A Novel Approach for the Recognition of a wide Arabic Handwritten Word Lexicon

Imen Ben Cheikh, Abdel Belaïd, Afef Kacem

A Novel Approach for the Recognition of a wide Arabic Handwritten Word Lexicon

I. Ben Cheikh[1], A. Belaïd[2], and A. Kacem[3]

[1,3]*UTIC-ESSTT, 5 Avenue Taha Hussein, BP56 Mnara,Tunis, Tunisie*
Imen.becheikh@gmail.com, Afef.kacem@esstt.rnu.tn
[2]*LORIA, Campus scientifique, B.P. 239, 54606 Vandoeuvres-Lès-Nancy, France*
abelaid@loria.fr

Abstract

This paper introduces a novel approach for the recognition of a wide vocabulary of Arabic handwritten words. Note that there is an essential difference between the global and analytic approaches in pattern recognition. While the global approach is limited to reduced vocabulary, the analytic approach succeeds to recognize a wide vocabulary but meets the problems of word segmentation especially for Arabic. Combining the neuronal approach with some linguistic characteristics of the Arabic, it is expected that we become able to recognize better and to handle a large vocabulary of Arabic handwritten words. The proposed approach invokes two transparent neuronal networks, TNN_1 and TNN_2, to respectively recognize roots, schemes and the elements of conjugation from the structural primitives of the words. The approach was evaluated using real examples from a data base established for this purpose. The results are promising, and suggestions for improvements are proposed.

1. Introduction

The review of previous studies dealing with word recognition reveals that the global approaches remain limited to reduced vocabulary of words. In return, the analytic approaches permit the recognition of wide vocabulary and even an entire text but they front the problems of segmentation of Arabic which is cursive by nature [1]. Some studies have led to a new class of approaches, qualified by pseudo-global, which make use of the PAW (Piece of Arabic Word), an inherent entity of the Arabic writing [2]. This new paradigm allows the recognition of a slight wide vocabulary, compared to global approaches, but it is still limited to a lexicon of Paws, defined in advance.

As the handwritten words are difficult to segment, we have abandoned the use of analytic approaches. In addition, the global and pseudo-global approaches seem to be not sufficient to deal with a large vocabulary (more than 10000 words). It has also been observed that statistical, structural, neuronal and other traditional methods perform poorly for Arabic, prove not so effective to take into account all the morphological variations of Arabic and to overcome the difficulties encountered in the case of the handwritten. To remedy such weakness, we believe that the most promising approaches are those integrating linguistic information in various levels: lexical, syntactic and semantic levels to improve robustness of the recognition against the variation of morphology. We propose, in this paper, the use of a neuronal-linguistic approach first, to avoid segmentation and second, to deal with a wide lexicon. In this work, we will mainly focus on the recognition of a large canonic vocabulary composed of words deriving from tri-consonant healthy roots. The recognition of the rest words will be considered in a future work.

Our neuronal-linguistic approach is based on two transparent neural networks, equipped with linguistic knowledge, and specialized in the recognition of the root - from which the word derives - and the scheme (or template) that the word follows. The word is then automatically reconstituted from its root and scheme.

The rest of this paper is organized as follows. First of all, we will give a brief Arabic linguistic analysis focusing on the current application. Afterwards, we will discuss some problems which have to be overcome using neuronal-linguistic approach. Next,

we will explain how our approach works through use of an illustrative example. A detailed experiment is carried out and successful recognition results are reported. We will finish this paper with some concluding remarks.

2. Arabic linguistic analysis

This section highlights the characteristics of the Arabic language which should offer solutions to the problem of vocabulary size. The Arab language is known for the richness of its morphological structure in sense of stable linguistic concepts specific to this language. Note that, an Arabic word is either decomposable or non-decomposable. عشرة, فرنسا and دكتور are example of non-decomposable words. A decomposable word is composed of morphemes: a prefix, a radical and a suffix. The radical or the verbal core is the derivation of a root according to a given scheme by introducing infix letters. A root is either tri-consonant (three letters) or quadri-consonant (four letters). Besides that, a root is either healthy or non-healthy (contains a vowel at least). The prefix and the suffix correspond to:
- The conjugation:
  . Time (accomplished/unaccomplished)
  . Aspect (indicative/apocopate/subjunctive/energetic)
  . Mode (real/potential)
  . Voice (passive/active)
- The person:
  . Subject (1st person/2nd person/3rd person)
  . Kind (female/male)
  . Number (singular/duel/plural)
- The fact that the root is healthy or not healthy.
- The fact that the noun is definite or indefinite (eg. مدرسة, المدرسة , the school, a school).

