

W-NUT 2026

**The 11th Workshop on Natural User-generated Text (W-NUT
2026)**

Proceedings of the Workshop

October 28, 2026

©2026 Association for Computational Linguistics

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
317 Sidney Baker St. S
Suite 400 - 134
Kerrville, TX 78028
USA
Tel: +1-855-225-1962
acl@aclweb.org

ISBN 979-8-89176-470-5

Introduction

The 11th Workshop on Natural User-generated Text (W-NUT) addresses the evolving challenges at the intersection of noisy, user-generated written content and modern Large Language Models (LLMs). As LLMs are increasingly deployed in real-world applications, they must contend with the messy, informal, and diverse nature of human communication found on social media, forums, messaging platforms, and other user-generated sources. Recent research shows that while LLMs demonstrate impressive capabilities across many NLP tasks, their robustness to noisy and unstructured text remains a critical challenge. This workshop bridges traditional NLP research on noisy text with contemporary challenges in building, evaluating, and deploying robust LLMs that can handle the complexities of natural language as it actually appears in the wild. We have received a total of 51 main workshop submissions, including ARR, of which 25 are included in the proceedings. The workshop will be held in hybrid in-person and virtual modes. We have three invited speakers: Hinrich Schütze, Haneul Yoo, and Maria Antoniak, who have generously agreed to share their ongoing research work. We are very thankful to have them in our workshop. We would like to thank the Program Committee members who reviewed the papers, as well as all of the workshop participants for submitting their work.

Organizing Committee

General Chair

JinYeong Bak, Sungkyunkwan University

Program Chair

Hyeju Jang, Indiana University Indianapolis

Rob van der Goot, IT University of Copenhagen

Alan Ramponi, Fondazione Bruno Kessler

Co-Organizers

Weerayut Buaphet, Vidyasirimedhi Institute of Science and Technology

Wei Xu, Georgia Institute of Technology

Alan Ritter, Georgia Institute of Technology

Program Committee

Reviewers

Piyush Agarwal, Khandaker Mamun Ahmed, Onat Dogan Akca, Emily Allaway, Avinash Goutham Aluguvelly, Antonios Anastasopoulos, Rajeel Ahmad Ansari, Maria Antoniak, Noor Mairukh Khan Arnob

Sivaraaman Balakrishnan, Cooper Ballard, Raghvi Baloni, Ankit Bhattacharjee, Kritiapriya Bhau-mik, Eduardo Blanco

Nathanael Chambers, Xuanjing Chen, Seongwoo Choi, Manuel R. Ciosici, Danilo Croce

Souvik Das, Anay Abhijit Dombé, Jane Arleth dela Cruz

Micha Elsner

Advitya Gemawat

Anna-Carolina Haensch, Marko Haralovic, Muhammad Mubashir Hassan, Monica Hegde, Mar-kus J. Hofmann, YeongJun Hwang

Diana Inkpen

Aditya Jain, Eshaan Jain, Yangfeng Ji, Chao Jiang, Ishan Jindal

Tomoyuki Kajiwar, Hari Kang, HyunJin Kim, Sai Tulasi Kolapudi

Jaehyeok Lee, Xinyu Li, Lucy H. Lin, Danni Liu, Nikola Ljubešić, Elena Lloret

Bolei Ma, Julina Maharjan, Bitan Majumder, Ruslana Margova, Tanmoy Mazumder, Alexander Mehler, Julia Mendelsohn, Yasuhide Miura, Manuel Montes

Günter Neumann, Vincent Ng, Phuong-Anh Nguyen-Le

Naoki Otani

Rajveer Singh Pall, Gianmarco Pappacoda, Suhyeon Park, Yuval Pinter, Alistair Plum, Alba Pérez-Montero

Nelu D. Radpour, Saman Rahbar, Shreyash Rawat, Federico Ruggeri

Iñaki San Vicente, Danae Sanchez Villegas, Albina Sarymsakova, Bogdan Savelyev, Nils Schwa-ger, Mirco Schönfeld, Anirban Sen, Ashfaq Ali Shafin, Vishal Shah, Prakriti Shankar Shetty, Vincent Siddons, Tanvir Ahmed Sijan, Abhai Pratap Singh, Amr Sobhy, Sangmin Song, Richard Sproat, Man-Ling Sung, Fayeel Jeelani Syed

Mariët Theune, Yashdeep Thorat

Sai P Vallurupalli

ChuanHsin Wang, Claire Wang

Sameer Yadav, Linsey C. Yang

Mike Zhang, Ted Zhang, Zihao Zheng, Xiliang Zhu

Keynote Talk
**Language Identification in the (Pretraining) Data Processing
Pipeline – State of the Art and Open Problems**

Hinrich Schütze
University of Munich
2026-10-28 09:30:00 – Room: Room G15-16

Abstract: TBA

Bio: Hinrich Schuetze is Professor for Computational Linguistics and co-director of the Center for Information and Language Processing at the University of Munich (LMU Munich). He received his PhD from Stanford in 1995 for his early research on embeddings (ACL Test-of-Time Paper Award). He then worked on natural language processing and information retrieval technology at Xerox PARC, at several Silicon Valley startups and at Google 1995-2004 and 2008/9. Hinrich is a coauthor of Foundations of Statistical Natural Language Processing (with Chris Manning) and Introduction to Information Retrieval (with Chris Manning and Prabhakar Raghavan). He was awarded a European Research Council Advanced Grant in 2017. Hinrich served as president of the Association for Computational Linguistics in 2020. He is a member of Academia Europaea and a fellow of ACL, ELLIS, ELIZA and HessianAI.

