




Morphological Disambiguation of Texts Based on Analogical Proportions

Bilel Elayeb^{1,2}^a, Myriam Bounhas³^b and Mohamed Firas Ettih⁴^c

¹RIADI Research Laboratory, ENSI, Manouba University, Tunisia

²LARODEC Research Laboratory, ISG Tunis, Tunis University, Tunisia

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

Keywords: Morphological Disambiguation, Analogical Proportions, Machine Learning Algorithms, Deep Learning Algorithms, Feature Selection, Classification.

Abstract: The Arabic language is known for its complexity, which encompasses extensive morphological and orthographic variations, as well as significant syntactic and semantic diversity. These unique characteristics often result in morphological ambiguity in Arabic. In this paper, we tackle the challenge of morphological disambiguation in Arabic texts. We frame this task as a classification problem, where the possible values of morphological features represent the classes, and a classification algorithm is used to assign the appropriate class to each word based on its context. Specifically, we investigate the effectiveness of an analogy-based classifier for morphological disambiguation in Arabic texts. Analogical Proportions (AP) are statements that express the relationship between four elements A, B, C, and D such that "A differs from B as C differs from D". Leveraging Analogical Proportions-based inference, the AP classifier predicts the fourth, unknown element (D), given that the first three (A, B, and C) are known. We evaluate this analogical classifier using a corpus of Classical Arabic texts. The average disambiguation rate (74.80%) of the AP classifier outperforms that of a set of well-established machine-learning and deep learning-based classifiers.

1 INTRODUCTION


Morphological disambiguation in the Arabic language involves analyzing and clarifying ambiguities in the structure and meaning of words by determining their accurate morphological attributes, such as root, stem, part of speech, and grammatical features (e.g., gender, number, and tense). Given Arabic's complex morphology, this process is essential for natural language processing (NLP) tasks (Elayeb et al., 2009; Elayeb and Ben Khiroun, 2023), including machine translation, text-to-speech systems, and information retrieval.


Arabic Morphological Disambiguation (AMD) still faces several challenges, such as: (i) morphology: Arabic words often include prefixes, suffixes, and infixes that convey various grammatical details, (ii) Complex word structures: words can


exhibit intricate combinations of affixes, clitics, and roots, which result in more difficult analysis, (iii) absence of vowels: written Arabic, particularly Modern Standard Arabic, frequently omits diacritical marks, which complicate the identification of a word's correct form, and (iv) ambiguity: the context-sensitive nature of Arabic leads to multiple possible morphological interpretations for many words.

Morphological disambiguation poses a significant challenge for languages rich in morphology and ambiguity, such as Arabic. A non-vocalized Arabic word can have more than 12 possible interpretations (Elayeb and Bounhas, 2016; Elayeb, 2019). For example, the non-vocalized Arabic word "ضرب" can be interpreted as a verb meaning "to strike" (ضَرَبَ) or as a noun meaning "a strike" (ضَرْبٌ).

Some Arabic words are homographs, meaning they are written identically but have different

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

^b <https://orcid.org/0000-0001-8320-6722>

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

meanings. Processing these words depends on recognizing morphemes, which form the basic units of meaning.

The purpose of morphological analysis is to determine the various criteria or functions of each lexical unit or word. Examples of these criteria include Part-Of-Speech (POS), which identifies whether a word is a noun, verb, or particle, as well as features such as number, gender, and information about clitics. Ambiguity arises during the analysis, especially when contextual clues do not align with a word's intended interpretation.

This paper explores a method for addressing morphological ambiguity in Arabic texts. We approach this task as a classification problem, where the possible morphological feature values represent the classes, and a classification algorithm determines the correct class for each word based on its context. Specifically, we investigate the efficiency of an analogy-based classifier for AMD. This type of classifier has recently demonstrated its effectiveness in classifying Arabic texts (Bounhas et al., 2024).

We begin by introducing Analogical Proportions (AP), which establish a relationship between four elements, A , B , C , and D , such that " A is to B as C is to D ". Then, based on the analogical inference, the AP classifier proposed by Bounhas et al. (2017), that we use, can predict the fourth, unknown element D when the first three elements (A , B , and C) are provided. This algorithm has not yet been tested before in the context of Arabic text morphological disambiguation. We evaluate the performance of this analogical classifier on a corpus of Classical Arabic texts and compare it to both ML and DL classifiers.

