Prospects for NLP in Algeria

BOUHADIBA Farouk
Université d'Oran

Résumé

Le traitement automatique des langues, communément connu sous l'acronyme TAL et TALN –Traitement Automatique des Langues Naturelles- est un domaine de recherche pluridisciplinaire qui implique des linguistes, des informaticiens, des logiciens, des psychologues, des lexicographes, des traducteurs etc. Il fait partie du domaine de l'Intelligence Artificielle (I.A.). Certaines tâches bien délimitées mettant en jeu le langage peuvent donner lieu à des programmes satisfaisants, mais dans la plupart des cas l'obtention d'une qualité identique à celle de la traduction par l'homme nécessite une intervention humaine en amont (pré-édition, simplification, etc.) ou en aval (post édition; post correction).

Le TAL est également porteur d'une réalité socio-économique. Il est plus que nécessaire de nos jours de lui 'faire une place' dans la formation et la recherche en Algérie qui demeure en retard dans ce domaine par rapport à ses voisins et par rapport à l'automation de la langue arabe à travers le Monde.

1. Introduction

Natural Language Processing (NLP) has long been looked at as 'incompatible' with the Arabic Language. As early as the 50's, a number of studies were conducted especially in the US to try to find ways to handle this language for computing and machine translation. Goldsmith's (1976) auto-segmental apparatus for example and other generative models have shown that Arabic, as any other language, can be segmented and analysed by the machine. Today's web sites in

Arabic, SMS messages, online dictionaries, automatic translators etc. show clearly that the Arabic script, syntax and morphology can be handled by algorithmics and other computing devices to put this powerful derivational language into information and data collection networks.

It is paramount for Algeria to motivate studies, training and research in this particular field. As a first step, this can be done by encouraging researchers to investigate this avenue of exploration and students by offering them the necessary environments (LMD, NLP Departments, NLP Research Centres …) for studies and research in this vein. It is more than necessary for us today to consider implementing programs and courses that meet the needs of our society in a thriving and full motion world. This is particularly true in the field of communication across boundaries where different languages are at work for swift information gathering and data processing.

Compared to the last two decades or so, the Arabic language has finally gained momentum in the world of information networks, translation and transfer. One of the first reasons for this development of Automatic Arabic Translation is its presence in the United Nations since 1974 when it became the official language of this international organization.

Nowadays, many socio-economic sectors rely on natural language processing and data mining. Automation by means of controlled languages has proven very useful in sectors like Health, Security, Trade and Commerce, Banking, Tourism etc. NLP will offer job opportunities for its degree holders in the labour market.

Language processing has evolved considerably in recent years especially in the areas of research and development in industry and international trade. For example, consider the case of an engineer who works in the Deep South in Hassi Messaoud, Hassi R'mel, Roud Ennous or even in Algiers, Oran, Annaba, etc., He wants to order spare parts for a mud pump from Houston, London or Paris. To do this, he has to find a human translator who knows how to use the proper terminology for the request in question. Assuming that such a translator is available and who masters the specific field of hydrocarbons - which unfortunately is often not the case – the latter has to find technical terms in a paper dictionary to formulate the request in English for instance, send the request – by email or by fax, etc. -, and wait for an answer to his order from the supplier or manufacturer on the other side of the world. The receiver of the order has in turn to understand the message he received from the client, sometimes he has to decipher the message and then send a reply (in English or any other language) that is in turn read and understood by the customer who then decides whether it meets his specific needs for the requested spare parts. In case the message of the sender is not understood by the supplier (which is sometimes the case because the English of the translator was not up to the standards), there is a considerable waste of time and money, the price of the requested spare part may have risen in the mean time or it may even be out of stock.

Such miscommunication problems and transaction difficulties and hurdles can be avoided thanks to NLP in the field of hydrocarbons in this case. By using a controlled language, the engineer of the Deep South will simply select the appropriate dictionary or machine translator on his computer, his laptop or even his mobile phone (for the necessary technical terms, ISO references types of spare parts, etc.) to formulate his request to the supplier. An electronic dictionary or automatic translator can be designed for this purpose. The user can also select a protocol for technical writing and send the report to receive a swift response.

