
Rules 3: Applys only for some pronouns and the flectional forms of some particles as "ليس" (not)

Table 2

Examples of morpho-syntactic attributes for a sample sentence

Writing/Reading Direction ←			
Word / **Attributes**	**الأثْوابَ**	**فاطِمَةَ**	**وغَسَلَت**
Aspect	N/A	N/A	Perfect
Case	Accusative	Nominative	N/A
State	DET	NOUN_PROP	N/A
Gender	M	F	F
Lemma	ثَوْب (vawob /cloth)	فاطِمَة (fATimap / Fatima)	غَسَل (ghasal /to wash)
Mood	N/A	N/A	N/A
Number	P	S	S
Person	N/A	N/A	1
POS	NOUN	NOUN_PROP	VERB
Stem	أثْواب (>vowAb)	فاطِمَة(fATimap)	غَسَلَت (ghasalat)
Translation	clothes	Fatima	washed
Voice	N/A	N/A	Active
CONJ	N	N	Y
PREP	N	N	N
PART	N	N	N
PRON	N	N	N

An application on a sample yields the values of these attributes for the words of the sentence " وغَسَلَت فاطِمَة الأثْوابَ" (wgasalat fATimap Al>vowAba /And Fatima washed the clothes). In this sentence, "وغَسَلَت" (wgasalat /and she washed) is a perfect verb conjugated in the first feminine person (هِيَ/hiy /she) and contains a conjunction (و/w/and, CONJ=Y)

4. CAMS-KG construction and mining

Based on text mining techniques, we capture the morpho-semantic structure of a word into a more formal structure which can be represented as a knowledge graph. According to Soudani et al. (2016) [70] and Chen (1999) [19], the integration of such resources into IR systems allows to overcome the so-called vocabulary mismatch in IR and to uncover with statistical techniques the latent semantic structure of documents that is often obscured by words chosen in a retrieval process [19, 70].

We propose a framework allowing a deep analysis of linguistic units of a corpus based on a hierarchical text mining process. Starting from the highest level of the corpus (i.e. the document) and ending to the less granular layer (i.e. the morphemes). Figure 1 shows an abstract architecture of our system for CAMS-KG construction and mining.

4.1. Morphological knowledge extraction

We start by studying related works in the field of Arabic NLP to choose the tools which will be integrated in our system.

4.1.1. Existent tools

Arabic morphological analyzers and stemmers differ in terms of output. For instance, Khoja stemmer [82] extracts the root of each word, while Larkey et al. [49, 50] and Darwish [41] retrieve the light stem. However, Ghwanmeh et al. [81] stemmer extracts the stem, the root and the verbed pattern. Outputs produced by these tools are non-vocalized whether input words are vocalized or not. Moreover, these tools do not perform a complete morphological analysis and are unable to recognize morpho-syntactic attributes. BAMA and its successor SAMA [22, 87] return the stems and the lemmas only. Alkhalil 1.0 returns both the root and the stem, but does not recognize the lemma. However, AlKhalil 2.0 is an enhanced version of Alkhalil 1.0 fetching the lemmas [57].

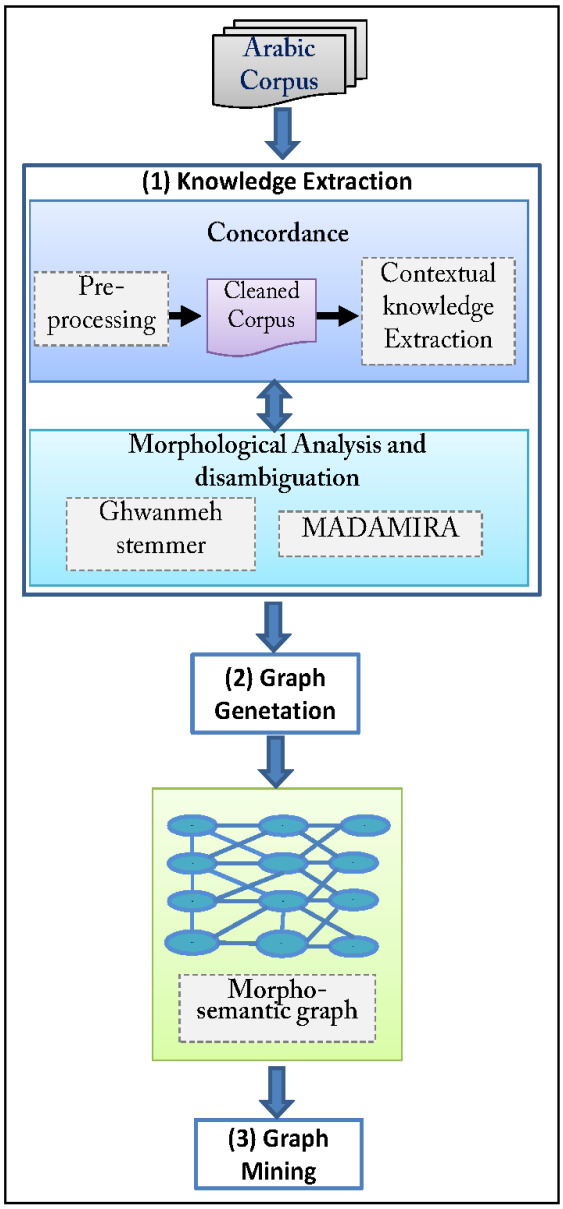

Fig 1. CAMS-KG construction and mining process.

In all these tools, the words are analyzed out of context and for a unique word, all the possible solutions are returned. Hence, a disambiguation step is required to retain only one solution from the suggested list. MADAMIRA [11] is a state-of-the art tool which combines SAMA [22], MADA [68] and AMIRA [59] to identify the correct morphological solution based on contextual information. Besides, morpho-syntactic attributes like POS, gender and number are returned. Recently, Darwish et al. [45] proposed QATARA which is a stemmer based on stem templates. It allows to tag words with POS gender and number. In another work [46], Darwish et al. proposed a novel segmenter called FARASA which performs tokenization, stemming, diacritization, named entity recognition and POS tagging.

Our system encompasses several modules which will be detailed in the following sections. We mainly focus on tools used to process corpora at the morphological and semantic levels (cf. Section 4.1 and 4.2). Besides, CAMS-KG is minded compute similarities between terms and enhance IR (cf. Section 4.3).

4.1.2. Comparative study

We compare existent Arabic NLP tools in terms of their ability to extract morphological knowledge and serve text mining applications. For stemming, recent comparative studies carried by Al-Kabi et al. (2011) [53] and Hadni et al. (2012) [62] show that Ghwanmeh stemmer [81] achieved the best results. This tool was also evaluated by Ben Guirat et al. (2016) [77].

However, even if some works (e.g. [77]) compared different tools and indexing units, lemma-based IR is not yet well investigated except in [2] and [70]. That's why, other studies proposed to get a morphology-based IR in order to reach better results. For instance, Roussey et al. (2010) used BAMA with no disambiguation utility [21]. Through their experiments, they compared their approach with light10 [49]. Findings reveal that morphological analysis improves considerably Mean Average Precision (MAP) and Precision at N (P@N) for short queries. Nevertheless, light10 reaches better results for long queries. This is explained by the fact that although morphological analysis engenders more ambiguities with long queries, light 10 convenes for long texts.

In deep morphology-based IR, some works [43] reveal that both Sebawi and IBM-LM analyzers enhance IR results better than light stemming. However, IBM-LM outperforms Sebawi. This may be explained by the fact that the former studies the correlation between morphemes by the use of language models and Prefix-Suffix filters. What's more, IBM-LM segments compound affixes like "والـ" (wAl; and the) and in contrast Sebawi does not. Additionally, a contextual approach to handle unknown stems is implemented by IBM-LM.

Authors in [40] used AMIRA [58] that exerts context to segment Arabic words and improves IR effectiveness. Soudani et al. (2016) [70] concluded that MADAMIRA outperforms all the other NLP tools essentially the stemmers of Khoja and Garside (1999) [82], Ghwanmeh et al. (2009) [75], Fraser et al., (2002) [8] and Darwish et al. (2009) [42] stemmers.

From previous cited works [77], we glean that Ghwanmeh stemmer achieves the best results compared to other light or root-based stemmers. Moreover, the experiments by Sebawi, IBM-LM and AMIRA report contradictory results. However, they reveal that the context or the statistical filtering of morphological solutions improves results. Nevertheless, most of these works use poor and insufficient morphological knowledge about the extracted lexical units. A host of them do not yield the morpho-syntactic attributes of the morphemes. They also remove short vowels, thus inducing ambiguity. Besides, only few works studied the effect of lemma on IR compared to the other indexing units as detailed by Soudani et al. (2016) [70]. Their results uncover that the use of a morphological disambiguation tool (i.e. MADAMIRA) and lemma-based indexing enhances remarkably and in an optimal way the IR system performance compared to other tools.

The work of Abdelali et al. (2016) [2] shows that FARASA have approximately an equal performance compared with MADAMIRA but both tools are better than QATARA. Nevertheless, FARASA outperforms MADAMIRA and QATARA in terms of Speed. According to our tests in classical Arabic text processing, MADAMIRA performs a better segmentation than FARASA. Indeed, the latter was trained only on modern Arabic texts. Besides, it returns non-vocalized lemmas, while the output of MADAMIRA is fully vocalized. For example, the expression "اللغة العربية الفصحى" (Allgp AlErbyp AlfSHY; Standard Arabic) is lemmatized by MADAMIRA as follows: "لُغَة عَرَبِيّ فُصْحَى", that is recognizing three vocalized lemmas. FARASA generates the following non-vocalized output for the same expression: "لغة عربي فصحى". Besides, MADAMIRA returns much morpho-syntactic attributes (cf. table 1) compared to FARASA which handles only POS, gender and number.

