POORVI ACHARYA

Fairfax, VA | LinkedIn | GitHub
poorvi.acharya@gmail.com | (760) 532-0087

PhD candidate in NLP specializing in machine translation, psycholinguistics, and low-resource languages. Eager to contribute to innovative language technology solutions in multilingual and low-resource settings.

Education

Ph.D. Computer Science, George Mason University — Jan 2024 -
Advisor: Antonios Anastasopoulos

B.S. Electrical Engineering and Computer Science, UC Berkeley — Aug 2015 - Dec 2019

Publications

Kai North, **Poorvi Acharya**, Marcos Zampieri, and Antonios Anastasopoulos. INTO Mason Dataset, NAACL'25.

Research Experience

National Language Translation Mission
Aug 2022 - Present
Machine Translation Researcher — *IIT Kharagpur (Remote)*

- Developed low-resource neural machine translation strategies for Indian languages, integrating recent advancements in deep learning and linguistic knowledge.
- Implemented an interpretable MT framework between Hindi, Kannada, and Sanskrit, using a Pāṇinian-based interlingua. Designed a joint Hindi-Kannada neural dependency parser using Trankit, an extension of **pyTorch** (README, code).
- Gained a deeper understanding of tradeoffs involved in designing language representations.
- Introduced professional software coding practices to the project (e.g. regular code reviews, detailed documentation, and testing). Mentored junior colleagues to encourage clean and maintainable code.

UC Berkeley - Political Science Department
Aug 2019 - Dec 2019
NLP Researcher — *Berkeley, CA*

- Used topic modeling of questions posed by EU parliament members to predict euro nationalism, Euroscepticism, and populism of questions posed by Western European vs post-soviet countries (Github), using **Scikit-learn**, **Pandas**, **Numpy** and **Keras**
- Became familiar with the tradeoffs between various NLP models for shorter length texts. Read through NLP papers to find novel approaches to try such as Hidden Markov models.

UC Berkeley - Sociology Department
Aug 2018 - Dec 2018
NLP Researcher — *Berkeley, CA*

- Classified Sociology papers based on discipline (e.g., sociology, political science, management, public health) through a combination of NLP models.
- Mapped terms in concept dictionaries (containing unigrams, bigrams, and trigrams) onto n-gram files derived from journal articles. Used word embeddings (e.g., word2vec) to assess semantic similarities between concepts.

Industry Experience

Arch Systems
May 2023 - Present
Senior Software Engineer, Data Team — *Remote*

- Successfully deployed and maintained an ETL platform for industrial manufacturing in **Python**, with an uptime of **99.9%**, integrating data from various machine types to enable real-time and predictive analytics. Ingested millions of streaming **IoT** data points daily.
- Simplified complex data streaming models while exposing necessary flexibility, handling high-volume machine data and time series data structures with InfluxDB.
- Utilized **Docker** to ship software as containerized applications, facilitating seamless communication between off-the-shelf and custom components like **RabbitMQ**, **Kafka**, **Django**, **Redis**, **Postgres**, and **InfluxDB**.
- Managed development operations, implemented CI/CD pipelines using GitHub Actions, and deployed containers with **AWS ECR** and **Portainer**. Maintained telemetry systems with **Grafana** for monitoring production instances and ensuring system reliability
- Gained a deep understanding of problem domain (machine data analytics in the electronics industry) to propose scalable technical solutions to complex challenges in an industrial context
- Functioned as a technical lead, mentoring junior engineers and collaborating with cross-functional teams to deliver high-quality software solutions. Engaged in system design and architecture discussions to drive technical excellence.

Rune Labs
Sep 2020 - Dec 2022
Backend Engineer, Data Engine Platform — *NYC/SF*

- Designed, implemented, and maintained data pipeline with up-time of **99.8%**, ingesting millions of rows of **PHI** (Protected Health Information) data from different data sources using **Go** and **Python**, leveraging **AWS** infrastructure (e.g. **S3**, **SNS**, **SQS**, etc.)
- Implemented a scalable distributed tracing pipeline (**Datadog**) that ingests millions of spans across multiple back-ends, resulting in a **5%** improvement in production uptime and stability through thorough code instrumentation.
- Participated in on-call rotation. Conducted retrospectives/post-mortems of infrastructure incidents e.g. AWS outages. Redesigned incident response template to include job failure metrics, resulting in successful retry of **thousands** of jobs post-incident.
- Built admin portal with **GraphQL** API to enable restricted operations on patient data, using AWS **Cognito** for authentication and elastic load balancer (**ELB**) for scalability, increasing clinician portal ease-of-use by **10X**.
- Learned the theoretical and practical tradeoffs of NoSQL stores, specifically **InfluxDB** and **DynamoDB**. Designed and developed internal (**gRPC**) and external (**REST**) APIs to leverage these high-performing key-value and time series data stores and interface with front-end systems.
- Devised unit, integration, and load test case plans based on real-world use case scenarios to produce high-quality results and improvement in the development timeline. Improved development tools and processes (**CircleCI**, **Canary**).