The remainder of this paper is organized as follows: Section 2 provides a summary and discussion of existing works on morphological disambiguation in Arabic texts. Section 3 summarizes a background on analogical proportions. In Section 4, we describe the proposed analogy-based classifier for AMD. Section 5 presents the experimental results and includes a comparative analysis of the AP classifier with both ML-based and DL-based classifiers. Finally, Section 6 concludes the study and suggests directions for future research.

2 RELATED WORKS

Many studies have reduced ambiguity by identifying the POS feature of Arabic text. POS attribute disambiguation is the process of determining the grammatical class of a word in a particular context. Morphological disambiguation is known as a

classification problem in which a set of POS values represents a class, and a single classification method is used to assign each word occurrence to a class based on sentence context.

One of the most important steps in disambiguation is choosing an appropriate classification method. Several automatic classification methods have been used, leveraging ML techniques to train classifiers from sentences annotated with POS values. In literature, disambiguation methods are commonly organized into three categories: (i) rule-based methods, (ii) statistical methods, and (iii) hybrid methods that combine these two approaches, and (iv) ML and DL-based approaches.

Rule-based or linguistic methods use a set of rules created by linguists to assign labels to different morphological attributes. Several WSD systems have been described by Daoud (2009), mainly involving heuristic, textual, and non-textual rules. Besides, these rules typically fall into grammatical, structural, and logical classes. Daoud and Daoud (2009) proposed a specialized parser called "EnConverters," developed in UNL and a rule-based programming language. The Daouds defined several types of disambiguation rules that merge syntactic and morphological context associations.

Statistical methods build one or more learning models from annotated corpora. These methods often utilize statistical models such as the Hidden Markov Model, which assumes a Markov process with unknown parameters. Classification methods such as SVM calculate the probabilities for each grammatical class of a word. Moreover, MADA is a tool designed by Habash and Rambow (2007) and is widely used to resolve morphological ambiguities in Arabic texts. Then, MADAMIRA (Pasha et al., 2014) combines MADA and the Arabic tokenizer AMIRA (Diab, 2007).

A hybrid method combines statistical information with linguistic rules to address morphological ambiguity. Khoja (2001) is among the researchers who adopted this approach, implementing it with the Viterbi algorithm. She calculated two probabilities from an annotated corpus of 50,000 words: (i) contextual probability, the likelihood that one label precedes or follows another, and (ii) lexical probability, the likelihood that a word has a certain morphological attribute. Based on these statistics, a set of grammatical rules was developed, achieving an accuracy of over 90% (Khoja, 2001).

Belguith and Chaâben (2006) also proposed a method for morphological analysis and disambiguation, categorized as a statistical approach that incorporates rule-based elements. This method

involves five steps (Ayed et al., 2018b): (i) text segmentation into words, (ii) morphological preprocessing by removing clitics based on a predefined list, (iii) affixal analysis, distinguishing the root and affixes of each word, (iv) morphological analysis using MORPH2, and (v) post-processing to group words using lexicons and a rule set. This method calculates the morphological attributes of each word using a term dictionary.

Bousmaha et al. (2016) proposed a hybrid disambiguation approach focusing on the selection of diacritical marks across different analysis levels. This method combines multi-criteria decision-making with a linguistic approach and offers an alternative solution for morphological ambiguity. For evaluation, the authors used an online shape analyzer, achieving promising results with F-measures above 0.8. Bounhas et al. (2015a) proposed three possibilistic classifiers for AMD: (i) the first classifier relies on the possibility measure, (ii) the second on the necessity measure, and (iii) the third combines these two measures. They enhanced these classifiers with information gain scores, serving as weights for classification attributes, to optimize space requirements for resolving contextual ambiguity and thereby simplify the disambiguation process.

Later, Bounhas et al. (2015b) introduced a hybrid possibilistic approach that integrates the possibilistic classifier with linguistic rules to assign labels to various morphological attributes, which improved Arabic text disambiguation rates. They also addressed “out-of-vocabulary” words lacking known morphological analysis. These possibilistic and hybrid classifiers were evaluated on the Arabic “Kunuz”⁴ dataset and compared with three machine-learning (ML) classifiers: SVM, Naïve Bayes, and Decision Tree.

Besides, Ayed et al. (2018b) explored possibilistic morphological disambiguation for structured Hadith texts in Arabic, incorporating semantic knowledge. Using AIKhalil analyzer, they conducted training and testing of morphological attributes, leveraging the XML format of “Kunuz” Hadith texts to integrate available semantic information. By including semantic attributes in their possibilistic classifiers, they achieved improved disambiguation rates in their experiments.

