W-NUT 2025

The Tenth Workshop on Noisy and User-generated Text (W-NUT 2025)

Proceedings of the Workshop

May 3, 2025

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 979-8-89176-232-9

Introduction

The W-NUT 2025 workshop focuses on a core set of natural language processing tasks on top of noisy and user-generated text, such as those found on social media, web forums and online reviews. The internet has democratized content creation leading to an explosion of informal user-generated text, publicly available in electronic format, motivating the need for NLP on noisy text to enable new data analytics applications. We have received a total of 18 main workshop submissions, of which 16 are included in the proceedings. The workshop will be held in hybrid in-person and virtual modes. We have two invited speakers: Su Lin Blodgett and Verena Blaschke, 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

Co-Organizers

Rob van der Goot, IT University of Copenhagen
Weerayut Buaphet, Vidyasirimedhi Institute of Science and Technology
Alan Ramponi, Fondazione Bruno Kessler
Wei Xu, Georgia Institute of Technology
Alan Ritter, Georgia Institute of Technology

Program Committee

Reviewers

Sweta Agrawal, Hamed Alhoori, Emily Allaway, Antonios Anastasopoulos, Maria Antoniak

Eduardo Blanco

Tommaso Caselli, Paul Cook, Danilo Croce

Micha Elsner

Yoshinari Fujinuma

YeongJun Hwang, Mika Hämäläinen

Kokil Jaidka, Aditya Jain, Chao Jiang, Ishan Jindal

HyunJin Kim, Suyoung Kim, Sachin Kumar

Jaehyeok Lee, Jing Li, Lucy H. Lin, Nikola Ljubešić

Yasuhide Miura, Manuel Montes, W. Graham Mueller

Günter Neumann, Vincent Ng

Naoki Otani

Rahul Raja, Alan Ramponi, Shubhashis Roy Dipta

Iñaki San Vicente, Danae Sanchez Villegas, H. Schwartz, Mirco Schönfeld, Vishal Shah, Vincent Siddons, Dan Simonson, Abhai Pratap Singh, Andreas Spitz

Joel R. Tetreault

Sai P Vallurupalli, Daniel Varab

Xiaojun Wan, Dustin Wright

Mike Zhang

Keynote Talk

Beyond "noisy" text: How (and why) to process dialect data

Verena Blaschke
LMU Munich & MCML
2025-05-03 09:30:00 – Room: **25 - Navajo/23 - Nambe**

Abstract: Processing data from non-standard dialects links two lines of research: creating NLP tools that are robust to "noisy" inputs, and extending the coverage of NLP tools to underserved language communities. In this talk, I will describe ways in which processing dialect data differs from processing standard-language data, and discuss some of the current challenges in dialect NLP research. For instance, I will talk about strategies to mitigate the effect of infelicitous subword tokenization caused by ad-hoc pronunciation spellings. Additionally, I argue that we should not only consider *how* to tackle dialectal variation in NLP, but also *why*. To this end, I will highlight perspectives of some dialect speaker communities on which language technologies should (or should not) be able to process or produce dialectal in- or output.

Bio: Verena Blaschke is a final-year PhD student at LMU Munich. She currently researches NLP for non-standard dialects and other low-resource language varieties, investigating how robust language models are towards language variation (and how to make them more robust). Her research is supervised by Barbara Plank and co-supervised by Hinrich Schütze. She also completed a research internship at Apple where she worked on multilingual NLP, and she previously developed software for machine-assisted historical linguistics at the University of Tübingen.

Keynote Talk

What Can We Learn from Perspectives on Noisy User-Generated Text?

Su Lin Blodgett
Microsoft Research Montréal
2025-05-03 16:00:00 – Room: **25 - Navajo/23 - Nambe**

Abstract: As language technologies become increasingly ubiquitous, research has shown that they struggle with real-world language variation and use. How can we expand the set of perspectives that inform our (and thus our technologies') engagement with such variation and use, and what can we learn by doing so? First, I will describe work on minoritized language varieties: building on work using quantitative methods to illustrate technologies' poor performance for such varieties, in this work we interview speakers of African American Language to better understand their experiences with language technologies and the impacts on them when technologies fail. I will discuss what this means for how we might design and assess language technologies to handle language variation, including the limits of quantitative methods for understanding people's experiences. Second, I will discuss disagreement in people's expectations and preferences—as technologies are increasingly designed to adapt to language variation, how do people think they should behave? I will describe work on natural language generation systems showing that people's expectations can vary widely, highlighting the importance of taking into account people's complex beliefs about language and technology, and raising questions about how to decide what constitute desirable system behaviors, when engaging with real-world language variation and use.

Bio: Su Lin Blodgett is a researcher in the Fairness, Accountability, Transparency, and Ethics (FATE) group at Microsoft Research Montréal. Her research examines the ethical and social implications of language technologies, focusing on the complexities of language and language technologies in their social contexts, and on supporting NLP practitioners in their ethical work. She completed her Ph.D. in computer science at the University of Massachusetts Amherst, where she was supported by the NSF Graduate Research Fellowship, and has been named as one of the 2022 100 Brilliant Women in AI Ethics.

Table of Contents

Program

Saturday, May 3, 2025

09:15 - 09:30 *Opening Remarks*

09:30 - 10:30 *Invited Talk - Verena Blaschke*

10:30 - 11:00 *Break*

11:00 - 12:30 *Presentation - Oral*

12:30 - 14:00 *Lunch and Networking*

14:00 - 15:30 *Presentation - Poster*

15:30 - 16:00 *Break*

16:00 - 17:00 *Invited Talk - Su Lin Blodgett*

17:00 - 17:30 *Best Paper Presentation*

17:30 - 17:40 *Closing*