4.1.3. MorphToolKit

From our comparative study, we may conclude that MADAMIRA stands out by its ability to produce high quality vocalized output and to recognize morpho-syntactic attributes. Even if it is criticized in terms of speed [2], we may tolerate this limit as our goal is to produce high quality knowledge graphs. Once these graphs are built, they may be used without rerunning MADMAIRA.

Nevertheless, MADAMIRA does not return the root nor the verbed pattern. That's why we integrate Ghwanmeh stemmer [81] as related works show its contribution compared to other stemmers [77, 62]. That is, we developed a tool called MorphToolKit which combines MADAMIRA and Ghwanmeh stemmer. From the XML output of MADAMIRA, we compute the morpho-syntactic attributes detailed in table 1. Then, words are processed with Ghwanmeh stemmer to extract roots and verbed patterns.

Furthermore, we build our resource from vocalized corpora. That is, we optimize the output of MADAMIRA by filtering the morphological solutions of a given word using its original short diacritics, which reduces ambiguities.

4.2. Semantic knowledge extraction

To study semantic dependencies between Arabic tokens, we are based on contextual information extracted by our concordancer (cf. section 4.2.1). Then graphs are generated and compressed as detailed in section 4.2.2

4.2.1. Concordancer

The goal of a concordancer is to extract the terms of a corpus with their corresponding contexts and frequencies. Some Arabic existing concordancers as the open source tool aConCorde [12] do not eliminate non Arabic characters from Arabic words and numerical characters are often concatenated with Arabic letters. Besides, aConCorde does not consider punctuation when identifying contexts. For example, in several contexts, words from the previous sentence or the next one are included in the current context.

To avoid such errors, segment a corpus and represent contextual knowledge, our tool performs as follows. A cleaning process is performed by eliminating non-Arabic characters and separating words from glued punctuation signs. Next, tokenization is done by cutting up paragraphs in each file into sentences. After that, sentences are segmented into words. All information about files, sentences and words are stored. For each

file, the sentences are stored with their order, frequencies, begin and end positions. For each word, we save the position in the sentence, the frequency, and the type (Arabic word, number, punctuation, etc.). Then, these information are enriched by the morphological knowledge captured by MorphToolKit.

4.2.2. Graph construction and compression

The contextual and morphological knowledge generated by the concordancer and MorphToolKit is transformed into a graph representation to allow inferring new relations between tokens and computing similarity measures.

From CAMS-KG, it is possible to extract a whole view of all the tokens types in the same graph, thus considering both morphological and co-occurrence links or to represent only some specific types of tokens and/or edges. In this step, we can filter tokens by any of morpho-syntactic attribute (e.g. removing empty words by a POS filter).

For graph representation, processing and mining, we exploit the WebGraph framework [73], which includes a set of algorithms to compress, store and manage large graphs with lazy methods [73]. In the case of a co-occurrence, we store in each edge the number of co-occurrence of the corresponding tokens. In the same manner, we represent morphological relations between tokens of different types (e.g. lemmas and root). In this case, the binary values are stored in edges (i.e. 1 if there is a morphological relation and 0 elsewhere).

4.3. Similarity calculus

The goal of this tool is to compute the similarity between the nodes of a graph whatever are their types. In the case of co-occurrence graphs, this calculus allows computing the similarity between terms; that is proposing tokens which are similar to an initial set of nodes.

Similarity measures are used for different purposes in Arabic NLP and IR; including Word Sense Disambiguation (WSD), synonym extraction, semantic indexing and query expansion [14, 79, 83]. In all these tasks, several measures are used to evaluate the similarity of words in context. We may cite the theorem of Bayes [79], Rocchio algorithm [83] and LESK [14, 55], Pointwise Mutual Information (PMI) [36], LLR, Dice Factor and T-Score [35]. In this paper, we propose to use Okapi BM25 [14, 39, 84] which is used in several works for term-term or document-document similarity and query expansion [75, 88, 30, 71]. This measure stands out by the following aspects. On the first hand, it allows both one-to-one and one-to-many associations. On the second hand, The INF factor (cf. formula 3) allows to evaluate the discriminative power of terms. That is terms which co-occur with many other terms are penalized. On the last hand, BM25 has three parameters which may be tuned to enhance results (cf. formula 2).

We formalize similarity calculus in graphs as follows. Let *C* be a set of input weighted nodes (n_i, w_i) provided as constraints:

$$C = \{(n_1, w_1), \dots, (n_m, w_m)\} \quad (1)$$

where n_i is the node number *i* and w_i is its weight.

Given that, we would like to mine the graph to retrieve all the similar nodes to the set of constraints *C*. The result is also a set of weighted nodes having the same structure as *C*. Weights are computed by Okapi BM25 [47] which showed to be efficient in several IR studies and campaigns [84-80, 88-84]. That is, we follow other research works which used IR models to compute similarities between queries and terms [16, 17]. Indeed, BM25 is not only a retrieval model used to compute query-document similarity, but it is also used to compute term-term or document-document similarities [30, 75, 88]. Our tool implements BM25 to compute term similarity in co-occurrence graphs.

Given the set of constraints *C*, we evaluate a candidate node n_c as given by formula (2).

$$\mathrm{BM25}(\mathrm{C}, \mathrm{n_c}) = \sum_{i=1}^{m} INF(n_i) * \frac{(k1+1)*e(n_i,n_c)}{k1((1-b)+b*\left(\frac{sumin(n_c)}{avgsumin}\right)+e(n_i,n_c)} * \frac{(k3+1)*w_i}{K3+w_i} \quad (2)$$

Where $INF(n_i)$ is the Inverse Node Frequency of the node n_i and it is analogue to Inverse Document Frequency (*IDF*). It is given by:

$$INF(n_i) = log\frac{N-out(n_i)+0.5}{out(n_i)+0.5} \quad (3)$$

K1, *b* and *K3* are the usual Okapi BM25 parameters, taking by default the values 1.2, 0.75 and 8.0 respectively. Besides, these formulae use several parameters from the graph:

- $e(n_i, n_c)$: the weight of the edge whose source is the node n_i and destination is n_c. This is analogue to the frequency of a term in a document.
- $out(n_i)$: the out-degree of the node n_i, that is the number of nodes in the graph having n_i as desti-

nation. In term-document graphs, this parameter is equivalent to the number of documents a given term (n_i) appears in.

- $sumin(n_i)$: the sum of the weights of the edges having n_i as destination. This is equivalent to the document length parameter in query-document matching.
- *avgsumin*: the average of the previous parameter (i.e. $\text{sumin}(n_i)$) over all the possible destinations in the graph. This is equal to the average document length in the classical version of Okapi BM25.
- *N*: the number of all the possible destinations in the graph, which is analogue to the number of documents in the collection.

With this generic implementation and given that our morpho-semantic resource is represented by means of graphs, we experiment several scenarios of graph-based Arabic IR, which are detailed in the following section.

5. CAMS-KG evaluation

We evaluate CAMS-KG in an objective and automatic manner in the context of Arabic IR and QE. For full surveys of both fields, we refer the readers to the works [15], [37] and [38]. In our case, as edges are typed and weighted by co-occurrence frequencies, we can combine several types of relations and apply similarity measures to propose terms to add to queries. The richness of the graph allows assessing the different combinations of morphology and/or semantic-based expansion (cf. section 5.2). Indeed, existent approaches of Arabic QE do not combine nor compare morpho-semantic knowledge for query expansion. In fact, it is worth noting that the impact of morphology on the semantic closeness of Arabic words has not been well examined. A marriage of morphology and semantics can be considered and applied either at the query formulation and reformulation steps or the indexing step. For instance, an important research question could be poring over the impact of morphological relations (e.g. the relation between lemmas derived from the same root) in selecting the closest candidate words to the initial query.

Besides, the fact that CAMS-KG is vocalized helps to morphologically and semantically disambiguate words. Our work allows lemma-based IR while related works used mainly roots, stems and light stems [15].

5.1. Data collections

Two principle corpora are used for experiments; namely Tashkeela [89] and ZAD [39]. The former is used to build CAMS-KG, while the latter is used as a standard IR test collection. The choice of Tashkeela is justified by the fact that it is a high coverage vocalized corpus collected from the Web, as we want to build our morpho-semantic resources from a huge corpus. To the best of our knowledge and based on surveys of existent Arabic corpora [7, 90], it is the most suitable corpus. Moreover, IR experiments are carried on the ZAD test collection which is of the same type of Tashkeela and it is semi-vocalized. Both Tashkeela and ZAD are classified as classical Arabic corpora.

Tables 3 and 4 provide general statistics about Tashkeela and ZAD respectively.

Table 3

General Statistics about Tashkeela

# books	# lines	# words	#sentences
84 books	1679609	73.786.005	1604510

Table 4

General Statistics about ZAD

#queries	Average query size	#Documents	#Relevant documents per query
25	5.5 words	2730	([0,72])

Detailed statistics about the two corpora are provided by table 5. A morphological analysis with MorphToolKit (Ghwanmeh stemmer and optimized MADAMIRA) (cf. 4.1.3) is applied to extract roots, stems and lemmas. We compute the number of vocalized and non-vocalized words based on MADAMIRA output. The number of distinct lemmas and stems is calculated for the four main categories of words (e.g. nouns, proper nouns, verbs and adjectives).

5.2. Experimental scenarios

Both queries and documents are processed by our text mining tool suite. However, for the IR tasks, we filter tokens by POS to remove stop-words. Only nouns, adjectives, verbs and proper nouns are considered.