More recently, Elayeb et al. (2022) experimented with a range of ML algorithms to address the challenge of morphological disambiguation in Arabic texts using various morphological features. The authors evaluated these algorithms on the Kunuz test collection (Ben

Khiron et al., 2012a; 2014; Ayed et al., 2018ab), which comprises classical, vocalized Arabic texts, specifically Hadith attributed to the Prophet Muhammad (PBUH). This corpus has attracted considerable research interest due to its linguistic, semantic, and social depth, as well as its structured organization. By contrast, the TREC collections (2001 and 2002), which include Arabic newspaper articles, lack vowel markers in their texts, contributing to potential word sense ambiguities and difficulties in determining POS tags and syntactic roles.

Moreover, Analogical Proportions, first developed by Prade et al. (2010), have demonstrated their efficiency when applied in several domains, such as information retrieval (IR) (Bounhas and Elayeb, 2019), Arabic text summarization (Elayeb et al., 2020), Arabic text classification (Elayeb et al., 2023; Bounhas et al., 2024), as well as the classification of structured data (Bounhas et al., 2017; Bounhas and Prade, 2023; 2024). This paper aims to investigate the performance of such proportions in disambiguating Arabic text data collections.

3 ANALOGICAL PROPORTIONS

An analogical proportion (AP) is a relationship between 4 items $A, B, C, D \in X$. This relationship states that “ A differs from B as C differs from D ” which enables us to compare the pair (A, B) to the pair (C, D) (in terms of similarities and differences). An Analogical proportion is usually denoted $A : B :: C : D$. This proportion also holds when “ $A : C :: B : D$ ” (by central permutation), “ $A : B :: A : B$ ” (by reflexivity) and “ $C : D :: A : B$ ” (by symmetry) (Prade and Richard, 2010).

Boolean Setting: In the Boolean context, there are only six valid valuations where “ $A : B :: C : D$ ” holds true or “ $A : B :: C : D = 1$ ”. These six valid valuations, or six possible assignments, are the only configurations among the 16 possible configurations that make an analogical proportion true. In particular, a 4-tuple (A, B, C, D) is in analogical proportion if it’s in one of those particulars’ assignments: $(0,0,0,0)$ $(1,1,1,1)$ $(0,0,1,1)$ $(1,1,0,0)$ $(0,1,0,1)$ $(1,0,1,0)$. As can be seen, the proportion remains valid for these six patterns even when items are negatively coded.

Nominal Extension: In the nominal case, the analogical proportion can only be true in only three possible assignments out of the six. Otherwise, it is

⁴ <https://github.com/bounhasibrahim/Kunuz/tree/Kunuz-IR>

no longer true. $A : B :: C : D$ is valid if it is of the form: (r, r, r, r) or (r, s, r, s) or (r, r, s, s) such that: $r \neq s$.

Multiple-Valued Extension: To handle numerical attributes the domain has to be in the interval $[0,1]$, which means that the analogical proportion $A : B :: C : D$ linking the 4 items A, B, C and D takes values in the interval $[0,1]$ (Dubois et al., 2016). It is then a matter of obtaining a high or low AP value depending on whether the values are closer to 1 or 0. For examples:

- $(0, 0.2, 0, 1) \rightarrow$ we expect that $A : B :: C : D$ has a low value close to 0 since 0.2 is closer to 0.
- $(0, 0.9, 0, 1) \rightarrow$ we expect that $A : B :: C : D$ has a high value close to 1 since 0.9 is closer to 1.

Normalization: Numerical attributes need to be normalized when dealing with AP classifiers since it helps reduce the time required to classify, enhances the classifier's performance and standardizes the data, which improves the overall process.

Inference: As introduced above, analogical proportions have been recently formalized within Boolean, nominal, and numerical frameworks.

In these latter, given a valid analogical proportion $A : B :: C : D$, the inference principle helps to derive one component of the four-part proportion from the other three. More formally, if the four objects A, B, C, D build a valid AP and if the first three objects A, B, C are known then it is possible to compute the fourth object D by solving the analogical equation: finding a value X such that $A : B :: C : X = I$.

