Journal of King Saud University – Computer and Information Sciences 29 (2017) 156–163

Contents lists available at ScienceDirect

Journal of King Saud University – Computer and Information Sciences

journal homepage: www.sciencedirect.com

Morphological, syntactic and diacritics rules for automatic diacritization of Arabic sentences

Amine Chennoufi *, Azzeddine Mazroui

Department of Mathematics and Computer Science, Faculty of Sciences, University Mohamed First, B-P 717, 60000 Oujda, Morocco

ARTICLE INFO

Article history:
Received 13 January 2016
Revised 25 May 2016
Accepted 23 June 2016
Available online 9 July 2016

Keywords:
Arabic language
Automatic diacritization
Arabic diacritical marks
Morphological analysis
Smoothing techniques
Hidden Markov model

ABSTRACT

The diacritical marks of Arabic language are characters other than letters and are in the majority of cases absent from Arab writings. This paper presents a hybrid system for automatic diacritization of Arabic sentences combining linguistic rules and statistical treatments. The used approach is based on four stages. The first phase consists of a morphological analysis using the second version of the morphological analyzer Alkhalil Morpho Sys. Morphosyntactic outputs from this step are used in the second phase to eliminate invalid word transitions according to the syntactic rules. Then, the system used in the third stage is a discrete hidden Markov model and Viterbi algorithm to determine the most probable diacritized sentence. The unseen transitions in the training corpus are processed using smoothing techniques. Finally, the last step deals with words not analyzed by Alkhalil analyzer, for which we use statistical treatments based on the letters. The word error rate of our system is around 2.58% if we ignore the diacritic of the last letter of the word and around 6.28% when this diacritic is taken into account.

1. Introduction

The diacritical mark is a sign accompanying a letter to modify the corresponding sound or to distinguish the word from another homonym word. Diacritical marks are widely used in Semitic languages including Arabic, Hebrew and other languages like Urdu. The purpose of these signs is to clarify the morphological structure, the grammatical function, the semantic meaning of words and other linguistic and voice features (Debili and Achour, 1998). Diacritical marks in the Arabic texts are often absent (Farghaly and Shaalan, 2009), unlike Latin languages like French, where the presence of vowels in the texts is mandatory (the vowels in Latin languages play in most cases the same function as diacritical marks in Arabic language). Indeed, according to Habash (2010), diacritical marks are absent in 98% of Arabic texts, and an undiacritized word can have several potential diacritizations in over 77% of cases (Boudchiche and Mazroui, 2015).

Arabic diacritical marks are classified into three groups (Zitouni et al., 2006):

1) The first group consisting of three single short diacritics: “ ◌َ ” fatha, “ ◌ُ ” damma and “ ◌ِ ” kasra. Thus, by adding any of these signs with the letter “ م ” /m[1]/, we obtain the following respective sounds: “ مَ ” /ma/, “ مُ ”/mu/ and “ مِ ”/mi/.
2) The second group represents the doubled case ending diacritics (called tanween): “ ◌ً ” tanween fatha, “ ◌ٌ ” tanween damma and “ ◌ٍ ” tanween kasra. These diacritical marks are reserved only for the last letter of nominal words (nouns, adjectives and adverbs). This phenomenon, called nunation”, has the phonetic effect of adding an N” sound after the corresponding short vowel at the word ending. Thus, the letter “ م ” /m/ with these three signs gives the following sounds: “ مًا ” /mF/ (man), “ مٌ ” /mN/ (mon) et “ مٍ ” /mK/ (min).
3) The third group is called syllabification marks and composed of “ ◌ّ ” shadda (geminate: consonant is doubled in duration) and “ ◌ْ ” sukun. This last group indicates the absence of a short vowel, and reflects a glottal stop while shadda reflects the doubling of a consonant and is always followed by a single diacritic or by a tanween. With the letter “ م ” /m/ and the diacritical mark fatha, we get “ مَّ ” /m∼a/.

* Corresponding author.
E-mail addresses: chennoufi.amin@gmail.com (A. Chennoufi), azze.mazroui@gmail.com (A. Mazroui).
Peer review under responsibility of King Saud University.

Production and hosting by Elsevier

[1] Buckwalter transliteration.

http://dx.doi.org/10.1016/j.jksuci.2016.06.004
1319-1578/

The diacritization operation of Arabic words occurs at two levels: morphological and syntactic levels (Diab et al., 2007). The morphological (lexical diacritics) consists of the internal diacritization of the word (the stem of the word without the last letter) and clarifies the meaning of the word. The syntactic level (casual diacritics) is interested in diacritization of the last letter of the stem and it is used to identify the syntactic role of words in the sentence. Lexical diacritics do not change with the position of the word in the sentence while the casual diacritic depends on the position of the word in the sentence. Thus, the Arabic-speaking reader should understand the Arabic text before reading it properly (Elshafei et al., 2006). This is a difficult for readers who do not have extensive knowledge of the Arabic language. Indeed, Hermena et al. (2015) studied the reaction of the readers facing the diacritized and undiacritized Arabic texts in eye-tracking experience. The results show that readers have benefited from the lifting of the ambiguity of words when diacritical marks are present.

