International Journal of Engineering and Advanced Technology Studies
Vol.3, No.7, pp.55-63, September 2015
Published by European Centre for Research Training and Development UK (www.eajournals.org)

INNOVATIVE APPROACH TO TEACHING ARABIC WRITING BY USING A HANDWRITING RECOGNITION SYSTEM

Abdelkarim Mars

Laboratory LIDILEM, the University of Grenoble Alpes

ABSTRACT *This article presents the development of an Arabic online handwriting recognition system based on neural networks approach. It offers solutions for most of the difficulties linked to Arabic writing recognition. Secondly, our proposed system will be integrated in a computer assisted languagelearning tool to create educational activities and to generate adequate feedback. This environment uses tools from natural language processing, mainly our handwriting recognition system, to create various educational activities. The developed system will be used as a tool for learning and teaching Arabic as a foreign language.*

KEYWORDS: Handwriting, Recognition, Neural, Language, Learning, Activities

INTRODUCTION

Currently a computer assisted language learning (CAL L) system must contribute to the revolution and innovation of teaching methods. On the other hand, a CALL system should be intelligent to explain the different mistakes and to give the proper feedback to the learner. These increases the possibilities of using these systems in various learning situations and make them more effective and more autonomous. However, current CALL system is still far from the expectations of teachers and learners and still does not respond to their needs. Current CALL systems are rather test environments of learner knowledge and looks to a learning support. In addition, the feedback offered by these systems remains basic and may not be suitable for independent learning because it should be able to diagnose the problems of a student (in spelling, grammar, etc.). It should intelligently generate suitable feedback, according to the situation of learning. For this reason, we suggest as a solution using NLP tools capabilities to provide solutions to the limitations of CALL systems in the goal of developing a complete and autonomous CAL L system.

In this context, our work focuses on the development of natural language processing (NL P) tools for use in an environment to learn the Arabic writing. For this, we would initially develop an online handwriting recognition system for Arabic, then in a second time we would integrate it into a CALL system.

The recognition of handwriting is an innovative natural language processing technology; it is capable of processing digital ink to convert it into digital text. This digital ink is obtained when you write with a pointing device (mouse, stylus, tablet, etc.) or directly with your index finger on a touch surface.

The handwriting recognition can be classified into two types: on-line and off-line recognition. Online recognition is used with the digital pen or touch interface. However, offline recognition is used in the scanned writing, verifying signatures on bank checks, writing on pictures, etc.

ISSN 2054-6297(Print), ISSN 2054-6300(Online)

LITERATURE

The CALL was emerged with the appearance of computers. Since the beginning of the years 1960s, a few CALL systems have been designed and implemented. In addition, CALL involves several branches L evy (1997) in the areas of cognitive sciences such as linguistics, psycholinguistics, computational linguistics and computer science. On the other hand, the term is widely used to describe the field of technology and teaching of second languages Chapelle (2001). In addition, according to Beatty, the CALL is a process in which a learner uses a computer and, therefore, improves his learning level in a given language Beatty (2003).

The growing importance of CALL caused the development of automated tools for most of the European languages (English, French, German, Italian, etc.). This need is not marginal for Arabic that knows a considerable delay in its automatic processing.

The Arab learning systems are rare. While there are some achievements in this direction such as the classical tools such as Hotpotatos, Netquiz, and other tools using natural language processing (NL P) such as ArabVISL, "Arabic CALL", Nielsen (2003) and Shaalan (2006). However, they both have common limits; variety of activities, interaction, and quality of the feedback. Another major drawback of these systems is the rigidity since the data used are often static and defined in advance. In addition, current systems are not able to manage these essential points during a learning situation. On the other hand, the learners commit several types of errors that it is sometimes difficult to anticipate.

According to Mars (2014), the integration of NLP tools in the field of CALL can solve the CALL system problems and limitations. First, the use of NLP allows learners to work independently on the system because the generating of activities becomes autonomous and does not require the presence of the language teacher. Then, NL P tools and approaches can enrich CALL systems by a wide variety of activities, such as writing exercises, oral expression exercises, dictation, pronunciation exercises, etc. In addition, CALL system becomes able to detect, explain and correct errors automatically. Finally, NLP allows true customization and adaptation of activities to the learner according to the level and the progress of the learner. Although there are several works and resources available for the Latin language, but, there are only a few works and resources available in Arabic Mars (2014). Furthermore, given the complexity of Arabic scripts, Arabic NLP tools are rare and their quality is bad. Therefore, a part of our work will be devoted to the construction of an online handwriting recognition (HWR) system.

