Diacritization: A Challenge to Arabic Treebank Annotation and Parsing

Mohamed Maamouri, Ann Bies, Seth Kulick

Linguistic Data Consortium, University of Pennsylvania, USA

Presenter Name: Al-Elaiwi Moh'd.

Table of Contents

Arabic Diacritization

- **Arabic diacritization** (referred to sometimes as vocalization(اللفظ) or vowelling (حروف العلة)), defined as the full or partial representation of short vowels, *shadda* (*consonantal (حرف صحيح) length* or germination(الازدواج)), tanween (*nunation or definiteness(الوضوح)), and hamza (the glottal stop حبس) (الصوت and its* support letters), is still largely understudied in the current NLP literature.

Keywords

- *Arabic NLP,*
- *Arabic diacritics (العلامة الصوتية المميزة),*
- *Diacritization,*
- *Modern Standard Arabic (MSA),*
- *Treebank's,*
- *Linguistic Annotation, (التذييل)*
- *Parsing,*
- *Orthographic (املائي)*

1. Introduction

- This paper focuse on two major challenges, not necessarily shared with many other natural languages:
  1. complex linguistic structure
  2. the specific features of its orthographic system.

- The Arabic orthographic system uses superscript and subscript diacritical marks (or diacritics) for the representation of the three short vowels (a, i, u), and four letters (ا **'alif,** ى **'imaala,** و **waaw, and** ي **yaa'**) **to** mark vocalic length.

- Short vowels are also used to indicate mood, aspect and voice endings for verbs and case endings for nouns. Moreover, long vowels are mostly used in derivation and word formation: as in kataba 'to write' vs. kAtaba 'to correspond with'. The shadda (consonantal length or gemination) is another important diacritic which is used for the derivation of new words. The hamza is used just to mark the existence of the glottal stop.

- The present paper will focus on the role and impact of diacritization on Arabic Treebank annotation and Arabic parsing.

2.Reality of Arabic Speech & Text

- The issue of diacritization in Arabic arises as the result of a mismatch between the orthographic conventions that have developed for written MSA and the Arabic language itself, including spoken MSA, with respect to the amount of linguistic information represented. MSA is generally written without diacritics, but the language itself, and also spoken MSA, of course includes all of the features that the diacritics would represent (short vowels, consonantal gemination, etc.).

--Reality of Arabic Speech & Text

1. Importance of Diacritics
2. Diacritics and Ambiguity
3. ‘Real-World’ Arabic Text Data

2.1 Importance of Diacritics

- The use of diacritics is extremely important in setting up grammatical functions leading to acceptable text understanding and correct reading or analysis, diacritical markings are rarely present in real-world/life situations. It is true that they are rarely visible in out-of-school written documents and they do not appear in most printed materials in the Arab region.
- It is to be noted that diacritized MSA text does exist outside of the Koran in numerous sources, such as the rich and important heritage Arabic literature books.

2.2 Diacritics and Ambiguity

- (a) The loss of the internal diacritics (such as short vowels or shadda) leads to the following types of ambiguity, as exemplified in a given MSA lemma: علم **Elm. The situation of this specific** form is as follows:
  - An ambiguity within 'core' POS tags, which distinguishes between the different lexical senses of the same 'core' POS tag. Example: The bare form علم **Elm can be diacritized as** عِلم**Eilm (a noun meaning** 'science, learning') or عَلَم**Ealam, another noun meaning 'flag'.**

- (b) A second type of ‘core’ POS tag ambiguity distinguishes between different lexical senses leading to different core POS tags. The same bare form علم **Elm**, can additionally be diacritized as three different verb forms, all lexically and semantically connected. Example:
  1. عَلِمَ**Ealima for 3rd Person Masculine, Singular, Perfective Verb (MSA Verb Form I)** meaning ‘he learned/knew’;
  2. عُلِمَ**Eulima for 3rd Person Singular, Passive Verb (MSA Verb Form I)** meaning ‘it/he was learned’ and,
  3. عَلَّمَ**Eal~ama for the Intensifying, Causative, Denominative Verb (MSA** Verb Form II) meaning ‘he taught.’

- (c) Finally, a huge amount of ambiguity occurs at the structural/grammatical level, where the use of short vowels is correlated with case (nominal) and mood/aspect (verbal) information. This information is rendered by the use of one of six possible diacritics. we have the following:
  1. **عِلْمٌ/عِلْمٌEilmu/EilmN (NOM Noun + Definite and Indefinite), عِلْمَ/عِلْماًEilma/EilmAF (ACCU Noun +**
  2. Definite and Indefinite) and **عِلْمٍ/عِلْمٍEilmi/EilmK (GEN Noun + Definite and Indefinite).**
- The loss of diacritics often leads to a significant increase of linguistic ambiguity (both structural and lexical), which can only be resolved by contextual information and an enough knowledge of the language.

2.3 'Real-World' Arabic Text Data

- When we look at the availability of Arabic text data, the situation break down to the following:

1. Unvocalized/non-diacritized Arabic text for MSA (and even for newly written dialectal Arabic) seems to be the most available material for the speech research community and the main data source for all other NLP research needs (mostly in newswire form).

2. Since non-diacritized text prevails, the Arabic NLP community seems to have accepted using it as the de facto ‘real world’ information material without feeling an obligation to question its choice/use, even espousing the idea sometimes that the robustness of software algorithms can deal with the problem and reduce the negative effect of the missing information on their research.

3. The excessive cost and the usually unequal and questionable quality of human/manual diacritization have led the scientific Arabic NLP community and its sponsors to focus more on volume of unvoweled data so far.

4. Most NLP Arabic research – even research dealing with diacritization – makes use of text-based information only and makes little use of diacritics even when they exist. No significant use is made of diacritics in the audio data – even when work starts from a speech source.

3. PARSER DEVELOPMENT: How Does Diacritization Impact Parsing?

- In general, the role of diacritics in a NLP pipeline that includes parsing is very much an open question.
- There are two aspects to the problem of how the parser might utilize diacritic information. One question concerns what diacritic information might be useful for the parser. While the earlier work, used the bare text, there has been very little work examining whether a parser can make use of vocalized text.

- However, in addition to exploring which diacritic information is useful for the parser, we must also be concerned with what might be available to the parser outside the context of these experiments and outside the context of Treebank research.

4. Conclusion

- The role of diacritization in the annotation process for the Arabic Treebank is now firmly established, and this data has been available and quite useful to the scientific community. In general, however, the correct way to utilize diacritization in various Natural Language Processing tasks is more of an open question.