Mohamed I. Eldesouki Mohamed

Publishing as "Mohamed Eldesouki"
https://github.com/disooqi/
(last update Oct. 2018)

Qatar Computing Research Institute (QCRI), Doha, Qatar
(+974) 33-542583
mohamohamed@qf.org.qa disooqi@gmail.com

EXPERIENCE

Research Associate Dec. 2015 - Present
Qatar Computing Research Institute (QCRI), Doha, Qatar

- Working on Dialectal Arabic Processing where I
  - currently work in collaboration with MIT-CSAIL on spoken language processing in QMDIS Project (QCRI-MIT Advanced Dialect Identification System) to classify between 5 Arabic dialects namely Egyptian, Levantine, Gulf, and MSA. Our best results gives us an accuracy of 78% overall accuracy across the five dialects using the 2017 Multi-Genre Broadcast challenge (MGB-3) data. In order to achieve a robust dialect identification, we explored using Siamese neural network models to learn similarity and dissimilarities among Arabic dialects, as well as i-vector post-processing to adapt domain mismatches. Both acoustic and linguistic features were used.
  - Led the work in the DSL Shared Task 2016 of the VarDial 2016 workshop for Arabic dialect identification and achieved the BEST ACCURACY among 18 participants,
  - Compare between two approaches namely bidirectional LSTM and SVM to build the state-of-the-art word segmentation for 4 different dialects, namely Egyptian, Levantine, Gulf, and Maghrebi[1] using an inhouse dataset.
  - Observed that there are shared pan-dialectal linguistic phenomena that allow computational models for dialects to learn from each other. Accordingly, I built a unified segmentation model where the training data for different dialects are combined and a single model is trained.
  - Currently involved in using Multi-task sequence-to-sequence learning to perform segmentation and POS in one step as a way to alleviate the error propagation in the pipeline approach.
- Developed ClassStrength[2]; a multi-lingual tweet classifier tool that classifies tweets into 14 general-purpose classes using distant supervision approach.

Research Assistant Jul. 2009 - Jun. 2012
Institute of Statistical Studies and Research (ISSR), Cairo University, Cairo, Egypt

Responsibilities:

- Teaching courses:
  CS-507: Introduction to Programming using Python
  CS-503: Fundamentals of Natural Language Processing

- Developed Text Preprocessing Cog (TPC); is a Python package that does stemming, tokenization, sentence breaking, segmentation, normalization, and POS tagging for Arabic language.
- WebCS; a news aggregator that crawl more than 120 Arabic newswires and automatically recognize the title, body, author(s) and publish date.

Senior Research Software Engineer Jul. 2012 - Dec. 2015
Taya IT Company, Cairo, Egypt

[1] https://github.com/qcri/dialectal_arabic_tools
[2] ClassStrength website, http://alt.qcri.org/class_strength/

Participated in developing:
TAPS (Taya Arabic Processing Suite); that includes Named Entity Recognition and Auto-complete and spelling correction,
TESE (Taya Enterprise Search Engine), Taya RecSys, and Greetings Studio.

PROJECTS

The following are some projects that I have participated in:

QMDIS Project (QCRI-MIT Advanced Dialect Identification System) in collaboration with MIT-CSAIL on spoken language processing to classify between 5 Arabic dialects namely Egyptian, Levantine, Gulf, and MSA.

- Design the architecture of the system
- Build the system where the communication between the client (web browser) and the server is using Websocket-based client-server protocol. The browser capture the audio through the microphone API and send the raw audio to the server
- Technologies used: Kaldi, Tornado (asynchronous networking library), Tensorflow, and Python.
- My work was recently accepted for a demo-paper in ICASSP 2018.

FARASA project the state-of-the-art Arabic language segmenter, POS tagger, diacritizer, NER, dependency parser and constituency parser.

- Design the system based on Microservice Architecture and expose all the tools of Farasa as Web API
- Technologies used are Jersey (JAX-RS), Flask web framework, Tensorflow, Java, Python

TAPS - Taya Arabic Processing Suite, handling many tasks namely; language detection, named-entity recognition, light Stemming, Arabic morphological analysis, Spell Correction.

Buzzdiggr project is a monitoring software that listens to all social media platforms and the entire web for mentions of your brand then provides an array of powerful features for your analysis.

- Building a dataset for training classifiers, and assuring that the data is covering certain criteria such as particular dialects, balanced in its sentiment (so that not all data are neutral), covering particular regions and countries
- Managing the full annotation process by follow up the annotators, assure the quality of annotation, developing tools for annotation and providing their annotation batches and finally analyze and report their results
- Developing a tool that uses the active learning part of VW as a core and use that tool to accelerate the annotation process
- Participated in the migration process of the NLP engine of buzzdiggr from Java to Python.

EDUCATION

Master (M.Sc) in Computer Science, Cairo University Jan. 2012
(Master by Research) - thesis + 3 publications through the Master's period.
Rank: Top of my class in pre-master year (with grade 86.2%)
Thesis Title: AN INTELLIGENT AGENT FOR ARABIC WEB INFORMATION RETRIEVAL
Supervisors: Dr. Kareem Darwish, QCRI/HBKU, Dr. Mervat Gheith, and Dr. Waleed Arafa, Cairo University, Egypt.
Specialized in Web Information Retrieval and Web Personalization systems for Arabic language.

