Int J Speech Technol (2016) 19:175–176
DOI 10.1007/s10772-016-9344-6

Note from the Guest Editors: Special issue on Arabic Natural Language Processing and Speech Recognition: A study of algorithms, resources, tools, techniques, and commercial applications

Mohammad A. M. Abushariah[1] · Amy Neustein[2] · Bassam H. Hammo[1]

Published online: 4 May 2016

Arabic language is a Semitic language and one of the six official languages of the United Nations (UN). It is one of the most widely spoken languages in the world. It is considered the first language of approximately 300 million native speakers situated mainly in three geographical regions, Levant, Gulf and Africa. Non-native speakers spread all over the world can reach four times the number of native speakers.

Arabic language consists of three major forms namely Classical Arabic (CA), Modern Standard Arabic (MSA), and Dialectal Arabic (DA), whereby each form has its own distinctive characteristics.

Classical Arabic (CA) is treated as the most formal and standard form of Arabic mainly because it is the language of The Holy Qur'an, religious instructions of Islam, and classical literature. It is also referred to as the Qur'anic Arabic language.

Modern Standard Arabic (MSA) is the current formal linguistic standard of Arabic language, which is widely taught in schools and universities, used in the office, the media, newspapers, formal speeches, courtrooms, and any kind of formal communication. Arabic language researchers classify MSA as being the only acceptable form of Arabic language for all native speakers. In addition, there is a tight relationship between CA and MSA forms, where the latter is syntactically, morphologically, and phonologically based on the earlier. However, MSA is a lexically more modernized version of CA.

It is important to note that neither CA nor MSA forms can be treated as the natural spoken language for all Arabic native speakers. However, Dialectal Arabic (DA) also known as Colloquial Arabic is the natural spoken language in everyday life. It varies from one country to another and includes the daily spoken Arabic, which deviates from the standard Arabic and sometimes more than one dialect can be found within a country. From writing and publishing perspectives, DA cannot be used as a standard form of Arabic language and does not have any commonly accepted standard for the writing system, because each dialect has its own characteristics that can be different from all other dialects and even from the MSA form, which affect the pronunciation, phonology, vocabulary, morphology, and syntax of Arabic language. Although there are many dialects for Arabic language, researchers mostly categorize them into two major categories namely (1) Western Arabic, which includes the Moroccan, Tunisian, Algerian, and Libyan dialects, and (2) Eastern Arabic, which includes the Egyptian, Gulf, and Levantine dialects.

In the last decade, Arabic language research has grown up significantly worldwide due to its importance and international recognition. Indeed, various research efforts are devoted to Arabic Natural Language Processing (ANLP) and Arabic Automatic Speech Recognition (AASR). In addition, several contributions and improvements have been achieved for various language processing components including morphological analysis, parsing, named entities recognition, audio transcription, acoustic modeling, language modeling, phonetic dictionaries, continuous speech recognition, speaker identification, etc., which involve all forms of Arabic language. However, in spite of the importance of Arabic language and

✉ Mohammad A. M. Abushariah
m.abushariah@ju.edu.jo

[1] Computer Information Systems Department, King Abdullah II School for Information Technology, The University of Jordan, Amman, Jordan

[2] Linguistic Technology Systems, Inc., Fort Lee, NJ, USA

contributions achieved, its research in ANLP and AASR is lacking in many aspects and research efforts are still required worldwide. In fact, more collaboration should be formed in order to contribute to the state-of-the-art and overcome various issues that face the research community of ANLP and AASR.

This special issue on ANLP and AASR found its seeds of germination in a very special session titled "Arabic Natural Language Processing: Algorithms, Resources, Tools, Techniques and Applications." This session was organized by The University of Jordan, The American University of Sharjah, and University of Leeds, in conjunction with the IEEE First International Conference on Communications, Signal Processing, and their Applications (ICCSPA'13), February 2013, Sharjah, UAE. We extended this initiative and proposed a special issue in a specialized high quality and reputed journal, choosing the International Journal of Speech Technology (IJST) to bring together researchers worldwide with research interests on ANLP and AASR in order to raise awareness of different perspectives and practices, and to identify some common themes.

This special issue focuses on providing an overview of the state-of-the-art. It explores new directions and emerging trends regarding ANLP and AASR and their relevant applications. Moreover, it evaluates methodologies and tools as well as ongoing and planned activities. New algorithms, resources, technologies and applications for processing text, speech, and multimodalities were strongly encouraged. In addition, papers integrating speech and/or multimodal with written resources were welcomed as well. The papers included expand on the conference papers by at least 45 %, as required for submission to a peer-reviewed journal.

It presents 23 original research articles written by 64 authors representing 11 different countries: Algeria, China, Egypt, Jordan, Malaysia, Morocco, Qatar, Saudi Arabia, Tunisia, United Kingdom, and United States. The special issue is divided into two main parts. Part I is Arabic Natural Language Processing, which is structured in 5 thematic areas ordered as follows: Semantics, Morphology, Part-Of-Speech Tagging, Parsing, and Corpora and Tools Evaluation. Part II is Arabic Automatic Speech Recognition, which is structured in 3 thematic areas ordered as follows: Speech Recognition, Speaker Recognition, and Speech Synthesis.

It is important to mention that the International Journal of Speech Technology (IJST) represented by its Editor-in-Chief, Dr. Amy Neustein accepted our proposal and generously provided us the direction, space, and platform to display the state-of-the-art and contributions for ANLP and AASR. We are indeed very grateful and thankful to Dr. Neustein, IJST, and Springer Editorial Officers who have been very committed to help us produce this special issue in the best manner and shape, because without their efforts this would not have been successfully achieved.

We would also like to thank all authors and reviewers for their commitment, dedication, and contribution to this special issue.