Statement of Purpose

zarif98sjs@gmail.com

University X | PhD in CS

Md Zarif Ul Alam | zarif98sjs.github.io

Humans develop an understanding of the environment through multiple modalities. In contrast, language is crafted for the purposes of communication. The current understanding of natural language focuses on learning statistical language models from text-corpora. That, too, is biased towards only high-resource languages. However, humans acquire language by communicating and interacting in the real world rather than learning the meaning of a word purely based on its relationship to other words. My research goal is to bridge this gap between language and perception, enabling the creation of intelligent agents that comprehend, interact with, and reason about the multimodal world through natural language. I also aim to address this challenge in the context of low-resource settings. My research interest is in **Natural Language Processing (NLP)**. In particular, I am interested in **language grounding**, **multi-modal learning** and NLP in **low-resource** setting.

Grounded language models in different modalities

Natural language, which exists in abundance, is often complemented by the surrounding context in other modalities. For intelligent agents to comprehensively understand the world around us, they need to be able to process and associate input signals from multiple modalities, especially **non-text** ones like image, video, speech, maps, etc. While I was working at **Chaldal**, a Dhaka and California based startup backed by Y Combinator, I observed that the current state of large language models (LLMs) was mostly constrained to having simple texts as knowledge base. The research problem that I was working on was to create an address-matching platform for house addresses. This can be thought similar to Google Maps, but whose sole purpose is to find buildings only from user-provided plain-text addresses, a task at which platforms like Google Maps fail miserably. In the absence of accurate GPS data, and in some cases where the user opts out of GPS due to security concerns, plain text addresses are the only thing that can be used to locate user addresses. These addresses, when collected through text-only fields, can sometimes become very hard to locate. For example, think of an address like this - "*5th house of the 2nd road past Park Street Starbucks, yellow building*". These kinds of addresses are very hard to map to the existing canonical addresses. To solve this, I am currently working on developing a **Geospatial language model** that enhances the understanding of geo-entities with spatial context. It will help us identify user addresses within reasonable proximity, even in case of incomplete or partially wrong addresses or cases when users refer to their building in terms of landmarks in the vicinity. After taking over the project, I improved the existing Lucene-based search using **fine-tuned LLM**, **retrieval augmented generation (RAG)** and **re-ranking**. It pushed the top-1 accuracy to over 60%, which was less than 10% before. Our work is still in progress, and we plan to submit the first part of our work in **ACL '24 - Industry Track**.

Low-resource machine learning and NLP for low-resource languages

The fact that the majority of NLP's recent discoveries tend to disproportionately favor resource-rich languages, like English, is a significant issue. Although spoken by millions worldwide, languages with limited resources—like Bangla, my mother tongue—tend to lag behind considerably. These low-resource languages lack useful training attributes such as supervised data, number of native speakers or experts, etc. [1]. Throughout my journey in **low-resource machine learning**, I became increasingly aware of the problem of having limited data, which is not confined to NLP only. I was first aware of this problem when I started working on my undergraduate thesis. Our goal was to develop a non-invasive method for the early detection of Parkinson's disease (PD), given a very limited number of labeled data. Parkinson's disease comes with a stigma, and on top of that, the number of neurologists is scarce even in high-income countries, making the data collection process very challenging. A practical solution would need to take this into account instead of just assuming that we will have all the data required to train deep and large architectures. So we developed a method for detecting PD from webcam-recorded videos using **positive unlabeled (PU) learning**. PU learning enables a classifier to learn only from positive and unlabeled samples, eliminating the need for reliable negative samples, unlike supervised learning. We show when data is very limited, positive unlabeled learning performs better than its supervised counterpart in a statistically significant way. This **first-authored** work [2] is currently under review in **Phase 2** of **AAAI '24**. This work was done under the supervision of **Prof. Rahman (BUET)**, and in collaboration with **Prof. Ehsan (University of Rochester)** and **Md. Saiful Islam (Assistant Professor (on leave), BUET)**.

