

Experimenting Machine-Learning Algorithms for Morphological Disambiguation of Arabic Texts

Bilel Elayeb^{1,2}^a, Mohamed Firas Ettih³^b and Raja Ayed^{2,4}^c

¹*Liwa College of Technology, P.O. Box 41009, Abu Dhabi, U.A.E.*

²*RIADI Research Laboratory, ENSI, Manouba University, Tunisia*

³*Université Paris-Est Créteil, Paris 12 Val de Marne, France*

⁴*Faculty of Economics and Management of Nabeul, Carthage University, Tunisia*


Keywords: Morphological Disambiguation, Arabic Text, Machine-Learning Algorithms, Data Transformation, Morphological Feature, Classification.


Abstract: Arabic language is characterized by its complexity and its morphological and orthographic variations including syntactic and semantic diversity of a word. This specificity may cause Arabic morphological ambiguity. We present in this paper a new architecture for morphological disambiguation of Arabic texts. The latter can be treated as a classification problem where the set of morphological features' values represent classes, and a classification algorithm is used to assign a class to each word's occurrence based on the context. The first step consists of identifying the correct morphological analysis of a non-vocalized Arabic word using the morphological dependencies extracted from the corpus of vocalized texts. Then, we propose a method of transforming imperfect training datasets into perfect data having precise attributes and certain classes. We experiment this architecture on a set of machine-learning classifiers using a corpus of classic Arabic texts. Results highlight some statistically significant improvement of SVM and Naïve Bayes classifiers in terms of disambiguation rate.


1 INTRODUCTION

Morphology is the field that studies how the smallest meaningful units, called morphemes, combine to form lemmas which are the autonomous units forming the language lexicon. The morphology systems are useful to support NLP tools such as analyzers and information retrieval systems (IRS). Morphological analysis only reveals the various potential of words' vowels in a text and interprets their structures. However, a morphological analyzer can be used to display all forms of verbs in Arabic. It can also display multiple forms if the user chooses to specify not only the root but other morphological attributes like gender, number, and mode. Morphological analyzer displays all possible values of these attributes. It can analyze any form of a given word with a certain coverage rate.

Numerous applications in the field of Arabic NLP must deal with the complex morphology of this language (Elayeb and Bounhas, 2016; Elayeb, 2019). Morphological analysis is an important step in automatic speech recognition, Arabic texts phonetization and automatic Arabic text summarization (Elayeb et al., 2020). Besides, information retrieval applications should index documents and extract relevant characteristics of their significant entities (Elayeb, 2009; Elayeb et al., 2009; Bounhas et al., 2011b). Indeed, Information Retrieval and Knowledge Extraction Systems (IRKES) require the recognition of useful entities in texts such as words, phrases, and concepts. The basic level concerns the word's structure, i.e. the morphological level. Indeed, a given word can have several morphological interpretations, which makes it ambiguous. For example, the word "أكل" can be interpreted as a noun meaning "food" (أَكْلٌ; >ak°luN)

^a <https://orcid.org/0000-0002-5050-2522>

^b <https://orcid.org/0000-0002-2237-7146>

^c <https://orcid.org/0000-0003-3339-8388>

and as a verb meaning "to eat" (أَكَلَ; >akala). This phenomenon represents a challenge for morphologically rich languages such as Arabic (Diab et al., 2004). Thus, a non-vocalized Arabic word can have more than 12 interpretations (Habash and Rambow, 2007; Habash et al., 2009b).

The information retrieval process begins with an analysis and pre-processing step that aims to index documents and extract their knowledge (Elayeb, 2009, 2018). Indexing is an essential phase, aiming at ensuring that documents and queries are represented by keywords. These terms are standardized forms obtained by a morphological analysis of words representative of documents and queries. Arabic language is characterized by its morphological and orthographic variations including syntactic and semantic diversity of a word. This morphological richness causes ambiguity and difficulty in identifying the appropriate analysis, which can affect the definition of indexing terms and change the queries' meanings. Adding short vowels to a word can decrease its ambiguity but does not eliminate it. Besides, morphological disambiguation is essential to facilitate and improve the IR indexing task.

Existing Arabic morphological disambiguation approaches mainly disambiguate the Part-Of-Speech (POS) feature of words and only cover a particular type of texts, namely modern Arabic texts. POS disambiguation consists of determining the grammatical category of a word in a particular context. It can also be considered as a classification problem where the set of POS values represent classes, and a classification method is used to assign a class to each occurrence of a word based on the context. One of the important steps in the disambiguation task is the suitable selection of the classification method. Supervised automatic classification methods were applied. They have used training techniques to learn a classifier from training sets annotated with the values of the POS class.

We experiment, analyze and compare in this paper a series of machine-learning algorithms in order to solve the problem of morphological disambiguation of Arabic texts using several morphological attributes. We test these algorithms on the Kunuz¹ test collection (Ben Khiroun et al., 2014; Ayed, 2017; Ayed et al., 2018ab) containing classical vocalized Arabic texts. This corpus contains Hadith texts assigned to the Prophet of Islam Mohammad (PBUH). It has been the subject of several research works due to its linguistic, semantic, social richness and its well-organized structure. However, the TREC

collections (2001 and 2002) of Arabic newspapers articles suffer from the absence of vowels in their texts, which could generate a certain word sense ambiguity and a difficulty in identifying its POS feature as well as its function in the sentence.

The remaining of this paper is organized as follows: existing works on morphological disambiguation of Arabic texts are summarized and discussed in Section 2. The main sources of morphological ambiguity are presented in Section 3. Section 4 presents the architecture of the proposed approach as well as the training and testing data preparation phases. Section 5 introduces the experimental results as well as a comparative study between the ML classifiers used in the morphological disambiguation of Arabic texts. We conclude our work in Section 6, and we suggest some perspectives for future research.

2 RELATED WORK

In the literature, disambiguation approaches can be classified into three categories, namely rule-based approaches, statistical approaches, and their combination known as hybrid approaches. We briefly detail and discuss these works in this section.

Rule-based or linguistic approaches have used a knowledge base of rules proposed by linguists to assign labels to different morphological attributes. Systems dealing with linguistic disambiguation approaches are described by (Othman et al., 2004; Daoud, 2009). These techniques are based on heuristics, contextual and non-contextual rules. These rules are often classified into grammatical, structural, and logical categories (Al-Ansary, 2005). Daoud and Daoud (2009) have proposed a specific type of analyzer called En-Converters written in UNL (Universal Networking Language) and EnCo² which is a rule-based programming language. The authors have defined several types of disambiguation rules combining morphological and syntactic contextual dependencies. However, the authors did not perform any experimental evaluation, making it difficult to assess their approach in terms of coverage, reusability, and precision. Some researchers, such as (Othman et al., 2004), have exploited full syntactic parsing for morphological disambiguation.

Statistical approaches are based on learning models from annotated corpora. They incorporate (i) statistical models such as the HMM (Hidden Markov Model) in which the modelled system is assumed to

¹ <http://www.jarir.tn/kunuzcorpus>

² <http://libraries.unl.edu/>

be a Markovian process of unknown parameters or (ii) classification methods such as SVM to compute probabilities of each value resulting from a given word POS. A model can be used to automatically classify other texts by referring to the already calculated probabilities.