Since the AP can be valid only for six possible assignments of the 4-tuples, there are cases where the equation, in the Boolean case, $A : B :: C : X = I$ has no solution. Indeed, the equations $1 : 0 :: 0 : X = I$ and $0 : 1 :: 1 : X = I$ have no solution.

It has been proven that the above analogical equation is solvable if and only if $(A \equiv B) \vee (A \equiv C)$ holds. In that case, the unique solution X is $X = A \equiv (B \equiv C)$; thus X is either equal to B (if $A = C$) or X is equal to C (if $A = B$). In this paper, we mainly focus on the Boolean setting since, as seen later in the experimental study, the datasets we tested contain only Boolean features.

4 ANALOGY-BASED CLASSIFIER FOR AMD

Analogical inference coincides with AP classification in which the class for an object D can be predicted based on the known classes of the three others: A, B and C . The classification of D (unknown variable) is

only possible if the equation on the class, $Class(A) : Class(B) :: Class(C) : X$, is solvable. (Bounhas et al., 2017). In classification problems, we assume that items are no longer defined as simple variables but rather as vectors of n attribute values, i.e. $\vec{A}=(a_1, \dots, a_n)$ where a_i is the value of attribute i for item \vec{A} , similarly $\vec{B}=(b_1, \dots, b_n)$, $\vec{C}=(c_1, \dots, c_n)$ and $\vec{D}=(d_1, \dots, d_n)$. We also assume implicitly that the four items $\vec{A}, \vec{B}, \vec{C}$ and \vec{D} are represented in terms of the same set of attributes. Then an AP: $\vec{A} : \vec{B} :: \vec{C} : \vec{D}$ is valid if and only if $\forall i \in [1, n], a_i : b_i :: c_i : d_i$.

Analogical classifiers, which are essentially based on the above analogical inference, operate by identifying triplets of examples $(\vec{A}, \vec{B}, \vec{C})$ in the training set that form an AP with the item to be classified (\vec{D}), on all or a maximum number of features. These classifiers also ensure that the corresponding analogical proportion equation for the class has a valid solution.

In its basic formulation, the analogical classifier (Bounhas et al., 2017) applies this principle to determine a solution for the class of \vec{D} that we denote $Class(\vec{D})$. When the analogical equation for the attributes holds, it increments the corresponding score by 1 and assigns the class label with the highest score to \vec{D} . This classifier systematically explores all possible triplets in the training set.

In contrast, the AP classifier, also introduced by Bounhas et al. (2017), does not consider all possible triplets from the training set when classifying a new item \vec{D} . Instead, it restricts the search scope to a smaller subset of candidate triplets. The AP classifier first identifies examples most similar to the item to be classified and narrows its search to pairs of examples exhibiting the same degree of dissimilarity (measured using Manhattan distance) as that between the new item \vec{D} and one of its nearest neighbors. This approach implicitly constructs triplets that are analogically proportional to the new item across all attributes.

Classification with the AP classifier involves an additive aggregation of the truth values associated with the pairs that can be analogically related to those formed by the target item \vec{D} and its nearest neighbor. Only pairs that yield a solvable analogical equation for the classes are considered. The algorithm proposed by Bounhas et al. (2017) achieves comparable performance to earlier analogical classifiers while demonstrating reduced average computational complexity in both nominal and numerical contexts. However, its evaluation has been limited to UCI benchmark datasets for nominal and

numerical data. This study aims to extend its application by evaluating the AP classifier's effectiveness on real datasets for Arabic text morphological disambiguation. Additionally, we conduct a comparative analysis with several competitive ML and DL algorithms.

The basic procedure of the AP classifier can be summarized by the following steps:

- Search for triplets of examples $(\vec{A}, \vec{B}, \vec{C})$ in the training set s.t: \vec{C} is the nearest neighbor of \vec{D} .
- Solve the equation: $Class(\vec{A}):Class(\vec{B})::Class(\vec{C}):X$.
- If the previous analogical equation on classes has a solution ℓ and if the analogical proportion $\vec{A}:\vec{B}::\vec{C}:\vec{D}$ is valid in a componentwise manner for each attribute, then increment the score of ℓ as $score(\ell) = score(\ell) + 1$.
- Assign to \vec{D} the class label ℓ having the highest score $score(\ell)$ as $Class(\vec{D}) = \text{argmax}_{\ell}(score)$.

Algorithm : AP classifier (Bounhas et al., 2017).