Table 5

Statistics about the retrieved tokens from Tashkeela and ZAD

Corpus / Lexical unit	Tashkeela	ZAD
Words		
Vocalized	857305	71244
Non-vocalized	036609	37309
Roots		
Roots	11879	8556
Stems		
Nouns	40911	11754
Proper Nouns	02028	00702
Verbs	55356	10481
Adjectives	13767	02398
Total	**112062**	**25335**
Lemmas		
Nouns	14853	07192
Proper Nouns	01763	00657
Verbs	07888	03719
Adjectives	03602	01341
Total	**28106**	**12909**

As graph-based representation allows several ways of mining and returning relevant documents, we define some experimental scenarios which allow us to tune some important parameters in Arabic IR and evaluate the impact of CAMS-KG. Indeed, we compare four approaches as follows.

5.2.1. Baseline approach

This approach matches documents to queries without any expansion. However, we evaluate the impact of the indexing unit on the performance of the IR system. This approach will be referenced by ***BM25 (unit)*** as our ranking model is Okapi BM25. The parameter "unit" may be replaced by any token type (i.e. ***BM25 (lemma)*** means using the Okapi BM25 ranking model with lemma-based indexing.

5.2.2. Morphological expansion

In this proposal, we expand a given query by units having some morphological links with its initial query terms. That is, ***BM25 (stem, lemma)*** means using BM25 with stem-based indexing and expanding the initial query by all the stems having the same lemmas in CAMS-KG. This will allow using several morphological relations between stems sharing the same lemmas, words sharing the same stem, etc.

5.2.3. Semantic expansion

This approach consists in adding to an initial query composed of *m* terms, the most *m* similar nodes in the graph. We recall that the added terms are similar to the entire initial query, thus taking into account its whole sense and the contextual relationships between tokens. For instance, this is performed by applying the BM25 similarity measure in a co-occurrence graph. However, the latter may be extracted from the test collection (i.e. ZAD) or CAMS-KG. That is, ***BM25 (lemma, co-oc [ZAD])*** means applying semantic expansion in a lemma-lemma co-occurrence graph extracted from ZAD. The expanded query includes the original terms for which we assign a weight equal to 1. The added terms are weighted according to a normalized similarity score computed by Okapi BM25.

5.2.4. Morpho-Semantic expansion

This approach combines both morphological and semantic expansion. That is, the nodes added to the initial query are similar to the initial terms at the morphological (i.e. having morphological relation) and semantic (i.e. having a high similarity score in the graph) levels. For example, ***BM25 (lemma, root+co-oc [CAMS])*** stands for adding to the initial query, the lemmas which share the same roots with them. The candidate lemmas are weighted according to BM25 in CAMS-KG. If the set of the lemmas having the same roots is empty, the initial query is expanded by other lemmas (i.e. applying only semantic expansion).

5.3. Experimental results

In the following sub-section, we discuss the results of the different indexing and query expansion scenarios.

5.3.1. Morphology-based indexing and expansion

. Different indexing and morphological expansion approaches are tested. They consist in three baseline approaches (i.e. ***BM25 (lemma)***, ***BM25 (stem)***, ***BM25 (vocalized_word)***) and three morphology-based expansion approaches (i.e. ***BM25 (stem, lemma)***, ***BM25 (vocalized_word, lemma)***, ***BM25 (vocalized_word, stem)***).The values of the diverse standard IR metrics, namely MAP (Mean-Average Precision), Recall, F-score, and precision at several values of rank position (5, 10 15 and 20) are reported by Table 6. Fig. 2 draws Recall-Precision rates curves of these approaches.

Table 6

Experimental results for morphology-based indexing and expansion

Approach \ Measure		MAP	Recall	F-Score	P@5	P@10	P@15	P@20
Baseline	BM25 (vocalized word)	0.302	0.505	0.3779	0.456	0.312	0.274	0.240
	BM25 (stem)	0.288	0.548	0.3775	0.440	0.328	0.263	0.220
	BM25 (lemma, MADAMIRA)	0.459	0.674	0.5461	0.688	0.528	0.424	0.352
	BM25 (lemma, MorphToolKit)	**0.524**	**0.74**	**0.614**	**0.664**	**0.500**	**0.399**	**0.338**
Morphological Expansion	BM25 (vocalized word, stem)	0.193	0.631	0.2955	0.240	0.196	0.154	0.162
	BM25 (vocalized word, lemma)	**0.249**	**0.694**	**0.3665**	**0.352**	**0.268**	0.210	0.178
	BM25 (stem, lemma)	0.202	0.640	0.307	0.272	0.244	**0.229**	**0.194**

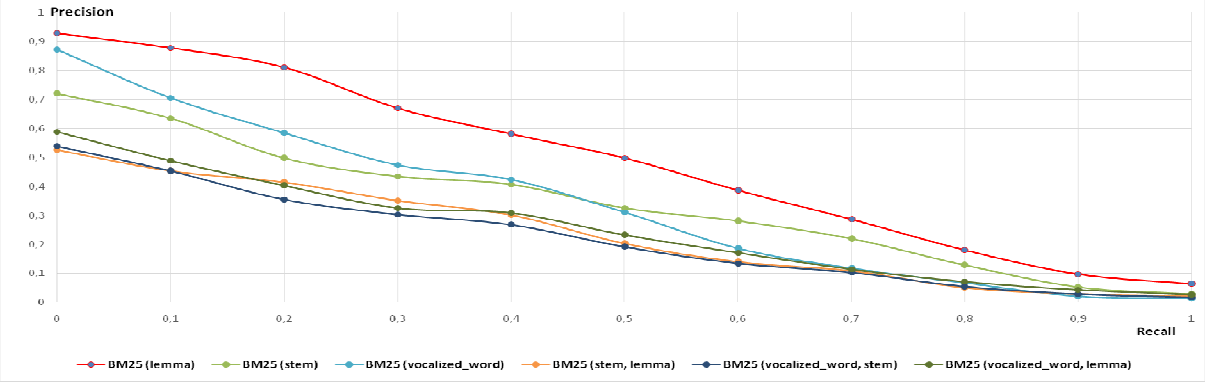

Fig. 2. Recall/Precision curves of morphology-based retrieval.

From these results, we observe that lemma-based indexing outperforms all the other approaches. A MAP equal to 45.9% is yielded by using such unit for indexing compared respectively to 28.8% and 30.2% for stem-based indexing ***BM25 (stem)*** and vocalized word-based indexing ***BM25 (vocalized word)***. Hence, enhancements of +17.1% and +15.7% are recorded by indexing with lemmas. We argument this by the fact that lemma is the most canonical unit which expresses formally and rigidly the semantics and reduces morphological ambiguity better than the stem or the surface word. By using MorphToolKit, results raise further to reach a MAP of 52.4% with an improvement rate of about +14.16% compared to MADAMIRA.

Moreover, these results show that indexing by surface words yields better results than stem-based indexing. This is confirmed in QE as ***BM25 (vocalized word, lemma)*** provides better results than ***BM25 (stem, lemma)***. Nevertheless, the fact that the former approach performs better than the latter is explained by the fact that it retrieves more words from CAMS-KG.

5.3.2. Semantic and morpho-semantic expansion

As previous results which showed that the lemma-based indexing enhances IR, the remaining experiments are performed with lemma-based indexing. This implies varying two parameters. On the one hand, we may extract co-occurrence graphs from CAMS-KG or build them from ZAD. On the other hand, as we are in the lemma level, morphological expansion may be based on roots (adding to the original query all the lemmas having the same roots) or on verbed patterns. In other words, we have six possible combinations according to the resources used for expansion (CAMS-KG or ZAD) and the expansion strategy (semantic only, root-based morpho-semantic expansion and verbed pattern-based morpho-semantic expansion) (cf. table 7). Table 8 provides the values of standard IR metrics, while figure 3 draws recall-precision curves.

Table 7

Semantic and morpho-semantic expansion approaches.

Approach	**Expansion**	**Knowledge source**	**Label**
Baseline (no expansion)	None	None	*BM25(lemma)*
Semantic Expansion	Co-occurrence	Tashkeela	*BM25 (lemma, co-oc [CAMS])*
		ZAD	*BM25 (lemma, co-oc [ZAD])*
Root-based morpho-semantic Expansion	Root + Co-occurrence	Tashkeela	*BM25 (lemma, root + co-oc [CAMS])*
		ZAD	*BM25 (lemma, root + co-oc [ZAD])*
Verbed pattern-based morpho- semantic Expansion	Verbed pattern + Co-occurrence	Tashkeela	*BM25 (lemma, verbed pattern + co-oc [CAMS])*
		ZAD	*BM25 (lemma, verbed pattern + co-oc [ZAD])*

Table 8

Experimental results for semantic and morpho-semantic expansion approaches

Measure / Approach	**MAP**	**Recall**	**F-measure**	**P@5**	**P@10**	**P@15**	**P@20**
BM25 (lemma)	0.524	0.743	0.614	0.66	0.5	0.399	0.338
BM25 (lemma, co-oc [ZAD])	0.524	0.804 (+8.20%)	0.629 (+2.44%)	0.64 (-3.03%)	0.5	0.399	0.338
BM25 (lemma, root+co-oc [ZAD])	0.515 (-1.72%)	0.832 (+12%)	0.636 (+3.85%)	0.64 (-3.03%)	0.5	0.399	0.338
BM25 (lemma, VP+co-oc [ZAD])	0.535 (+2.10%)	0.831 (+11.84%)	0.650 (+5.86%)	0.67 (+1.52%)	0.5	0.397 (-00.50%)	0.334 (-1.18%)
BM25 (lemma, co-oc [CAMS])	0.531 (+1.34%)	0.773 (+4.03%)	0.634 (+3.25%)	0.67 (+1.52%)	0.53 (+6.00%)	0.439 (+10.00%)	0.36 (+6.51%)
BM25 (lemma, root+co-oc [CAMS])	0.534 (+1.91%)	0.832 (+12%)	0.650 (+5.86%)	0.65 (-1.52%)	0.52 (+4.00%)	0.426 (+6.77%)	0.354 (+4.73%)
BM25 (lemma, VP+co-oc [CAMS])	0.506 (-3.44%)	0.833 (+12.11%)	0.629 (+2.44%)	0.64 (-3.03%)	0.5	0.399	0.338

