Journal of King Saud University – Computer and Information Sciences (2014) 26, 372–378

King Saud University

Journal of King Saud University – Computer and Information Sciences

www.ksu.edu.sa
www.sciencedirect.com

ADAM: Analyzer for Dialectal Arabic Morphology

Wael Salloum [a,*], Nizar Habash [b]

[a] Center for Computational Learning Systems, Columbia University, United States
[b] New York University Abu Dhabi, United Arab Emirates

Available online 2 October 2014

KEYWORDS

Arabic natural language processing;
Dialectal Arabic;
Arabic morphology;
Machine translation

Abstract While Modern Standard Arabic (MSA) has many resources, Arabic Dialects, the primarily spoken local varieties of Arabic, are quite impoverished in this regard. In this article, we present ADAM (Analyzer for Dialectal Arabic Morphology). ADAM is a poor man's solution to quickly develop morphological analyzers for dialectal Arabic. ADAM has roughly half the out-of-vocabulary rate of a state-of-the-art MSA analyzer and is comparable in its recall performance to an Egyptian dialectal morphological analyzer that took years and expensive resources to build.

1. Introduction

Arabic dialects, or the primarily spoken local varieties of Arabic, have recently received increased attention in the field of natural language processing (NLP). An important challenge for work on these dialects is to create morphological analyzers, or tools that provide for a particular written word all of its possible analyses out of context. While Modern Standard Arabic (MSA) has many such resources (Graff et al., 2009; Smrž, 2007; Habash, 2007), Dialectal Arabic (DA) is quite impoverished (Habash et al., 2012b). Furthermore, MSA and the dialects are quite different morphologically: Habash et al., 2012b reported that only 64% of Egyptian Arabic words are analyzable using an MSA analyzer. Thus, using MSA resources to process the dialects will have limited value. Additionally, as for any language or dialect, developing good large-scale coverage lexicons and analyzers can require much time and effort.

In this article, we present ADAM (Analyzer for Dialectal Arabic Morphology). ADAM is a poor man's solution for developing a quick and dirty morphological analyzer for dialectal Arabic. ADAM can be used as is or can function as the first step in bootstrapping analyzers for Arabic dialects. It covers all part-of-speech (POS) tags just as any other morphological analyzer; however, because we use ADAM mainly to process text, we do not model phonological differences between Arabic dialects and we do not evaluate the differences in phonology. In this work, we apply ADAM extensions to MSA clitics to generate proclitics and enclitics for different Arabic dialects. This technique can also be applied to stems to generate dialectal stems; however, that is outside the scope of this work.

In Section 2, we review some of the challenges of processing Arabic in general and Arabic dialects in particular. We discuss related work in Section 3, and we outline and detail our approach in Section 4. Finally, in Section 5, we present several detailed evaluations using a variety of metrics and compare against state-of-the-art analyzers of MSA and Egyptian Arabic.

* Corresponding author.
E-mail addresses: wael@ccls.columbia.edu (W. Salloum), nizar.habash@nyu.edu (N. Habash).
Peer review under responsibility of King Saud University.

Production and hosting by Elsevier

http://dx.doi.org/10.1016/j.jksuci.2014.06.010
1319-1578

2. Arabic language facts and challenges

In this section, we discuss the challenges of processing Arabic in general and dialectal Arabic (DA) in particular.

2.1. Arabic linguistic challenges

The Arabic language is quite challenging for NLP. Arabic is a morphologically complex language that includes rich inflectional morphology, expressed both templatically and affixationally, and several classes of attachable clitics. For example, the Arabic word وسيكتبونها (w + s + y − ktb − wn + hA[1], 'and they will write it') has two proclitics (+ و w+, 'and,' and +س s+, 'will'), one prefix (–ي y–, '3rd person'), one suffix (ون- –wn, 'masculine plural') and one pronominal enclitic (ها+ +hA, 'it/her'). Additionally, Arabic is written with optional diacritics that specify short vowels, consonantal doubling and the nunation morpheme. The absence of these diacritics together with the language's rich morphology lead to a high degree of ambiguity: e.g., the Buckwalter Arabic Morphological Analyzer (BAMA) (Buckwalter, 2004) produces an average of 12 analyses per word. Moreover, some Arabic letters are often spelled inconsistently, which leads to an increase in both sparsity (multiple forms of the same word) and ambiguity (the same form corresponding to multiple words), e.g., variants of Hamzated Alif, أ Â or إ Ă, are often written without their Hamza (ء '): ا A; and the Alif-Maqsura (or dotless Ya), ى ý, and the regular dotted Ya, ي y, are often used interchangeably in word final position (ElKholy and Habash, 2010). Arabic complex morphology and ambiguity are handled using tools for analysis, disambiguation and tokenization (Habash and Rambow, 2005; Diab et al., 2007). In this article, we focus on the problem of morphological analysis, which is concerned with identifying all and only the possible readings (or analyses) for a word out of context (Habash, 2010).