Input: $k > 1$, S a training set, $\vec{D} \notin S$ a new instance to be classified

For each label ℓ **Do**

$Score(\ell) = 0$

EndFor

For each \vec{C} in $N_k(\vec{D})$ **Do**

For each pair (\vec{A}, \vec{B}) in $|S|^2$ **Do**

If $(Class(\vec{A}) : Class(\vec{B}) :: Class(\vec{C}) : X$ has solution ℓ)

and $(\vec{A}:\vec{B}::\vec{C}:\vec{D})$ **then** $Score(\ell) = Score(\ell) + 1$

Endif

EndFor

EndFor

$Score^* = \max(Score(\ell))$

if $(Score^* \neq 0)$ **then**

if $(\text{unique}(Score^*, Score(\ell)))$ **then**

$Class(\vec{D}) = \text{argmax}_{\ell}(Score(\ell))$

else

Majority vote

Endif

else

unclassified

Endif

return $Class(\vec{D})$

The AP classifier's success depends on the presence of class-solvable triplets in the training set that form an analogical proportion with the new item to be classified. If such triplets are absent, the classifier fails. This limitation is more likely to occur with small training sets. However, in our analysis of the Arabic text morphological disambiguation datasets, this issue did not arise. In cases where multiple candidate labels are found i.e., the predicted label is not unique (see the algorithm below), the ambiguity is resolved through a majority vote among all possible candidate labels.

The disambiguation process for a given ambiguous word is based on the context of the Arabic text. For instance, we assume that the POS is the morphological feature (MF) to be disambiguated. The classification process relies on two preceding attributes (POS-1 and POS-2) and two succeeding attributes (POS+1 and POS+2) of this word.

5 EXPERIMENTAL RESULTS

In this section, we provide a brief overview of the test collection in Section 5.1. The experimental scenario is detailed in Section 5.2, followed by a comparative study in Section 5.3, which highlights the efficiency of the AP classifier in comparison to various ML and DL classifiers.

5.1 Test Collection

We primarily aim to train the AP, ML and DL classifiers by capturing morphological dependencies using vocalized texts, and then we test these models on non-vocalized texts. During training, we use the morphological analyzer ARAMORPH⁵ on vocalized Arabic texts from Hadith to extract values for 14 morphological features. During the data pre-processing step, a data transformation technique is applied to convert imperfect data into perfect data suitable for classical ML and DL classifiers (Elayeb et al., 2022). The selected classifiers are subsequently trained on vocalized Arabic texts and tested on non-vocalized ones.

Table 1: Overview of the Arabic dataset.

Morphological Feature (MF)	Size (Ko)	Attributes	Instances
POS	8.806	1961	1516
ADJECTIVE	470	105	1502
ASPECT	937	209	1501
CASE	931	209	1501
CONJUNCTION	472	105	1501
DETERMINER	1.186	266	1503
GENDER	696	157	1501
MODE	703	157	1501
NUMBER	926	209	1500
PARTICLE	938	209	1507
PERSON	932	209	1502
PREPOSITION	472	105	1502
VOICE	704	157	1501
PRONOUN	14.551	3329	1477

Table 1 presents an overview of the Arabic dataset ("Kunuz" corpus of Hadith texts (Bounhas et al., 2010; 2011ab)) in terms of data size for the 14

⁵ <https://www.nongnu.org/aramorph/english/index.html>

morphological features, including their total number of attributes (having Boolean values) and instances.

5.2 Experimental Scenario

We apply a 10-fold cross-validation technique across three application domains derived from six Hadith books to assess the performance of the AP classifier, three ML (SVM, Naïve Bayes, and Decision Tree) and three DL classifiers (GRU, CNN and LSTM). For each morphological feature, we compute the average disambiguation rate over the (9+1) iterations.

To get the disambiguation rates, we follow these steps: (i) we first analyze the vocalized texts and record the correct morphological solutions; (ii) second, we remove short vowels from the same texts; (iii) third, we disambiguate the resulting texts using a given classifier and we store the results; and (iv) finally, we compare the two sets of results to compute the disambiguation rate.

We experiment with the three ML classifiers (SVM, NB and DT) using their optimized parameters detailed in (Elayeb et al., 2022). Besides, the structure of the Convolutional Neural Network (CNN) model includes a dropout layer, followed by three CNN layers with a kernel size of 5 and 128 filters. These are followed by global max-pooling with default parameters and another dropout layer. We also used both LSTM and GRU algorithms. The LSTM algorithm contains a single LSTM layer, while the GRU algorithm consists of two GRU layers. This configuration was determined through experimentation to achieve optimal accuracy.