It is in the specific areas of safety and health, among others, that controlled languages - and thus NLP - have proved to be necessary and reliable in cases of emergency. Many accidents, fires and other hazards that threaten our security could be avoided in Algeria and elsewhere. For example the fire disaster that ravaged the Railway station of “Gare de Lyon” in France, resulted on June 27, 1988 in 56 dead victims among passengers and staff. The final report on this incident mentions that: "The Committee also noted a defective typographical presentation [...] which led to misinterpretation [of the fire accident]." (My translation).

On the same line of thought, according to a joint study by Pharmaciens Sans Frontières, a large part of the 4000 tons of medicine sent after the Tsunami in the Indonesian province of Aceh (2 million people) had to be withdrawn quickly from supply and delivery because 60% of these medical supplies were written in a language that the health personnel and the Tsunami victims were not able to understand for a proper use of the medicine in question.

With controlled languages and technical protocols, error messages of all kinds, misinterpretations and misunderstandings can be avoided in seconds and at the press of a button on the computer. This being so because the everyday language is full of ambiguities with sometimes particular twists, polysemic forms and structures, images, onomatopoeia, direct speech formulations and indirect speech formulations, etc. This means that a distortion of the message between the sender and the receiver is may occur and end up with dramatic results. This is mainly true in cases of telephone instructions, orders, or recommendations and in written reports. For example: We may say to someone: *'Il y a le feu à l'Amphi 4'* ('There's fire in Amphi 4') where the word 'feu' (fire) is used in an imagery form to mean that there is trouble or a problem of some sort in Amphi 4[1]. Although the ambiguity does not seem to appear in the English version, it is clearly present in the French utterance. Similarly, an utterance such as *'He told me about his love'* may mean talking about 'one's feelings' or about 'one's second half'.

To meet today's quality and safety measures, the processing of information is of paramount importance and necessity. A Controlled Language (CL) is a set of editorial writings tailored to the information needs of the user. These facilities can simplify and standardize the language used in any procedural text to meet, among others, the following objectives:

- To ensure a rapid and effective transfer of information.
- To ensure a full understanding of the dispatched text.
- To prevent misinterpretation due to natural language ambiguities.
- To reduce costs attributed to misinterpretation (misuse of a product, failure to comply with orders, etc.

[1] . This utterance was produced by a colleague who was invigilating during an exam to mean that he was not able to invigilate on his own so many students in Amphi 4.

- To reduce the time and costs associated with the processing
  and validation of technical text translations.
- To alleviate stress in sending emergency messages.

2. The economic impact and importance of NLP in Algeria

Algeria is de facto involved in the process of globalization. The country has thus to provide educational facilities, among other update changes, which help for a better handling of the transition to a market economy. Globalisation requires, as it were, the use of modern communication technologies, ICTs and technical electronic reporting at all levels and in all sectors of the Algerian Economy. It is of paramount importance today to motivate Algerian executives and other socio-economic partners not only to meet the standard requirements of globalization, but also to handle by means of language and information processing economic objectives in terms of technological transfer and information gathering.

In the educational sector, training for NLP in the general field of Language Sciences is but one component of these necessary update changes for Algeria. To do this, Departments or Research Centres in NLP must be established together with the elaboration of appropriate teaching programs in Natural Language Processing at University level.

The University of Oran, among other universities in Algeria, has made some substantial progress in this vein. However, these efforts are under the form of scattered research in some accredited research laboratories. The latter will have to concentrate their research objectives in NLP around focal points or Research Centres for the development of language engineering and programming in Controlled Languages and to assess by the same token the role of the Arabic language in information gathering and networks.

The following suggestions in the automatic processing of three languages: Arabic, French and English can be integrated into a global system of automatic translation between languages.

The first step is to analyze the systems of these genetically different languages to perform a more systematic analysis of the morpho-syntaxic, lexical and semantic components in order to develop rules between these languages and formalize them under the form of algorithms and rules of transfer from one language to another.

The next step is to proceed to the annotation and modelling of the Arabic Language by means of language segmentation, Arabic vowel recognition, micro systemic analysis, lexicography, etc. and most of all sorting out the problem of the Arabic writing directionality compared to French and English. More importantly, research work has to be conducted on the basis of the transfer where the first input language is Arabic as research has long been led on the basis where the input language is French, English or any other language like Spanish, Russian, etc. The point is that the Arabic language is particularly difficult to treat by computer language. This issue on the processing of this language has not been sufficiently addressed in the global research on NLP (data mining, information gathering and language automation and control in particular). Slow progress has been made in this field compared to other languages such as English, French, Russian, German, Chinese etc.

