Possibilistic Morphological Disambiguation of Structured Hadiths Arabic Texts using Semantic Knowledge

Raja Ayed[1,2], Bilel Elayeb[1,3] and Narjès Bellamine Ben Saoud[1]
[1]*RIADI Research Laboratory, ENSI, Manouba University, 2010, Tunisia*
[2]*Faculty of Sciences of Gabes, Gabes University, 6072, Tunisia*
[3]*Emirates College of Technology, P.O. Box 41009, Abu Dhabi, U.A.E.*

Keywords: Possibility Theory, Arabic Morphological Disambiguation, Hadiths Structured Corpus, Classification Attributes.

Abstract: We propose, in this paper, a possibilistic morphological approach to disambiguate hadiths Arabic texts using semantic knowledge. The disambiguation is considered as a classification problem. The possibilistic approach uses vocalized texts to train a possibilistic classifier in order to classify non-vocalized texts as they are more ambiguous. Morphological attributes are used for training and test. Hadiths are structured in XML format that provides semantic information. We enlarge the classification attributes' set by adding semantic attributes extracted from the hadiths structure. We prove that the possibilistic approach gives the best rates using AlKhalil analyzer to prepare the training and the test sets. Our proposed possibilistic approach enhances disambiguation rates of Arabic hadiths' texts when it includes semantic knowledge.

1 INTRODUCTION

Arabic belongs to the Semitic languages and is considered as the major language of the Middle East. It is written from right to left. It has many special properties and a complex method of words' construction especially for the classical Arabic. Morphological disambiguation approaches deal with the complex morphology of Arabic language to resolve ambiguity. Vocalized texts are considered less ambiguous than non-vocalized ones. Hadiths corpus presents an interesting data source of classical and vocalized texts that can be used to evaluate a morphological disambiguation approach. As the disambiguation is considered as a classification task, we propose the morphological possibilistic approach that learns morphological knowledge from vocalized texts and exploit this knowledge to disambiguate non-vocalized ones. To do so, we organize this paper as follows: we give, in section 2 and section 3, an overview of the Arabic morphological ambiguity and the morphological disambiguation approaches. In section 4, we illustrate the possibilistic disambiguation approach. Section 5 describes the hadiths corpus and focuses on the socio-semantic aspects of hadiths. In section 6, we describe our approach that takes advantage of the hadiths structure and their semantic knowledge to disambiguate morphological features using the possibilistic disambiguation approach. Subsequently, a set of experimentations results are presented in section 7. Finally, we summarize in section 8 our findings in the conclusion and propose future perspectives.

2 ARABIC MORPHOLOGICAL AMBIGUITY

The morphological analysis tries to identify possible solutions, of a given word, of some morphological features like POS, gender, aspect, case, etc. Some morphological analysis tools give the stem (Hajic, 2000) or the lemma (Attia, 2008) of a word to determine correctly its morphological features. They give, also, a segmented solution specifying the prefixes, affixes and suffixes. A word is considered ambiguous if it has more than just one solution. Disambiguation assigns, to the ambiguous word, the most appropriate analysis given the context of the word (Ayed et al., 2012). Disambiguation becomes a challenging field of many NLP researches (Ayed et al., 2012). This issue is closely related to the

Ayed, R., Elayeb, B. and Saoud, N.
Possibilistic Morphological Disambiguation of Structured Hadiths Arabic Texts using Semantic Knowledge.
DOI: 10.5220/0006651105650572
In *Proceedings of the 10th International Conference on Agents and Artificial Intelligence (ICAART 2018) - Volume 2*, pages 565-572
ISBN: 978-989-758-275-2
Copyright © 2018 by SCITEPRESS – Science and Technology Publications, Lda. All rights reserved

complexity of the morphology of Arabic language.

Ambiguity is mainly caused by the agglutination and the lack of diacritics (short vowels). So, the same word could have many interpretations (Habash et al., 2009). For example, the word “أستتفكروننا” is agglutinated. It means in English “Will you remember us?”. In this case, morphological analyzer aims to remove the prefixes and suffixes to extract the root of the word and give the right morphological feature. Some prefixes and suffixes can be homographic with each other (Attia, 2008). Without diacritics, the word “قرأت” (qrAt) can mean “ قَرَأْتُ“ (qaraAtu; I read) or “قَرَأْتَ” (qaraAta; you read). Some other meanings can be identified by adding more short vowels, or diacritics. Prefixes and suffixes can by coincidence, produce a form that is homographic with another full form. For instance, the vocalized word “أَكْرَمُ” (>ak°ramu) can be, at the same time, (i) the proper noun “Akram” in some grammatical cases and (ii) the adjective “more generous”.

