Lucy Li

Natural language processing, computational sociolinguistics, & computational social science

lucy3_li@berkeley.edu | lucy3.github.io

Education

University of California, Berkeley *Berkeley, CA*

PhD Information Science *Aug 2019 - present*

- Advisor: David Bamman
- Committee members: Isaac Bleaman, Niloufar Salehi, Dan Jurafsky
- Berkeley AI Research (BAIR)

Stanford University *Stanford, CA*

BS Symbolic Systems, MS Computer Science *Sept 2014 - June 2019*

- A coterminal degree, w/ a concentration in language and depth in artificial intelligence.
- Study abroad at University of Oxford, Winter 2017.

Experience

Allen Institute for Artificial Intelligence *Seattle, Washington*

Research Intern *June 2023 - Present*

- Mentor: Jesse Dodge
- Analyzing large language models' data curation practices on the AllenNLP team, in collaboration with Luca Soldaini, Emma Strubell, Suchin Gururangan, and Lauren F. Klein.

Allen Institute for Artificial Intelligence *Seattle, Washington*

Research Intern *May 2022 - Dec 2022*

- Mentors: Katie Keith, Jesse Dodge
- Mapping scientific domains on the Semantic Scholar and AllenNLP teams. Awarded "Outstanding Intern of the Year."

Microsoft Research *Montreal, Canada*

Research Intern *May 2021 - Aug 2021*

- Mentors: Alexandra Olteanu, Su Lin Blodgett
- Auditing natural language generation systems on the Fairness, Accountability, Transparency, and Ethics (FATE) team, in collaboration with Milad Shokouhi and Hanna Wallach.

Stanford Computer Science *Stanford, CA*

Research Assistant *Jan 2019 – Dec 2019*

- Advisors: Dan Jurafsky, Patricia Bromley.
- Investigated the framing and representation of underrepresented groups in history textbooks with linguistics PhD student Dora Demszky.

École Polytechnique Fédérale de Lausanne *Lausanne, Switzerland*

Research Intern *July 2018 – Sept 2018*

- Advisor: Robert West (Data Science Lab)
- Operationalized and analyzed behavioral trends in a political quote dataset using Apache Spark, emotion lexicons, Stanford CoreNLP parsers, and social networks.

Stanford Computer Science *Stanford, CA*

Research Assistant *April 2017 – June 2018*

- Advisors: David Jurgens, Jure Leskovec (Stanford Network Analysis Project), Dan Jurafsky (NLP group)
- Used language and social network features to classify fictional and real relationships with scikit-learn, NLTK, and Keras.

Papers

*indicates equal contribution.

Journals & Conferences

Li Lucy, Jesse Dodge, David Bamman, Katherine A. Keith. Words as Gatekeepers: Measuring Discipline-specific Terms and Meanings in Scholarly Publications. *Findings of the Association of Computational Linguistics (ACL)*, 2023.

Li Lucy, Divya Tadimeti, David Bamman. Discovering Differences in the Representation of People using Contextualized Semantic Axes. *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2022

Li Lucy, David Bamman. Characterizing English variation across social media communities with BERT. *Transactions of the Association of Computational Linguistics (TACL)*, 2021.

Li Lucy*, Dora Demszky*, Patricia Bromley, Dan Jurafsky. Content Analysis of Textbooks via Natural Language Processing: Findings on Gender, Race, and Ethnicity in Texas U.S. History Textbooks. *AERA Open*, 2020. [Best paper at American Educational Research Association (AERA) Educational Data Science Conference.]

Workshops

Li Lucy, David Bamman. Gender and Representation Bias in GPT-3 Generated Stories. *Workshop on Narrative Understanding (WNU) at the North American Association for Computational Linguistics (NAACL)*, 2021.

Emma Lurie, **Li Lucy**, Masha Belyi, Sofia Dewar, Daniel Rincón, John Baldwin, Rajvardhan Oak. Investigating Causal Effects of Instructions in Crowdsourced Claim Matching. *Computation + Journalism Symposium (C+J)*, 2020. [non-archival.]

Li Lucy, Julia Mendelsohn. Using sentiment induction to understand variation in gendered online communities. *Society for Computation in Linguistics (SCiL)*, 2019.

Li Lucy, Jon Gauthier. Are distributional representations ready for the real world? Evaluating word vectors for grounded perceptual meaning. *Language Grounding for Robotics (RoboNLP) Workshop at the Association for Computational Linguistics (ACL)*, 2017.

Working Papers

Luca Soldaini, Akshita Bhagia, Rodney Kinney, Dustin Schwenk, Russell Authur, Khyathi Chandu, **Li Lucy**, Xinxi Lyu, Ian Magnusson, Aakanksha Naik, Matthew E. Peters, Abhilasha Ravichander, Zejiang Shen, Emma Strubell, Nishant Subramani, Oyvind Tafjord, Evan Pete Walsh, Hannaneh Hajishirzi, Noah A. Smith, Luke Zettlemoyer, Iz Beltagy, Jesse Dodge, Dirk Groeneveld, Kyle Lo. Dolma: An Open Corpus of 3 Trillion Tokens for Language Model Pretraining Research. 2023.

