Journal of King Saud University – Computer and Information Sciences 29 (2017) A1–A3

Contents lists available at ScienceDirect

Journal of King Saud University – Computer and Information Sciences

journal homepage: www.sciencedirect.com

Arabic Natural Language Processing: Models, systems and applications

At present, language technologies are instrumental to millions of people, who use them every day with little if any awareness of their existence and role. Popular machine translation systems or web search engines rely more and more on levels of linguistic information automatically overlaid by batch processing tools using language technologies. This development has not only had a huge impact on our daily life, but has also deeply affected the way we think about language as an object of scientific inquiry. The present special issue of JKSU is intended to explore the contribution of computational language models to a better understanding of linguistic, psycholinguistic, sociolinguistic, and literary issues of the Arabic language and culture. The wide array of contributions offered here, ranging from text diacriticisation to psycho-computational modelling of Arabic lexical organisation, bears witness to the maturity of the field and highlight a few general lessons we can learn from current research on Arabic Natural Language Processing.

Computational models of language are, primarily, models of language usage. They focus on those aspects of language performance that are involved in, but are not limited to, language acquisition, lexical access, speech and optical character recognition, text translation, text reading, text understanding, knowledge and ontology extraction, sentiment analysis. What all these tasks have in common is their use of language as a means of conveying information to respond to specific communicative needs and goals. Dealing with language performance ultimately requires bringing language variety and subjectivity to inter-subjective invariance, to a shared representation of its content and structure. From this perspective, emotional undertones, variations in style, speed, pitch, handwriting, topic or dialect are superadded signal complications, which are nonetheless inseparable from language performance. On reflection, in real language-based communication, noise is not simply overlaid on the message, but is actually PART of the message. Upon hearing the voice of a speaker or reading a text, one can get a lot of information about the speaker/writer's own sex, age, bodily features, cultural level, social status, personal attitude, political bias and even ethnicity.

Arabic language processing happens to throw all these performance-related aspects in sharp relief. A first, significant level of variation can already be found in spelling, where underspecified (i.e. non-diacriticized) written words are amenable to a large variety of fully spelled forms. Contrary to many languages, where the process of decoding spelling into sounds typically precedes text understanding, in Arabic there is virtually no reading without prior understanding. This poses new challenges to received language processing architectures for highly resourced languages, where levels of linguistic analysis, from morphology (word-level) to syntax (phrase-level) and semantics, are classically processed in a strictly serial fashion, with one level feeding the ensuing one. In Arabic, this approach is hopelessly inaccurate and inefficient, due to a unique combination of three factors: underspecified spelling, rich inflection and productive derivation. It is thus not surprising that considerable effort in Arabic NLP currently focuses on the proper treatment of morphological processing, and on assessing its impact on further levels of linguistic analysis. The contributions presented in this volume make no exception.

In "AlKhalil Morpho Sys 2: A robust Arabic morpho-syntactic analyser", Mohamed Boudchiche and colleagues show significant improvements in the performance of an existing Arabic word analyser, augmented with a larger and richer morphological database, and a battery of fine-grained morpho-syntactic constraints over word constituents. Improvements involve lexical coverage (percentage of analysed words), accuracy (amount of morpho-syntactic and morpho-lexical information output for each analysed word), and, with some qualifications, also time of execution. Apportioning linguistic information at the right time in the processing assembly line appears to make NLP systems more efficient and robust.

Dialectal variety adds a further dimension of complexity to morpho-syntactic disambiguation of standard modern Arabic. "Morphological Disambiguation of Tunisian Dialect", by Inès Zribi and colleagues, addresses Tunisian part-of-speech (pos) tagging as a bootstrapping task from a severely under-resourced language. A spoken corpus is first transcribed and sentence boundaries are detected automatically. The resulting text is then analysed by a general-purpose standard Arabic morpho-syntactic analyser. Finally, a propositional rule-based learner is trained to single out, for each word, one contextually relevant pos out of the whole range of solutions output at the previous step. Results are compared with those obtained by two other machine-learning classifiers, based on decision trees and support vector machines respectively. Overall performance is extremely encouraging, and provides a significant jumping-off point for further improvements on both Tunisian and other Arabic dialects.

