The 31st International Conference on Computational Linguistics

AbjadNLP

The 1st Workshop on NLP for Languages Using Arabic Script

Editors

Mo El-Haj Amal Haddad Cynthia Amol

Sina Ahmadi Hugh Paterson III

Ignatius Ezeani Saad Ezzini Paul Rayson

Proceedings of the Workshop

January 19, 2025

https://wp.lancs.ac.uk/abjad

©2025 The Association for Computational Linguistics

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
209 N. Eighth Street
Stroudsburg, PA 18360
USA
Tel: +1-570-476-8006
Fax: +1-570-476-0860
acl@aclweb.org

ISBN 979-8-89176-200-8

Preface

It is my pleasure to welcome you to the inaugural edition of the AbjadNLP Workshop, an event dedicated to advancing research and applications for languages that use the Arabic script in Natural Language Processing (NLP). This workshop serves as a vital platform to unite researchers and practitioners addressing the linguistic, cultural, and computational challenges inherent to these languages, fostering innovative solutions for low-resource and historically underrepresented languages.

We received 27 submissions, of which 16 papers were accepted, resulting in an acceptance rate of approximately 59%. This strong start underscores the growing recognition of the significance of Arabic-script languages within the NLP community.

The submissions to this workshop showcased remarkable diversity, reflecting the linguistic and cultural richness of languages that utilise the Arabic script. Contributions span a variety of languages, including Wolofal, a form of Ajami script used for the Wolof language in West Africa; Persian, which, despite its long tradition of NLP research, continues to face challenges related to limited resources in stylistic and dialectal variations; and Turkish, focusing on the integration of Arabic-origin words into the Turkish language and its linguistic evolution. Additionally, research on Perso-Arabic scripts, used in languages such as Persian, Urdu, Pashto, and others, underscores the flexibility and adaptability of the Arabic script across diverse linguistic families and regions.

The workshop also includes studies on Arabic, exploring fundamental challenges and methodologies that can inform research on other low-resource Abjad and Ajami languages. Collectively, these contributions highlight the breadth of research addressing both resource-rich and low-resource languages, advancing our understanding of the unique computational challenges posed by Arabic-script languages.

I am deeply grateful to our co-organisers, reviewers, and authors for their invaluable contributions to this workshop. Their dedication and expertise have ensured the high quality of this year's programme, and their work paves the way for future innovations in this field.

This workshop represents a significant step towards building a vibrant and collaborative community dedicated to Arabic-script languages in NLP. We look forward to inspiring presentations, engaging discussions, and continued growth in the coming years.

Thank you for joining us, and I hope you find this workshop as inspiring as we do.

Mo El-Haj
Chair Organiser
AbjadNLP Workshop

Organizing Committee

Mo El-Haj
Hugh Paterson III
Saad Ezzini
Ignatius Ezeani
Mahum Hayat Khan
Muhammad Sharjeel
Sina Ahmadi
Cynthia Amol
Amal Haddad Haddad
Jaleh Delfani
Ruslan Mitkov
Paul Rayson

Table of Contents

The Best of Both Worlds: Exploring Wolofal in the Context of NLP
Ngoc Tan Le, Ali Mijiyawa, Abdoulahat Leye and Fatiha Sadat 1

MultiProp Framework: Ensemble Models for Enhanced Cross-Lingual Propaganda Detection in Social Media and News using Data Augmentation, Text Segmentation, and Meta-Learning
Farizeh Aldabbas, Shaina Ashraf, Rafet Sifa and Lucie Flek 7

Towards Unified Processing of Perso-Arabic Scripts for ASR
Srihari Bandarupalli, Bhavana Akkiraju, Sri Charan Devarakonda, Harinie Sivaramasethu, Vamshiraghusimha Narasinga and Anil Vuppala 23

In-Depth Analysis of Arabic-Origin Words in the Turkish Morpholex
Mounes Zaval, Abdullah İhsanoğlu, Asım Ersoy and Olcay Taner Yıldız 29

DadmaTools V2: an Adapter-Based Natural Language Processing Toolkit for the Persian Language
sadegh jafari, Farhan Farsi, Navid Ebrahimi, Mohamad Bagher Sajadi and Sauleh Eetemadi 37

Developing an Informal-Formal Persian Corpus: Highlighting the Differences between Two Writing Styles
Vahide Tajalli, Mehrnoush Shamsfard and Fateme Kalantari 44

Boosting Sentiment Analysis in Persian through a GAN-Based Synthetic Data Augmentation Method
Masoumeh Mohammadi, Mohammad Ruhul Amin and Shadi Tavakoli 54

Psychological Health Chatbot, Detecting and Assisting Patients in their Path to Recovery
sadegh jafari, Mohammad Erfan Zare, Amireza Vishte, Mirzae Melike, Zahra Amiri, Sima Mohammadparast and Sauleh Eetemadi 64

A Derivational ChainBank for Modern Standard Arabic
Reham Marzouk, sondos krouna and Nizar Habash 78

Sentiment Analysis of Arabic Tweets Using Large Language Models
Pankaj Dadure, Ananya Dixit, Kunal Tewatia, Nandini Paliwal and Anshika Malla 88

Evaluating Large Language Models on Health-Related Claims Across Arabic Dialects
Abdulsalam obaid Alharbi, Abdullah Alsuhaibani, Abdulrahman Abdullah Alalawi, Usman Naseem, Shoaib Jameel, salil kanhere and Imran Razzak 95