The number of schemes can go up to 70 (eg. فعل, 808 منفعل, افتعال, مفعول, استفعال, مفاعل, مفعال, تفاعل). healthy tri-consonant roots can generate a lexicon of 98000 words [1]. On average, 80 frequently used words can derive from a given root in various schemes [3].

A decomposable word follows a given scheme depending on whether it is a verb (كتب:فعل, to write), a noun of agent (كاتب: إسم فاعل, an author), a noun of patient, (مكتوب : إسم مفعول, written), a noun of machine (كتاب: إسم آلة) , كتابة : مصدر) a book), a verbal noun (writing), a noun of place (مكتبة: إسم مكان, a library), a broken plural (كتب : جمع غير سالم, books).

We would like to mention that not any root fits with any scheme. Some coherence rules must be checked before using a scheme for one root. For example, the similar adjective (صفة مشبهة) and the superlative (تفضيل اسم) can not be used with the verb كتب.

3. Neuronal model: How to benefit from the Arabic fertility and linguistic stability?

In our approach, we adopted a technique quite different from the previous ones. Our study is started with a corpus of decomposable words deriving from tri-consonant healthy roots and showing large variety in writing.

To benefit from the properties of Arabic, we initially conceived a Transparent Neural Network, named TNN_0, which enables, from a word structural primitives description, to find out its root and conjugated scheme. Notice that we have already experiment the use of TNN on a reduced vocabulary as reported in [4]. We concluded that such network provides fast analysis and correction of results. It also permits a better modeling of root and scheme concepts. To not burden the output layer of TNN_0, by the big number of conjugated schemes, we reduced their corresponding neurons knowing that a conjugated scheme (eg. يتفاعلون) is defined by a brief scheme (a non-conjugated one, eg. تفاعل) and the elements of its conjugation (time, aspect, mode, voice, subject, kind, number, definite/indefinite and type (verb/noun…)). Thus, TNN_0 has as many layers as decomposition levels in Arabic (primitives, letters, PAWs, words). As shown in figure 1 (a), the output layer of TNN_0 has as many neurons as used roots, brief schemes and conjugation elements. Note that the number of roots depends on the treated lexicon, but the number of schemes and conjugation elements does not exceed 80.
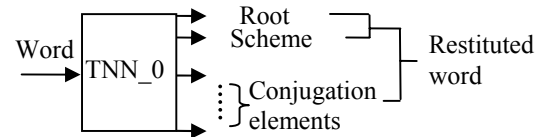
(a) TNN_0 architecture
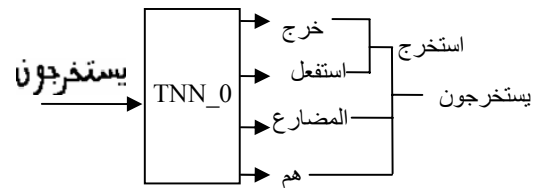
(b) A training example by TNN_0
Figure 1: TNN_0, from the word its root and conjugated scheme

Given a lexicon of 10000 words, generated from 200 roots (50 words per root), such network succeed in reducing the output layer neurons from 10000 to 280 (200 root neurons and 80 scheme ones). But it is convenient to note that the lexicon is generated by too

many PAWs. The question now: is there a possibility to decrease the number of PAWs especially when it might easily exceed tens of thousands? The response is no because PAWs do not have any linguistic meaning that allows their factorization. For that reason, we propose now, to divide TNN_0 into TNN_1 and TNN_2 to train then recognize, respectively, words roots and schemes. We believe in dissociating roots from schemes since these two entities do not require the same information to be learned and recognized. For instance, the information about PAWs of a word, are useless for the training of its root but useful for the training of its scheme.

3.1. TNN_1: Roots training from words structural primitives

TNN_1 is a three-layer network (primitives: 70 neurons, letters: 117 neurons and roots: 200 neurons). As shown in the figure 2, TNN_1 learns 1) how to ignore prefixes, infixes and suffixes and 2) how to take into account only of the significant letters of the word that correspond to root letters.