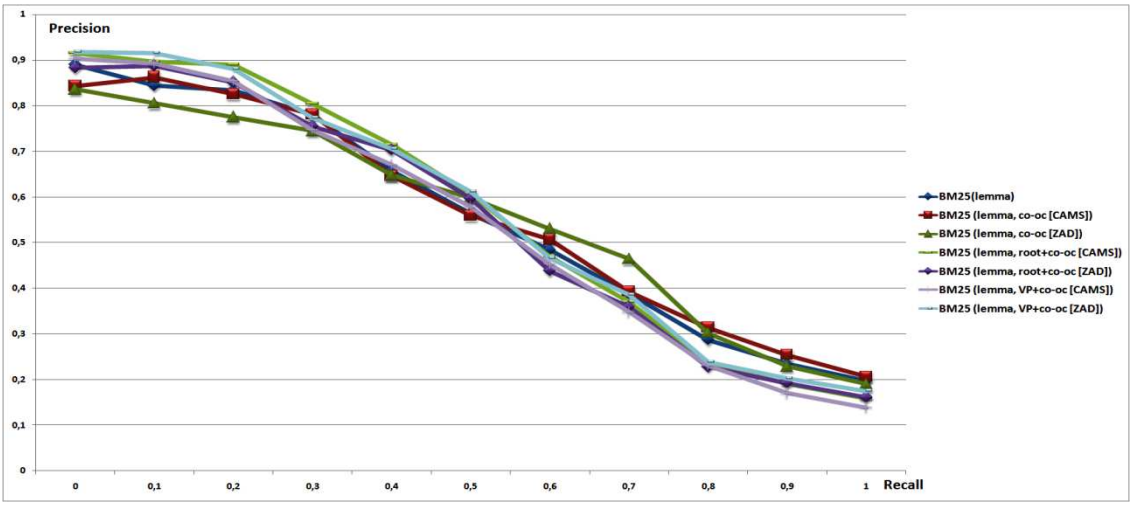

Fig. 3. Recall/Precision curves for semantic and morpho-semantic expansions.

Semantic expansion from ZAD gives better results in terms of recall and more relevant documents are retrieved, but CAMS-KG-based semantic expansion is better in terms of MAP. This implies that the lemmas added from ZAD trend to co-occur with the initial query terms in relevant documents. However, the lemmas added from CAMS-KG are semantically close to the original ones, thus decreasing the score of irrelevant documents. This shows the quality of CAMS-KG which seems to cover ZAD topics.

For the root-based morpho-semantic expansion in ZAD and CAMS-KG, both approaches increase the three IR metrics and CAMS-KG performs better than ZAD.

The rates in VP-based expansion confirm that expansion always improves recall for both ZAD and CAMS-KG. However, CAMS-KG-based expansion reaches better recall and worse MAP. In fact, recall rates for ZAD are lower than CAMS-KG, which may be explained by the specificity and the finest thematic granularity of the ZAD corpus. This leads to the fact that the retrieved documents are less than those returned by the CAMS-KG -based approach.

Indeed, this can be then justified by the fact that the Tashkeela vocabulary is bigger than which of ZAD. Therefore, the CAMS-KG is semantically and morphologically richer than ZAD. Thus, expansion performed on ZAD suggests less morphologically related lemmas and fewer documents are retrieved.

5.3.3. Global evaluation

Globally, semantic and morpho-semantic expansion enhance IR results compared with the baselines. Nevertheless, the performance of expansion and its impact on the IR process depend on the corpus type, topic and size as well as the query properties and the used tool for linguistic analysis.

Table 9 shows that all the semantic and morpho-semantic expansion approaches are statistically significant (p-value <0.05) compared to ZAD-based semantic expansion. This shows that the average of differences between the runs is significant.

According to the experiments, semantic expansion improves results in both ZAD and Tashkeela graphs in terms of recall and performs better than morphological expansion. ZAD-based semantic expansion slightly precedes Tashkeela-based one given that contexts in the ZAD collection are naturally semantically closer to the original queries.

Moreover, as observed, the VP-based morpho-semantic expansion based on Tashkeela gives the best recall rate (83.3%) among all the morpho-semantic approaches while the VP-based morpho-semantic expansion based on ZAD reaches the best MAP value (53.5%). Nevertheless, for the three types of expansion (i.e. semantic (1), root-based (2) and VP-based (3)), IR based on Tashkeela has the best MAP rates (53.1% (1), 53.4% (2)).

Fig. 4 highlights the precision values for the N first retrieved documents. It is clear that for the top first 5 documents CAMS-KG-based expansion especially the VP-based expansion outperforms ZAD-based expansion. Then, from the tenth retrieved documents, CAMS-KG root-based and semantic expansion have close performance. In general, we confirm the importance of the use of vocalized linguistic resources for expansion against a non vocalized one and its pivotal impact on the retrieval results.

Table 9

Test of significance for the different expansion approaches

	BM25(lemma, co-oc [ZAD]) vs BM25 (lemma, co-oc [CAMS])	BM25 (lemma, co-oc [ZAD]) vs BM25 (lemma, root+co-oc [ZAD])	BM25 (lemma, co-oc [ZAD]) vs BM25 (lemma, VP+co-oc [ZAD])	BM25 (lemma, co-oc [ZAD]) vs BM25 (lemma, root+co-oc [CAMS])	BM25 (lemma, co-oc [ZAD]) vs BM25 (lemma, VP+co-oc [CAMS])
Wilcoxon Signed Rank Test	0.000018	0.000024	0.000021	0.000021	0.000021
Test of Student	0.000002	0.000005	0.000004	0.000002	0.000003

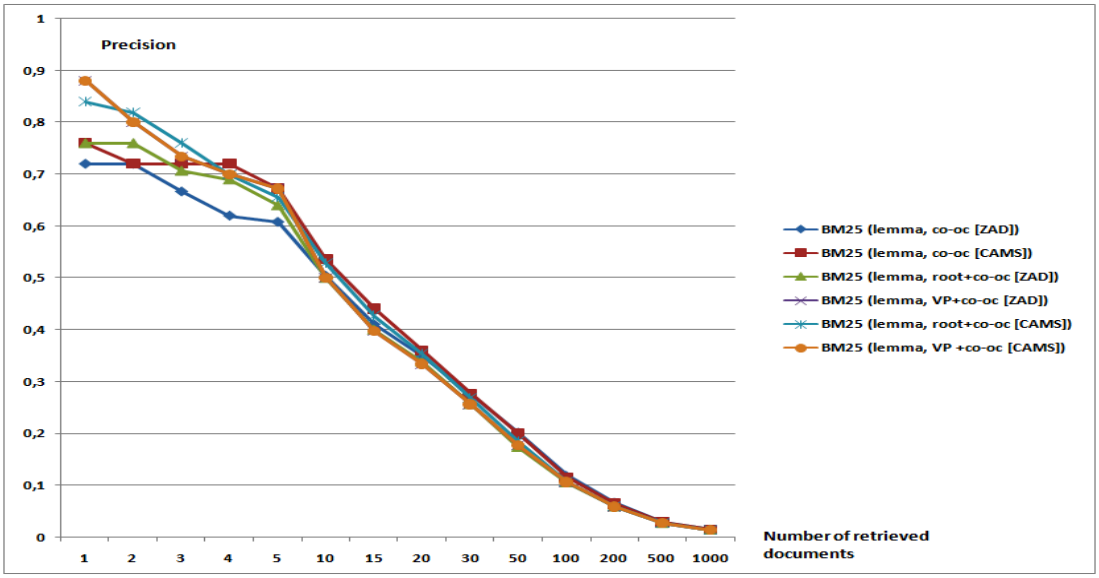

Fig. 4. Precision rates at the *N* first retrieved documents.

To better understand these results, we provide a detailed evaluation. Generally, ZAD queries are short which means that they are not semantically rich. Indeed, they contain many general words which are not domain relevant. For instance, the terms "أحكام (>HkAm /Provisions)" and " حُكم(Hukm/Provision)" exist in more than 30% of ZAD queries. Such frequent words do not add any semantic information relatively to the query domain. Thus, queries will be expanded with unrelated terms which have great co-occurrence values with such terms. On the other hand, several ZAD queries include many named entities such as proper nouns compared to the total size of the query. Morpho-semantic expansion of such words has no significance and adds to the original query terms which are not semantically related. For morpho-semantic expansion based on CAMS-KG, root-based approach reaches better results than the VP-based one, which is not the case for ZAD. This demonstrates that a root-based morpho-semantic expansion enhances the results in a more optimal way than the VP-based one does.

CAMS-KG-based expansion gives the best outcomes taking into consideration the fact that there are more probable contexts with more frequent morphologically related terms. As pointed out before, the VP-based morpho-semantic expansion yields better results in ZAD compared with CAMS-KG. This can be explained by the weak performance and ability of Ghwanmeh stemmer in identifying verbed patterns. Given the bigger size of CAMS-KG, the percentage of errors is more important. Therefore, semantic enrichment relied on this word form will be greatly poor.

