Arabic natural language processing: handwriting recognition

Abdel Belaïd

Arabic natural language processing

A. Belaïd, Loria

Introduction

The automatic recognition of Arabic writing is a very young research discipline with very challenging and significant problems. Indeed, with the air of the Internet, of Multi-media, the recognition of Arabic is useful to contributing like its close disciplines, Latin writing recognition, speech recognition and Vision processing, in current applications around digital libraries, document security and in numerical data processing in general.

Arabic is a Semitic language spoken and understood in various forms by millions of people throughout the Middle East and in Africa, and it is used by 234 million people worldwide. Furthermore, Arabic gave rise to several other alphabets like Farsi or Urdu increasing much the interest of this script. Farsi is the main language used in Iran and Afghanistan, and it is spoken by more than 110 million people, concerning also some people in Tajikistan, and Pakistan. Urdu is an Indo-Aryan language with about 104 million speakers. It is the national language of Pakistan and is closely related to Hindi, though a lot of Urdu vocabulary comes from Persian and Arabic, which is not the case for Hindi. Urdu has been written with a version of the Perso-Arabic script since the 12th century and is normally written in Nastaliq style.

We intend to address here only the problem of cursive handwriting of Arabic script (Farsi and Urdu also when pertinent examples allow it) through which, we would like to expose some processing techniques and strategies. We will explain how the characteristics of theses scripts impose choices different from Latin. Lastly, although under certain aspects one can find similarities between Arabic and Latin, the approaches are different. Indeed, being Arabic is calligraphic and non syllabic, where the responsibility of the writer is always engaged, that does not allow us to follow the traditional recognition schema: top down or bottom-up with a classical segmentation in lexical components.

We will show how the morphological analysis of the Arabic writing can initially improve the precision of the recognition methods. Then, we will show that the cognitive approaches, close to the human reading models, are mostly capable to bring the best solutions. We mainly focus on the Interactive Activation (IA) Model of McClelland & Rumelhart [1] which assumptions concerning the visual analysis make it close to the human brain reasoning. The architecture, proposed for its deployment, is similar to a neural network which facilitates its implementation. Indeed, the recognition methods are initially classified, then compared to the IA reading models and some advices are given for their rapprochement.

The last part of the paper addresses the linguistics aspect. The natural Arab processing could provide at the same time what missed in managing the error correction during the phase of post-processing and also with the representation of the lexicons and the use of large vocabularies.

New research frontiers in Arabic Handwriting Recognition

1. Segmentation in regularities and singularities rather than in letters or graphemes

Due to local variability of Arabic, It is widely accepted that Arabic word segmentation in letters is very delicate and not always ensured. Even in Latin, this idea becomes real.

Usually in most attempts, Arabic word is segmented into graphemes (copied on Latin). This is also an error in most cases. This is why researchers oriented their segmentation approaches towards more language peculiarities as singularities and regularities.

In fact, letters within a word are joined even in machine-print. Letter shape and whether or not to connect depend on the letter and its neighbours. The connection length is variable providing to a word a variable length allowing it occupying more or less writing spaces. However, some markings (regularities) offer a sort of help in situating the letters in the word. First, diacritics and accents take priority when deciphering letters which have similar base shapes. Second, markings indicate doubled consonants like "hamza," "shadda," and "madda".

At side of the marks, morphology can also contribute to the segmentation. In fact, letters are connected at the same relative height. The "baseline" is the line at the height at which letters are connected. Letters are wholly above it except for descenders and some markings. For handwriting, the baseline is an ideal concept (separating, the zone of singularities in the bottom and the zone of the regularities in the top) and a simplification of actual writing. In practice, connections occur near, but not necessarily on, such a line. Arabic has several standard ligatures, which are exceptions to the above rules for joining letters. Most common is "laam-alif," the combination of "laam" and "alif," and others include "yaa-miim" and "laam-miim." [2]

2. Cognitive model rather than classical recognition approach

Word recognition implies the processing of visual data and its interpretation at the linguistic level. Psychologists call "mental lexicon access", the process by which the human associates the image of a word to its significance. Three models emerged [3]. The first and the second are representative of a traditional school, while the last is an adaptation of "brain-style" simulation models and fits in the neural approaches. This model, named "Interactive Activation Model" [1], is a model where the perception takes place in a multilevel processing system with three levels of representation: visual features, letters and words represented by neurons. Neurons have excitatory connections for those having the feature, inhibitory weights else. The processing combines both bottom-up and top-down information allowing the reader to use his knowledge to complete.

