WiNLP 2022

The Sixth Widening NLP Workshop

Proceedings of the Workshop

December 7, 2022

The WiNLP organizers gratefully acknowledge the support from the following sponsors.



©2022 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 978-1-959429-28-9

Organizing Committee

Organizers

Shaily Bhatt, Google Research India
Sunipa Dev, Google Research
Bonaventure Dossou, Jacobs University Bremen and Mila Quebec AI Institute
Tirthankar Ghosal, Institute of Formal and Applied Linguistics, Charles University, Czech Republic
Hatem Haddad, iCompass, Tunisia
Haley M. Lepp, Wikimedia Foundation
Fatemehsadat Mireshghallah, University of California, San Diego
Surangika Ranathunga, University of Moratuwa
Xanda Schofield, Harvey Mudd College
Isidora Tourni, Boston University
Weijia Xu, University of Maryland

Program Committee

Chairs

Shaily Bhatt, Google Research Sunipa Dev, Google Research Bonaventure F. P. Dossou, Mila Tirthankar Ghosal, Institute of Formal and Applied Linguistics, Charles University Hatem Haddad, iCompass Haley Lepp, Educational Testing Service (ETS) Fatemehsadat Mireshghallah, UC San Diego Surangika Ranathunga, university of moratuwa Alexandra Schofield, Harvey Mudd College Isidora Tourni, Boston University

Program Committee

Gavin Abercrombie, Heriot Watt University Fernando Alva-manchego, Cardiff University Maria Antoniak, Allen Institute for Artificial Intelligence Tania Avgustinova, Saarland University Marco Baroni, ICREA Emily M. Bender, University of Washington Joachim Bingel, Hero I/S Chloé Braud, IRIT - CNRS - ANITI Jose Camacho-collados, Cardiff University Marie Candito, LLF (Univ. de Paris / CNRS) Kai-wei Chang, UCLA Zhuo Cheng, University of North Carolina at Charlotte Kenneth Church, Northeastern University Shay B. Cohen, University of Edinburgh Ryan Cotterell, ETH Zürich Nisansa De Silva, University of Moratuwa Rahul Divekar, Educational Testing Service Esin Durmus, Stanford University Ignatius Ezeani, Lancaster University Mareike Hartmann, Saarland University Julia Hirschberg, Columbia University in the City of New York Hitoshi Isahara, Toyohashi University of Technology Cassandra L. Jacobs, University at Buffalo Sneha Jha, Imperial College Neha Nayak Kennard, University of Massachusetts Amherst Huda Khayrallah, Microsoft Rebecca Knowles, National Research Council Canada Philipp Koehn, Johns Hopkins University Rik Koncel-kedziorski, Amazon Valia Kordoni, Humboldt-Universität zu Berlin Lun-wei Ku, Academia Sinica Jonathan Kummerfeld, University of Sydney Junhui Li, Soochow University, Suzhou

Adam Lopez, University of Edinburgh Arya D. Mccarthy, Johns Hopkins University Ryan Mcdonald, ASAPP Kourosh Meshgi, Riken AIP Max Müller-eberstein, IT University of Copenhagen Nikita Nangia, New York University Vincent Ng, University of Texas at Dallas Joakim Nivre, Research Institutes of Sweden Debora Nozza, Bocconi University Kemal Oflazer, Carnegie Mellon University Atul Kr. Ojha, Data Science Institute, Unit for Linguistic Data, University of Galway Eda Okur. Intel Labs Alexis Palmer, University of Colorado Boulder Siyao Peng, Georgetown University Valentina Pyatkin, Bar-Ilan University Dragomir Radev, Yale University Gabriela Ramirez-de-la-rosa, Universidad Autónoma Metropolitana Ellen Riloff, University of Utah Brian Roark, Google Inc. Leanne Rolston, LivePerson Inc. Frank Rudzicz, St Michael's Hospital; University of Toronto, Department of Computer Science Neville Ryant, Linguistic Data Consortium Sashank Santhanam, University of North Carolina at Charlotte/ Apple Brandon Sepulvado, NORC at the University of Chicago Ravi Shekhar, Queen Mary University of London Karin Sim Smith, Iconic Mark Steedman, University of Edinburgh Suzanne Stevenson, University of Toronto Jeniya Tabassum, Amazon Mokanarangan Thayaparan, University of Manchester Christoph Tillmann, IBM Research Sara Tonelli, FBK Elena Tutubalina, HSE University, Russia and Kazan Federal University, Russia and AIRI, Russia and Insilico Medicine Hong Kong, Hong Kong Aline Villavicencio, University of Sheffield, UK Bonnie Webber, University of Edinburgh Weijia Xu, University of Maryland Olga Zamaraeva, Universidade da Coruña Luke Zettlemoyer, University of Washington; Meta

