Use of Neural Networks in Arabic Text Transliteration

Professor Mohammed Zeki Khedher
Jordan University
Email: khedher@fet.ju.edu.jo

1. Neural Networks for Syntax Analysis

Certainly one of the most important questions for the study of human language is: How do people unfailingly manage to acquire such a complex rule system? A system so complex that it has resisted the efforts of linguists to date to adequately describe in a formal system.

One way of trying to test that is to use various recurrent neural network architectures for classifying natural language sentences as grammatical or ungrammatical, thereby exhibiting the same kind of discriminatory power provided by the Principles and Parameters linguistic framework. The computational power of Elman networks has been shown to be of good results.It has been found that the main issue is a training issue and not a representational issue. Backpropagation-through-time is an iterative algorithm that is not guaranteed to find the global minima of the cost function error surface. An important question is :Are the networks learning the grammar? The hierarchy of architectures with increasing computational power (for a given number of hidden nodes) give an insight into whether the increased power is used to model the more complex structures found in the grammar. It has been shown that both Elman and W&Z recurrent neural networks are able to learn an appropriate grammar for discriminating between the sharply grammatical/ungrammatical pairs. But these results are only for a very limited amount of data and it is expected that increased difficulty will be encountered in training the models as more data is used. It is clear that there is considerable difficulty in scaling the models.(1)

Most of the research on neural architectures for syntax analysis has focused on the investigation of neural networks that are designed to learn to parse particular clases of syntactic structure (e.g. strings from deterministic context-free languages or natural language sentences constructed using limited vocabulary).(2)

PARSEC(2) is a modular neural parser consisting of six neural-network modules. It transforms a semantically rich and therefore fairly complex English sentence into three output representations produced by its perspective output modules. The three output modules are role labeler which associates case-role labels with each phrase block in each clause, inter-clause labeler which indicates subordinate and relative clause relationships and the mood labeler which indicates the overall sentence mood. Each of these modules is trained individually by variation of backpropagation algorithm. The input is a sequence of syntactically as well as semantically tagged words in the form of binary vectors and is sequentially presented to PARSEC, one word at a time. A test on 117 sentences was reported giving 78% correct labeling when trained with 240 sentences.

Another proposed (2) neural architecture for syntax analysis is obtained through systematic and provably correct composition of a suitable set of component symbolic functions which are altimately realized using neural associative processor modules. The

neural associative processor is essentially a 2-layer perceptron which can store and retrieve arbitrary binary pattern associations. This allowed massive parallelism to support applications that require syntax analysis to be performed in real time. The proposed neural network for syntax analysis is capable of handling sequentially presented character strings of variable length, and it is assembled from neural-network modules for lexical analysis, stack processing, parsing, and parse tree construction(2,3).

There has been substantial progress in modelling lexical representations using neural networks. However, the question of lexical-semantic representation has been mostly overlooked. In particular there are few computational models of where semantic representations come from, and how they can be acquired Each concept is assigned a node, and connection strengths reflect the amount of conceptual relevance each node has to its partner. The stronger, or shorter, connections represent a high degree of semantic relatedness. Weaker, or longer, connections hold between less related nodes.

Distributed Models: An alternative model of lexical semantic representation describes a word not as a single node, but as a pattern of activity across many nodes. Representing meaning as a set of values over a large number of dimensions is also appealing, since it allows a computational interpretation of feature-based semantic theories. In a feature-based theory, each concept is represented by a distribution of numerical values over a set of semantic features (usually perceptual properties possessed by instances of the concept). The values ascribed to each feature can be treated as a vector that determines the location of the concept in feature space. Distributed semantic representations can be used as target values in a multi-layer perceptron trained with a gradient-descent algorithm.

The Self-organizing Map (SOM) is a neural network intended to explain how the cerebral cortex can become organized topographically. Topographic organization is present when the implicit spatial structure of an input signal is retained in a low-dimensional neural representation, such that nearby points in the neural representation refer to nearby points in the input space (4,5,6,7,8).

2. Arabic Text Syntax Analysis:

Any arabic morphological system must be capable of dealing with classical arabic language. There must be a clear boarder between the morphological information and the algorithms to be used for manipulating these data.(9)

In dealing with arbic language processing a number of points have to be kept in mind:
(I) Arabic words may contain one or more than one sub-words. In fact these subwords may be independant words which hold one of the grammatical important positions in the sentence. Hence the word (which is the series of characters between two blank characters) is either a simple word or a compound one. The latter may contain two or more sub-words. Sub-words may be a pronoun, a sign of the grammatical state (e.g. state of “Jarr”, “Nasib”, “Rafi” or “Sikoon”, etc.), or some other type.

