An - Najah Univ. J. Res. (H. Sc.) Vol. 21(4), 2007
Short Communication *مراسلة قصيرة*

The Treatment of Some Deviation Cases in Non-Standard Data Using Artificial Intelligence and Natural Language Processing Applied to Arabic Language

معالجة بعض حالات الانحراف في قواعد البيانات غير المعيارية باستخدام الذكاء الصناعي ومعالجة اللغات الطبيعية وتطبيقاتهم على اللغة العربية

Bilal Alshafei

بلال الشافعي

Department of France, College of Art, An-Najah National University,

Nablus, Palestinian Territories

E-mail: bshafi2@hotmail.com

Received: (6/3/2006), Accepted: (5/8/2007)

Abstract

The aim of this study is to try to investigate the problem of deviation in non-standard data generated by inquiry into Arabic language, using research engines or some Natural Language Processing programs (NLP) applied to Arabic. The deviation could be corrected by using expert systems and some neural networking. It will try to analyze these issues from both linguistic point of view and technological point of view in order to find the best way to build a model of work that could find solutions for some deviant cases using Artificial Intelligence (Expert Systems and neural networks). Above all, this work could serve as a starting point for farther research on the deviant cases.

ملخص

تهدف هذه الدراسة إلى تقديم بعض ظواهر الانحراف في قواعد البيانات الالكترونية غير المعيارية ومحاولة معالجتها، وذلك من خلال الطرح أو الإجابة على الاستفسارات المقدمة لأدوات البحث ووسائل معالجة المعلومات بطريقة أتوماتيكية، وذلك باستخدام تقنيات معالجة اللغات الطبيعية "Natural Language Processing" والذكاء الصناعي (AI) "Artificial Intelligence" المعتمد على الأنظمة الخبيرة "Expert Systems" والشبكات العصبية "Neural Networks". وتعد هذه الدراسة جزءاً من الدراسات التي تعالج قواعد البيانات باللغة العربية باستخدام الحاسوب.

Introduction

This paper focuses on studying and analyzing the solutions and technologies that help in surmounting the complex Arabic language automatic treatment in analyzing methods and techniques used in AI to solve some problems generated by deviant cases.

This study will help us to build a model of automatic treatment for these cases that may provide us with the opportunity to publish data and make it searchable and retrievable. This could be done by using the advanced research in the field of Natural Language Processing search and retrieval technologies and utilizing the linguistic features of both Arabic and English languages in the indexing and search processes. This method could be utilized to perform morphological analysis on each word in the document to provide search types. Efforts will be focused on analyzing the deviant cases to understand this phenomenon and to find how to build logical and formal model to deal with these cases automatically.

Deviation definition

It is unusual to talk about deviation for computer treatment considering the fact that this pertains to human behavior and can not be applied to software used by computers based on natural language processing. In fact, the human behavior is totally different when it comes to deviation. A human being will eliminate all aspects that are not related to a communicative situation and will focus on the elements that relate directly to the psychological representation to elicit a meaningful communication between sender and receiver. The definition of *deviation* is *"noticeable difference from what is expected, especially from accepted standards of behavior: sexual deviation | a slight deviation from the original plan*"[(1)].

(1) Longman Interactive American Dictionary.

The terminology used by computational linguistics is an **ill-formed input**. This term describes input that violates the constraints of a "normative" system (whether pragmatic, semantic, or syntactic)[(2)] .

The term "deviation" could be used to compare human beings' behavior to a machine behavior. This way of thinking can describe many aspects that the **ill-formed input** cannot do such as: jokes, sarcasm, imagination, metaphor …etc.

As it has been mentioned, the term *deviation* could hardly be used to describe the phenomena. We can explain the term "deviation" as the noticeable difference from what is expected when using non-standard data from accepted solutions given by the NLP methods. Based on this point of view, efforts will be focused on studying and analyzing these phenomena of deviation as a part of computational linguistics analysis. This way of thinking will put on the track different types of problems related to the Arabic language: diacritics, research tools, translation,… etc. The main aim here is to describe the problem, study existing solutions, find how similar problems have been solved in different linguistic systems and what could be done in the case of the complex Arabic script…

"Deviation" is defined, from the point of view of artificial intelligence, as the gap between the original script, frame or algorithm and the scheme recognized by a society or some donating persons. This gap could have the following aspects:

- *Isolated* when a new script is interfering in our original script.
- *Repeated* when a new script takes the place of the old one.
- *Fused* when many scripts are mixed and the result is an unstable script.

(2) Sondheimer W. (October 1980). "A Rule-Based Approach to Ill-Formed Input". Proceeding of the Eighth International Conference on Computational Linguistics. 46-54. Tokyo, Japan.

To deal with the problem of human deviation, two types of tools have been found instrumental to enhance NLP, expert systems and neural networks.

