

# Wei Xu

✉ wei.xu@cc.gatech.edu   🌐 <https://cocoxu.github.io/>   📍 CODA Tech Square, Atlanta, GA, 30308 U.S.A.  
✉ @cocoweixu   📄 <https://scholar.google.com/citations?user=Bf0dG-oAAAAJ>

---

<b>RESEARCH</b>	Natural Language Processing, Machine Learning, Social Media	
<b>CITIZENSHIP</b>	United States	
<b>ACADEMIC APPOINTMENTS</b>	<b>Associate Professor, Georgia Institute of Technology</b> , Atlanta, GA <i>College of Computing, School of Interactive Computing</i>	Aug 2023 – Present
	<b>Assistant Professor, Georgia Institute of Technology</b> , Atlanta, GA <i>College of Computing, School of Interactive Computing</i>	Aug 2020 – July 2023
	<b>Assistant Professor, The Ohio State University</b> , Columbus, OH <i>Department of Computer Science and Engineering</i>	Aug 2016 – July 2020
	<b>Visiting Faculty, Carnegie Mellon University</b> , Pittsburgh, PA <i>Language Technologies Institute</i> (Host: Graham Neubig)	Summer 2019
	<b>Postdoctoral Researcher, University of Pennsylvania</b> , Philadelphia, PA <i>Computer Information and Science Department</i> (Advisor: Chris Callison-Burch)	Feb 2014 – Aug 2016
<b>EDUCATION</b>	<b>Ph.D. in Computer Science</b> , <b>New York University</b> , New York, NY Advisor: Ralph Grishman; Committee: Satoshi Sekine, Ernest Davis, Bill Dolan (Microsoft Research), Luke Zettlemoyer (University of Washington / Meta)	2014
	<b>B.S./M.S. in Computer Science</b> , <b>Tsinghua University</b> , Beijing, CHINA	2004/2007
<b>SELECTED AWARDS</b>	<b>Best Social Impact Award, ACL</b> , 2024 <b>Best Paper Award Honorable Mention, ACL</b> , 2023 <b>NSF CAREER Award</b> , 2022 <b>NSF CRII Award</b> , 2018 <b>Best Paper Award, COLING</b> , 2018 <b>Criteo Faculty Research Award</b> , 2018 <b>CrowdFlower AI for Everyone Award</b> , 2018 <b>NYU MacCracken PhD Fellowship</b> , 2007 – 2012	
<b>PROFESSIONAL SERVICES</b>	<b>Best Paper Award Committee:</b> EMNLP (2018); <b>Senior Area Chair:</b> EMNLP (2024, 2022), NAACL (2022, 2021), ACL (2020); <b>Area Chair:</b> COLM (2024), ACL (2023, 2019), EMNLP (2021, 2020, 2018, 2016), AAAI (2020), NAACL (2019), COLING (2018); <b>Action Editor:</b> ACL Rolling Review (2021-2022); <b>Program Committee:</b> ACL (2021, 2018, 2017, 2015, 2014, 2013), EMNLP (2017, 2015, 2014), NAACL (2015), WWW (2017, 2016, 2015), AAAI (2016, 2015, 2012), KDD (2015), COLING (2014); <b>Publicity Chair:</b> EMNLP (2019), NAACL (2018, 2016); <b>Workshop Chair:</b> ACL (2017); <b>Journal Reviewer:</b> Transactions of the Association for Computational (TACL), Journal of Artificial Intelligence Research (JAIR); <b>Organizer:</b> Workshop on Text Simplification, Accessibility, and Readability at EMNLP (2022); Workshop on Natural Language Generation, Evaluation, and Metrics at ACL (2021); Workshop on Noisy User-generated Text at EACL (2023), EMNLP (2021, 2020, 2019, 2018, 2017), COLING (2022, 2016), ACL (2015); Mid-Atlantic Student Colloquium on Speech, Language and Learning (2016).	
<b>PUBLICATIONS</b>	(Underline is used to indicate student advisees.) <i>Having Beer after Prayer? Measuring Cultural Bias in Large Language Models</i> <u>Tarek Naous</u> , <u>Michael J. Ryan</u> , Alan Ritter, Wei Xu ACL 2024, long paper (acceptance rate 21.3%) ( <b>Best Social Impact Award</b> , selection ratio $\leq 0.25\%$ ; <b>Press Coverage by VentureBeat</b> ) <i>Reducing Privacy Risks in Online Self-Disclosures with Language Models</i> <u>Yao Dou</u> , Isadora Krsek, <u>Tarek Naous</u> , Anubha Kabra, Sauvik Das, Alan Ritter, Wei Xu ACL 2024, long paper (acceptance rate 21.3%) <i>NEO-BENCH: Evaluating Robustness of Large Language Models with Neologisms</i> <u>Jonathan Zheng</u> , Alan Ritter, Wei Xu ACL 2024, long paper (acceptance rate 21.3%) <i>Meta-Tuning LLMs to Leverage Lexical Knowledge for Generalizable Language Style Understanding</i> Ruohao Guo, Wei Xu, Alan Ritter ACL 2024, long paper (acceptance rate 21.3%)	