Can LLMs Verify Arabic Claims? Evaluating the Arabic Fact-Checking Abilities of Multilingual LLMs
Ayushman Gupta, Aryan Singhal, Thomas Law, Veekshith Rao, Evan Duan and Ryan Luo Li . 104

Can LLMs Translate Cultural Nuance in Dialects? A Case Study on Lebanese Arabic
silvana yakhni and Ali Chehab 114

Automated Generation of Arabic Verb Conjugations with Multilingual Urdu Translation: An NLP Approach
Haq Nawaz, Manal Elobaid, Ali Al-Laith and Saif Ullah 136

Evaluation of Large Language Models on Arabic Punctuation Prediction
Asma Ali Al Wazrah, Afrah Altamimi, Hawra Aljasim, Waad Alshammari, Rawan Al-Matham, Omar Elnashar, Mohamed Amin and Abdulrahman AlOsaimy 144

Evaluating RAG Pipelines for Arabic Lexical Information Retrieval: A Comparative Study of Embedding and Generation Models
Raghad Al-Rasheed, Abdullah Al Muaddi, Hawra Aljasim, Rawan Al-Matham, Muneera Alhoshan, Asma Al Wazrah and Abdulrahman AlOsaimy 155

Conference Program

Sunday, January 19, 2025

9:00–9:05 Welcome and Opening Remarks

09:05–09:25 *The Best of Both Worlds: Exploring Wolofal in the Context of NLP*
Ngoc Tan Le, Ali Mijiyawa, Abdoulahat Leye and Fatiha Sadat

09:25–09:45 *MultiProp Framework: Ensemble Models for Enhanced Cross-Lingual Propaganda Detection in Social Media and News using Data Augmentation, Text Segmentation, and Meta-Learning*
Farizeh Aldabbas, Shaina Ashraf, Rafet Sifa and Lucie Flek

09:45–10:05 *Towards Unified Processing of Perso-Arabic Scripts for ASR*
Srihari Bandarupalli, Bhavana Akkiraju, Sri Charan Devarakonda, Harinie Sivaramasethu, Vamshiraghusimha Narasinga and Anil Vuppala

10:05–10:25 *In-Depth Analysis of Arabic-Origin Words in the Turkish Morpholex*
Mounes Zaval, Abdullah İhsanoğlu, Asım Ersoy and Olcay Taner Yıldız

Coffee Break

11:00–11:20 *DadmaTools V2: an Adapter-Based Natural Language Processing Toolkit for the Persian Language*
sadegh jafari, Farhan Farsi, Navid Ebrahimi, Mohamad Bagher Sajadi and Sauleh Eetemadi

11:20–11:40 *Developing an Informal-Formal Persian Corpus: Highlighting the Differences between Two Writing Styles*
Vahide Tajalli, Mehrnoush Shamsfard and Fateme Kalantari

11:40–12:00 *Boosting Sentiment Analysis in Persian through a GAN-Based Synthetic Data Augmentation Method*
Masoumeh Mohammadi, Mohammad Ruhul Amin and Shadi Tavakoli

12:00–12:20 *Psychological Health Chatbot, Detecting and Assisting Patients in their Path to Recovery*
sadegh jafari, Mohammad Erfan Zare, Amireza Vishte, Mirzae Melike, Zahra Amiri, Sima Mohammadparast and Sauleh Eetemadi

Lunch Break

13:20–13:40 *A Derivational ChainBank for Modern Standard Arabic*
Reham Marzouk, sondos krouna and Nizar Habash

13:40–14:00 *Sentiment Analysis of Arabic Tweets Using Large Language Models*
Pankaj Dadure, Ananya Dixit, Kunal Tewatia, Nandini Paliwal and Anshika Malla

14:00–14:20 *Evaluating Large Language Models on Health-Related Claims Across Arabic Dialects*
Abdulsalam obaid Alharbi, Abdullah Alsuhaibani, Abdulrahman Abdullah Alalawi, Usman Naseem, Shoaib Jameel, salil kanhere and Imran Razzak

14:20–14:40 *Can LLMs Verify Arabic Claims? Evaluating the Arabic Fact-Checking Abilities of Multilingual LLMs*
Ayushman Gupta, Aryan Singhal, Thomas Law, Veekshith Rao, Evan Duan and Ryan Luo Li

14:40–15:00 *Can LLMs Translate Cultural Nuance in Dialects? A Case Study on Lebanese Arabic*
silvana yakhni and Ali Chehab

15:00–15:20 *Automated Generation of Arabic Verb Conjugations with Multilingual Urdu Translation: An NLP Approach*
Haq Nawaz, Manal Elobaid, Ali Al-Laith and Saif Ullah

Coffee Break

16:00–16:20 *Evaluation of Large Language Models on Arabic Punctuation Prediction*
Asma Ali Al Wazrah, Afrah Altamimi, Hawra Aljasim, Waad Alshammari, Rawan Al-Matham, Omar Elnashar, Mohamed Amin and Abdulrahman AlOsaimy

16:20–16:40 *Evaluating RAG Pipelines for Arabic Lexical Information Retrieval: A Comparative Study of Embedding and Generation Models*
Raghad Al-Rasheed, Abdullah Al Muaddi, Hawra Aljasim, Rawan Al-Matham, Muneera Alhoshan, Asma Al Wazrah and Abdulrahman AlOsaimy

16:40–16:45 *Closing Remarks and Wrap-Up*
Dr Mo El-Haj