Hussein SOORI, Jan PLATOS, Vaclav SNASEL

VSB-Technical University of Ostrava

Analysis of the Arabic using neural networks: an overview

Abstract. *This paper is a quick review of some of the scholarly work aiming at solving various problems of the Arabic language using neural networks. It includes some research work concerning online recognition of handwritten Arabic characters, speech recognition, offline character text recognition, text categorization and recognition of printed text. This paper concludes that more research should be conducted in this area considering the importance of the Arabic language, the rapid growth of internet users in the Arab world, and the widespread usage of Arabic characters by many languages other than Arabic.*

Streszczenie. *W artykule przedstawiono metody analizy języka arabskiego z wykorzystaniem sieci neuronowych. Analizowano możliwości rozpoznawania pisma odręcznego, drukowanego jak i mowy. (**Analiza języka arabskiego z wykorzystaniem sieci neuronowych**)*

Keywords: neural networks, Arabic language, natural language processing, speech recognition, character recognition, text recognition, text Categorization, handwritten city-names recognition**.**
Słowa kluczowe: język arabski, rozp[oznawanie znaków, sieci neuronowe.

Introduction

With the growing number of people using Arabic language around the world, there is a growing demand for more research in the area of computer science and information technology for the Arabic language. This importance can be listed here under some reasons: the rapidly growing number of computer and internet users in the Arab world and the fact that Arabic language is the sixth most used language in the world today. Another important factor is, after the Latin alphabet, Arabic alphabet is the second-most widely used alphabet around the world. Arabic script has been used and adopted to such diverse languages as the Slavic tongues (also known as Slavic languages), Spanish, Persian, Urdu, Turkish, Hebrew, Amazigh (Berber), Swahili, Malay (Jawi in Malaysia and Indonesia), Hausa, Mandinka (in West Africa), Swahili (in East Africa), Sudanese, and some other languages [13].

Neural Network

The term neural network had been used for some time to refer to a network or circuit of biological neurons. However, currently this term is used in neurocomputing to refer to artificial neural networks, which are composed of artificial neurons or nodes. According to Hecht-Nielsen, a neural network is a parallel, distributed information processing structure consisting of processing elements interconnected via unidirectional signal channels called, connections [17]. These elements can possess a local memory and carry out localized information processing operations. Moreover, each of these processing elements has a signal output connection that fans out into as many collateral connections as required. The idea came up as opposed to the idea of making computer software programs that are made of logic and sets of rules. In brief, the idea has to do mainly with the computational simulation of the human brain in terms of processing to solve problems, learning from past experiences, and adapting to different situations as per the different problems and the new challenges to come.

Neural Networks and Arabic Language Applications

In the last few decades, there has been some research work in the field of natural language processing, data mining, text categorization, and speech synthesis regarding Arabic language by using neural networks. This paper reviews some of the most important works that has to do with handwritten Arabic characters, speech recognition, offline character text recognition, text categorization, and recognition of printed text.

On-line Recognition of Handwritten Arabic Characters Using Kohonen Neural Network

This study was conducted by Mezghani *et al.* [12]. According to them, their work motivation is twofold: the features of representation are Fourier coefficients for the x(t) and y(t) components of pen positions, and their recognition was carried out by a Kohonen memory developed using a database of 7400 samples. They use Fourier descriptors as computed in Kuhl and Gardina's (1982). As far as data is concerned, they ignore all diactricals marks that do not carry important information about shape, and they work in 18 shapes out of the 29 isolated characters in Arabic. Their experiment results show that they run three experiments varying one parameter at a time: the dimension of feature vector, the number of iterations of the training algorithm and the number of nodes in a Kohonen memory. According to them, their results led to a memory that gives about 88.38% recognition rate on characters written without constraints on the writer. Their final experimental results show that the network successfully recognizes both clearly and roughly written characters with good performance.

