ANLP 2014

The EMNLP 2014 Workshop on Arabic Natural Language Processing

Proceedings of the Workshop

October 25, 2014
Doha, Qatar

©2014 The Association for Computational Linguistics

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
209 N. Eighth Street
Stroudsburg, PA 18360
USA
Tel: +1-570-476-8006
Fax: +1-570-476-0860
acl@aclweb.org

ISBN 978-1-937284-96-1

Introduction

Welcome to Arabic Natural Language Processing Workshop at EMNLP 2014 in Doha, Qatar.

There has been a lot of progress in the last 15 years in the area of Arabic Natural Language Processing (NLP). In particular, the TIDES, GALE, and BOLT programs provided a significant boost to Arabic NLP, both in generating new language and speech resources on a large scale, and in advancing the state-of-the-art in morphological processing, parsing, named entity recognition, information retrieval, speech recognition, and machine translation. The substantial investment done through these projects reflect the fact that the Middle East continues to grow in its political and economic importance. We also observe that countries in the Middle East invest substantially into higher education and into building an ecosystem, which fosters new research initiatives. This creates the hope that our own research field, NLP, and especially Arabic NLP will continue to grow, will continue to attract students both in the region and in top international universities, and that new job opportunities will open up not only in the well established language service providers, but also through start-ups offering innovative solutions.

A number of Arabic NLP (or Arabic NLP-related) workshops and conferences have taken place, both in the Arab World and in association with international conferences. The Arabic NLP workshop at EMNLP 2014 follows in the footsteps of these previous efforts to provide a forum for researchers to share and discuss their ongoing work. The Arabic NLP workshop also includes a shared task on Automatic Arabic Error Correction, which was designed in the tradition of high profile NLP shared tasks such as CONLL's grammar/error detection and correction shared tasks in 2011-2013, and numerous machine translation campaigns by NIST/WMT/MEDAR, among others. The challenge chosen for the shared task is highly relevant, not only to spelling correction while composing a text, but also to developing techniques for automatically correcting errors in the far-from-perfect outputs of NLP technologies such as speech recognition or machine translation.

We are happy to have received 40 submissions. Unfortunately, not all the papers could be included in the workshop due to time limitations. The acceptance rate is 50%. The papers cover a wide range of topics: building language resources for standard and dialectal Arabic, language identification, sentiment analysis, named entity disambiguation, and machine translation for dialectal Arabic, etc. Twelve papers were selected for oral presentation and were organized under the general topics Corpora (four papers), Text Mining (four papers), Translation & Transliteration (three papers) and one paper describing the shared task. The remaining eight papers were selected to be presented in a poster session. There is no difference in quality between the oral and poster presentations.

The shared task was a success. We received 18 systems submissions from nine teams in six countries, representing a diverse set of approaches. Nine shared task system description (short) papers are included in the proceedings to document the shared task systems, but were not reviewed with the rest of the papers of the main workshop. These papers will be presented as posters.

The quantity and quality of the contributions to the main workshop, as well as the shared task, are strong indicators that there is a continued need for this kind of dedicated Arabic NLP workshop. We would like to acknowledge all the hard work of the submitting authors and thank the reviewers for their diligent work and for the valuable feedback they provided. We are also thankful to the work of the shared task committee, website committee and the publication co-chairs.

It has been an honor to server as program co-chairs. We hope that the reader of these proceedings will find them stimulating and beneficial.

Nizar Habash and Stephan Vogel, Arabic NLP Workshop, EMNLP 2014.

Organizers:

Program Co-Chairs

Nizar Habash, New York University Abu Dhabi
Stephan Vogel, Qatar Computing Research Institute

Publication Co-chairs

Nadi Tomeh, Université Paris 13, Sorbonne Paris Cité
Houda Bouamor, Carnegie Mellon University Qatar

Website Committee

Noura Farra, Columbia University
Kareem Darwish, Qatar Computing Research Institute

Shared Task Committee

Behrang Mohit (co-chair), Carnegie Mellon University Qatar
Alla Rozovskaya (co-chair), Columbia University
Wajdi Zaghouani, Carnegie Mellon University Qatar
Ossama Obeid, Carnegie Mellon University Qatar
Nizar Habash (advisor), New York University Abu Dhabi

Program Committee:

