Jason Wei Statement of Purpose

In recent years, pretraining language models on large amounts of unsupervised text has driven substantial progress on natural language processing (NLP) benchmarks. When combined with computational scale, language models gain additional benefits such as the ability to learn from fewer training examples and positive cross-task transfer in multitask settings. These high-level observations have drawn me to two broad research directions:

(1) How might we build and use large language models for downstream NLP tasks?

(2) How do we accurately characterize the behavior of such models?

During my AI residency at Google, I have been fortunate to conduct research studying these questions. This initial work and plans for future research are outlined below.

Useful Language Models at Scale

The success of GPT-3 (Brown et al., 2020) spurred the now-popular prompting paradigm, in which intentionally constructed prompts allow language models to perform various tasks via inference-time interactions only. My first project in this area (Reif et al., 2021) proposed a method that we call "augmented zero-shot prompting", which prompts a model with demonstrations of related tasks and then asks the model to perform a new zero-shot task. We explored augmented zero-shot prompting for text style transfer, where it allowed the model to successfully perform zero-shot style transfer to arbitrary styles such as "more melodramatic" or "in the style of Stephen King." This initial work demonstrated a new zero-shot capability of language models and was implemented in an internal AI-assisted writing tool that generated appropriate text re-writes given arbitrary styles specified by a user.

I found this zero-shot ability of language models on tasks such as style transfer striking given the unsupervised nature of their pretraining objective. This raised a natural question—could a small amount of supervision improve their zero-shot performance on a broad range of NLP tasks? In my most recent project, I explored a method that we call "instruction tuning", which finetunes a language model on a collection of tasks phrased as instructions, allowing it to perform unseen tasks in a zero-shot setting (Wei et al., 2021a). We took a large language model and instruction-tuned it on over 60 NLP tasks phrased as instructions. We found that the resulting model, called FLAN (for Finetuned Language Net), outperforms its unmodified counterpart in almost all zero-shot evaluation settings and compares favorably against zero-shot GPT-3. I am excited by these positive results on cross-task generalization, which motivate further research on generalist models.

While these two projects demonstrate two use cases of language models at scale, I believe there is substantial headroom for further applications. As future work, I am excited about using language models to generate entire synthetic datasets (Schick et al., 2021; Wang et al., 2021). One idea is to explore whether finetuning language models to specialize in generating datasets via instructions can allow them to generate high-quality datasets for novel tasks. Such synthetic datasets could be used to train task-specific downstream classifiers with low inference cost, and can also serve as augmented data for improving models such as FLAN. More broadly, I would also like to work on adding multi-lingual and multi-modal capabilities into these large models, as well as improving their bias, fairness, and factuality.

Characterizing the Behavior of Language Models

Though large language models achieve strong performance on NLP benchmarks, it is not clear why they do so. For instance, the striking performance of language models on targeted linguistic evaluations has raised the question of whether they acquire symbolic rules (Goldberg, 2019, and others). Prior work, however, generally does not analyze how performance depends on specific properties of training data. This motivated a case study I performed (Wei et al., 2021b) of BERT on the syntactic task of subject-verb agreement (SVA). We first evaluated the generalization ability of BERT on subject-verb pairs that never occurred in the same sentence in training, finding promising performance suggestive of symbolic learning. Using causal manipulations of pretraining data, we further qualify that BERT's SVA ability also depends on both sufficient frequency of words in the training data, as well as frequency balance between different forms of words. These results highlight how certain dataset properties facilitate desired syntactic behavior of language models.

For future work, I hope to further characterize the conditions in which language models exhibit desired generalization behavior. For instance, large language models such as FLAN show promising abilities to respond purely to instructions describing a task, even on rule-based tasks such as "rewrite this sentence in passive voice." I hope to perform causal manipulations using an instructional generalization dataset (e.g., BIG-bench) to elucidate the conditions that allow for generalization to unseen instructions. I believe that an in-depth analysis of how these models achieve compelling generalization can drive insights on how to advance instructions-based NLP.

Future Plans

My career aspiration after completing the PhD is to become a professor, as an academic career offers unique mentorship opportunities not available in industry. Reflecting on my past, I realize that I have been profoundly influenced by my advisors, who not only gave me project-specific advice but also shaped the way I think about research. I hope to pay forward this kind of positive influence.

I am interested in doing my PhD at Stanford because it is a top NLP program, and I want to join one of the largest communities of talented students and faculty who are interested in the same research directions as I am. There are many advisors at Stanford who I would be excited to potentially work with. For instance, I would be interested in working with Chris Manning, Percy Liang, or Tatsunori Hashimoto on deep learning methods for NLP. Dan Jurafsky and Christopher Potts are other advisors who I would be interested in working with, given some of my prior papers on cognitive science and computational linguistics. Finally, I am especially excited about Stanford's Center for Research on Foundation Models.

Works Cited

- T. Brown, B. Mann, N. Ryder, et al. 2020. Language models are few-shot learners. NeurIPS 2020.
- Y. Goldberg. 2019. Assessing BERT's syntactic abilities. arXiv.
- E. Reif, D. Ippolito, A. Yuan, A. Coenen, C. Calison-Burch, and J. Wei. 2021. A recipe for arbitrary text style transfer with large language models. arXiv.
- T. Schick and H. Schutze. 2021. Generating datasets with pretrained language models. EMNLP 2021.
- Z. Wang, A. Yu, O. Firat, and Y. Cao. 2021. Towards zero-label language learning. arXiv.
- J. Wei, M. Bosma, V. Zhao, K. Guu, A. Yu, B. Lester, N. Du, A. Dai, and Q. Le. 2021a. Finetuned language models are zero-shot learners. arXiv.
- J. Wei, D. Garrette, T. Linzen, and E. Pavlick. 2021b. Frequency effects on syntactic rule-learning in transformers. EMNLP 2021.