Back Propagation Neutral Network of Arabic Characters Classification Module Utilizing Microsoft Word

This study sets two objects: summarizing the main characteristics of Arabic language writing style and suggesting a neural network recognition circuit. It uses neural network with back propagation training mechanism for classification. This was designed and trained by this research work to recognize any set of character combinations, sizes or fonts used in Microsoft word. One of the results that this study came out with is that the proposed network recognition behaviours were compared to a perceptron-like net that combines perceptron with ADALINE features. Those circuits were tested for three character set combinations; 28 basic Arabic characters plus 10 numerals set, 52 Latin characters and 10 numerals only. Hamza [15] concludes that the method was robust and flexible and can be easily extended to any character set. Moreover, the network exhibited recognition rates approaching 100% with reasonable noise tolerance.

Arabic Speech Recognition Using Recurrent Neural Networks

El Choubassi *et al*. [10] developed a system which is implemented by modular recurrent Elman neural networks (MRENN). They reviewed different approaches that were adopted in speech recognition one of which is hidden

Markov model (HMM) and neural networks (NN). According to them, HMMs have been the most popular and most commonly used approaches while NN have not been used for speech recognition until recently. However, according to them, the novelty in their approach is the use of a small RNN for each word in the vocabulary set instead of a unique large RNN for the entire set.

There are many distinctive features in their speech recognition system: it is implemented using neural networks and designed for Arabic language recognition, it recognizes a limited set of isolated words, it is female speaker-independent and performs favourably for male speakers and it is tolerant to moderate noise.

They implement their system in two stages. In the first stage of the design, speech is appropriately processed to the input to the neural networks. By this, they imply feature extraction achieved through modelling the human vocal tract using linear predictive coding which is then converted to the more robust cepstral coefficients. To compress those features, vector quantization is used, and a codebook is created using the K-means algorithm. In the second stage, they train the system for different utterances of the words in the vocabulary set. These utterances should constitute a good sample set of the various conditions and situations in which the word may he pronounced. According to them, this training was implemented on Elman neural networks using the back propagation algorithm with momentum and variable learning rate. In the third and last stage, the system is tested under different conditions: noisy and clean environments, speakers who trained the system and new speakers.

They come up with the result that the approach they adopted gave a promising recognition rate that can match, if not compete, with the ones usually obtained by HMM based approach.

An Approach to Offline Arabic Character Recognition Using Neural Networks

This work introduces a technique for the automatic recognition of Arabic Characters (Nawaz *et al.*) [22]. The technique in this work is based on Neural Pattern Recognition Approach. The main features of the system are preprocessing of the text, segmentation of the text to individual characters, feature extraction using centralized moments technique and recognition using RBF Network. The system used is implemented in Java under Windows Environment. It is designed for a single font multi size character set.

According to Nawaz *et al.* [22], only few papers have addressed the problem of Arabic character recognition [2]. One of the main reasons for this is that the characteristics of the Arabic language do not allow direct implementation of many algorithms used for other languages having English or Chinese-like characters [6, 7].

It is worth mentioning here that the connectivity and variant shape of characters in different word positions creates problems in recognition [14]. One of the problems with Arabic characters has to do with the different shapes of the same character. This depends on whether the character comes at the beginning, middle, or end of the word because most Arabic characters take different shapes accordingly. An example of the letter (ع) -pronounced as Ayin- is given in Table 1 below.

The processing stage of the proposed OCR system involves tacking two problems of noise: removal of isolated pixels and skew detection and correction. The paper also discusses how the segmentation of text to individual characters is performed. This includes three steps: segmentation of text to lines, segmentation of lines to words and segmentation of words to individual characters. In addition to that, this paper explains the feature extraction technique employed. Feature extraction represents the character image by a set of numerical features. These features are used by the classifier to classify the data. In their work, they utilize moments and other shape descriptors by Hu [18] to build the feature space. As far as recognition of character is concerned and in implementing the RBF Network Architecture, they used the Brain Construction Kit tool (BCK). BCK is a Java package developed in the public domain that enables users to create, train, modify and use artificial neural networks (ANNs). It consists of three layers, an input layer of dummy neurons, a hidden layer of radial neurons and an output layer of linear neurons.