I also faced a similar problem during my internship at the **Xu Lab** of **Prof. Min Xu (Carnegie Mellon University)**. In the lab, my research was on **3D object detection** for Cryo-electron tomography. Annotation on this type of data is difficult and costly to obtain because of the minuscule size of the particles. The very large size of the 3D tomograms, added with a signal-to-noise ratio as low as 0.1, makes the task of 3D particle picking a very challenging vision problem. We developed a novel **semi-supervised method** for this problem, which requires a very small amount of annotated data. Here, I was responsible for both the idea development and implementation. The work is currently in the manuscript preparation phase [3].

Through these experiences, I have come to appreciate the critical importance of addressing challenges in low-resource settings. By focusing on low-resource NLP, I aim to contribute to the development of inclusive and accessible technologies. As part of my machine learning course project, I co-created the first **Bangla Plagiarism Dataset** [4]. Apart from that, my work on **Bangla Complex Named Entity Recognition** [5] won **Bangladesh's first NLP hackathon**, organized by **Amazon Web Service**. In my senior year, I, along with a few of my classmates, launched the very first **Bengali Automatic Speech Recognition** competition [6]. Through this, researchers got to work on the largest sentence-level automatic speech recognition corpus for Bangla [7]. **Kaggle** recognized our efforts, and our competition won the **Community Competition Creator Prize** worth $5000$$. Coming from a country where computational resource is scarce, we decided to donate the prize to start a new machine learning lab at our university. I also made **open-source contributions** to the most popular NLP toolkit for Bangla language, **github/bnlp**. My pursuit of low-resource NLP and low-resource ML in general, aligns with my commitment to making meaningful contributions to fields where data constraints have traditionally posed obstacles.

Research Vision & Why a PhD?

My research experiences and current interests have drawn me toward two main research directions:

- GPT (Generative Pretrained Transformer) based systems are by definition "generative". They produce artificial content based on probabilities of co-occurrences i.e they make stuff up. They are also by definition "pre-trained", which means almost all of their knowledge is non-adaptive, not acquired interactively or in real-time. Due to their correlational nature LLMs fundamentally lack robust reasoning [8]. They are comparable to "*System 1*" [1] thinking, which is fast, automotic, and unconscious, unlike "*System 2*" thinking, which is slower, logical, and reflective. Can we make language models capable of interacting with the real world with proper reasoning, allowing models to comprehend all the intricacies present across different modalities in the real world? Is it possible for models to learn interactively and continuously from these interactions and circumvent the "Narrow AI" [2] approach? Can we build a generalized multimodal multitask model for real-world interaction?
- Do the recent advancements in language processing also hold when the amount of training data is very small? If not, how can we have language models with similar capabilities for low-resource languages?

Multimodal foundation models may well be the key to artificial general intelligence (AGI) [9]. If we are able to incorporate multimodal knowledge bases, we will have a much more powerful model capable of more generalized tasks. For instance, enabling language models to grasp maps and spatial context at scale opens the possibility of employing them in searching and navigation tasks. A simple query like "*Hey Siri, show me the shortest route to my sister's home with a gas station and a gift shop for babies on the left-hand side.*" – shouldn't be an impossible query for an intelligent system, right? Current AI systems struggle with tasks requiring high-level reasoning abilities like causal understanding, logical deduction, and counterfactual reasoning, even though they achieve impressive performance on exam benchmarks [10]. My academic pursuits aside, I personally study cognitive science out of my fascination with the subject, and I look forward to exploring reasoning ability of language models inspired from it.