*FactPICO: Factuality Evaluation for Plain Language Summarization of Medical Evidence*  
Sebastian Antony Joseph, Lily Chen, Jan Trienes, Hannah Louisa Göke, Monika Coers, Wei Xu, Byron C Wallace, Junyi Jessie Li  
ACL 2024, long paper (acceptance rate 21.3%)

*InfoLossQA: Characterizing and Recovering Information Loss in Text Simplification.*  
Jan Trienes, Sebastian Joseph, Jörg Schlötterer, Christin Seifert, Kyle Lo, Wei Xu, Byron Wallace, Jessie Li  
ACL 2024, long paper (acceptance rate 21.3%)

*Constrained Decoding for Cross-lingual Label Projection*  
Duong Minh Le, Yang Chen, Alan Ritter, Wei Xu  
ICLR 2024 (acceptance rate 30.8%)

*Design and Evaluation of an Automatic Text Simplification Prototype with Deaf and Hard-of-hearing Readers*  
Oliver Alonzo, Sooyeon Lee, Akhter Al Amin, Mounica Maddela, Wei Xu, Matt Huenerfauth  
ASSETS 2024 (acceptance rate 30.0%)

*Stanceosaurus 2.0 - Classifying Stance Towards Russian and Spanish Misinformation*  
Anton Lavrouk, Ian Ligon, Jonathan Zheng, Tarek Naous, Wei Xu, Alan Ritter  
EACL W-NUT Workshop 2024, long paper

*Dancing Between Success and Failure: Edit-level Simplification Evaluation using SALSA*  
David Heineman, Yao Dou, Mounica Maddela, Wei Xu  
EMNLP 2023, long paper (acceptance rate 23.3%)

*Multilingual Simplification of Medical Texts*  
Sebastian Joseph, Kathryn Kazanas, Keziah Reina, Vishnesh Ramanathan, Wei Xu, Byron Wallace, Junyi Jessie Li  
EMNLP 2023, long paper (acceptance rate 23.3%)

*Thresh: A Unified, Customizable and Deployable Platform for Fine-Grained Text Evaluation*  
David Heineman, Yao Dou, Wei Xu  
EMNLP 2023, system demonstration (acceptance rate 25%)

*A Computational Interface to Translate Strategic Intent from Unstructured Language in a Low-Data Setting*  
Pradyumna Tambwekar, Lakshita Dodeja, Nathan Vaska, Wei Xu, Matthew Gombolay  
Findings of EMNLP 2023, long paper (acceptance rate 46.2%)

*LENS: A Learnable Evaluation Metric for Text Simplification*  
Mounica Maddela, Yao Dou, David Heineman, Wei Xu  
ACL 2023, long paper (acceptance rate 23.5%)

*Distill or Annotate? Cost-Efficient Fine-Tuning of Compact Models*  
Junmo Kang, Wei Xu, Alan Ritter  
ACL 2023, long paper (acceptance rate 23.5%)

*Improved Instruction Ordering in Recipe-Grounded Conversation*  
Duong Minh Le, Ruohao Guo, Wei Xu, Alan Ritter  
ACL 2023, long paper (acceptance rate 23.5%)

*Revisiting non-English Text Simplification: A Unified Multilingual Benchmark*  
Michael Ryan, Tarek Naous, Wei Xu  
ACL 2023, long paper (acceptance rate 23.5%) – **Best Paper Award Honorable Mention**

*Human-in-the-loop Evaluation for Early Misinformation Detection: A Case Study of COVID-19 Treatments*  
Ethan Mendes, Yang Chen, Wei Xu, Alan Ritter  
ACL 2023, long paper (acceptance rate 23.5%)

*Frustratingly Easy Label Projection for Cross-lingual Transfer*  
Yang Chen, Chao Jiang, Alan Ritter, Wei Xu  
Findings of ACL 2023, long paper

