Extending the Reach of NLP: Overcoming the Data Bottleneck

Abstract:

Transformer-based models have revolutionized natural language processing (NLP) and significantly improved various NLP tasks. However, many researchers make implicit assumptions about their training setups, assuming that the train and test sets are drawn from the same distribution. This assumption can limit the applicability of these models across different languages and domains. The high cost of training state-of-the-art NLP models using various languages and domains has resulted in training them for only a subset of languages and domains, leading to a significant performance gap in excluded domains and languages. This performance gap marginalizes many individuals from accessing useful models. This talk will address the challenges, approaches, and opportunities for democratizing NLP across different languages and domains. Finally, we will explore future directions for making these models accessible to a broader audience.