3. IA and Arabic Recognition

Arabic writing fits very well the reading principle of IA clearly: it privileges the superiority of the Word, and it exploits the local perceptual information to help the word understanding. But the corresponding model needs to be adapted to consider the PAW intermediate reading level and letter distortions. PAWs introduce an intermediate global level of information while letter shape variations make more complex their localization and their modelling. Hence, there is a perfect similarity if it is adapted.

This model was applied to the recognition of numerical amounts of cheques by Côté in 1998 [4], by Pasquier in 2000 [5] with a more significant contextualisation of information within the three hierarchy levels. The concept, set up within the framework of this model, is interesting owing to the fact that the mechanisms of interaction between the levels of interpretation (contextual) can make it possible to exploit more "easily" contextual knowledge at various levels.

We also used this model in the system of Maddouri in 2002 [6] for the recognition of Tunisian city names. However, this first model wasn't trained but comprises a first tentative of correction of the feature level brought by the word level in case of ambiguity. This version was improved twice: one by Rangoni in 2007 [7] by proposing the notion of input selection,

context return, perceptive cycles and dynamic reasoning, and one other now by Ben Cheik [8] by incorporating over the word level linguistic levels allowing the model working on large vocabulary.

4. Arabic Recognition Methodologies Regarding the IA Model

Considering human perception of Arabic writing with the particularity of PAW, we revised the literature approaches at a vision level and examined their proximity with IA model:

- **Global Vision Classifier**: In this holistic approach, the word is regarded as a whole, allowing correlations to the totality of the pattern. Researchers adapt features used for Latin and uses either Neural Networks or Hidden Markov Models [9]. The non specialisation of the features leads to make many local arrangements and interpretations. Although it clearly accredits the word superiority principle of IA, it moves away them from model IA. Its use remains limited to small vocabularies because its complexity grows linearly with the amount of word models. Furthermore, only two levels are considered: input features and word which needs more precision in the feature extraction. Finally, there needs adaptation when high level features are used because it is necessary to use the characteristics of the language.

- **Semi-Global-Based Vision Classifier**: In this approach, the analysis is reported on an intermediate level between letter and word. This fits well in Arabic thanks to the PAW considered as a natural segmentation of the word allowing refining the analysis by reducing the basic vocabulary and extending by consequence the lexicon. Ben Amara, and al. illustrates this fact in [10] where the PAW-level deals with a moderate vocabulary of city names, usually not treatable with a global word approach. Curiously, few works have concerned the PAW-level in the literature, may be, due to the influence of Latin works with the tendency to recognize whole words.

- **Local-Based-Vision Classifier**: In this vision level the objective focuses on letters or smaller entities for their interpretation. The process is to gather, bind and confront these entities to identify the word. The Sayre dilemma practised in Latin is asked again. Several approaches [9] give good results showing that the analytic approach can perform well. But the usual drawbacks of over and under-segmentation of this approach are accentuated in the case of Arabic. Here, it is obvious that the IA model is not respecting the principle of "Word Superiority Effect".

- **Hybrid-Level Classifier**: This vision combines several strategies. This combination better fits the human reading proved to firstly analyze global word shapes and secondly to search for local information only to discriminate ambiguous cases. Maddouri and al. proposed a combination of global and local models based on a Transparent NN [6]. This model stems from the model IA adapted by Côté [4] for Latin recognition. In this model, Maddouri added a PAW level but fixed by hand the weights between neurons. More recently, Ben Cheikh [11] introduced a training step from examples and reinforced the interaction between the three levels.

Contribution of Natural Language Processing

Usually, we keep in reserve the NLP to the last step of the recognition system to correct the errors by providing some considerations stemmed from the writing language [12]. Arabic is characterized by a complex morphological structure. It is a highly inflected language where prefixes and suffixes are appended to the stem of a word to indicate tense, case, gender, number, etc. For that reason, it seems that words are not the best lexical units in this case and, perhaps, sub-word units would be a better choice. In the literature, few

researches are engaged in that direction. However, the tendency is to choose morphemes as sub-word units.

An Arabic word is decomposable or not in morphemes: (prefixes, radical and suffix). As noticed in [8] a decomposable word follows a given scheme or model of derivation, depending on whether it is a verb (كتب, فعل, to write), a noun of agent (إسم فاعل: كاتب, an author), a noun of patient, (مكتوب : إسم مفعول, written), a noun of machine (كتاب: إسم آلة, a book), a verbal noun (كتابة : مصدر, writing), a noun of place (مكتبة: إسم مكان, a library), a broken plural (كتب : جمع غير سالم, books). Te radical or the verbal core is the derivation of a root according to a given scheme (template) by introducing infix letters. A root is either tri-consonant (three letters) or quadri-consonant (four letters). The number of schemes does not exceed 75. 800 triconsonnatic roots can generate a lexicon of size higher than 90000 [13]. About 80, frequently used, words can derive from the same root.