*Teaching the Pre-trained Model to Generate Simple Texts for Text Simplification*  
Renliang Sun, Wei Xu, Xiaojun Wan  
Findings of ACL 2023, short paper

*Improving Large-scale Paraphrase Acquisition and Generation*  
Yao Dou, Chao Jiang, Wei Xu  
EMNLP 2022, long paper (acceptance rate 22.1%)

*arXivEdits: Understanding the Human Revision Process in Scientific Writing*  
Chao Jiang, Wei Xu, Sam Stevens  
EMNLP 2022, long paper (acceptance rate 22.1%)

*Stanceosaurus: Classifying Stance Towards Multicultural Misinformation*

Jonathan Zheng, Ashutosh Baheti, Tarek Naous, Wei Xu, Alan Ritter

EMNLP 2022, long paper (acceptance rate 22.1%)

*A Dataset of Word-Complexity Judgements from Deaf and Hard-of-Hearing Adults for Text Simplification*

Oliver Alonzo, Sooyeon Lee, Mounica Maddela, Wei Xu and Matt Huenerfauth

EMNLP TSAR Workshop 2022, long paper

*Extracting a Knowledge Base of COVID-19 Events from Social Media*

Shi Zong, Ashutosh Baheti, Wei Xu, Alan Ritter

COLING 2022, long paper (acceptance rate 31.2%)

*BiSECT: Learning to Split and Rephrase Sentences with Bitexts*

Joongwon Kim\*, Mounica Maddela\*, Reno Kriz, Wei Xu, Chris Callison-Burch (\*equal contribution)

EMNLP 2021, long paper (acceptance rate 23.3%)

*Pre-train or Annotate? Domain Adaptation with a Constrained Budget*

Fan Bai, Alan Ritter, Wei Xu

EMNLP 2021 (acceptance rate 23.3%)

*WIKIBIAS: Detecting Multi-Span Subjective Biases in Language*

Yang Zhong, Jingfeng Yang, Wei Xu, Diyi Yang

Findings of EMNLP 2021

*Neural semi-Markov CRF for Monolingual Word Alignment*

Wuwei Lan\*, Chao Jiang\*, Wei Xu (\*equal contribution)

ACL 2021, long paper (acceptance rate 21.2%)

*Controllable Text Simplification with Explicit Paraphrasing*

Mounica Maddela, Fernando Alva-Manchego, Wei Xu

NAACL 2021, long paper (acceptance rate 28%)

*The GEM Benchmark: Natural Language Generation, its Evaluation and Metrics*

Sebastian Gehrmann, Tosin Adewumi, Karmanya Aggarwal, Pawan Sasanka Ammanamanchi, Anuoluwapo Aremu, Antoine Bosselut, Khyathi Raghavi Chandu, Miruna-Adriana Clinciu, Dipanjan Das, Kaustubh Dhole, Wanyu Du, Esin Durmus, Ondřej Dušek, Chris Chinenye Emezue, Varun Gangal, Cristina Garbacea, Tatsunori Hashimoto, Yufang Hou, Yacine Jernite, Harsh Jhamtani, Yangfeng Ji, Shailza Jolly, Mihir Kale, Dhruv Kumar, Faisal Ladhak, Aman Madaan, Mounica Maddela, Khyati Mahajan, Saad Mahamood, Bodhisattwa Prasad Majumder, Pedro Henrique Martins, Angelina McMillan-Major, Simon Mille, Emiel van Miltenburg, Moin Nadeem, Shashi Narayan, Vitaly Nikolaev, Andre Niyongabo Rubungo, Salomey Osei, Ankur Parikh, Laura Perez-Beltrachini, Niranjan Ramesh Rao, Vikas Raunak, Juan Diego Rodriguez, Sashank Santhanam, João Sedoc, Thibault Sellam, Samira Shaikh, Anastasia Shimorina, Marco Antonio Sobrevilla Cabezudo, Hendrik Strobelt, Nishant Subramani, Wei Xu, Diyi Yang, Akhila Yerukola, Jiawei Zhou (alphabetically ordered)

ACL 2021 Workshop on Natural Language Generation, Evaluation, and Metrics

(project website: <https://gem-benchmark.com/>)

*An Empirical Study of Pre-trained Transformers for Arabic Information Extraction*

Wuwei Lan, Yang Chen, Wei Xu, Alan Ritter

EMNLP 2020, short paper (acceptance rate 16.7%)

*WNUT-2020 Task 1 Overview: Extracting Entities and Relations from Wet Lab Protocols*