The handwriting recognition is the transcription of handwritten text into digital data for use by the computer. This area has been addressed since the 1950s. There are many types of handwriting, such as printed characters or cursive handwriting. These varieties of writing affect the recognition quality of the system and they cause many problems during the phase of segmentation of the word in characters.

Currently, there are basic applications for writing learning using Multimedia technologies (video, flash, etc.). Therefore, these applications are very limited and static. In addition, they do not have an advanced feedback that adapts to various errors and different learning situations. However, we do not find a complete application for writing learning based on NLP tools. Therefore, the integration of a handwriting recognition system in a CALL system can be considered as an innovation in the field of language learning. In addition, it can be considered as an added value to our research.

ISSN 2054-6297(Print), ISSN 2054-6300(Online)

METHODOLOGY

In order to solve some limitations of CALL systems, we are developing a computerassisted language learning system, designed for learners of the Arabic language as a second or foreign language, using a handwriting recognition system.

Before starting the description of the realization step of the HWR system, we will start with the study of the Arabic language and its complexity.

Arabic script

Arabic language is the mother tongue of more than 300 million people Aljlayl (2002). The Arabic writing is consonantal and cursive; it is composed of 28 basic letters, 12 special extra letters (ﺀ, ﺃ, ﺇ, ﺆ) and eight diacritics Bousalma (1998). Arabic is written from right to left. Automatic processing of the Arabic language is a very difficult task, especially in the field of handwriting recognition for several reasons:

- The majority of the letters change their shapes depending on their position in the word; initial, middle, end, or just isolated.
- Many Arabic letters contain dots and dashes glued to the letter such as the letter Alif Mad "ﺁ". They are usually added by the writer at the end of handwritten words.
- In Arabic handwriting, the style of writing varies from one writer to another.
- An Arabic word must be written recursively and characters must be connected to another one existing character in the middle of the word.

Moreover, Arab writers have the habit of neglecting the diacritics when writing Amin (1997), Märgner (2008), because the word can be deducted from the context.

Neural network approach

There are many ways to train a handwriting recognition system. Among the existing methods, we mention the hidden Markov Model (HMM), neural networks, expert system, the k-nearest neighbor and other techniques or a combination of these techniques. Some researchers divide these methods into two main classes:

- Syntactic: the class that involves the description of the character shapes in an abstract way.
- Statistics: where the system learned from the data directly without specified the structure of the knowledge system.

For our system, we have adopted the approach of the neural network.

Ideally, a neural network is a combination of elementary functions called neurons (or neural). Each artificial neuron realizes, at each time t, a nonlinear function, which, in the network, represents the outputs of neurons. The figure below shows a graphical representation of a neural network.

ISSN 2054-6297(Print), ISSN 2054-6300(Online)

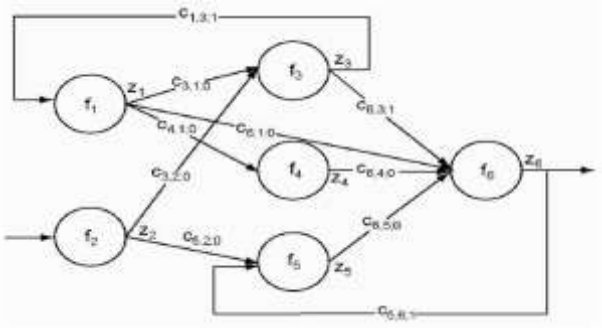

Figure 1. Neural network represented by a directed graph.

A neural network is composed of two parts, one is called multi-layer perceptron (MLP) and the second one is called Time Delay Neural Networks (TDNN). A PMC contains three layers: input layer, one or more hidden layers and an output layer. The neuron in each layer is connected to all the other neurons of the lower layer, in this case, the network is called a fully connected network Bridle (1990).