Fig. 5 provides a meticulous experimental drawing on how average precision varies in a waved manner for the different expansion approaches for each query. This depends on the characteristics of queries and the nature of their terms.

Tables 10 and 11 detail respectively the average precision (AVGP) rates of the queries Q4 and Q6 for the different expansion strategies compared to the baseline. They also unfold the terms added to expand the initial queries, the number of retrieved documents (#RET) and the number of relevant retrieved documents (#RET_REL). We have chosen these queries given that they have distinctive features; namely they are short or contain proper nouns (e.g. بدر;bdr) or non-discriminant words (e.g. أحكام; provisions and أحداث; events). We notice that query number 4 (Q4, Table 10) has only two relevant documents and they are returned by all the approaches, while Q6 (Table 11) has 54 relevant documents.

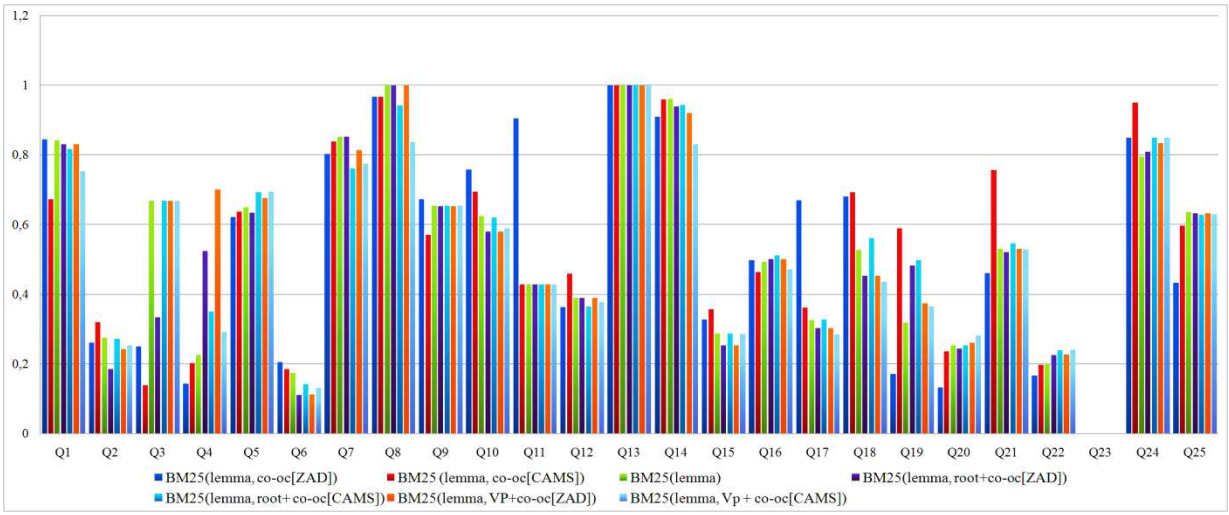

Fig. 5. Per query AVG precision rates for the expansion approaches.

Table 10

Results of query number 4 (Q4): الحلف(AlHlf/Swearing) بالطلاق(bAlTlAq/with divorce)

Approach	Query	AVGP	#RET
BM25 (lemma)	Q4 = حِلْف/NOUN (Hilof/Alliance) +طَلاق/NOUN (TalAq/ divorce)	0.22	211
BM25(lemma, co-oc [ZAD])	Q4 +حَنِث/VB (Haniv/to break one's oath) + حِنْث/NOUN (Hinov/breaking one's oath)	0.14	215
BM25 (lemma, co-oc [CAMS])	Q4 +ناشَد/VB (nA$ad/to adjure) + جاهِلِيَّة /NOUN (jAhiliy~ap/ pre-Islam)	0.20	248
BM25(lemma, root+co-oc[ZAD])	Q4 +طالِق/ADJ (Taliq/ divorced) +طَلَّق/VB (TAllaq/ to divorce)	0.52	232
BM25 (lemma, root+co-oc[CAMS])	Q4 +حالَف/VB (HAlaf/ to ally with) +حَلْف/NOUN (Halof / swearing)	0.35	212
BM25(lemma, VP+co-oc[ZAD])	Q4 +طالِق/ADJ (TAliq/divorced) +حَلَّف/VB (Hal~af/ to swear)	0.70	232
BM25(lemma, VP+co-oc[CAMS])	Q4 +حَلْف/NOUN (Halof/swearing) +حَلَّف/VB (Hal~af/to swear)	0.29	228

We remark that the analysis of the non vocalized word "الحلف" in this query returns the lemma حِلْف (Hilof), which is a noun meaning "Alliance" while the correct analysis is حَلْف/NOUN (Halof /swearing). Given that the input term is not vocalized, MorphToolKit retains the solution of MADAMIRA. That is, AVG Precision decreases compared to the baseline (0.22) when applying a semantic expansion on ZAD (0.14) or on CAMS-KG (0.20). However, they increase with morpho-semantic expansions (0.52, 0.35, 0.7 and 0.29) for root-based expansion on both corpora, but ZAD outperforms CAMS-KG. The decreased values are due to the fact that the added terms are not related to the original ones. For semantic expansion on ZAD, the added terms (e.g. حَنِث; "to break one's oath" and حِنْث; "breaking one's oath") are close to the query term "حَلْف" (swearing). However, they are not related to the query topic. Then,

many irrelevant documents are retrieved and they are ranked at the top results. The same case is with semantic expansion from CAMS-KG where the verb ناشَد(to adjure) and جاهِلِيَّة(pre-Islam) are added, because they frequently co-occur with the noun حِلْف (Hilof/Alliance).

Unlike semantic expansion, morpho-semantic enrichment remarkably enhances AVG Precision for both ZAD and CAMS-KG, as the added terms are relevant for Q4. That is, it achieves good rates especially for VP (0.7 compared to 0.52 for root) by which the added terms are related to both terms of the query. Furthermore, morpho-semantic expansion hinged on CAMS-KG, gives less precision values (0.35, 0.29) but greater than the baseline (0.22). For the root case, only one added term is semantically related to the query context (e.g. "حَلْف"; Halof; swearing), which is equal to the correct lemma of the first term of the query. Hence, morpho-semantic expansion adjusts the initial analysis returned by MADMAIRA. Succinctly, we remark that the erroneous lemma (i.e. حِلْف; Hilof; Alliance) influenced the expansion process in the case of CAMS-KG root-based expansion, where the verb " حالَف; HAlaf; to ally with)" was added.

In the following, we focus on Q6 (cf. table 11). This query is longer than Q4 (5 words vs. 2), but it contains many frequent words which are the terms أحداث(>HdAv/events) and أحكام(>HkA/provisions). The obtained rates show that semantic expansion outperforms all the other approaches with average precision reaching 0.20 and 0.18 respectively for ZAD and CAMS-KG. For the baseline, the proper noun بَدْر(bdr) has been analyzed and interpreted as a common noun making then a complete loss of the original meaning.

Table 11

Results of query number 6 (Q6): أحداث(>HdAv/events) وأحكام(w>HkA/ and provisions)
بدر (bdr; proper noun) الكبرى(AlkbrY/ the greatest) غزوة (gzwp)

Approach	Query	AVGP	#RET	#RET_REL
BM25(lemma)	Q6=حَدَث/NOUN)Hadav/event) +حُكْم/NOUN (Hukom/provision) +غَزْو/NOUN (gazow/invasion) +بَدْر/NOUN (bador/full moon) +كُبْرِي/NOUN (kobri/cabaret)	0.17	0699	40
BM25(lemma, co-oc [ZAD])	Q6 +حَقّ/PROPER_NOUN (Haq~) +حَرْب/NOUN (Harob/war) +حالِم/NOUN (HAlim/dreamy)	0.20	0766	44
BM25(lemma, co-oc [CAMS])	Q6 +بَنِي/PROPER_NOUN (baniy/sonsof) +نَاحِيَة/NOUN (nAHiyap/side) +خَنْدَق/NOUN (xanodaq/trench) +حَدِي/VB (Hadiy/to stay at)	0.18	0747	40
BM25 (lemma, root+co-oc[ZAD])	Q6 +حَدَث/VB (Hadav/tohappen) +حَدِيث/NOUN (Hadiyv/tale) +حاكِم/NOUN (HAkim/ruler) +حُدُوث/NOUN (Huduwv/happening)	0.11	1290	41
BM25 (lemma, root+co-oc[CAMS])	Q6 +حَدِيث/ADJ (Hadiyv/recent) +أَحْدَث/VB (>aHodav/to create) +حَدِيث/NOUN (Hadiyv/tale) +بادَر/VB (bAdar/to overtake)	0.14	1292	41
BM25 (lemma, VP+co-oc[ZAD])	Q6 +حَدِيث/ADJ (Hadiyv/recent) +حَدَث/VB (Hadav/tohappen) +حَدِيث/NOUN (Hadiyv/tale) +حُدُوث/NOUN (Huduwv/happening)	0.11	1301	41
BM25 (lemma, VP+co-oc[CAMS])	Q6 +حَدِيث/ADJ (Hadiyv/recent) +حَدَث/VB (Hadav/to happen) +حَدِيث/NOUN (Hadiyv/tale) +حَدَث/NOUN (Hadav/mishap)	0.13	1331	41

Moreover, AVG Precision decreases for the morpho-semantic enrichment, as these approaches return more documents. Moreover, the added terms for the case of morpho-semantic expansion for both ZAD and CAMS-KG are not relevant. The noun حَدِيث(Hadiyv/tale), the adjective حَدِيث (Hadiyv/recent) and the verbs أَحْدَثَ(>aHodav/to create) and بادَر(bAdar/to overtake) deviate the original semantics of the query. From all these examples, the gist of the experiments is that many parameters affect the IR results. On the one hand, analysis and disambiguation errors affect retrieval results. On the other hand, the shortness of queries and the nature of their terms are key elements influencing system performance. That is, the nature of terms and their frequencies and relationships in the training corpus are important factors.

