NLP for Arabic and Related Languages

Mona Diab* — **Nizar Habash**** — **Imed Zitouni*****

* *The George Washington University, USA*
mtdiab@gwu.edu
** *New York University Abu Dhabi, UAE*
nizar.habash@nyu.edu
*** *Microsoft Research, USA*
izitouni@microsoft.com

1. On Arabic and Natural Language Processing

Arabic natural language processing (NLP) is a challenging field of research. This is due to many factors including Arabic's complex and rich morphology, its high degree of ambiguity as well as the presence of a number of dialects that vary quite widely. Furthermore, Arabic has many important geopolitical connections and is spoken by over 400 million people in countries with varying degrees of prosperity and stability. Arabic in its standard form, known as Modern Standard Arabic (MSA), is one of the 6 official languages of the United Nations. It is the primary language of the latest world refugee problem affecting the Middle East and Europe. The opportunities that are made possible by working on this language and its dialects cannot be underestimated in their consequence on the Arab World, the Mediterranean Region and the rest of the world.

Apart from its geopolitical significance, Arabic poses interesting challenges to NLP in general. Given its complexity when considered with its dialects, the language pushes the boundaries of NLP as it forces researchers to think of creative solutions posed by the inherent nature of the language. The use of Arabic is diglossic, the standard language used in formal settings and in education is significantly different from the spoken vernaculars. The spoken vernaculars are quite diverse depending on whether they are spoken in cities or rural areas, settled communities or Bedouin environments. The variational dimensions are quite pronounced. We find social variations such as educated versus lay, male vs. female, urban vs. rural, re-

TAL. Volume 58 – n°3/2017, pages 9 à 13

ligious variants. The differences are not only in the accent but actually dialectal variants that are reflected in the lexical choice, morpho/phonological variations. To add to the complexity of the linguistic situation, Arabic spoken and written modalities typically code switch within utterance. In the written modality, we see pervasive code switching between the vernaculars and MSA; in the spoken variety, we observe rampant code switching with French, English, Italian and Spanish depending on the country. Moreover, in the written modality, even for MSA, Arabic is underspecified for short vowels leading to even more ambiguity over and above natural lexical/syntactic/phonological/semantic/pragmatic ambiguity present in natural language.

Before the onslaught of social media, Arabic in the digital world was perceived as strictly MSA with potential contrast to Classical Arabic (the language pertaining to old historical books dating back to the 6th-19th century). However with the ubiquity of current technology from SMS to chat rooms in the context of social media, we note the pervasive presence of written/spoken dialectal. The challenge with processing such resource is manyfold. Mainly, we have no written standard orthographies for these dialectal varieties that not only vary across Arab countries but actually within the same country along geographical and social continua. Accordingly, the nature of such linguistic expressions is quite low resource by definition. Therefore building resources and solutions that adopt domain adaptation techniques by default is necessary, especially if we need to scale solutions beyond one variety of Arabic to cater to all Arabics. Moreover robust solutions devised for Arabic processing could serve as solutions for other languages with multiple varieties living side by side such as Indonesian and Malay. Given the importance of Arabic, there has been a lot of progress in the last fifteen years in the area of Arabic NLP. This special edition of *TAL* intends to provide a forum for researchers to share and discuss their work.

2. Summaries of Articles

There are four contributions in this special issue. All of the articles have a common focus on computational semantics, although with different approaches and tasks. The first article presents a survey on Arabic sentiment analysis, a very popular area of research that has grown very fast in the last few years. The second article is a specific Arabic sentiment analysis study targeting the Algerian dialect of Arabic. The third article approaches computational semantics for Arabic from a Frame Semantics point of view, and describes the challenges of building an Arabic version of FrameNet. Last but not least, the fourth article presents a framework for information extraction in Arabic that supports Arabic's complex and rich morphology.

2.1. *Modern Trends in Arabic Sentiment Analysis: A Survey (Mulki, Haddad and Babaoğlu)*

The growth of Arabic textual content on social media platforms, together with the continuous crises in the Arab World, have evoked the need to analyze the opinions of

the public regarding ongoing events. As a result, Arabic Sentiment Analysis (ASA) has become the focus of many recent NLP studies. With several Arabic NLP resources being publicly available along with the emergence of deep learning techniques, researchers could handle the complex nature of Arabic language more efficiently. In the last decade, various ASA systems have been built. Yet, their achievements have not been investigated or compared against each other. This survey covers the ASA research carried out during the past five years. The survey compares and evaluates the performances and gives insight into the ability of the created resources to support future ASA research.