In the field of handwriting recognition, we use TDNN (Time Delay Neural Networks) to spatial move. The TDNN is a network often used in time for sequential data, so it is well suited for the recognition of on-line handwriting. The aims of using a TDNN are to recognize all extracted features independent of time and usually form a larger training database of a pattern or ink.

Data acquisition

A handwriting recognition engine requires a database of ink containing reference information regarding the shape of each character or each word. This reference information helps the recognition system to interpret the ink. In addition, these bases used to allow us to validate our approach based on neural network solution developed and to test the performance of our engine.

To be generic and be able to interpret as many variations as possible, the writers should be diverse as possible. For the Arabic script, many databases have been produced, but less accessible to the public. In fact, researchers have developed their private databases generally representative of small dictionaries with limited vocabulary or just isolated forms of letters and / or numbers. The number of writers is also limited. Actually, there is no basic full and robust data ink developed for an Arabic handwriting recognition system.

After studying existing ink databases and view the incompatibility (in size, diversity of writers, content) of these bases to our needs, we decided to build our own bases.

Recognition process

A handwriting recognition system takes as input an ink from an online equipment (Tablet, digital pens). To convert this ink on digital text, several treatments are performed by the system. Among these components, the system should have a module for pre-processing and normalization. Once the ink is pre-treated, another module extracts the necessary characteristics, which are used after extraction by the classification module based on the neural network to give a hypothesis of the word.

ISSN 2054-6297(Print), ISSN 2054-6300(Online)

Therefore, the system receives as input a file with ink that is in the form of point coordinates [x, y] in UNIPEN format. Then, the ink should be re-sampled in fixed points before the normalization. Then, from the normalized ink, the system should extract a set of vectors (we should extract seven characteristics for all points). Finally, the system will present these points to a neural network that will provide in the end a probability vector associated with each class.

Our handwriting recognition algorithm consists of four steps:

- Preprocessing and normalization: Pretreatment of the ink is a necessary step to improve its quality, which can lead to improving the function of representation and recognition. Some problems, such as repetitive points, imperfections in the scanning process must be eliminated. The re-sampling procedure is very important for improving the quality of the ink K avallieratou (2002). The normalization of the ink is a standard procedure in most recognition systems. In our case, the normalization done on character ink and word ink is primarily intended to standardize the initial ink so as to make it invariant to the translation, the size of the character and style of writing.

- Segmentation: To generate candidates for the characters, the input word is segmented into graphemes. The figure below shows the maximum

  segmentation of the Arabic word “Alhob”.

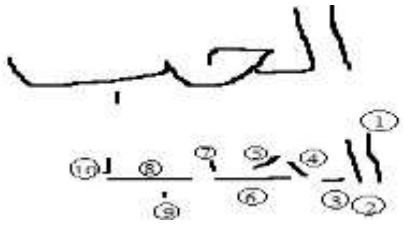

Figure 2. Example Maximum segmentation of the word “Alhob”.

The highest segmentations are performed based on the minimum and maximum coordinates along the axis (y) of the word. This is considered a basic method of word segmentation. In the previous figure, the word "Alhob" is divided into ten slots as shown above. For a longer word, there will be more slices. The slices are combined to form a character hypothesis. The total number of hypotheses affects the complexity of our training process and the accuracy of the recognition. This number of hypotheses depend on the total number of units that should be included in the case of the character.

- Feature extraction: For training a recognition system, the resultant ink after the pretreatment phase and the normalization is used to extract characteristics to use it in the training of models or for recognition. In our case, seven characteristics values were extracted for each item resulting. The characteristic values for each point x (n), y (n) are:

  i) Normalized x (n) between -1 and 1. ii) Normalized y (n) between -1 and 1.

  iii) The cosine of the angle made by the line between the point x (n + 1), y (n + 1) and the point x (n-1), y (n-1) and the x-axis.

