ICTA'07, April 12-14, Hammamet, Tunisia

Learning applications based on Multi-layer perception system for automatic tagging

Mohsen Maraoui
LIDILEM Laboratory, Stendhal University, Grenoble (France)
mohsen.maraoui@u-grenoble3.fr
Georges Antoniadis
LIDILEM Laboratory, Stendhal University, Grenoble (France)
georges.antoniadis@u-grenoble3.fr
Mounir Zrigui
UTIC Laboratory (Monastir unit), Faculty of Science, Monastir (Tunisia)
mounir.zrigui@fsm.rnu.tn

Abstract

Within the framework of our research, which consists of the integration of the Arab language in platform MIRTO (Interactive Multi-trainings by Research on Texts and the Oral examination), we need a reliable analyzer and by consequence an electronic dictionary which is tagged and complete. For this purpose, we began the realization of a dictionary, generated and tagged automatically, having as starting point the Arab roots and by using the patterns and their meanings. The first task consists on tagging these roots, which pushed us to carry out a tag-maker for the Arab roots. We will present initially the architecture of our tool and afterwards its didactic utility.

1. Introduction

Working out CALL (Computer Assisted Language Learning) exercises by using NLP (Natural Language Processing) technologies is realizable, and it is already the case for platform MIRTO [1] for some languages like the French and the English for whom the results relating to grammatical tagging run up against the 95% bar of recognition rate [2], but it is not the case for the Arab which is considered as a language difficult to control automatically [3] [4].

The morphological analyzer represents the base of the NLP applications, especially in the field of the learning of the languages because in this kind of applications, the output cannot be wrong: the application can not give to its user an approximate result. For this reason we need analyzer with a very high level of reliability, i.e. which gives answers fast and correctly, things which do not exist at the moment for the Arab language. The inexistence of tagged complete dictionary is one of causes of this failure. In fact, the dictionary represents the knowledge base of the analyzer, so if it contains gaps, the analysis is defective. However, all the existing Arab electronic dictionaries until now are: not labelled, incomplete or very specific to an application or a particular linguistic model [5].

To be reliable, Arabic CALL applications need a robust morphological analyzer, i.e. able to give at least a tag for a word which belongs to the language, whose core is a complete dictionary (which contains all the words of the Arab language and their corresponding tags). A complete dictionary of Arabic contains thousands of the words, and for implementing it manually we would need a lot of time. This explains the automatic need for treatment, especially because most of the Arabic lexicon[1] can be easily derivable. For this purpose, we began the generation of a automatically tagged dictionary, starting from the Arab roots and by using their patterns and their meanings, while having as basis the conditions of morphemic structure of Arabic.

2. Arab electronic dictionary

The majority of the verbs and the nouns are the combinations of a root[2] (generally with three radical consonants [6]) and of a pattern[3].

A word family can thus be generated automatically of the same semantic concept starting from only one root using various patterns, moreover even the Arab roots can be generated automatically [7].

With electronic dictionary, we can easily determine the roots of the words, Thereafter these roots can be used to help the learners of Arabic to know the synonyms and the meanings of words. Because in order to find a definition or a synonym of an Arab word ,and in the absence of a

[1] The Arab lexicon contains three categories of units: verbs, nouns and particles.
[2] The whole of two, three or four consonants which represent a well defined concept. Example: KTB كتب ⇔ to write.
[3] A specific model that represents in a diagramed way a language structure or verbal behavior of the speakers.

human help, the learners of Arabic will use the paper dictionary of synonyms (ordered alphabetically by roots for the derivable words) to face the incomprehensible texts [8]. To identify a word in the dictionary we must known its root, which is not obvious for a person who learns Arabic.

The goal of our work is to have a complete dictionary composed of several modules:
-tagging of roots,
-derivation and tagging of the names,
-derivation and tagging of the verbs,
-conjugation and tagging of the conjugated verbs

In this article, we are focusing on the tagging of roots part which is very important, because this part will determine the nature of root through the tag, and by consequence the type of the anomaly which will lead to the specific derivation procedure. Some roots cannot be employed with the whole defined designs in the language because they are not compatible with their anomalies. This is why a false tagging of a root gives a false generation and a false tagging of the family of derived words from this root.