(ii) Transliteration is an important asset in syntax analysis. There are seven diacritics, which are signs for transliteration. They are : “Fatha”, “Dhammah” , “Kasrah” , “Sikoon” , “Tanween Fatih” , “Tanween Kasir” , “Tanween Dhamm”.
There is also the “Shaddah” which stands for doubling the last character and can be combined with the first three signs above.
(iii) As arabic language is one of the most advanced living natural languages, it has some hidden parts which are to be understood by the experienced reader , or the person who hear it spoken. Assumption of such hidden parts are necessary.
(iv) Arabic words are of two types: Derived words and Underived words. The derived words belong to certain root (either composed of 3, 4 or 5 characters). Such roots are listed in fomous arabic references. The derivation takes place in accordance to defined rules called “Wazin”. Hence a derived word may be defined by its root and wazin together. Some additions may be found with such words, namely : prefix, suffix, and diacritics.
Underived words are the words like: “Harf”, Adverb or names of no meanings or which has no origin of derivation.
The subject of arabic derivations of words had recieved extensive studies(10). Derivation of a word from its root is called “Derivation” (Ishtiqaq). This needs a reference to a dictionary of roots categorized into various characteristics. These categories lead to the selection of the rule to be used in the derivation. Some of these are:
(I) List of verbs , types and their characteristics.
(ii) Rules of combination of verbs with pronouns.
(iii) Rules of singular and twin (Muthannah) and plural (with its 3 types).
(iv) Rules of combining characters (Idgham) , changing vowels (Ibdal).
(v) Rules related to the “Hamzah” character.
(vi) Table of sources (Masadir)
(vii) List of irregular words.
With all above information classified in proper manner, it would be possible to treat the arabic text “Sarf” in the computer systematically.
The production of arabic dictionaries on the other hand has been attempted (11) , in order to use them for basic linguistic nucleous as well as extentions. Extensive studies of the prefix an suffix uses in relation to the mid-part (the nucleous) have been done.
Take an example the word “سيمنحانكها “. This word consists of a first prefix “س “ to denote future tense, a second prefix “ي” , a nucleous “منح”, a first suffix “ا” to denote twin , a second suffix “ن “ , a third suffix “ك “ which is a pronoun and a fourth suffix “ ها“ a second pronoun. In other cases the diacritic may play a basic role in the word at the end of the nucleous or at the end of the whole word.
The segregation of components of sub-words may be performed by different means. The resulting structure of such components may be put in a form such as the one shown in Table(1)

Table I

Jarr	pre-1	mid-1	post-1	Majroor	pre-2	mid-2	N	sign	Num	pre-2
من		من		العالمين	ال	عالم	)	ين	7	
وفي	و	في		أموالهم		أموال	)		1	هم
وفيهم	و	في	هم							

The codes N and Num shall be referred to latter as in Tables II and III

3. Arabic Text Transliteration:

There has been some attempts to do automatic transliteration of arabic text by SAKHAR Company . An experiment was made in automatic translitration based on the background mentioned in Table I above (10,11). The experiment showed a moderately promising results as shown by the statistics presented(12).

3.1 Proposal for Use of Neural Networks in Arabic Language Text Processing:

The use of neural networks in syntax analysis need a lot of research and what we see from previous sections of this paper, we are at the begining of some promising attempts in this direction.

In arabic language, there are quite a large number of topics which may be suitable for the use of neural networks. However, quite a good amount of extra processing and preparations are necessary. Inclusion of some other artificial intelligence treatments (such as expert system, fuzzy logic and genatic algorithms) in such processing may also be necessary.

Here are few topics , where neural networks seem to be useful. They are just examples.:

(I) Parsing of compund arabic words into their components of subwords.

(ii) Investigations to find hidden words in arabic text.

(iii) Identifying the various cases : “Rafi’”, “Nasib” , “Jarr” , “Jazim”

(iv) Identifying the words which are of fixed shape “Mabni” and those which are liable for case changes according to position in the sentence “Mu’rab”

(v) Categorizing words into : Noun, Verb and Preposition (Harf)

The main reason why it seems that the arabic text processing seems to be suitable for neuaral network application is that people from their early age are trained to talk properly. Why neural networks can not be trained similarly? Of course, proper and enough data is necessary.

3.2 Neural Network for “Jarr” and “Majroor”

3.2.1 Statement of the Problem

As an example for arabic text transliteration, the following problem is presented here. Only one type of operation is selected for this purpose, that is the “Al-Jarr wal Majroor” "الجار والمجرور" for the five most common articles of “Jarr” which are “من في عن على إلى“ .