In "Morphological, syntactic and diacritics rules for automatic diacritization of Arabic sentences", by Amine Chennoufi and Azzeddine Mazroui, the AlKhalil morphological analyser is used as the basis for full vowelization of a written text. The output of AlKhalil undergoes further hybrid processing steps, whereby rule-based constraints are merged with a statistical Markov model.

Peer review under responsibility of King Saud University.

Production and hosting by Elsevier

http://dx.doi.org/10.1016/j.jksuci.2017.04.004

The reported accuracy improvement over other comparable systems not only defines a new benchmark for Arabic NLP technology, but also provides an algorithmic assessment of the complexity of the task of reading Arabic non-diacriticized texts.

Assessment of technological improvements in system development requires consensual input/output representations, system comparability and de facto benchmarks. This is a lively research issue in the NLP community, and a precondition for progress in any technological area. Developers of Arabic tools are increasingly aware of the need for shared benchmark resources, standard representation formats, and established evaluation practices. In "Enhancing Arabic stemming process using resources and benchmarking tools", Younes Jaafar and colleagues deal with this issue by making interesting suggestions on the evaluation of morphological analysers. Their evaluation approach allows a balanced assessment of system accuracy and efficiency, as a function of the specific application the system is intended to serve. Their proposal establishes rigorous practices that will hopefully be followed in other domains as well as other languages.

Hierarchical levels of specification of pos categories are focused on in "Towards a standard part of speech tagset for the Arabic language", where Imad Zeroual and colleagues propose a new categorical ontology for Arabic Natural Language Processing, and investigate its implications both in principle and experimentally. The main point made by the authors is that a linguistically-sound tagset for Arabic pos tagging is bound to affect the performance of NLP systems. Better categories are, in fact, ultimately easier to process in context. Although we agree that language-specific issues can possibly call for substantive rethinking of traditional linguistic categories, and that different tagsets may ultimately underpin a different way of carving out language generalizations, more effort should be put into an attempt to discover underlying universal principles for pos tagging.

In the end, the idea that Arabic NLP requires a parallel processing architecture, where more levels of information can be invoked and made interact as early as possible, is in line with recent acquisitions in the neuro-biology of language, blurring the traditional divide between lexical resources and grammar rules. A Temporal Self-organising Map, capable of learning a complex inflectional system like Arabic conjugation, is illustrated in "Arabic word processing and morphology induction through adaptive memory self-organisation strategies" by Claudia Marzi and colleagues. Based on principles of Hebbian learning, the network memorises Arabic word forms as symbolic time series, and organises them as a function of gradient levels of morphological redundancy. After training, the map exhibits processing sensitivity to word frequency and structure regularity, and shows non-concatenative morphological effects as patterns of co-activation between paradigmatically-related, fully-stored forms.

Text classification consists in assigning each item in a document repository a thematic category (e.g., Health, Economy or Education) from a predefined set. The task represents an important preliminary step in the process of indexing and accessing documents by their content. In "Toward an enhanced Arabic text classification using cosine similarity and Latent Semantic Indexing", Fawaz S. Al-Anzi and Dia AbuZeina explore the use of Latent Semantic Indexing, a technique for representing the meaning of words as a frequency distribution of the content words the target word typically co-occurs with in real contexts, for text classification. Instead of grouping documents that contain the same words, the algorithm cluster together documents that tend to contain semantically similar words. Authors show the effectiveness of relying on "latent" semantic information, coupled with a simple document similarity metric such as document cosine distance.

Clustering web search results (snippets) in meaningful thematic classes also facilitates browsing the web, as it provides a principled way for human users to tell relevant results from noise. In "Formal Concept Analysis for Arabic Web Search Results Clustering", Issam Sahmoudi and Abdelmonaime Lachkar illustrate the use of Formal Concept Analysis as a flexible, context-sensitive algorithm for snippet clustering. The algorithm is based on the intuition that snippets can be grouped into natural classes by looking at the number of "concepts" (nominal roots) they happen to share. The most relevant roots are used as concept labels for clusters, and clusters are arranged hierarchically, with each subconcept in the hierarchy containing a subset of the snippets in the concept above it. The final result looks like a text classification task, with the important difference that, here, class nodes and relations may vary as a function of the query words and retrieved snippets.

