Journal of King Saud University – Computer and Information Sciences 29 (2017) 147–155

Contents lists available at ScienceDirect

Journal of King Saud University – Computer and Information Sciences

journal homepage: www.sciencedirect.com

Morphological disambiguation of Tunisian dialect

Inès Zribi [a,*], Mariem Ellouze [a], Lamia Hadrich Belguith [a], Philippe Blache [b]

[a] ANLP Research group, MIRACL, University of Sfax, Tunisia
[b] LPL, CNRS, Aix-Marseille University, France

ARTICLE INFO

Article history:
Received 14 June 2016
Revised 9 January 2017
Accepted 17 January 2017
Available online 29 January 2017

Keywords:
Tunisian dialect
Spoken language
Morphological analysis
Morphological disambiguation

ABSTRACT

In this paper, we propose a method to disambiguate the output of a morphological analyzer of the Tunisian dialect. We test three machine-learning techniques that classify the morphological analysis of each word token into two classes: *true* and *false*. The class label is assigned to each analysis according to the context of the corresponding word in a sentence. In failure cases, we combine the results of the proposed techniques with a bigram classifier to choose only one analysis for a given word. We disambiguate the result of the morphological analyzer of the Tunisian Dialect *Al-Khalil-TUN* (Zribi et al., 2013b). We use the Spoken Tunisian Arabic Corpus *STAC* (Zribi et al., 2015) to train and test our method. The evaluation shows that the proposed method has achieved an accuracy performance of 87.32%.

1. Introduction

Morphological analysis (MA) is a crucial stage in a variety of natural language processing (NLP) applications (information retrieval, question answering, etc.). The analysis of languages with complex and rich morphology handicaps the performance of these applications due to the large number of analyses produced for each word independently of the context in which the word occurs. Therefore, a morphological disambiguation module is required.

Morphological disambiguation (MD) (also called Part-of-Speech (POS) tagging) consists in determining one correct POS tag among a set of POS tags that are assigned to a word, by taking into account the word's context.

In the literature, many techniques/systems have been developed for POS tagging modern standard Arabic (MSA). They follow two principal approaches to developing a tagger: a handcrafted rule-based approach, and a statistical approach. The handcrafted rule-based approach may be a practicable solution, but it requires a considerable investment of human effort. The most referenced works are done by Al-Taani and Al-Rub (2009) and Tlili-Guiassa (2006). Statistical approaches prove to be able to learn tagging from tagged data on the basis of a sufficient quantity of tagged documents. The most referenced works are carried out by Diab et al. (2004), Habash and Rambow (2005), Khoja (2001).

Contrariwise, Arabic dialects have not received much attention due to the scarcity of resources (corpus and lexicon) and tools (morphological analyzers, tokenizers, etc.). In addition, Arabic dialects are a spoken variety. Tagging a spoken language is typically harder than tagging a written one, due to the effect of disfluencies, incomplete sentences, etc. (Duh and Kirchhoff, 2005).

In this paper, we present the Tunisian Arabic Morphological DisAmbiguation System (TAMDAS). This system uses the output of a Tunisian Dialect (TD) morphological analyzer (Al-Khalil-TUN) (Zribi et al., 2013b) and a TD corpus (the STAC corpus) (Zribi et al., 2015), to morphologically disambiguate TD annotated transcriptions.

TAMDAS tests three different classifiers and combines their results with a bigram module in failure cases. We build a classifier based on feature vectors, which are generated from the morphologically annotated corpus, and then use it to classify the possible analyses of each word into *correct* and *false* classes.

This paper has seven main sections. Section 2 presents an overview of previous works that studied TD and the POS tagging of dialectal Arabic. Section 3 presents the characteristics of TD. In Section 4, we present the challenge of tagging a spoken language, especially in the case of TD. In Section 5, we describe the TD resources, then, in Section 6, we present our method. Finally, we give the results of the system evaluation, and discuss some errors.

* Corresponding author.
E-mail addresses: ineszribi@gmail.com (I. Zribi), Mariem.ellouze@planet.tn (M. Ellouze), l.belguith@fsegs.rnu.tn (L.H. Belguith), philippe.blache@lpl-aix.fr (P. Blache).

Peer review under responsibility of King Saud University.

Production and hosting by Elsevier

http://dx.doi.org/10.1016/j.jksuci.2017.01.004
1319-1578/

2. Related works

Some studies have been conducted in the field of Dialectal Arabic (DA) processing with a variety of approaches and at different degrees of linguistic depth. Most of the approaches tend to develop dialectal data (Al-Badrashiny et al., 2014; Al-Shargi et al., 2016; Khalifa et al., 2016; Maamouri et al., 2014; Samih and Maier, 2016), and tools (Darwish, 2014; Habash et al., 2012b; Habash and Rambow, 2006; Salloum and Habash, 2014, 2011) to treat a specific Arabic dialect. The most referenced works are carried out by Habash et al. (2013, 2005), Habash and Rambow (2005), Rambow et al. (2006).

In fact, few research studies treated the POS tagging task of Arabic dialects. Most of them dealt with Levantine and Egyptian Arabic. They treated these dialects as written varieties of Arabic languages (no characteristic of speech are considered). However, the automatic processing of Tunisian Dialect (TD) and its spoken varieties has not received much attention.

