Arabic character diacritization using DNN

Ikbel Hadj Ali, Zied Mnasri, Zied Lachiri
Signal, Image and Technology of Information Laboratory, University Tunis El Manar, Tunisia
https://doi.org/10.36505/ExLing-2018/09/0011/000344

Abstract

In this paper, automatic Arabic character diacritization is more accurately achieved using deep neural networks. Actually, though diacritic signs represent short vowels and/or indicate gemination on consonants, they are omitted in modern standard Arabic (MSA). However, most speech processing applications like speech synthesis and machine translation need such marks to convey the right meaning. Therefore in this work, automatic diacritization accuracy is enhanced using feedforward DNN. The results show that using more significant and Arabic-specific input features increases the prediction accuracy of diacritic signs.
Key words: Arabic characters, diacritic signs, feedforward DNN, input features.

Introduction

Modern standard Arabic (MSA) is natively spoken by more than 300 million people in the MENA region (Middle East and North Africa). Therefore, it's urgent to catch up the development of NLP (Natural Language Processing) to keep Arabic present in novel speech technology products. In Arabic, short vowels and geminated consonants are indicated by diacritic signs. However, in contrary to classical literary Arabic, diacritic signs are mostly omitted in modern standard Arabic MSA. Arabic speech processing applications, especially speech synthesis and machine translation, need accurate diacritization to convey the right meaning. This problem was addressed a particular attention, in many related works using linguistic and grammatical rules. Therefore, in this paper, machine learning, and especially deep neural networks are used to increase the accuracy of Arabic character diacritization.

The rest of this paper is organized as follows: section 2 introduces the Arabic phonology and linguistics; section 3 presents a brief description of the deep neural networks used in this work; section 4 shows the speech material, the conducted experiments and the yielding results. Finally the findings are discussed and commented.

Arabic diacritics

Arabic is a Semitic language which has the advantage of having a single literary version. Though there are many dialects (colloquial Arabic) which differ, not only from one country to another, but also from one region to another in the same country, there's only one standard version, which is used in literature,

journalism and science. This standard version, called MSA (Modern Standard Arabic) had inherited from Classic Arabic, which used to be the literary version since the middle ages.

Arabic diacritic signs

Arabic has 28 consonants and three vowels. One of the specific characteristics of Arabic is the ability to double all consonants (gemination) and to lengthen all vowels. However, this leads to changing the word meaning, e.g. the word *"darasa"* دَرَسَ means (to study), whereas with a geminated *"r"* it becomes *"darrasa"* دَرَّسَ (to teach) and with a long final "a" it changes to *"darasa:"* دَرَسَا (to study with somebody else).

Arabic alphabet does not include special letters for short vowels. However, these vowels are marked on consonants using three diacritic signs, i.e. *"fatha"* for /a/, *"dhamma"* for /u/ and *"kasra"* for /i/. Doubling theses diacritics at the end of a noun indicates the indeterminate form (cf. Table 1). Besides, *"sukun"* and *"shaddah"* are used to indicate stop and gemination, respectively.

In addition, different ways of diacritization of the same word change the meaning, e.g. the word consonant-based root ("d,r,s" د,ر,س) may be pronounced *"darasa"* دَرَسَ (to study) or *"durisa"* دُرِسَ (to be studied) depending on the diacritics representing the vowels introduced after each consonant

Arabic diacritization systems

Many techniques were used to achieve this task. However, all of them can be divided into three main categories, i.e. rule-based, model-based and data-driven techniques. Rule-based techniques rely on the implementation of linguistic rules to determine the diacritic sign of each character (Halabi 2016), whereas model-based techniques use n-gram language models to predict the diacritic sign of each character (Habash 2009). More recently, with the development of data-driven prediction tools, probabilistic learning techniques like HMM (hidden Markov models) were applied to predict the diacritic sign based on a set of contextual features (Rashwan 2011).

DNN in speech processing

DNN are nowadays used in most speech processing applications, such as speech synthesis, speech recognition and machine translation. In speech processing, DNN can be used either for regression, to predict continuous values, such as segment duration or fundamental frequency (pitch) values, or for classification tasks, such as segment voicing decision or diacritization sign prediction. Therefore, it has been recently performed using DNN (Rebai and Ben Ayed 2015). However results need to be enhanced using more significant characteristic features.

