Data & Knowledge Engineering 100 (2015) 240–254

Contents lists available at ScienceDirect

Data & Knowledge Engineering

journal homepage: www.elsevier.com/locate/datak

A hybrid possibilistic approach for Arabic full morphological disambiguation

Ibrahim Bounhas [a,*,1], Raja Ayed [b,1], Bilel Elayeb [b,c,1], Narjès Bellamine Ben Saoud [b,d]

[a] LISI Laboratory of Computer Science for Industrial Systems, Carthage University, Tunisia
[b] RIADI Research Laboratory, ENSI, Manouba University, 2010 Tunisia
[c] Emirates College of Technology, P.O. Box: 41009, Abu Dhabi, United Arab Emirates
[d] Higher Institute of Informatics (ISI), Tunis El Manar University, 2080 Ariana, Tunisia

ARTICLE INFO

Article history:
Received 15 January 2015
Received in revised form 18 June 2015
Accepted 21 June 2015
Available online 27 June 2015

Keywords:
Arabic morphological disambiguation
Possibilistic classification
Imperfect data
Linguistic rules
Out-of-Vocabulary word analysis

ABSTRACT

Morphological ambiguity is an important phenomenon affecting several tasks in Arabic text analysis, indexing and mining. Nevertheless, it has not been well studied in related works. We investigate, in this paper, new approaches to disambiguate the morphological features of non-vocalized Arabic texts, combining statistical classification and linguistic rules. Indeed, we perform unsupervised training from unlabelled vocalized Arabic corpora. Thus, the training and testing sets contain imperfect instances (i.e. having ambiguous attributes and/or classes). To handle imperfect data, we compare two approaches: i) a possibilistic approach allowing to handle imperfection in a direct manner; and, ii) a data transformation-based approach permitting to convert an imperfect dataset to a perfect one, thus allowing to exploit classical classifiers. We also present an approach dealing with unknown (Out-of-Vocabulary) words. The experiments focus mainly on classical texts, which were not sufficiently studied in related works. We show that the possibilistic approach performs better than the transformation-based one. Besides, we report encouraging results as far as i) the role of linguistic rules in enhancing the disambiguation rates; and, ii) the accuracy of our approach for full morphological disambiguation of unknown words.

1. Introduction

1.1. Arabic morphological ambiguities: causes and effects

In Arabic language, various words may have the same morphological form, because of the lack of short vowels in some texts, which represent a challenge for many Arabic Natural Language Processing (NLP) tasks [1]. Indeed, a non-vocalized word may have more than 10 interpretations. For instance, if we add short vowels to the form “كتب”, we can obtain “كُتُبٌ” (books) or “كَتَبَ” (he wrote) and so on. This example shows the impact of derivation on morphological ambiguity. Certainly, there are other sources of ambiguity caused by the agglutinative character of Arabic. For example, the word “وضوء” (wDw') has three possible interpretations: namely “وُضُوء” (wuDuw': ablution),“وَضُوء” (waDuw': water for ablution) and “ضَوْء” (Dw': light). Thus, the letter “و” may or may not be considered as a conjunction agglutinated to the lemma. Another source of morphological ambiguity is the “free word order phenomenon” [2]. We

[*] Corresponding author at: Higher Institute of documentation, La Manouba University, Campus Universitaire de la Manouba, 2010, Manouba, Tunisia. Tel.: +216 96563257.

E-mail addresses: bounhas.ibrahim@yahoo.fr (I. Bounhas), ayed.raja@gmail.com (R. Ayed), bilel.elayeb@riadi.rnu.tn (B. Elayeb), narjes.bellamine@ensi.rnu.tn (N. Bellamine Ben Saoud).

[1] Jarir: Joint group for Artificial Reasoning and Information Retrieval (www.jarir.tn).

http://dx.doi.org/10.1016/j.datak.2015.06.008
0169-023X/© 2015 Elsevier B.V. All rights reserved.

talk about a case where the structure of the sentences or the expressions in Arabic may affect morphological disambiguation. For example, the expression "الوَلَدُ كَذِبَ" (the boy lied) may be replaced by "كَذِبَ الوَلَدُ" having the same meaning. This may mislead morphological disambiguation statistical algorithms, as it changes the frequencies of contextual dependencies.

