Journal of King Saud University – Computer and Information Sciences (2014) **26**, 355–356

King Saud University

Journal of King Saud University – Computer and Information Sciences

www.ksu.edu.sa
www.sciencedirect.com

EDITORIAL

Introduction to the special issue on Arabic NLP: Current state and future challenges

Overview of the articles

Arabic natural language processing (NLP) is still in its initial stage compared to the work in English and other languages. NLP is made possible by the collaboration of many disciplines including computer science, linguistics, mathematics, psychology and artificial intelligence. The results of which is highly beneficial to many applications such as Machine Translation, Information Retrieval, Information Extraction, Text Summarization and Question Answering.

This special issue of the Journal of King Saud University – Computer and Information Sciences (CIS) synthesizes current research in the field of Arabic NLP. A total of 56 submissions was received, 11 of which were finally accepted for this special issue. Each accepted paper has gone through three rounds of reviews, each round with two to three reviewers. The content of this special issue covers different topics such as: Dialectal Arabic Morphology, Arabic Corpus, Transliteration, Annotation, Discourse Relations, Sentiment Lexicon, Arabic named entities, Arabic Treebank, Text Summarization, Ontological Relations and Authorship attribution. The following is a brief summary of each of the main articles in this issue.

"arTenTen: Arabic Corpus and Word Sketches" by Nizar Habash et al. presented a web-crawled corpus of Arabic, gathered in 2012. arTenTen comprises 5.8 billion words. A chunk of it has been lemmatized and POS-tagged with the MADA tools and has then been loaded into Sketch Engine, a leading corpus query tool, where it is open for all to use. We have also created 'word sketches': one-page, automatic, corpus-derived summaries of a word's grammatical and collocational behaviour. We show, with examples, what the corpus can show us about Arabic words and phrases, and how this can support lexicography and inform linguistic research. The paper also presents the 'sketch grammar' (the basis for the word sketches) in some detail; describes the process of building and processing the corpus; and considers the role of the corpus in further research on Arabic.

"ADAM: Analyzer for Dialectal Arabic Morphology" by Wael Salloum and Nizar Habash presented ADAM (Analyzer for Dialectal Arabic Morphology), a poor man's solution to quickly develop morphological analyzers for dialectal Arabic. ADAM has about half the out-of-vocabulary rate of a state-of-the-art MSA analyzer, and is comparable in its recall performance to an Egyptian dialectal morphological analyzer that took years and expensive resources to build.

"Transliteration Normalization for Information Extraction and Machine Translation" by Yuval Marton and Imed Zitouni identified and clustered name spelling variants using a Statistical Machine Translation method: word alignment. The variants are identified by being aligned to the same "pivot" name in another language (the source-language in Machine Translation settings). Based on word-to-word translation and transliteration probabilities, as well as the string edit distance metric, the target-language names with similar spellings are clustered. Then, they are normalized to a canonical form. With this approach tens of thousands of high precision name transliteration spelling variants are extracted from sentence aligned bilingual corpora in Arabic and English (in both languages). When these normalized name spelling variants are applied to Information Extraction tasks, improvements over strong baseline systems are observed. When applied to Machine Translation tasks, a large improvement potential is shown.

"Arabic Web Pages Clustering and Annotation Using Semantic Class Features" by Alghamdi et al. presented a system to retrieve a machine-understandable data with the help of Web content mining technique to detect covert knowledge within these data. They propose an approach to achieve clustering with semantic similarities. This approach comprises integrating k-means document clustering with semantic features extraction and document vectorisation in order to group the Arabic web pages according to semantic similarities and then show the semantic annotation. The document vectorisation helps to transform text documents into semantic class probability distribution or semantic class density. In order to reach semantic similarities, the approach extracts the semantic class features and integrates them into the similarity weighting schema. The quality of the clustering result was evaluated

Peer review under responsibility of King Saud University.

ELSEVIER Production and hosting by Elsevier

http://dx.doi.org/10.1016/j.jksuci.2014.10.001

using purity and mean intra-cluster distance (MICD) evaluation measures. They also evaluated the proposed approach on a set of common Arabic news web pages.

''Learning Explicit and Implicit Arabic Discourse Relations'' by KESKES et al. proposed a supervised learning approach to identify discourse relations in Arabic texts. They used the Discourse Arabic Treebank corpus (D-ATB) which is composed of newspaper documents extracted from the syntactically annotated Arabic Treebank v3.2 part3 where each document is associated with complete discourse graph according to the cognitive principles of Segmented Discourse Representation Theory (SDRT). Their list of discourse relations is composed of a three-level hierarchy of 24 relations grouped into 4 top-level classes. To automatically learn them, they used state of the art features whose efficiency has been empirically proved. They investigated how each feature contributes to the learning process. Then they reported their experiments on identifying fine-grained discourse relations, mid-level classes and also top-level classes. Finally, they compared their approach with three baselines that are based on the most frequent relation, discourse connectives and the features used by Al-Saif and Markert. The obtained results were very encouraging and outperform all the baselines with an *F*-score of 78.1% and an accuracy of 80.6%.

