Available online at www.sciencedirect.com

ScienceDirect

Computer Speech and Language 33 (2015) 67–87

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

Experimenting a discriminative possibilistic classifier with reweighting model for Arabic morphological disambiguation☆

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

[a] *LISI Laboratory of Computer Science for Industrial Systems, Higher Institute of Documentation (ISD), Manouba University 2010, 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] *Informatics Research Institute of Toulouse (IRIT), 02 Rue de Charles Camichel, 31071 Toulouse Cedex 7, France*
[e] *Higher Institute of Informatics (ISI), Tunis El Manar University, 2080 Ariana, Tunisia*

Received 14 January 2014; received in revised form 19 September 2014; accepted 17 December 2014
Available online 3 January 2015

Abstract

In this paper, we experiment a discriminative possibilistic classifier with a reweighting model for morphological disambiguation of Arabic texts. The main idea is to provide a possibilistic classifier that acquires automatically disambiguation knowledge from vocalized corpora and tests on non-vocalized texts. Initially, we determine all the possible analyses of vocalized words using a morphological analyzer. The values of their morphological features are exploited to train the classifier. The testing phase consists in identifying the accurate class value (i.e., a morphological feature) using the features of the preceding and the following words. The appropriate class is the one having the greatest value of a possibilistic measure computed over the training set. To discriminate the effect of each feature, we add the weights of the training attributes to this measure. To assess this approach, we carry out experiments on a corpus of Arabic stories and on the Arabic Treebank. We present results concerning all the morphological features and we discern to which degree the discriminative approach improves disambiguation rates and extract the dependency relationships among the features. The results reveal the contribution of possibility theory for resolving ambiguities in real applications. We also compare the success rates in modern versus classical Arabic texts. Finally, we try to evaluate the impact of the lexical likelihood in morphological disambiguation.

Keywords: Morphological analysis; Morphological disambiguation; Discriminative possibilistic classifier; Reweighting model

1. Introduction and related work

Many applications in the field of Arabic Natural Language Processing (ANLP) need to deal with the complex morphology of this language. Morphological analysis and disambiguation is an important step in Automatic Speech

☆ This paper has been recommended for acceptance by E. Briscoe.
* Corresponding author. Tel.: +216 71520164.
E-mail addresses: Bounhas.ibrahim@yahoo.fr (I. Bounhas), ayed.raja@gmail.com (R. Ayed), bilel.elayeb@riadi.rnu.tn (B. Elayeb), fabevrard@free.fr (F. Evrard), narjes.bellamine@ensi.rnu.tn (N. Bellamine Ben Saoud).

http://dx.doi.org/10.1016/j.csl.2014.12.005
0885-2308/

Recognition (ASR) (Diehl et al., 2012; Kirchhoff et al., 2006), Arabic text phonetization (El-Imam, 2004) and summarization (Azmi and Al-Thanyyan, 2012). Besides, information access applications need to index documents and extract relevant features about their meaningful entities (Bounhas et al., 2011b; Elayeb, 2009; Elayeb et al., 2009, 2011, 2014). Indeed, Information Retrieval and Knowledge Extraction Systems (IRKES) require recognizing useful entities in texts such as words, expressions and concepts. The basic level concerns the structure of words; i.e., the morphological level. Indeed, a given word may have many interpretations at this level, what is called morphological ambiguity. This phenomenon is more challenging with morphologically rich languages such as Arabic (Diab et al., 2004). Thus, a non-vocalized Arabic word may have more than 12 interpretations (Habash and Rambow, 2007; Habash et al., 2009b).

In this paper, we study existent morphological disambiguation approaches applied for Arabic texts. Then, we present our framework, which allows to, automatically, learn contextual knowledge required for disambiguation from vocalized texts. This framework tries to avoid the limits of existent systems, which require manually encoded knowledge or labeled corpora. It is also an attempt to consider Arabic classical texts; because most of the existent tools were trained and assessed on modern corpora (cf. Section 1.2). Another important concern is lexical likelihood, which differs from one type of text to another. This issue is carefully studied in this paper; we examine, therefore, the effect of this factor on morphological disambiguation. Our framework is illustrated through examples (cf. Section 3) and assessed through experimental results (cf. Section 4). Indeed, the basic version of our possibilistic classifier was presented in two conference papers (Ayed et al., 2012a,b). This new contribution stands out by the following aspects.

First, we present a reweighting model which tries to evaluate the discriminative power of attributes. We also take into account the discriminative power of the values of each attribute. Indeed, it is the first time that the necessity measure is being used in possibilistic classification, as the state-of-the art possibilistic classifier used only the possibility measure (Haouari et al., 2009). We also propose a new version of information entropy adapted for attribute reweighting in imprecise data. In the whole, we obtain six different classifiers (combining possibility, necessity and entropy). Besides, we compute, in the training phase, the lexical likelihood to take into account the dependencies between a given word and its features.

Second, we fully re-experiment these classifies, thus assessing the impact of these discriminative weights and the lexical likelihood. In addition, this paper is a fully revised version which provides a more detailed interpretation of results. In fact, we employ the Wilcoxon Matched-Pairs Signed-Ranks Test (Demsar, 2006) to assess our results, besides computing the disambiguation rates.

Finally, the experiments in Ayed et al. (2012a,b) were performed in a non-standardized traditional corpus. In this paper, we assess our model on the Arabic Treebank, thus showing its performance in a modern standard corpus.

1.1. Arabic morphological ambiguity

Morphological analysis tools, like Hajic's analyzer (Hajic, 2000), allow to recognize the stem of a given word and its flectional marks. The analyzer interprets a given word out of context and returns a set of possible solutions (analyses), each having different morphological features.[1] A word is ambiguous if it has more than one solution. Disambiguation is the task of choosing, among these solutions, the most appropriate given context of the word (Ayed et al., 2012a, 2014a,b). However, this task is not easy to achieve, because of the complexity of the Arabic language morphology (Kirchhoff et al., 2006).

We analyze the main sources of ambiguities in the Arabic language and their consequences as follows. In fact, this language is agglutinative, derivational and inflectional. For example, the word "وعوض" (wDw′) may be analyzed as "وعُوضُ" (wuDuw′: ablution),"وعُوضَ" (waDuw′: water for ablution) or وضوء (Dw′: light). In this example, the letter "و" is interpreted either as a conjunction or as the first letter of the lemma. Even in the second case, we obtain two possible lemmas diacriticized differently. In fact, the main source of ambiguity is the lack of diacritics in most existing Arabic texts. Morphological ambiguities make it difficult to extract simple terms, because the morphological analyzers enumerate for each word many possible lemmas (Bounhas et al., 2011b).

[1] POS (verb, noun, particle, etc.), number (singular, dual or plural), gender (male or female), voice (active or passive), mode of the verb (indicative, subjunctive, etc.), person (first, second or third), aspect (perfect or imperfect), etc.