Postgraduate Diploma in Computer Science, Cairo University May 2006

Bachelor of Computers and Information, Cairo University May 2003
Majoring in: Information Systems

PUBLICATIONS (Google Scholar citations: 51 - h-index: 4)

2018

- **Mohamed Eldesouki**, Suwon Shon, and Ahmed Ali, (2018), *QCRI-MIT Live Arabic Dialect Identification System*, ICASSP, Calgary, Canada [DEMO]
- Kareem Darwish, Hamdy Mubarak, **Mohamed Eldesouki**, Ahmed Abdelali, Younes Samih, Randah Alharbi, Mohammed Attia, Walid Magdy, and Laura Kallmeyer, (2018), *Multi-Dialect Arabic POS Tagging: A CRF Approach*, In 11th edition of the Language Resources and Evaluation Conference (LREC), 7-12 May 2018, Miyazaki (Japan).

2017

- Salvatore Romeo, Giovanni Da San Martino, Yonatan Belinkov, Alberto Barrón-Cedeño, **Mohamed Eldesouki**, Kareem Darwish, Hamdy Mubarak, James Glass, and Alessandro Moschitti, (2017), *Language Processing and Learning Models for Community Question Answering in Arabic*, Information Processing & Management (In Press), (https://doi.org/10.1016/j.ipm.2017.07.003).
- Younes Samih, **Mohamed Eldesouki**, Mohammed Attia,Kareem Darwish, Ahmed Abdelali, Hamdy Mubarak and Laura Kallmeyer, (2017), *Learning from Relatives: Unified Dialectal Arabic Segmentation*. In Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017), Vancouver, Canada, 432-441.
- Walid Magdy, and **Mohamed Eldesouky**, (2017), *ClassStrength: A Multilingual Tool for Tweets Classification*, Proc. of the 2017 IEEE/ACM International Conference on Advances in social networks analysis and mining (ASONAM), Sydney, Australia, 593-596.
- **Mohamed Eldesouki**, Younes Samih, Ahmed Abdelali, Mohammed Attia, Hamdy Mubarak, Kareem Darwish and Kallmeyer Laura, (2017), *Arabic Multi-Dialect Segmentation: bi-LSTM-CRF vs. SVM*, http://arxiv.org/abs/1708.05891.
- Younes Samih, Mohammed Attia, **Mohamed Eldesouki**, Ahmed Abdelali, Hamdy Mubarak, Laura Kallmeyer and Kareem Darwish, (2017), *A Neural Architecture for Dialectal Arabic Segmentation*. In Proc. of The 3rd Arabic Natural Language Processing Workshop (WANLP-2017) co-located with EACL 2017, Valencia, Spain, pages 46-54.
- Kareem Darwish, Hamdy Mubarak, Ahmed Abdelali and **Mohamed Eldesouki**, (2017), *Arabic POS Tagging: Don't Abandon Feature Engineering Just Yet*, In Proc. of The 3rd Arabic NLP Workshop (WANLP-2017) co-located with EACL 2017, Valencia, Spain, P. 130.

2016

- **Mohamed Eldesouki**, Fahim Dalvi, Hassan Sajjad, and Kareem Darwish, (2016), *QCRI @ DSL 2016: Spoken Arabic Dialect Identification Using Textual Features*, Proc. of the 3rd Workshop on NLP for Similar Languages, Varieties and Dialects, (VarDial 3), Osaka, Japan, P. 221.

2012

- **Mohamed I. Eldesouki**, *An Intelligent Agent for Arabic Web Information Retrieval*, (2012), Master's Thesis, Cairo University.

2011

- **Mohamed Eldesouki**, Waleed Arafa, Kareem Darwish, Mervat H. Gheith, (2011), *Representing Arabic Documents Using Controlled Vocabulary Extracted from Wikipedia*, In Proc. of The 11th Conference on Language Engineering (ESOLEC'11), Cairo, Egypt.
- **Mohamed Eldesouki**, Waleed Arafa, Kareem Darwish, Mervat Gheith, (2011), *Using Wikipedia for Retrieving Arabic Documents*. In Proceedings of Arabic Language Technology International Conference (ALTIC 2011), Alexandria, Egypt.

2009

- **Mohamed I. Eldesouki**, Waleed M. Arafa, Kareem M. Darwish,(2009), *Stemming techniques of Arabic Language: Comparative Study from the Information Retrieval Perspective*, The Egyptian Computer Journal. 36(1):30-49.

TECHNOLOGY

Proficient and familiar with a vast array of programming languages, concepts and technologies, including:

Programming Languages:
Proficient in Python, C/C++, Java, Prolog and Lisp and familiar with C# .NET and PHP.
Machine learning & Scientific packages; Tensorflow, Keras, Scikit-learn, Octave, CRF++, Numby, Scipy, matplotlib, Jupyter notebook, YASMET, YamCha.
NLP & IR packages; NLTK, Indri (Lemur project), Solr (Lucene), FARASA toolkit, MADAMIRA, SRILM (SRI Language Model), Kaldi.
Technologies and Tools:
Web Platforms: Java EE, Django, and Flask
Web Frontend Technologies: HTML, XHTML, CSS, JavaScript, HTML DOM, Ajax, XML, XML DOM, Web sevices, JSON and Bootstrap framework.
Other tech.: Git, conda, virtual environment, virtual machines, Docker, Python Packaging.
OS Platforms: Linux, Windows, Mac

OTHER ACTIVITIES

- Reviewing for: ACL 2018, BJIT, RANLP 2017 http://lml.bas.bg/ranlp2017/pc.php,
- Continuing Education: Achieved several certificates through MOOC courses including machine learning, deep learning, and General AI[3].

[3]For the full list of certificates please visit my linkedIn profile https://www.linkedin.com/in/disooqi/ and look for Accomplishments-Certifications section