As mentioned in Cheriet [12], the question of incorporating NLP in a recognition system is a non resolved problem. Up to day, it is not definitely clear where the incorporation of NLP is more profitable for a writing recognition system. S. Kanoun [14] uses several linguistic concepts (affix restriction, semantic restriction, etc.) to filter the assumptions and to guide the recognition process. At first, the step of word segmentation provides affixes, suffixes and radicals. Next, by decomposition of radicals in roots and infixes, using a lexical restriction, then by an affix and semantic restrictions, it was be possible to hypotheses on the word to be recognized. In W. kammoun [53], after affix decomposition and semantic filtering, she deduced roots and words. By lexical filtering of the roots, decomposable and not decomposable words are separated.

In Ben Cheikh work [8], we incorporate the morphology analysis within the IA Model. In fact, in order to not learn all the lexicon words, we propose to train the roots and the schemes in the model. For the root training, we propose a model with three layers: primitive, letter and root. This network learns how to neglect the prefixes, infixes and suffixes and learns how to just consider consonants of the roots. For scheme training, we use another IA model composed of four layers: primitive, letter, PAW of scheme and scheme) for three reasons: 1) a word PAW does not have, in itself, as much importance as its form, 2) the possibility of reducing the number of neurons, representing PAWs of words, from ten of thousands to less than one hundred of neurons representing the PAWs of schemes and 3) several PAWs of schemes contribute in the same scheme. Contrary to the first model, this network learns how to ignore the letters of root and just focus on how to determine the scheme of the word. Consequently, for the training of a vocabulary, of size 15000 for example, thanks to this neuron-linguistic approach, we will just need to learn 200 roots and less than 80 schemes. Thus, the increase of the vocabulary size does not necessary imply the increase of the corpus size since this model is able to even recognize a word that it have never been learned; it is enough to learn its root and scheme via other words.

References

[1] J. L. McClelland and D. E. Rumelhart, "An interactive activation model of context effects", Letter perception in Psychological Review, 88: pp. 375-407, 1981.

[2] L. M. Lorigo and V. Govindaraju, "Offline Arabic Handwriting Recognition: A survey", IEEE Trans. on Pat. Anal. and Mach. Int. (PAMI), vol. 28, n. 5, pp. 712-724, may 2006.

[3] Jacobs, Arthur M. & Grainger, Johnathan (1994) Models of visual word recognition: sampling the state of the art. Journal of Experimental Psychology: Human Perception and Performance 20 (6) 1311-1334.

[4] M. Côté, E. Lecolinet, M. Cheriet and C. Y. Suen, "Building a perception based model for reading cursive script", in ICDAR, vol. II, pp. 898-901, 1995.

[5] Pasquier L., Anquetil E., Lorette G., Modèle iterative d'interprétation multi-contextuelle pour la lecture d'écriture manuscrite, 1éème Congrès RFIA, Paris, vol. 3, p. 347-356, 2000

[6] S. Snoussi Maddouri, H. Amiri, A. Belaïd and Ch. Choisy, "Combination of Local and Global Vision Modeling for Arabic Handwritten Words Recognition", in 8th IWHFR, pp. 128-132. 2002.

[7] Y. Rangoni and A. Belaïd , Document Logical Structure Analysis Based on Perceptive Cycles, Lecture Notes, H. Bunke and L. Spitz editors, Nelson, New Zealand, Feb 2006.

[8] I. Ben Cheikh, A. Belaïd, and A. Kacem, A Novel Approach for the Recognition of a wide Arabic Handwritten Word Lexicon, ICPR, Tampa, USA, Dec. 2008.

[9] A. Belaïd and Ch. Choisy, Human Reading Based Strategies for off-line Arabic Word Recognition SACH'06, Summit on Arabic and Chinese Handwriting, Univ. Of Maryland, College Park, Sept. 27-28, 2006, 2006

[10] N. Ben Amara, and A. Belaid, "Printed PAW recognition based on planar hidden Markov models", Proceedings of the 13th International Conference on Pattern Recognition, vol.2, pp. 220 - 224, 25-29 Aug. 1996.

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

[12] M. Cheriet, Visual Recognition of Arabic Handwriting: Challenges and New Directions, SACH 06, College Park, MD, USA, Sept. 2006, pp. 129-136.

[13] 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.

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

[15] Wady Kammoun and Abdel Ennaji. Reconnaissance de textes arabes à vocabulaire ouvert. In 8`eme colloque international francophone sur l'écrit et le document
(CIFED'2004), June 2004.