Morphological ambiguities can reach even the syntactic and the semantic levels of analysis. At the syntactic level, the sentence may give different grammatical functions if it contains one or more ambiguous words (Bounhas et al., 2011). For example, all the words composing the sentence " ذهب وحيد الرجل" (*hb wHyd Alrjl) are ambiguous. ”ذهب ” can be “ذَهَبُ” (*ahabu; gold) or “ذَهَبَ” (*ahaba; he goes). “وحيد” can be the proper noun “وَحِيدٌ“ (waHiydN; Wahid) or “وَحِيدُ” (waHiydu; single). “الرجل” can be “الرَّجُلُ” (Alrrajulu; the man) or “الرِجْلِ” (Alrrijli; the leg). The combination of the solutions, of each word, may give multiple meanings of this sentence. This affects, even, the semantic level of ambiguity. The sentence may be interpreted as, for example, “ذَهَبُ وَحِيدِ الرِجْلِ" (the gold of the man with a single leg). These illustrations show that diacritics, or shorts vowels, are essential in determining the grammatical category, the morphological feature and the significance of some words. Therefore, vocalized texts are considered less ambiguous than non-vocalized texts.

3 ARABIC MORPHOLOGICAL DISAMBIGUATION

Researches, in Arabic NLP, are interested in disambiguation of Arabic words by identifying their morphological features depending on their context. Several classification methods are used to disambiguate 1morphological words’ features (Habash et al., 2009; Roth et al., 2008). Hence, the disambiguation is considered as a classification problem where a morphological feature of an ambiguous word is the class that we aim to identify.

The classification methods are based on learning techniques to train classifier from the datasets where their morphological features are known.

The existing approaches in disambiguation can be divided into three categories that are: (i) linguistic approaches, (ii) statistical approaches and (iii) hybrid approaches.

Linguistic approaches are, also, called rule-based approaches. They use rules written by linguists to tag the different morphological features (Diab et al., 2007). We talk about heuristics, contextual and non-contextual rules (Elshafei et al., 2002).

Statistical approaches incorporate different classification methods such as Support Vector Machine (SVM) and Hidden Markov models (HMM) to calculate probability of each value of a grammatical category of a word (Vapnik, 1999). SVM was used by the disambiguation tools MADA (Habash et al., 2013) and MADAMIRA (Pasha et al., 2014) that combines MADA and AMIRA (Habash et al., 2009). To define the grammatical category of a word, MADAMIRA converts the Arabic text to the Buckwalter (2004) transliteration. Then, it analyzes the text morphologically and gives all the possible POS for each word. Finally, it applies SVM and language models to choose the right POS from all proposed POS values after the morphological analysis.

The hybrid approaches combine the two last approaches. We talk about linguistic and statistical approaches to disambiguate the words in their morphological level. The approach of Tlili-Guiassa (2006) is based on the MBL approach (Memory based learning). It analyzes grammatical and inflectional affixes and grammatical rules. This approach is used to classify a collection of Quranic and educational texts. (Zribi et al., 2006) combine rule-based approach with a trigram HMM tagger. Texts with 6000 words are used to train the trigram classifier and heuristic rules were applied to select from the proposed results.

According to (Hoceini et al., 2011), the linguistic approaches for disambiguation are more reliable than the statistical approaches. The linguistic approaches, which are based on a specific set of rules, require only a linguist to define these rules. While for the statistical approaches, the statistics calculated for training are the same used for any test domain. The training phase is necessary for both statistical and hybrid approaches in order to learn the required settings for disambiguation.

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS

The hybrid approaches combine the linguistic and statistical approaches so they take advantage of each approach. As a result, they are considered as the most effective and coherent in terms of analysis.

4 THE POSSIBILISTIC MOPHOLOGICAL DISAMBIGUATION APPROACH

We present the possibilistic disambiguation approach based on the possibility theory (Zadeh, 1978). This approach was proposed by Bounhas et al. (2015). The possibilistic theory is unlike the probability theory. It discerns between uncertainty and imprecision which describe an incomplete information (Ayed et al., 2012).

Imprecision is revealed when a reality state is defined by variables with multi-values. Uncertainty appears when we are unable to provide or to know a statement to determine the real value of a proposal (Dubois and Prade, 2010).

