The 4th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT) with Shared Task on Offensive Language Detection

http://edinburghnlp.inf.ed.ac.uk/workshops/OSACT4/

Workshop Description:

Given the success of the first, second, and third workshops on Open-Source Arabic Corpora and Corpora Processing Tools (OSACT) in LREC 2014, LREC 2016 and LREC 2018, the fourth workshop comes to encourage researchers and practitioners of Arabic language technologies, including computational linguistics (CL), natural language processing (NLP), and information retrieval (IR) to share and discuss their research efforts, corpora, and tools. The workshop will also give special attention on Human Language Technologies based on AI/Machine Learning, which is one of LREC 2020 hot topics. In addition to the general topics of CL, NLP and IR, the workshop will give a special emphasis on Offensive Language Detection shared task.

Shared Task on Offensive Language Detection:

Offensive speech (vulgar or targeted offense), as an expression of heightened polarization and discourse in society, has been on the rise. This is due in part to the large adoption of social media platforms that allow for greater polarization. The shared task attempts to detect such speech in the realm of Arabic social media.

In this task, we will use a combination of the SemEval 2020 Arabic offensive language dataset (OffensEval2020), which contains 10,000 tweets that were manually annotated (testing data), and a new dataset of tweets that were annotated in a semi-supervised fashion. The later dataset specifically identifies users who reply often to specific accounts, where the majority of their replies were deemed offensive using a preliminary classifier (training data). The intuition is that the remaining replies are likely to be offensive also. So, the purpose of this shared task is to intensify research on the identification of offensive content in Arabic language Twitter posts. One goal of the workshop is to define shared challenges using this dataset. We encourage submissions describing experiments for research tasks on the dataset.

Motivation and Topics of interest:

In the NLP, CL, and IR communities, Arabic is considered to be relatively resource-poor compared to English. This situation was thought to be the reason for the limited number of corpus-based studies in Arabic. However, the past years witnessed the emergence of new considerably free Modern Standard Arabic (MSA) corpora and to a lesser extent Arabic processing tools.

This workshop follows the footsteps of previous editions of OSACT to provide a forum for researchers to share and discuss their ongoing work. This workshop is timely given the continued rise in research projects focusing on Arabic Language Resources.

Topics of interest

Corpora:

- Surveying and criticizing the design of available Arabic corpora, their associated and processing tools.
- Availing new annotated corpora for NLP and IR applications such as named entity recognition, machine translation, sentiment analysis, text classification, and language learning.
- Evaluating the use of crowdsourcing platforms for Arabic data annotation.

Tools and Technologies:

- Language education, e.g., L1 and L2.
- Language modeling and pre-trained models.
- Tokenization, normalization, word segmentation, morphological analysis, part-of-speech tagging, etc.
- Sentiment analysis, dialect identification, and text classification
- Dialect translation
- Fake news detection
- Web and social media search and analytics

Issues in the design, construction and use of Arabic LRs: text, speech, sign, gesture, image, in single or multimodal/multimedia data

- Guidelines, standards, best practices and models for LRs interoperability
- Methodologies and tools for LRs construction and annotation
- Methodologies and tools for extraction and acquisition of knowledge
- Ontologies, terminology and knowledge representation
- LRs and Semantic Web (including Linked Data, Knowledge Graphs, etc.)

Important Dates

Submission deadline: 20 February 2020

Notification of acceptance: 10 March 2020

Camera Ready of manuscripts: 25 March 2020

Workshop date: 12 May 2020.

Submission guidelines

The language of the workshop is English and submissions should be with respect to LREC 2020 paper submission instructions (https://lrec2020.lrec-conf.org/en/submission2020/authors-kit/). All papers will be peer reviewed possibly by three independent referees. Papers must be submitted electronically in PDF format to the STAR system.

When submitting a paper from the STAR page, authors will be asked to provide essential information about resources (in a broad sense, i.e. technologies, standards, evaluation kits, etc.) that have been used for the work described in the paper or are a new result of your research. Moreover, ELRA encourages all LREC authors to share the described LRs (data, tools, services, etc.), to enable their reuse, replicability of experiments (including evaluation ones).

Identify, Describe and Share your LRs!

- Describing your LRs in the LRE Map is now a normal practice in the submission procedure of LREC (introduced in 2010 and adopted by other conferences). To continue the efforts initiated at LREC 2014 about “Sharing LRs” (data, tools, web-services, etc.), authors will have the possibility, when submitting a paper, to upload LRs in a special LREC repository. This effort of sharing LRs, linked to the LRE Map for their description, may become a new “regular” feature for conferences in our field, thus contributing to creating a common repository where everyone can deposit and share data.

- As scientific work requires accurate citations of referenced work so as to allow the community to understand the whole context and also replicate the experiments conducted by other researchers, LREC 2020 endorses the need to uniquely Identify LRs through the use of the International Standard Language Resource Number (ISLRN, www.islrn.org), a Persistent Unique Identifier to be assigned to each Language Resource. The assignment of ISLRNs to LRs cited in LREC papers will be offered at submission time.

Organizing Committee:

Hend Al-Khalifa, King Saud University, KSA

Walid Magdy, University of Edinburgh, UK

Kareem Darwish, Qatar Computing Research Institute, Qatar

Tamer Elsayed, Qatar University, Qatar

Hamdy Hussein, Qatar Computing Research Institute, Qatar

Programme Committee

Nizar Habash, New York University Abu Dhabi, UAE

Wajdi Zaghouani, Carnegie Mellon University, Qatar

Mahmoud El-Haj, Lancaster University, UK

Wassim El-Hajj, American University of Beirut, Lebanon

Irina Temnikova, Qatar Computing Research Institute, Qatar

Abeer Aldayel, University of Edinburgh, UK

Raghad Alshalan, Imam Abdulrahman Bin Faisal University, KSA

Shahad Alshalan, Imam Abdulrahman Bin Faisal University, KSA

Luluh Aldhubayi, King Saud University, KSA

Nora Al-Twairish, King Saud University, KSA

Khaled Shaalan, The British University in Dubai, UAE

Fethi Bougares, Université du Maine, Avenue Laënnec, France

Hazem Hajj, American University of Beirut, Lebanon

Nadi Tomeh, LIPNUniversity of Paris 13, Sorbonne Paris Cité Paris, France

Samhaa R. El-Beltagy, Nile UniversitySheikh Zayed, GizaEgypt

Muhammad Abdul-Mageed, The university of British Columbia, Canada

Lamia Hadrich Belguith, University of Sfax, Tunisia

Reem Suwaileh, Qatar University

Maram Hasanain, Qatar University

Mucahid Kutlu, TOBB University, Turkey

More names to come . . .