The DA POS tagging techniques follow two principal approaches. The ***first approach*** suggests using MSA resources and a few DA resources to create a POS tagger. In this context, (Duh and Kirchhoff, 2005) used the Buckwalter Morphological Analyzer (Buckwalter, 2004) designed for MSA, the LDC MSA Treebank corpus and some dialectal resources (the CallHome Egyptian Colloquial Arabic corpus, the LDC Levantine Arabic corpus) in combination with unsupervised learning algorithms in order to develop a POS tagger for Egyptian Arabic. The authors proposed to bootstrap the HMM tagger using POS information from the morphological analyzer. They improved the tagger by integrating additional data from other dialects (Duh and Kirchhoff, 2005). They reported a POS accuracy of 70.9%. Likewise, Rambow et al. (2006) explored MSA data and resources to develop a POS tagger for Levantine dialect. They adapted an MSA POS-tagger for Levantine data. They suggested that leveraging the existing resources is a viable option. Rambow et al. (2006) developed a bilingual small lexicon MSA/Levantine dialect. Combining information from this lexicon and a parameter renormalization strategy based on minimal linguistic knowledge, Rambow et al. (2006) noticed the biggest improvement in the tagger. Moreover, Habash et al. (2013) developed a morphological analysis and disambiguation for Egyptian Arabic based on an existing tool for MSA (the MADA tool, Habash and Rambow (2005) and Roth et al. (2008). MADA uses an existing morphological analyzer of MSA and applies a set of models (support vector machines and N-gram language models) to produce a per word in-context prediction. A ranking component computes the scores of the analysis produced by the morphological analyzer using a tuned weighted sum of matches with the predicted features (Habash et al., 2013). The top-scoring analysis is chosen as the best prediction of the tool (Habash et al., 2013).

The ***second approach*** of POS tagging DA intends to start from scratch. No MSA resources are used in this approach. Al-Sabbagh and Girju (2012) implemented Brill's Transformation-Based tagging algorithm (Brill, 1994) for the task of POS tagging Egyptian Arabic. For training, they used the manually annotated Twitter-based corpus. They reported an 87.6% accuracy on POS tagging.

Only two studies dealt with the POS tagging of the Tunisian dialect. They adopted the first approach. Boujelbane et al. (2014)) have retrained the MSA Stanford POS tagger (Toutanova and Manning, 2000). To retrain their system, they used a corpus derived from a translation of the MSA Treebank into TD. An accuracy of 78.5% in POS tagging of Tunisian transcribed texts was reported. Hamdi et al. (2015) proposed three steps for POS tagging TD. Their method is based on MSA resources. They convert a TD sentence into a MSA lattice, which is disambiguated to produce MSA target sentences. Finally, the MSA tagger assigns to each word its POS tag. This system achieved an accuracy of 89%.

3. Tunisian dialect

The Tunisian Dialect (TD) is the dialect of the Arabic language spoken in Tunisia. It is considered as a low variety given that it is neither codified nor standardized, even though it is the mother tongue daily spoken by everyone (Saidi, 2007). The regional varieties of TD are the Tunis dialect (Capital), the Sahel dialect, the Sfax dialect, the Northwestern Tunisian dialect, the Southwestern Tunisian dialect and the South-Eastern Tunisian dialect (Gibson, 1998; Talmoudi, 1980).

There are a lot of different and similar points between TD and MSA (Zribi et al., 2013a). In order to compare these two varieties of Arabic language, we focus on four levels: namely the phonological level, the morphological level, the lexical level and the syntactic level.

3.1. The phonological level

The vocalic system of TD is reduced (Tilmatine, 1999). Some short vowels are not overtly rendered, especially if they are located at the end of the word (Mejri et al., 2009). The MSA verb [1]شَرِبَ <šariba> /šariba/[2] 'he drank' is pronounced شْرِبْ /šrib/ (note the deletion of the vowels located at the beginning and the end of the verb). Moreover, TD has a long vowel/e:/ which does not exist in MSA (Zribi et al., 2014).

The consonant system also includes some phonetic differences (Mejri et al., 2009). In some cases, the Arabic consonant ق <q> /q/ is pronounced /g/. The MSA word بَقَرَة <baqaraħ> /baqara/ 'cow' is pronounced in TD /bagra/. In addition, some consonants in TD have multiple pronunciations. For example, the consonant غ <γ> /γ/ and ج <j> /j/ can also be pronounced respectively as /x/ and /z/.

3.2. The morphological level

The main difference between MSA and TD is on the affix level. We can notice the presence of new dialectal affixes and the deletion of others. Dual suffixes ان <An> and ين <yn> are generally absent. They are replaced by the numeral زوز <zwz> 'two' located after or before the plural form of the noun. However, some words in TD can be agglutinated to the suffix ين <yn> to express duality. In verb conjugation, TD is characterized by the absence of the dual (feminine and masculine) and the feminine in the plural. It has witnessed many simplifications in its affixation system (Ouerhani, 2009). Indeed, new affixes appeared. The first one is the negation clitic ش <š>. It is agglutinated to the end of the verb that must be preceded by the negation particle ما <mA>(*e.g.*, ما كليتش <mA klyt-š> 'I don't eat') (Mejri et al., 2009). The interrogation prefix of MSA أ <Â> is transformed in TD into the suffix شي <-šy> (*e.g.*, خرجشي, <xrj-šy>, 'Did he go out?'). Likewise, the future prefix سـ <s-> is replaced by the particle باش <bAš> 'will'. In addition, we note the absence of the dual clitics in TD.

3.3. The lexical level

Historical events have made the linguistic situation in Tunisia rather complex. The prolonged Ottoman Turkish political domination of North Africa roughly from the mid-fifteenth to the late nineteenth century and the French colonization from 1830 had an impact on the absorption of foreign vocabulary into the lexicon of local Arabic dialects (Holes, 2004).

In addition to Turkish and French, we find many examples of European language lexical elements in TD. We can identify a signif-

[1] We follow the CODA-TUN convention (Zribi et al., 2014) when writing examples of words in TD.

[2] Transliteration is coded following Buckwalter transliteration. For more details about it, see (Habash et al., 2007).