Experiments

Implementation

A feedforward DNN using 2 hidden layers and sigmoid activation function was used for Arabic character diacritization. In the preprocessing phase, the selected input features (cf. Table 1), were transformed into a specific code for each, whereas output targets (diacritic signs) were encoded using one-against-all code, since the task is multi-class classification. During the learning process, early stopping option was used to prevent over-fitting.

Table 1. Features classification and coding.

Feature	Value	Coding	Nodes
Identity (of previous/ current/next letter)	/a/ا , /b/ب, /t/ت …etc.	One against all	36
Type (of previous/ current/next letter)	Plosive, fricative, nasal, trill, lateral, semi-vowel…	One against all	8
Gemination (of previous/ current/next letter)	Yes/No	Binary	1
Relative position of current letter in the word	beginning/middle/ end	Coarse coding	3
Relative position of current word in the sentence	beginning/middle/ end	Coarse coding	3
Content word	Yes/no	Binary	1

A database of 28737 sentences fully-diacritized containing ca. 1.6 million characters was used. 80% of the database was allocated for training whereas the remaining 20% were used for validation and test.

Results and discussion

To assess the accuracy of DNN-based diacritization, two measures were used, i.e. total accuracy rate (TAR), calculated all over the characters in the validation set, and class-wise accuracy rate (CAR), which is calculated for each single diacritic sign using the confusion matrix. The best DNN model, tested on the validation set, has given a TAR of 84.4%. Furthermore, the matrix confusion was calculated to extract the CAR for each diacritic sign. For some signs, like *sukun (a stop on a consonant)* the CAR has reached more than 90% (cf. Table 2). Also, for letters which need no diacritic signs (like long vowels, e.g. /a:/ ا, /u:/ و and /i:/ ي) the CAR related to the class *None* was very high, 92.7% (cf. Table 2).

Table 2. Accuracy results of DNN-based diacritic signs prediction model.

Diacritic sign	Tested samples	Recognized samples	Accuracy
Fatha /a/ (ـَ)	62596	43134	68.9 %
Dhamma /i/ (ـِ)	22068	15509	70.2 %
Kasra /u/ (ـُ)	112449	98873	87.9 %
Fathaten /an/ (ـً)	269	109	40.5 %
Dhammaten /un/ (ـٌ)	14	9	64.3 %
Kasraten /in/ (ـٍ)	1170	888	75.9 %
Sukun (stop) (ـْ)	41188	37339	90.6 %
None	80246	74438	92.7 %
Total	320000	270299	84.4 %

However, the prediction accuracy needs to be enhanced for some classes like *kasra* /i/, *fatha* /a/ and *dhamma* /u/ that are essential to understand the meaning of the word. Also, less abundant diacritics like *fathaten* /an/ (ـً), *dhammaten* /un/ (ـٌ) and *kasraten* /in/ (ـٍ) need to be better modelled to enhance their prediction accuracy. Therefore, mono- and bi-directional long short term memory (LSTM and B-LSTM) deep neural networks might be used to take advantage of the recurrent aspect of speech.

References

Habash, N., Rambow, O., Roth, R. 2009. MADA+ TOKAN: A toolkit for Arabic tokenization, diacritization, morphological disambiguation, POS tagging, stemming and lemmatization. In Proceedings of the 2nd international conference on Arabic language resources and tools (MEDAR), Cairo, Egypt (Vol. 41, p. 62).

Halabi, N., Wald, M. 2016. Phonetic inventory for an Arabic speech corpus. In Proceedings of the Tenth International Conference on Language Resources+ and Evaluation (LREC 2016), Slovenia, 734-738.

Rashwan, M.A., Al-Badrashiny, M.A., Attia, M., Abdou, S.M., Rafea, A. 2011. A stochastic Arabic diacritizer based on a hybrid of factorized and unfactorized textual features. IEEE Transactions on Audio, Speech, and Language Processing, 19(1), 166-175.

Rebai, I., BenAyed, Y. 2015. Text-to-speech synthesis system with Arabic diacritic recognition system. Computer Speech & Language, 34(1), 43-60.