COLING 2020

The Fifth Arabic Natural Language Processing Workshop WANLP 2020

Proceedings of the Workshop

December 12, 2020 Barcelona, Spain (Online) Copyright of each paper stays with the respective authors (or their employers).

ISBN 978-1-952148-38-5

Preface

Assalamu 3alaykum, e-greetings! Welcome to The Fifth Arabic Natural Language Processing Workshop (WANLP 2020) held jointly with COLING 2020 online.

This workshop aims at providing opportunities for researchers and engineers working on Arabic NLP to share and discuss their ongoing work. Similar to COLING 2020, we will not be traveling to Barcelona, Spain to attend the workshop. The location for this year is different than in the past; WANLP 2020 would be hosted online, a shift due to the unprecedented challenges caused by the COVID-19 virus. We want this workshop to offer a sign of inclusion for everyone around the globe to come share their knowledge and learn from each other, in a safe and welcoming environment. We will be attending online from anywhere in the world, including my living room. This is our first time having WANLP as a virtual workshop. This workshop is the fifth in a series, following the First workshop on Arabic NLP held at EMNLP 2014 in Doha, Qatar; the Second workshop on Arabic NLP held at ACL 2015 in Beijing, China; the Third workshop on Arabic NLP held at EACL 2017 in Valencia, Spain; and the Fourth workshop on Arabic NLP held at ACL 2019 in Florence, Italy.

We received 44 main workshop submissions, out of which 20 were accepted. Papers submitted to the main workshop were reviewed by at least three reviewers. This workshop also includes a shared task on Nuanced Arabic Dialect (NADI). NADI is the first shared task to target naturally-occurring fine-grained dialectal text at the sub-country level, as opposed to previous shared tasks that focused on regional level dialect labeling. The shared task is a success with 18 teams participating. The shared task system descriptions papers were reviewed by two reviewers each. They are included in the proceedings and will be presented during the workshop. A long paper describing the shared task is accepted as part of the main workshop. The selection process was competitive and we believe it resulted in a balanced and diverse program that is appealing to the audience. We observed a high quality of contributions that resulted in an acceptance rate of 45% in the main workshop. The higher number of submissions to the main workshop relative to previous instances and the success of the shared task are strong indicators that provide compelling evidence for the continued need for a dedicated Arabic NLP workshop.

We would like to thank everyone who submitted a paper to the workshop. We would also like to express our gratitude to the members of the Program Committee, who worked very hard to provide reviews supporting a tight schedule. Organizing WANLP 2020 is a team effort.

Imed Zitouni, General Chair, on behalf of the workshop organizers.

Website of the workshop: http://wanlp2020.arabic-nlp.net/

Workshop Organizers

General Chair:

Imed Zitouni, Google, USA

Program Chairs:

Muhammad Abdul-Mageed, UBC, Canada Houda Bouamor, Carnegie Mellon University in Qatar Fethi Bougares, University of Le Mans, France Mahmoud El-Haj, Lancaster University, England

Publication Chair:

Nadi Tomeh, LIPN, Université Sorbonne Paris Nord

Publicity Chair:

Wajdi Zaghouani, Hamad Bin Khalifa University, Qatar

Advisory Committee:

Muhammad Abdul-Mageed, UBC, Canada Ahmed Ali, Qatar Computing Research Institute, Qatar Hend Alkhalifa, King Saud University, Saudi Arabia Houda Bouamor, Carnegie Mellon University in Qatar Fethi Bougares, Le Mans University, France Khalid Choukri, ELDA, European Language Resource Association, France Kareem Darwish, Qatar Computing Research Institute, Qatar Mona Diab, George Washington University, USA Mahmoud El-Haj, Lancaster University, UK Samhaa El-Beltagy, Nile University, Egypt Wassim El-Hajj, American University of Beirut, Lebanon Nizar Habash, New York University Abu Dhabi, UAE Lamia Hadrich Belguith, University of Sfax, Tunisia Hazem Hajj, American University of Beirut, Lebanon Walid Magdy, University of Edinburgh, Scotland Khaled Shaalan, The British University in Dubai, UAE Kamel Smaili, University of Lorraine, France Nadi Tomeh, Université Sorbonne Paris Nord, France Wajdi Zaghouani, Hamad Bin Khalifa University, Qatar Imed Zitouni, Google, USA