The main idea of the possibilistic approach is to provide a possibilistic classifier (Haouari et al., 2009) that obtains disambiguation knowledge from vocalized texts and tests on non-vocalized texts. Training and test phases involve instances that use classification attributes and classes. Each instance is associated to a word *w*. For the possibilistic disambiguation approach, classes are the 14 morphological features (MF) presented by MADA (Roth et al., 2008). They are, exclusively, POS, conjunction, particle, determiner, pronoun, person, voice, aspect, gender, number, case, preposition, mode and adjective (Ayed et al., 2012). Each morphological feature may have one or more possible values. The training and the test attributes designate the morphological features (MF) of the two following and the two preceding words of *w* (MF±i, $i \in \{1, 2\}$). For instance, we get the attributes POS-2, POS-1, POS+1 and POS+2 which designate the POS (Part-Of-Speech) of the two previous and following words of *w*. POS may be the class of *w*.

Every instance, in the training set, is related to a vocalized word. Table 1 gives an example of a training set where the training attributes are (POS±i, i=1) and the class is POS. Table 2 presents an example of a testing set that corresponds to a non-vocalized word whose POS is ambiguous and needs to be determined. The attributes of the test set must be the same of the training set.

In table 1 and table 2, both training and test sets are imperfect since they contain uncertain and imprecise instances. In fact, the second instance, of the training set, is uncertain as it gives more than one class' value. The first instance, of the training set, is imprecise since it contains attribute that provide more than just one value; i.e. POS+1. The test instance is also imprecise. Indeed, its POS+1 attribute has two possible values. The possibility theory deals with the problem of imperfection of training and test instances. Arabic vocalized words are less ambiguous but, in some cases, they give ambiguous values which justify the imperfection of particular instances of the training sets.

Disambiguation consists in identifying the right values, of the morphological features (POS, conjunction, case, etc.), among values given by a morphological analyzer.

Table 1: Example of instances from a training dataset.

Word	POS-1	POS+1	POS (class)
شَرِبَ	VERB_ PERFECT	{INTERROG_ PART; REL_ PRON}	VERB_ PERFECT
خَلِيل	VERB_ PERFECT	VERB_ PERFECT	{NOUN; NOUN-PROP}

Table 2: Example of an instance from a test dataset.

Word	POS-1	POS+1	POS (class)
وقف	VERB_ PERFECT	{INTERROG_ PART; REL_ PRON}	?

4.1 The Training Phase

To prepare training datasets, we analyze vocalized texts and rearrange instances as shown in table 1. Morphological analysis tools identify the different features of a given word out of its context (Hajic, 2000). Possibilistic disambiguation approach uses the updated version of BAMA 1.2.1 (Buckwalter, 2004), AraMorph that treats vocalized texts. For each AraMorph identifies the prefixes, the suffixes and the stem of each word and associates values of the 14 morphological features (Habash et al., 2009; Roth et al., 2008). Possibilistic disambiguation uses a possibilistic classifier where the class MF is one of the morphological features (MF ϵ {POS, conjunction, particle, determiner, pronoun, person, voice, aspect, gender, number, case, preposition, mode and adjective}) and the attributes are the set of (MF±i, $i \in \{1, 2\}$). We denote c_i a value of a class MF and a_{jL} a value of an attribute MF±i (that we note A_j). In the training phase, we compute the frequency

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS

measure of an attribute value (*a_{jL})* for a class *c_i* (Elayeb et al., 2009). We determine this measure for each couple (attribute value, class) in the training set. The frequency is computed as follows:

$$Freq\,(a_{jL}, c_i) = \frac{Occ(a_{jL}, c_i)}{Max_{L=1}^{|\mathrm{Aj}\,|} Occ(a_{jL}, c_i)} \qquad (1)$$

Where $Occ(a_{jL}, c_i)$ is the number of instances that have the class c_i and a_{jL} values for the attribute *A_j*. $\left|\mathrm{A_j}\right|$ is the number of possible values of *A_j*. The *Max* operator is used to normalize the frequency (Elayeb et al., 2009).

$$Occ(a_{jL}, c_i) = \sum_{k=1}^{|T|} \frac{\emptyset_{\mathrm{ijkL}}}{|A_{jk}| * |C_k|} \qquad (2)$$

T is the training set, |T| is the number of instances in the training set, *I_k* is an instance from T. $|A_{jk}|$ is the number of the attribute *A_j* values in *I_k* and $|C_k|$ is the class values of *I_k*. The product $|A_{jk}|*|C_k|$ decodes the imperfection of an instance. In fact, if the instance *I_k* is imprecise than one or more of its attributes have more than one value. Consequently, $|A_{jk}|$ is not equal to 1 for the imprecise attribute *A_{jk}* and $|C_k|$ is not equal to 1 for the uncertain class. Hence, if the instance is perfect, than the product is equal to 1. *$\emptyset_{ijkL}$* is equal to 1 if the value a_{jL} belongs to the possible values of *A_j* in the instance *I_k* and *c_i* is one of the possible classes of *I_k*. If not, *$\emptyset_{ijkL}$* is equal to 0.