Keynote Talk: Panel, Finding a cool job in NLP

Dr. Hussein AL-NATSHEH, Roshna Omer Abdulrahman, Kinda Altarbouch, Felix Laumann,

Erika Varis Doggett

N/A

Abstract: Panel: Finding a cool job in NLP

Bio: Dr. Hussein AL-NATSHEH

Dr AL-NATSHEH joined Beyond Limits in early 2021 and is a recognized AI leader in the Middle East, being a member of several program committees in NLP and a speaker in many AI regional events. He brings a total experience of 18+ years in digital transformation and technology business innovation with a focus on AI solutions. Prior to joining Beyond Limits, he was the Chief AI Officer at Mawdoo3 Limited, the world's most visited group of Arabic websites. He joined Mawdoo3 as an executive in 2016 after acquiring his data mining start-up where he founded its AI division that developed its state-of-theart Arabic NLP/NLU SaaS toolkit with the biggest data science team in the region at that time. There, he also co-founded 2 AI subsidiaries; the first Arabic voice digital assistant (Salma.ai), which is like "Siri" but understands Arabic dialects, and (Oalam.ai) which is like "Grammarly" but specialized in the Arabic language. Hussein holds a Ph.D. degree from the University of Lyon, France in Data Science, in addition to 2 European Master's degrees in the same field and a master's degree in computer engineering from Jordan. He made the later master's thesis in machine learning at the European Organization for Nuclear Research (CERN) in Switzerland. In 2009, he founded (Ciapple), a data mining and predictive analytics technology startup. His startup received several regional technology business innovation awards. Previously in 2007, he served as the technology commercialization program manager at Queen Rania Center for Entrepreneurship (QRCE) in Jordan the director of its innovation division (TRIZmatic) helping El Hassan Science City in Jordan (The higher umbrella of 4 organizations at that time; QRCE, PSUT, RSS, and HCST) in building its knowledge economy strategy for the country and establishing its national Technology Transfer Office (TTO).

Roshna Omer Abdulrahman

Roshna Omer Abdulrahman is an Information Management Associate at UNHCR Iraq. She has over five years of experience working as a software engineer. She has published research on Low resource NLP for Kurdish as a part of her master's degree and is trying to keep up with NLP outside of her day job. She has BSc in Computer Science and an MSc in Software Engineering. In her free time, you will find her taking a walk with an audiobook in her ears.

Kinda Altarbouch

Kinda Altarbouch co-founder and CEO at Lableb which develops digital services and Arabic language technology products. Kinda was listed in MIT Innovators Under 35 (MENA region) in 2020. She was chosen by the Arab Youth Center as a Young Arab Pioneer in 2017 and 2018. Also, She is a member of the Board of Advisory Council of Popular Science Magazine (Arabic Edition) since 2017 and a member of the Syrian Society for Scientific Research (SySSR) since 2014. She has served as a teaching assistant at Damascus University and scientific supervisor for the Robotic Club in the same faculty from 2015 to 2018. She holds a bachelor's degree in Computer Science and Artificial Intelligence from Damascus University, Syria.

Felix Laumann

Felix Laumann holds a PhD in Statistics from Imperial College London. He is the CEO at NeuralSpace, a platform to build Natural Language Processing (NLP) solutions in a no-code web interface and through simple APIs. NeuralSpace offers 80+ languages spoken in India, the Middle East and Africa, such as 7 Arabic dialects and 11 Indian languages.

