Idyouss Lahousseine

On the 2nd International Conference on Arabic Language Resources and Tools

Arabic Natural Language Processing seems to be moving forward faster than I thought before. The Second International Conference on Arabic Language Resources and Tools, Held the 22nd and 23rd April 2009 by MEDAR consortium in Cairo (Egypt), has given me a good picture about how far this field has gone up to now. I was indeed fortunate to attend this conference, from which I have learned a lot about the existing Arabic language resources, players and products, and during which I followed with interest the participants presenting their interesting projects and tools for Arabic.

The conference was truly a golden opportunity for me to learn about some of the additional obstacles facing Arabic NLP in particular, compared to other languages. The specific additional layer of complexity which attracted my attention relates to Arabic standardization with respect to vowels and diacritics, a problem which received more emphasis especially during the tutorials in the Faculty of Computers and Information, Cairo University which took place the 21st of April (one day ahead of the conference). This issue seems to interestingly figure as a serious problem in almost every presentation even during the conference. It is not easy to say if the problem could be resolved at all without calling for more standardization of the language. However, even if this constitutes an obstacle in many Arabic NLP tasks, yet the field is moving ahead steadily. In the following paragraphs, I am going to shed light on some of what I gained from attending this conference.

In order for me to demonstrate the basis for my positive evaluation, I would like to say few words about my present activities, which encouraged me to apply for the MEDAR conference.

I am currently in the process of defining my PhD project, which I plan to work on under the co-supervision of Professor Geeraerts and Dr. Steurs at Lessius/KLUeven University, Belgium. My interest has led me to do my research in the field of semantic similarity assessment, a technique which is widely used in a number of Natural Language Processing tasks such as text categorization, information retrieval ETC. The two common methods for assessing semantic similarity are the lexical-based approach—using thesauri or WordNet—and the corpus-based practice. Among the points I am planning to deal with in my research, which there is not enough space to fully present, is using both methods for Arabic to see which one can produce better results. One of the famous models used for semantic modelling, or more particularly for semantic similarity assessment, is the Word Space Model, which is a model of semantic similarity which uses statistics of distributional properties of words to assess proximity and thus similarity between words. It is worth mentioning here that the research unit at Leuven, within which I will hopefully be working, requires a corpus with a minimum size of 20 million words. I probably have to point out here that this is not the exact thesis I mean to elaborate upon, but I have selected these points so as to illustrate the relevance of the conference in Cairo for my research, as I will demonstrate shortly.

As can perhaps easily be deduced from this shortish initiation into my research, my primary motivation behind attending the conference is searching for Arabic lexical resources and corpora. I wanted to learn as much as possible about the availability and accessibility of these resources, their quantity and quality in terms of size and annotation, some of the particular challenges that they face in comparison with other languages, besides the copyright issue ETC.

I am happy to report that the conference as a whole enabled me to answer almost all the questions I had. In this respect, I would like to highly value my meeting with Dr. Christopher Cieri, the executive director at LDC, with whom I exchanged contacting information for future cooperation with regards to corpora. He kindly calmed down my worries about the copyright problem and promised that once I am settled on a thesis, I can contact LDC for the procedure of obtaining access to GigaWord, a huge corpus of 600 million words. This will obviously solve the minimum size of 20 million words required by the Word Space Model for the type of research I will undertake, as I mentioned above.

As for the WordNet, I was fortunate to meet very interesting people with whom I exchanged the most recent stages of the development of this interesting lexical resource. Certainly, it could have been possible to search for these resources and all the related information via the internet and contacting the people concerned by means of the "rich contacting information they provide in their web pages". However, I believe that making direct contacts with the people in the field is much more enriching and better than sending "unknown" messages to "unknown" people, to which replies are unfortunately not guaranteed. Fortunately now with almost all my questions answered thanks to the conference, I am heading closer to the final stage in defining my research topic.

In addition to these direct practical considerations, I have also learned a great deal from the talks during both the tutorials and the conference about Arabic natural language processing, which I will certainly need in the process of my research in the near future. I am truly fortunate to have followed the talks, presentations and interventions by Dr. Habash, from whom I learned a lot. I also just cannot think or talk about this conference without remembering the special impact of Professor Anty way's presentations on my understanding of the statistical approach to machine translation. The interesting thing here is that although the discussion of the professor was mainly on machine translation, I could easily make sense and use of a number of statistical notions which are used in other NLP areas. This helped me understand a lot about some of the concepts I have come across in my readings about the Word Space model. This is needless to mention that the conference was an opportunity for me to meet some of the most active researchers in Arabic NLP, most of whom I was already familiar with their writings during both my passive and active reading stages in my research.

To sum up, I would like to state it clearly that I learned during this conference more than I expected I would. As I have shown, joining the conference has served me in two ways. On the one hand, it has beautifully responded to my immediate needs for my research with respect to Arabic resources, and on the other, it contributed to deepening my understanding of Arabic NLP in general. Before closing this little contribution, I would like to express my heartfelt thanks for MEDAR for having made it possible for me to attend the conference. I convey my sincere congratulations to MEDAR and to all those who have put a great deal of efforts towards organizing and preparing for the conference on the success of the event. My profound respect goes for all the participants, especially the organizers—those who did all they can to make the whole process as smooth as it was….