2.2. *Une approche fondée sur les lexiques d'analyse de sentiments du dialecte algérien (Guellil, Azouaou, Saâdane and Semmar)*

The paper presents a tool for sentiment analysis of the Algerian dialect, combining the use of lexicons of sentiments and the processing of agglutination. The proposed approach starts by building a lexicon of Algerian-dialect sentiments leveraging an English lexicon. Then, a morphological analysis for sentiment analysis is conducted, where the valence and intensity of the sentiment in the input text is defined. This is different from most of sentiment analysis tools that are currently available since they are capable of processing only MSA. Experiments are conducted using two sentiment lexicons and a test corpus of 700 messages. Results show improvement of performance at each processing step.

2.3. *L'analyse et l'annotation à base de FrameNet: contribution à l'étude contrastive des événements de mouvement en arabe et en anglais (Lakhfif and Laskri)*

The paper describes a computational approach based on Frame Semantics for Arabic language processing. This approach is based on the adaptability of Berkeley FrameNet database and the transferability of FrameNet tools for Arabic, a language that differ typologically from English. The paper describes an attempt to build an equivalent Arabic FrameNet where it shows the use of such a semantic resource for Arabic text semantic analysis, representation and annotation. A frame based contrastive study of motion-events expressions in bilingual text (English-Arabic) is presented, using FrameNet based tool for semantic annotation. Results conducted on a corpus of motion events expressions confirm the cross-linguistic nature of Frame Semantics approach and the suitability of the theory for Arabic processing.

2.4. *Morphology-based Entity and Relational Entity Extraction Framework for Arabic (Jaber and Zaraket)*

Rule-based techniques and tools to extract entities and relational entities from documents allow users to specify desired entities using natural language questions, finite state automata, regular expressions, structured query language statements, or proprietary scripts. These techniques and tools require expertise in linguistics and programming. They lack support of Arabic morphological analysis which is key to process Arabic text. This work presents MERF, a morphology-based entity and relational entity extraction framework for Arabic text. MERF provides a friendly interface where the user, with basic knowledge of linguistic features and regular expressions, defines tag types and interactively associates them with regular expressions defined over Boolean formulae. Boolean formulae range over matches of Arabic morphological features, and synonymity features. Users define relations with tuples of subexpression matches and can associate code actions with subexpressions. MERF computes feature matches, regular expression matches, and constructs entities and relational entities from the user-defined relations. MERF is evaluated with several case studies and compared with existing application-specific techniques. The results show that MERF requires shorter development time and effort compared to existing techniques and produces reasonably accurate results within a reasonable overhead in run time.

Acknowledgments

We want to thank the *TAL journal* editors and committee as well as the specific scientific committee. We are particularly grateful to the reviewers for their time and effort to improve this special issue.

We dedicate this issue to the memory of Isabelle Tellier.

Specific scientific committee (by alphabetical order): Tiba Zaki Abdulhameed (Western Michingan University, USA); Muhammad Abdulmageed (University of British Columbia, Canada); Motasem Alrahabi (University of Sorbonne Abu Dhabi, UAE); Mohammad Attia (Google Inc., USA); Eric Atwell (Leeds University, UK); Yassine Benajiba (Symanto Group, USA); Houda Bouamor (Carnegie Mellon University in Qatar, Qatar); Tim Buckwalter (University of Maryland, USA); Violetta Cavalli-Sforza (Al Akhawain University, Morocco); Khalid Choukri (ELRA , France); Mona Diab (The George Washington University, USA); Joseph Dichy (University Lyon 2, France); Heba Elfardy (Columbia University, USA); Yannick Estève (University of Le Mans, France); Mahmoud Ghoneim (The George Washington University, USA); Nizar Habash (New York University Abu Dhabi, UAE); Bassam Haddad (University of Petra, Jordan); Kais Haddar (University of Sfax, Tunisia); Lamia Hadrich-Belguith (University of Sfax, Tunisia); Hazem Hajj (American University of Beirut, Lebanon); Denis Jouvet (INRIA, France); Omar Larouk (ENSSIB Lyon, France); Mohsen Rashwan (Research and Development International (RDI), Egypt); Khaled Shaalan (British University of Dubai, UAE); Khalil Sima'an (University of

Amsterdam, Netherlands); Kamel Smaïli (University of Lorraine, France); Abdelhadi Soudi (École Nationale Supérieure des Mines, Morocco); Nadi Tomeh (University of Paris 13, France); Stephan Vogel (Qatar Computing Research Institute, Qatar); Wajdi Zaghouani (Carnegie Mellon University in Qatar, Qatar); Aya Zirikly (The George Washington University, USA); and Imed Zitouni (Microsoft, USA).