This paper came out with the result that all modules showed good performance when tested separately. The system worked fine and showed a recognition rate of about 76%. It has been noticed also that the extracted features of the images produced from segmentation module deviate a lot from the respective results in the training set.

Table 1: Different shapes of the letter (ع)

Initial	Medial	Final	Separate
عـ	ـعـ	ـع	ع

A Comparative Study of Neural Networks Architectures on Arabic Text Categorization

Using Feature Extraction, Harrag *et al.* [16] present a model based on the Neural Network (NN) for classifying Arabic texts. According to them, many researchers have been working on text categorization in English and other European languages. However, few researchers have worked on text categorization for the Arabic language.

Harrag *et al.* [16] review some works related to Arabic text categorization. For example, El-Kourdi *et al.* [12] used Naive Bayes classifier for automatic Arabic document classification where the average accuracy reported was about 68.78%. Sawaf *et al.* [25] used statistical classification methods such as maximum entropy to classify and cluster News articles. However, the best classification accuracy they reported was 62.7%. El-Halees [11] described a method based on association rules to classify Arabic documents. The classification accuracy reported was 74.41%. Duwairi [9] proposed a distance-based classifier for categorizing Arabic text. The average accuracy reported was 0.62 for the recall and 0.74 for the precision. Syiam *et al.* [30] experimental results show that the suggested hybrid method of statistical and light stemmers is the most suitable stemming algorithm for Arabic language and gives generalization accuracy of about 98%. Mesleh [20] described a Support Vector Machines (SVMs) based text classification system for Arabic language articles. The system effectiveness for Arabic data set in term of F-measure is 88.11. Al-Harbi *et al.* [1] evaluated the performance of two popular classification algorithms (SVM and C5.0/C4.5) on classifying Arabic text using seven Arabic corpora. The SVM average accuracy was 68.65%, while the average accuracy for the C5.0 was 78.42%.

Harrag et al. [16], propose the use of Singular Value Decomposition (SVD) as a preprocessor of NN with the aim of further reducing data in terms of both size and dimensionality. According to them, the use of SVD makes data more amenable to classification and the convergence training process faster. The effectiveness of the Multilayer Perceptron (MLP) and the Radial Basis Function (RBF) classifiers are implemented. They conducted experiments using an in-house corpus of Arabic texts. They used precision, recall and F-measure to quantify categorization effectiveness. Their results show that the proposed SVD-

Supported MLP/RBF ANN classifier is able to achieve high effectiveness. Moreover, experimental results show that the MLP classifier outperforms the RBF classifier and that the SVD-supported NN classifier is better than the basic NN, as far as Arabic text categorization is concerned.

Recognition of Printed Arabic Text Using Neural Networks

In this work, Amin and Mansoor [3] focus on the automatic recognition of Arabic printed text using artificial neural networks in addition to conventional techniques. According to them, this approach has a number of advantages: it combines rule-based (structural) and classification tests; feature extraction is inexpensive, and execution time is independent of character font and size. They divide their technique into three major steps: The first step is preprocessing in which the original image is transformed into a binary image utilizing a 300 dpi scanner and then forming the connected component. In the second step, global features of the input Arabic word are extracted such as number of sub-words, number of peaks within the sub-word, number and position of the complementary character, etc. The third step involves the use of an artificial neural network for character classification.

They concluded that although almost a third of a billion people worldwide, in several different languages, use Arabic characters for writing, little research progress (both online and off-line) have been achieved towards the automatic recognition of Arabic characters. According to them, this is a result of the lack of adequate support in terms of funding, other utilities such as Arabic text database, dictionaries, and to the cursive nature of its writing rules.

Rule Based Neural Networks Construction for Handwritten Arabic City-Names Recognition

In this research work, Souici *et al.* [27] deal with a knowledge based artificial neural network for handwritten Arabic city-names recognition.

They mention some of their previous research where they used neural nets with distributed representation for character recognition [24] and local representation for word recognition [28]. However, in this work [27], they deal with a combination of both approaches (symbolic and connectionist) in building a knowledge based artificial neural network (KBANN) system for Arabic handwriting city-names recognition.