Emphasis will be mainly placed on the expert systems because of their good performance in treating Arabic language script and their facility of construction.

Could Artificial Intelligence produce systems without deviation?

To answer this question, it should be kept in mind that the expert systems or neural networks used in AI were made in a way that the problem of deviation could not be produced. The analyst programmers take all possibilities into consideration to produce the expected results from this software, despite the fact that some systems could have random methods. If a solution cannot be found or if a problem of deviation takes place, the system will give the answer "syntax error".

Some automatic correction could be easily found. But if we don't respect the algorithm or have an unexpected result, this would not find a solution or it would put an end to the process.

Example:

هذا البيوت -- هذا البيت - هذه البيوت،

وترجل الفرس بجانب النهر

Automatic correction will find the first sentence incorrect and the user can intervene to correct the sequence (the error may be placed at the 2 words). In the second one, the automatic correction will not be useful (a test made using many types of software has attested this sentence as correct – fig 1). In this case we can use an expert system to identify the verb [*Tarajjal*a] and connect it to a human being's activity.

Fig (1): Does AI have the tools to analyze deviation?(3)

Expert Systems can analyze and treat deviant cases because they use logic treatment based on predefined algorithms and schemes. We usually build ES from the output, which means that the results we get are already known. They can measure the gap from the norm. To have a good analysis of a deviant problem, we have to take into consideration different aspects of NLP applied to Arabic language:

1. Different forms (vowels or not)
2. Right to left script
3. Agglutination
4. Different levels of Arabic
5. Regional languages
6. Different periods and
7. Continuous text level

(3) Madec, H. (1993). "Systèmes expert et réseaux neuronaux : a propos de déviance". Actes des rencontres Besançon- Neûchatel. France – Suisse.

To accomplish this, we have to consider the use of large electronic dictionaries that contain all forms: vowels or non-vowels, grammatical classes, ambiguous forms …etc

The choice of these dictionaries could be explained by the large amount of forms that the Arabic language could produce. In fact, the forms without vowels could reach 500, 000 entries and the forms with vowels will be about 1,000, 000. At the same time, if we add all the agglutinated forms, we will find more than 100 million forms. We can also add grammatical classes to these forms which include the following categories [(4)]:

∑verb at the 3rd person masculine singular/active,

∑ Verb at the 3rd person masculine singular/passive,

∑substantive masculine plural,

∑substantive masculine singular and

∑imperative verb 2nd person masculine singular.

The AI has two ways to deal with the deviation in non-standard databases: expert systems and neural networks. The first can be very helpful to build a model of treatment that can stimulate the behavior of an expert facing the problem of recognizing indicators to find out the deviation then diagnose its forms and evaluate the gap from the norm. The second can be helpful in explaining the individual behavior related to a deviant case and the evolution of such phenomena.

These two tools cannot work if we do not provide the essential dictionaries mentioned above. These tools could be completed with full analysis of deviant cases treated with the two elements of AI.

(4) Debili. F. (2001). Traitement automatique de l'arabe voyelle ou non. CNRS: ITAAF, Paris, France.

The indicators that could help to understand the deviation

- *Indicators given by the subject of deviant case*

We have to know the reasons behind talking about the subject in the deviance case to understand all the elements that could be helpful in building the expert analysis. We can take the case of someone who is producing a (written or spoken) sentence about the weather with a deviance problem (using the word غيث to mean وحل as some Moroccans do). The elements that will give us the indicators to understand this deviant case could be:

- Circumstances in which the speech was produced (context): in a class, in the street, with friends …etc.
- Communication aspects: who is speaking to whom, when, why, what message s/he wants to convey and in which language …etc.
- Situation: This includes political, economic, social, cultural, religious, regional, historical ….etc. It is very important to understand the choice of the subject of speech.
- Motivation: why he chose this subject, what idea may s/he have, what the purpose of the speech is, what s/he means ….etc
- Socio-cultural aspects: Is there any cultural or social interference? Does the whole community use this kind of word or is it an individual case? ….

- *Indicators given by actor of the case*

When a subject produces oral or written speech containing a deviant case, we have to focus our attention to elicit the maximum information about him or her in order to collect all indicators that may help us in studying the nature of the deviant. For example, if we have a sentence containing a plural form with a wrong form [tifl → ?atāfīl] (طفل --- اطافيل) we can conclude that the subject who produces this sentence could be:

a child under 5 or 6 years old;

an adult who is imitating the language of children;

a foreigner who can't master the rule of plural nouns in Arabic;

an illiterate person;

an adult who belongs to a minority not using Arabic as a mother tongue;

others

- *Indicators that will generate the forms*

It is important to know that this case may generate other forms that can be added to the previous dictionaries: under which circumstances? How many? The frequency of this kind of errors? In the case of the plural noun we know that the use of plural nouns in Arabic is very intricate especially, when the root of the word has to be changed.