Our major concern for the development of this type of research on electronic dictionaries and automatic translators in a field like that of hydrocarbons in Algeria is motivated primarily by an ever increasing demand by users in this area. The latter (Field Engineers in Exploration and Production Divisions, Technicians in Maintenance and Safety, etc.) receive documents in English on the technical and production characteristics of spare parts of a mud pump, a gas turbine, a crown block or drill pipes, etc..) but they have tremendous difficulties understanding them in English. They thrive for this type of online Arabic, English and French dictionaries to help them do their job properly. This is mainly true for those who have some basic knowledge in English and are good at Arabic and/or French. The same translation and comprehension difficulties of technical documents which are not written in Arabic emerge when these engineers, technicians or any other user in this field have access to guidelines, application notices and other technical documents in English for their job. These documents cover a wide range of hydrocarbon activities in areas such as Start-up (start-up procedures), Exploration (analysis of geological structures, seismic graphs, maps, etc.). Production, Development and Process (drilling, estimated oil and gas reserves, quantitative and qualitative analyses, reserves estimation, uncertainty, water saturation, decline curve analysis, etc..). These users have to some extent the mastery of the basics of the English language, but they are faced with two major problems:

a. Try to translate and understand the document using a paper dictionary.

b. Call a human translator who sometimes does not master the terminology in the field of Hydrocarbons.

In both cases, this results in a considerable waste of time, especially when it comes to responding within very tight deadlines to a document in English (received by fax, email, or any other means of communication). Erroneous if not false translation in this case is much due to the abundance of the technical terms that the document contains rather than to the syntax of the language in which the document is written. A case in point would be the expression “*dog leg drilling*” in the area of oil field exploration and drilling which is translated as “forage par pied / jambe de chien” in French or “المنعطف الحفر” in Arabic. According to the engineers in drilling and the exploration of oil wells, both technical terms mean almost nothing to them when it comes to this particular way of drilling to find the ‘trap’ or ‘oil source’ in an oil field.

NLP and Controlled Languages (Arabic, English, French in this case) impose themselves as inevitable tools, devices and facilities for a better performance of the job, a better safety and a better control of the equipment reliability. In doing so, a better production is achieved which in turn generates better economic benefits for the oil companies and for Algeria in general.

References

ALSHAWI, H. (1992). *The Core Language Engine.* ACL-MIT Press series in Natural Language Processing.

ARNOLD,D.J.,(2003), *Why Translation is Difficult for Computers*. In Harold Somers, Eds. Pp. 119–142. John Benjamins, Amsterdam.

BIROCHEAU, G. (2000). Morphological Tagging to Resolve Morphological Ambiguities. In *Proceedings of the Second International Conference on Language Resources and Evaluation*

BOGACKI, Krzysztof, (2009), *Controlled languages and Machine Translation,* ISMTCL Proceedings, International Review Bulag, PUFC, ISSN 0758 6787, ISBN 978-2-84867-261-8, p. 49-55.

BOUHADIBA, Farouk, (2004), *L'infixation en Arabe et la concaténation en TA : questionnement d'un linguiste,* 2èmes Journées sur la Traductologie, 5-6 mai 2004, Besançon, France

BOUHADIBA, Farouk, (2007), *La langue arabe et le TAL : étude de cas*, RML 5, 2007, Actes du 1er Colloque International en Traductologie et TAL, 9-11 avril 2007, Université d'Oran, Algérie.

CARDEY, S., (2003), *Modélisation, systémique, traductibilité*, BULAG, 28, Université de Franche-Comté, Besançon, 2003.

CARDEY, Sylviane, (2009), *Machine Translation of Controlled Languages for More Reliable Human Communication in Safety Critical Applications*, The 11th International Symposium on Social Communication, ACTAS, Santiago de Cuba, p. 330-335.

GAVIEIRO-VILLATTE E., SPAGGIARI L., (1999), *Open ended Overview of Controlled Language*, in BULAG n°24, pp. 89-100

GOLDSMITH, J. (1976) : Auto-segmental Phonology
(Outstanding Dissertations in Linguistics), Jorge Hankamer Ed. Garland Series, New York & London, 1979.

HUTCHINS, W. J., SOMERS, H. L., (1992), An *Introduction to Machine Translation*, London, Academic Press.

HUTCHINS, W.J., (1979), *Linguistic Models in Machine Translation.* UEA Papers in Linguistics 9, pp. 29-52.