Statement of Purpose: Master's in Language Technologies (MLT)

We live in an exciting age where Artificial Intelligence (AI) is radically transforming our lives. Today, we can find any information using search engines, have conversations with digital assistants, use translation services to comprehend most languages, and have email clients compose our emails. This promise of AI integrated into our daily lives, automating mundane tasks and empowering us to do much more, is what drew me to Machine Learning (ML) and Natural Language Processing (NLP). It has been exciting to see the recent breakthroughs in NLP, with large language models (LMs) like GPT-3 single-handedly writing code, composing narratives, and summarizing any text. Despite these advances, NLP techniques still struggle with factual correctness, lack common-sense reasoning abilities, often propagate societal biases, are less performant on low-resource languages, and have large carbon footprints. These problems limit the full potential of NLP, and tackling them will require new leaps in thinking. I wish to pursue a Master's in Language Technologies (MLT) at CMU's LTI to gain the broad knowledge and in-depth research experience necessary to address these challenges for the wider adoption of ML and NLP. I believe that my expertise in building real-world NLP systems has given me the skill-set to excel in graduate studies and has equipped me with a novel perspective to bring to the incoming class of future technological leaders at CMU.

Working as an Applied Scientist at Microsoft for over two years has helped me experience how the challenges faced by real-world applications of NLP limit the adoption of cutting-edge research. This became evident as I worked on the Suggested Replies dialog system, which assists users by providing fully-formed response suggestions to their email and chat messages. To keep training and inference times tractable, this system utilized a biLSTM model trained from scratch, and thus could not reap the advantages of transfer learning from large pre-trained models. Motivated to leverage pre-trained models, I experimented with several model compression approaches to bring down their fine-tuning and inference times and found low-cost techniques to be surprisingly effective, without needing the added computational costs characteristic of popular methods such as distillation. This work helped us deploy a fine-tuned pre-trained model, and it was exciting to see the resulting gains in user engagement. Through further experiments, I was then able to show that the size of the dataset, use of a pre-trained model, and domain adaptation of the model affect the efficacy of compression techniques and found that the large volumes of data available in the industry can make low-cost compression approaches very competitive. These findings were published in Microsoft's AI Journal and demonstrated that industrial settings of NLP problems could have optimal solutions different from those in academia. I want to carry this perspective with me to grad school to explore robust solutions that can be seamlessly adopted in both academic and industrial settings.

Leveraging user information to provide delightful, personalized experiences is crucial for the large-scale adoption of NLP. With strict user data privacy requirements and many active users with varying amounts of personal data, the industry has unique challenges and opportunities in developing novel personalization solutions. To gain exposure to this impactful area, I incorporated personalization into our GPT-2 based dialog system to tailor the model's reply suggestions to user writing styles. I represented users in the training data using unique user embeddings and then trained these embeddings as a prefix to condition GPT-2's reply generation – similar to the recent work on prefix-tuning. Since this personalized GPT-2 model could only work for the users seen at training time, I solved the cold-start problem by augmenting the model with a projection network and training it to project the sparse n-gram features from users' emails to the dense user embedding space. This projection network enabled us to generate user embeddings for new users on the fly using the n-grams extracted from a single one of their emails. Together these approaches helped us suggest personalized replies to all users. However, the large number of users in the industry makes it expensive to periodically train and update their user embeddings. Motivated by the use of prompts to guide models like GPT-3, I experimented with replacing the trained user embeddings with non-trainable user-specific prompts to induce personalized outputs and found these prompts to show superior performance. This work led to a paper submission at an upcoming NLP conference and

taught me how to utilize the distinct constraints of the industry to research novel solutions. To enable more fine-grained control of personalization of LMs, I am now exploring an extension to the low-rank adaptation of large LMs (LoRA) technique to allow direct personalized updates to GPT-2's attention weights. Despite our promising progress, open challenges remain, such as theoretically bounding the parameter updates required for different downstream tasks, developing optimal parameter-efficient transfer learning methods only tuning these minimal required weights, and building personalized models robust to adversarial attacks. The expertise I've gained through my work in analyzing the weaknesses of cutting-edge techniques and developing innovative solutions will be invaluable in making impactful research contributions to such areas during my Master's.

My broad research interests are in machine learning for NLP and greatly coincide with those of the LTI's faculty. The areas that particularly interest me are helping machines approach more human-like intelligence by grounding language and improving common-sense reasoning, as well as building more efficient, robust ML models to encode language. Language has meaning through the impact it has on the thoughts, emotions, and actions of others. Therefore grounding language in perceptual knowledge and multi-agent interactions is crucial for modeling more nuanced and meaningful semantic representations. I would love to work with Prof. Daniel Fried in this area by exploring RSA pragmatics-based speaker-listener models to hierarchically approach tasks like document-based question answering or reading comprehension, expanding pragmatics models to scenarios with multiple listeners guided by a speaker to collaboratively solve problems, and using pragmatics to model non-collaborative or adversarial language interactions. I hope to collaborate with Prof. Yonatan Bisk on further modeling theory of mind to also allow listeners to build mental representations of speakers, exploring the use of large LMs like GPT-3 to few-shot learn theory of mind models, and improving vision-and-language models for tasks such as question answering.

Commonsense reasoning remains an open problem in NLP and will be critical in allowing machines to better understand and communicate with humans, so I also wish to study ways to better encode such knowledge. My interest in commonsense reasoning aligns with Prof. Maarten Sap's recent work and I aspire to collaborate on developing stronger benchmarks for commonsense reasoning, building on work such as SocialIQA and Event2Mind, and helping machines reason about more complex social situations, such as those involving moral dilemmas, using theory of mind models. I am also deeply invested in interpreting and improving state-of-the-art (SOTA) machine learning modeling and training techniques for NLP, so I would be excited to further research areas like parameter-efficient transfer learning techniques, novel pre-training objectives, and extensions or alternatives to existing Transformer models. My work on parameter-efficient techniques like prefix-tuning and LoRA and my research on using prompts to retrieve personalized outputs relate to Prof. Graham Neubig's recent work on studying what unifies parameter-efficient transfer learning techniques and on optimizing prompts to discover what LMs really know. I would love to work with Prof. Neubig on further interpreting and improving NLP models. My research interests also align with Prof. Yiming Yang's work on the fundamentals of ML for NLP. I aspire to build on Prof. Yang's work on novel pre-training methods, such as jointly pre-training knowledge graphs and language models, and more efficient Transformer-based models, such as MobileBERT and Funnel-Transformer.

While my current role as an Applied Scientist in a product team offers opportunities to innovate and gain experience with SOTA NLP technologies, the constraint of needing to fulfill an immediate business need through every innovation doesn't provide me the flexibility and environment that I need to grow as a leader in NLP. I strive to develop NLP solutions that significantly advance the field and are generalizable across applications. This will involve thinking beyond the immediate needs of a given product. Therefore, my long-term career objective is to pursue a Ph.D. and be part of a research lab in academia or industry, working on such visionary research investments. The MLT program at CMU's LTI will enable me to gain a broader understanding of the field in a structured way, conduct in-depth research mentored by world-class faculty, and be surrounded by brilliant, motivated peers, and is thus the best next step for me to move towards my ambitious goals.