4.2 The Test Phase

To prepare test datasets, we analyze non-vocalized texts and rearrange instances as shown in table 2. Morphological disambiguation of a non-vocalized word consists in determining the accurate values of its morphological features. In other terms, morphological disambiguation determines the class value of the test instance.

We reduce the ambiguity of some Arabic words by including linguistic rules defined in (Diab et al., 2007). Linguistic rules decrease the number of the test instances. We present each rule as a function of the attributes used by the possibilistic classifier. For instance, we experiment the rule "A preposition cannot follow a preposition" by: if (POS-1="PREP") then POS≠"PREP". Hence, we remove the "PREP" value from the possible POS values. For the remaining ambiguous test instances, we use the possibilistic morphological disambiguation approach to select the morphological feature value.

The possibilistic morphological disambiguation approach proposes measures to distinguish the accurate class; i.e. necessity and possibility (Bounhas et al., 2015). Each of these measures is computed, over the training set, for each possible value of the class, to determine among the accurate value. A morphological feature's value is accorded, to a non-vocalized word, if it has the maximum value of possibility and necessity.

The possibility and the necessity measures are given, respectively, by the formulas (3) and (4).

$$\Pi(\mathrm{c_i}|\mathrm{I_k}) = \prod_{\mathrm{j=1}}^{\mathrm{m}}\prod_{\mathrm{L=1}}^{|\mathrm{A_{jk}}|} Freq\,(a_{jL}, c_i) * \beta_{\mathrm{jk}} \qquad (3)$$

$$N(c_i|\mathrm{I_k}) = 1 - \prod_{j=1}^{\mathrm{m}}\prod_{\mathrm{L=1}}^{|\mathrm{A_{jk}}|}\left(1 - \frac{\lambda\mathrm{ijL} * Freq\,(a_{jL}, c_i)}{\beta_{\mathrm{jk}}}\right) \qquad (4)$$

We denote the test instance *I_k*. *m* is the number of test attributes. *c_i* is a class value. *A_j* is a test attribute and *$|A_j|$* is the number of the attribute *A_j* values in the instance *I_k*. *a_{jL}* is a value of *A_j*. The frequencies $Freq\,(a_{jL}, c_i)$ are previously calculated over the training set. *λ*ijL is equal to *$log_{10}(P/nc_{jL})$* where *P* is the number of possible class values and *nc_{jL}* is the number of the class values having a non-null frequency with the attribute *a_{jL}* i. e. *Freq (a_{jL}, c_i)*≠0. β_{jk} is a factor that we add for imprecise attributes. In fact, if an attribute has 4 possible values, we compute the product of the frequencies of these 4 values and we multiply (for the possibility) or divide (for the necessity) each of these frequencies by 1/4 (β_{jk}). The selected class value of the test instance *I_k* corresponds to the value $c*$. It is the value that has the highest score among all class values:

$$\mathrm{c}^* = \arg\max_{\mathrm{c_i}} \left(\Pi(\mathrm{c_i}|\mathrm{I_k}) + \mathrm{N}(\mathrm{c_i}|\mathrm{I_k})\right) \qquad (5)$$

The possibilistic approach is considered as a hybrid approach as it combines possibilistic measures with linguistic rules.

5 THE HADITHS ARABIC CORPUS

The hadiths Arabic corpus represents Islamic texts, said by the prophet Mohamed, which discusses several real-life concerns. Hadiths are written in classical Arabic with vocalized words. This corpus was the topic of some research fields as knowledge extraction, texts classification (Harrag et al., 2009; Harrag et al., 2013) and information retrieval (Bounhas et al., 2010; Bounhas et al., 2011). Ben

SCITEPRESS
SCIENCE AND TECHNOLOGY PUBLICATIONS

Khiroun et al. (2014) attempt to create, from the hadiths texts, a standard test collection for Arabic information retrieval.

The corpus of hadiths is one from the rare vocalized Arabic corpora. It contains about 2.5 million of words dispersed on more than 100 books of hadiths. The most known and used books are Sahih Muslim, Sahih Al Bukhari, Sunan Annasaii, Sunan Ibn Majah, Sunan Abi Dawud, and Sunan Ettermidhi (Al-Echikh, 1998). Besides, the corpus is entirely structured into chapters and sub-chapters which provide an appropriate source of contextual data (Bounhas et al., 2011). Hadiths corpus can be studied along several social and semantic axes which make it a reliable resource for information retrieval and knowledge extraction.

