Augmenting Conversational Data with Generative Conversational Networks

Alexandros Papangelis, Amazon Alexa AI

Conversational Artificial Intelligence (ConvAI) has seen leaps of progress in the recent past, partly due to the popularity of commercial conversational agents (Alexa, Siri, Google Assistant, and many others) that are interacting with real customers every day and partly due to advances in hardware that enabled training larger and larger models ("Language Models" or LMs). Such large LMs have revolutionized the way we do research; we rely more and more on general-purpose "pre-trained LMs (PLM)" and focus on fine-tuning their behavior to address our needs. To achieve good performance on unseen applications, however, PLMs typically require large amounts of data, which can be very costly or even unavailable. Deploying large PLMs poses another challenge, due to the large memory footprint and inference time. These are some of the reasons that have led researchers to investigate data-efficient approaches ("few-shot") and data augmentation approaches.

In our work (called Generative Conversational Networks or GCN), therefore, we are looking into controlled data generation approaches that can automatically synthesize high quality application-specific data and adapt to changes over time. We prove that for various ConvAI tasks (intent detection, slot filling, open-domain conversations, and knowledge-grounded conversations), GCN can achieve similar performance as our baselines using two orders of magnitude less data (“seed data”). We achieve that by using the seed data to generate high quality synthetic data that are used to train the downstream models. While data augmentation works well for well-defined tasks (e.g. intent detection), when it comes to generating entire synthetic conversations, we need to address an open challenge: how do we evaluate the quality of a conversation. We address this by using Reinforcement Learning to train our data generator using a variety of signals, each of which evaluates a different aspect of a conversation. An added advantage that models deployed with GCN have is that we can directly use human ratings to continuously improve the data generation / model training process.

During this talk, I will show how we use this method to create diverse and useful data for various ConvAI tasks and I will discuss challenges and current efforts on synthetic data generation.