Souici *et al.* [27] preview other works dealing with mail sorting. According to them, postal address reading is considered the most successful applications of pattern recognition. They are also writer-independent and off-line systems processing large vocabularies (several thousands of words). Nevertheless, the redundancy existing between the city name and the zip code can be exploited to improve the performance. An enormous amount of mail is processed every day by the postal services in the world. As an example, in 1997 the US Postal Service processed about 630 millions of letters per day [29]. With this huge amount of mail, the use of automatic systems for postal addresses sorting is therefore primordial. The text to extract can involve: zip code, city name, street number and name, postal box, etc. These data can be in printed, isolated handwritten or cursive forms. Components of mail sorting system are: image acquisition, address field localization, text reading and finally the determination of the mail destination. The localization of the address field includes a preprocessing step: thresholding and binarization, segmentation in lines [4.5,8,29], in bounding boxes [19] and a segmentation-detection step of the way numbers, names, apartment number, postal box, city and country names, etc. are written. The address determination, *i.e.,* final customer or local delivery point, includes recognizing localized units such as characters, words, punctuations, etc., and to interpret the recognized symbols according to the database of valid addresses. For character and word recognition, all known techniques of pattern recognition can be applied.

However, more investigations seems to have taken place for statistical methods such as K nearest neighbours (KNN), hidden Markov models (HMM), artificial neural nets (ANN), etc. They also mention that Srihari in [29], implemented several classifiers: a structural, a hierarchical, a neuronal and a template matching based recognizers individually and in combination for character recognition, then he developed an HMM module for word recognition. In [5], HMM letters are combined in HMM words for the recognition of way names in French postal addresses.

According to Souici *et al.* [27], all the above techniques and many others are operational in commercialized systems all around the world but they are generally dedicated to Roman or Asian script processing. However, no such system exists for Arabic script, in spite of the fact that more than 20 countries use it.

In their article [27], they propose a KBANN based system for handwritten city-names recognition in Algerian postal addresses. Addresses are acquired in grey levels at 200DPI (dots per inch) resolution. On each image, they apply a median filter with 3X3 window as structuring element in order to eliminate the background and reduce the noise. A mean local thresholding operation transforms the image to a binary matrix (1 for the text and 0 for the background). As for city name, localization and extraction, they extract the writing of the line which contains the city name and the zip code (it is generally the line before the last one). To localize the city name, they use a technique based on word gap statistics [8]. They also compute gaps between connected components extracted by vertical projections of the whole address. The gap mean value is used as a segmentation threshold. Gaps greater than the threshold represent the word-gaps.

Arabic text is written and read from right to left and a city name can be composed of several words. According to them, the more closely connected component to the left corresponds to the zip code, while the right ones correspond to the city name, which is the main concern of their work.

Arabic is an agglutinative language. Souici *et al.* [27], describe feature extraction as follows: Among the 28 basic Arabic letters, 6 are not connectable with the succeeding letter. Thus, an Arabic word may be decomposed into sub-words, high or low secondary components exist for 15 letters which have one to three dots. Some of these characters have the same primary component and differ only by the position and/or the number of dots. They also could find loops, ascenders and descenders in Arabic writing.

Their previous experiments with holistic Arabic literal amounts recognition [26, 28] prove the usefulness of perceptual features and diacritical dots. In this work, the same set of features is retained and extracted by contour tracing and Freeman chain coding [23]. These features are used to describe the 55 words of Algerian city names lexicon with their number of possible occurrences.

Souici et al. [27], designed their classifier inspired by the KBANN approach (Knowledge Based Artificial Neural Network) developed by Towell in 1991 which was tested in the fields of molecular biology, DNA sequence analysis and process control. They used the theoretical knowledge expressed by rules to determine the initial topology of the

neural network. Their network is then refined using standard neural learning algorithm and a set of training examples.

Conclusion

The above reviewed works are significant in terms of touching current issues in information technology by using neurocomputing, but it is obvious that the amount of research in this area does not do justice to the importance of the Arabic language and to the rapid growth in the number of people using Arabic language in the world today. With the Arabic characters being used by almost a third of a billion people worldwide, there is an eminent need for more research in this area.