Program Committee:

Mourad Abbas, CRSTDLA, Algeria Ahmed Abdelali, Qatar Computing Research Institute, HBKU, Qatar Muhammad Abdul-Mageed, The University of British Columbia, Canada Bayan Abu Shawar, Al Ain University, UAE Wafia Adouane, University of Gothenburg, Sweden Haithem Afli, Cork Institute of Technology, Ireland Hussein Al-Natsheh, Mawdoo3 Limited, Jordan Almoataz Al-Said, Cairo University, Egypt Bashar Alhafni, New York University Abu Dhabi, UAE Ahmed Ali, Qatar Computing Research Institute, HBKU, Qatar Hend Alkhalifa, King Saud University, KSA Chafik Aloulou, Univeristé de Sfax, Tunisia Areeb Alowisheq, Imam University, KSA Mohammed Attia, George Washington University Gilbert Badaro, American University of Beirut, Lebanon Riadh Belkebir, New York University Abu Dhabi, UAE Houda Bouamor, Carnegie Mellon University in Qatar Karim Bouzoubaa, Mohammad V University, Morocco Shammur Chowdhury, Qatar Computing Research Institute, HBKU, Qatar Kareem Darwish, Qatar Computing Research Institute, HBKU, Qatar Mahmoud El Haj, Lancaster University, UK Wassim El-Hajj, American University of Beirut, Lebanon Shady Elbassuoni, American University of Beirut, Lebanon Mariem Ellouze, University of Sfax, Tunisia Tamer Elsayed, Qatar University, Qatar Sahar Ghannay, LIUM Laboratory, France Nada Ghneim, Higher Institute for Applied Sciences and Technology, Syria Nizar Habash, New York University Abu Dhabi, UAE Bassam Haddad, University of Petra, Jordan Lamia Hadrich Belguith, University of Sfax, Tunisia Hazem Hajj, American University of Beirut, Lebanon Salima Harrat, École Normale Supérieure (Bouzaréah), Algeria Maram Hasanain, Qatar University, Qatar Go Inoue, New York University Abu Dhabi, UAE Mustafa Jarrar, Bir Zeit University, Palestine Ganesh Jawahar, The University of British Columbia, Canada Salam Khalifa, New York University Abu Dhabi, UAE Walid Magdy, University of Edinburgh, Scotland Azzeddine Mazroui, University Mohamed I, Morocco Seif Mechti, University of Sfax, Tunisia Salima Medhaffar, Le Mans University, France Karima Meftouh, Badji Mokhtar University, Algeria Hamdy Mubarak, Qatar Computing Research Institute, HBKU, Qatar El Moatez Billah Nagoudi, The University of British Columbia, Canada Preslav Nakov, Qatar Computing Research Institute, HBKU, Qatar Alexis Nasr, University of Marseille, France Joshi Praveen, Cork Institute of Technology, Ireland Younes Samih, Heinrich Heine Universität Düsseldorf, Germany Khaled Shaalan, The British University in Dubai, UAE Khaled Shaban, Qatar University, Qatar Peter Sullivan Sullivan, The University of British Columbia, Canada Reem Suwaileh, Qatar University, Qatar. Nadi Tomeh, Université Sorbonne Paris Nord, France Omar Trigui, University of Sousse, Tunisia

Wajdi Zaghouani, Hamad Bin Khalifa University, Qatar Nasser Zalmout, Amazon Inc., USA Taha Zerrouki, University of Bouira, Algeria Chiyu Zhang, The University of British Columbia, Canada

Invited Speaker:

Perslav Nakov, Qatar Computing Research Institute, HBKU, Doha, Qatar

Table of Contents

<i>German-Arabic Speech-to-Speech Translation for Psychiatric Diagnosis</i> Juan Hussain, Mohammed Mediani, Moritz Behr, M. Amin Cheragui, Sebastian Stüker and Alexan- der Waibel
Hate Speech Detection in Saudi Twittersphere: A Deep Learning Approach Raghad Alshaalan and Hend Al-Khalifa
Is it Great or Terrible? Preserving Sentiment in Neural Machine Translation of Arabic Reviews Hadeel Saadany and Constantin Orasan
Deep Diacritization: Efficient Hierarchical Recurrence for Improved Arabic DiacritizationBadr AlKhamissi, Muhammad ElNokrashy and Mohamed Gabr
A Semi-Supervised BERT Approach for Arabic Named Entity Recognition Chadi Helwe, Ghassan Dib, Mohsen Shamas and Shady Elbassuoni
<i>Empathy-driven Arabic Conversational Chatbot</i> Tarek Naous, Christian Hokayem and Hazem Hajj
 Machine Generation and Detection of Arabic Manipulated and Fake News El Moatez Billah Nagoudi, AbdelRahim Elmadany, Muhammad Abdul-Mageed and Tariq Alhindi 69
<i>Transliteration of Judeo-Arabic Texts into Arabic Script Using Recurrent Neural Networks</i> Ori Terner, Kfir Bar and Nachum Dershowitz
NADI 2020: The First Nuanced Arabic Dialect Identification Shared Task Muhammad Abdul-Mageed, Chiyu Zhang, Houda Bouamor and Nizar Habash
Multi-dialect Arabic BERT for Country-level Dialect IdentificationBashar Talafha, Mohammad Ali, Muhy Eddin Za'ter, Haitham Seelawi, Ibraheem Tuffaha, MostafaSamir, Wael Farhan and Hussein Al-Natsheh
On the Importance of Tokenization in Arabic Embedding Models Mohamed Alkaoud and Mairaj Syed119
Tracing Traditions: Automatic Extraction of Isnads from Classical Arabic Texts Ryan Muther and David Smith
<i>Embed More Ignore Less (EMIL): Exploiting Enriched Representations for Arabic NLP</i> ahmed younes and Julie Weeds
MANorm: A Normalization Dictionary for Moroccan Arabic Dialect Written in Latin Script Randa Zarnoufi, Hamid Jaafar, Walid Bachri and Mounia Abik
A Unified Model for Arabizi Detection and Transliteration using Sequence-to-Sequence Models Ali Shazal, Aiza Usman and Nizar Habash
Multi-Task Sequence Prediction For Tunisian Arabizi Multi-Level Annotation elisa gugliotta, Marco Dinarelli and Olivier Kraif 178
AraWEAT: Multidimensional Analysis of Biases in Arabic Word Embeddings Anne Lauscher, Rafik Takieddin, Simone Paolo Ponzetto and Goran Glavaš

Parallel resources for Tunisian Arabic Dialect TranslationSaméh Kchaou, Rahma Boujelbane and Lamia Hadrich-Belguith
<i>Bert Transformer model for Detecting Arabic GPT2 Auto-Generated Tweets</i> Fouzi Harrag, Maria Dabbah, Kareem Darwish and Ahmed Abdelali
Contextual Embeddings for Arabic-English Code-Switched Data Caroline Sabty, Mohamed Islam and Slim Abdennadher
<i>Improving Arabic Text Categorization Using Transformer Training Diversification</i> Shammur Absar Chowdhury, Ahmed Abdelali, Kareem Darwish, Jung Soon-Gyo, Joni Salminen and Bernard J. Jansen
<i>Team Alexa at NADI Shared Task</i> mutaz younes, Nour Al-khdour and Mohammad AL-Smadi
Comparison between Voting Classifier and Deep Learning methods for Arabic Dialect Identification GHOUL dhaou and Gaël Lejeune
Simple vs Oversampling-based Classification Methods for Fine Grained Arabic Dialect Identification in Twitter Mohamed Lichouri and Mourad Abbas
Semi-supervised Fine-grained Approach for Arabic dialect detection task Nitin Nikamanth Appiah Balaji and Bharathi B
Arabic Dialect Identification Using BERT-Based Domain Adaptation Ahmad Beltagy, Abdelrahman Abouelenin and Omar ElSherief
Weighted combination of BERT and N-GRAM features for Nuanced Arabic Dialect Identification Abdellah El Mekki, Ahmed Alami, Hamza Alami, Ahmed Khoumsi and Ismail Berrada 268
Arabic dialect identification: An Arabic-BERT model with data augmentation and ensembling strategy Kamel Gaanoun and Imade Benelallam
Faheem at NADI shared task: Identifying the dialect of Arabic tweet Nouf AlShenaifi and Aqil Azmi 282
Identifying Nuanced Dialect for Arabic Tweets with Deep Learning and Reverse Translation Corpus Extension System Rawan Tahssin, Youssef Kishk and Marwan Torki
<i>The QMUL/HRBDT contribution to the NADI Arabic Dialect Identification Shared Task</i> Abdulrahman Aloraini, Massimo Poesio and Ayman Alhelbawy
Arabic Dialects Identification for All Arabic countries Ahmed aliwy, Hawraa Taher and Zena AboAltaheen
Arabic Dialect Identification Using BERT Fine-Tuning Moataz Mansour, Moustafa Tohamy, Zeyad Ezzat and Marwan Torki
LTG-ST at NADI Shared Task 1: Arabic Dialect Identification using a Stacking Classifier Samia Touileb