Erika Varis Doggett

Erika is a Sr. Research Scientist at Disney Research Studios focusing on a variety of topics within machine learning, with special attention to computer vision, generative imagery, and natural language

processing techniques. Recent areas of interest include natural language generation, inpainting-based anomaly detection, neural style transfer, attention and saliency prediction, entity extraction and classification, speech recognition, machine translation, and other sequencing models. Prior to Disney Research, Erika managed an applied RnD artificial intelligence/machine learning team within the StudioLAB at Walt Disney Studios. Erika received my Ph.D. in Linguistics from the University of Southern California and lectured for two years at the University of Nevada Reno before transitioning to industry in NLP for social media data.

Keynote Talk: N/A

Prof. Houda Bouamor Carnegie Mellon University, Qatar

Abstract:

Bio: Houda Bouamor is an Assistant Teaching Professor in the Information Systems department at CMU-Q. She is also part of the Natural Language Processing Lab and works on Arabic NLP with Nizar Habash and Kemal Oflazer. Her Ph.D. is from Paris-Sud University, France, where she worked on paraphrasing in the framework of Computational Linguistics, advised by Anne Vilnat and Aurélien Max. Her M.Sc. is in Computer Science from the Paris-Est Marne-La-Vallée University and her bachelors degree in Computer Science from the University of Manouba, Tunisia. She worked on different projects including resolving different NLP issues. Her main research interest revolves around Statistical Machine Translation.

Keynote Talk: Being a researcher in Arabic NLP

Haithem Afli, Nedjma Ousidhoum, Mustafa Jarrar

N/A

Abstract: Panel: Being a researcher in Arabic NLP

Bio: Haithem Afli

Dr Haithem Afli is a leading expert in Natural Language Processing and Artificial Intelligence and a rising star in the area of applied Artificial Intelligence in Healthcare, Life-science, and Fintech. Dr Afli is lecturing AI within the Computer Science Department of Munster Technological University (MTU) in Ireland and leading the MTU Human Centred AI Research Group, HAI. Dr Haithem Afli is Science Foundation Ireland funded investigator at ADAPT Centre where he is responsible for leading one of the three challenge leads of the DCT strand and he is a member of the ADAPT Executive Management Committee, representing MTU. Haithem's research interest is primarily focuses in the areas of Machine Translation, Sentiment Analysis, Natural Language Processing and Machine Learning. Dr Afli has published over 70 peer reviewed research articles and is serving as Editor, Program Chair, Program Committee Member and advisor in many international research conferences and journals. Dr Haithem Afli was a key contributor to the recently established MTU Masters programmes on Computational Biology and Artificial Intelligence. As an academic researcher, Haithem is keen to commercialise his research with industry partnerships and is actively involved in managing academia-industry partnership projects including co-founding LinguAnalysis.ai.

Nedjma Ousidhoum

Nedjma Ousidhoum is a Postdoctoral Research Associate at the University of Cambridge. Nedjma works on automated fact-checking for journalists. Prior to this, Nedjma was a PhD student at the Hong Kong University of Science and Technology (HKUST) where they worked on toxic content detection in NLP. Nedjma is also one of the co-organisers of the Afrisenti Semeval shared task, we have built sentiment analysis datasets for African language including Algerian Arabic.

