LREC-COLING 2024

The 6th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT) with Shared Tasks on Arabic LLMs Hallucination and Dialect to MSA Machine Translation

Workshop Proceedings

Editors
Hend Al-Khalifa, Kareem Darwish,
Hamdy Mubarak, Mona Ali
and Tamer Elsayed

25 May, 2024
Torino, Italia

Proceedings of The 6th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT) with Shared Tasks on Arabic LLMs Hallucination and Dialect to MSA Machine Translation

ISBN 978-2-493814-36-4
ISSN 2951-2093 (COLING); 2522-2686 (LREC)

Jointly organized by the ELRA Language Resources Association
and the International Committee on Computational Linguistics

Preface

Following the success of five editions of the of Open-Source Arabic Corpora and Corpora Processing Tools (OSACT) workshop collocated with LREC 2014, LREC 2016, LREC 2018, LREC 2020, and LREC2022, the sixth workshop comes to enable researchers and practitioners of Arabic language technologies to present their research with associated data and tools and to push the boundaries of their work in computational linguistics (CL), natural language processing (NLP), and information retrieval (IR). The sixth iteration gives special attention to areas of timely interest to the community, namely Large Language Models (LLMs), Generative AI, and dialectal translation, with two dedicated shared tasks on detecting LLM hallucinations and dialects to Modern Standard Arabic (MSA) translation.

OSACT6 had an acceptance rate of 43%, where we received 23 regular papers from which 10 papers were accepted, in addition to 6 shared task papers. We believe that the accepted papers are of high quality and present a mixture of interesting topics.

This year, we introduced the Shared Task on Dialectal Arabic (DA) to Modern Standard Arabic (MSA) Machine Translation, which attracted many teams from different countries in the Middle East, Europe, and the US. For this shared task, 29 teams signed up, and six teams made submissions to the competition's leaderboard, with five of them submitting their system description papers.

The other shared task aimed to address hallucinations (generation of false or misleading content) in Arabic Large Language Models (LLMs), such as GPT-3.5 and GPT-4. It features a dataset of 10,000 sentences from these LLMs annotated for factuality and correctness. There were two subtasks: A) detecting if a given sentence is factually correct, incorrect, or non-factual without additional information; and B) detecting the accuracy using the model's name, input word, part-of-speech (POS), and readability level. Only one team signed up and submitted a system paper.

Finally, we would like to thank everyone who in one way or another helped in making this workshop a success. Our special thanks go to the members of the program committee, who did an excellent job in reviewing the submitted papers, and to the LREC-COLING-2024 organizers. Finally, we would like to thank our authors and the workshop participants.

This volume documents the Proceedings of the 6th Workshop on Open-Source Arabic Corpora and Processing Tools, held on 25 May 2024 as part of the LREC-COLING-2024 conference.

Hend Al-Khalifa, Kareem Darwish, Hamdy Mubarak,
Mona Ali and Tamer Elsayed
OSACT6 Organizing Committee

Organizing Committee

- Hend Al-Khalifa, King Saud University, KSA
- Hamdy Mubarak, Qatar Computing Research Institute, Qatar
- Kareem Darwish, aiXplain Inc., US
- Tamer Elsayed, Qatar University, Qatar
- Mona Ali, Northeastern University, Canada

Programme Committee

- Ganesh Jawahar, University of British Columbia, Canada
- Go Inoue, Mohamed bin Zayed University of Artificial Intelligence, UAE
- Bassam Haddad, University of Petra, Jordan
- Hamada Nayel, Banha University, Egypt
- Ibrahim Abu Farha, The University of Sheffield, UK
- Imed Zitouni, Google, USA
- Almoataz B. Al-Said, Cairo University, Egypt
- Mourad Abbas, Assistant Secretary-General of Al-Tnall Al-Arabi in Algeria
- Nada Ghneim, Arab International University, Syria
- Omar Trigui, University of Sousse, Tunisia
- Salima Harrat, École Normale Supérieure de Bouzaréah (ENSB), Algeria
- Salima Mdhaffar, Avignon University (LIA), France
- Kamel Smaili, University of Lorraine, France
- Violetta Cavalli-Sforza, Al Akhawayn University, Morocco
- Wassim El-Hajj, American University of Beirut, Lebanon
- Wissam Antoun, ALMAnaCH - INRIA Paris, France
- Nada Almarwani, Taibah University, KSA
- Samah Aloufi, Taibah University, KSA
- Imene Bensalem, Constantine 2 University, Algeria
- Abdelkader El Mahdaouy, Mohammed VI Polytechnic University, Morocco
- Amr Keleg, University of Edinburgh, UK
- Wajdi Zaghouani, Hamad Bin Khalifa University, Qatar
- Amr El-Gendy, Arab Academy, Egypt