Jeniya Tabassum, Sydney Lee, Wei Xu, Alan Ritter

EMNLP 2020 Workshop on Noisy User-generated Text (shared-task overview)

*Neural CRF Model for Sentence Alignment in Text Simplification*

Chao Jiang, Mounica Maddela, Wuwei Lan, Yang Zhong, Wei Xu

ACL 2020, long paper (acceptance rate 25.2%)

*An Empirical Study of Named Entity Recognition in StackOverflow*

Jeniya Tabassum, Mounica Maddela, Wei Xu, Alan Ritter

ACL 2020, long paper (acceptance rate 25.2%)

*Generalizing Natural Language Analysis through Span-relation Representations*

Zhengbao Jiang, Wei Xu, Jun Araki, Graham Neubig

ACL 2020, long paper (acceptance rate 25.2%)

*Learning Relation Entailment with Structured and Textual Information*

Zhengbao Jiang, Jun Araki, Donghan Yu, Ruohong Zhang, Wei Xu, Yiming Yang, Graham Neubig

AKBC 2020, long paper

*Discourse Level Factors for Sentence Deletion in Text Simplification*

Yang Zhong, Chao Jiang, Wei Xu, Junyi Jessy Li  
 AAAI 2020, long paper (acceptance rate 20.6%; oral presentation)

*Multi-task Pairwise Neural Ranking for Hashtag Segmentation*  
Mounica Maddela, Wei Xu, Daniel Preotiuc-Pietro  
 ACL 2019, long paper (acceptance rate 25.7%)

*A Word-Complexity Lexicon and A Neural Readability Ranking Model for Lexical Simplification*  
Mounica Maddela, Wei Xu  
 EMNLP 2018, long paper (acceptance rate 25.8%; oral presentation)

*Neural Network Models for Paraphrase Identification, Semantic Textual Similarity, Natural Language Inference, and Question Answering*  
Wuwei Lan, Wei Xu  
 COLING 2018, long paper (**Best Paper Award**; selection rate  $8/888 = 0.90\%$ )

*An Annotated Corpus for Machine Reading of Instructions in Wet Lab Protocols*  
Chaitanya Kulkarni, Wei Xu, Alan Ritter, Raghu Machiraju  
 NAACL 2018, short paper (acceptance rate 29%)

*Character-based Neural Networks for Sentence Pair Modeling*  
Wuwei Lan, Wei Xu  
 NAACL 2018, short paper (acceptance rate 29%)

*A Continuously Growing Dataset of Sentential Paraphrases*  
Wuwei Lan, Siyu Qiu, Hua He, Wei Xu  
 EMNLP 2017, long paper (acceptance rate 25.8%)

*From Shakespeare to Twitter: What are Language Styles all about?*  
 Wei Xu  
 EMNLP 2017 Workshop on Stylistic Variation

*A Minimally Supervised Method for Recognizing and Normalizing Time Expressions in Twitter*  
Jeniya Tabassum, Alan Ritter, Wei Xu  
 EMNLP 2016, long paper (acceptance rate 26%; oral presentation)

*Optimizing Statistical Machine Translation for Simplification*  
 Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, Chris Callison-Burch  
 TACL 2016, long paper (oral presentation at ACL 2016)

*Discovering User Attribute Stylistic Differences via Paraphrasing*  
 Daniel Preotiuc-Pietro, Wei Xu, Lyle Ungar  
 AAAI 2016, long paper (acceptance rate 26%; oral presentation)

*Results of the WNUT16 Named Entity Recognition Shared Task*  
 Benjamin Strauss, Bethany Toma, Alan Ritter, Marie-Catherine de Marneffe, Wei Xu  
 COLING 2016 Workshop on Noisy User-generated Text (shared-task overview)

*Problems in Current Text Simplification Research: New Data Can Help*  
 Wei Xu, Chris Callison-Burch, Courtney Napoles  
 TACL 2015, long paper (oral presentation at EMNLP 2015)

*Cost Optimization for Crowdsourcing Translation*  
 Mingkun Gao, Wei Xu, Chris Callison-Burch  
 NAACL 2015, long paper (acceptance rate 29%)

*SemEval-2015 Task 1: Paraphrase and Semantic Similarity in Twitter*  
 Wei Xu, Chris Callison-Burch, William B. Dolan  
 SemEval 2015, long paper (shared-task overview)