Mustafa Jarrar

A full professor of Computer Science, and director of the PhD program at Birzeit University in Palestine. He was a Fulbright visiting professor at the University at Buffalo in the USA (2016-2017), a Marie Curie fellow at the University of Cyprus (2007-2009), and a senior research scientist at Vrije Universiteit Brussel (1999-2007), where he completed his Masters (2000) and PhD (early 2005). He has won several prestigious awards including, among others, the Shoman Arab Researchers Award in Technology, Mohammed Bin Rashid Award for the Arabic Language, and Google Faculty Research Award. Prof. Jarrar has published 80+ articles in the areas of Natural Language Processing, Ontology Engineering, Semantic Web, and Graph Databases. He chaired 23 international workshops, a PC member of 150+ journals/conferences, a coordinator/manager of 25 large EU projects, a full member of the IFIP2.6 on Database Semantics, the IFIP2.12 on Web Semantics, and the UN ESCWA Technology Centre Board of Governors, among others. Prof. Jarrar is also the founder of both Sina Institute for Knowledge Engineering and Language Technologies, and the Palestinian eGovernment Academy, and advisor at the Palestinian ministry of Telecom and IT for e-government topics, where he developed and chaired the Palestinian e-Government Interoperability Framework (called Zinnar).

Keynote Talk: Fireside chat, Dr. Klaus Zechner

Dr. Klaus Zechner

Educational Testing Service

Abstract: Fireside chat: Dr. Klaus Zechner

Bio: Klaus Zechner obtained his Ph.D. in Language Technologies from Carnegie Mellon University, and joined Educational Testing service (ETS) as a research scientist in 2002. He currently works in the Artificial Intelligence Research Lab at ETS on automated scoring of speech. He has been pioneering research and development of technologies for automated scoring of non-native speech at ETS for over 20 years, leading large RnD projects dedicated to the continuous improvement of these technologies. He holds more than 20 patents related to SpeechRater®, an automated speech scoring system developed at ETS under his leadership. SpeechRater is currently used operationally as contributory scoring system, along with human raters, for the TOEFL® iBT Speaking assessment. Klaus Zechner has authored more than 100 publications, most of which appeared in peer-reviewed journals, conference and workshop proceedings. He also edited the first comprehensive volume on automated speaking assessment (published by Routledge in 2019).

Table of Contents

Program

Wednesday, December 7, 2022

08:45 - 09:45 Poster Session

Lexical methods for bias exploration from a Latin American perspective Luciana Benotti, Laura Alonso Alemany and Lucia Gonzalez

Low Resourced Multilingual Neural Machine Translation for Ometo-English

Improving neural machine translation for low-resource languages using related language resources Atnafu Lambebo Tonja

Transformer Based Amharic Headline Generation using Sub-word2Vec Representation Mahlet Taye, Yaregal Assabie and Abebaw Eshetu

DistillEmb: Distilling word embeddings via contrastive learning Amanuel Mersha and Stephen Wu

Perturbation-based Active Learning for Question Answering Fan Luo and Mihai Surdeanu

Short Comparative Analysis on Pretrained BART and RoBERTa in Detecting Hate Speech on YouTube and Reddit Platforms Dinuja Perera and Nisansa De Silva

Synthesis and Evaluation of a Domain-specific Large Data Set for Dungeons & Dragons

Akila Peiris and Nisansa De Silva

Amharic Fake News Detection on Social Media Using Feature Fusion

MBTI Personality Prediction Approach on Persian Twitter

Samin Fatehi, Zahra Anvarian, Yasmin Madani, Mohammadjavad Mehditabar and Sauleh Eetemadi

Afaan Oromo Hate Speech Detection and Classification on Social Media

Amharic-Kistanigna Bi-directional Machine Translation using Deep Learning Mengistu Negia and Rahel Tamiru

Exploiting Available Resources for the Training of Manglish Language Models Meisin Lee and Lay-ki Soon

Towards a general purpose machine translation system for Sranantongo Just Zwennicker and David Stap

Transfer Learning and Word Sense Disambiguation for Low-resource Language, the Case of Amharic Neima Ahmed and Million Meshesha

Boosting the Performance of Gender Subspace in Domain-Specific Gender Bias Analysis

Yanqin Tan, Cassandra L. Jacobs, Mimi Zhang, Marvin Thielk and Yi Chu

Contextual Embeddings Can Distinguish Homonymy from Polysemy in a Human-Like Way

The BERT Walked Down the Garden Path Assigned Semantic Roles

The Effect of Normalization for Bi-directional Amharic-English Neural Machine Translation