This work was supported by the SGS in VSB – Technical University of Ostrava, Czech Republic, under the grant No. SP2013/70, and has been elaborated in the framework of the IT4Innovations Centre of Excellence project, reg. no. CZ.1.05/1.1.00/02.0070 supported by Operational Programme 'Research and Development for Innovations' funded by Structural Funds of the European Union and state budget of the Czech Republic and by the Bio-Inspired Methods: research, development and knowledge transfer project, reg. no. CZ.1.07/2.3.00/20.0073 funded by Operational Programme Education for Competitiveness, co-financed by ESF and state budget of the Czech Republic.

REFERENCES

[1] Al-Harbi, S., Almuhareb, A., Al-Thubaity, A., Khorsheed, M. S., Al-Rajeh, A.: Automatic Arabic Text Classification. In: Proceedings of 9es Journées internationales d'Analyse statistique des Données Textuelles, JADT'08, pp. 77–83. France (2008)

[2] Amin, A., Off-Line Arabic Character Recognition System: State of the Art, Portern Recognition, Vol. 31, No. 5, pp 517-530 (1998)

[3] Amin, A., Mansoor, W.: Recognition of Printed Arabic Text using Neural Networks. In: Proceedings of the 4th International Conference on Document Analysis and Recognition IEEE Computer Society Washington, DC, USA © (1997)

[4] Bennasri A., Zahour A., Taconet B.: Arabic script preprocessing and application to postal addresses. ACIDCA'2000, Vision & Pattern Recognition, Monastir, 74-79. Tunisia (2000)

[5] Bertille J.M., Gilloux M., El Yacoubi A. : Localisation et reconnaissance conjointes de noms de voies dans les lignes distribution des adresses postales. SRTP/RD/Traitement automatique ligne distribution, Transition n°7,16-25. (1994)

[6] Chinveerphan, S., Zidouri, A. B. C., Sato, M.: Modified MCR Expression of Binary Document Images. IEICE Trans. Inf: & Syst., Vol. E78 -D, No. 4, pp. 503-507, April (1995)

[7] Chinveerphan, S., Zidouri, A. B. C., Sato, M.: Recognition of Machine Primed Arabic Character and Numerals Based on MCR., IEICE Trim. Inf & Syst., Vol. E78 -D, No. 12, pp. 1649-1655, Dec. (1995)

[8] Dowton A.C., Leedham C.G.: Preprocessing and presorting of envelope images for automatic sorting using OCR. Pattern Recognition, Vol. 23, n°. 3/4, 347-362. (1990)

[9] Duwairi, R. M.: A Distance-based Classifier for Arabic Text Categorization. In: Proceedings of the International Conference on Data Mining, Las Vegas USA. (2005)

[10] El Choubassi, M. M., El Khoury, H.E., Alagha, C. E. J., Skaf, J. A., Al-Alaoui, M. A.: Arabic Speech Recognition Using Recurrent Neural Networks. In: Proceedings of the 3rd IEEE International Symposium on Signal Processing and Information Technology, 2003. ISSPIT (2003)

[11] El-Halees, A., Arabic Text Classification Using Maximum Entropy. The Islamic University Journal (Series of Natural Studies and Engineering),Vol. 15(1), pp. 157-167. (2007)

[12] El-Kourdi, M., Bensaid, A, Rachidi, T.: Automatic Arabic Document Categorization Based on the Naive Bayes Algorithm. In: Proceedings of the 20th International Conference on Computational Linguistics, Geneva, August (2004)

[13] Encyclopedia Britannica Online: Alphabet. Online (Feb 2011), http:// www.britannica.com/EBchecked/topic/17212/alphabet

[14] Fakir, M. and Hassani, M. M.: Automatic Arabic Characters recognition by moment invariants. Colloque international de telecommunications, Fes, Morocco, pp 100-103.(1997)

[15] Hamza, Ali A.: Back Propagation Neural Network Arabic Characters Classification Module Utilizing Microsoft Word. In: Journal of Computer Science 4 (9): 744-751, 2008 ISSN 15493636 Science Publications (2008)