*Shared Tasks of the 2015 Workshop on Noisy User-generated Text: Twitter Lexical Normalization and Named Entity Recognition*  
 Timothy Baldwin, Marie Catherine de Marneffe, Bo Han, Young-Bum Kim, Alan Ritter, Wei Xu  
 ACL 2015 Workshop on Noisy User-generated Text (shared-task overview; author ordered alphabetically)

*Data-driven Approaches for Paraphrasing Across Language Variations*  
 Wei Xu  
 Ph.D. Thesis 2014

*Extracting Lexically Divergent Paraphrases from Twitter*  
 Wei Xu, Alan Ritter, Chris Callison-Burch, William B. Dolan, Yangfeng Ji  
 TACL 2014, long paper (oral presentation at NAACL 2015)

*Infusion of Labeled Data into Distant Supervision for Relation Extraction*

Maria Pershina, Bonan Min, Wei Xu, Ralph Grishman  
 ACL 2014, short paper (acceptance rate 25.2%; oral presentation)  
*Filling Knowledge Base Gaps for Distant Supervision of Relation Extraction*  
 Wei Xu, Raphael Hoffmann, Le Zhao, Ralph Grishman  
 ACL 2013, short paper (acceptance rate 24%)  
*Gathering and Generating Paraphrases from Twitter with Application to Normalization*  
 Wei Xu, Alan Ritter, Ralph Grishman  
 ACL 2013 Workshop on Building and Using Comparable Corpora  
*A Preliminary Study of Tweet Summarization using Information Extraction*  
 Wei Xu, Ralph Grishman, Adam Meyers, Alan Ritter  
 NAACL 2013 Workshop on Language Analysis in Social Media  
*Paraphrasing for Style*  
 Wei Xu, Alan Ritter, Bill Dolan, Ralph Grishman, Colin Cherry  
 COLING 2012, long paper (acceptance rate 25%)  
*Exploiting Syntactic and Distributional Information for Spelling Correction with Web-Scale N-grams Models*  
 Wei Xu, Joel Tetreault, Martin Chodorow, Ralph Grishman, Le Zhao  
 EMNLP 2011, long paper (acceptance rate 23.7%)  
*New York University 2011 System for KBP (Knowledge Base Population) Slot Filing*  
 Ang Sun, Ralph Grishman, Wei Xu, Bonan Min  
 TAC 2011 (best performance system in NIST KBP-2011 evaluation)  
*Passage Retrieval for Information Extraction using Distant Supervision*  
 Wei Xu, Ralph Grishman, Le Zhao  
 IJCNLP 2011, long paper (acceptance rate 36%)  
*Who, What, When, Where, Why? Comparing Multiple Approaches to the Cross-Lingual 5W Task*  
 Kristen Parton, Kathleen McKeown, Bob Coyne, Mona Diab, Ralph Grishman, Dilek Hakkani-Tür, Mary Harper, Heng Ji, Weiyun Ma, Adam Meyers, Sara Stolbach, Ang Sun, Gokhan Tur, Wei Xu, Sibel Yaman  
 ACL 2009, long paper (acceptance rate 21%; oral presentation)  
*A Parse-and-Trim Approach with Information Significance for Chinese Sentence Compression*  
 Wei Xu, Ralph Grishman  
 ACL Workshop on Language Generation and Summarisation 2009  
*Transducing Logical Relations from Automatic and Manual Annotation*  
 Adam Meyers, Michiko Kosaka, Heng Ji, Nianwen Xue, Mary Harper, Ang Sun, Wei Xu, Shasha Liao  
 ACL Workshop on Linguistic Annotation 2009  
*Automatic Recognition of Logical Relations for English, Chinese and Japanese in the GLARF Framework*  
 Adam Meyers, Michiko Kosaka, Nianwen Xue, Heng Ji, Ang Sun, Shasha Liao, Wei Xu  
 SemEval 2009, long paper  
*Extractive Summarization using Inter- and Intra- Event Relevance*  
 Wenjie Li, Wei Xu, Mingli Wu, Chunfa Yuan, Qin Lu  
 ACL 2006, long paper (acceptance rate 23%; oral presentation)  
*Using Non-Local Features to Improve Named Entity Recognition Recall*  
 Xinnian Mao, Wei Xu, Yuan Dong, Haila Wang  
 PACLIC 2007, long paper  
*Deriving Event Relevance from the Ontology Constructed with Formal Concept Analysis*  
 Wei Xu, Wenjie Li, Mingli Wu, Wei Li, Chunfa Yuan  
 CICLing 2006, long paper (acceptance rate 30.4%; oral presentation)  
*Building Document Graph for Text Summarization: An Event-based Approach*  
 Wei Xu, Wenjie Li, Mingli Wu, Wei Li, Chunfa Yuan  
 ICCPOL 2006  
*The THU/PolyU System at MSE 2006: An Event-relevance based Approach*  
 Wei Xu, Chunfa Yuan, Mingli Wu, Wenjie Li  
 MSE 2006