2.2. Dialectal Arabic challenges

Contemporary Arabic is a collection of varieties: MSA, which has a standard orthography and is used in formal settings, and DAs, which are commonly used informally and with increasing presence on the web but do not have standard orthographies. There are several DA varieties that vary primarily geographically, e.g., Levantine Arabic, Egyptian Arabic, and so on (Habash, 2010). DAs differ from MSA phonologically, morphologically and, to a lesser degree, syntactically. The differences between MSA and DAs have often been compared to those between Latin and the Romance languages (Habash, 2006). The morphological differences are most noticeably expressed in the use of clitics and affixes that do not exist in MSA. For instance, the Levantine and Egyptian Arabic equivalent of the MSA example above is وحيكتبوها, (*w + H + y − ktb − w + hA*, 'and they will write it').[2] The optionality of vocalic diacritics helps hide some of the differences resulting from vowel changes; compare the diacritized forms: *wHayuktubuwhA* (Levantine), *waHayiktibuwhA* (Egyptian) and *wasayaktubuwnahA* (MSA) (Salloum and Habash, 2011). It is important to note that Levantine and Egyptian differ significantly in phonology, but the orthographical choice of dropping short vowels bridges the gap between them. For extended discussion about the difference between the two dialects, we refer the reader to the following books: Omar, 1976; Abdel-Massih et al., 1979; Cowell, 1964. In this work, we focus on processing text, and therefore, we do not model short vowels.

All of the NLP challenges of MSA described above are shared by DA. However, the lack of standard orthographies for the dialects and their numerous varieties poses new challenges (Habash et al., 2012a). Additionally, DAs are rather impoverished in terms of available tools and resources compared to MSA; e.g., there are very few parallel DA-English corpora and almost no MSA-DA parallel corpora. The number and sophistication of morphological analysis and disambiguation tools for DA are very limited in comparison to those of MSA (Duh and Kirchhoff, 2005; Habash and Rambow, 2006; Abo Bakr et al., 2008; Habash et al., 2012b). MSA tools cannot be effectively used to handle DA: Habash and Rambow, 2006 reported that less than two-thirds of Levantine verbs can be analyzed using an MSA morphological analyzer and Habash et al., 2012b reported that only 64% of Egyptian Arabic words are analyzable using an MSA analyzer.

Salloum and Habash (2011) reported that 26% of out-of-vocabulary (OOV) terms in dialectal corpora have MSA readings or are proper nouns. The rest, 74%, are dialectal words. They classify the dialectal words into two types: words that have MSA-like stems and dialectal affixational morphology (affixes/clitics) and those that have dialectal stems and possibly dialectal morphology. The former set accounts for almost half of all OOVs (49.7%) or almost two-thirds of all dialectal OOVs. In this article, like Salloum and Habash, 2011, we only target dialectal affixational morphology cases, as they are the largest class involving dialectal phenomena that do not require extension to stem lexica.

3. Related work

There has been a large amount of works on Arabic morphological analysis with a focus on MSA (Beesley et al., 1989; Kiraz, 2000; Buckwalter, 2004; Al-Sughaiyer and Al-Kharashi, 2004; Attia, 2008; Graff et al., 2009; Altantawy et al., 2011; Attia et al., 2013). In comparison, only a few efforts have targeted DA morphology (Kilany et al., 2002; Habash and Rambow, 2006; Abo Bakr et al., 2008; Salloum and Habash, 2011; Mohamed et al., 2012; Habash et al., 2012b; Hamdi et al., 2013).

Efforts for modeling dialectal Arabic morphology generally fall in two camps. First are the solutions that focus on extending MSA tools to cover DA phenomena. For example, Abo Bakr et al., 2008 and Salloum and Habash, 2011 extended the BAMA/SAMA databases (Buckwalter, 2004; Graff et al., 2009) to accept DA prefixes and suffixes. Such efforts are interested in mapping DA text to some MSA-like form; as such, they do not model DA linguistic phenomena. These solutions are fast and cheap to implement.

The second camp is interested in modeling DA directly. However, the attempts at doing so are lacking in coverage in one dimension or another. The earliest effort on Egyptian that

[1] Arabic transliteration is in the Habash-Soudi-Buckwalter scheme (Habash et al., 2007).

[2] A spelling variation for this Egyptian Arabic word is وهيكتبوها *w + h + y − ktb − w + hA*.