Automatic Speech Recognition using Self-Supervised Learning Approach

Prosody Based Automatic Speech segmentation for Amharic

Before and beyond MeToo, Measuring changes in power and agency within sexual abuse news stories over time Gyulim Kang and Hope Schroeder

Challenges of Amharic Hate Speech Data Annotation Using Yandex Toloka Crowdsourcing Platform

Abinew Ali Ayele, Tadesse Destaw Belay, Seid Muhie Yimam, Skadi Dinter, Tesfa Asfaw and Chris Biemann

Question Answering Classification for Amharic Social Media Community Based Questions

Tadesse Destaw Belay, Seid Muhie Yimam, Abinew Ali Ayele and Chris Biemann

On Genitalia, Reproduction and Pleasure: Biases in the Representation of Sexes Sadhi Vornberger and Peter Bourgonje

An Unsupervised Learning Approach for Categorising Research Proposals and Recommending Papers Annie Lee, Mariia Ponomarenko and Peiyuan Zhou

Detecting Depression on Twitter with a Time-Aware Multimodal Transformer

Detecting Adverse Drug Events from social media: A brief literature review Imane Guellil, Nidhaleddine Chenni, Yousra Berrachedi, Massinissa Abboud, Jinge Wu, Beatrice Alex and Honghan Wu

Data Augmentation to Address the Out-of-Vocabulary Problem in Low-Resource Sinhala-English Neural Machine Translation Aloka Fernando and Surangika Ranathunga

Evaluating Gender Bias in Pre-trained Indic Language Models Neeraja Kirtane, V Manushree and Aditya Kane

Tackling Gender Microaggressions in Hindi Text Vishakha Agrawal

ParsVQA-Caps: A Benchmark for Visual Question Answering and Image Captioning in Persian

Shaghayegh Mobasher, Ghazal Zamaninejad, Maryam Hashemi, Melika Nobakhtian and Sauleh Eetemadi

HERDPhobia: A Dataset for Hate Speech Detection against Fulani Herdsmen in Nigeria Saminu Aliyu, Gregory Wajiga, Muhammad Murtala, Shamsuddeen Muhammad, Idris Abdulmumin and Ibrahim Ahmad Experiments on Generalizability of BERTopic on Short Text Muriël De Groot, Mohammad Aliannejadi and Marcel Haas Domain-Specific Lexicon-Based Sentiment Analysis using Contextual Shifter Patterns Shamsuddeen Muhammad and Idris Abdulmumin Generate Answer to Visual Questions with Pre-trained Vision-and-Language **Embeddings** Hadi Sheikhi, Maryam Hashemi and Sauleh Eetemadi 09:45 - 10:00 Welcome 10:00 - 10:50 Panel: Finding a cool job in NLP 10:50 - 11:00 Coffee Break 11:00 - 11:50 Keynote and Q&A: Prof. Houda Bouamor Break (Noon Prayer)

- 11:50 12:30
- 12:30 13:30 Sponsors' Lunch
- 13:30 14:20 Panel: Being a researcher in Arabic NLP
- 14:20 14:30 Coffee Break
- 14:30 15:10 Second Poster Session

Lexical methods for bias exploration from a Latin American perspective Luciana Benotti, Laura Alonso Alemany and Lucia Gonzalez

Low Resourced Multilingual Neural Machine Translation for Ometo-English

Improving neural machine translation for low-resource languages using related language resources Atnafu Lambebo Tonja

Transformer Based Amharic Headline Generation using Sub-word2Vec Representation Mahlet Taye, Yaregal Assabie and Abebaw Eshetu

DistillEmb: Distilling word embeddings via contrastive learning Amanuel Mersha and Stephen Wu

Perturbation-based Active Learning for Question Answering Fan Luo and Mihai Surdeanu

Short Comparative Analysis on Pretrained BART and RoBERTa in Detecting Hate Speech on YouTube and Reddit Platforms Dinuja Perera and Nisansa De Silva

Synthesis and Evaluation of a Domain-specific Large Data Set for Dungeons & Dragons Akila Peiris and Nisansa De Silva