The absence of diacritical marks is a source of complexity for automatic processing systems of the Arabic language that cannot easily determine the meaning of the sentence (Said et al., 2013). Therefore, the need for an automatic diacritization tool of Arabic is more than necessary to remove ambiguity and improve the performances of automatic processing of Arabic applications such as machine translation (Vergyri and Kirchhoff, 2004) and speech recognition (Messaoudi et al., 2004). The introduction of diacritical marks in Arabic dialect speech corpus Levantine[2] (BBN/AUB Babylon DARPA) has helped to increase its reliability and efficiency (Alotaibi et al., 2013).

In addition, the lack of diacritical marks in Arabic sentences represents the main cause of the confusion encountered during its analysis (Boudchiche and Mazroui, 2015) and (Debili and Achour, 1998). The study of Bouamor et al. (2015) showed that the automatic text diacritization increases quality manual tagging of the corpus.

The objective of this paper is to present an automatic Arabic diacritization system combining linguistic rules and statistical treatments. This article is structured as follows: the second paragraph presents the previous works on this area. The third paragraph is devoted to the presentation of the different steps of our system. Indeed, we describe the morphological analysis adopted in the first part of the system. Then, we explain the syntactic control used in the second part and some diacritical rules. We conclude this section by presenting the statistical model adopted in the third and fourth steps of the system. The fourth paragraph deals with the experimentation and evaluation system. We end this paper by a conclusion and some perspectives.

2. Related work

Automatic diacritization approaches can be classified into four categories. The first one includes approaches based only on statistical processing. The second category includes hybrid approaches using a morphological analysis followed by a statistical processing. The third category consists of hybrid approaches using morphological analysis, syntactic rules and statistical processing. The last one contains the automatic diacritization systems developed by commercial companies. Approaches based solely on the rules are rarely used because of their complexities due to the high level of ambiguity and the large number of morphosyntactic rules (Debili and Achour, 1998).

2.1. Statistics-based models

Gal (2002) was one of the first to use an approach based on hidden Markov models (HMM) for the vocalization of Semitic texts. He has tested his method on the Quran as Arabic texts and the Old Testament for the Hebrew language. The developed application does not extend to all Arabic diacritical marks. Emam and Fischer (2005) extended the statistical processing of diacritization based on examples for Statistical Machine Translation (SMT). Alghamdi et al. (2010) introduced a method based on the quad-gram at the letters. Recently, the researcher (Hifny, 2013) presented a statistical method based on n-gram and compared some smoothing techniques to treat the case of unseen transitions. More recently, Abandah et al. (2015) used a training phase based on recurrent neural networks (RNN) for automatically adding diacritical marks to Arabic text without relying on any prior morphological or contextual analysis. The diacritization is solved as a sequence of transcription problem. Their approach uses a deep bidirectional long short-term memory network that builds high-level linguistic abstractions of text and exploits long-range context in both input directions.

2.2. Morphological hybrid approaches

These approaches use both morphological analysis and statistical processing. The works of Vergyri and Kirchhoff (2004) are among the first to use these approaches. Thus, diacritical marks in the Arab conversations are restored by combining morphological and contextual information with a statistical model labeling (acoustic signal). However, they did not model the Shadda diacritic. Similarly, Nelken and Shieber (2005) presented a system that uses an automatic finite state probability, and incorporated a trigram model based on words, a quad-gram language model based on letters and an extremely simple morphological model to identify the prefix and the suffix of word. Zitouni et al. (2006) combined a statistical model based on maximum entropy with the classification of words. The input parameters of this model are the simple letter of the word and the morphological segments and the syntactic state. Habash and Rambow (2007) use the outputs of the morphological analyzer BAMA (Buckwalter, 2004) and individual taggers to choose among these outputs the most selected by these taggers. Diab et al. (2007) were inspired by the machine translation system (SMT), and they introduced six different diacritization schemes developed from observations of the naturally relevant diacritical marks. For these schemes, the morphological analyzer used was MADA (Habash et al., 2013). Recently, Bebah et al. (2014) exploited the morphological analyzer Alkhalil Morpho Sys (Bebah et al., 2011) in a process based on hidden Markov models.

2.3. Morphosyntactical hybrid approaches

These methods use both morphological and syntactic rules, and statistical processing. The architecture of the automatic diacritization system proposed by Shaalan et al. (2009) combines three approaches: automatic segmentation, part-of-speech (POS) tagging and the chunk parsing. This method is based on the lexicon of extraction, the bi-gram model and the support vector machines (SVM). The syntactic information is used to treat for each word the diacritical mark of its last letter in a separate final process. The solution, proposed by Rashwan et al. (2011) uses in the first step morphological and syntactic information from ArabMorp[3] and ArabTagger[4] tools, and then an n-gram model and the A* algo-

[2] https://catalog.ldc.upenn.edu/LDC2005S08.

[3] http://www.rdi-eg.com/technologies/Morpho.aspx.

[4] http://www.rdi-eg.com/technologies/POS.aspx.