5.4. Comparative study

In this paragraph we compare our results to related works which used the ZAD test collection; namely Ben Guirat et al. (2016) [77], Darwish and Oard (2002) [39] and Mahgoub et al. (2014) [13]. Ben Guirat et al. (2016) [77] implemented a hybrid indexing system using PL2 as ranking model combined with a variety of NLP tools; namely Larkey, Ghwanmeh and Khoja stemmers. They experimented and combined three indexing units (i.e. roots, verbed patterns and stems). Darwish and Oard explored the effectiveness of information retrieval with BM25, using different indexing units; namely (i) non vocalized words; (ii) stems returned by Sebawi; (iii) lightly stemmed words; and, (iv) Sebawi roots.

Mahgoub et al. (2014) [13] introduced a semantic approach for query expansion where two strategies are investigated. The first strategy (single query expansion) consists in adding terms which are similar to the whole query, while the second (multiple query expansion) expands each term separately.

Table 12 displays a comparative study between our IR results and those of related works [13, 39, 77]. These results show that our approaches outperform related works for all the metrics. Through all these works, we confirm that the indexing unit has a remarkable impact on IR performance. Thus, indexing documents by surface word forms (non vocalized words) reaches better results (MAP=0.45) [39] than using vocalized words (MAP=0.302). For stem-based indexing, Sebawi reaches better results (MAP=0.46) [39] than the Ghwanmeh stemmer (MAP=0.288).

Table 12

Experimental results of ZAD-based IR related works

	Approach	Recall	MAP	Precision	P@5	P@10
[77]	PL2(stem, Larkey)	0.350	0.290	0.280	0.43	0.275
	PL2(stem, Ghwanmeh)	0.420	0.370	0.330	0.52	0.362
	PL2(VP, Ghwanmeh)	0.410	0.330	0.320	0.53	0.366
	PL2(Root, Khoja)	0.290	0.180	0.190	0.38	0.258
	PL2(Root, Ghwanmeh)	0.490	0.320	0.310	0.55	0.358
	PL2 (Root+VP+stem, Ghwanmeh)	0.640	0.410	0.380	0.52	0.408
[39]	BM25(non vocalized word)	x	0.450	x	x	x
	BM25(stem, Sebawi)	x	0.460	x	x	x
	BM25(Light stem)	x	0.480	x	x	x
	BM25(Root, Sebawi)	x	0.440	x	x	x
[13]	SVM (stem, Light stem)	0.508	x	x	0.50	0.380
	SVM(stem, Light stem, single query expansion)	0.580	x	x	0.57	0.436
	SVM(stem, Light stem, multiple query expansion)	0.594	x	x	0.56	0.444
Our results	BM25(vocalized word, MADAMIRA)	0.505	0.302	0.325	0.45	0.312
	BM25(stem, MADAMIRA)	0.548	0.288	0.314	0.44	0.328
	BM25(lemma, MADAMIRA)	0.674	0.459	0.428	0.68	0.528
	BM25(lemma, MorphToolKit)	0.743	0,524	0.512	0.66	0.516
	BM25(lemma, co-oc [ZAD], MorphToolKit)	0.804	0,524	0.487	0.60	0.504
	BM25(lemma, co-oc [CAMS], MorphToolKit)	0.773	0,531	0.485	0.67	0.536
	BM25(lemma, root+co-oc[ZAD], MorphToolKit)	0.832	0,515	0.524	0.64	0.500
	BM25(lemma, root+co-oc[CAMS], MorphToolKit)	0.832	0,534	0.519	0.65	0.528
	BM25(lemma, VP+co-oc[ZAD], MorphToolKit)	0.831	0,535	0.521	0.67	0.500
	BM25(lemma, VP+co-oc[CAMS], MorphToolKit)	0.833	0,506	0.491	0.64	0.500

In the same vein, by applying a light stemming, the IR system effectiveness is improved (MAP=0.48) [39]. Then, Darwish and Oard submit that the probabilistic BM25 model improves strongly the IR results compared to the PL2 model [77] or the vector space model [70]. However, lemma outperforms all the other indexing units.

6. Conclusion

In this paper, we tried to contribute in building morpho-semantic classical Arabic resources. A host of efforts have been invested to analyze and disambiguate huge vocalized Arabic corpora. These corpora are transformed into rich graph-based representations storing all required morphological and semantic knowledge. To the best of our knowledge none of existent works tried to model the Arabic lexicon with all its levels (from roots to vocalized words) and morpho-syntactic attributes. To evaluate our resource, we assessed several scenarios of document indexing and query expansion. That is, we examined the impact of Arabic morphology on the search process by experimenting different indexing approaches of the test collection. Thus, we showed that the IR system behavior and performance depend outstandingly on this parameter both for indexing and query expansion. Lemma-based IR, which has not been well investigated in previous works, seems to contribute notably to the system effectiveness. Lemma gives noteworthy results whereby the IR system retrieves more relevant documents than the other indexing units. Furthermore, a comparative study with similar works using the same IR collection shows the contribution of CAMS-KG. Nevertheless, some worse results are recorded due to the relevance judgment of the test collection, which considers only documents containing all the query terms as relevant.

Although these encouraging results, several optimizations and improvements would be appended. On the first hand, we envisage to experiment our approaches on Kunuz which is a standard test collection being constructed from classical Arabic documents [72]. On the second hand, we plan to enlarge the coverage of our resource by mining Arabic vocalized sites on the Web. As our resource is represented through extensible and scalable compressed graphs, it is easy to integrate other corpora or lexical resources. On the third hand, we aim to integrate other semantic resources like LMF normalized dictionaries [69]. Consequently, synonyms, definitions and diverse contextual knowledge will be added and will permit graph-based word sense disambiguation, given that some Arabic lemmas are ambiguous at the semantic level. On the fourth hand MADAMIRA is slow, which is time consuming in processing large corpora and Ghwanmeh stemmer produces erroneous results for some words. We aim to contribute in the field by building a disambiguation module based on AlKhalil 2.0, which will allow us extract all the lexical entris in one swoop. Finally, the WebGraph framework [73] allows to experiment several graph mining algorithms. Some of these algorithms are already implemented in the framework, but it is possible to integrate other techniques of text mining. For instance word embeddings seems to contribute in Arabic IR [6] and the semantic spaces generated by such techniques may be transformed into graphs and integrated into our IRS.

Acknowledgements

We would like to thank Dr. Kareem Darwish for providing ZAD Test Collection.

References

[1] A. Abbache, F. Meziane, G. Belalem and F. Z. Belkredim, Arabic Query Expansion Using WordNet and Association Rules, International Journal of Intelligent Information Technologies, 12(3), pp.51-64, 2016.

[2] A. Abdelali, K. Darwish, N. Durrani and H. Mubarak, Farasa: A Fast and Furious Segmenter for Arabic, Conference Paper · June 2016, DOI: 10.18653/v1/N16-3003

[3] A. AbdelRaouf, C. A. Higgins, T. Pridmore and M. Khalil, Building a multi-modal Arabic corpus (MMAC), International Journal on Document Analysis and Recognition (IJDAR), 13(4) 285–302, 2010.

[4] A. Bakhouche and Y. Tlili-Guiassa, Meaning representation for automatic indexing of Arabic texts. International Journal of Computer Science Issues (IJCSI). 9(6), 173–178, 2012.

[5] A. Chen and F. Gey, Building an Arabic stemmer for information retrieval. Proceedings of the Text Retrieval Conference TREC-11, Maryland, USA, 13-16 November, 631-639, 2001.

[6] A. El Mahdaouy, S. O. El Alaoui and E. Gaussier, Semantically Enhanced Term Frequency based on Word Embeddings for Arabic Information Retrieval. The fourth IEEE International Colloquium on Information Science and Technology (CIST), Tangier-Assilah, Morocco, 24-26 October, 2016.

[7] A. Farghaly and K. Shaalan, Arabic natural language processing: challenges and solutions. ACM Transactions on Asian and Low-Resource Language Information Processing. 8(4), 1–22, 2009.

[8] A. Fraser, J. Xu and R. Weischedel, TREC 2002 Cross-lingual Retrieval at BBN. Proceedings of the Text Retrieval

Conference TREC-11, 13-16 November Maryland, USA, 2002.

[9] A. M. Al-Zoghby and K. Shaalan, Semantic Search for Arabic. Proceedings of the twenty-Eighth International Florida Artificial Intelligence Research Society (FLAIRS) Conference, Hollywood, Florida. USA, 18–20 May, 524-529, 2015.

[10] A. M. Al-Zoghby and K. Shaalan, Conceptual Search for Arabic Web Content. 16th International Conference on Intelligent Text Processing and Computational Linguistics, Cairo, Egypt, 14–20 April 405-416, 2015.

[11] A. Pasha, M. Al-badrashiny, M. Diab, A. Kholy, R. El Eskander, N. Habash, M. Pooleery, O. Rambow and R. M. Roth, MADAMIRA: a fast, comprehensive tool for morphological analysis and disambiguation of Arabic. 9th Language Resources and Evaluation Conference, Reykjavik, Iceland, 26-31 May, 1094–1101, 2014.

