*Please write a statement of your current academic and future career plans*

In the coming years I seek to research the relationships of complex systems using unlabeled data to build efficient, robust, large scale multi-lingual language representations. I would focus on data and learning efficiency and then build on how these methods can be scaled to multimodal and orthogonal domains. Having been following the current bleeding edge research intensely its clear to me that current approaches to Natural Language Processing (NLP) have proven to be wildly successful but are not scalable. Current state of the art systems take weeks to create, need an corpus the size of the entire internet, are difficult to recreate, and are unusable for any situation where state of the art hardware is unavailable to serve. Independent of their impressive gains, current models have scaled to unmanageable sizes (11B for the T5 system) and use the entire internet as a training corpus. Improvements in novel performance have gone from exponential to logarithmic and since mode and corpus size can’t continue to grow infinitely. I believe there is much work to be done to understand how models can learn more efficiently and datasets can be made smaller and cleaner.

Goal: research the relationships of complex systems using unlabeled data, large scale language representation

* with a focus on data and learning efficiency
* shallow focus in language systems
* current approaches in NLP have proven successful but are not scalable
  + models take weeks to create, require a corpus of the whole internet, difficult to recreate.
  + Gains in the last few years have been impressive, but models have scaled to unmanaged sizes (11b parameters), use the entire internet corpus to train, and need supercomputers to train. Current methods are unlikely to continue to scale data efficiency improves.
* I believe that there is much work to be done on understanding how small we can make datasets and still have the model learn robust representations of langauge
* Build systems that use language in as a method of communication to convey information present across modalities. Language is a tool to use to achieve means, not the end result.
* Multimodal inference.
* further expand models that work with all languages by default and can have a firm understanding of multilingually.

Given my experience

Undergraduate degree built up a strong computer science base which I could build a specialization upon

* machine learning was the first specialized course I took that really grasped my interest
* research experience at Basis Tech exploring multilingual Named Entity Resolution, solidified interest in language given bilingual background
* Higher level classes in NLP and Network Systems focused on mathematical background on field changing research inspired and prepared(Watson, PAXOS, Backrub)

Work experience at Microsoft highlighted need for research on unlabeled data and efficient training methods

* neural machine translation as a productivity gain
* I learned the basics informally by reading relevant papers and every paper they cited till I could understand papers fully. 100s of papers.
* Training of Information Retrieval Ranking Engines using hyperscale datasets
* Teacher Student model distillation
* Build methods from which 60m/year supervised data labels were created across 16 languages
* Metrics as a source of group optimization, overfitting, and efficiency
* Datasets for the community
  + Started with a data labeling project called MSMARCO
  + Built to a family of datasets
    - MSMARCO paper cited >230 times
    - OpenKP at EMNLP 2019
    - TREC 2019-2020
* Teaching others
  + AFIRM 2019, 2020
  + Hackathon JHU
* Diversity
  + Recruiting in Mexico and RPI focused on diversity.

Master’s degree at UW built foundation in NLP

* Pushed me out of my comfort area with linguistic focused classes
  + Helped make me think of concepts in a formal logic system
  + Taught me value of hand tuned rules as theoretical bounds of a system and quick methods to work.
  + Hand constructed rules and grammars are difficult to maintain scale and prone to contradictions
* Fundamentals of Applied Statistics to understand (NLP) with a focus on Deep systems
  + Program focused on cutting edge research and tracing routes back to traditional non-Deep Learned methods.
  + Computational linguistics and elective courses have taught me how to explore qualities of various systems and implement strong baselines
* Solidified my desire to continue education and try to join academic ranks
* Thesis in progress focuses on multimodal inference in open domain question answering.

-