Abdelmajid Ben-Hamadou, University of Sfax, Tunisia
Abdelhadi Soudi, Ecole Nationale de l'Industrie Minérale, Morocco
Abdelsalam Nwesri, University of Tripoli, Libya
Achraf Chalabi , Microsoft Research, Egypt
Ahmed Ali, Qatar Computing Research Institute, Qatar
Ahmed Rafea, The American University in Cairo, Egypt
Alexis Nasr, University of Marseille, France
Ali Farghaly, Monterey Peninsula College, USA
Almoataz B. Al-Said, Cairo University, Egypt
Alon Lavie, Carnegie Mellon University, USA
Aly Fahmy, Cairo University, Egypt
Azadeh Shakery, University of Tehran, Iran
Azzeddine Mazroui, University Mohamed I, Morocco
Bassam Haddad, University of Petra, Jordan
Bayan Abu Shawar, Arab Open University, Jordan
Behrang Mohit, Carnegie Mellon University Qatar, Qatar
Eric Atwell, University of Leeds, UK
Farhad Oroumchian, University of Wollongong, Australia
Ghassan Mourad, Université Libanaise, Lebanon
Hassan Sawaf, eBay Inc., USA
Hazem Hajj, American University of Beirut, Lebanon
Hend Alkhalifa, King Saud University, Saudi Arabia
Houda Bouamor, Carnegie Mellon University Qatar, Qatar
Imed Zitouni, Microsoft Research, USA

Joseph Dichy, Université Lyon 2, France
Karim Bouzoubaa , Mohammad V University, Morocco
Karine Megerdoomian, The MITRE Corporation, USA
Katrin Kirchhoff, University of Washington, USA
Kemal Oflazer, Carnegie Mellon University Qatar, Qatar
Khaled Shaalan, The British University in Dubai, UAE
Khaled Shaban, Qatar University, Qatar
Khalil Sima'an, Universiteit van Amsterdam, Netherlands
Lamia Hadrich Belguith, University of Sfax, Tunisia
Michael Rosner, University of Malta, Malta
Mohamed Elmahdy, Qatar University, Qatar
Mohsen Rashwan, Cairo University, Egypt
Mona Diab, George Washington University, USA
Mustafa Jarrar, Bir Zeit University, Palestine
Nada Ghneim, Higher Institute for Applied Sciences and Technology, Syria
Nadi Tomeh, Université Paris 13, Sorbonne Paris Cité, France
Ossama Emam, IBM, USA
Otakar Smrž, Charles University , Czech Republic
Owen Rambow, Columbia University, USA
Preslav Nakov, Qatar Computing Research Institute, Qatar
Ramzi Abbes, TECHLIMED, France
Salwa Hamada, Cairo University, Egypt
Shahram Khadivi, Tehran Polytechnic, Iran
Sherri Condon , The MITRE Corporation, USA
Taha Zerrouki, University of Bouira, Algeria
Violetta Cavalli-Sforza, Al Akhawayn University, Morocco

Table of Contents

Conference Program

Saturday, October 25, 2014

Session 1: Corpora

9:00–9:20 *Using Twitter to Collect a Multi-Dialectal Corpus of Arabic*
Hamdy Mubarak and Kareem Darwish

9:20–9:40 *The International Corpus of Arabic: Compilation, Analysis and Evaluation*
Sameh Alansary and Magdy Nagi

9:45–10:05 *Building a Corpus for Palestinian Arabic: a Preliminary Study*
Mustafa Jarrar, Nizar Habash, Diyam Akra and Nasser Zalmout

10:05–10:25 *Annotating corpus data for a quantitative, constructional analysis of motion verbs in Modern Standard Arabic*
Dana Abdulrahim

10:30–11:00 ***Break / Poster setup***

Shared Task

11:00–11:30 *The First QALB Shared Task on Automatic Text Correction for Arabic*
Behrang Mohit, Alla Rozovskaya, Nizar Habash, Wajdi Zaghouani and Ossama Obeid

11:30–11:45 *Shared Task: 1-minute Summary for Shared Task Participants*
Shared Task participants

11:45–12:15 *Shared Task: Panel*
Group Discussion

12:15–12:30 *Main Workshop Poster Teaser 1-minute Summary*
Main Workshop participants

12:30–14:00 ***Lunch / Main and Shared Task Poster Session***

Saturday, October 25, 2014 (continued)