[12] A. Roberts, L. Al-Sulaiti and E. Atwel, aConCorde: Towards an Open-Source, Extendable Concordancer for Arabic. CORPORA journal. 1(1), 39-60, 2006.

[13] A. Y., Mahgoub, M. A. Rashwan, H. Raafat, M. A. Zahran and M. B. Fayek, Semantic Query Expansion for Arabic Information Retrieval. In EMNLP: The Arabic Natural Language Processing Workshop, Conference on Empirical Methods in Natural Language Processing, Doha, Qatar, 25 October, 87-92, 2014.

[14] A. Zouaghi, L. Merhbene and M. Zrigui, Combination of information retrieval methods with LESK algorithm for Arabic word sense disambiguation. Journal of Artificial Intelligence. 38, 257-269, 2012.

[15] B. Elayeb and I. Bounhas, Arabic Cross-Language Information Retrieval: A Review. ACM Transactions on Asian and Low-Resource Language Information Processing. 15(3), 44 pages, Article 18, 2015.

[16] B. Elayeb, I. Bounhas, O. Ben Khiroun and N. Ben Saoud, Combining Semantic Query Disambiguation and Expansion to Improve Intelligent Information Retrieval. in Duval B., van den Herik J., Loiseau S., and Filipe J. (eds.), ICAART 2014 Revised Selected papers, LNAI 8946, Springer International Publishing Switzerland, 280-295, 2015.

[17] B. Elayeb, I. Bounhas, O. Ben Khiroun, F. Evrard and N. Ben Saoud, A Comparative Study between Possibilistic and Probabilistic Approaches for Monolingual Word Sense Disambiguation. Knowledge and Information Systems. 44(1), 91-126, 2015.

[18] B. Hammo, A. Sleit and M. El-Haj, Enhancing Retrieval Effectiveness of Diacritisized Arabic Passages Using Stemmer and Thesaurus. 19th Midwest Artificial Intelligence And Cognitive Science Conference (Maics2008), Cincinnati, OH, USA, February 26, 189–196, 2008.

[19] C. Chen, Visualising Semantic Spaces and Author Co-Citation Networks in digital libraries. Information Processing and Management. 35(3), 401-420, 1999.

[20] C. Hirsch, J. Hosking and J. Grundy, Interactive Visualization Tools for Exploring the Semantic Graph of Large Knowledge Spaces. Proceedings of the IUI'09 Workshop on Visual Interfaces to the Social and the Semantic Web, 8 February, Sanibel Island, Florida, USA, 1-6, 2009.

[21] C. Roussey, F. Harrathi, L. Maisonnasse and S. Calabrett, Vers une approche statistique pour l'indexation sémantique des documents multilingues. INFORSID 2010: 127-141, 2010.

[22] D. Graff, M. Maamouri, B. Bouziri, S. Krouna, S. Kulick and T. Buckwalter, Standard Arabic Morphological Analyzer (SAMA) Version 3.1. Linguistic Data Consortium LDC2009E73, 2009.

[23] E. Mohamed and F. Sadat, Hybrid Arabic–French machine translation using syntactic re-ordering and morphological pre-processing, Computer Speech & Language Journal, 32 (1): 135-144, 2015.

[24] F. Baccar Ben Amar, B. Gargouri and A. Ben Hamadou, Generating core domain ontologies from normalized dictionaries. Engineering Applications of Artificial Intelligence. 51 (May 2016), 230–241, 2016.

[25] F. Diehl, M.J.F. Gales, M. Tomalin and P.C. Woodland, Morphological decomposition in Arabic ASR systems, Computer Speech & Language Journal, 26 (4): 229-243, 2012.

[26] F. Hieber, Translation-based Ranking in Cross-Language Information Retrieval, PhD thesis, Ruprecht-Karl- University, Heidelberg, Deutschland, 2014.

[27] F. Z. Belkredim and A. El-Sebai, An ontology based formalism for the Arabic language using verbs and their derivatives. Communications of the IBIMA. 11(5), 44-52, 2009.

[28] F. Z. Belkredim and F. Meziane, DEAR-ONTO:A DErivational ARabic Ontology Based on Verbs. International Journal of Computer Processing Of Languages. 21(3), 279–291, 2008.

[29] G. A. Abandah, A. Graves, B. Al-Shagoor, A. Arabiyat, F. Jamour and M. Al-Taee, Automatic diacritization of Arabic text using recurrent neural networks, International Journal on Document Analysis and Recognition (IJDAR), 18(2) 183–197, 2015.

[30] G. Maheshwari, P. Trivedi, H. Sahijwani, K. Jha, S. Dasgupta and J. Lehmann, SimDoc: Topic Sequence Alignment based Document Similarity Framework. arXiv preprint arXiv:1611.04822, 2016.

[31] H. Achour and M. Zouari, Multilingual learning objects indexing and retrieving based on ontologies. World Congress on IEEE 2013 of the Computer and Information Technology (WCCIT), Sousse, Tunisia, 22 - 24 June, 2013.

[32] H. Safi, M. Jaou and L. Hadrich Belguith, PIRAT: A Personalized Information Retrieval System in Arabic Texts Based on a Hybrid Representation of a User Profile. 21st International Conference on Applications of Natural Language to Information Systems (NLDB 2016), Salford, UK, 22-24 June, 326-334, 2016.

[33] I. Bounhas, R. Ayed, B. Elayeb, F. Evrard and N. Ben Saoud, A Hybrid Possibilistic Approach for Arabic Full Morphological Disambiguation. Data & Knowledge Engineering. 100(Part B), 240-254, 2015.

[34] I. Bounhas, R. Ayed, B. Elayeb, F. Evrard and N. Ben Saoud Experimenting a discriminative possibilistic classifier with reweighting model for Arabic morphological disambiguation. Computer Speech and Language Journal, 33(1): 67-87, 2015.

[35] I. Bounhas, B. Elayeb, F. Evrard and Y. Slimani, "ArabOnto: experimenting a new distributional approach for building Arabic ontological resources", International Journal of Metadata, Semantics and Ontologies (IJMSO), vol. 6, no. 2, pp. 81–95, 2011.

[36] J. Atwan, M. Mohd, H. Rashaideh and G. Kanaan, Semantically enhanced pseudo relevance feedback for Arabic information retrieval. Journal of Information Science. 42(2), 246-260, 2016.

[37] J. Atwan and M. Mohd, Arabic Query Expansion: A Review. Asian Journal of Information Technology, 16(10), 754-770. (2017).

[38] K. Darwish and W. Magdy. Arabic Information Retrieval. Foundations and Trends in Information Retrieval, 7(4), pp. 239–342, 2013. DOI: 10.1561/1500000031.

[39] K. Darwish and D. Oard, Term selection for searching printed Arabic. Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieva, 2002, New York, NY, USA l, August 11 – 15, 2002.