The relevance of conceptual information for text understanding and information extraction raises the problem of acquiring this information in the first place. Term extraction and semantic tagging are preliminary steps on the way to building an ontology fully automatically. "Arabic medical entity tagging using distant learning in a Multilingual Framework", by Viviana Cotik and colleagues, illustrates an interesting application of machine learning-driven medical term tagging, based on multilingual encyclopaedic resources and massive training. An under-resourced language like Arabic is shown to benefit considerably from online, mutually linked multilingual terminological resources, which are becoming increasingly available for languages like English or Spanish.

A pattern-based approach to legal ontology acquisition is described in "Deriving ontological semantic relations between Arabic compound nouns concepts", by Imen Bouaziz Mezghanni and Faiez Gargouri. The paper is focused on complex nominal constructions (named "compounds") that are found in the legal domain (e.g. "competent civil court" of "judicial police officer"). The approach is based on an analysis of the internal linguistic structure of complex nominals, and their position in legal texts (e.g. the specific articles where they happen to occur). The authors adopt a relational variant of Formal Concept Analysis, whereby taxonomic (ISA) relations between complex nominals are augmented with the transversal semantic relations holding between their noun constituents (e.g. a *police officer* is an *officer* that *belongs* to the *police*). Representing acquired knowledge through formal augmented taxonomies offers the benefit of flexibly querying this knowledge for legal text retrieval.

Last but not least, ontologies can also be instrumental for understanding people attitudes and inclinations, as illustrated in "Semantic Sentiment Analysis in Arabic Social Media", by Samir Tartir and Ibrahim Abdul-Nabi. Here, a sentiment ontology developed for the Jordan variant of modern Arabic classifies words by the feelings they express, thus offering a key to exploring subjective inclinations conveyed in both modern standard Arabic and dialects through social media.

Over the past few decades, computational models of language processing have considerably changed the way we conceive of verbal communication as an object of scientific inquiry. The XX century focus on language competence as a combinatorial system of meaningful building blocks, whose formal properties are studied independently of their use in real communication contexts, has been giving way to a mounting awareness that language is about conveying information. Little can be understood about language when, to use Wittgenstein's phrase, "language goes on holiday". This special issue on Arabic Natural Language Processing reminds us that there is a huge amount of available, unstructured information conveyed by language, and that this information is still, for most languages, waiting to be mined and made openly accessible through shared representation structures. We do not know how long it will take to circumvent the considerable challenges connected with language and dialect specificities/idiosyncrasies, and with the highly subjective and goal-oriented nature of the informa-

tion content conveyed by speakers/writers. What we know is that any attempt to cast language content into shared, accessible representations, will take a highly interdisciplinary effort, where each specialist can benefit from the insights of other scientists from more or less neighbouring disciplinary domains. From this perspective, the future of Natural Language Processing will very much depend on its capacity to function as a truly interface domain, fostering convergence between electrical and electronic engineering, computer sciences, human neuro-physiology, psychology, cognitive science and linguistics.

Acknowledgements

The original impulse of the present volume comes from the 1st International Workshop on Arabic Natural Language Processing, convened in Tetouan (Morocco) in Fall 2014, within the 3rd International IEEE Colloquium on Information Science and Technology, cosponsored by the IEEE Morocco Section and the IEEE Morocco Computer & Communication Joint Chapter. Thanks are due to the Steering, Organizing and Program Committees of the Colloquium for hosting the workshop, and in particular to Mohammed El Mohajir, Conference General Chair, for his constant support and encouragement. Our gratitude goes to all authors of the selected papers for submitting their work and undergoing a long reviewing process, and to all referees for contributing their valuable feedback and suggestions. We would also like to thank the Editor in Chief of JKSU – Computer and Information Sciences for his stamina, support and firm guidance.

Vito Pirrelli
Institute for Computational Linguistics, National Research Council, Pisa, Italy

Arsalane Zarghili
Faculty of Science and Technology, Sidi Mohamed Ben Abdellah University, Fez, Morocco