Personal Statement - Jason Wei

I'm a recent graduate of Dartmouth College and, since October 2020, an AI Resident at Google. I'm interested in NLP, and I currently study two big questions:

(1) How can we perform data augmentation on text, and what new methods does text augmentation facilitate?
(2) What insights can NLP reveal in the social sciences?

Curiously, these two questions are actually motivated by experiences outside of NLP: my background in medical image analysis (where data augmentation thrives) and my liberal arts training at Dartmouth. By the end of this personal statement, you'll have a sense of why I'm interested in these questions, how I've studied them, my future directions, and who I am as a researcher.

Text data augmentation

Although data augmentation has been successful in computer vision, its uses in NLP have been relatively limited, perhaps because it is challenging to generate diverse and semantically-invariant examples of language. Having experienced the benefits of data augmentation in my prior medical image analysis research, however, I was inspired to ask: how might we perform data augmentation for text? In my first NLP project, I proposed four simple token-perturbation operations for augmenting text data and evaluated them on five benchmark text classification datasets. Although my proposed techniques potentially introduced semantic and syntactic adversity, I found that they still led to noticeable improvements in training–on average, training with augmented data using only 50% of the training set achieved the same accuracy as normal training using all available data. My paper was published in EMNLP 2019 and has already been cited more than 100 times, according to Google Scholar.

My first paper benchmarked data augmentation for text classification tasks, and I have since realized that data augmentation can be used not only as an add-on to boost performance but also to facilitate specific machine learning techniques. For instance, data augmentation can elegantly facilitate curriculum learning, a paradigm from computer vision that trains models on examples in order of increasing difficulty. Many data augmentation techniques are dictated by a strength parameter which controls how much noise is injected, and since more-noisy examples are harder to learn, gradually increasing augmentation strength during training is essentially a natural form of curriculum learning. I tested this hypothesis in the context of triplet-loss classifiers and found that this type of curriculum training is both faster and more accurate than standard training (paper under review).

Moving forward, data augmentation's most substantial role could be in contrastive self-supervised learning (e.g., SimCLR), which is state-of-the-art in computer vision but still limited in NLP. Can self-supervised learning techniques leverage text augmentation to outperform the masked-language model objective currently used in BERT? In future work, I plan to investigate various data augmentation objectives for text pre-training and study how the choice of objective relates to what aspects of language are learned by these self-supervised models. Moreover, I hypothesize that curriculum learning may also play an important role in improving training, since, in computer vision, harder pre-training objectives have been shown to lead to better downstream performance.

NLP for social science

During my liberal arts undergrad at Dartmouth, I took many non-major courses such as Middle East History, during which I would often wonder, can NLP techniques bring new insights into the social phenomena we're studying? One instance where my interdisciplinary aspirations came to fruition was studying bias in the Israeli-Palestinian conflict, where I learned that retellings of the same historical events often drastically differed based on whether the author was Israeli or Palestinian. I collected a dataset of historical narratives from books and newspaper articles, and I found that the phrases most indicative of narrative origin for ConvNet classifiers corroborated historians' intuitions of religious/community-based Israeli motivations and emphathy/suffering-based Palestinian claims (FLAIRS 2020 paper).

In a similar spirit, I became interested in NLP for mental health after taking a class with a psychology professor who studies the Complexity Loss Paradox, which posits that animals exhibit reduced complexity in behavior under stress. In an ongoing project, I am studying complexity loss in online therapy conversations, specifically whether the anxiety levels of therapy clients correlate with linguistic complexity measures such as lexical diversity, syntax, readability, and prototypicality. As a future direction, I plan to investigate how these cues in linguistic behavior, as detected by NLP models, can help us understand what types of behavioral therapy strategies are effective. More broadly, I have been personally inspired by the importance of mental health and hope to one day work on building NLP-driven online therapy interventions that are personalized and understand empathy.

Most recently, I conversely have been exploring how inductive biases from cognitive science can help us train better machine learning models. In a current work with Ryan Cotterell, I am studying how uniform information density (UID) theory, which hypothesizes that humans prefer information that is spread more uniformly throughout language, can be operationalized to train better language models.

Why I want to do a PhD

I have been excited about machine learning research since I first started in the field as a second-year undergraduate. When I'm actively designing a study or running experiments, I often feel a state of flow, and sometimes, even when I'm doing other things such as working out or trying to fall asleep, my mind can't seem to get away from the problems I'm working on and ideas that I'm interested in. I take this as an indicator that I will enjoy and thrive during my PhD. I expect my PhD to be a challenging yet rewarding time of rapid growth and personal self-discovery in which I will develop into a creative and mature NLP researcher.

Why research drives me

I recently found out that the techniques from my paper on data augmentation in NLP were taught in Stanford's Natural Language Understanding course (CS 224U). In fact, a group of graduate students there had developed a similar technique for question answering and had cited me in their paper, saying that my techniques had inspired them. As someone who watched my fair share of Stanford lectures when I first began studying deep learning, it was motivating to see my own research being taught in a course and inspiring others in the field. Seeing how I, even as an undergraduate student, could impact the field of NLP drives me to work harder, and I hope to continue working on projects that push the field forward and inspire new ways of thinking.

Future goals

After completing my PhD, my dream is to ultimately become a professor. While doing research as an undergraduate at Dartmouth, I was advised by several exceptional professors who met with me many times a week to advise me on research directions, teach me how to design experiments, and even help me debug low-level issues. Reflecting back, I am profoundly inspired by how these student-advisor relationships helped me grow not only as a scholar but also as an individual, and I hope to one day be able to have the same kind of impact on future students.

Why Stanford?

Stanford has a top NLP program, and I want to join one of the largest communities of talented students and professors who are interested in the same research directions as I am. Specifically, at Stanford, I could work with Noah Goodman on operationalizing inductive biases from psychology or Dan Jurafsky on developing new NLP techniques for studying social and psychological phenomena. Regarding my research on data augmentation, I would hope to bring new insights into Tatsunori Hashimoto's research on robust NLP, since data augmentation can improve model robustness to linguistic adversity.