For example, Mansour et al. (2007) have combined probabilities calculated on Arabic and Hebrew learning sets to classify the words POS in Arabic texts. ElHadj et al. (2009) have presented a POS tagger system that combines morphological analysis and the Markov model. The tagging process is based on the Arabic sentence structure. First, the text is fully morphologically analyzed to reduce the number of possible POS values. Second, the HMM statistical model, based on the structure of the Arabic sentence, is used to assign each word the exact POS value. ElHadj et al. (2009) have used their annotated corpus which is composed of classical Arabic books. The total of words in this corpus is approximately 21.000 words. Diab et al. (2004) have developed a morphological classifier using SVM. They have trained and tested the classifier using an Arabic Treebank of 4.000 training sentences and 100 test sentences. The most widely used tool, denoted MADA, is developed by (Habash and Rambow, 2005) to solve the morphological ambiguity of Arabic texts. This tool achieves over 86% accuracy in anticipation of diacritics. MADA have prioritized complete analysis in terms of overall accuracy. From its version 2.1, MADA has used the ARAGEN version of BAMA (Habash, 2007) which can generate morphological analysis even for unknown words that are not covered by the lexicon.

A hybrid approach combines linguistic rules with statistical information in order to resolve morphological ambiguity. In (Tlili-Guiassa, 2006), the author has proposed an approach that analyses grammatical and inflectional affixes and grammatical rules based on the MBL (Memory Based Learning) approach. It is applied to classify a collection of Quranic and educational texts. Zribi et al. (2006) have combined the rule-based approach with an HMM trigram-based tagger. The training of the trigram classifier has been performed using texts containing 6.000 words. Heuristic rules were applied to select an analysis among the proposed results.

Khoja (2001) has implemented a hybrid approach based on the Viterbi algorithm. The proposed technique computed two probabilities on an annotated corpus composed of 50.000 words: (i) a lexical probability, which is the probability that a word has a certain value of a morphological attribute; and (ii) a contextual probability, which is the

probability that one tag precedes or succeeds another. A list of grammatical rules is prepared from these statistics in order to achieve an accuracy greater than 90%.

Belguith and Chaâben (2006) have proposed a method for Arabic morphological analysis and disambiguation. It is classified as a statistical approach, but it also includes rules. This approach is based on five steps: (i) segmentation of the text into words; (ii) morphological pre-processing which consists in removing the clitics based on a predefined list; (iii) affixal analysis which recognizes the basic elements of a word, namely the root and the affixes; (iv) morphological analysis based on MORPH2 (Belguith and Chaâben, 2006); and (v) post-processing which identifies word groupings based on a lexicon and a set of rules. This approach has computed the morphological attributes of each word using a determined vocabulary.

Bousmaha et al. (2016) have suggested a hybrid disambiguation approach based on diacritics selection at different levels of analysis. This hybrid approach has combined a linguistic approach with a multicriteria decision and could be considered as an alternative to solve the problem of morpho-lexical ambiguity. During the assessment process, the authors have relied on the online morphological analyzer and obtained encouraging results with an F-measure of over 80%.

Bounhas et al. (2015a) have proposed three possibilistic classifiers for Arabic morphological disambiguation: (i) the first classifier is based on possibility measure, (ii) the second one is based on necessity measure and, (iii) the third one is based on the combination of these two measures. The authors have enriched these classifiers with the information gain scores, useful as weights for the classification attributes, to reduce the required space for resolving the contextual ambiguity, which simplify the disambiguation process.

Later, Bounhas et al. (2015b) have suggested a hybrid possibilistic approach that combines the possibilistic classifier with linguistic rules to assign labels to different morphological attributes. This approach has improved the disambiguation rate of Arabic texts. The authors have also presented an approach dealing with "out-of-vocabulary" words whose morphological analysis is unknown. These possibilistic and hybrid classifiers were tested, in terms of disambiguation rate, on the Arabic "Kunuz" collection and compared to the three ML classifiers SVM, Naïve Bayes and Decision Tree.

More recently, Ayed et al. (2018) have investigated possibilistic morphological

disambiguation of structured Hadiths Arabic texts using semantic knowledge. Training and testing steps required morphological attributes and have been performed using AlKhalil analyzer. The authors have taken advantage of the XML format of the structured Hadiths texts of "Kunuz" collection to benefit from the available semantic information. They have involved semantic attributes in their possibilistic classifiers. Experimental results showed an improvement in the disambiguation rates of possibilistic classifiers when considering semantic knowledge.

3 ARABIC MORPHOLOGICAL AMBIGUITY

Arabic words are often ambiguous in their morphological analysis because of the richness of the Arabic affixation and clitic systems (Elayeb and Bounhas, 2016; Elayeb, 2018; Elayeb, 2019; Elayeb, 2021). Besides, the omission of short vowels in standard orthography can amplify the ambiguity of certain words. We determine in the following the main factors of ambiguity in Arabic texts, namely (i) agglutination ambiguity; (ii) derivational and inflectional ambiguity; and (iii) ambiguity of non-vocalized texts.

3.1 Agglutination Ambiguity

In Arabic, some prepositions and pronouns stick to the nouns, verbs, adjectives and particles to which they are linked. Agglutination is one of the problems found in processing the Arabic language since an agglutinated word can be translated into a sentence in English. The Arabic word "أَسْتَسُونَنَا" (>asatan^osuwnanA) can be translated as "Are you going to forget us?". Then the Arabic language analyzer must segment this word to recognize the root. Moreover, sentences in Arabic do not follow an exact structure such as in French or English: Subject + Verb + Complement, which makes it difficult to process these texts.

3.2 Derivational and Inflectional Ambiguity

Some grammatical factors such as verbs' conjugation or nouns' declension generate inflection of the words forms. The word "يَتَكَلَّمُونَ" (they speak; yatakal-*amuwna*) is the result of the concatenation of the prefix "يَ" (*ya*) indicating the present tense and

the suffix "ونَ" (*wna*) indicating the masculine plural of the verb "تَكَلَّمَ" (he talk; takal-*ama*). Many inflectional operations produce a slight change in pronunciation without an explicit effect on spelling due to the lack of short vowels. A recurring example is the ambiguity of active vs passive vs imperative forms. As an indication, the form "أرسل" (>rsil) becomes "أرسلَ" (he sent; >ar^osala) in the active voice, "أُرْسِلَ" (he is sent; >ur^osila) in the passive voice and "أرسلْ" (sends; >ar^osil) in the imperative. Besides, some affixes can be homographic. For example, the prefix "ت" (*t*) can indicate both male or female person. For example, the word "تأكلُ" (*ta* >^okulu) can be translated into "she eats" and "you eat".