Keynote Talk

Enhancing Multilingual LLMs with Code-Switching

Haneul Yoo

New York University

2026-10-28 14:00:00 – Room: Room G15-16

Abstract: Despite recent advances, large language models (LLMs) exhibit uneven performance across languages and often struggle with code-switching—the natural alternation between multiple languages. While code-switching is a common and meaningful mode of human communication, LLMs remain brittle to code-switching inputs, treating them as noise. This talk reframes code-switching as a valuable linguistic signal, exploring how it can be leveraged across the LLM lifecycle, from evaluation, to training, to inference. I will begin with output language alignment (OLA), highlighting systematic failures to respond in the expected language in code-switching contexts. I will also introduce code-switching red-teaming (CSRT), which reveals vulnerabilities in safety and multilingual understanding. I will then present code-switching curriculum learning (CSCL), a data-efficient approach to cross-lingual transfer inspired by human second-language learning. Finally, I will introduce code-switching in-context learning (CSICL), which uses gradual transitions from a target language to English to improve multilingual generalization without additional training.

Bio: Haneul Yoo is an Assistant Professor/Faculty Fellow in the Department of Computer Science, Courant Institute School of Mathematics, Computing, and Data Science at New York University. She is also a Research Fellow with the Social Innovation and Special Projects Team in the Division of Policy, Evaluation and Training, Department of Peace Operations at the United Nations. She received her PhD from the School of Computing at KAIST. With a background in machine learning (ML) and natural language processing (NLP), she has worked on (1) developing language models and methods to improve generalization across languages, (2) designing resources and evaluation frameworks for underrepresented languages and cultures, and (3) applying AI in high-stakes, real-world domains for social good. She has published papers in major NLP conferences, including ACL, EMNLP, and NAACL, and has been named a Rising Star at KAIST exploreCSR supported by Google. She has served as a Social Chair at NeurIPS 2026 and co-organized the Multilingual and Equitable Language Technologies (MELT) workshop at COLM 2025, the Women in Machine Learning (WiML) symposium at ICML 2026, and the Generative AI for the World (GenAI4World) workshop at COLM 2026.

Keynote Talk
**When Noise Is the Point: What Humanistic Tasks Reveal
About Our Data Pipelines**

Maria Antoniak
University of Colorado Boulder
2026-10-28 16:00:00 – Room: Room G15-16

Abstract: In computational studies of humanistic tasks such as storytelling and poetry, we often encounter datasets whose features frustrate standard NLP pipelines. In this talk, I'll draw on three settings where my collaborators and I have run into this friction. First, I'll discuss user–chatbot conversations like WildChat, where creative requests are long, detailed, and frequently revised, making these logs a rich but also complicated view of how people write with chatbots. Second, I'll describe our development of a fine-grained narrative annotation scheme for very large-scale, heterogeneous web text in pretraining data, where operationalizing a humanities concept is complicated by data diversity. Third, I'll turn to poetry, where the linearization systems used to build training corpora routinely flatten or delete formatting that poets use deliberately. Across these cases, I'll argue that what counts as noise depends on who is reading and for what purpose, and that taking humanistic perspectives seriously can help us build data practices, annotation schemes, and evaluations that better reflect how people actually use language.

Bio: Maria Antoniak is an Assistant Professor of Computer Science at the University of Colorado Boulder where she is also affiliated with the Department of Information Science and a member of the Boulder NLP Group. Previously, she was a Young Investigator at the Allen Institute for AI and a postdoc at the Pioneer Centre for AI at the University of Copenhagen. She completed her PhD in Information Science at Cornell University and her master's degree in Computational Linguistics at the University of Washington. She has also spent time at places like ETH Zürich, Microsoft Research FATE, Twitter Cortex, and Facebook Core Data Science. Her research is in natural language processing and cultural analytics.

