The 5th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT)

with Shared Tasks on Quran QA and Fine-Grained Hate Speech Detection

Website: https://osact-lrec.github.io/

Workshop Description:

Given the success of the first, second, third, and fourth workshops on Open-Source Arabic Corpora and Corpora Processing Tools (OSACT) in LREC 2014, LREC 2016, LREC 2018 and LREC 2020, the fifth 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 latest research efforts, corpora, and tools. The workshop will give a special emphasis on two shared tasks, namely: Quran QA and Fine-Grained Hate Speech Detection.

Quran QA Shared Task

Reading comprehension (RC) task, viewed as a type of questions-answering (QA) tasks, is perceived as the ideal method to evaluate language understanding by computer systems. Given a passage of text, a machine reading comprehension system is required to answer a set of questions over the given passage. We propose a shared task of Arabic Reading Comprehension over the Holy Qur'an, aiming to trigger state-of-the-art reading comprehension research on a book that is sacredly held by more than 1.8 billion people across the world.

The Holy Qur'an is composed of 114 chapters (Suras) and 6,236 verses that comprise more than 80k words in Classical Arabic. The participating systems are expected to provide answers to questions on given passages (sets of consecutive verses) from the Holy Qur'an, where the answers are spans of text extracted from the given passages. Our dataset for the shared task (developed based on the AyaTEC dataset) is composed of 1,348 QA tuples of passage-question-answer triplets for 169 different questions. The dataset adopts the same format of the SQuAD v1.1 dataset. A question might have more than one answer in the passage; therefore, the system is expected to extract all of them and return a ranked list of answer spans. To evaluate the system performance, multiple evaluation measures, that adopt exact and partial matching of spans, will be used.

Fine-grained detection of hate speech on Arabic Twitter Shared Task

Detecting offensive language and hate speech has gained an increasing interest from researchers in NLP and computational social sciences communities in the past few years. For example, at the ACL 2021 main conference, there wereare 3 papers about offensive language, and 10 papers about hate speech (https://2021.aclweb.org/program/accept/). Additionally, there was a dedicated workshop on online abuse and harm with a shared task on hateful memes (https://www.workshopononlineabuse.com/home#h.czplpodomjyq). Detecting offensive language and hate speech is very important for online safety, content moderation, etc. Studies show that the presence of hate speech may be connected to hate crimes (Hate Speech Watch, 2014).

Given the success of the shared task on Arabic offensive language detection at OSACT 2020 (https://edinburghnlp.inf.ed.ac.uk/workshops/OSACT4/), we decided to continue our effort to enhance detection of offensive language and hate speech on Arabic Twitter. We share with the research community the largest annotated Arabic tweets that don't have bias towards specific topics, genres or dialects. Each tweet is judged by 3 annotators using crowdsourcing for offensiveness. Offensive tweets were classified to one of hate speech types: Gender, Race, Ideology, Social Class, Religion, and Disability. Also, annotators judged whether a tweet has vulgar language or violence.

Hate speech is defined as any kind of offensive language (insults, slurs, threats, encouraging violence, etc.) that targets a person or a group of people based on common characteristics such as race, gender, religion and belief, etc.

The corpus contains ~13K tweets in total: 35% are offensive and 11% are hate speech. Vulgar and violence tweets represent 1.5% and 0.7% of the whole corpus.

Ratios of offensive language and hate speech in the corpus are the highest among other corpora without using pre-specified keywords or selecting a specific domain.

More details will be published soon.

<u>We will have 3 shared subtasks:</u>

Subtask 1: Detect whether a tweet is offensive or not.

Subtask 2: Detect whether a tweet has hate speech or not.

Subtask 3: Detect fine-grained type of hate speech.

The same evaluation platform (Codalab) used in OSACT 2020 shared task will be used in the shared tasks.

Data will be split into: 70% for training, 10% development, and 20% for testing.

We encourage participants to use this data and/or any other external data (previous datasets, lexicons, in-house data, etc.), and try to explain model behaviour and study model generalization.

Topics of interest

Language Resources:

- Pre-trained Arabic language models and their applications.
- Surveying and evaluating 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.
- Open source Arabic processing toolkits.

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.)

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 due: April 10, 2022

Notification of acceptance: May 1, 2022

Camera-ready papers due: May 25, 2022

Workshop date: June 20, 2022

Submission guidelines

The language of the workshop is English and submissions should be with respect to LREC 2022 paper submission instructions (https://lrec2022.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 START page, authors will be asked to provide essential information about resources (in a broad sense, i.e. also 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 and replicability of experiments (including evaluation ones).

Organizing Committee:

Hend Al-Khalifa, King Saud University, KSA

Walid Magdy, University of Edinburgh, UK

Tamer Elsayed, Qatar University, Qatar

Hamdy Mubarak, Qatar Computing Research Institute, Qatar

Kareem Darwish, AiXplain, inc. USA

Abdulmohsen Al-Thubaity, KACST, KSA

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

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, LIPN University of Paris 13, Sorbonne Paris CitéParis, 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, Qatar

Maram Hasanain, Qatar University, Qatar

Mucahid Kutlu, TOBB University, Turkey

Abdulrahman Almuhareb, King Abdulaziz City for Science and Technology, KSA

Waleed Alsanie, King Abdulaziz City for Science and Technology, KSA

Sakhar Alkhereyf, King Abdulaziz City for Science and Technology, KSA

More names to come . . .