The ambiguity is also caused by the derivation. The Arabic word is the result of a combination of root, vowels, prefixes, infixes, suffixes and a morphological scheme. Prefixes and suffixes can, accidentally, produce a form that is homographic with another word form. For example, the non-vocalized form "أسد" (>sd) can give the meaning of the verb "أَسَدُ" (>asud-*u*) ("أُ + سُدُّ": I block) or the noun "أَسَدٌ" (lion; >asaduN). Likewise, clitics can possibly produce a form that is homographic with another whole word. For example, the form "نَفْسِي" (*nfsy*) can be "نَفْسِي" (*naf^osiy*) which means "psychological" or "myself" which is the combination of "نفس + ي".

3.3 Ambiguity of Non-vocalized Texts

A word is less ambiguous if it is presented in its vocalized form. Non-vocalized words generate many solutions for morphological analysis. A form of a vocalized word can give various morphological interpretations (Habash and Rambow, 2007; Habash et al., 2009b) by adding short vowels. For example, the non-vocalized form "أخرج" (>xrj) can accept about thirty analyses. Among them we cite "أَخْرَجُ" (I go out; >ax^oruju), "أَخْرَجَ" (he brought out; >ax^oraja) and "أَخْرَجَ" (he got out; >ux^oraja). Thus, orthographic alternation operations often produce inflected forms which may belong to several different lemmas or stems. Two words are different just because one of its middle characters is duplicated (chedda). For the previous example, "أَخْرَجُ" (I make out; >x^oriju) is different from the other words only by adding the double "خ" (*xa*) character.

3.4 Discussion

Morphological ambiguities affect other levels of analysis and mislead IRKES results. Syntactically, it is difficult to identify the grammatical function of a word in a sentence. For example, the expression "بحث الرجل" (*bHv Alrjl*) can be interpreted as "بَحَثَ الرَّجُلُ" (*baHava Alrajulu*) which is a whole sentence meaning "The man sought" where the first word "بَحَثَ" (*baHava*) is the verb of the sentence. It can also be read as "بَحْثُ الرَّجُلِ" (*baH°vu Alr~ajuli*) which represents a compound noun meaning "The search for man".

Likewise, the structure of phrases or expressions in Arabic can affect morphological disambiguation. We are mainly talking about a phenomenon commonly recognized in Arabic texts, known as "free word order". For example, the previous expression "بحث الرجل" can be replaced by "الرجل بحث" without changing the meaning. At the semantic level, and because of the agglutination, the derivation and the inflection of the Arabic language, the word "وضوء" (*wDw°*) can have several meanings depending on its morphological interpretation. It can be analyzed as "وَضُوءٌ" (ablution; *wuDw°*), "وَضُوءٌ" (water for ablution; *waDuw°*) or "ضَوْءٌ" (light; *Daw°uN*). In this example, the letter "و" is interpreted as either a conjunction, or like the first letter of the lemma. Even in the second case, we get two possible lemmas diacritized differently.

The example of the sentence "ذهب وحيد الرجل" (**hb wHyd Alrjl*) matches the syntactic and semantic level. Indeed, the words of this sentence are all ambiguous. "ذهب" (**Hb*) can be "ذَهَبٌ" (gold; **ahabuN*) or "ذَهَبَ" (he's gone; **ahaba*). The word "وحيد" (*WHyd*) can be the proper name "وَحِيدٌ" (*Wahid*; *waHiduN*) or "وَحِيدٌ" (alone; *waHid*) or "وَحَيْدٌ" (and he neutralized; *wa Hay~ada*) where the word is an agglutination of the conjunction "و" with the verb "حَيْدَ". The word "الرجل" can be "الرَّجُلُ" (man) or "الرَّجْلُ" (leg). The combination of the solutions of each word gives multiple meanings of this sentence. The phrase can be interpreted as "ذَهَبُ الرَّجُلِ" (one-legged man's gold). From these examples, it is clear that short vowels have great importance in understanding the POS, function and meaning of words. Thus, vocalized texts are less ambiguous than non-vocalized ones. Several disambiguation approaches have been developed to overcome the problem of morphological ambiguity.

We detail in the following our main contributions to deal with the morphological ambiguity of Arabic texts.

4 THE PROPOSED APPROACH

We present our approach for machine-learning morphological dependencies to disambiguate Arabic texts. This approach avoids manual user intervention in the training phase by exploiting vocalized texts which are less ambiguous. We model the disambiguation task as a classification problem as it is already used in several existing systems (Khoja, 2001; Diab et al., 2004; Habash and Rambow, 2005; Habash and Rambow, 2007; Roth et al., 2008; Habash et al., 2009b; Bounhas et al., 2015ab; Ayed, 2017; Ayed et al., 2018).

We present, in Figure 1, the general architecture of the proposed approach. We use vocalized Arabic texts in both training and testing steps. Firstly, ML classifiers are trained using vocalized texts. Secondly, we remove short vowels from these texts to perform the testing step and compute disambiguation rates. A morphological analyzer gives the different analysis values of each word. An analysed word corresponds to the values of the 14 morphological features (POS, Adjective, Aspect, Case, Conjunction, Determiner, Gender, Mode, Number, Particle, Person, Preposition, Voice, Pronoun). Training and testing datasets are presented in a unified format. Indeed, to resolve the ambiguity of a morphological feature (*MF*), we first define the appropriate attributes that describe each instance of the training set. These attributes are also those of the test set.

An attribute, or a morphological characteristic of a given word, is related to the characteristics of its preceding and following words. We specify a window that controls the number of words considered as attributes describing the class of an instance. In many existing approaches, the window size is equal to 2 (Habash and Rambow, 2005). Thus, to classify the *MF_i* of a specific word *w*, we define the attributes *MF_i-2*, *MF_i-1*, *MF_i+1* and *MF_i+2* if the window size is 2 (Ayed et al., 2012b). We can spread the classification attributes to 56 forming the attributes' values of *MF_i ± p* with $i \in [1, 14]$; where 14 is the number of attributes and $p \in \{1, 2\}$ (Ayed et al., 2012a). They indicate, respectively, the morphological attributes of the two preceding words and the two following words of *w*. Thus, we use the other morphological characteristics of the two neighbouring words to determine the class value of the current word. The class value is known in the training set. It corresponds

to the value of the analysis of the MF_i attribute of the word w . However, in the test set, the class is unknown, and the values of the morphological attributes can be ambiguous since they correspond to the analysis of non-vocalized words. This is because the building of test instances differs from the building of the training instances only in the input data which are non-vocalized words. Therefore, the class is the morphological feature to be disambiguated.