The primitives are characteristics combinations. A characteristic is a shaft (H), a down stroke (J), a buckle (B), high diacritical points (P), low diacritics (Q), nothing of what precedes (R) and a position in a PAW: beginning (D), middle(M), end (F) or isolated(I). Let us explain the approach considering the figure 2. In this figure, TNN_1 outputs the root: كثر from the primitives sequence: PD-HM-HF-PD-JF of the word: تكاثر.
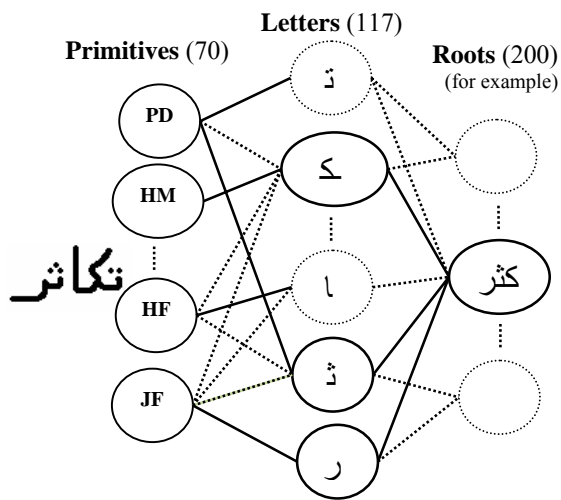

Figure 2: TNN_1, from structural primitives to roots

3.2 TNN_2: Schemes and conjugation elements training from words structural primitives

TNN_2 is a four-layer network (primitives, letters, schemes' PAWs and schemes). As presented in figure 3, the 3rd layer models the PAWs of Arabic schemes and not the PAWs of Arabic words. We choose to conceive neurons for schemes' PAWs for two reasons. First, in this network, a word PAW in itself does not have as much importance as its form. Second, this choice allows reducing necessary neurons from tens of thousands to 80. Contrary to TNN_1, TNN_2 learns how to ignore the root letters and focus on the PAWs templates so that to find out the scheme the word follows. The symbol: *, in the neurons of schemes' PAWs layer, is used to replace any letter respecting the context of the neuron. For example, the neuron: ذ*ا replaces all words parts having this form. Consequently, * cannot be any letter, but a letter in the middle since it is attached to ا and ذ. For the example of the word: تكاثر, the scheme: تفاعل and the conjugation elements: male and singular in accomplished time are given by TNN_2 as shown in the figure 3.

4. Word recognition

After this empirical training stage using real examples, the networks must recognize new handwritten words. Structural primitives of words are provided to the TNN_1 and TNN_2 entries which will elect the corresponding root and scheme. Notice that TNN_1 can easily output ten or so of candidate roots since several roots are composed of the same letters (see figure 4). An analysis of coherence between schemes and the candidate roots should eliminate the bad propositions. Obviously, perceptive cycles could be taken in place to remove ambiguities. As shown in figure 4, with the proposed root and scheme, words are constituted and the distance between their structural primitives and those of the word, given in entry, can be calculated which can guide toward the right candidate.
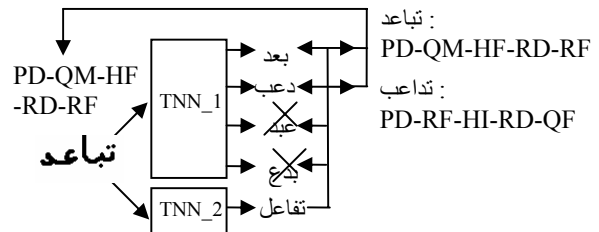