5.1 Social Aspect in Hadiths

Each hadith is composed of two parts (i) the Sanad (السند) or the chain of narrators through which the hadith was transmitted and (ii) the Matn (المتن) or the text told by the prophet. An example of a hadith is given by figure 1.

> Mussad told us from Yahia from Shu’bah from Qatadah from Anas that the prophet said: None of you will be a true believer until he loves for his brother that which he loves for himself. "Al Bukhari recited it ".
> حَدَّثَنَا مُسَدَّدٌ قَالَ حَدَّثَنَا يَحْيَى عَنْ شُعْبَةَ عَنْ قَتَادَةَ عَنْ أَنَسٍ عَنْ النَّبِيِ قَالَ: لاَ يُؤْمِنُ أَحَدُكُمْ حَتَّى يُحِبَّ لِأَخِيهِ مَا يُحِبُّ لِنَفْسِهِ. "رَوَاهُ البُخَارِي”

Figure 1: An example of a vocalized hadith.

The Sanad is represented by حَدَّثَنَا مُسَدَّدٌ قَالَ حَدَّثَنَا يَحْيَى عَنْ شُعْبَة عَنْ قَتَادَة عَنْ أَنَسٍ عَنْ النَّبِيِ قَالَ and the Matn gives the replied text لاَ يُؤْمِنُ أَحَدُكُمْ حَتَّى يُحِبَّ لِأَخِيهِ مَا يُحِبَ لِنَفْسِهِ. The narrators are separated by the word عن (En ; from). The verb قال (qAl ; said) announces the beginning of the text transmission. The expression رواه البخاري (Al Bukhari recited it) means that the hadith exists in "Sahih Al Bukhari". The sheikh (شيخ; $yx) is an Islamic scholar. The sheikh Al Bukhari learned the hadith text from his sheikh Musadd (مُسَدٌّ) who heard it from his sheikh Yahya (يَحْيَى). The latter learned the text from his sheikh Shu’bah (شُعْبَة) who heard it from Qatadah (قَتَادَة) who was a disciple of one of the prophet companions Anas (أَنَسٍ). Anas transmitted the hadith text from the prophet.

Social relationships can be extracted from the Matn and the Sanad. We can identify the sheikh relationship commonly through the word عن. Also, family relationships can be identified, for instance, by the words إبن (<bn; son of) or أبو (>bw; father of). A narrator can also report what his brother said using, for example, حدّثني أخي (my brother told me) or what his grandfather said using, for example, حدّثني جدّي (my grandfather told me). The narrators in the Sanad can be related through other social attributes such as place of residence and communities’ membership.

5.2 Semantic Aspect in Hadiths

The hadiths’ books are structured by theme, except for some books that are arranged by narrators. Descriptive data of several hadiths were added by the scholars to facilitate the understanding of the texts.

The hadiths are structured in XML format in such a way that they are classified according to specific themes. A hadith can cover a multitude of themes and a huge amount of information in the Matn as well as in the Sanad. From a hadith, we can extract information from the knowledge conveyed by the content or by the titles of the chapters and the sub-chapters and even the comments provided by the narrators and scholars of the hadith. The XML tags add semantic information to the hadiths. In figure 2, we present an example of a structured XML file from Al Bukhari book. The <THEME> tag indicates the theme. <S> designates the section where the hadith appears. The <S> element is followed by all its hadiths tags. <DOC> describes a hadith where <R> indicates its Sanad and <TEXT> includes its Matn. The hadith’s theme is that of its section.

```
<BOOK id="146">
…
<S l="1">
        <THEME> بَاب بَدْءُ الْوَحْيِ </ THEME>
        …
</S>
<DOC>
        <DOCID>(1)-[1]</DOCID>
        …
        <R ID="4698" S="عبد الله بن الزبير بن عيسى بن
    عبيد الله بن أسامة بن عبد الله بن حميد بن زهير بن الحارث
    بن أسد بن عبد العزى" TP="F">
         الْحُمَيْدِيُّ عَبْدُ اللَّهِ بْنُ الزُّبَيْرِ
        </R>
        …
        <TEXT>    إِنَّمَا الْأَعْمَالُ بِالنِّيَّاتِ، وَإِنَّمَا لِكُلِّ امْرِئٍ مَا
    نَوَى، فَمَنْ كَانَتْ هِجْرَتُهُ إِلَى دُنْيَا يُصِيبُهَا أَوْ إِلَى امْرَأَةٍ
    "يَنْكِحُهَا، فَهِجْرَتُهُ إِلَى مَا هَاجَرَ إِلَيْهِ
        </TEXT>
</DOC>
```

Figure 2: An example of a structured XML hadith.

