ANLP Workshop 2015

The Second Workshop on Arabic Natural Language Processing

Proceedings of the Workshop

July 30, 2015
Beijing, China

©2015 The Association for Computational Linguistics
and The Asian Federation of Natural Language Processing

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 978-1-941643-58-7

Foreword

Assalamu 3alaykum wa nín hăo! Welcome to the Second Arabic Natural Language Processing Workshop held at ACL 2015 in Beijing, China.

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 ACL 2015 follows in the footsteps of these previous efforts to provide a forum for researchers to share and discuss their ongoing work. As in the first Arabic NLP workshop held at EMNLP 2014 in Doha, Qatar, this workshop 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 numerous machine translation campaigns by NIST/WMT/MEDAR, among others.

We received 23 main workshop submissions and selected 15 (65%) for presentation in the workshop. Nine papers will be presented orally and six as part of a poster session. The presentation mode is independent of of the ranking of the papers. The papers cover a diverse set of topics from designing orthography conventions and annotation tools to speech recognition and deep learning for sentiment analysis.

The shared task was a success with eight teams from six countries participating. The shared task system descriptions (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. A long paper describing the shared task will be presented orally.

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 serve as program co-chairs. We hope that the reader of these proceedings will find them stimulating and beneficial.

Nizar Habash, Stephan Vogel and Kareem Darwish

Organizers:

Program Co-chairs:

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

Publication Co-chairs:

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

Publicity chair:

Wajdi Zaghouani, Carnegie Mellon University Qatar

Shared Task Committee:

Alla Rozovskaya (co-chair), Columbia University
Houda Bouamor (co-chair), Carnegie Mellon University Qatar
Behrang Mohit, Ask.com
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
Abdelsalam Nwesri, University of Tripoli, Libya
Achraf Chalabi , Microsoft Research, Egypt
Ahmed Ali, Qatar Computing Research Institute, Qatar
Ahmed El Kholy, Columbia University, USA
Ahmed Rafea, The American University in Cairo, Egypt
Alberto Barrón Cedeño, Qatar Computing Research Institute, Qatar
Alexis Nasr, University of Marseille, France
Ali Farghaly, Monterey Peninsula College, USA
Almoataz B. Al-Said, Cairo University, Egypt
Aly Fahmy, Cairo University, Egypt
Azzeddine Mazroui, University Mohamed I, Morocco
Bassam Haddad, University of Petra, Jordan
Emad Mohamed, Suez Canal University, Egypt
Fransisco Guzman, Qatar Computing Research Institute, Qatar
Ghassan Mourad, Université Libanaise, Lebanon
Hamdy Mubarak, Qatar Computing Research Institute, Qatar
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

Kareem Darwish, Qatar Computing Research Institute, Qatar
Karim Bouzoubaa , Mohammad V University, Morocco
Kemal Oflazer, Carnegie Mellon University Qatar, Qatar
Khaled Shaalan, The British University in Dubai, UAE
Khaled Shaban, Qatar University, Qatar
Khalid Choukri, ELDA, European Language Resource Association, France
Lamia Hadrich Belguith, University of Sfax, Tunisia
Mohamed Elmahdy, Qatar University, Qatar
Mohamed Maamouri, Linguistic Data Consortium, USA
Mona Diab, George Washington University, USA
Mustafa Jarrar, Bir Zeit University, Palestine
Nada Ghneim, Higher Institute for Applied Sciences and Technology, Syria
Nadi Tomeh, University Paris 13, Sorbonne Paris Cité, France
Nizar Habash, New York University Abu Dhabi, UAE
Otakar Smrž, Džám-e Džam Language Institute, Czech Republic
Owen Rambow, Columbia University, USA
Preslav Nakov, Qatar Computing Research Institute, Qatar
Ramy Eskander, Columbia University, USA
Salwa Hamada, Cairo University, Egypt
Samantha Wray, Qatar Computing Research Institute, Qatar
Shahram Khadivi, Tehran Polytechnic, Iran
Sherri Condon , The MITRE Corporation, USA
Stephan Vogel, Qatar Computing Research Institute, Qatar
Taha Zerrouki, University of Bouira, Algeria
Wael Salloum, Columbia University, USA
Walid Magdy, Qatar Computing Research Institute, Qatar

Table of Contents

Classifying Arab Names Geographically
Hamdy Mubarak and Kareem Darwish 1

Deep Learning Models for Sentiment Analysis in Arabic
Ahmad Al Sallab, Hazem Hajj, Gilbert Badaro, Ramy Baly, Wassim El Hajj and Khaled Bashir Shaban 9