The sub-clauses containing examples for these articles were taken from the text of the “ Holly Quran”

3.2.2 The Data

It has been found that there are 6337 sub-clauses in the Quran, which contain one of these five articles.They may be divided into two categories:

(I) One word "Jarr wa Majroor"; in which the "Majroor" is a pronoun connected to the "Jarr" e.g. "عليهم منهم فيها عنّا إلينا" "There are 2048 cases of this kind in the Quran which contains the above five "Jarr" articles.
(ii) Separate words for the "Jarr" and "Majroor". As examples for this case "ومن " , " في آذانهم" , " وفي الأرض " , " لألى الله " , " إخوانهم" . There are 4289 cases of this kind in the Quran. It is presumed that these cases are of quite enough variety of cases for our purpose and may cover nearly all aspects related to this problem in arabic language.

3.2.3 Segmentation of arabic words:

(I) The "Jarr" words : There are two types of connecting subwords with the words of "Jarr": The prefix subwords and the suffix subwords.
Examples for the prefix: ومن ، ففي ، لإلى
Examples for the suffix: فينا ، منهم ، عليهما
The pronouns which are the suffixes in all the second examples are actually the "Majroor"
There are some words which contain both prefix and suffix , e.g. ففيها ، ومنهم
(iii)
a- "Majroor" word which may be "Ma'rifa" or "Nakira" and contains one of the following articles :
Prefix which may be "الـ ", hence it is "Ma'rifa".
Main body of the word, the nucleous (noun usually)
The sign of "Jarr" which may be "Tashkeel sign" (diacretics) or a number of characters
Suffix which may contain one or two parts as well as "Tashkeel"
The "Majroor" word may be in a position of "Mudhaff" followed by "Mudaff Ilaihi" Hence it is "Ma'rifa"
b- "Majroor" subclause : This will start usually by an article (not a noun) followed by a noun. The sub-clause all together is the Majroor without an appearant article for the transliteration.
The following arabic grammatic rules are of relevance to the above topic:
1- The "Jarr " على ، إلى end with the character "alif maqsura" ى . This character is converted into ي once a pronoun suffix subword is connected to the "Jarr" e.g. عليهم ، إليهم
2- When some pronuns are connected as prffix subword with the "Jarr" an interaction called "Idgham" may occur for similar letters e.g. " عنّا = نا + عن The use of the "ّ" "shaddah" is to indicate this when transliteration is indicated.
3- Words which are called "Mamno' min Al-Sarf" are to take "Fatha" in case of "Jarr" instead of normal "Kasra" for normal words e.g. "إبراهيم "
4- Nouns ending with "ى " cannot take a sign e.g. "موسى "
5- Words ending with "ى " when it is "Nakira" then the word will take "Tanween Fatih" instead of "Tanween Kasir", e.g. " فيه هدىً "
6- Plaural signs in case of "Jarr" are : " ين "for "جمع المذكر السالم", "Kasra" for "جمع المؤنث السالم " , "Fatha" for "جمع التكسير" in special cases and "يـْن " for twin plaural "المثنى "
7- The sign for "Jarr" hence are shown in Table (2)

Table II

sign	كسرة	تنوين كسر	فتحة	تنوين فتح	ي	يـْن	ين
code	1	2	3	4	5	6	7

8- The codes of the type of the "Majroor" word are given in Table III

Table III

Code	Type
0	الحالة العاما
1	الممنوع من الصرف
2	شبه جملة : إن
3	(إسم موصول أو إسم إشارة (مبني
4	شبه جملة : حيث
5	الأسماء الخمس
6	معتل الآخر
7	شاذ
8	جمع مذكر سالم
9	مثنى
11	جمع تكسير

3.2.4 Implementation:

The data used was prepared in a data base form using "Nafitha" arabization program and Foxpro Data base engine.

The data base contains the following items:

1. The "Jarr" word
2. The "Majroor" word
3. The prefix subword
4. The "Jarr" body
5. The suffix subword
6. The prefix subword of "Majroor"
7. The nucleous of the "Majroor"
8. The suffix subword of the "Majroor"
9. The sign of the "Jarr"
10. The coding of the type of the "Majroor" as in Table II, which is the required output of the neural network.
11 The codings of the input data contains the following

(I) Type of the "Majroor" word as given by Table III

(ii) A binary digit for "Jarr" prefix: 0 = without and 1= with.

(iii) A binary digit for "Jarr" suffix: 0 = without and 1= with

(iv) A binary digit for "Majroor" prefix: 0 = without and 1= with

(v) A binary digit for nucleous of the "Majroor" : 0 = normal and 1 for those ending with ى

(vi) A binary digit for "Majroor" suffix : 0 = without and 1= with

(vii) A binary digit for "Majroor" suffix 0= "Ma'rifa" and 1 = "Nakira"

12. The data are fed to neural network and training is performed.

13. It has been tested them with the same date and the reslults showed quite good performance.

14. Some cases related to the “Tashkeel” of some signs of the “Jarr” were ignored, e.g. the ending of ي in Table II covers the cases of no sign or “Fatha” or “Kasra”. Another case is for ين ending with “Fatha” or “Kasra”

3.2.5 Results:

Input data which were represented into binary numbers were converted into decimal when fed to the neural network and were input twice for further emphasis.