(حجر—حجارة\ كيس—أكياس).

It can be said that this case will generate many forms that have to be taken into consideration. We can find a way to generate automatically all forms related to this grammatical deviation that happens regularly. We can also take this case and study all relevant details to help human experts to understand and analyze all the aspects concerning this case in order to build a model of treatment for other similar cases.

- *Classification of indicators*

The classification of results from all the indicators that we have treated must be understandable, precise, accurate, methodological and practical. It must also be open, i.e., allows future improvements. The data gathered from all these indicators are so huge that we have to treat them automatically, i.e., software that does not ask for treatment of indicators which have already been classified (except for ambiguous cases that would have been tagged).

- *Treatment of indicators*

The treatment of indicators must be done by a human expert who will take into consideration a two-level analysis: from a linguistic-legal point of view and a technological point of view, in order to find the best

way to build a logical hierarchy, then the algorithms that will form the model of work. A human expert must try to produce an ES that can deal with the deviant cases on the basis of stimulating the behavior of the expert, recognizing indicators that can lead to finding out the deviation then diagnosing its forms and evaluating the gap from the norm. A model of neural network can be added to explain individual behavior related to a deviant case and its evolution.

How can we explain indicators?

- *Describing a collection of rules and conditions*

From the work done on the previous items we can search for rules and conditions that can reduce the huge amount of data. At the same time, the results of this work explain mechanisms of deviant cases in the Arabic language. These mechanisms can also improve our comprehension of the complex behavior of deviation. Some of these rules or conditions could be used as standards for neologisms or other linguistic norms.

- *Taking into consideration all the other factors*

Delimiting all the factors of the case is very hard to achieve. For this reason algorithms must allow the intervention of the expert and the evolution of the scheme in a way that the improvements can take place. As previously mentioned the ES and or neural network must be open and accept the intervention of the expert. It must also be capable of learning - using the feedback to learn from repeated cases.

- *Explaining indicators by a human expert*

The work of a human expert is very important to put the diverse indicators in relation; i.e., to be classified. Consequently, s/he must explain these relations and how they operate. But how can we understand the complexity of these cases? We can do so by explaining all the aspects concerning their function and the task of rules and conditions that we might find, and also by clarifying the methodology by which these rules and conditions can be employed to explain how they work in our

linguistic system. It should be understood why a human being can solve the deviant cases without problems (s/he can choose from a speech what s/he considers as relevant and correct deviant cases in a simple and natural way – the use of some psychological methods may explain this behavior). The result of this work may lead to a clearer and more correct way of understanding deviant cases and their mechanisms.

- *Constructing a model for the treatment of the indicators*

The last step in this analysis is to construct a model of automatic treatment for the deviant cases that can be used and to improve the different tools used in NLP (e.g., electronic dictionaries).

Conclusion

It is not easy to deal with deviant cases in the non-standard Arabic, especially when this work has to be done without any connection to the ongoing research in this field. This study tried to highlight issues related to deviant cases and how the AI can be useful in dealing with these issues. At the same time, it should be kept in mind that this work will remain incomplete without the help of other studies in the different fields of NLP applied to Arabic. So we have to combine our efforts with other researchers in this field to find a framework that may be the starting point for extensive research on the deviant cases and the linguistic problems related to them.

The use of AI to solve the problems generated by the deviant cases illustrates the powerful structure that it can offer for the possibility of using different formal analyses to handle the problem of deviant cases.

This study could empower Arabic in international solutions and enable the end user to search and retrieve information from Arabic and non-Arabic documents. Finally, it may be very helpful for other advanced research that would build on these results and try to offer a piece of evidence to support them.

References

- Dardin, J. (1987). Systèmes experts et sciences humaines. Eyrolles, Paris, France.
- Debili, F. (2001). Traitement automatique de l'arabe voyelle ou non, CNRS: ITAAF, Paris, France.
- Dutta S. (1993). Knowledge Processing and Applied Artificial Intelligence. Oxford, England: Butterworth.
- Garnham G. (Aug 1995). "Parsing in Context: Computational and Psycholinguistic Approaches to Resolving Ambiguity during Sentence Processing". Language and Cognitive Processes. 10 (3-4). 377-81.
- Lehnert R. (1982). Strategies for Natural Language Processing. L.E.A, London, England.
- Madec, H. (1993). "Systèmes expert et réseaux neuronaux : a propos de déviance". Actes des rencontres Besançon-Neûchatel. France – Suisse.
- Sondheimer W. (October 1980). "A Rule-Based Approach to Ill-Formed Input", Proceeding of the Eighth International Conference on Computational Linguistics. Tokyo, Japan.
- Voyer, K. (1987). Moteurs de systèmes experts. Eyrolles, Paris, France.