Marc Marone

marcmarone.com • mmarone1@jhu.edu • Google Scholar

Education

- **Johns Hopkins University** — **Baltimore, MD**
  Ph.D. in Computer Science, Advisor: Benjamin Van Durme — *08/2020 – Present*
- **Georgia Institute of Technology** — **Atlanta, GA**
  B.S. in Computer Science, Highest Honors — *08/2015 – 12/2018*

Experience

My research interests center around language understanding, especially translation and knowledge grounding. Lately I've become interested in efficient datastructures and tooling for understanding large language models and corresponding datasets. Check out dataportraits.org

- **Johns Hopkins** — **Baltimore, MD**
  Research Assistant — *08/2020 – Present*
  Research towards my PhD. Topics include efficient dataset documentation [1], code generation models [2], large language model factuality [3], federated learning [4], cross & multilingual semantic understanding [5], [6].
- **Microsoft - Semantic Machines** — **Bellevue, WA**
  Research Intern — *05/2021 – 08/2021*
  Researched encoders for mutilingual semantic parsing in task oriented dialogue. Used grammars and text to code models to improve multilingual performance.
- **Microsoft - Machine Translation Research** — **Redmond, WA**
  Researcher — *02/2019 – 08/2020*
  Worked on research projects involving production quality transformer models, semi-supervised machine learning, and data to text generation [7] under Hany Hassan. Built recipes for efficient domain adaptation of transformer models to customer specific data.
- **Microsoft** — **Bellevue, WA**
  SWE Intern — *05/2018 – 08/2018*
  Built a real time analysis system for diagnosing Bing service performance anomalies using Spark, Redis, d3.js
- **Quantlab Financial** — **Houston, TX**
  SWE Intern — *05/2017 – 08/2017*
  Built an anomaly detection service for analyzing quantitative trading system data using Elasticsearch, Jupyter
- **Georgia Tech** — **Atlanta, GA**
  Research Assistant — *08/2017 – 12/2018*
  Undergraduate Research Assistant with Jacob Eisenstein. Researched document level translation and summarization. Developed methods for analyzing learned representations in multilingual tagger models [8].

Teaching

- **AI Safety and Security** — **Johns Hopkins University**
  Instructor — *Present*
  Designed and taught a seminar course on AI Safety and Security. Topics include hallucinations in ChatGPT, data privacy, AI Art, and adversarial attacks. Planned lectures accessible to students with minimal computer science backgrounds.
- **Teaching Assistant** — **Georgia Institute of Technology**
  Intro. to OOP, Intro. to AI — *7 Semesters*
  Taught recitations, held office hours, and designed assignments for two courses: Intro. AI, Object Oriented Programming

---

Grants and Awards

- **Hopkins Engineering Applications & Research Tutorials** **2023**
  Designed course on AI Safety and Security topics
- **Amazon Initiative for Interactive Artificial Intelligence, grant co-author** **2022**
  Rapid Multilingual Dataset Creation with Automatic Projection and Human Supervision
- **President's Undergraduate Research Award, Georgia Tech** **2018**
- **Outstanding Sophomore, Georgia Tech** **2017**

Service and Organizing

External

- **BigCode:** Open source initiative to build public billion parameter large language models for code [2]. I worked on building the dataset and related attribution tools. We built a plagiarism detection system for real users of an AI code completion extension.

Internal

- **JHU Pre-Application Support:** Mentor PhD applicants by providing application guidance
- **CLSP Director Search Committee:** Student representative on committee to select a center director
- **Recruiting Committee:** Organize prospective PhD student recruiting visits
- **Dean Search Committee:** Sole undergraduate representative on committee to select the Dean of Computing
- **AI/ML Club:** Ran weekly presentations and workshops to club members
- **HackGT:** Organizer for a 1000+ person hackathon, presented workshops attended by 100+ students

---

Skills

Programming Languages: Python, Bash, Java, OCaml, HTML/CSS, Javascript

Software: PyTorch, Huggingface, Redis, Marian, fairseq, DyNet

---

Selected Publications

[1] M. **Marone** and B. Van Durme, "Data portraits: Recording foundation model training data," in *NeurIPS Datasets and Benchmarks*, 2023.

[2] R. Li, L. B. Allal, Y. Zi, N. Muennighoff, D. Kocetkov, C. Mou, M. **Marone**, C. Akiki, J. Li, J. Chim, *et al.*, "Starcoder: May the source be with you!" *Transactions on Machine Learning Research*, 2023.

[3] O. Weller*, M. **Marone***, N. Weir, D. Lawrie, D. Khashabi, and B. V. Durme, ""According to ..." Prompting language models improves quoting from pre-training data," *arXiv preprint*, 2023.

[4] O. Weller*, M. **Marone***, V. Braverman, D. Lawrie, and B. Van Durme, "Pretrained models for multilingual federated learning," in *ACL*, Seattle, United States, Jul. 2022.

[5] M. Yarmohammadi, S. Wu, M. **Marone**, H. Xu, S. Ebner, G. Qin, Y. Chen, J. Guo, C. Harman, K. Murray, A. S. White, M. Dredze, and B. Van Durme, "Everything is all it takes: A multipronged strategy for zero-shot cross-lingual information extraction," in *EMNLP*, Nov. 2021.

[6] S. Behzad, S. Ebner, M. Marone, B. Van Durme, and M. Yarmohammadi, "The effect of alignment correction on cross-lingual annotation projection," in *ACL: Proceedings of the 17th Linguistic Annotation Workshop (LAW-XVII)*, 2023.

[7] L. Miculicich*, M. **Marone***, and H. Hassan, "Selecting, planning, and rewriting: A modular approach for data-to-document generation and translation," in *EMNLP: 3rd Workshop on Neural Generation and Translation*, Nov. 2019.

[8] Y. Pinter, M. **Marone**, and J. Eisenstein, "Character eyes: Seeing language through character-level taggers," in *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, Florence, Italy, Aug. 2019.