CAREER: Towards a Data-Centric Revolution in Natural Language Processing

PI: Swabha Swayamdipta **Email:** swabhas@usc.edu
Affiliation: Assistant Professor, University of Southern California

Overview

Despite the rising ubiquity of models of natural language across diverse applications, there is mounting evidence that such language models are not reliably accurate and safe to deploy. As these models are built with an almost singular focus on *scale*: models with trillions of parameters trained on terabytes of web corpora, they tend to not take into consideration the key role that *language data* plays in building reliable, safe and high performing models. The PI's long-term career focus is to develop data-centric models of natural language, where the focus is centered on how the model interacts with, learns from and is evaluated with respect to high quality data. Concretely, evaluation schemes must take into consideration the role that quality plays in the deployment of safe and reliable applications where models, for instance, do not hallucinate false information in the name of generalization, and have an interpretable internal world representation to ensure reliability of generated information.

The key challenges to be addressed are two-fold: (i) the creation of high quality datasets, either via selection from web-based corpora or via natural language generation that incorporates domain expertise, and (ii) building measures of the quality of data to be trained on, as well as evaluation measures for generated language capable to discern fine-grained distinctions towards safe deployment in reliable and trustworthy applications.

Keywords Natural language processing; Language Generation; Evaluation of Language Models; Measuring Linguistic Data Quality

Intellectual Merits

This project will advance principles of data-efficiency in training models of natural language, achieve higher generalization in performance across several tasks in natural language understanding, result in high quality datasets that incorporate domain expertise, and measures for evaluating the safety and capabilities of language models.

Broader Impacts

Our proposal could lead to the design of practical data-efficient algorithms for training large language models for industrial and academic settings, as well as evaluation metrics which will enable safe deployment of the same. The data and tools we build will be made available open source towards enabling further scientific development. Beside broader impacts in research, we will pursue educational and outreach activities intimately tied together with our research goals, towards the development of the next generation of developers intimately familiar with AI-enabled technology. Much of this research will be done in collaboration with PhD students and undergraduate researchers, and we will take concrete steps towards building a diverse research group, with members from historically underrepresented groups in computer science. We additionally aim to reach out to high-school students through partnerships in the university and the Los Angeles County Office of Education.