Milind Agarwal

magarwa@gmu.edu
https://milind-agarwal.github.io/
Updated Sept 26, 2024

EDUCATION

George Mason University
Ph.D. in Computer Science — Aug 2022 - *present*
Research Advisor: Antonis Anastasopoulos — GPA: 4.0

Johns Hopkins University
MSE in Computer Science — 2019-2021, GPA: 3.89
BS in Computer Science — 2017-2021, GPA: 3.77
Additional Major: Applied Mathematics & Statistics
Upsilon Pi Epsilon - CS Honor Society

RESEARCH EXPERIENCE

Natural Language Processing and Machine Learning

- **Information Extraction for Low-Resource Languages** — Aug 2022 - *present*
  Advisor: Prof. Antonis Anastasopoulos, George Mason University
  - Developing robust methods for scalable data modeling in data-constrained settings and languages - specifically improving and leveraging languageID and OCR methods to create high-quality datasets and close the data gap for endangered North/South American languages
- **Lexical Discovery in 1800+ Languages** — Summer 2020
  Advisor: Prof. David Yarowsky, Johns Hopkins University
  - Constructed a lexical database of pronouns indexed by cross-lingual dimensions of meaning, to help address the data-scarcity problem in low-resource NLP, primarily using the Bible
- **VowPal Wabbit, Microsoft Research** — Summer 2020
  Mentor: Marco Rossi, Principal Data Scientist
  - Implemented a visualziation library to help Microsoft researchers analyze in-house contextual bandit customer data and forecast reinforcement learning policies.
- **COVID-19 Chatbot, Johns Hopkins CLSP** — Spring 2020
  Advisors: Prof. Joao Sedoc, Adam Poliak
  - Systematically analyzed healtchare provider and local stores and pharmacy websites to model real-time critical updates for the COVID QA Data Collection Effort

Computational Biology and Genomics (Johns Hopkins University)

- **Rare Immune Disorder Detection** (Prof. Janet Markle) — Fall 2018 - Summer 2019
  - Designed a workflow, WebSeq, to analyze multimodal data (whole exome sequencing, clinical information, pedigree charts) to greatly increase the speed and robustness of identifying disease-causing genes for rare immunological disorders
- **Multiview Computational Genomics** (Prof. Alexis Battle) — Academic Years 2018-2020
  - Developed predictive models, simulation tools, and classification systems to analyze extremely multimodal breast cancer healthcare data, and cell type deconvolution data
- **Epileptic Seizure Visualization** (Dr. Nathan Crone) — Summer 2018
  - Implemented a visualization tool to help understand seizures in the emergency unit patients (from Apple Watch data), and to identify markers for early-prediction

PEER-REVIEWED PAPERS

M. Agarwal, A. Anastasopoulos.A Concise Survey of OCR for Low-Resource Languages. *Americas-NLP @NAACL 2024* [Paper]

M. Agarwal, M. Alam, and A. Anastasopoulos. LIMIT: Language Identification,Misidentification, and Translation using Hierarchical Models in 350+ Languages. *EMNLP 2023.* [Paper]

QueerInAI et al. Queer In AI: A Case Study in Community-Led Participatory AI. *FAccT 2023*. [Paper] (Best Paper Award)

S. Ahmadi, **M. Agarwal**, and A. Anastasopoulos. PALI: A Language Identification Benchmark for Perso-Arabic Scripts. *VarDial 2023 @ EACL 2023*. [Paper]

M. Agarwal, K. Ghimire, J. Cogan, Undiagnosed Disease Network, J. Markle. WebSeq: A Genomic Data Analytics Platform for Monogenic Disease Discovery. Journal of Bioinformatics and Systems Biology 6 (2023): 01-09. [Paper]

M. Agarwal. Are We There Yet? – Building an equitable future with low-resource and endangered language research. *QinAI @ NeurIPS 2021.*

A. Poliak et al. Collecting Verified COVID-19 Question Answer Pairs. *NLP-COVID Workshop @ EMNLP 2020* [Paper]

RESEARCH AWARDS

Stanford SILICON Practitioner Award Academic Year 2024-25
Research award by Stanford SILICON to promote research for digitially disadvantaged languages
Project: OCR and AI Development for Kwak'wala and other indigeous Canadian languages