## TUTORIALS

### Automatic and Human-AI Interactive Text Generation

ACL 2024 (<https://acl2024-text-generation-tutorial.github.io/>)

August 2024

### Social Media and Text Analytics

NASSLLI 2015

July 2015

INVITED  
TALKS

<b>Cultural Biases, World Languages, and Privacy Protection in Large Language Models</b>	
Bloomberg, CTO Data Science Speaker series	Oct 2024 (expected)
MIT, Embodied Intelligence Seminar	Sep 2024 (expected)
<b>Human-AI Collaboration in Evaluating Large Language Models</b>	
Northeastern University	Sep 2024 (expected)
HEAL Workshop at CHI 2024	May 2024
<b>Cultural Biases and Multilingual Capabilities of Large Language Models</b>	
Megagon Labs	Aug 2024
<b>Amazing Multilingual Capabilities and Concerning Cultural Biases in Large Language Models</b>	
University of California, Los Angeles	Mar 2024
University of South California	Mar 2024
Google Research, India	Mar 2024
<b>Amplifying LLM's Cross-lingual Ability with Label Projection</b>	
Brainlink, South Korea	Dec 2023
<b>GPT-3 vs Humans: Rethinking Evaluation of Natural Language Generation</b>	
Johns Hopkins University (CLIP Seminar), Baltimore, MD	Feb 2023
<b>Capturing Human Language Diversity and (Mis-)Information Spreading Online</b>	
VinAI, Vietnam	Apr 2023
University of Chicago	Feb 2023
Columbia University, New York, NY	Oct 2022
<b>Importance of Data and Controllability in Neural Language Generation</b>	
Cornell Tech (LMSS Seminar), New York, NY	Sep 2022
Dataminr, New York, NY	Nov 2021
Nanjing University, Nanjing, China	Oct 2021
SimpleText workshop at CLEF 2021	Sep 2021
Stanford University (NLP Seminar), Stanford, CA	Aug 2021
University of California, Los Angeles (Big Data and ML Seminar)	Jun 2021
<b>Importance of Data and Linguistics in Neural Language Generation</b>	
New York University, New York, NY (NLP and Text-as-Data Speaker Series)	May 2021
Carnegie Mellon University, Pittsburgh, PA (LTI Colloquium)	Nov 2020
Google Research	Oct 2020
<b>Natural Language Understanding for Noisy Text</b>	
University of Sheffield, Sheffield, United Kingdom (NLP Seminar)	Oct 2020
USC Information Sciences Institute, Los Angeles, CA (NLP Seminar)	Oct 2020
<b>Automatic Text Simplification</b>	
University of Pittsburgh, Pittsburgh, PA (NLP Seminar)	Oct 2020
<b>Understanding and Generating Human Language</b>	
Emory University, Atlanta, GA (CS Department Seminar)	Sep 2020
University of Maryland, College Park, MD (CS Colloquium)	Feb 2020
University of Massachusetts, Amherst, MA	Jan 2020
Georgia Institute of Technology, Atlanta, GA	Dec 2019
<b>Learning for Unlimited Human Language</b>	
Peking University, Beijing, China	Dec 2018
<b>Learning Large-scale Paraphrases for Natural Language Understanding and Generation</b>	
Midwest Machine Learning Symposium, Chicago, IL	Jun 2018
Facebook, Menlo Park, CA	May 2018
Stanford Research Institute, Menlo Park, CA	May 2018
Twitter, San Francisco, CA	May 2018
IBM Thomas J. Watson Research Center, New York, NY	Nov 2017
<b>How does AI Understand Language?</b>	
Women in Analytics Conference, Columbus, OH (Main Stage Panel)	Mar 2018
<b>Can Paraphrase be a Ultimate Solution for NLU and NLG?</b>	
Google Research, New York, NY	Jul 2017
<b>Paraphrase <math>\approx</math> Monolingual Translation</b>	
Amazon, Berlin, Germany	Aug 2016
<b>Multiple Instance Learning from Unlimited Text</b>	