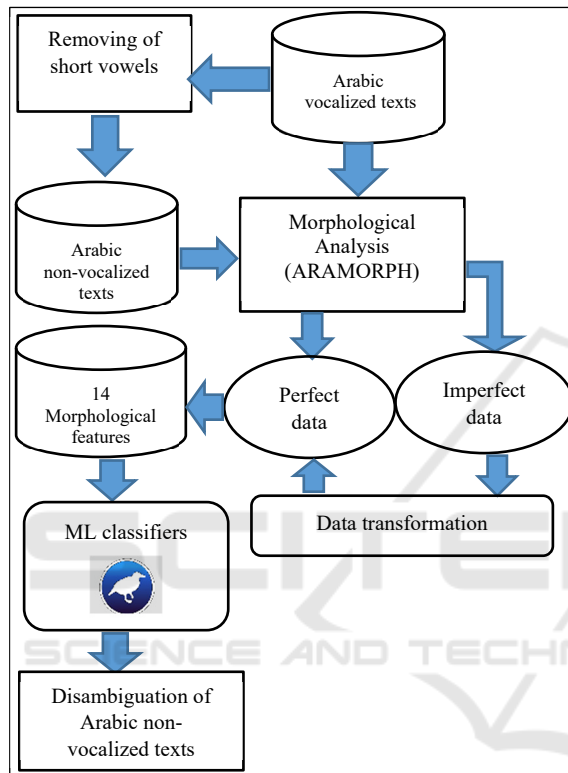


Figure 1: Overview of the approach architecture.

Unlike existing tools (Diab et al., 2004; Habash and Rambow, 2005; Habash and Rambow, 2007; Habash et al., 2009b) whose learn from an annotated corpus, we build our architecture-based ML classifiers from unannotated texts. This training method is widely used for unsupervised disambiguation. In our case, the context used to resolve the ambiguity of a given word is itself ambiguous since an attribute in a training set can have several values which correspond to the possible analysis of a vocalized word. Indeed, a vocalized word itself can be ambiguous. This kind of problem can be considered as a case of imprecision (Bounhas et al., 2015ab).

ML classifiers used for morphological disambiguation accept only perfect data with training and test instances involving complete and exact

information. For this purpose, we propose a method of transforming imperfect data into perfect data, with the aim at aligning them with the input format of ML classifiers. The main idea of our approach is to transform training and/or test instances containing imperfect data into data with precise attributes and classes. We describe, in the following, the operating mode of the proposed approach to transform data, in order to overcome the problem of imperfection.

Table 1: Example of imperfect instances of the training set.

POS-1	POS+1	POS (class)
NOUN	VERB-PERFECT	{NOUN ; NOUN-PROP}
{NOUN ; VERB-PERFECT}	NOUN-PROP	NOUN
NOUN-PROP	NOUN	NOUN
NOUN	VERB-PERFECT	VERB-PERFECT

We illustrate, through Table 1, an example of an imperfect training dataset. We assume that the POS is the morphological feature (MF) to be disambiguated. We also assume that the attributes are POS-1 and POS+1. The training set includes 4 instances. The first instance is uncertain since it provides two possible values of the class (NOUN and NOUN-PROP). The second instance is imprecise because it gives two possible values of the POS-1 attribute (NOUN and VERB-PERFECT).

We transform this base of instance to obtain a perfect dataset without loss of information. To solve the imprecision problem, we denote the values of attribute A as $A_i = \{a_1, a_2, \dots, a_n\}$. Firstly, we associate attribute A with each of its a_i values to form a new attribute denoted " A_{a_i} ". Secondly, and given that the POS-1 attribute has 3 possible values in the previous dataset (see Table 1), we associate 3 new attributes with it (POS-1_NOUN, POS-1_VERB-PERFECT and POS-1_NOUN-PROP). Thirdly, we assign binary values (0 or 1) to the new attributes. For a given instance, if a_i belongs to the values of attribute A , then the attribute " A_{a_i} " is equal to 1. Finally, we generate a new training set with precise attributes given by Table 2. To overcome the problem of class uncertainty, we propose the decomposition of an instance into other instances with a single class value. For an instance with n possible values of a class $\{c_1, c_2, \dots, c_n\}$, we get n instances associated with it. Each of these instances has the same attribute values and class c_i . For the example of Table 2, we produce a training set whose attributes and classes are completely perfect (cf. Table 3). Thus, this transformation method overcomes the problem of imperfect data, and enables us to run ML classifiers for Arabic texts disambiguation since training and testing steps require precise and certain instances.

Table 2: Sample of a transformed training set with precise attributes.

POS-1_NOUN	POS-1_VERB-PERFECT	POS-1_NOUN-PROP	POS+1_NOUN	POS+1_VERB-PERFECT	POS+1_NOUN-PROP	POS (class)
1	0	0	0	1	0	{NOUN ; NOUN-PROP}
1	1	0	0	0	1	NOUN
0	0	1	1	0	0	NOUN
1	0	0	0	1	0	VERB-PERFECT

Table 3: A training set transformed with certain classes.

POS-1_NOUN	POS-1_VERB-PERFECT	POS-1_NOUN-PROP	POS+1_NOUN	POS+1_VERB-PERFECT	POS+1_NOUN-PROP	POS (class)
1	0	0	0	1	0	NOUN
1	0	0	0	1	0	NOUN-PROP
1	1	0	0	0	1	NOUN
1	1	0	0	0	1	NOUN
0	0	1	1	0	0	NOUN
0	0	1	1	0	0	NOUN
1	0	0	0	1	0	VERB-PERFECT
1	0	0	0	1	0	VERB-PERFECT

5 EXPERIMENTATION AND RESULTS

We present in this section a brief description of the test collection (cf. Section 5.1), the experimental scenario and results (cf. Sections 5.2 and 5.3, respectively). Finally, we discuss a comparative study highlighting the efficiency of each ML classifiers. The statistical significance improvements of the best classifiers are also investigated in Section 5.4.

5.1 Test Collection

The main objective of our approach is to train ML classifiers (i.e., acquire morphological dependencies) using vocalized texts, then the testing task is performed using non-vocalized texts. The training process of the morphological dependencies of the Hadith Arabic texts has been performed through the morphological analyzer of vocalized text "ARAMORPH" to obtain the values of the 14 morphological features. Then, a step of eliminating short vowels is essential to be able to test on non-vocalized texts.

Furthermore, we consider the classical Arabic texts, which have been ignored in previous works. Therefore, we use a collection of Arabic stories namely the "Kunuz" corpus of Hadith texts, which have been studied in several works such as (Harrag et al., 2009b; Bounhas et al., 2010; Bounhas et al., 2011b; Harrag et al., 2013; Bounhas et al., 2020). The Hadiths cover religious knowledge as well as

common and universal knowledge. Moreover, the Kunuz corpus is one of the rare vocalized Arabic corpora.

We use the six well-recognized encyclopedic books organized by theme namely "صحيح البخاري", (Sahih Al-Bukhari) "صحيح مسلم" (Sahih Muslim), "سنن أبي داود" (Sunan Abi Dawud), "سنن الترمذي" (Sunan Ettermidhi), "سنن ابن ماجه" (Sunan Ibn Majah) and "سنن النسائي" (Sunan Annasai) (Al-Echikh, 1998). We limit our experiments to three sub-corpora corresponding to the following areas of interest: "الأشربة" ($Al>\$RBP$; drinks), "الزواج" ($AlzwaJ$; marriage) and "الطهارة" ($AlThArp$; purification) (Ayed et al., 2012b).

These areas were chosen because they are generic and exist in the various books of Hadith. Table 4 presents the number of words in the three sub-corpora for the six books. While Table 5 summarizes the data size for the 14 morphological features with their corresponding total number of attributes and instances.