Amharic Fake News Detection on Social Media Using Feature Fusion

MBTI Personality Prediction Approach on Persian Twitter

Samin Fatehi, Zahra Anvarian, Yasmin Madani, Mohammadjavad Mehditabar and Sauleh Eetemadi

Afaan Oromo Hate Speech Detection and Classification on Social Media

Amharic-Kistanigna Bi-directional Machine Translation using Deep Learning Mengistu Negia and Rahel Tamiru

Exploiting Available Resources for the Training of Manglish Language Models Meisin Lee and Lay-ki Soon

Towards a general purpose machine translation system for Sranantongo Just Zwennicker and David Stap

Transfer Learning and Word Sense Disambiguation for Low-resource Language, the Case of Amharic Neima Ahmed and Million Meshesha

Boosting the Performance of Gender Subspace in Domain-Specific Gender Bias Analysis

Yanqin Tan, Cassandra L. Jacobs, Mimi Zhang, Marvin Thielk and Yi Chu

Contextual Embeddings Can Distinguish Homonymy from Polysemy in a Human-Like Way

The BERT Walked Down the Garden Path Assigned Semantic Roles

The Effect of Normalization for Bi-directional Amharic-English Neural Machine Translation

Automatic Speech Recognition using Self-Supervised Learning Approach

Prosody Based Automatic Speech segmentation for Amharic

Before and beyond MeToo, Measuring changes in power and agency within sexual abuse news stories over time Gyulim Kang and Hope Schroeder

Challenges of Amharic Hate Speech Data Annotation Using Yandex Toloka Crowdsourcing Platform

Abinew Ali Ayele, Tadesse Destaw Belay, Seid Muhie Yimam, Skadi Dinter, Tesfa Asfaw and Chris Biemann

Question Answering Classification for Amharic Social Media Community Based Questions

Tadesse Destaw Belay, Seid Muhie Yimam, Abinew Ali Ayele and Chris Biemann

On Genitalia, Reproduction and Pleasure: Biases in the Representation of Sexes Sadhi Vornberger and Peter Bourgonje

An Unsupervised Learning Approach for Categorising Research Proposals and Recommending Papers

Annie Lee, Mariia Ponomarenko and Peiyuan Zhou

Detecting Depression on Twitter with a Time-Aware Multimodal Transformer

Detecting Adverse Drug Events from social media: A brief literature review Imane Guellil, Nidhaleddine Chenni, Yousra Berrachedi, Massinissa Abboud, Jinge Wu, Beatrice Alex and Honghan Wu

Data Augmentation to Address the Out-of-Vocabulary Problem in Low-Resource Sinhala-English Neural Machine Translation Aloka Fernando and Surangika Ranathunga

Evaluating Gender Bias in Pre-trained Indic Language Models Neeraja Kirtane, V Manushree and Aditya Kane

Tackling Gender Microaggressions in Hindi Text Vishakha Agrawal

ParsVQA-Caps: A Benchmark for Visual Question Answering and Image Captioning in Persian

Shaghayegh Mobasher, Ghazal Zamaninejad, Maryam Hashemi, Melika Nobakhtian and Sauleh Eetemadi

HERDPhobia: A Dataset for Hate Speech Detection against Fulani Herdsmen in Nigeria

Saminu Aliyu, Gregory Wajiga, Muhammad Murtala, Shamsuddeen Muhammad, Idris Abdulmumin and Ibrahim Ahmad

Experiments on Generalizability of BERTopic on Short Text Muriël De Groot, Mohammad Aliannejadi and Marcel Haas

Domain-Specific Lexicon-Based Sentiment Analysis using Contextual Shifter Patterns Shamsuddeen Muhammad and Idris Abdulmumin

Generate Answer to Visual Questions with Pre-trained Vision-and-Language Embeddings Hadi Sheikhi, Maryam Hashemi and Sauleh Eetemadi

- 15:10 15:30 Break (Afternoon Prayer)
- 15:30 16:20 Fireside chat: Dr. Klaus Zechner
- 16:30 16:50 Closing Session

Author Index