Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 117 (2017) 169–174

Procedia Computer Science

www.elsevier.com/locate/procedia

3rd International Conference on Arabic Computational Linguistics, ACLing 2017, 5–6 November 2017, Dubai, United Arab Emirates

Automatic minimal diacritization of Arabic texts

Rehab Alnefaie[a], Aqil M. Azmi[b,*]

[a]College of Computer Engineering and Sciences, Prince Sattam Bin Abdulaziz University, Al-Kharj 16278, Saudi Arabia
[b]Department of Computer Science, College of Computer & Information Sciences, King Saud University, Riyadh 11543, Saudi Arabia

Abstract

Modern Standard Arabic (MSA) is typically written without short vowels, which helps in clarifying the sense and meaning of the word. The short vowels are omitted since experienced Arabic readers can infer the meaning through the context. But there are cases where even the native Arabic speakers cannot resolve. The process of restoring the diacritical marks (short vowels) is known as diacritization. Most of the developed algorithms for diacritization fully restores all the markings, many of which are trivial or unnecessary. In this paper, we present a system that restores the diacritical markings where it is mostly needed, resolving the ambiguity. This is a more challenging problem than fully restoring all the diacritics. The system combines morphological analyzers and context similarities. The goal of the morphological analyzers is to generate all word candidates for the diacritics, and the model eliminates word ambiguity through a statistical approach and context similarities. Out of 80 paragraphs our system resolved 57 cases.

Peer-review under responsibility of the scientific committee of the 3rd International Conference on Arabic Computational Linguistics.

Keywords: Arabic language; Automatic vowelization; Ambiguity; Diacritization; Morphological analysis; Statistical methods

1. Introduction

Arabic can be broadly classified as classical and modern. The Classical Arabic is the pure language used by the Arabs, the language the Holy Qur'an was revealed in. The Modern Standard Arabic (MSA) is an evolving language that is set to meet the modern challenges. It is the language that we read in modern magazines, newspapers, and the Internet.

Diacritics (short vowels) are marking that are placed either above or below the letter. These markings are used to indicate the phonetic information associated with each letter, which helps in clarifying the sense and meaning of the word. A simple Arabic word علم could mean *flag*, *knowledge*, *teach* etc. MSA is typically written without short vowels. Not because it is not needed, but because the natives can easily decipher (or disambiguate) the meaning of the undiacritized word through the context. This is not the case with non-native speakers, children, those learning Arabic as a second language and computer applications, e.g. speech synthesis. However, there are cases when even the natives fail to disambiguate [3]. Consider for example (درس محمد الدرس), this short sentence could either mean "Mohammad studied the lesson", or "Mohammad taught the lesson". Our goal is to use the least number of diacritics

* Corresponding author. Tel.: +966-11-467-6574 ; fax: +966-11-467-5423.
E-mail address: r.alnefaie@psau.edu.sa (RA), aqil@ksu.edu.sa (AMA)

1877-0509 © 2017 The Authors. Published by Elsevier B.V.
Peer-review under responsibility of the scientific committee of the 3rd International Conference on Arabic Computational Linguistics.
10.1016/j.procs.2017.10.106

to complete the task of resolving the ambiguity for the native speaker. In the previous example it can be resolved using a single marking (درّس محمد الدرس), which fully resolves the ambiguity and the reader knows that the second interpretation was intended.

Many Arabic applications, such as Arabic Machine Translation (MT), may get better result when the text is vocalized. In a study, [5] noted that full diacritization scheme degrades the performance of statistical MT. However, the authors found that none of the partial diacritization impacted the performance of MT when compared to plain Arabic text (i.e. no diacritization). None of the four partial diacritization schemes considered where meant to resolve the ambiguity. We believe that, have they used our scheme, it would have positively impacted the performance of MT.

In another study on native Arab children, [8] found that the full diacritical marking had an adverse effect on both reading speed, as well as text comprehension. Though no reasoning was mentioned, but we thinks it is owed to the crowded space which puts heavy strain on the eyes. Again, our partial diacritization scheme where only ambiguity is concerned will not clutter the space, and it may help the comprehension. This we need to confirm in a separate study.

Consider the fully diacritized text, (دَرَّسَ مُحَمَّدٌ الدَّرْسَ) which means "Mohammad taught the lesson". With our diacritization scheme, it will be (درّس محمد الدرس), a much pleasant for the eyes. It is true that a fully diacritized text slows down the typist as s/he has to press SHIFT after every letter just to put the diacritical marking, our scheme slows down the typist as well. Typing is a mechanical process with little thinking involved, however, for our scheme the typist has to constantly think and struggle where to place the diacritical markings.

The aim of this work is to develop an automatic minimal diacritization system for Arabic language. The system focuses on determining the least number of diacritical markings to resolve ambiguity. This paper is organized as follows. In Section 2, we go over some of the earlier works. Our proposed algorithm is in Section 3. In Section 4, we evaluate the system, and conclude in Section 5.

2. Literature survey

There are plenty of works on Arabic diacritization. There are several categories of automatic diacritization, and these are associated with letters to be diacritized. Ahmed [2] categorizes them as full diacritization, half diacritization, and partial diacritization. In half diacritization, only the morphological-independent letters are diacritized; this is full diacritization less the end cases if it depends of the syntactic analysis, e.g. (دَرَّسَ مُحَمَّدْ الدَرس). Partial diacritization is a scheme which provides less diacritic information than half diacritization. Azmi and Almajed [3] proposed a fourth category, minimalist diacritization, with two objectives in mind. One of the objectives was to place the diacritics on certain letters so that ambiguity is fully resolved.

The techniques used to restore the diacritical marking can be classified as: rule based, statistical approach, and the hybrid approach [3]. As we said before, there is large volume of work on Arabic diacritization, so we will go over few of them. For a good survey on the subject, the reader is advised to see [3].

The rule-based often exploit human knowledge to solve the problem intelligently. The main disadvantage of this approach is its huge list of rules [9].

The statistical approach builds the model depending on the probability distribution of a sequence of words, or characters, or a segment like term frequency in the corpus. This approach requires a large corpus that is fully diacritized. It is possible the corpus may suffer from data sparseness which is due to the inflection property of Arabic, and this causes many words to remain unseen in the corpora. On the other hand, this scheme does not need any linguistic knowledge. Gal [6] proposed a system for Arabic and Hebrew diacritization using Hidden Markov Models. The system builds a bi-gram model at the word level of the training data. Then it finds the most likely path transitions of a segment using