[40] K. Darwish and A. Ali, Arabic retrieval revisited: Morphological hole filling. 50th Annual Meeting of the Association for Computational Linguistics (ACL), Jeju Island, Korea, 8 - 14 July, 2012.
[41] K. Darwish, Building a shallow morphological analyzer in one day. ACL-2002 Workshop on Computational Approaches to Semitic Languages, 11 July, Pennsylvania, USA, 2002.
[42] K. Darwish, W. Arafa and M. I. Eldesouki, Stemming techniques of Arabic Language: Comparative Study from the Information Retrieval Perspective. The Egyptian Computer Journal. 36(1), 30-49, 2009.
[43] K. Darwish, H. Hassan and O. Emam, Examining the effect of improved context sensitive morphology on Arabic information retrieval. ACL-2005 Workshop on Computational Approaches to Semitic Languages, Michigan, USA, 29 June, 25-30, 2005.
[44] K. Darwish and D. Oard, Adapting morphology for Arabic information retrieval. In: Soudi, A., Van den Bosch, A., Neumann, G. (Eds.) Arabic Computational Morphology, Springer Verlag, 245–262, 2007.
[45] K. Darwish, A. Abdelali and H. Mubarak, Using stem-templates to improve Arabic pos and gender/ number tagging. In LREC-2014, 2014.
[46] K. Darwish and H. Mubarak, Farasa: A New Fast and Accurate Arabic Word Segmenter, 2014.
[47] K. Spärck Jones, S. Walker and S. E. Robertson, A probabilistic model of information retrieval: Development and comparative experiments: Part 1. Information Processing & Management. 36 (6), 779–808, 2000.
[48] L. Abouenour, K. Bouzouba and P. Rosso, An evaluated semantic query expansion and structure-based approach for enhancing Arabic question/answering. International Journal on Information and Communication Technologies. 3(3) 37-51, 2010.
[49] L. S. Larkey, L. Ballesteros and M. E. Connell, Light Stemming for Arabic Information Retrieval. In: Soudi, A., Van den Bosch, A., Neumann, G. (Eds.) Arabic Computational Morphology, Springer Verlag, pp. 221-243, 2007.
[50] L. S. Larkey, L. Ballesteros and M. E. Connell, Improving stemming for Arabic information retrieval: light stemming and co-occurrence analysis. In Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval, Tampere, Finland, 11 - 15 August, 275-282, 2002.
[51] L. S. Larkey and M. E. Connell, Arabic information retrieval at UMass in TREC-10. Text Retrieval Conference, 13-16 November, Maryland, USA, 2001.
[52] M. Aljlayl and O. Frieder, On Arabic search: improving the retrieval effectiveness via a light stemming approach. 11th Conference on Information and Knowledge Management, Washington, DC, USA, 04 - 09 November, 340-347, 2002.
[53] M. Al-Kabi, Q. Al-Radaideh and K. Akawi, Benchmarking and assessing the performance of Arabic Stemmers. Journal of Information Science. 37(2), 1–12, 2011.
[54] M. A. Abderrahim, M. E. A. Abderrahim and M. A. Chikh, Using Arabic WordNet for semantic indexation in information retrieval system. International Journal of Computer Science Issues. 10(1), 327–332, 2013.
[55] M. A. Abderrahim, M. Dib, M. E. Abderrahim and M. A. Chikh, Semantic indexing of Arabic texts for information retrieval system. International Journal of Speech Technology. 19(2), 229–236, 2016.
[56] M. Anizi and J. Dichy, Assessing word-form based Search for Information in Arabic: Towards a New Type of Lexical Resource, Second International Conference on Arabic Language Resources and Tools, Cairo, Egypt, 21-23 April, 2009.
[57] M. Boudchiche, A. Mazroui, M. Ould Abdallahi Ould Bebah, , A. Lakhouaja and A. Boudlal, AlKhalil Morpho Sys 2: A robust Arabic morpho-syntactic analyzer. Journal of King Saud University – Computer and Information Sciences. 29(2), 141-146, 2016. doi: http://dx.doi.org/10.1016/j.jksuci.2016.05.002
[58] M. Diab, Second generation tools (AMIRA 2.0): Fast and robust tokenization, POS tagging, and base phrase chunking. 2nd International Conference on Arabic Language Resources and Tools, 21-23 April, Cairo, Egypt, 2009.
[59] M. Diab, H. Kadri and J. Daniel, Automated methods for processing Arabic text: from tokenization to base phrase chunking. Workshop on Computational Approaches to Semitic Languages: Common Issues and Resources, Czech Republic, 28 June, Prague, 89-96, 2007.
[60] M. E. A. Abderrahim, Utilisation Des Ressources Externes Pour la Reformulation des Requêtes Dans un Système de Recherche d'Information. The Prague Bulletin of Mathematical Linguistics. 99(1), 85–97, 2013.
[61] M. Gasmi, Utilisation des ontologies pour l'indexation automatique des sites Web en Arabe. Mémoire de mastère, Université Kasdi Merbah Ouargla, Maroc, 2009.
[62] M. Hadni, A. Lachkar and S. Alaoui Ouatik, A New and Efficient Stemming Technique for Arabic Text Categorization. International Conference on Multimedia Computing and Systems, 10-12 May, Tangier, Morocco, 791 – 796, 2012.
[63] M. Hattab, B. Haddad, M. Yaseen, A. Duraidi, A. Abu Shmais and A. Addaall, Arabic search engine: Improving search based on combination of morphological analysis and generation considering semantic pattern". In Proceedings of the 2nd International Conference on Arabic Language Resources and Tools, Cairo, Egypt, 22-23 April, 159-162, 2009.
[64] M. Shaheen and A. M. Ezzeldin, Arabic Question Answering: Systems, Resources, Tools, and Future Trends. Arabian Journal for Science and Engineering. 39(6) 4541-4564, 2014.
[65] M. A. Sherif and A. C. N. Ngomo, Semantic Quran: A Multilingual Resource for Natural-Language Processing, Semantic Web Journal, 6(4): 339-345, 2015.
[66] N. Fareed, H. Mousa and A. Elsisi, Enhanced Semantic Arabic Question Answering System Based on Khoja Stemmer and AWN. Proceedings of the 9th International Conference on Computer Engineering, (ICENCO-2013), Cairo, Egypt, 28-29 December, 85-91, 2013.
[67] N. Habash, Introduction to Arabic Natural Language Processing. ISBN: 9781598297959, Ed. Graeme Hirst, University of Toronto, 2010.
[68] N. Habash, O. Rambow and R. Roth, MADA+ TOKAN: a toolkit for Arabic tokenization, diacritization, morphological disambiguation, POS tagging, stemming and lemmatization. Proc. Second International Confernce on Arabic Language Resources and Tools, Cairo, Egypt, 22-23 April, 102–109, 2009.
[69] N. Soudani, I. Bounhas, B. Elayeb and Y. Slimani, Toward an Arabic Ontology for Arabic Word Sense Disambiguation Based on Normalized Dictionaries. In R. Meersman et al. (Eds.), Proceedings of The 13th International Conference on Ontologies, DataBases, and Applications of Semantics (ODBASE), Springer-Verlag Berlin Heidelberg, LNCS 8842, Amantea, Italy, 27-30 October, 655-658, 2014.
[70] N. Soudani, I. Bounhas and Y., Slimani, Semantic Information Retrieval: A Comparative Experimental Study of NLP Tools and Language Resources for Arabic. 28th International Conference on Tools with Artificial Intelligence (ICTAI), San Jose, Canada, 06-08 November, 879-887, 2016.
[71] N. Soudani, I. Bounhas and Y. Slimani, "A hybrid approach for standardized Dictionary-based knowledge extraction for

Arabic morpho-semantic retrieval", in Proceedings of the 2nd IEEE International Workshop on Arabic and derived Script Analysis and Recognition (ASAR 2018), The Alan Turing Institute, London, United Kingdom, 2018.

[72] O. Ben Khiroun, R. Ayed, B. Elayeb, I. Bounhas, N. Ben Saoud and F. Evrard, Towards a New Standard Arabic Test Collection for Mono- and Cross-Language Information Retrieval. in Natural Language Processing and Information Systems - Proceedings of The 19th International Conference on Applications of Natural Language to Information Systems, NLDB'2014, Montpellier, France, June 18-20, 168–171, 2014.

[73] P. Boldi and S. Vigna, The WebGraph Framework I: Compression Techniques. Proceedings of WWW 2004, New York, USA, 17-22 May, 595-602, 2004.

[74] R. Mohamed and M. Ragab, Al-Bayan : A Knowledge-based System for Arabic Answer Selection, Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval-2015), Denver, Colorado, USA, 4-5 June, 2015.

[75] R. M. Alguliyev, R. M. Aliguliyev and N. R. Isazade, A New Similarity Measure and Mathematical Model for Text Summarization. Problems of Information Technology, 6(1), 42-53, 2015.

[76] R. Xie, Z. Liu and M. Sun, Representation Learning of Knowledge Graphs with Hierarchical Types, in Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-16), pp.2965-2971, 2016.

[77] S. Ben Guirat, I. Bounhas and Y. Slimani, Combining Indexing Units for Arabic Information Retrieval. International Journal of Software Innovation (IJSI). 4(4), 1-14, 2016.

[78] S. Elkateb, W. J. Black, P. Vossen, H. Rodriguez, A. Pease, M. Alkhalifa and F. Christiane, Building a WordNet for arabic. In The 5th Conference on Language Resources and Evaluation (LREC), 29-34, 24-26 May, Genoa, Italy, 2006.

[79] S. Elmougy, H. Taher and H. Noaman, Naïve Bayes Classifier for Arabic Word Sense Disambiguation. Proceedings of the 6th International Conference on Informatics and Systems, Cairo, Egypt, Mars 27 - 29, pp.16-21, 2008.

[80] S. E. Robertson, S. Walker and M. Hancock-Beaulieu, Okapi at TREC-7. Proceedings of the Seventh Text REtrieval Conference. Gaithersburg, USA, 9-11 November, 1998.

[81] S. Ghwanmeh, S. Rabab'ah, R. Al-Shalabi and G. Kanaan, Enhanced Algorithm for Extracting the Root of Arabic Words. Sixth International Conference on Computer Graphics, Imaging and Visualization, IEEE Computer Society, 11-14 August, Tianjin, China, 388-391, 2009.

[82] S. Khoja and S. Garside, Stemming Arabic Text. Technical report. Computing department, Lancaster University, U.K., 1999.
http://www.comp.lancs.ac.uk/computing/users/khoja/stemmer.ps.

[83] S. M. Eid, A. B. Al-Said, N. M. Wanas, M. A. Rashwan and N. H. Hegazy, Comparative Study of Rocchio Classifier Applied to supervised WSD Using Arabic Lexical Samples. Proceedings of the tenth conference of language engineering (SEOLEC'2010), Cairo, Egypt, 15-16 December, 2010.

[84] S. Robertson and H. Zaragoza, The Probabilistic Relevance Framework: BM25 and Beyond. Foundation of Trends of Information Retrieval. 3(4), 333–389, 2009.

[85] S. Yang, L. Zou, Z. Wang, J. Yan and J.R. Wen, Efficiently Answering Technical Questions- A Knowledge Graph Approach, in Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17), pp. 3111- 3118, 2017.

[86] S. Zhao and Y. Zhang, Tailor knowledge graph for query understanding: linking intent topics by propagation, in Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1070–1080, Doha, Qatar, October 25-29, 2014.

[87]T. Buckwalter, Arabic Morphological Analyzer Version 1.0. Linguistic Data Consortium. LDC2002L49, 2002.

[88] T. Elsayed, J. Lin and D. W. Oard, Pairwise document similarity in large collections with MapReduce. In Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies, Columbus, Ohio, USQ, June 16, 2008, Short Papers, pp.265-268, 2008.

[89] T. Zarrouki and A. Balla, Tashkeela: Novel corpus of Arabic vocalized texts, data for auto-diacritization systems. International journal Data in Brief. 11, 147–151, 2017.

[90] W. Zaghouani, Critical survey of the freely available arabic corpora. International Conference on Language Resources and Evaluation (LREC'2014), OSACT Workshop, Reykjavik, Iceland, 26-31 May, 1–8, 2014.

[91] Y. S. Lee, K. Papineni, S. Roukos, O. Emam and H. Hassan, Language model based Arabic word segmentation. 41st Annual Meeting of the Association for Computational Linguistics, Sapporo, Japan, 7 July, 399-406, 2003.