A Light Lexicon-based Mobile Application for Sentiment Mining of Arabic Tweets
Gilbert Badaro, Ramy Baly, Rana Akel, Linda Fayad, Jeffrey Khairallah, Hazem Hajj, Khaled Shaban and Wassim El-Hajj 18

The Second QALB Shared Task on Automatic Text Correction for Arabic
Alla Rozovskaya, Houda Bouamor, Nizar Habash, Wajdi Zaghouani, Ossama Obeid and Behrang Mohit 26

Natural Language Processing for Dialectical Arabic: A Survey
Abdulhadi Shoufan and Sumaya Alameri 36

DIWAN: A Dialectal Word Annotation Tool for Arabic
Faisal Al-Shargi and Owen Rambow 49

POS-tagging of Tunisian Dialect Using Standard Arabic Resources and Tools
Ahmed Hamdi, Alexis Nasr, Nizar Habash and Nuria Gala 59

A Conventional Orthography for Algerian Arabic
Houda Saadane and Nizar Habash 69

A Pilot Study on Arabic Multi-Genre Corpus Diacritization
Houda Bouamor, Wajdi Zaghouani, Mona Diab, Ossama Obeid, Kemal Oflazer, Mahmoud Ghoneim and Abdelati Hawwari 80

Annotating Targets of Opinions in Arabic using Crowdsourcing
Noura Farra, Kathy McKeown and Nizar Habash 89

Best Practices for Crowdsourcing Dialectal Arabic Speech Transcription
Samantha Wray, Hamdy Mubarak and Ahmed Ali 99

Joint Arabic Segmentation and Part-Of-Speech Tagging
Shabib AlGahtani and John McNaught 108

Multi-Reference Evaluation for Dialectal Speech Recognition System: A Study for Egyptian ASR
Ahmed Ali, Walid Magdy and Steve Renals 118

Arib@QALB-2015 Shared Task: A Hybrid Cascade Model for Arabic Spelling Error Detection and Correction
Nouf AlShenaifi, Rehab AlNefie, Maha Al-Yahya and Hend Al-Khalifa 127

CUFE@QALB-2015 Shared Task: Arabic Error Correction System
Michael Nawar 133

GWU-HASP-2015@QALB-2015 Shared Task: Priming Spelling Candidates with Probability
Mohammed Attia, Mohamed Al-Badrashiny and Mona Diab 138

QCMUQ@QALB-2015 Shared Task: Combining Character level MT and Error-tolerant Finite-State Recognition for Arabic Spelling Correction
Houda Bouamor, Hassan Sajjad, Nadir Durrani and Kemal Oflazer 144

QCRI@QALB-2015 Shared Task: Correction of Arabic Text for Native and Non-Native Speakers' Errors
Hamdy Mubarak, Kareem Darwish and Ahmed Abdelali 150

SAHSOH@QALB-2015 Shared Task: A Rule-Based Correction Method of Common Arabic Native and Non-Native Speakers' Errors
Wajdi Zaghouani, Taha Zerrouki and Amar Balla 155

TECHLIMED@QALB-Shared Task 2015: a hybrid Arabic Error Correction System
Djamel MOSTEFA, Jaber ABUALASAL, Omar ASBAYOU, Mahmoud GZAWI and Ramzi Abbès 161

UMMU@QALB-2015 Shared Task: Character and Word level SMT pipeline for Automatic Error Correction of Arabic Text
Fethi Bougares and Houda Bouamor 166

Robust Part-of-speech Tagging of Arabic Text
Hanan Aldarmaki and Mona Diab 173

Answer Selection in Arabic Community Question Answering: A Feature-Rich Approach
Yonatan Belinkov, Alberto Barrón-Cedeño and Hamdy Mubarak 183

EDRAK: Entity-Centric Data Resource for Arabic Knowledge
Mohamed H. Gad-elrab, Mohamed Amir Yosef and Gerhard Weikum 191

Workshop Program

July 30, 2015

09:00–10:00 Main Workshop Papers - Oral Presentations - Session 1

09:00–09:20 *Classifying Arab Names Geographically*
Hamdy Mubarak and Kareem Darwish

09:20–09:40 *Deep Learning Models for Sentiment Analysis in Arabic*
Ahmad Al Sallab, Hazem Hajj, Gilbert Badaro, Ramy Baly, Wassim El Hajj and Khaled Bashir Shaban