Conference Program

Saturday December 12, 2020 - Time zone · Central European Time (CET)

- 14:00–14:10 Opening Remarks Imed Zitouni
- 14:10–15:00 *Keynote Speaker* Perslav Nakov
- 15:00–15:20 Discussion: Arabic Language Technology Association
- 15:20–15:25 Short Break
- 15:25–17:45 Session 1: Main Workshop Papers
- 15:25–15:40 *German-Arabic Speech-to-Speech Translation for Psychiatric Diagnosis* Juan Hussain, Mohammed Mediani, Moritz Behr, M. Amin Cheragui, Sebastian Stüker and Alexander Waibel
- 15:40–15:55 *Hate Speech Detection in Saudi Twittersphere: A Deep Learning Approach* Raghad Alshaalan and Hend Al-Khalifa
- 15:55–16:10 Is it Great or Terrible? Preserving Sentiment in Neural Machine Translation of Arabic Reviews Hadeel Saadany and Constantin Orasan
- 16:10–16:25 Deep Diacritization: Efficient Hierarchical Recurrence for Improved Arabic Diacritization
 Badr AlKhamissi, Muhammad ElNokrashy and Mohamed Gabr

16:25–16:45 Break

- 16:45–17:00 *A Semi-Supervised BERT Approach for Arabic Named Entity Recognition* Chadi Helwe, Ghassan Dib, Mohsen Shamas and Shady Elbassuoni
- 17:00–17:15 *Empathy-driven Arabic Conversational Chatbot* Tarek Naous, Christian Hokayem and Hazem Hajj

Saturday December 12, 2020 - Time zone · Central European Time (CET) (continued)

- 17:15–17:30 *Machine Generation and Detection of Arabic Manipulated and Fake News* El Moatez Billah Nagoudi, AbdelRahim Elmadany, Muhammad Abdul-Mageed and Tariq Alhindi
- 17:30–17:45 *Transliteration of Judeo-Arabic Texts into Arabic Script Using Recurrent Neural Networks* Ori Terner, Kfir Bar and Nachum Dershowitz
- 17:45–18:15 Session 2: NADI Shared Task: Summary and Winner
- 17:45–18:00 *NADI 2020: The First Nuanced Arabic Dialect Identification Shared Task* Muhammad Abdul-Mageed, Chiyu Zhang, Houda Bouamor and Nizar Habash
- 18:00–18:15 Multi-dialect Arabic BERT for Country-level Dialect Identification
 Bashar Talafha, Mohammad Ali, Muhy Eddin Za'ter, Haitham Seelawi, Ibraheem
 Tuffaha, Mostafa Samir, Wael Farhan and Hussein Al-Natsheh
- 18:15–18:20 Short Break
- 18:20–18:53 Session 3: Main Workshop Poster Boasters
- 18:20–18:23 *On the Importance of Tokenization in Arabic Embedding Models* Mohamed Alkaoud and Mairaj Syed
- 18:23–18:26 *Tracing Traditions: Automatic Extraction of Isnads from Classical Arabic Texts* Ryan Muther and David Smith
- 18:26–18:29 Embed More Ignore Less (EMIL): Exploiting Enriched Representations for Arabic NLP ahmed younes and Julie Weeds
- 18:29–18:32 MANorm: A Normalization Dictionary for Moroccan Arabic Dialect Written in Latin Script Randa Zarnoufi, Hamid Jaafar, Walid Bachri and Mounia Abik
- 18:32–18:35 A Unified Model for Arabizi Detection and Transliteration using Sequence-to-Sequence Models
 Ali Shazal, Aiza Usman and Nizar Habash