''Building an Arabic Sentiment Lexicon Using Semi-Supervised Learning'' by Mahyoub et al. presented an Arabic Sentiment Lexicon that assigns sentiment scores to the words found in Arabic WordNet. Starting from a small seed list of positive and negative words, they used semi-supervised learning to propagate the scores on Arabic WordNet by exploiting the synset relations. Their algorithm assigned a positive sentiment score to more than 800, a negative score to more than 600 and a neutral score to more than 6000 words in the Arabic WordNet. The lexicon was evaluated by incorporating it into a machine learning based classifier. The experiments were conducted on several Arabic sentiment corpora and achieved 96% classification accuracy.

''A hybrid method for extracting relations between Arabic named entities'' by Boujelben et al. presented a hybrid approach to extract relation between Arabic named entities. Given Arabic is a rich morphological language, they built linguistic and learning model to predict the position of word that express a semantic relation within a clause. The main idea is to employ linguistic modules in order to ameliorate the results obtained from machine learning based method. Their empirical results indicate that the hybrid approach outperforms both rule-based and ML-based approaches when applied to AnerCorp corpus.

''An Arabic CCG Approach for Determining Constituent Types from Arabic Treebank'' by El-taher et al. described the required preprocessing step of the Treebank and how to determine Arabic constituents' types. They conducted an experiment on parts 1 and 2 of the Penn Arabic Treebank aiming at converting the Treebank into an Arabic CCGbank. The performance of our algorithm when applied to ATB1v2.0 and ATB2v2.0 has achieved 99% identification of head nodes' and 100% coverage over the Treebank data.

"Minimum Redundancy and Maximum Relevance for Single and multidocument Arabic Text Summarization'' by Oufaida et al. proposed a novel statistical summarization system for Arabic texts. Their system used a clustering algorithm and an adapted discriminant analysis method: mRMR(minimum Redundancy and Maximum Relevance) to score terms. Through mRMR analysis, terms are ranked according to their discriminant and coverage power. Second, they proposed a novel sentence extraction algorithm which selects sentences with top ranked terms and maximum diversity. Their system used minimal language-dependant processing: sentence splitting, tokenization and root extraction. Experimental results on EASC and TAC 2011 MultiLingual datasets showed that their proposed approach is competitive to the state of the art systems.

"Automatic Extraction of Ontological Relations from Arabic Text'' by Mohammed AlZamil and Qasem Al-Radaideh proposed a methodology that extracts ontological relationships from Arabic text. The goals of their research were: to extract semantic features of Arabic text, propose syntactic patterns of relationships among concepts, and propose a formal model of extracting ontological relations. The proposed methodology has been designed to analyze the Arabic text using lexical semantic patterns of Arabic language according to a set of features. Next, the features have been abstracted and enriched with formal description for the purpose of generalizing the resulted rules. The rules, then formulated a classifier that accepts Arabic text, analyzes it, and then displays related concepts labeled with its designated relationship. Moreover, to resolve ambiguity of having homonyms, a set of Machine Translation, text mining, and part of speech tagging algorithms have been reused. They performed extensive experiments to measure the effectiveness of the proposed tools. The results indicated that their proposed methodology is promising for automating the process of extracting ontological relations.

Finally, ''Naïve Bayes classifiers for authorship attribution of Arabic texts'' by Alaa Altheneyan and Mohamed El Bachir Menai studied the use of naïve Bayes classifiers for Arabic authorship attribution, taking into account different event models, namely simple Naïve Bayes (NB), Multinomial Naïve Bayes (MNB), Multi-variant Bernoulli Naïve Bayes (MBNB) and Multi-variant Poisson Naïve Bayes (MPNB). They evaluated their performance on a large Arabic dataset extracted from books of 10 different authors and compared them with other existing methods. The experimental results show that MBNB provides the best results and was able to attribute the author of a text with an accuracy of 97.43%. Comparison results with related methods show that MBNB and MNB are appropriate for authorship attribution.

Acknowledgments

This special issue would not have been possible without the contributions of many people. We wish to thank our elite guest editors: Dr. Eric Atwell from Language research group, I-AIBS Institute for Artificial Intelligence and Biological Systems, School of Computing, Faculty of Engineering, UNIVERSITY OF LEEDS. Prof. Khaled Shaalan from the Faculty of Computers & Information, Cairo Univ. (on Secondment to The British University in Dubai), and Dr. Imed Zitouni – PhD, Principal Researcher at Microsoft, Member of the Relevance and Measurement team of Microsoft. We also extend our gratitude to the authors for submitting their work to this special issue and all referees for their expertise and dedication in providing valuable feedback and suggestions.

Hend S. Al-Khalifa
Information Technology Department, College of Computer and Information Sciences, King Saud University, Saudi Arabia