Table 4: The number of words in the three sub-corpora for the six books.

Hadith Book	Drinks	Marriage	Purification	Total
Sahih Al-Bukhari	02766	11521	11016	25303
Sahih Muslim	09117	06693	05063	20873
Sunan Abi Dawud	02672	05780	15319	23771
Sunan Ettermidhi	01835	05910	09291	17036
Sunan Ibn Majah	01748	06539	13179	21466
Sunan Annasai	06703	08741	12554	27998
Total	24841	45184	66422	136447

Table 5: Summary of the data size for the 14 morphological features.

Morphological feature	Size (Mo)	Attributs	Instances
POS	215	1961	38304
ADJECTIVE	11.3	105	37562
ASPECT	22.6	209	37567
CASE	22.5	209	37564
CONJUNCTION	11.3	105	37562
DETERMINER	33.7	313	37631
GENDER	16.8	157	37562
MODE	16.9	157	37562
NUMBER	22.4	209	37563
PARTICLE	22.6	209	37713
PERSON	22.5	209	37571
PREPOSITION	11.4	105	37562
VOICE	17.0	157	37562
PRONOUN	358	3329	37615

5.2 Experimental Scenario

We have conducted a set of experimental tests using 7 types of classic classifiers based on 20 machine-learning algorithms such as: (1) "Bayes classifiers" include *Naïve Bayes* and *Bayes Net* algorithms. (2) "Function classifiers" are based on *SVM* (Support Vector Machine), *Logistic* and *SMO* (Sequential Minimal Optimization) algorithms. (3) "Lazy classifiers" involve the algorithms *IBK* (*k*-nearest neighbors (*k*-NN)), *KSTAR* (*K**) and *LWL* (Locally Weighted Learning) algorithms. (4) "Meta classifiers" contain *ClassificationViaRegression*, *FiltredClassifier* and *Vote* algorithm. (5) "MISC classifiers" incorporate *InputMappedClassifier* and *SerializedClassifier*. (6) "Rules classifiers" are based on *DecisionTable*, *ZeroR* and *OneR* classifiers. (7) "Trees classifiers" integrate *J48*, *RandomForest*, *RandomTree* and *DecisionStump* algorithms.

We have applied a 10-fold cross-validation technique (Kohavi, 1995) on the three domains of application extracted from the six books of Hadith to estimate the performance of the 20 ML classifiers. For each morphological feature, we calculate the average disambiguation rate over the (9 + 1) iterations. To benefit from these rates, we have proceeded as follows: (i) we analyze the vocalized texts and we store the correct morphological solutions; (ii) we remove short vowels from the same texts; (iii) we disambiguate the texts obtained with the 20 ML classifiers, then we store the results; and (iv) we compare the two results to compute the disambiguation rate.

Firstly, we have experimented these ML classifiers with their default parameters in WEKA⁶ tool. Then, we have optimized these parameters (Pedro and Pazzani, 1997) to enhance the disambiguation rate of each classifier. We have performed the optimisation process of these ML classifiers using some WEKA meta-classifiers such as *Grid search*, *Threshold selector* and *CVParameter selection*.

We have used in our optimisation process the *CVParameter selection* which is the most popular and efficient. We have fixed the Confidence Factor *C* (e.g., *C* ranged from 0.1 to 0.5 with 5 steps classifiers options) and modify the Minimum Object *M* and vice-versa until achieving the best disambiguation rate. A comparative study between our default and optimized results is discussed in Section 5.3. Then an investigation of the statistical significance improvement of each optimized classifier is presented in Section 5.4.

5.3 Experimental Results

We assess the classical ML classifiers in terms of morphological disambiguation rate. We compare the results with those given by the known efficient classifiers SVM and Naïve Bayes for the 14 morphological features. Imperfect test instances require a transformation process to align with the input format of classical ML classifiers. These instances have imperfect attributes and classes. We illustrate, in Tables 6-9, the morphological disambiguation rates of the 14 morphological features given by the above mentioned 20 ML classifiers using their default and optimal parameters.

The experiments show that the SVM, Naïve Bayes and Decision Tree classifiers, with their optimal parameters, have the best average disambiguation rates of 81.45%, 80.75% and 80.51%, respectively. For some morphological feature, we obtain the same results by certain classifiers such as SVM and SMO having very similar algorithms. This can be explained by the fact that the associated morphological features have few (less than 6) possible values. On the other hand, some other features generate different results by different classifiers. For example, the attribute PRONOUN has 64 possible values. The results highlight that the good optimization of the parameters of any ML classifier improve its precision when disambiguating classical Arabic texts.

We note here that some ML classifiers are lacked by the data size of some morphological features (e.g.,

⁶ <https://www.cs.waikato.ac.nz/ml/weka/>

POS and PRONOUN), whose cannot be handled even with the maximum size of WEKA memory (2020 Mo). This problem has been solved by considering some parts of these data (randomly selected) rather than working with the full data size. This process decreases the disambiguation rate of these big-sized morphological features if compared to others.

5.4 Comparison and Discussion

The goal is to investigate the statistically significant improvements of the 20 optimized ML classifiers in terms of disambiguation rate. For this purpose, we use Wilcoxon Matched-Pairs Signed-Ranks Test proposed by (Demsar, 2006). It is a non-parametric alternative to the paired t -test that allows us to compare two classifiers based on multiple features. The p -values are calculated by comparing the best

SVM classifier with the 19 other remaining ones (cf. Table 10). Then, the second-best classifier Naïve Bayes is also compared to the 18 remaining classifiers (cf. Table 11). The statistically significant improvement of a given classifier over a second one is confirmed if the computed p -value < 0.05 .

The p -values of Tables 10 and 11 are less than 0.05 (indicated by *) for the classifiers *IBK*, *KSTAR*, *LWL*, *InputMappedClassifier*, *Stacking* and *OneR*. We note here that the difference between the median of the reference algorithm SVM (respectively Naïve Bayes) and that of the remaining algorithms is statistically significant. Whereas the p -values are greater than 0.05 for the other remaining algorithms. In this case, we do not have sufficient evidence to conclude that the median of SVM classifier (respectively Naïve Bayes) is statistically different from the median of the remaining algorithms.

Table 6: Disambiguation rates of 14 morphological features using default and optimal classifiers' parameters (1/4).