Saturday December 12, 2020 - Time zone · Central European Time (CET) (continued)

18:35–18:38	<i>Multi-Task Sequence Prediction For Tunisian Arabizi Multi-Level Annotation</i> elisa gugliotta, Marco Dinarelli and Olivier Kraif
18:38–18:41	AraWEAT: Multidimensional Analysis of Biases in Arabic Word Embeddings Anne Lauscher, Rafik Takieddin, Simone Paolo Ponzetto and Goran Glavaš
18:41–18:44	Parallel resources for Tunisian Arabic Dialect Translation Saméh Kchaou, Rahma Boujelbane and Lamia Hadrich-Belguith
18:44–18:47	Bert Transformer model for Detecting Arabic GPT2 Auto-Generated Tweets Fouzi Harrag, Maria Dabbah, Kareem Darwish and Ahmed Abdelali

- 18:47–18:50 *Contextual Embeddings for Arabic-English Code-Switched Data* Caroline Sabty, Mohamed Islam and Slim Abdennadher
- 18:50–18:53 *Improving Arabic Text Categorization Using Transformer Training Diversification* Shammur Absar Chowdhury, Ahmed Abdelali, Kareem Darwish, Jung Soon-Gyo, Joni Salminen and Bernard J. Jansen
- 18:53–19:00 Concluding Remakrs (Underline) Imed Zitouni

19:00–20:00 Session 4: Main Workshop Posters and NADI Shared Task Papers

Team Alexa at NADI Shared Task mutaz younes, Nour Al-khdour and Mohammad AL-Smadi

Comparison between Voting Classifier and Deep Learning methods for Arabic Dialect Identification GHOUL dhaou and Gaël Lejeune

Simple vs Oversampling-based Classification Methods for Fine Grained Arabic Dialect Identification in Twitter Mohamed Lichouri and Mourad Abbas

Semi-supervised Fine-grained Approach for Arabic dialect detection task Nitin Nikamanth Appiah Balaji and Bharathi B

Saturday December 12, 2020 - Time zone · Central European Time (CET) (continued)

Arabic Dialect Identification Using BERT-Based Domain Adaptation Ahmad Beltagy, Abdelrahman Abouelenin and Omar ElSherief

Weighted combination of BERT and N-GRAM features for Nuanced Arabic Dialect Identification

Abdellah El Mekki, Ahmed Alami, Hamza Alami, Ahmed Khoumsi and Ismail Berrada

Arabic dialect identification: An Arabic-BERT model with data augmentation and ensembling strategy Kamel Gaanoun and Imade Benelallam

Faheem at NADI shared task: Identifying the dialect of Arabic tweet Nouf AlShenaifi and Aqil Azmi

Identifying Nuanced Dialect for Arabic Tweets with Deep Learning and Reverse Translation Corpus Extension System Rawan Tahssin, Youssef Kishk and Marwan Torki

The QMUL/HRBDT contribution to the NADI Arabic Dialect Identification Shared Task

Abdulrahman Aloraini, Massimo Poesio and Ayman Alhelbawy

Arabic Dialects Identification for All Arabic countries Ahmed aliwy, Hawraa Taher and Zena AboAltaheen

Arabic Dialect Identification Using BERT Fine-Tuning Moataz Mansour, Moustafa Tohamy, Zeyad Ezzat and Marwan Torki

LTG-ST at NADI Shared Task 1: Arabic Dialect Identification using a Stacking Classifier Samia Touileb

20:00–21:00 Panel Discussion (Zoom)