ABSTRACT

Title of proposal: Gathering Language Data Using Experts

Denis Peskov, 2021

Dissertation directed by: Professor Jordan Boyd-Graber
Department of Computer Science
College of Information Studies
Language Science Center
Institute for Advanced Computer Studies

Natural language processing needs substantial data to make robust predictions. Automatic methods, unspecialized crowds, and domain experts can be used to collect this prerequisite data. We curate large-scale conversational and question answering NLP datasets using these various methods.

A low-cost, high-output approach to data creation is *automation*. We create and analyze a large-scale audio question answering dataset through text-to-speech technology. Additionally, we create synthetic data from templates to identify limitations in machine translation. We conclude that the cost-savings and scalability of automation come at the cost of data quality and naturalness.

Human input can provide this degree of naturalness, but is limited in scale. Hence, large-scale data collection is frequently done through *crowd-sourcing*. A question-rewriting task, in which a long information-gathering conversation is used as source material for many stand-alone questions, shows the limitation of using this methodology for *generating* data. Certain users provide low-quality rewrites—

removing words from the question, copy and pasting the answer into the question—if left unsupervised. We automatically prevent unsatisfactory submissions with an interface, but the quality control process requires manually reviewing 5,000 questions.

However, certain users provide more reliable annotations and generations than others, which can be used to improve the quality control process. *Hybrid* solutions pair potentially unreliable and unverified users in the crowd with experts. As an example, Amazon customer service agents are used for curation and annotation of goal-oriented 81,000 conversations across six domains. By grounding the conversation with a reliable conversationalist—the Amazon agent—we create untemplated conversations and reliably identify low-quality conversations. But, the sentences generated from crowd workers are less natural and diverse than those from experts.

Therefore, we posit that exclusively using domain *experts* for data generation can create novel and reliable NLP datasets. First, we introduce computational adaptation, which adapts, rather than translates, entities across cultures. We work with native speakers in two countries to generate the data, since the gold label for this is subjective and paramount. Furthermore, we hire professional translators to assess our data. Last, in a study on the game of Diplomacy, community members generate a corpus of 17,000 messages that are self-annotated while playing a game about trust and deception. The language is varied in length, tone, vocabulary, punctuation, and even emojis. Additionally, we create a real-time self-annotation system that annotates deception in a manner not possible through crowd-sourced or automatic methods. The extra effort in data collection will hopefully ensure the longevity of these datasets and galvanize other novel NLP ideas.