ISSN 2054-6297(Print), ISSN 2054-6300(Online)

iv) The sinus of the angle made by the line between the point x (n + 1), y (n + 1) and the point x (n-1), y (n-1) and the x-axis.

v) The cosine of the angle of curvature between the points x (n + 2), y (n + 2) and the point x (n-2), y (n-2) to x (n), y (n).

vi) The sinus of the angle of curvature between the points x (n + 2), y (n + 2) and the point x (n-2), y (n-2) to x (n), y (n). vii) Fixed the binary value 1 when the pen is down, or -1 when the pen is up. The characteristics (iii) and (iv) provide direction information and characteristics (v) and (vi) provide curvature information. The Figure shows in more detail, the four related characteristics curves in mode (iii), (iv), (v) and (vi). For recognition, the characteristic values for a single character are used.

- Training and recognition: The training of the neural networks was made using all of the characters recovered from 170 forms. In this section, we present the implementation and the different algorithms associated with training step and recognition step. In the implementation phase, we selected two main parts: extraction and classification. The extraction process is done at lower layers, and the realization of the second process is done by a conventional multilayer perceptron (MLP). This perceptron receives as input all outputs that come from the extraction of the TDNN. In addition, it is possible to change these two processes. In the training phase, the choice of the learning method presents the important points in the implementation of a solution based on neural network approach according to Hérault (1994). Ideally, the method should allow the network to converge quickly to the global minimum calculated by the cost function. Depending on the complexity of optimization problems, several methods can be selected. In our system, we used a method based on the algorithm of the back propagation of gradient, which is able to converge faster than the overall gradient method Sarle (1997). Finally, the last phase is the recognition. In this step, we mainly use the TDNN approach. After the stage of the segmentation and feature extraction, we get in the output of the TDNN the number of classes by letter and the number of cases of the ink. Then, from this output, we calculate the probability of each word in the dictionary. In a TDNN, words are represented as a sequence of characters where each character is modeled by one or more states. The Therefore, the TDNN can be regarded a hybrid device recognition that combined the features of neural networks. Each neuron in the input layer represents a state of a character (in The Beginning, in the middle, at the end or isolated). Generally, the score of a character is calculated by finding the optimal alignment path through each state and summing the activated states in this direction. Similarly, the word score is calculated by finding the optimal alignment path through the states of characters in the word. The Therefore, the final score is calculated by Adding activated all states in this direction.

E valuation of the recognition process

In this section, we present results that correspond to the evaluation of our system of Arabic handwriting recognition. This evaluation is made on the database of test resulting from the collection. We used the correct recognition rate to evaluate the performance of our system.

ISSN 2054-6297(Print), ISSN 2054-6300(Online)

Data	Size of database test	Accuracy
Characters	6090	98,5 %
Words	1080	96,9 %
Sentences	300	91,7 %

Table 1. Results table.

This table shows the results of the evaluation of our system using testing databases. We can see that our results are effective because we can know almost characters.

During the recognition phase, our system is almost in real-time.

Integration of the HWR system

We will now integrate our handwriting recognition system for the Arabic language. For this, we decided to adapt the output of our recognition system needs linguists and didacticians.

Generally, in a writing learning process, the learner must start by writing characters, then writing words and finally writing sentences. Therefore, we will integrate in our system the three recognition models; characters model, words model and sentences model. The task of the generation of an activity is composed of three steps. First, the teacher must set an educational context for script application. Then, they must choose the recognition model to use (characters, words or sentences). Finally, the teacher should set the instructions and validate the creation of the activity.

RESULTS AND DISCUSSION

At the end of the implementation phase of the complete architecture of the system, it missing only the test and evaluation step of the system. However, evaluating the architecture, the contribution of TAL, and the contribution of CALL on the quality of language learning are very problematic.

The development and the integration of an NLP tool in a language learning system are very expensive in terms of resources, development, implementation, time and effort. Furthermore, these tools have a major problem related to their performance. Although most researchers and educationalists argue that, this method of learning increases learners' motivation and present a very effective tool to help novice learners. In addition, recent research has shown that human language is more complex than previously believed. Therefore, the development and the use of computer-assisted language learning often require the participation of scientists, computer engineers, and linguists, experts in artificial intelligence, cognitive psychologists, mathematicians, and developers, among others Ellis (2004).