3. Tagging of roots

The great part of the Arab lexicon is structured around the three letters roots (we use notation C1C2C3 for designating the three letters of the root). The written Arab uses a different notation from the currently one used in Europe: in fact the model of a root presents itself by: فعـل fa (F), ayn (`), lam (L): three consonants which constitute a real root of the language, expressing the idea "to act, make".

According to consonants which compose the roots, we distinguish two classes of roots : *healthy* and *sick*.

1. Healthy roots or "sahih (صـــحيح)" : They don't contain any of the sick letters alif (ا), waw (و) and yaa (ي). They are divided in three sub-classes:
-"Moudhaaf (مضـاعف)" [r3s1] : which contains the doubling of a letter.
-"Mahmouz (مهموز)" [r3s2] : which contains in its letters, the letter hamza.
-"Salem (سـالم)" [r3s3] : which does not contain doubly or hamza on her letters.

2. Sick roots or "Mouatall (معتـــلّ)" : They contain one or two sick letters. They are divided in five sub-classes:
-"Mithal (مثـال)" [r3m1] : its first letter is sick.
-"Ajouaf (أجوف)" [r3m2] : its second letter is sick.
-"Nakes (نـاقص)" [r3m3] : its third letter is sick.
-"Lafif Mafrouk (لفيـــف مفـــروق)" [r3m4] : its first and third letters are sick.
-"Lafif Makroun (لفيـــف مقـــرون)" [r3m5] : its second and third letters are sick.

We signal that we have used a more detailed representation, to give more information on the roots and by consequence we obtained 28 labels which will be used for tagging.
Example: root (AKALA أكل ⇔to eat) will have the tag "three letter root Sahih Mahmouz", root (BAKAYA بكـــي ⇔to cry) will have the tag "three letter root Mouatall Nakes"

4. Realization

For the automatic tagging of these roots, we have used a neural network (NN), which is a calculation model whose design is very schematically inspired by the operation of real neurons (human or not). NN used in computer are part of the artificial intelligence, which showed its relevance in several disciplines of data-processing field like the visual recognition, filtering of the sounds and robotics in general. They were frequently used as classifiers, e.g. for the identification of anomalies.

Neural classification rests on the use of a discriminating network which makes it possible to take into account all the information of each class in order to be able to separate the inputs. There are several categories of the neural networks, among which we are interested on the multi-layer perception architecture.

A multi-layer perception is a network with static architecture, all the neurons of a layer being connected to each one of the preceding layer and having a transition course between the entry and the exit known as of "hidden". The method of the discriminating network thus rests on the use of a single network of neurons giving as output the class of membership (in our case the tag) of an unknown root in entry.

Among the advantages that can be presented by the neural networks in this field can present, there is a certain simplicity of implementation, and certainly an easy parallelization. Indeed, the computing time necessary to do the identification is related only to the complexity of the neural network. But even with many cells on the hidden layer, the propagation of the entry towards the exit is done in a quasi-instantaneous way (in ms). The size of the stored data is also strongly reduced, since the network (a few KB) replaces any need for a database.

This mechanism is ideal to carry out the automatic tagging of the thousands of roots with a phenomenal speed, thanks to the procedure of training. Tagging is composed of two phases:

1. The training: the fact of memorizing in a dense form the examples. In our case the base of training contains approximately hundred examples, each one representing a well defined particular tag of one of the subclasses of the roots. The training is made during the creation of the program, reason why the user (who is not judicious being

computer programming specialist) can use the software directly. We used the same method of training used by [9] which is a supervised method.
2. Generalization: the fact of being able, thanks to the learned examples, to treat distinct examples, which are not met yet, but similar. In our case the network is able to identify any three letter roots and to give it a single tag (to test the application we applied it to the 3822 three letter roots of the "Lisan Al Arab"[4] dictionary and it gave 100% of right results).