- Maha Alamri, AlBaha University, KSA
- Saied Alshahrani, Clarkson University, USA
- Lubna Alhenaki, Majmaah University, KSA
- Fatimah Alqahtani, Jazan University, KSA
- Eman Albilali, King Saud Univeristy, KSA
- Ahmed Abdelali, SDAIA, KSA
- Mohamed Al-Badrashiny, aiXplain Inc., US
- Firoj Alam, QCRI, Qatar
- Norah Alzahrani, SDAIA, KSA
- Nadir Durrani, QCRI, Qatar
- Ashraf Elneima, aiXplain Inc., US
- Nizar Habash, NYU-AD, UAE
- Walid Magdy, University of Edinburgh, UK
- Zaid Alyafeai, KFUPM, KSA
- Injy Hamed, NYU-AD, UAE
- Fouzi Harrag, Ferhat Abbas University, Algeria

Table of Contents

Workshop Program

Saturday 25 May 2024

Session 1: Main Workshop

9:00–9:10 ***Workshop Opening***

9:10–9:50 *Keynote Talk: Towards Arab-Centric Large Language Models*
Muhammad Abdul-Mageed

9:50–10:10 *AraTar: A Corpus to Support the Fine-grained Detection of Hate Speech Targets in the Arabic Language*
Seham Alghamdi, Youcef Benkhedda, Basma Alharbi and Riza Batista-Navarro

10:10–10:30 *CLEANANERCorp: Identifying and Correcting Incorrect Labels in the ANERcorp Dataset*
Mashael AlDuwais, Hend Al-Khalifa and Abdulmalik AlSalman

Session 2: Main Workshop (Cont.)

11:00–11:20 *Munazarat 1.0: A Corpus of Arabic Competitive Debates*
Mohammad M. Khader, AbdulGabbar Al-Sharafi, Mohamad Hamza Al-Sioufy, Wajdi Zaghouani and Ali Al-Zawqari

11:20–11:40 *Leveraging Corpus Metadata to Detect Template-based Translation: An Exploratory Case Study of the Egyptian Arabic Wikipedia Edition*
Saied Alshahrani, Hesham Haroon Mohammed, Ali Elfilali, Mariama Njie and Jeanna Matthews

11:40–12:00 *A Novel Approach for Root Selection in the Dependency Parsing*
Sharefah Ahmed Al-Ghamdi, Hend Al-Khalifa and Abdulmalik AlSalman

12:00–12:20 *AraMed: Arabic Medical Question Answering using Pretrained Transformer Language Models*
Ashwag Alasmari, sarah alhumoud and Waad Alshammari

12:20–12:40 *The Multilingual Corpus of World's Constitutions (MCWC)*
Mo El-Haj and Saad Ezzini

12:40–13:00 *TafsirExtractor: Text Preprocessing Pipeline preparing Classical Arabic Literature for Machine Learning Applications*
Carl Kruse and Sajawel Ahmed

Saturday 25 May 2024 (continued)

Session 3: Main Workshop (Cont.)

14:00–14:20 *Advancing the Arabic WordNet: Elevating Content Quality*

Abed Alhakim Freihat, Hadi Mahmoud Khalilia, Gábor Bella and Fausto Giunchiglia

14:20–14:40 *Arabic Speech Recognition of zero-resourced Languages: A case of Shehri (Jibbali) Language*

Norah A. Alrashoudi, Omar Said Alshahri and Hend Al-Khalifa

Session 4: Shared Tasks

14:40–14:55 *OSACT6 Dialect to MSA Translation Shared Task Overview*

Ashraf Hatim Elneima, AhmedElmogtaba Abdelmoniem Ali Abdelaziz and Kareem Darwish

14:55–15:10 *OSACT 2024 Task 2: Arabic Dialect to MSA Translation*

hanin atwany, Nour Rabih, Ibrahim Mohammed, Abdul Waheed and Bhiksha Raj

15:10–15:25 *ASOS at OSACT6 Shared Task: Investigation of Data Augmentation in Arabic Dialect-MSA Translation*

Omer Nacar, Abdullah Alharbi, Serry Sibaee, Samar Ahmed, Lahouari Ghouti and Anis Koubaa

15:25–15:50 *LLM-based MT Data Creation: Dialectal to MSA Translation Shared Task*

AhmedElmogtaba Abdelmoniem Ali Abdelaziz, Ashraf Hatim Elneima and Kareem Darwish

15:50–16:00 *Sirius_Translators at OSACT6 2024 Shared Task: Fin-tuning Ara-T5 Models for Translating Arabic Dialectal Text to Modern Standard Arabic*

Salwa Saad Alahmari

Saturday 25 May 2024 (continued)

Session 5: Shared Tasks (Cont.)

16:30–16:45 *AraT5-MSAizer: Translating Dialectal Arabic to MSA*

Murhaf Fares

16:45–17:00 *ASOS at Arabic LLMs Hallucinations 2024: Can LLMs detect their Hallucinations :)*

Serry Taiseer Sibaee, Abdullah I. Alharbi, Samar Ahmed, Omar Nacar, Lahouri Ghouti and Anis Koubaa

17:00–17:05 ***Workshop Closing***