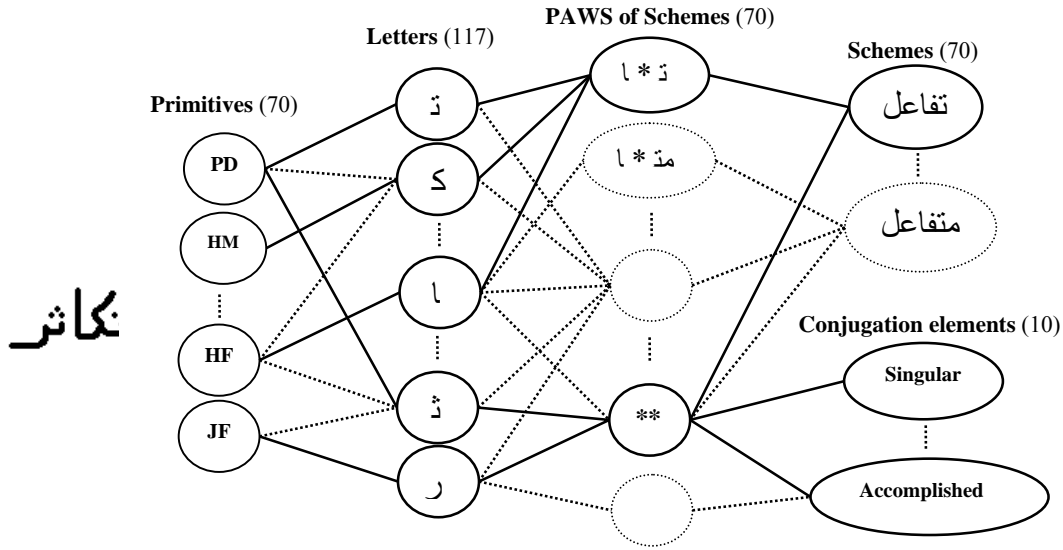**Figure 4: word restituting and ambiguity resolution**

Figure 3: TNN_2, from structural primitives to schemes and conjugation elements

5. Experimentations

As we do not have yet a large base of Arabic handwritten words, we studied 30 different Arabic fonts. The chosen fonts show so large variety in writing that words seem to be handwritten. To validate our approach, we worked, in this paper, on the training of a vocabulary composed of 1500 words derived from 50 roots using 25 schemes and considering the conjugation elements. The learning step of our system consists to train TNN_1 and TNN_2 separately. For TNN_1, the objective is to train 50 roots from their structural primitives. For the samples, chosen to learn roots, we used the different words from which they derive rather than their different scripts as commonly done. We trained TNN_1 on 1500 words: 30 samples per root. At the same time, the training of TNN_2 consists of learning 25 schemes from the structural primitives of words. Training samples are words following those schemes and deriving from different roots. The same training corpus is used for the both networks. For the recognition step, we used other corpus composed of 750 words. Following the proposed approach, recognition rates of **87%** and **75%** have been respectively achieved for TNN_1 and TNN_2 in the present experiment. That is without considering coherence rules and perceptive cycles.

6. Conclusion and future work

This paper described shortly a neuronal-linguistic approach for the recognition of a wide lexicon of Arabic handwritten words. It used features some of which were not explored in the previous studies [5]. The proposed approach provides, both of networks, TNN_1 and TNN_2, with linguistic knowledge and reduces the training stage which remains necessary because of styles and writing conditions variability. Experiments, involving a test set of considerable size show encouraging results. To end this paper, we would like to mention the importance of the perceptive cycles to improve the obtained recognition rate. For the same purpose, we suggest to use TNN_1 and TNN_2 just to return the radical of the word (eg. the radical of the word يتكاثر ون is تكاثر). Next, having the radical, we will only need to dissociate it from the prefix and suffix parts of the word in entry. A subsequent treatment consists in recognizing those parts by checking some coherence rules used in the Arabic language. These subjects are left to future research work.

References

[1] S. Kanoun. Identification et Analyse de Textes Arabes par Approche Affixale. PHD, Univ. de Sciences et techniques de Rouen, 2002.

[2] A. Belaïd and C. Choisy. Human Reading Based Strategies for off-line Arabic Word Recognition.

[3] A. Ben Hamadou. Vérification et Correction Automatiques par Analyse Affixale des Textes Ecrits en Langage Naturel. PHD, Univ. de Sciences, des Techniques et de Médecine de Tunis, 1993.

[4] I. Ben Cheikh and A. Kacem. Neural Network for the Recognition of Handwritten Tunisian City Names. ICDAR'07, pp 1108-1112, September 2007.

[5] M. Cheriet and M. Beldjehem. Visual Processing of Arabic Handwriting: Challenges and New Directions. SACH'06, Sepember 2006.