Doctoral Research Scholar Academic Year 2024-25
Research award by Mason's Provost Office to advance towards dissertation completion
Project: Digitization of the Archive of the Indigenous Languages of Latin America (AILLA)

Summer PhD Research Award Summer 2023
$8000+ grant by Mason CS for original and innovative PhD research
Project: Digitization of the Archive of the Indigenous Languages of Latin America (AILLA)

Best Project Award, Intuitive Surgical Fall 2020
$600 by Intuitive Surgical for best Deep Learning course research project
Project: Self-supervised Contrastive Image Classification with Image Sentences

Microsoft Open Source Fest Award Summer 2020
$10000 funding by Microsoft to conduct research work on VowPal Wabbit
Project: Contextual Bandits Data Visualization using VowPal Wabbit

Joseph C. Pistritto Research Fellowship Academic Year 2019-2020
$6000 grant by JHU Dept. of Computer Science for data visualization research
Project: Visualization research for multi-view data (Battle Lab)

Provost Undergraduate Research Award Summer 2019
$4000 research award to fund summer research at Markle Lab, BSPH.
Project: Developing WebSeq to find genetic causes of rare autoimmune disorders (Markle Lab)

UNDER REVIEW

M. Agarwal, J. Otten, A. Anastasopoulos. SAGE: Script-Agnostic Language Identification [Poster]
Best Paper Award at MASC-SLL 2024, Johns Hopkins University

M. Agarwal, A. Anastasopoulos. AILLA-OCR: A First Textual and Structural Post-OCR Dataset for 8 Indigenous Languages of Latin America

SHARED TASK PAPERS

M. Agarwal et al. Findings of the IWSLT 2023 Evaluation Campaign. *IWSLT @ ACL 2023* [Paper]

A. McCarthy et al. The SIGMORPHON 2022 Shared Task on Cross-lingual and Low-Resource Grapheme-to-Phoneme Conversion. *SIGMORPHON @ ACL 2023* [Paper]

PRESENTATIONS

(Invited Talk) Low-Resource Language Identification. *March 2024, Notre Dame University*

Milind Agarwal, Joshua Otten, Antonios Anastasopoulos. Script-Agnostic Language Identification. *SouthNLP 2024, MASC-SLL 2024* (Best Paper Award)

Milind Agarwal, Sina Ahmadi, Mahfuz Ibn Alam, and Antonis Anastasopoulos. Confusion-Resolution Hierarchical Models. *MASC-SLL 2023*

Milind Agarwal, Paul Mineiro, Marco Rossi. Contextual Bandit Data Visualization with VowPal Wabbit. *NeurIPS 2020.*[Slides]

Milind Agarwal, Alexis Battle. Multiplotlib: a python library for multi-view data visualization. *Richard Tapia Conference, 2020.* [Poster]

Milind Agarwal, Janet Markle. Web-Based Analytics for Genomic Data. *Johns Hopkins University CARES Symposium, 2019.* [Poster]

OTHER AWARDS

Advanced Language Processing School (*Univ. Grenoble Alpes and Naver Labs Europe*)	Winter 2022
Princeton University Prospective PhD Preview Scholar	Fall 2021
ICML Diversity and Inclusion Fellowship	Summer 2021
Johns Hopkins Life Design Fellowship	Summer 2021
Dorothy F. Sheppard Award for Outstanding Service to JHU Residential Life	Spring 2021
oSTEM/Queer in AI Fellowship for Ph.D. Applications	Fall 2020
Richard Tapia Conference Scholarship	Summer 2020

SERVICE

Reviewer: National Council of Undergraduate Research (2020, 2021), EMNLP (2023), ARR (Oct'24, Feb'24), IWSLT (2024), Queer in AI (NAACL 2024), COLM (2024)

Organizer IWSLT: Dialectal and Low-resource Track @ ACL 2023, MTMA 2023: Machine Translation Marathon in the Americas, MASC-SLL 2023, SIGMORPHON 2022, Queer in AI Conference Social Organizer (2023, 2024)