In order to test the different components of the platform before putting it in a real learning environment, we carried out some experimental tests by teachers and learners. These tests allowed us to view the benefits of our CAL L system to teachers and learners, such as feedback, indexing texts, automatic generation of activities, etc. Moreover, these tests allowed us to detect some errors in our implementation. Once the tests and experiments are completed, we decided to put the platform in a real environment. For this, we participated with our system to a large international project led by the AUF (Agence Universitaire de la Francophonie). This allows us to install our CAL L system locally in two European universities, the University of Tirana

ISSN 2054-6297(Print), ISSN 2054-6300(Online)

and the University of Moldova. Therefore, these experiments should allow a real and comprehensive evaluation of our platform.

Once the current experiments and tests of our system in the Albanian University and Moldavian University are completed, we plan to implement the environment on the Web and make it available to the NL P community and the CALL community. This allows us to conduct collaborative work between teachers and learners. In addition, the launch of the platform enriches the resources and evolve the system.

CONCLUSION

Currently, the market is promising for learners who are looking for a smart way to learn Arabic because the domain of Arabic CALL is still restricted to a few prototypes.

We consider that the open aspect of our system is a great advantage for the future to evolve our platform. In this perspective, we can always add to the system an infinite number of activities, integrating multiple NL P tools and resources; adapt the platform to the needs of teachers and educationalists.

REFERENCES

Aljilayl, M. and Frieder, O. (2002) *On Arabic search: improving the retrieval effectiveness via a light stemming approach*. In ACM CIK M 2002 International Conference on Information and K nowledge Management, McL ean, V A, USA, pp. 340–347.

Amin, A. (1997) *Off line Arabic character recognition*. Fourth International Conference Document Analysis and recognition (ICDAR'97), p. 596.

Bouslama, F. and Amin, A. (1998) *Pen-based recognition system of Arabic character utilizing structural and fuzzy techniques*. In Proceedings of Second International Conference on K nowledge-Based Intelligent Electronic Systems, pp. 76–85.

Beatty, K . (2003) *Teaching and researching computer-assisted language learning*. New Y ork : L ong-man, p 7 et p 8.

Bridle, J. S. (1990) *Training Stochastic Model Recognition Algorithms as Networks Can Lead to Maximum Mutual Information Estimation of Parameters*. In Advances in Neural Information Processing Systems 2, pages 211-21.

Chapelle, C. A. (2001) *Computer applications in second language acquisition*. New Y ork : Cambridge p 3.

Ellis, R. (2004) *The study of second language acquisition*. Oxford applied linguistics. OUP.

Hérault, J. and Jutten, C. (1994) *Réseaux neuronaux et traitement du signal*. Hermes.

K avallieratou, E., Fakotakis, N. and K okkinakis, G. (2002) *An unconstrained handwriting recognition system*. International Journal on Document Analysis and Recognition 4, 226–242.

L evy, M. (1997) *CALL: Context and conceptualization*. Oxford: Oxford University Press, p 1.

Märgner, V. and El-Abed, H. (2008) *Arabic and Chinese handwriting recognition.* Databases and Competitions: Strategies to Improve Arabic Recognition

ISSN 2054-6297(Print), ISSN 2054-6300(Online)

Systems", vol. 4768. Springer, LNCS.

Mars, A. and Antoniadis, G. (2014) *Arabic language learning system using NLP tools and pedagogically indexed texts*. International Conference on Computer Science and Engineering 2014 (ICCSE'2014), 13-15 June, Hammamet, Tunisia.

Nielsen, H. and Carlsen, M. (2003) *Interactive Arabic grammar on the Internet: Problems and solutions*. Computer Assisted L anguage L earning (CAL L): An International Journal, 95 –112.

Sarle, W. S. (1997) *Neural Network FAQ, part 2: Learning, What are batch, incremental, on-line, off-line, deterministic, stochastic, adaptive, instantaneous, pattern, constructive, and sequential learning?* Periodic posting to the Usenet newsgroup comp.ai.neural-nets.

Shaalan, K . and Talhami, E.H. (2006) *Arabic Error Feedback in an Online Arabic Learning System*. Advances in Natural L anguage Proccesing, Research in Computing Science 18, pp. 203-212.

ISSN 2054-6297(Print), ISSN 2054-6300(Online)