6 POSSIBILISTIC MORPHOLOGICAL DISAMBIGUATION OF HADITHS

Many research works in information retrieval are carried out on hadiths texts (Harrag et al., 2009; Jozi et al., 2012). Morphological disambiguation is crucial to get pertinent results of an information retrieval system (Bessou and Touahria, 2014). It's used to disambiguate both documents and queries which are written, commonly, in non-vocalized texts.

We take advantage of the hadiths structure and their provided semantic data to disambiguate a non-vocalized hadith text using the possibilistic approach. To do, we perform a morphological analysis of the hadiths texts (an example is given in figure 2) using a morphological analyzer to determine the various values of the 14 morphological features. The analysis gives the different possibilities without affecting the exact value of a morphological attribute.

We add other classification attributes to provide more sense to morphological data. Classification attributes include:

- A semantic attribute that designates the theme of the analyzed word. In fact, if the word belongs to a hadith text, than its theme is that of this hadith. In other terms, a word that appears in a <DOC> element gets the theme of the hadith presented by this <DOC> element.
- An attribute, that we denote "dimension", which indicates whether a word belongs to a Sanad or a Matn. In fact, if a word appears in the <R> element content, than it belongs to Sanad and if a word appears in the <TEXT> element content, than it belongs to Matn.
- The morphological attributes (i.e. MF±i, i∈{1, 2}) used by the possibilistic morphological disambiguation approach.

Training and test instances use the same classification attributes. Table 3 shows an example of an ambiguous non-vocalized word. To simplify the example, we use only the POS-1 and POS+1 morphological attributes. The non-vocalized word إبن (<bn) is ambiguous even if we add short vowels. In fact, the vocalized word إِبْنِ (<ib°ni) can mean (i) "son of" which is a noun and (ii) "build" which is a verb. Thus, the possible POS values set is {VERB_IMPERATIVE; NOUN}. The theme given to this word is الإيمان (Al<ymAn; the faith). This word is a part of Sanad. In a narrative context, the word إبن cannot be an imperative verb. Based on the necessity and the possibility measures computed over the training set, the class NOUN gives the maximum sum value.

Table 3: Example of a test instance using morphological and semantic attributes.

Word	POS-1	POS+1	theme	Dim-ension	POS (class)
إبن	NOUN_PROP	NOUN_PROP	الإيمان	Sanad	?

7 EXPERIMENTAL RESULTS

We describe, in this section, the evaluation method used to experiment the possibilistic disambiguation approach. We present, also, disambiguation results of hadiths using morphological and semantic information. We study variation effect of the analyzers on the disambiguation results.

7.1 The Evaluation Method

To assess the performance of the possibilistic approach on the hadiths disambiguation, we use the 5-fold cross-validation. In other terms, 80% of the corpus texts are used for training and 20% are used for test. We obtain five possible combinations. The training texts are vocalized. We omit vowels of the test set. Hence, we obtain non-vocalized texts. We compute the average success rates from all the 4+1 combinations. To get these rates, we analyze the vocalized texts and we save their morphological solutions. Then, we omit the short vowels of the same texts. Finally, we apply the possibilistic classifier to disambiguate these texts and we save the results. If the two obtained results of a word are similar, then this word is correctly classified.

7.2 Comparing Morphological Analyzers for Possibilistic Disambiguation

We propose to vary several morphological analyzers in order to determine the best one for the possibilistic disambiguation approach. We disambiguate about 10000 words from hadiths texts using (i) Aramorph, (ii) BAMA analyzer of MADAMIRA (Pasha et al., 2014) and (iii) AlKhalil analyzer (Bousmaha et al., 2013). Table 4 illustrates the results of the possibilistic disambiguation approach using these analyzers. The results are given for the POS morphological feature. The

SCITEPRESS
SCIENCE AND TECHNOLOGY PUBLICATIONS

disambiguation based on AlKhalil gives the highest rate followed by the MADAMIRA analyzer. Aramorph gives a respectable disambiguation rate but it doesn't overcome Alkhalil and MADAMIRA. Hence, the morphological disambiguation approach depends, closely, on the analyzer used to prepare training and test instances.

Table 4: Disambiguation rates using different morphological analyzers.