	Microsoft Research Asia, Beijing, China	Dec 2016
	University of Delaware, Newark, DE	Sep 2016
	University of Edinburgh, Edinburgh, United Kingdom	May 2016
	Ohio State University, Columbus, OH	Apr 2016
	University of North Carolina, Chapel Hill, NC	Apr 2016
	Arizona State University, Tempe, AZ	Mar 2016
	Vanderbilt University, Nashville, TN	Mar 2016
	Imperial College London, London, United Kingdom	Mar 2016
	University of Waterloo, Waterloo, ON, Canada (CS Seminar)	Mar 2016
	Indiana University, Bloomington, IN (Computer Science Colloquium Series)	Feb 2016
	Washington University, St Louis, MI (Computer Science & Engineering Colloquia Series)	Feb 2016
	Simon Fraser University, Vancouver, BC, Canada	Feb 2016
	University of Alberta, Edmonton, AB, Canada	Feb 2016
	Yale University, New Haven, CT (CS Talk)	Feb 2016
	University of Maryland, College Park, MD (CLIP Colloquium)	Oct 2015
	Ohio State University, Columbus, OH (Clippers Seminar)	Oct 2015
	<b>Large-scale Paraphrase Acquisition from Twitter</b>	
	DARPA's DEFT Project Meeting, Boulder, CO	May 2015
	<b>Learning and Generating Paraphrases from Twitter and Beyond</b>	
	Carnegie Mellon University, Pittsburgh, PA	Apr 2015
	Columbia University, New York, NY (NLP Talk)	Apr 2015
	Johns Hopkins University, Baltimore, MD (CLSP Colloquium)	Feb 2015
	<b>Paraphrases in Twitter</b>	
	Twitter, San Francisco, CA	Feb 2015
	<b>Modeling Lexically Divergent Paraphrases in Twitter (and Shakespeare!)</b>	
	The City University of New York, New York, NY (NLP Seminar)	Mar 2015
	IBM Research - Almaden, San Jose, CA	Feb 2015
	University of California, Berkeley, CA	Feb 2015
	The University of Texas, Austin, TX (Forum for Artificial Intelligence)	Feb 2015
	Yahoo!, New York, NY	Dec 2014
	Carnegie Mellon University, Pittsburgh, PA (CL+NLP Lunch Seminar)	Nov 2014
	Microsoft Research, Seattle, WA (Visiting Speaker Series)	Aug 2014
	<b>Incremental Information Extraction</b>	
	Stanford Research Institute, Palo Alto, CA	Apr 2012
	IARPA's KDD Project Meeting, San Diego, CA	May 2011
	<b>Event-based Summarization</b>	
	Thomson Reuters, Eagan, Minnesota, MN	Nov 2009
<b>RESEARCH GRANTS</b>	<b>NSF Grant</b> <i>The Cost of AI: A Comparative Study of Machine Learning Training Methods</i>	2024 – 2028
	co-PI (25%), total \$399,634	
	<b>NSF CAREER</b> <i>An Integrated Framework for Controllable Text Generation</i>	2022 – 2027
	PI (100%), total \$537,527	
	<b>NSF AI Institute</b> <i>Collaborative Assistance and Responsive Interaction for Networked Groups</i>	2021 – 2026
	Senior Personnel, total \$19,995,808	
	<b>IARPA Grant</b> <i>Human Interpretable Attribution of Text Using Underlying Structure</i>	2022 – 2024
	co-PI (50%), total \$491,467	
	<b>IARPA Grant</b> <i>Better Extraction from Text Towards Enhanced Retrieval</i>	2019 – 2023
	co-PI (50%), total \$850,000	
	<b>NSF Grant</b> <i>Collaborative Research: Automatic Text-Simplification and Reading-Assistance to Support Self-Directed Learning by Deaf and Hard-of-Hearing Computing Workers</i>	2018 – 2022
	PI (100%), total \$375,732	
	<b>NSF CRII RI: Learning a Timely Semantic Resource from Social Media Data</b>	2018 – 2021
	PI (100%), total \$183,000	
	<b>DARPA Research Grant</b> <i>Computational Simulation of Online Social Behavior</i>	2017 – 2021
	co-PI (50%), total \$600,000	
<b>TEACHING</b>	<i>CS 8803-LLM Large Language Models (graduate level)</i>	
	Teaching eval: in progress Fall 2024	

*CS 7650 Natural Language Processing (graduate level)*  
Teaching eval: 4.7/5.0 Spring 2024, 4.3/5.0 Fall 2022, 4.3/5.0 Fall 2021

*CS 8803 Advanced Natural Language Processing (graduate level)*  
Teaching eval: 4.9/5.0 Fall 2023