While new models keep coming out every week after the rise of LLMs, I want to focus on **high-level research questions** that will stay relevant in the long term. My experience in both **NLP** and **computer vision** gave me a unique perspective on the research gaps in multimodal research and low-resource machine learning. And that motivated me to pursue a PhD. After the COVID-19 pandemic, my grades took a hit due to personal health and family loss. But my research was one of the things that kept me going in this depressing time. It was my love for open-ended problems that motivated me to take on a research-focused job after my graduation and continue my research. Interning at two US labs at **Auburn** and **Carnegie Mellon University**, and collaborating with another one (**University of Rochester**) gave me a solid understanding and foundation for doing research. And now that I have a good sense of existing research gaps, I feel compelled to pursue a PhD. Coming from a country that is severely under-represented in the ML community, especially in NLP, I have always put an effort into making contributions to the very small research community I was part of. I believe a PhD will be a stepping stone for my goal to lead an NLP research lab in the industry and also to put my language, Bangla and my country, Bangladesh on the global machine learning landscape.

Department Fit & Why University X?

...

[1] https://en.wikipedia.org/wiki/Thinking,_Fast_and_Slow

[2] https://en.wikipedia.org/wiki/Weak_artificial_intelligence

References

[1] Alexandre Magueresse, Vincent Carles, and Evan Heetderks. "Low-resource languages: A review of past work and future challenges". *arXiv preprint arXiv:2006.07264*, 2020.

[2] **Md Zarif Ul Alam**, Md Saiful Islam, Ehsan Hoque, and Mohammad Saifur Rahman. "PULSAR: Graph based Positive Unlabeled Learning with Multi Stream Adaptive Convolutions for Parkinson's Disease Recognition". [project page], [preprint].

[3] Ajmain Yasar Ahmed*, **Md Zarif Ul Alam***[3] , Mostofa Rafid Uddin, and Min Xu. "TomoPicker: An Automated Particle Picking Pipeline for cryo-electron tomograms with low False-positives". In *manuscript preparation phase*.

[4] **Md Zarif Ul Alam**, Ramisa Alam, and Md. Tareq Mahmood. "Bangla plagiarism dataset". `https://huggingface.co/datasets/zarif98sjs/bangla-plagiarism-dataset`, 2023.

[5] HAZ Sameen Shahgir, Ramisa Alam, and **Md Zarif Ul Alam**. "BanglaCoNER: Towards Robust Bangla Complex Named Entity Recognition". *arXiv e-prints*, 2023. [pdf], [code].

[6] **Md Zarif Ul Alam**, Najibul Haque Sarker, Zaber Ibn Abdul Hakim, Mashiat Mustaq, Md. Shahrin Nakkhatra, Ramisa Alam, Sheikh Azizul Hakim, Asif Sushmit, Tahsin Reasat, and Mobassir Hosen. "DL Sprint - BUET CSE Fest 2022". `https://kaggle.com/competitions/dlsprint`, 2022.

[7] Samiul Alam, Asif Sushmit, Zaowad Abdullah, Shahrin Nakkhatra, MD Ansary, Syed Mobassir Hossen, Sazia Morshed Mehnaz, Tahsin Reasat, and Ahmed Imtiaz Humayun. "Bengali common voice speech dataset for automatic speech recognition". *arXiv preprint arXiv:2206.14053*, 2022.

[8] Karthik Valmeekam, Matthew Marquez, Alberto Olmo, Sarath Sreedharan, and Subbarao Kambhampati. "PlanBench: An Extensible Benchmark for Evaluating Large Language Models on Planning and Reasoning about Change". In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023.

[9] Nanyi Fei, Zhiwu Lu, Yizhao Gao, Guoxing Yang, Yuqi Huo, Jingyuan Wen, Haoyu Lu, Ruihua Song, Xin Gao, Tao Xiang, et al. "Towards artificial general intelligence via a multimodal foundation model". *Nature Communications*, 13(1):3094, 2022.

[10] Google Deepmind. "Gemini: A Family of Highly Capable Multimodal Models". `https://storage.googleapis.com/deepmind-media/gemini/gemini_1_report.pdf`, 2023.

[3] *Equal contribution.