Minju Choi*, **Li Lucy***. "Othering" through War: Depiction of Asians/Asian Americans in U.S. History Textbooks from California and Texas. 2022.

For Non-Research Audiences

Maria Antoniak, **Li Lucy**, Maarten Sap, Luca Soldaini. Using Large Language Models With Care. 2023. [Link.]

David Jurgens, **Li Lucy**. A Look inside the Pedagogy of Natural Language Processing. 2018. [Link.]

Awards, Fellowships, & Grants

Rising Star in EECS, Rising Stars Academic Career Workshop *2023*
Meta Research PhD Fellowship Finalist, Meta *2023*
AI2 Outstanding Intern of the Year Award, Allen Institute for Artificial Intelligence *2022*
Human-Centered Artificial Intelligence Seed Grant, Stanford HAI (PI: Patricia Bromley) *2021*
Graduate Research Fellowship, National Science Foundation *2019*
K. Jon Barwise Award for Distinguished Contributions, Stanford Symbolic Systems *2018*
Undergraduate Advising & Research (UAR) Small Grant, $1500, Stanford University *2018*
Grants for Education and Research, $1145, Stanford Symbolic Systems *2017*
Phi Beta Kappa, Stanford University *2017*

Presentations

Measuring Depictions and Expressions of Social Groups with NLP
May 2023. "Stanford NLP Seminar," Stanford University.

Context-Dependent Depictions of People Across Three Domains.
March 2023. "NLP for Social Science: From Language Models to Social Structures," Columbia University.

What big data and big models bring to the table.
February 2023. "Roundtable Series: Learning How to Play with the Machines," University of California, Berkeley.

Social NLP.
April 2022. Guest lecture, "Natural Language Processing," University of California, Berkeley.

Characterizing English variation across social media communities with BERT.
Oct 2022. Guest lecture, "Practical Approaches to Data Science with Text," Emory University.
June 2021. Guest lecture, "Computational Text Analysis," Barnard College.
Nov 2021. Guest lecture, "Practical Approaches to Data Science with Text," Emory University.

Content Analysis of Textbooks via Natural Language Processing.
Sept 2022. McGill Narrative and Society Conference, Montreal.
Oct 2021. 103rd Anniversary of the School of Information, Berkeley.
Feb 2021. Guest lecture, "Doing Digital History", Stanford.
Feb 2021. Stanford Human-Computer Interaction Lunch Seminar.
May 2021. Guest lecture, "Using Data to Describe the World," Stanford.
May 2020. Guest lecture, "Using Data to Describe the World," Stanford.
Oct 2019. 10th Annual New Directions in Analyzing Text as Data (TADA). Stanford, CA.

Teaching Experience

Undergraduate Advisees: Sabrina Baur (Current), Claire Wang (2022-Current), Aryia Dattamajumdar (2023), JJ Kim-Ebio (2022-2023), Sebastian Orozco (2022), Divya Tadimeti (2021-2022), Nikhil Mandava (2021).

Berkeley INFO 290, NLP & Social Interaction, Instructor (upcoming) *Summer 2024*
Stanford CS 224U, Natural Language Understanding, Course Assistant (top 5% in CS) *Spring 2019*
Symbolic Systems Program, Advising Fellow *2016 - 2017, 2019*
Stanford EE/CME 103, Introduction to Matrix Methods, Course Assistant *Fall 2017*

Service

Professional

Reviewer: ACL Rolling Review (2021-Present), ACL (2021-Present), EMNLP (2022-Present), NeurIPs (2023), AI & HCI Workshop at ICML (2023), The Web Conference (2023), NLP for Positive Impact Workshop (2022), COLING (2022), NAACL Student Research Workshop (2022), NAACL Workshop on Understanding Implicit and Underspecified Language (2022), SCiL (2022), CHI (2022), CSCW (2022), ACL Workshop on NLP for Positive Impact (2021), EMNLP (2021), AERA Open (2020), NeurIPS Human and Machine in-the-Loop Evaluation and Learning Strategies Workshop (2020).

Organizing committee: Teaching NLP (2021) at NAACL.

Advisory board: BERT for Humanists.

Community

General: Diaries of Social Data Research (Podcast Host, 2021-Present), Sociologists of Digital Things (Admin, 2021), NLP+CSS PhD Summer Reading Group (2020).

Students: Berkeley Undergraduate Research Apprentice Program (2021-Present), BAIR Mentoring Program (2020-2021), CS Kickstart (Speaker; 2020), UC Berkeley Girls in Engineering (Leader; 2020), Berkeley AI4ALL (Mentor; 2019), Stanford AI4All (Mentor; 2019), Girls Teaching Girls to Code (Mentor/Lead; 2018, 2019).

Skills

Computer Languages: Python, Julia, SQL

Natural Languages: English, Mandarin Chinese

Tools: NLTK, Stanford CoreNLP, SpaCy, scikit-learn, Apache Spark, MTurk, Figure Eight, Keras, TensorFlow, PyTorch.