Main and Shared Task Poster Session

12:30–14:00 *Main Workshop Posters*
Main Workshop participants

A Framework for the Classification and Annotation of Multiword Expressions in Dialectal Arabic
Abdelati Hawwari, Mohammed Attia and Mona Diab

Al-Bayan: An Arabic Question Answering System for the Holy Quran
Heba Abdelnasser, Maha Ragab, Reham Mohamed, Alaa Mohamed, Bassant Farouk, Nagwa El-Makky and Marwan Torki

Automatic Arabic diacritics restoration based on deep nets
Ahmad Al Sallab, Mohsen Rashwan, Hazem M. Raafat and Ahmed Rafea

Combining strategies for tagging and parsing Arabic
Maytham Alabbas and Allan Ramsay

Named Entity Recognition System for Dialectal Arabic
Ayah Zirikly and Mona Diab

Semantic Query Expansion for Arabic Information Retrieval
Ashraf Mahgoub, Mohsen Rashwan, Hazem Raafat, Mohamed Zahran and Magda Fayek

Transliteration of Arabizi into Arabic Orthography: Developing a Parallel Annotated Arabizi-Arabic Script SMS/Chat Corpus
Ann Bies, Zhiyi Song, Mohamed Maamouri, Stephen Grimes, Haejoong Lee, Jonathan Wright, Stephanie Strassel, Nizar Habash, Ramy Eskander and Owen Rambow

Tunisian dialect Wordnet creation and enrichment using web resources and other Wordnets
Rihab Bouchlaghem, Aymen Elkhlifi and Rim Faiz

12:30–14:00 *Shared Task Posters*
Shared Task participants

A Pipeline Approach to Supervised Error Correction for the QALB-2014 Shared Task
Nadi Tomeh, Nizar Habash, Ramy Eskander and Joseph Le Roux

Arabic Spelling Correction using Supervised Learning
Youssef Hassan, Mohamed Aly and Amir Atiya

Saturday, October 25, 2014 (continued)

Autocorrection of arabic common errors for large text corpus
Taha Zerrouki, Khaled Alhawiti and Amar Balla

Automatic Correction of Arabic Text: a Cascaded Approach
Hamdy Mubarak and Kareem Darwish

CMUQ@QALB-2014: An SMT-based System for Automatic Arabic Error Correction
Serena Jeblee, Houda Bouamor, Wajdi Zaghouani and Kemal Oflazer

Fast and Robust Arabic Error Correction System
Michael Nawar and Moheb Ragheb

GWU-HASP: Hybrid Arabic Spelling and Punctuation Corrector
Mohammed Attia, Mohamed Al-Badrashiny and Mona Diab

TECHLIMED system description for the Shared Task on Automatic Arabic Error Correction
Djamel MOSTEFA, Omar ASBAYOU and Ramzi ABBES

The Columbia System in the QALB-2014 Shared Task on Arabic Error Correction
Alla Rozovskaya, Nizar Habash, Ramy Eskander, Noura Farra and Wael Salloum

Session 2: Text Mining

14:00–14:20 *A Large Scale Arabic Sentiment Lexicon for Arabic Opinion Mining*
Gilbert Badaro, Ramy Baly, Hazem Hajj, Nizar Habash and Wassim El-Hajj

14:20–14:40 *Evaluating Distant Supervision for Subjectivity and Sentiment Analysis on Arabic Twitter Feeds*
Eshrag Refaee and Verena Rieser

14:45–15:05 *Arabic Native Language Identification*
Shervin Malmasi and Mark Dras

15:05–15:25 *AIDArabic A Named-Entity Disambiguation Framework for Arabic Text*
Mohamed Amir Yosef, Marc Spaniol and Gerhard Weikum

15:30–16:00 ***Break***

Saturday, October 25, 2014 (continued)

Session 3: Translation & Transliteration

16:00–16:20 *Domain and Dialect Adaptation for Machine Translation into Egyptian Arabic*
Serena Jeblee, Weston Feely, Houda Bouamor, Alon Lavie, Nizar Habash and Kemal Oflazer

16:25–16:45 *Unsupervised Word Segmentation Improves Dialectal Arabic to English Machine Translation*
Kamla Al-Mannai, Hassan Sajjad, Alaa Khader, Fahad Al Obaidli, Preslav Nakov and Stephan Vogel

16:50–17:10 *Arabizi Detection and Conversion to Arabic*
Kareem Darwish

Closing Session

17:10–18:00 *Workshop Group Discussion*
Group Discussion