Morphological feature	SVM		Naïve Bayes		Decision Tree		Bayesian Net		Vote	
	Default	Optimal	Default	Optimal	Default	Optimal	Default	Optimal	Default	Optimal
POS	89.98 %	92.71%	43.35%	79.54%	71.61%	76.81%	42.97%	52.02%	29.96%	38.25%
ADJECTIVE	96.51%	97.72%	96.51%	97.71%	96.51%	97.72%	96.51%	97.72%	96.51%	97.72%
ASPECT	71.20%	77.56%	71.20%	77.56%	71.20%	77.56%	71.20%	77.56%	71.21%	77.56%
CASE	56.12%	68.21%	56.12%	68.21%	56.12%	68.21%	56.12%	68.21%	56.12%	68.21%
CONJUNCTION	83.03%	87.62%	83.03%	87.62%	83.03%	87.62%	83.03%	87.62%	83.03%	87.62%
DETERMINER	64.12%	67.67%	64.12%	67.67%	64.16%	67.67%	64.12%	67.67%	64.12%	67.67%
GENDER	57.15%	63.77%	57.15%	63.77%	57.15%	63.77 %	57.15%	63.77%	57.15%	63.77%
MODE	99.32 %	99.38%	99.32%	99.38%	99.32%	99.38%	99.32%	99.38%	99.32%	99.38%
NUMBER	85.18%	93.43%	85.18%	93.43%	85.18%	93.43%	85.18%	93.43%	85.18%	93.43%
PARTICLE	96.65%	98.81%	96.65%	98.81%	96.65%	98.81%	96.65%	98.81%	96.65%	98.81%
PERSON	60.22%	67.55%	60.22 %	67.55%	60.22%	67.55%	60.22%	67.55%	60.22%	67.55%
PREPOSITION	82.87%	85.12%	82.87%	85.12%	82.87%	85.12%	82.87%	85.12%	82.87%	85.12%
VOICE	71.21%	77.92%	71.21%	77.92%	71.21%	77.92%	71.21%	77.92%	71.21%	77.92%
PRONOUN	56.88 %	62.81%	59.06 %	66.24%	62.38%	65.61%	61.08%	67.70%	58.94%	63.33%
Average	76.46%	81.45%	73.29%	80.75%	75.54%	80.51%	73.40%	78.89%	72.32%	77.60%

Table 7: Disambiguation rates of 14 morphological features using default and optimal classifiers' parameters (2/4).

Morphological feature	SMO		J48		RandomForest		DecisionStump		Logistic	
	Default	Optimal	Default	Optimal	Default	Optimal	Default	Optimal	Default	Optimal
POS	43.02%	52.85%	58.64%	64.68%	29.96%	34.58%	29.96%	33.61%	27.09%	35.12%
ADJECTIVE	96.51%	97.72%	96.51%	97.72%	96.51%	96.58%	96.51%	96.58%	96.51%	96.58%
ASPECT	71.20%	77.56%	71.20%	77.56%	71.20%	77.27%	71.20%	77.27%	71.20%	77.27%
CASE	71.20%	68.21%	56.12%	68.21%	56.12%	68.30%	56.12%	68.30%	56.12%	68.30%
CONJUNCTION	83.03%	87.62%	83.03%	87.62%	83.03%	86.60%	83.03%	86.60%	83.03%	86.60%
DETERMINER	64.16%	67.67%	64.16%	67.67%	64.16%	65.64%	64.12%	68.81%	64.12%	68.81%
GENDER	57.15%	63.77%	57.15%	63.77%	57.15%	65.72%	57.15%	65.72%	57.15%	65.72%
MODE	99.32%	99.38%	99.32%	99.38%	99.32%	99.40%	99.32%	99.40%	99.32%	99.40%
NUMBER	85.18%	93.43%	85.18%	93.43%	85.18%	88.59%	85.18%	88.59%	85.18%	88.59%
PARTICLE	96.65%	98.81%	96.65%	98.81%	96.65%	96.80%	96.65%	96.80%	96.65%	96.80%
PERSON	60.22%	67.55%	60.22%	67.55%	60.22%	69.82%	60.22%	69.82%	60.22%	69.82%
PREPOSITION	82.87%	85.12%	82.87%	85.12%	82.87%	85.71%	82.87%	85.71%	82.87%	85.71%
VOICE	71.21%	77.92%	71.21%	77.92%	71.21%	80.02%	71.21%	80.02%	71.21%	80.02%
PRONOUN	61.58%	66.67%	61.58%	66.67%	62.39%	64.55%	60.93%	64.05%	61.58%	64.52%
Average	74.52%	78.88%	74.56%	79.72%	72.57%	77.11%	72.46%	77.23%	72.30%	77.38%

Table 8: Disambiguation rates of 14 morphological features using default and optimal classifiers' parameters (3/4).

Morphological feature	FilteredClassifier		DecisionTable		Classification ViaRegression		ZeroR		IBK (KNN)	
	Default	Optimal	Default	Optimal	Default	Optimal	Default	Optimal	Default	Optimal
POS	55.79%	60.79%	48.68%	52.29%	43.10%	50.10%	29.96%	37.19%	61.36%	62.36%
ADJECTIVE	96.51%	96.58%	96.51%	96.58%	96.51%	96.58%	96.51%	96.58%	96.51%	96.55%
ASPECT	71.20%	77.27%	71.20%	77.27%	71.20%	77.27%	71.20%	77.27%	71.20%	73.20%
CASE	56.12%	68.30%	56.12%	68.30%	56.12%	68.30%	56.12%	68.30%	56.12%	57.12%
CONJUNCTION	83.03%	86.60%	83.03%	86.60%	83.03%	86.60%	83.03%	86.60%	83.03%	83.50%
DETERMINER	64.12%	68.81%	64.12%	68.81%	64.12%	68.81%	64.12%	68.81%	64.18%	65.18%
GENDER	57.15%	65.72%	57.15%	65.72%	57.15%	65.72%	57.15%	65.72%	57.15%	58.15%
MODE	99.32%	99.40%	99.32%	99.40%	99.32%	99.40%	99.32%	99.40%	99.32%	99.33%
NUMBER	85.18%	88.59%	85.18%	88.59%	85.18%	88.59%	85.18%	88.59%	85.18%	87.18%
PARTICLE	96.65%	96.80%	96.65%	96.80%	96.65%	96.80%	96.65%	96.80%	96.65%	96.65%
PERSON	60.22%	69.82%	60.22%	69.82%	60.22%	69.82%	60.22%	69.82%	60.22%	65.22%
PREPOSITION	82.87%	85.71%	82.87%	85.71%	82.87%	85.71%	82.87%	85.71%	82.87%	84.87%
VOICE	71.21%	80.02%	71.21%	80.02%	71.21%	80.02%	71.21%	80.02%	71.21%	73.21%
PRONOUN	61.75%	64.27%	62.30%	65.51%	58.68%	63.12%	58.94%	62.41%	61.67%	63.67%
Average	74.37%	79.19%	73.90%	78.67%	73.24%	78.35%	72.32%	77.37%	74.76%	76.16%

Table 9: Disambiguation rates of 14 morphological features using default and optimal classifiers' parameters (4/4).

