Sarah A. Wiegreffe

✉ *saw@gatech.edu*
🖱 *sarahwie.github.io*

Education

2017–2022 **Georgia Institute of Technology**, *Ph.D. in Computer Science*.
Advisor: Professor Mark Riedl.
Committee: Professors Alan Ritter, Wei Xu, Noah Smith (University of Washington), and Sameer Singh (University of California Irvine).

2017-2020 **Georgia Institute of Technology**, *M.S. in Computer Science*.
Specialization: Machine Learning.
Relevant coursework: Computational Statistics, Statistical Machine Learning, Deep Learning, Natural Language Processing.

2013-2017 **Honors College at the College of Charleston**, *B.S. in Data Science*.
Summa Cum Laude.
Awarded Data Science Major of the Year and Departmental Honors.
Minors in Mathematics and International Studies.

2015 **University of Tartu**, *Estonia*.
Visiting student in the Faculty of Mathematics and Computer Science.
Coursework: Cryptology, Computational Neuroscience, Advanced French (European scale B2→C1).

Publications

Acceptance rates listed where known. * denotes equal contribution.

PhD Dissertation

13. **Sarah Wiegreffe**. *Interpreting Neural Networks for and with Natural Language*. 2022.

Preprints

12. Kaige Xie, **Sarah Wiegreffe**, Mark Riedl. *Calibrating Trust of Multi-Hop Question Answering Systems with Decompositional Probes.* 2022.

Peer-reviewed, Archival

11. **Sarah Wiegreffe**, Jack Hessel, Swabha Swayamdipta, Mark Riedl, Yejin Choi. *Reframing Human-AI Collaboration for Generating Free-Text Explanations.* Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT) 2022. Seattle, WA. Acceptance rate 21.96%.

10. **Sarah Wiegreffe***, Ana Marasović*. *Teach Me to Explain: A Review of Datasets for Explainable Natural Language Processing.* Conference on Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track 2021. Online. Acceptance rate 38%.

9. **Sarah Wiegreffe**, Ana Marasović, Noah A. Smith. *Measuring Association Between Labels and Rationales.* Conference on Empirical Methods in Natural Language Processing (EMNLP) 2021. Punta Cana, Dominican Republic. Acceptance rate 23.4% (8.8% oral presentations).

8. Sarthak Jain, **Sarah Wiegreffe**, Yuval Pinter, Byron C. Wallace. *Learning to Faithfully Rationalize by Construction.* Annual Meeting of the Association for Computational Linguistics (ACL) 2020. Online. Acceptance rate 22.7%.

7. **Sarah Wiegreffe***, Yuval Pinter*. *Attention is not not Explanation.* Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) 2019. Hong Kong. Acceptance rate 24% (7% oral presentations).

6. **Sarah Wiegreffe**, Edward Choi, Sherry Yan, Jimeng Sun, Jacob Eisenstein. *Clinical Concept Extraction for Document-Level Coding.* Biomedical Natural Language Processing Workshop (BioNLP) at the Annual Meeting of the Association for Computational Linguistics (ACL) 2019. Florence, Italy.

5. James Mullenbach, **Sarah Wiegreffe**, Jon Duke, Jimeng Sun, Jacob Eisenstein. *Explainable Prediction of Medical Codes from Clinical Text.* Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT) 2018. New Orleans, LA. Acceptance rate 31% (oral presentation).

Peer-reviewed, Non-archival (poster presentations)

4. Xiangyu Peng*, Siyan Li*, **Sarah Wiegreffe**, Mark Riedl. *Inferring the Reader: Guiding Automated Story Generation with Commonsense Reasoning.* Narrative Understanding Workshop at the North American Chapter of the Association for Computational Linguistics (NAACL) 2021. Online.

3. Xiangyu Peng*, Siyan Li*, **Sarah Wiegreffe**, Mark Riedl. *Improving Neural Storytelling with Commonsense Inferences.* Women in Machine Learning (WiML) workshop at the Conference on Neural Information Processing Systems (NeurIPS) 2020. Online.

2. **Sarah Wiegreffe***, Yuval Pinter*. *Attention is not not Explanation.* Women in Machine Learning (WiML) workshop at the Conference on Neural Information Processing Systems (NeurIPS) 2019. Vancouver, Canada.

1. **Sarah Wiegreffe**, Jihad Obeid, Paul Anderson. *Can Classification of Publications by Translational Categories be Automated?* American Medical Informatics Association (AMIA) Translational Bioinformatics Summit 2017. San Francisco, CA.

Selected Honors and Awards

2020 **Outstanding Intern**, *Allen Institute for Artificial Intelligence*.
Gift of $10,000 and returning offer. Awarded to 2-3 interns per year by research mentor nomination.

2018 **Graduate Cohort Member**, *ACM Computing Research Association*.
Sponsored to attend the Association for Computing Machinery (ACM)'s national workshop for female computing PhD students.

2017 **Graduate Fellowship**, *Phi Kappa Phi Honor Society*.
Gift of $5,000. Awarded to 51 students nationwide beginning doctoral studies.

2017 **Data Science Major of the Year**, *College of Charleston*.
One student selected per academic year.

2016 **Grace Hopper Scholar**, *Anita Borg Institute*.
Sponsored by Intel to attend the largest annual conference for women in computing.