To ensure a fair comparison across all tested classifiers, we also optimize the parameter k for the AP classifier (k representing the number of nearest neighbors \vec{C} considered for classifying an item). Specifically, within each fold of the outer 10-fold cross-validation, we first extract the training set. Then, an inner 5-fold cross-validation is performed on this training set to determine the optimal value of k . The selected value of k from this initial step is subsequently used to classify the test examples in the corresponding outer fold. This procedure is repeated for all folds in the outer cross-validation.

The classification results for the various classifiers, presented in Table 2, correspond to the optimal value of each tuned parameter.

5.3 Experimental Results and Discussion

Table 2 summarizes the disambiguation rates of the 14 morphological features using AP, ML, and DL

classifiers. The results confirm that AP and ML classifiers outperform DL classifiers in disambiguating the following morphological features with similar efficiency: ADJECTIVE (96.54%), ASPECT (71.29%), CASE (56.16%), GENDER (57.16%), MODE (99.40%), NUMBER (85.27%), and VOICE (71.29%). Moreover, the AP classifier achieved an almost identical rate (96.48%) if compared to SVM and NB (96.75%) when disambiguating the morphological feature PARTICLE. However, DL classifiers achieved the best classification results for CONJUNCTION (84.70%), DETERMINER (68.40%), and PREPOSITION (84%). Furthermore, CNN and LSTM emerged as the best classifiers for disambiguating POS (78.00%) and PERSON (61.70%), respectively. Conversely, the AP classifier achieved the highest rate for disambiguating PRONOUN (67.21%). Overall, the average disambiguation rate of the AP classifier across the 14 morphological features is 74.80%, outperforming all ML and DL classifiers. Notably, the data size of certain morphological features poses challenges for some classifiers (e.g., POS for ML classifiers and PRONOUN for DL classifiers). For example, POS includes 1,961 attributes, while PRONOUN includes 3,329 attributes. ML classifiers, in particular, struggle to process this data, even with WEKA's maximum memory allocation of 2,020 MB. To address this limitation, randomly selected subsets of the data are used instead of the entire dataset. However, this method leads to a reduction in disambiguation accuracy for these extensive morphological features if compared to smaller ones.

Furthermore, we observe that certain classifiers produced similar or identical outcomes for specific morphological features. This can be explained by the limited number of possible values for these features (fewer than six) (see for example the results for the Adjective feature with only two possible classes). In contrast, other features yield diverse results across different classifiers. For example, the feature PRONOUN includes 64 possible class values. These observations highlight that effectively optimizing classifier parameters plays a crucial role in improving their performance when disambiguating Classical Arabic texts. Moreover, managing small corpora and text collections with a large number of attributes remains one of the primary challenges of existing DL classifiers. For instance, the GRU algorithm demonstrates a low disambiguation rate (20.20%) for the morphological feature PRONOUN, which comprises 3,329 attributes. These findings align with the conclusions of Elnagar et al. (2020), who also tested DL algorithms for Arabic text classification.

Table 2: Disambiguation rates of 14 morphological features using AP, ML and DL classifiers.

Morphological Feature (MF)	ML classifiers			DL classifiers			Analogical classifier
	SVM	NB	DT	GRU	CNN	LSTM	AP
POS	29.67%	36.21%	48.94%	73.00%	78.00%	31.20%	56.01%
ADJECTIVE	96.54%	96.54%	96.54%	96.00%	96.10%	96.00%	96.54%
ASPECT	71.29%	71.29%	71.29%	69.70%	70.10%	73.00%	71.29%
CASE	56.16%	56.16%	56.16%	21.50%	51.50%	58.80%	56.16%
CONJUNCTION	83.08%	83.08%	83.08%	84.70%	84.70%	84.70%	83.08%
DETERMINER	64.20%	64.20%	64.20%	68.40%	68.40%	68.40%	64.13%
GENDER	57.16%	57.16%	57.16%	55.10%	55.10%	55.10%	57.16%
MODE	99.40%	99.40%	99.40%	99.30%	99.30%	99.30%	99.40%
NUMBER	85.27%	85.27%	85.27%	45.20%	45.20%	85.60%	85.27%
PARTICLE	96.75%	96.75%	96.68%	43.00%	93.00%	94.30%	96.48%
PERSON	60.25%	60.25%	60.25%	60.80%	60.80%	61.70%	60.25%
PREPOSITION	82.89%	82.89%	82.89%	84.00%	84.00%	84.00%	82.89%
VOICE	71.29%	71.29%	71.29%	13.00%	70.90%	73.00%	71.29%
PRONOUN	63.04%	60.19%	62.56%	20.20%	58.10%	58.10%	67.21%
Average	72.64%	72.90%	73.97%	59.60%	72.10%	73.10%	74.80%