Analyzer	Aramorph	MADAMIRA	AlKhalil
disambiguation rate	93.93%	94.09%	94.67%

7.3 Evaluating the Possibilistic Disambiguation Approach

Morphological disambiguation consists in determining the accurate values of the morphological features. We perform experiments on the book Al Bukhari using AlKhalil analyzer to provide classification instances. The disambiguation rates of hadiths' texts give 76.36%, 76.46% and 76.23% respectively using Decision Tree, SVM and Naïve Bayesian classifiers (Bounhas et al., 2015). Comparing to these rates, our approach using the possibilistic classifier gives better results. In fact, the possibilistic morphological disambiguation approach provides an average rate of 86.37% for the 14 morphological features using only the morphological attributes MA (i.e. MF±i, i∈{1, 2}) for classification. The average increases by 0.29% if we use both the morphological attributes and the semantic attributes (SA) in classification (see section 6). We notice that improvement does not affect all the morphological features. Indeed, only the disambiguation rates of POS, aspect, adjective, conjunction, gender, person, preposition and pronoun increase. This is explained by the fact that adding semantic attributes, for these features, reduces the training set size. The possibility and the necessity measures computing over the training instances give high results of the correct classes. For instance, for the POS class, some values of morphological attributes, such as NOUN_PROP (proper noun), appear most in narrators chains (Sanad). Thus, an ambiguous word that contains NOUN_PROP in its possible values of POS is probably a proper noun rather than another value of the POS class. Similarly, many Arabic proper nouns can have the same meaning as an adjective. We note, for example, "أسعد" (>sEd; more happy) or "جميل" (jmyl; beautiful). This fact may justify the enhancement of the disambiguation rate of the adjective feature using semantic attributes. Thus, the addition of semantic attributes reduces the number of possible values in a particular context. Semantic attributes restrict and filter the training sets which increase the possibility and necessity measures giving support to the accurate class.

Table 5: Possibilistic morphological disambiguation rates using morphological attributes and semantic attributes.

MF	Disambiguation rates (MA)	Disambiguation rates (MA & SA)
POS	95.13%	96.01%
ADJECTIVE	99.17 %	99.27%
ASPECT	81.53 %	81.58%
CASE	63.52 %	63.52%
CONJUNCTION	91.07 %	91.12%
DETERMINER	97.02 %	97.02%
GENDER	96.55 %	96.66%
MODE	99.96%	99.96%
NUMBER	93.10 %	93.10%
PARTICLE	98.88 %	98.88%
PERSON	66.06 %	67.02%
PREPOSITION	88.27 %	90.02%
VOICE	79.11 %	79.11%
PRONOUN	59.81 %	59.83%
Average	86.37 %	86.66%

8 CONCLUSION

We presented, in this paper, the possibilistic morphological disambiguation approach that uses morphological and semantic attributes to disambiguate 14 morphological features of hadiths texts. Disambiguation consists in choosing the accurate morphological feature value from the solutions proposed by a morphological analyzer. We proved that the possibilistic approach gave the best rates using AlKhalil analyzer. The possibilistic disambiguation rates that use morphological and semantic attributes are better than those that use only morphological attributes. Experiments showed an encouraging improvement of the possibilistic approach to deal with classical Arabic texts. We aim, as a future work, to improve the performance of an information retrieval system by presenting a queries and documents disambiguation phase. We consider that our work is a first step toward building an information retrieval system which treats both vocalized and non-vocalized documents and focuses on hadiths Arabic texts.

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS

REFERENCES

Attia, M., 2008. *Handling Arabic Morphological and Syntactic Ambiguity Within the LFG Framework with View to Machine Translation*. PhD Thesis, University of Manchester.

Ayed, R., Bounhas, I., Elayeb, B., Evrard, F., Bellamine Ben Saoud, N., 2012b. Arabic Morphological Analysis and Disambiguation Using a Possibilistic Classifier. *In: Proceedings of the 8th International Conference on Intelligent Computing (ICIC)*, July 25-29, 2012 (pp. 274–279). Springer Berlin Heidelberg.

Bessou, S. and Touahria, M., 2014. An Accuracy-Enhanced Stemming Algorithm for Arabic Information Retrieval. *Neural Network World Journal*, 24(2):117–128.

Bounhas, I., Elayeb, B., Evrard, F., Slimani, Y., 2011. Organizing Contextual Knowledge for Arabic Text Disambiguation and Terminology Extraction. *Knowledge Organization*, 38(6), 473–490.