Morphological feature	KSTAR		LWL		InputMapped Classifier		Stacking		OneR	
	Default	Optimal	Default	Optimal	Default	Optimal	Default	Optimal	Default	Optimal
POS	62.02%	63.02%	35.53%	36.53%	29.97%	30.97%	29.97%	30.97%	32.69%	33.69%
ADJECTIVE	96.51%	96.55%	96.51%	96.55%	96.51%	96.55%	96.51%	96.55%	96.51%	96.55%
ASPECT	71.20%	73.20%	71.20%	73.20%	71.20%	73.20%	71.20%	73.20%	71.20%	73.20%
CASE	56.12%	57.12%	56.12%	57.12%	56.12%	57.12%	56.12%	57.12%	56.12%	57.12%
CONJUNCTION	83.03%	83.50%	83.03%	83.50%	83.03%	83.50%	83.03%	83.50%	83.03%	83.50%
DETERMINER	64.12%	65.12%	64.12%	65.12%	64.12%	65.12%	64.12%	65.12%	64.12%	65.12%
GENDER	57.15%	58.15%	57.15%	58.15%	57.15%	58.15%	57.15%	58.15%	57.15%	58.15%
MODE	99.32%	99.33%	99.32%	99.33%	99.32%	99.33%	99.32%	99.33%	99.32%	99.33%
NUMBER	85.18%	87.18%	85.18%	87.18%	85.18%	87.18%	85.18%	87.18%	85.18%	87.18%
PARTICLE	96.65%	96.65%	96.65%	96.65%	96.65%	96.65%	96.65%	96.65%	96.65%	96.65%
PERSON	60.22%	65.22%	60.22%	65.22%	60.22%	65.22%	60.22%	65.22%	60.22%	65.22%
PREPOSITION	82.87%	84.87%	82.87%	84.87%	82.87%	84.87%	82.87%	84.87%	82.87%	84.87%
VOICE	71.21%	73.21%	71.21%	73.21%	71.21%	73.21%	71.21%	73.21%	71.21%	73.21%
PRONOUN	58.58%	60.58%	61.69%	63.69%	58.55%	60.55%	58.55%	60.55%	60.47%	63.47%
Average	74.58%	75.98%	72.91%	74.31%	72.29%	73.69%	72.29%	73.69%	72.62%	74.09%

Table 10: The p-values for the statistically significant improvements of SVM compared to the 19 classifiers.

SVM vs.	Naïve Bayes	Decision Tree	Bayesian Net	Vote	SMO	J48	Random Forest	Decision stump	Logistic	Filtred Classifier
	0.59	0.65	0.65	0.65	0.65	0.65	0.63	0.94	0.94	0.94
	Decision Table	Classification ViaRegression	ZeroR	IBK (KNN)	KSTAR	LWL	InputMapped Classifier	Stacking	OneR	
	0.89	0.94	0.75	0.001*	0.0009*	0.001*	0.0009*	0.0009*	0.001*	

Table 11: The p-values for the statistically significant improvements of Naïve Bayes compared to the 18 classifiers.

Naïve Bayes vs.	Decision Tree	Bayesian Net	Vote	SMO	J48	Random Forest	Decision stump	Logistic	Filtred Classifier	Decision Table
	0.28	1	0.28	1	1	0.36	0.59	0.68	0.63	0.72
	Classification ViaRegression	ZeroR	IBK (KNN)	KSTAR	LWL	InputMapped Classifier	Stacking	OneR		
	0.55	0.55	0.0009*	0.0009*	0.0009*	0.0009*	0.0009*	0.0009*	0.0009*	

6 CONCLUSION

We have focused in this paper on the problem of morphological disambiguation of Arabic texts. The latter have a very rich and complex morphology that has given rise to many challenges for the natural language processing tasks. The morphological disambiguation process still among the main challenge for the information retrieval systems. It is useful to identify the appropriate form of index terms in Arabic IR. In this context, we have proposed a new architecture for morphological disambiguation of Arabic terms using a series of classical ML algorithms. During the data pre-processing step, we have proposed and implemented a data transformation technique useful to transform imperfect data to perfect ones. Then, the selected classifiers are trained on vocalized Arabic texts and tested on non-vocalized ones. We have also performed an optimisation process to enhance the efficiency of each classifier, a method that none has suggested before to improve the morphological disambiguation rate of Arabic texts.

We have experimented these ML classifiers using the "Kunuz" collection of classical Arabic texts in order to compare and discuss their efficiency. The SVM classifier seems to be the most efficient in the morphological disambiguation of Arabic texts. It achieved a statistically significant improvement over a few competing algorithms. Besides, the second efficient Naïve Bayes classifier has achieved some statistically significant improvements compared to some ML algorithms. Our short-term concern is to use Friedman's statistical test to compare the 20 ML classifiers together to more investigate the degree of statistically significant improvement of each algorithm. But our long-term concern consists in testing these ML classifiers using the modern Arabic texts collection TreeBank.

ACKNOWLEDGEMENTS

This work was funded by the Liwa College of Technology in Abu Dhabi (UAE) under research grant IRG-BIT-002-2020.

REFERENCES

- Al-Ansary, S. (2005). *Building a Computational Lexicon for Arabic: A corpus-based approach*. In Alhawary, M. T., Benmamoun, E., (eds.): *Current Issues in Linguistic Theory*, volume 267, pp. 173–193. John Benjamins Publishing Company, Amsterdam.
- Al-Echikh, A. (1998). *Encyclopedia of the six major citation collections*. Dar-esselem, Ryadh, KSA.
- Ayed, R. (2017). *Désambiguïation Morphologique de Textes Arabes à Base de Classification Possibiliste pour la Recherche d'Information Socio-Sémantique*. PhD Thesis, ENSI, Manouba University, Tunisia.
- Ayed, R., Bounhas, I., Elayeb, B., Evrard, F., Bellamine Ben Saoud, N. (2012a). Arabic Morphological Analysis and Disambiguation Using a Possibilistic Classifier. In *Proceedings of ICIC-2012*, pp. 274–279, Huangshan, China. Springer Berlin Heidelberg.
- Ayed, R., Bounhas, I., Elayeb, B., Evrard, F., Bellamine Ben Saoud, N. (2012b). A Possibilistic Approach for the Automatic Morphological Disambiguation of Arabic Texts. In *Proceedings of SNPD-2012*, pp. 187–194, Kyoto, Japan, IEEE Computer Society.
- Ayed, R., Chouigui, A., Elayeb, B. (2018a). A New Morphological Annotation Tool for Arabic Texts. In *Proceedings of AICCSA-2018*, pp. 1-6, Aqaba, Jordan, IEEE Computer Society.
- Ayed, R., Elayeb, B., Bellamine Ben Saoud, N. (2018b). Possibilistic Morphological Disambiguation of Structured Hadiths Arabic Texts Using Semantic Knowledge. In *Proceedings of ICAART-2018*, pp. 565-572, Funchal, Madeira, Portugal, SciTePress,.
- Belguith, L. H., Chaâben, N. (2006). Analyse et désambiguïation morphologiques de textes arabes non voyellés. In *Proceedings of TALN-2006*, pp. 493–501, Leuven, Belgique, ATALA.
- Ben Khiroun, O., Ayed, R., Elayeb, B., Bounhas, I., Bellamine Ben Saoud, N., Evrard, F. (2014). Towards a New Standard Arabic Test Collection for Mono- and Cross-Language Information Retrieval. In *Proceedings of NLDB-2014*, LNCS 8455, pp. 168–171, Montpellier, France, Springer International Publishing.
- Bounhas, I., Ayed, R., Elayeb, B., Bellamine Ben Saoud, N. (2015b). A hybrid possibilistic approach for Arabic full morphological disambiguation. *Data & Knowledge Engineering*, 100:240-254.
- Bounhas, I., Ayed, R., Elayeb, B., Evrard, F., Bellamine Ben Saoud, N. (2015a). Experimenting a discriminative possibilistic classifier with reweighting model for Arabic morphological disambiguation. *Computer Speech & Language*, 33(1):67-87.
- Bounhas, I., Elayeb, B., Evrard, F., Slimani, Y. (2010). Toward a computer study of the reliability of Arabic stories. *Journal of the American Society for Information Science and Technology*, 61(8):1686–1705.
- Bounhas, I., Elayeb, B., Evrard, F., Slimani, Y. (2011b). Organizing Contextual Knowledge for Arabic Text Disambiguation and Terminology Extraction. *Knowledge Organization*, 38(6):473–490.
- Bounhas, I., Soudani, N., Slimani, Y. (2020). Building a morpho-semantic knowledge graph for Arabic information retrieval. *Information Processing & Management*, 57(6): 102124