Morphological analyzers provide the stem and the flectional marks of one Arabic word [3]. A morphological analyzer provides the different values of morphological features (i.e. part-of-speech, gender, number, voice, etc.) independently of the context. It assigns to a given word its possible solutions. If the analyzer provides more than one solution for a given word, then this word will be considered ambiguous. So, we need to apply morphological disambiguation, which consists in selecting the most accurate value among the proposed solutions. Many works use classification-based approaches to resolve the morphological ambiguities [4–6]. Nevertheless, the complexity of morphologically rich languages like Arabic makes this task hard to achieve [7,8].

Resolving morphological ambiguities is essential for many other levels of text analysis, mining and retrieval. The morphological features are an important input for syntactic parsers, which aim to recognize the grammatical function of a word in a sentence. For example, the expression "كذب الحديث" may be read as "كَذِبُ الحَدِيث" (an annexation compound noun, which may be translated as "lied speech"). It may, also, be interpreted as "كَذَبَ الحَدِيثَ" (a whole sentence meaning "He lied in his speech") where the first word (صَدَقَ; to lie) is the verb of the sentence. It is clear from this example, that the short diacritics [9] have an important role in morphological disambiguation and may be useful for syntactic analysis; and that vocalized texts are less ambiguous than non-vocalized ones. The example of "وضوء" shows that the morphological disambiguation may help identifying the semantic meaning of a word.

Furthermore, several types of applications need to deal with the structure of words, such as Automatic Speech Recognition (ASR) [7,60,61], Arabic text phonetization [10] and summarization [11]. To retrieve documents or extract relevant knowledge, we need to recognize the useful entities (e.g. words, expressions, concepts, etc.). In these fields, the structure of words has a central role.

1.2. Notes on terminology

The goal of this section is to clarify some key concepts used in the paper. We mainly talk about the following notions.

Classical vs. modern standard Arabic: classical Arabic is the language used until the Middle Ages, while modern standard Arabic (MSA) is the Arabic used after this period. Indeed, MSA is a simplified version, which has some differences, especially in terms of style and lexicon, while there are not great changes in terms of syntax and morphology [35]. Thus, in MSA, we find words which represent modern and technical concepts. This may affect the lexical probability defined in Section 1.3.2.

Vocalized vs. non vocalized Arabic texts: diacritic or vocalized texts are composed of diacriticized Arabic words i.e. words accompanied with short diacritics (تشكيل; taskhil). For example, the word "كتب" is not vocalized, while "كَتَبَ" is.

Ambiguity, imprecision and uncertainty: we treat morphological disambiguation as a classification task with imperfect data. In possibility theory, imperfection covers both imprecise and uncertain data [34,64]. In our case, imperfection is caused by morphological ambiguities. For example, if a word has two possible values of POS, it causes imprecision if it is used as an attribute to disambiguate other words. It causes uncertainty if its POS is considered as a class.

Ambiguous word: a word which has more than one morphological analysis i.e. two or more possible interpretations of its structure.

Imprecision: a classification attribute is imprecise if it admits more than one possible value (multi-attribute or set-valued attribute). If we take the example of the attribute color, having "red", "blue" and "yellow" as domain, we say it is imprecise for a given instance, if we know for example that its value is either "red" or "blue".

Uncertainty: we say that a training instance is uncertain if it admits more than one possible class.

1.3. Related work

Arabic morphological disambiguation approaches use linguistic and/or statistical knowledge. Although this classification is not precise in this case, we will distinguish two kinds of approaches as follows.

1.3.1. Linguistics approaches

Rules written by linguists are exploited to label the morphological features. In related works, we talk about heuristics, contextual and non-contextual rules (cf. [12–14] for more reading). Daoud and Daoud [13] proposed a specific type of parsers called EnConverters written in UNL (Universal Networking Language) and EnCo[2]. Thus, they defined several types of disambiguation rules combining morphological and syntactic contextual dependencies. Nevertheless, the authors did not perform any experimental evaluation, thus making hard to assess their approach in terms of the coverage, reusability and accuracy. Some researchers (e.g. [14]) tried to exploit full syntactic parsing for morphological disambiguation.

1.3.2. Statistical approaches

Statistical approaches try to model fuzzy relationships between tokens, features and contextual information and to train the disambiguation models on large collections to enhance performance. Hence, statistical classifiers are widely used in this field. Support vector machines (SVM) [15] and several probabilistic models (e.g. Markov Models) are used for Arabic and other languages [1,7,8,16–18] to model fuzzy disambiguation contextual knowledge. As stated before, unlike linguistic approaches, which need only

[2] A rule-based programming language (http://libraries.unl.edu/).