Bounhas, I., Ayed, R., Elayeb, B., Bellamine Ben Saoud, N., 2015. A Hybrid Possibilistic Approach for Arabic Full Morphological Disambiguation. *Data & Knowledge Engineering*, vol. 100,Part B, pp. 240-254.

Bousmaha, K. Z., Abdoun, S. C., Belguith, L. H. et Rahmouni, M. K., 2013. Une approche de désambiguïsation morpho_lexicale évaluée sur l'analyseur morphologique Alkhalil. www.webreview.dz, 20(2): 32–46.

Buckwalter, T., 2004. Buckwalter Arabic Morphological Analyzer Version 2.0. Online Publication, Linguistic Data Consortium (LDC) catalog number LDC2002L49, ISBN 1-58563-257-0, from http://www.nongnu.org/aramorph/

Diab, M., Hacioglu, K., Jurafsky, D., 2007. Automated Methods for Processing Arabic Text: From Tokenization to Base Phrase Chunking. *In van den Bosch, A. and Soudi, A., editors, Arabic Computational Morphology: Knowledge-based and Empirical Methods. Kluwer/Springer*.

Dubois, D., Prade, H., 2010. Formal Representations of Uncertainty. *In D. Bouyssou, D. Dubois, M. Pirlot, & H. Prade (Eds.), Decision-making Process* (pp. 85–156). ISTE &Hoboken, London, UK, Wiley, USA.

Elshafei, M., Al-Muhtaseb, H., Al-Ghamdi, M., 2002. Techniques for high quality Arabic speech synthesis. *Information Sciences* 140(3), 255-267.

Elayeb, B., Evrard, F., Zaghdoud, M., Ben Ahmed, M., 2009. Towards an intelligent possibilistic web information retrieval using multiagent system. *Interactive Technology and Smart Education*, 6(1), 2009, 40–59.

Habash, N., Rambow, O., Roth, R., 2009. 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)*, Cairo, Egypt, April 22-23, pp. 102-109.

Habash, N., Roth, R., Rambow, O., Esk, R. et Tomeh, N. (2013). Morphological Analysis and Disambiguation for Dialectal Arabic. In Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies(NAACL-HLT), Atlanta, GA.

Hajic, J., 2000. Morphological Tagging: Data vs. Dictionaries. *In: Proceedings of the 1st North American Chapter of the Association for Computational Linguistics Conference* (pp. 94–101). Stroudsburg, PA, USA: Association for Computational Linguistics, April 29- May 24, 2000.

Haouari, B., Ben Amor, N., Elouedi, Z., Mellouli, K., 2009. Naïve possibilistic network classifiers. *Fuzzy Sets and Systems*, 160(22), 3224–3238.

Harrag, F., Hamdi-Cherif, A., Malik, A., Al-Salman, S. and El-Qawasmeh, E., 2009. Experiments in improvement of Arabic information retrieval. *In proceedings of the 3rd International Conference on Arabic Language Processing*, Rabat, Morocco.

Hoceini, Y., Cheragui, M. A., Abbas, M., 2011. Towards a New Approach for Disambiguation in NLP by Multiple Criterian Decision-Aid. *The Prague Bulletin of Mathematical Linguistics* 95, 19-32.

Jozi, H., Zadeh, A. R., Barati, E. et Minaei, B, 2012. A new framework for detecting similar texts in Islamic Hadith Corpora. *In The International Conference on Language Resources and Evaluation*, Istanbul, Turkey.

Pasha, A., Al-Badrashiny, M., Diab, M., El Kholy, A., Eskander, R., Habash, N., Pooleery, M., Rambow, O. and Roth, R. (2014). MADAMIRA: A Fast, Comprehensive Tool for Morphological Analysis and Disambiguation of Arabic. *In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland.

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 the Association for Computational Linguistics conference (ACL)*, Columbus, Ohio, USA, June 19-20, 2008

Tlili-Guiassa, Y., 2006. Hybrid Method for Tagging Arabic Text. *Journal of Computer Science* 2 (3): 245-248.

Vapnik, V. N., 1999. An overview of statistical learning theory. *IEEE Transactions on Neural Networks*, 10(5), 988–999. DOI:10.1109/72.788640

Zadeh, L.A., 1978. Fuzzy Sets as a Basis for a Theory of Possibility. Fuzzy Sets and Systems, 1, 3-28, 1978.

Zribi, C., Torjmen, A., Ben Ahmed, M., 2006. An Efficient Multi-agent System Combining POS-Taggers for Arabic Texts. *In Proceedings of 7th international conference of Computational Linguistics and Intelligent Text Processing, LNCS* Volume 3878, Springer, 121-131.