The AP classifier appears to be less sensitive to dataset size, even when managing a large number of attributes or classes. Notably, it achieves the highest disambiguation rate for the PRONOUN dataset compared to all other classifiers. This highlights, once again, the AP classifier's efficiency in handling multi-class classification tasks involving numerous attributes, even under conditions of data scarcity (Bounhas et Prade, 2023). This efficiency is achieved through the use of triplets of examples from the training set, which form an analogical proportion with the item being classified. Consequently, the triplet-based approach serves as a method for augmenting sparse data, allowing the classifier to draw reliable conclusions about the item by aggregating scores across various triplets.

6 CONCLUION

Arabic morphological disambiguation remains one of the major challenges in several domains, including machine translation, information retrieval, speech recognition and synthesis, educational tools, and text-to-speech systems. Solving the problem of ambiguity in Arabic texts with high accuracy can be highly beneficial for these application areas. For this purpose and given the success of analogical proportions in summarizing Arabic texts and in classifying both structured nominal or numerical data and unstructured data (such as Arabic text classification), we aim to investigate the efficiency of an AP classifier in disambiguating Classical Arabic texts. We compare the performance of the analogy-based algorithm to a set of well-established ML and DL classifiers. The results demonstrate the competitive performance of the AP classifier in terms of the average disambiguation rate across the 14 morphological features.

Furthermore, the AP classifier exhibits reduced complexity compared to analogical classifiers that take into account all triplets, while maintaining accuracy that is either better than or, in many cases, equivalent to some ML and DL classifiers (Bounhas et al, 2024.a).

Despite its efficiency, the AP classifier requires further investigation and enhancement. First, it is important to test it using modern Arabic text collections, such as TreeBank. Second, we aim to expand the AP classifier to accommodate regional dialects with unique morphological structures. Finally, we believe that the accuracy of Arabic morphological disambiguation can be improved by incorporating advanced linguistic context, including integration with syntax and semantics. Additionally, we propose leveraging unsupervised learning techniques to reduce reliance on annotated datasets, focusing on the development of unsupervised or semi-supervised methods.