Prosperata (Healthcare Analytics) June 2020 – Sep 2020
Data Science Consultant *US Remote*

- Prototyped automated detection of adverse drug reactions (ADR) from social media (Twitter and Facebook), using the FDA's adverse event reporting system (FAERS) as a baseline (**SQLite**).
- Early detection of ADR has crucial implications for health outcomes. In some instances, social media has been shown to predict ADR up to 6 months before corresponding data appeared in FAERS.
- Mined Twitter data using NER to find tweets mentioning a safety signal and corresponding disease. Looked at different statistical methods to predict presence of ADR or not.

DesignMind Business Solutions June 2019 – Aug 2019
Full Stack Intern *San Francisco, CA*

- Built a low cost database solution that automatically scrapes a company's social media data (i.e. Facebook, Google Analytics, etc.) to generate a Power BI report through Django, updated daily for machine learning purposes.

DesignMind Business Solutions June 2018 – Aug 2018
Big Data Intern *San Francisco, CA*

- Developed an optimized breast cancer prediction model using Spark, HDFS, and logistic regression with **93%** classification success rate (similar rate of false positives to false negatives).
- Built a data processing library using AWS cloudless computing services. Enabled the business developers to define custom data processing algorithms on files stored in **S3** and run through AWS **Lambda**.

Itron, Idea Labs (products and services for energy and water resource management) May 2017 – Aug 2017
IoT Intern *San Diego, CA*

- Programmed the boards that go into Itron's smart meters (gas, electric, water). Wrote **bash** scripts and tutorials for developers to read the GPIO, IC2, DAC, and ADC pins.
- Developed DSP algorithms in **C** for communication between boards.
- Presented product at IoT World conference, Austin Smart city, Microsoft deep learning hackathon.

Skills

Languages	Python, Go, C, Java, C++, R, BASH/shell, Linux (awk, sed)
Machine Learning & NLP Frameworks	PyTorch, TensorFlow, Keras, Hugging Face Transformers (NLP Models), NLTK, SpaCy, Gensim , Scikit-learn
Tools & Libraries	Pandas, NumPy, SciPy, Matplotlib, Seaborn, Plotly, Git, GitHub, Docker, Jupyter Notebook
NLP Techniques & Tasks	Tokenization, Lemmatization, POS Tagging, Named Entity Recognition (NER), Language Modeling (BERT, GPT, etc.), Word Embeddings (Word2Vec, GloVe, FastText), Sentiment Analysis, Text Classification, Summarization, Question Answering Systems, Machine Translation
Research & Statistical Methods	Probabilistic Models (Hidden Markov Models, CRFs), Statistical Analysis (t-tests, ANOVA, Regression Models), Data Preprocessing & Feature Engineering, Experimental Design and Evaluation Metrics (Precision, Recall, F1)
Databases	NoSQL, SQL, DynamoDB, S3, MongoDB, Elasticsearch, Data Scraping & Collection (BeautifulSoup, Scrapy)
Frameworks	REST API, GraphQL, gRPC
Cloud Computing & Deployment	AWS, Google Cloud, TensorFlow Serving, Flask (Model Deployment)
General software skills	unit testing, TDD, automation, design patterns, peer review, regex etc.
Human Languages	Proficient in Japanese, Korean, French, Hindi, Kannada, English (Fluent)

Current Independent Projects

Multiscript Dictionary OCR project

Sep 2020 - Present

- Designed web application to perform multilingual text recognition of images using the **Python Google Vision** API (Github).
- Presented at the UC Berkeley Google Cloud Meeting (YouTube).
- Aiming to create an application to search the dictionary using text and voice, which would be useful for language learners and linguists.

Awards

$500 prize in the NIH/AHRQ National AI/ML Challenge

- Implemented models for predicting average length of hospital stay for a county and number of inpatient stays for a county, given population statistics of that county. Minimized mean error for time series data using **scikit-learn**. Results with excellent accuracy.