- Bousmaha, K. Z., Rahmouni, M. K., Kouinef, B., Belguith, L. H. (2016). A Hybrid Approach for the Morpho-Lexical Disambiguation of Arabic. *Journal of Information Processing Systems*, 12(3):358–380.
- Daoud, D. (2009). Synchronized Morphological and Syntactic Disambiguation for Arabic. *Advances in Computational Linguistics*, 41:73–86.
- Daoud, D., Daoud, M. (2009). Arabic Disambiguation Using Dependency Grammar. In *Proceedings of TALN-2009*, Senlis, France, ATALA.
- Demsar, J. (2006). Statistical Comparisons of Classifiers over Multiple Data Sets. *Journal of Machine Learning Research*, 7:1–30.
- Diab, M., Hacıoglu, K., Jurafsky, D. (2004). Automatic Tagging of Arabic Text: From Raw Text to Base Phrase Chunks. In *Proceedings of HLT-NAACL 2004: Short Papers, HLT-NAACL-Short'04*, pp. 149–152, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Elayeb, B. (2009). *SARIPOD : Système multi-Agent de Recherche Intelligente POSSibiliste des Documents Web*. PhD. Thesis in Artificial Intelligence, INPT, Toulouse, France, 2009.
- Elayeb, B. (2018). *Recherche d'Information Possibiliste : De la Désambiguïsation et la Reformulation de Requêtes vers la Fiabilité de l'Information Recherchée*. HDR Thesis, ENSI, Manouba University, Tunisia.
- Elayeb, B. (2019). Arabic Word Sense Disambiguation: A Review. *Artificial Intelligence Review*, 52(4):2475–2532.
- Elayeb, B. (2021). Arabic Text Classification: A Literature Review. In *Proceedings of AICCSA-2021*, pp. 1-8, Tangier, Morocco, IEEE Computer Society.
- Elayeb, B., Bounhas, I. (2016). Arabic Cross-Language Information Retrieval: A Review. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 15(3):18:1-18:44.
- Elayeb, B., Chouigui, A., Bounhas, M., Ben Khiroun, O. (2020). Automatic Arabic Text Summarization Using Analogical Proportions. *Cognitive Computation*, 12(5): 1043-1069.
- Elayeb, B., Evrard, F., Zaghdoud, M., Ben Ahmed, M. (2009). Towards an intelligent possibilistic web information retrieval using multiagent system. *The International Journal of Interactive Technology and Smart Education*, 6(1):40–59.
- ElHadj, Y., Al-Sughayeir, I. A., Al-Ansari, A. M. (2009). Arabic part-of-speech tagging using the sentence structure. In *Proceedings of the 2nd International Conference on Arabic Language Resources and Tools*, pp. 241–245, Cairo, Egypt.
- Habash, N. (2007). *Arabic Morphological Representations for Machine Translation*. In Soudi, A., Bosch, A. v. d., Neumann, G., (eds.): *Arabic Computational Morphology*, Vol. 38, *Speech and Language Technology*, pages 263–285. Springer Netherlands.
- Habash, N., Rambow, O. (2005). Arabic Tokenization, Part-of-speech Tagging and Morphological Disambiguation in One Fell Swoop. In *Proceedings of ACL'05*, pp. 573–580, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Habash, N., Rambow, O. (2007). Arabic Diacritization Through Full Morphological Tagging. In *Proceedings of HLT-NAACL 2007: Short Papers, HLT-NAACL-Short'07*, pp. 53–56, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Habash, N., Rambow, O., Roth, R. (2009b). MADA+TOKAN: A toolkit for Arabic tokenization, diacritization, morphological disambiguation, POS tagging, stemming and lemmatization. In *Proceedings of the 2nd International Conference on Arabic Language Resources and Tools (MEDAR)*, pp. 102–109, Cairo, Egypt.
- Harrag, F., Alothaim, A., Abanmy, A., Alomaigan, F., Alsalehi, S. (2013). Ontology Extraction Approach for Prophetic Narration (Hadith) using Association Rules. *International Journal on Islamic Applications in Computer Science and Technology*, 1(2).
- Harrag, F., Hamdi-Cherif, A., Malik, A., Al-Salman, S., El-Qawasmeh, E. (2009b). Experiments in improvement of Arabic information retrieval. In *proceedings of the 3rd International Conference on Arabic Language Processing (CITALA)*, Rabat, Morocco.
- Khoja, S. (2001). APT: Arabic part-of-speech tagger. In *Proceedings of the NAACL-2001*, Pennsylvania, USA.
- Kohavi, R. (1995). A Study of Cross-validation and Bootstrap for Accuracy Estimation and Model Selection. In *Proceedings of IJCAI'95*, Volume 2, pp. 1137–1143, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.
- Mansour, S., Sima'an, K., Winter, Y. (2007). Smoothing a Lexicon-based POS Tagger for Arabic and Hebrew. In *Proceedings of the Workshop Semitic'07*, pp. 97–103, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Othman, E., Shaalan, K., Rafea, A. (2004). Towards resolving ambiguity in understanding Arabic sentence. In *Proceedings of the International Conference on Arabic Language Resources and Tools (NEMLAR)*, pp. 118–122, Cairo, Egypt.
- Pedro, D., Pazzani, M. (1997). On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning*, 29:103–137.
- Roth, R., Rambow, O., Habash, N., Diab, M., Rudin, C. (2008). Arabic Morphological Tagging, Diacritization, and Lemmatization Using Lexeme Models and Feature Ranking. In *Proceedings of HLT-NAACL-Short'08, HLT-Short'08*, pp. 117–120, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Tlili-Guiassa, Y. (2006). Hybrid Method for Tagging Arabic Text. *Journal of Computer Science*, 2(3):245–248.
- Zribi, C. B. O., Torjmen, A., Ahmed, M. B. (2006). An Efficient Multiagent System Combining POS-Taggers for Arabic Texts. In *Proceedings of CICLing-2006*, LNCS 3878, pp. 121–131, Mexico City, Mexico, Springer Berlin Heidelberg.