5. Examples of Teaching Applications

The development of computer science carried out a lot of improvements in the various fields of the education system. We are aware of a multitude of uses of data processing in the educational context. The computer assisted learning model proposed here belongs to the one of the most recognized uses, which is the natural language processing. This treatment requires the elaboration of automatic tools, among which the *tag-maker*, which is at the base of the majority of the NLP applications.

The CALL applications for the foreign languages are very recent. Here, we aimed that our roots tag-maker (developed for the person who is learning Arabic as foreign language) has the same effects as a tag-maker intended to treat the native language. For that, we adapted our system to the native language of the students (in our case French). The software application is composed of three interfaces:
1. The reception interface: through which we can choose one of other two interfaces.
2. The tag-making interface : the user selects any Arab three letter root and the program posts the Arabic label and the French code (this code makes it possible to give the definition and the explanations in French for the label chosen, where necessary). This interface is used to learn or check the tags of the roots; it replaces the teacher who is not any more the only holder of the knowledge.
3. Test Interface : by starting the application, the program chooses randomly a three letter Arab root. The user must make a choice and validate to see whether its answer is correct or not. This interface is used to evaluate knowledge of student.

6. Conclusion

Thanks to the use of neural networks, we realized a robust tag-maker of Arab roots who represents the first stage of the realization of a morphological analyzer of Arab text. This analyzer will be used in Arabic CALL applications, reason why it must have a great degree of reliability.

The next stage of our work consists in automatically generating of the other parties of the complete dictionary, witch will be the knowledge base of our analyzer.

7. References

[1] G. Antoniadis, S. Echinard, O. Kraif, T. Lebarbé, and C. Ponton, "Modélisation de l'intégration de ressources TAL pour l'apprentissage des langues : la plateforme MIRTO", *Alsic.org ou alsic.u-strasbg.fr*, Volume 8, article mis en ligne en novembre 2005,pp. 65-79.

[2] H. Achour, F. Débili and E. Souici, "La langue arabe et l'ordinateur : de l'étiquetage grammatical à la voyellation automatique", *Correspondances de l'IRMC*, N° 71, juillet-août 2002, pp. 10-28.

[3] M. Aljlayl and O. Frieder, "On Arabic Search: Improving the Retrieval Effectiveness via a Light Stemming Approach", *In 11th International Conference on Information and Knowledge Management (CIKM)*, Virginia (USA), November 2002, pp.340-347.

[4] L. Ballesteros, M.E. Connell and L.S. Larkey, "Improving Stemming for Arabic Information Retrieval: Light Stemming and Co-occurrence Analysis" , *In Proceedings of the 25th Annual International Conference on Research and Development in Information Retrieval* (SIGIR 2002), Tampere, Finland, August 2002, pp. 275-282.

[5] M. Maraoui, "Automatic tagging system of the Arab roots for learning", *Workshop "TEL in working context"*, Grenoble, 13-15 November 2006.

[6] M. Alissali, S. Baloul, M. Baudry and P.B. Mareüil, "Interface syntaxe-prosodie dans un système de synthèse de la parole à partir du texte en arabe", *24es Journées d'Étude sur la Parole*, Nancy 24-27 juin 2002, pp.329-332.

[7] M. Ben Ahmed, A. Haddad and M. Zrigui, "Un système de génération automatique de dictionnaires linguistiques de l'arabes", *TALN 2005*, Dourdan, 6-10 juin 2005.

[8] R. Zaafrani, "Un dictionnaire électronique pour apprenant de l'arabe (langue seconde) basé sur corpus", *JEP-TALN 2004, Traitement Automatique de l'Arabe*, Fès, 20 avril 2004.

[9] H. El Ayech, A. Mahfouf and A. Zribi, "Reconnaissance de la métrique des poèmes arabes par les réseaux de neurones artificiels", *TALN 2006*, Leuven (Belgique),10-13 avril 2006.

[4] http://www.angelfire.com/tx4/lisan/khamash.htm