Table of Contents

<i>A Precision-First Whitespace Restoration Pipeline for Noisy Urdu Tafseer Text</i> Muhammad Mubashir Hassan	1
<i>Are More Emotion Theories Better? Joint Prompting Improves Emotion Categorization Robustness</i> Albina Sarymsakova and Roman Klinger	10
<i>Beyond Clean Text: Evaluating Encoder and Decoder Robustness for Bangla Event Detection in Noisy Text</i> Tanvir Ahmed Sijan, S M Golam Rifat, Nayeemul Islam and Md Musfique Anwar	26
<i>CrisisKD: Five-Stage Knowledge Distillation for Aspect-Level Sentiment and Emotion Analysis in Crisis Discourse</i> Marko Haralovic, Onat Dogan Akca, Salih Eren Yüçetürk, Minsi Li and Mariët Theune	49
<i>Discourse-Structural Noise in LLM-Generated Explanations</i> Raghvi Baloni and Achim Rettinger	65
<i>Do LLMs Give Consistent Opinions? Evaluating Response Reliability Under Varying Likert-Scale Formulations in Survey-Style MCQA</i> Martin Kandlinger, Bolei Ma, Anna-Carolina Haensch and Matthias Aßenmacher	72
<i>Evaluating Token Probabilities as Confidence Estimators in Human Behavior Simulations: A Case Study in Luxembourgish Next Like Prediction</i> Nils Schwager, Simon Münker, Alistair Plum and Achim Rettinger	89
<i>From Errors to Precision: A Taxonomy-Guided Approach for LLM Flood Cause Extraction</i> Jane Arleth dela Cruz, Iris Hendrickx and Martha Larson	101
<i>GeNERscore: Semantic Alignment Metrics for Generative Named Entity Recognition</i> Gianmarco Pappacoda, Federico Ruggeri and Paolo Torrioni	117
<i>How Noisy is Automatic Annotation? Characterizing Label Noise from Generative Models in Multi-Dimensional Spanish Subjectivity Annotation</i> Alba Pérez-Montero, Elena Lloret and Paloma Moreda Pozo	141
<i>Increasing the Robustness of the Fine-tuned Multilingual Machine-Generated Text Detectors</i> Dominik Macko, Robert Moro and Ivan Srba	154
<i>LLMs are Easily Startled: Fragility Testing with Lookalike Character Noise</i> Peter deJonge and Nathanael Chambers	167
<i>Multi-author style detection with stylistic embeddings</i> Man-Ling Sung, Chak-Fai Li, Damianos Karakos, Frederick E. Bernardin III and Brian Ulicny	182
<i>NERQual: Evaluating the Robustness of Named Entity Recognition Models to Data Quality Issues</i> Denis Trautner and Philippe Thomas	195
<i>Not All Duplicates Are Coordination: Generic vs. Non-Generic Duplicate Campaigns in Information Operations</i> Ashfaq Ali Shafin and Khandaker Mamun Ahmed	210
<i>Personalized Emotion Analysis via Prompting Large Language Models with Writer Information</i> Kentaro Hanafusa, Daisuke Maekawa and Tomoyuki Kajiware	217

<i>Predicting Steering Vectors and Adapter Weights for Few-Shot Author-Style Transfer</i>	
Leonard Popp, Danni Liu, Supriti Sinhamahapatra and Jan Niehues	227
<i>QueryDiagramEval: Evaluating and Benchmarking LLM Diagram Generation under Noisy Conditions</i>	
Ted Zhang, Souvik Das, Kai Ni, Aditya Agrawal, Xin Guo and Denis Kochedykov	244
<i>Robust Dual-Signal Fusion: Hybrid Neuro-Symbolic Gating with Compressed Chain-of-Thought Refinement for Irony Detection in Social Media Texts</i>	
Ankit Bhattacharjee and Kriti Priya Bhaumik	266
<i>RSPC: A Benchmark for Modeling Stress and Psychiatric Conditions in Digitally Mediated Relationships using Psychiatrist Annotations</i>	
Parmitha Vangapandu, Sai Ganesh M, Sathwik Narkedimilli, Msvpj Sathvik, Timothy Liu, Simon See and Johannes C Eichstaedt	286
<i>Spoiler Court: Interpretable Spoiler Detection via Adversarial Multi-Agent LLM Judgement</i>	
Sangmin Song, Hari Kang, Suhyeon Park, Seongwoo Choi, Sungkyun Im and Sangoh Lee ..	313
<i>The Curse of Multilinguality in Lexical Normalization</i>	
Saman Rahbar	325
<i>Topic Matching in the Wild: Benchmark and Lessons from Real-World ASR Transcripts</i>	
Saman Rahbar, Xiliang Zhu, Irvin Cardoza and David Rossouw	332
<i>TurEngMix: A Text Corpus and Benchmark for Turkish-English Code-Mixed Language Identification and Named Entity Recognition</i>	
Ilayda Dogan, Phuong-Anh Nguyen-Le and Julia Mendelsohn	341
<i>Where Does Noise Robustness Live? Evidence from Ablation, Role Swaps, and Cross-Lingual Transfer in Code-Mixed Banking NLU</i>	
Piyush Agarwal	356

Program

Wednesday, October 28, 2026

09:00 - 09:30	<i>Opening Remarks</i>
09:30 - 10:30	<i>Invited Talk - Hinrich Schütze</i>
10:30 - 11:00	<i>Coffee Break & Networking</i>
11:00 - 12:30	<i>Presentation - Oral</i>
12:30 - 14:00	<i>Lunch and Networking</i>
14:00 - 15:00	<i>Invited Talk - Haneul Yoo</i>
15:00 - 16:00	<i>Presentation - Poster</i>
16:00 - 17:00	<i>Invited Talk - Maria Antoniak</i>
17:00 - 17:30	<i>Closing & Next Steps</i>