09:40–10:00 *A Light Lexicon-based Mobile Application for Sentiment Mining of Arabic Tweets*
Gilbert Badaro, Ramy Baly, Rana Akel, Linda Fayad, Jeffrey Khairallah, Hazem Hajj, Khaled Shaban and Wassim El-Hajj

10:00–10:30 Shared Task Talk

10:00–10:30 *The Second QALB Shared Task on Automatic Text Correction for Arabic*
Alla Rozovskaya, Houda Bouamor, Nizar Habash, Wajdi Zaghouani, Ossama Obeid and Behrang Mohit

10:30–11:00 ***Break***

11:00–12:00 Main Workshop Papers - Oral Presentations - Session 2

11:00–11:20 *Natural Language Processing for Dialectical Arabic: A Survey*
Abdulhadi Shoufan and Sumaya Alameri

11:20–11:40 *DIWAN: A Dialectal Word Annotation Tool for Arabic*
Faisal Al-Shargi and Owen Rambow

11:40–12:00 *POS-tagging of Tunisian Dialect Using Standard Arabic Resources and Tools*
Ahmed Hamdi, Alexis Nasr, Nizar Habash and Nuria Gala

12:00–14:00 ***Lunch + Poster Setup***

July 30, 2015 (continued)

14:00–15:30 Poster Session

\+ ***Posters: Main Workshop Papers***

A Conventional Orthography for Algerian Arabic
Houda Saadane and Nizar Habash

A Pilot Study on Arabic Multi-Genre Corpus Diacritization
Houda Bouamor, Wajdi Zaghouani, Mona Diab, Ossama Obeid, Kemal Oflazer, Mahmoud Ghoneim and Abdelati Hawwari

Annotating Targets of Opinions in Arabic using Crowdsourcing
Noura Farra, Kathy McKeown and Nizar Habash

Best Practices for Crowdsourcing Dialectal Arabic Speech Transcription
Samantha Wray, Hamdy Mubarak and Ahmed Ali

Joint Arabic Segmentation and Part-Of-Speech Tagging
Shabib AlGahtani and John McNaught

Multi-Reference Evaluation for Dialectal Speech Recognition System: A Study for Egyptian ASR
Ahmed Ali, Walid Magdy and Steve Renals

\+ ***Posters: Shared Task Papers***

Arib@QALB-2015 Shared Task: A Hybrid Cascade Model for Arabic Spelling Error Detection and Correction
Nouf AlShenaifi, Rehab AlNefie, Maha Al-Yahya and Hend Al-Khalifa

CUFE@QALB-2015 Shared Task: Arabic Error Correction System
Michael Nawar

GWU-HASP-2015@QALB-2015 Shared Task: Priming Spelling Candidates with Probability
Mohammed Attia, Mohamed Al-Badrashiny and Mona Diab

QCMUQ@QALB-2015 Shared Task: Combining Character level MT and Error-tolerant Finite-State Recognition for Arabic Spelling Correction
Houda Bouamor, Hassan Sajjad, Nadir Durrani and Kemal Oflazer

July 30, 2015 (continued)

QCRI@QALB-2015 Shared Task: Correction of Arabic Text for Native and Non-Native Speakers' Errors
Hamdy Mubarak, Kareem Darwish and Ahmed Abdelali

SAHSOH@QALB-2015 Shared Task: A Rule-Based Correction Method of Common Arabic Native and Non-Native Speakers' Errors
Wajdi Zaghouani, Taha Zerrouki and Amar Balla

TECHLIMED@QALB-Shared Task 2015: a hybrid Arabic Error Correction System
Djamel MOSTEFA, Jaber ABUALASAL, Omar ASBAYOU, Mahmoud GZAWI and Ramzi Abbès

UMMU@QALB-2015 Shared Task: Character and Word level SMT pipeline for Automatic Error Correction of Arabic Text
Fethi Bougares and Houda Bouamor

15:30–16:00 ***Break***

16:00–17:00 **Main Workshop Papers - Oral Presentations - Session 3**

16:00–16:20 *Robust Part-of-speech Tagging of Arabic Text*
Hanan Aldarmaki and Mona Diab

16:20–16:40 *Answer Selection in Arabic Community Question Answering: A Feature-Rich Approach*
Yonatan Belinkov, Alberto Barrón-Cedeño and Hamdy Mubarak

16:40–17:00 *EDRAK: Entity-Centric Data Resource for Arabic Knowledge*
Mohamed H. Gad-elrab, Mohamed Amir Yosef and Gerhard Weikum

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