[16] Harrag, F., Al-Salman, A. M. S., BenMohammed, M. A.: Comparative Study of Neural Networks Architectures on Arabic Text Categorization Using Feature Extraction. In: Int. Conf. on Machine and Web Intelligence (ICMWI), Algiers (2010)

[17] Hecht-Nielsen, Robert: Neurocomputing. Addison-Wesley Publishing Company, Inc (1990)

[18] Hu, M. K.: Visual Pattern Recognition by Moment Invariant. IRE Trans. on Information Theory, Vol. IT-8, 179-187. (1962)

[19] Mahadevan U., Srihari S.N.: Parsing and recognition of city, state and ZIP codes in handwritten addresses. ICDAR'99, 325-328. (1999)

[20] Mesleh, A. A.: Chi Square Feature Extraction Based SVMs Arabic Language Text Categorization System. Journal of Computer Science, Vol. 3(6), pp. 430-435. (2007)

[21] Mezghani, Neila, Mitiche, Amar, Cheriet, Mohamed: On-line recognition of handwritten Arabic characters using A Kohonen neural network. In: Proceedings of the Eighth International Workshop on Frontiers in Handwriting Recognition (IWFHR'02) 0-7695-16920, IEEE (2002)

[22] Nawaz, S. N., Sarfraz, M., Zidouri, A., Al-Khatib, W. G.: An Approach to Offline Arabic Character Recognition Using Neural Networks. In: Proceedings of the 2003 10th IEEE Int. Conf. on Electronics, Circuits and Systems, 2003. ICECS (2003)

[23] Pavlidis T.: Algorithms for Graphic and Image Processing", Rockville, MD: Computer science press (1982)

[24] Sari T., Souici L., Sellami M.: Off-line Handwritten Arabic Character Segmentation and Recognition System: ACSA-RECAM. IWFHR'2002, 8th International Workshop on Frontiers in Handwriting Recognition, Niagara-on-the-Lake, Ontario, Canada, August (2002)

[25] Sawaf, H,. Zaplo, J., Ney, H.: Statistical Classification Methods for Arabic News Articles. Workshop on Arabic Natural Language Processing, ACL'01, Toulouse, France, July (2001)

[26] Souici L., Aoun A., Sellami M.: Global recognition system for Arabic literal amounts. International Conference on Computer Technologies and Applications, ICCTA'99, Alexandria, Egypt, August (1999)

[27] Souici, L., Farah, N., Sari, T., Sellami, M.: Rule Based Neural Networks Construction for Handwritten Arabic City-Names Recognition. Artificial Intelligence: Methodology, Systems, and Applications. Lecture Notes in Computer Science, Volume 3192/2004, 331-340, DOI: 10.1007/978-3-540-30106-6_34 (2004)

[28] Souici-Meslati L., Rahim H., Zemehri M. C Sellami M. Systeme Connexionniste a Représentation Locale pour la Reconnaissance de Montants Littéraux Arabes. CIFED'2002, Conférence Internationale Francophone sur l'Ecrit et le Document, Hammamet, Tunisie, October (2002)

[29] Srihari S. N.: Recognition of handwritten and machine printed text for postal address interpretation. Pattern Recognition Letters: Special issue on postal processing and character recognition, Vol. 14, N°: 4, 291-302. April (1993)

[30] Syiam, M. M., Fayed, Z. T., Habib, M. B.: An Intelligent System for Arabic Text Categorization, IJICIS, 6(1), pp. 1-19. (2006)

***Authors**: Hussein Soori, VSB-Technical University of Ostrava, 17 listopadu 15,70833 Ostrava Poruba, E-mail:sen.soori@vsb.cz; Jan Platos, VSB-Technical University of Ostrava, 17. listopadu 15,70833 Ostrava Poruba, E-mail:jan.platos@vsb.cz; Vaclav Snasel, VSB-Technical University of Ostrava, 17. listopadu 15,70833 Ostrava Poruba, E-mail:vaclav.snasel@vsb.cz;*