*CS 4650 Natural Language Processing (undergraduate level)*  
Teaching eval: 4.7/5.0 Spring 2023, 4.6/5.0 Spring 2022, 3.7/5.0 Spring 2021

*CSE 5539 Social Media and Text Analytics (<http://socialmedia-class.org/>)*  
A new course integrated with research, covering from basic to state-of-the-art machine learning algorithms  
Teaching eval: 4.13/5.00 Fall 2019, 4.40/5.00 Fall 2017, 4.60/5.00 Fall 2016; 5.72/6.00 at NASSLLI 2015

*CSE 5522 Artificial Intelligence II: Advanced Techniques (mixed undergraduate and graduate level)*  
Teaching eval: 4.85/5.00 Fall 2018, 4.50/5.00 Spring 2018

*CSE 5525 Speech and Language Processing (mixed undergraduate and graduate level)*  
Teaching eval: 4.42/5.00 Spring 2020, 3.80/5.00 Spring 2017

**OUTREACH  
ACTIVITIES**

Judge, Hackathon event of Women @ College of Computing, Georgia Tech Mar 2024  
Panel Moderator, Women in Cable Telecommunications event, Atlanta, Georgia Nov 2023  
Mentor, ACL Mentorship Session, online, Sep 2023  
Mentor, Group Mentoring Sessions for undergraduate/master students at ACL 2020 Jul 2020  
Speaker/Judge, Ohio High School Hackathon Mar 2019  
Speaker, Franklin Friday art and science festival in Columbus Ohio Mar 2019  
Panelist, CogFest - Cognitive Science Festival Apr 2018  
Mentor, Women and Underrepresented Minorities in NLP Workshop Jun 2018  
Mentor, OSU's AI Hackathon Apr 2018  
Speaker/Panelist, Women in Analytics Conference Mar 2018  
Speaker, OSU's AI Club Feb 2018  
Judge, HackOhio Oct 2017  
Mentor, Women and Underrepresented Minorities in NLP Workshop Jul 2017  
Judge, Ohio High School Hackathon Mar 2017  
Presenter, Philadelphia Science Festival Apr 2015

**OPEN SOURCE  
CODE / DATA**

*Thresh: A Unified, Customizable and Deployable Platform for Fine-Grained Text Evaluation* Dec 2023  
<https://github.com/davidheineman/thresh>

*LENS: A Learnable Evaluation Metric for Text Simplification* Jul 2023  
<https://github.com/Yao-Dou/LENS>

*arXivEdits: A Dataset with Edit-level Annotations of Scientific Paperps* Jul 2023  
<https://github.com/chaojiang06/arXivEdits>

*ChattyChef: A Dataset of Cooking Dialogues* Jul 2023  
<https://github.com/octaviaguio/ChattyChef>

*MultiPIT: Multi-Topic Paraphrases in Twitter* Dec 2022  
<https://yao-dou.github.io/multipit/>

*HashtagMaster: A Semantic Analysis Tool for Hashtags* Jun 2019  
[https://mounicam.github.io/hashtag\\_master](https://mounicam.github.io/hashtag_master)

*Pairwise Neural Ranking Model and SimplePPDB++* Oct 2018  
[https://github.com/lanwuwei/SPM\\_toolkit](https://github.com/lanwuwei/SPM_toolkit)

*SPM Toolkit for Sentence Pair Modeling* Aug 2018  
[https://github.com/lanwuwei/SPM\\_toolkit](https://github.com/lanwuwei/SPM_toolkit)

*LanguageNet: Large-scale Paraphrase Corpus* Sep 2017  
<https://github.com/lanwuwei/paraphrase-dataset>

*Syntax MT-based Text Simplification System and SARI Evaluation Metric* May 2015  
<https://github.com/cocoxu/simplification/> (contribution to the Joshua Machine Translation Toolkit)

*NEWSELA Text Simplification Corpus* Sep 2015  
<https://newsela.com/data/> (widely adopted as the benchmark for text simplification research)

*Multiple-instance Learning Paraphrase Model* Dec 2014  
<https://github.com/cocoxu/multip>

*Twitter Paraphrase Corpus (shared-task at SemEval-2015)* Oct 2014  
<http://alt.qcri.org/semeval2015/task1/>

*Twitter Normalization Phrase Table* Oct 2014  
<https://github.com/cocoxu/twitterparaphrase/>

*Parallel Shakespeare Corpus and Model* Jul 2012  
<https://github.com/cocoxu/Shakespeare/>