Preface

This special issue of the International Journal on Information and Communication Technologies (IJICT) on “Advances in Arabic Language Processing” is dedicated to publishing a subset of the high quality papers of the International Conference on Arabic Language Processing (CITALA 2009), held in Rabat, Morocco, May 3-4, 2009. It focuses on recent trends on Arabic language processing in order to reflect the progress made in this field.

Indeed, the United Nations adopted the Arabic language as one of its 6 official languages and it is a language spoken by over 300 million people in the world. In addition, Arabic has become a major language for Human Language Technology. Therefore, we focus on specific issues that would help citizens living in Arab countries to have access to information and technologies (open source resources, dictionaries, search engines, grammar checkers, topic identification, etc.) in their mother tongues and therefore discusses requirements to customize existing technologies. This special issue identifies problems of common interest, and possible mechanisms to move towards solutions, such as sharing of resources, tools, standards, sharing and dissemination of information and expertise, adoption of current best practices, etc.

The international conference has received 56 (from 8 countries) submissions, which were peer refereed by the international program committee members. Among them, 28 papers (42%) were accepted for presentation at the conference covering Arabic Language Processing related topics, such as: resources, morphology, syntax, information retrieval, speech synthesis, learning, named-entity recognition, topic identification, and Arabic scripting.

For this special issue, the program committee suggested 10 articles from the CITALA 2009 proceeding to submit an extended version of their papers and have been more carefully and secondly reviewed by the special issue reviewers committees to make sure that they meet the journal standards. Finally, the judgment was to only accept 4 to be included in this special volume of the IJICT journal. In addition to these 4 papers, we are proud to have added two distinctive papers to our volume, which are also refereed: one of our CITALA 2009 invited talk conference paper and an invited paper. The authors come from various countries: Morocco, Algeria, Tunisia, Spain, Egypt, and USA.

Our special issue covers several issues about Arabic Language Processing: Arabic machine translation, approaches to follow, development and use of resources, syntax and topic identification.

The first article is the invited paper. A. Farghaly presents a historical survey of machine translation in general and Arabic machine translation in particular. The survey starts with the direct method and moves on the transfer and then to the most recent statistical and machine learning methods. He presents a description and a critique of each of the approaches adopted in the field to deal with Arabic machine translation. Then he focuses on the current dichotomy between symbolic based and statistical and machine learning approaches. While there are situations where it would be more relevant to adopt one approach than the other, he takes the position that with regard to the Arabic language, it would be more appropriate for research to pay more attention to the symbolic approach while not neglecting the advances made in machine learning.

The second article, presented in the CITALA 2009 conference as an invited talk, is about approaches to follow when dealing with an Arabic Language Processing problem. K. Shaalan distinguishes the rule-based and the statistical-based approaches and presents the advantages and disadvantages of each approach with a preference to the rule-based one because of many reasons such as the lack of resources of the Arabic language and the data sparseness. From his long experience with the rule-based approach, the author then details how it can be applied for different Arabic Language Processing tasks such as morphology, syntax, machine translation, named-entity (NE) recognition and computer assisted-language Learning.

The third article is about resources which is a very important issue in Arabic Language Processing. Indeed, the development of sophisticated applications in Arabic NLP suffers from the lack of available electronic resources. The Arabic WordNet (AWN) lexico-conceptual ontology is one of these few resources. The design of AWN presents many advantages for its use in the context of ANLP because it has the same structure as the Princeton WordNet and WordNets of other languages. H. Rodriguez and M. Alkhalifa briefly report on the current status of Arabic WordNet and focus on its rather limited NE coverage. For this purpose, authors expose the automatic extraction of Arabic Named Entities (NEs) from the Arabic Wikipedia, their automatic attachment to Arabic WordNet, and their automatic link to Princeton's English WordNet. The proposal is presented, applied, and evaluated.

The fourth article is about Question/Answering (Q/A) systems with a large use of the AWN resource. Research in the field of Q/A have known significant progress for languages such as English, Spanish, or Italian. In the context of the Arabic language there are few attempts for building Q/A systems and it is still a challenging task. There are several lines that may be considered for the improvement of Arabic Q/A systems such as the improvement of the question analysis task as well as the enhancement of the Passage Retrieval module. L. Abouennour, K. Bouzouba, and P. Rosso describe an approach for improving the re-ranking of passages for Arabic Q/A. This approach implements a process performing a semantic Query Expansion based on the AWN ontology with a structure-based PR based on the Distance Density N-gram model. Authors present experiments showing that the accuracy, the Mean Reciprocal Rank and the number of answered questions have been significantly improved.

The fifth article presents an approach to tackle the relative clauses phenomenon at the syntactic level. In any natural language, syntactic analysis is fundamental for other phases such as the semantic analysis. It is also necessary for several applications dealing with natural language such as human-machine dialogue systems, automatic translation and grammatical errors correction. Despite this importance, the syntactic analysis has not been properly explored in for the Arabic language, especially for complex phenomena such as relative clauses. For this purpose, K. Haddar, S. Boukedi and I. Zalila present an HPSG grammar based on a type hierarchy inspired from classic Arabic and respecting the Arabic language specificities. Authors present also experiments to demonstrate that obtained results are satisfactory.

The sixth and last article is about topic identification. This task has been sufficiently studied for Indo-European languages using text categorization methods such as Bayesian classifiers, decision tree, neural networks, kNN “k Nearest Neighbors” etc. Nevertheless, for Modern Standard Arabic, few works have been carried out. M. Abbas, K. Smaili, and D. Berkani evaluate two text categorization methods applied to Arabic documents, namely: the well-known kNN method and the TR-classifer, a new method based on triggers. Authors conducted experiments showing that TR-Classifier has the advantage to give best performances compared to kNN, by using much reduced sizes of Topic Vocabularies. TR-Classifier performance is enhanced by increasing jointly the number of triggers and the size of topic vocabularies. A general vocabulary is needed for kNN, and it is obtained by the concatenation of those used by the TR-Classifier.

Finally, on the occasion of this third CITALA conference, we thank the Scientific Committee, composed of colleagues from all over the world, for accepting to review the conference submissions and the selected papers of this special issue of the IJICT journal and feeding the authors with their expertise on Arabic language processing.

Very special thanks to all the authors, who provide the ‘substance’ to CITALA and to this special issue, and give us such a broad picture of the field. We would like to take the opportunity to thank all those who contributed so hard to making this conference a success. We express our big gratitude to all the sponsors that have believed in the importance of our conference, and have helped with economic support.

Guest Editors

Karim Bouzoubaa, Mohammadia School of Engineers, Morocco.
e-mail: karim.bouzoubaa@emi.ac.ma

Ali Farghaly, Oracle Coporation & Monterey Institute of International Studies, USA.
e-mail: ali.farghaly@oracle.com

Khaled Shaalan, The British University in Dubai, UAE.
e-mail: khaled.shaalan@buid.ac.ae

Editorial Board