REFERENCES

- Ayed, R., Bounhas, I., Elayeb, B., Evrard, F., Bellamine Ben Saoud, N. (2012a). Arabic Morphological Analysis and Disambiguation Using a Possibilistic Classifier. In *Proc. 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 *Proc. 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 *Proc. 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 *Proc. 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 *Proc. 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 *Proc. 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 Knowl. Eng.*, 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. *Comput. Speech Lang.*, 33(1):67-87.
- Bounhas, I., Elayeb, B., Evrard, F., Slimani, Y. (2010). Toward a computer study of the reliability of Arabic stories. *J. Assoc. Inf. Sci. Technol.*, 61(8):1686–1705.
- Bounhas, I., Elayeb, B., Evrard, F., Slimani, Y. (2011a). ArabOnto: Experimenting a New Distributional Approach for Building Arabic Ontological Resources. *Int. J. Metadata Semant. Ontologies*, 6(2):81-95.
- Bounhas, I., Elayeb, B., Evrard, F., Slimani, Y. (2011b). Organizing Contextual Knowledge for Arabic Text Disambiguation and Terminology Extraction. *Knowl. Org.*, 38(6):473–490.
- Bounhas, M., Elayeb, B. (2019). Analogy-based Matching Model for Domain-specific Information Retrieval. In *Proc. of ICAART-2019*, Vol. 2, pp. 496-505, Prague, Czech Republic, SciTePress.
- Bounhas, M., Elayeb, B., Chouigui, A., Hussain, A., Cambria, E. (2024). Arabic text classification based on analogical proportions. *Expert Syst. J. Knowl. Eng.* 41(10).
- Bounhas, M., Prade, H. (2023). Analogy-based classifiers: An improved algorithm exploiting competent data pairs. *Int. J. Approx. Reason.*, 158: 108923.
- Bounhas, M., Prade, H. (2024). Revisiting analogical proportions and analogical inference. *Int. J. Approx. Reason.* 171: 109202.
- Bounhas, M., Prade, H., Richard, G. (2017). Analogy-based classifiers for nominal or numerical data. *Int. J. Approx. Reason.*, 91: 36-55.
- Bousmaha, K. Z., Rahmouni, M. K., Kouninef, B., Belguith, L. H. (2016). A Hybrid Approach for the Morpho-Lexical Disambiguation of Arabic. *J. Inf. Process. Syst.*, 12(3):358–380.
- Daoud, D. (2009). Synchronized Morphological and Syntactic Disambiguation for Arabic. *Adv. Comput. Linguistics*, 41:73–86.
- Daoud, D., Daoud, M. (2009). Arabic Disambiguation Using Dependency Grammar. In *Proc. of TALN-2009*, Senlis, France, ATALA.
- Diab, M. T. (2007). Improved Arabic Base Phrase Chunking with a New Enriched POS Tag Set. In *Proc. of Semitic-2007*, pp. 89–96, Stroudsburg, PA, USA. ACL.
- Dubois, D., Prade, H., Richard, G. (2016). Multiple-valued extensions of analogical proportions, *Fuzzy Sets Syst.*, 292:193–202.
- Elayeb, B. (2019). Arabic Word Sense Disambiguation: A Review. *Artif. Intell. Rev.*, 52(4):2475-2532.
- Elayeb, B. (2021). Arabic Text Classification: A Literature Review. In *Proc. of AICCSA-2021*, pp. 1-8, Tangier, Morocco, IEEE Computer Society.
- Elayeb, B., Ayed, R. (2022). Socio-Semantic Information Retrieval of Structured Arabic Texts. In *Proc. of AICCSA-2022*, pp. 1-8, Abu Dhabi, UAE, IEEE Computer Society.
- Elayeb, B., Ayed, R. (2023). Analogical Text Mining: Application to Arabic Text Summarization and Classification. In *Proc. of AICCSA-2023*, pp. 1-8, Giza, Egypt, IEEE Computer Society.
- Elayeb, B., Ben Khiroun, O. (2023). SPEEDSER: A Possibilistic System for Query Disambiguation, Expansion and Translation. *Int. J. Inf. Technol. Decis. Mak.* DOI: 10.1142/S0219622023500499
- Elayeb, B., Bounhas, I. (2016). Arabic Cross-Language Information Retrieval: A Review. *ACM Trans. Asian Low Resour. Lang. Inf. Process.*, 15(3):18:1-18:44.
- Elayeb, B., Chouigui, A., Bounhas, M., Ben Khiroun, O. (2020). Automatic Arabic Text Summarization Using Analogical Proportions. *Cogn. Comput.*, 12(5):1043-1069.
- Elayeb, B., Ettih, M.F., Ayed, R. (2022). Experimenting machine-learning algorithms for morphological disambiguation of Arabic texts. In *Proc. of ICAART-2022*, Vol. 3, pp. 851-862, SciTePress.
- Elayeb, B., Evrard, F., Zaghdoud, M., Ben Ahmed, M. (2009). Towards an intelligent possibilistic web information retrieval using multiagent system. *Interact. Technol. Smart Educ.*, 6(1):40–59.
- Elnagar, A., Debsi, R. A., Einea, O. (2020). Arabic text classification using deep learning models. *Inf. Process. Manag.*, 57(1), 102121.
- Habash, N., Rambow, O. (2007). Arabic Diacritization Through Full Morphological Tagging. In *Proc. of HLT-NAACL 2007: Short Papers, HLT-NAACL-Short'07*, pp. 53–56, Stroudsburg, PA, USA. ACL.
- Khoja, S. (2001). APT: Arabic part-of-speech tagger. In *Proc. of the NAACL-2001*, Pennsylvania, USA.
- Pasha, A., Al-Badrashiny, M., Diab, M., El Kholly, A., Eskander, R., Habash, N., Pooleery, M., Rambow, O., Roth, R. (2014). MADAMIRA: A Fast, Comprehensive Tool for Morphological Analysis and Disambiguation of Arabic. In *Proc. of LREC-2014*, pp. 1094-1101, Reykjavik, Iceland, ELRA.
- Prade, H., Richard, G. (2010). Reasoning with logical proportions. In *Proc. of KR-2010*, pp.545–555, Toronto, AAAI Press.