2015 **Diploma of French Language Studies: Level B2**, *French Ministry of Education*.
Passed standardized oral and written exams. Recognized as having obtained fluency by the French government, sufficient for enrollment in French universities.

2013-2017 **Charleston Fellow**, *College of Charleston*.
Gift of $92,000 toward tuition and fees. Awarded by competitive interview process to less than 0.01% of students at the university.

Selected Talks

2022 **Reframing Human-AI Collaboration for Generating Free-Text Explanations**, *NLP Reading Group*, University of Oxford.

2021 **Can Large Language Models Explain their Predictions?**, *Allen Institute for AI Internal Company-Wide Meeting*.

2021 **Measuring Association Between Labels and Free-Text Rationales (pre-recorded video)**, *EMNLP 2021*.

2021 **Measuring Association Between Labels and Free-Text Rationales (live; recorded)**, *NLP with Friends seminar*, Online.

2020 **BlackBoxNLP: What are we looking for, and where do we stand? (live; recorded)**, *NLP/ISI seminar*, University of Southern California.

2019 **Attention is not not Explanation (live; recorded)**, *EMNLP 2019*.

2019 **Transformers and Natural Language Applications**, *Guest lecture*, graduate deep learning course at Georgia Tech.

2019 **Self Attention for Universal Representations of Clinical Events**, *Final internship presentation*, Google AI.

Professional Experience

Nonprofits and Industry

2022-present **Young Investigator**, *Allen Institute of Artificial Intelligence*.
Post-doctoral position on the Aristo team.

2021 **Research Intern**, *Allen Institute of Artificial Intelligence*.
Hosted by Drs. Jack Hessel and Swabha Swayamdipta, and Professor Yejin Choi. Worked on few-shot explanation generation and effective human evaluation.

2020 **Research Intern**, *Allen Institute of Artificial Intelligence*.
Hosted by Dr. Ana Marasović and Professor Noah Smith. Worked on interpretability of deep learning models for NLP. **Awarded outstanding intern award.**

2019 **Research Intern**, *Google AI Health (formerly/now Google Brain)*.
Hosted by Dr. Edward Choi (now assistant professor at KAIST), Gerardo Flores, and Dr. Andrew Dai. Improved outcome prediction for clinical time-series data using unsupervised pretraining. Resulted in unpublished short paper *Learning Bi-Directional Clinical Event Representations: a Comparison of Architectures* (available upon request).

2018 **Research Intern**, *Sutter Health*.
Hosted by Dr. Sherry Yan and Professor Jimeng Sun. Worked on deep learning methodology for disease prediction from clinical text.

Academia

2020-2022 **Research Assistant**, *Entertainment Intelligence/Human-Centered AI Lab*, Georgia Tech.
Advised by Professor Mark Riedl on research problems centered around interpreting NLP systems with applications to text generation and commonsense reasoning.

2017-2019 **Research Assistant**, *Computational Linguistics Lab*, Georgia Tech.
Advised by Professor Jacob Eisenstein on problems such as convex optimization for incorporating lexical semantics in word embeddings and representation learning for clinical notes.

2016-2017 **Research Assistant**, *Anderson Lab*, College of Charleston.
Advised by Professor Paul Anderson on word embeddings used directly as document-level classifiers. Resulted in Bachelor's Essay "Word2Vec Inversion Methods in Topic Recognition Tasks".

Teaching

Assistantships

Fall 2021 **Natural Language Processing (CS 7643)**, *Georgia Tech*, 91 students.

Spring 2021 **Deep Learning (CS 4803/7643)**, *Georgia Tech*, 170 students.

Fall 2019 **Deep Learning (CS 4803/7643)**, *Georgia Tech*, 215 students.

Spring 2019 **Machine Learning (CS 4641)**, *Georgia Tech*, 110 students.

Mentoring

Spring 2021- **Kaige Xie**, *Machine Learning PhD student at Georgia Tech*.
Spring 2022 Met weekly. Resulted in a full paper submission (2022).

Fall 2020- **Xiangyu Peng**, *Machine Learning PhD student at Georgia Tech*.
Spring 2021 **and Siyan Li**, *undergraduate student at Georgia Tech → M.S. student at Stanford*.
Met weekly. Resulted in two workshop presentations (2020, 2021) and a full paper submission (2022).

Academic Service

Organization

- Area Chair: *EMNLP 2022*
- Workshop Organizer: *BlackBoxNLP 2022*
- Publicity Chair: *NAACL 2021*
- Student Volunteer: *EMNLP 2019, FAT* 2019, NAACL 2018*

Conference/Journal Reviewing

- ARR: *November 2021 - present; monthly*
- NAACL: *2021*
- EMNLP: *2019, 2020, 2021*
- ACL: *2018 (subreviewer), 2019, 2020*

- Transactions on Interactive Intelligent Systems (TiiS): *2022*
- AMIA Informatics: *2018, 2019*

Workshop Program Committees

- Deep Learning Approaches for Low-Resource NLP (NAACL): *2022*
- Commonsense Representation and Reasoning (ACL): *2022*
- BlackBoxNLP (EMNLP): *2020, 2021*
- Machine Learning for Healthcare (NeurIPS): *2017, 2018, 2019*

Other

- Reviewer, Georgia Tech PhD Application Support Program for underrepresented applicants: *2021*
- Reviewer, Women in Machine Learning (WiML) Workshop: *2019*