MATLAB Neural network Tool Box was used and the back propagation algorithm was selected. 24 hidden neurons in one hidden layer were used.

Training was performed in 3600 epoches. The sigmoid shape of transfer function Neuorn Model was used.

Prelimenary results showed 91% agreement in the first trial. it is anticipated that results agreement about 100% should be achievable once all relevant rules are included into the database

4. Discussions and Conclusions:

Natural languages processing is one of the most difficult topics in artificial intelligence. It is progressing slowly and steadily. The use of neural networks is just at its begining in this area. Prelimenary results of pioneer researchers in the field showed some promising results. It is only natural and reasonable to expect future progress in the field as human judgement is only an accumulation of experience. Rules are usually concluded after experimentation of data.

The arabic language as one of the most advanced living languages, can utilize neural networks techniques in many aspects.

However other field of artificial intelligence may be used with neural networks , namely expert systems , fuzzy logic and genatic algorithms in this respect too. With the aid of proper data bases created in careful manner and with the use of these techniques, lexical and contexual anlaysis of arabic language may be tackeled successfully.

It is strongly recommended to consider seriousely the use of neural networks in research of arabic language processing.

The example use is by no means the final shape of the rules used and quite a good chances for improvement is possible. It is just to show the possibility of this wide area of application of the neural networks.

5. References

1. Natural Language Grammatical Inference with Recurrent Neural Networks
Steve Lawrence{1}, C. Lee Giles{1,2}, Sandiway Fong{1}
{1}NEC Research Institute, 4 Independence Way, Princeton, NJ 08540, USA.
{2}UMIACS, U. of Maryland, College Park, MD 20742, USA.
{lawrence,giles,sandiway}@research.nj.nec.com}

2. Chun-Hsien and Vasant Hanovar, A Neural-Network Architecture for Syntax Analysis, IEEE Trans. on Neural Networks, Vol. 10, No.1, January 1999, pp 94-114

3. A. N. Jain, A. Waibel, and D.S. Touretzky, “PARSE: A structured connectionist parsing system for spoken language “, IEEE Proc. Int. Conf. Accoustic, Speech , Signal Processing, San Franciso, CA March 1992, pp 205-208

4 J. A. Bullinaria and C. C. Huckle.
Modelling lexical decision using corpus derived semantic representations in

a connectionist network. Proceedings of the Fourth Neural Computation and Psychology Workshop, 1997.

5 C. C. Huckle.
Unsupervised Categorization of Word Meanings Using Statistical and
Neural Network Methods. PhD thesis, Centre for Cognitive Science, Edinburgh University, 1995.

6 T. K. Landauer and S. T. Dumais.
A solution to Plato's problem: the latent semantic analysis theory of
induction and representation of knowledge. Psychological Review, 104:211-240, 1997.

7 Steve Lawrence, Sandiway Fong, and C. Lee Giles.
Natural language grammatical inference: A comparison of recurrent neural
networks and machine learning methods.
In Stefan Wermter, Ellen Riloff, and Gabriele Scheler, Symbolic,
Connectionist, and Statistical Approaches to Learning for Natural Language
Processing, Lecture notes in AI, pages 33-47. Springer Verlag, New York,
1996.

8 L.R. Leerink and M. Jabri.
Learning the past tense of English verbs using recurrent neural networks.
Peter Bartlett, Anthony Burkitt, and Robert Williamson, editors, Australian Conference on Neural Networks, pages 222-226. Australian
National University, 1996.

9. T.A. El-Sadany and M.A. Hashish, An Arabic Morphological System, IBM Systems Journal, Vol. 28, No. 4, 1989, pp 600-612

10 مروان البواب و محمد حسان الطيان
المعالجة الصرفية العربية بالحاسوب
مجلة المنظمة العربية للتربية والثقافة والعلوم - كانون الثاني 1999 صفحة 7-17

11. عبد الفتاح إبراهيم وسالم غزالي
قاعدة البيانات المعجمية العربية - حصيلة وآفاق
مجلة المنظمة العربية للتربية والثقافة والعلوم - كانون الثاني 1999 صفحة 15-23

12. مالك غنيمة
مجلة المنظمة العربية للتربية والثقافة والعلوم - كانون الثاني 1999 صفحة 33-36