DIALOGPT : Large-Scale Generative Pre-training for Conversational Response Generation

Yizhe Zhang **Siqi Sun** **Michel Galley** **Yen-Chun Chen**
Chris Brockett **Xiang Gao** **Jianfeng Gao** **Jingjing Liu** **Bill Dolan**
Microsoft Corporation, Redmond, WA, USA *
{yizzhang,siqi.sun,mgalley,yenchen,chrisbkt,xiag,jfgao,jingjl,billdol}@microsoft.com

Abstract

We present a large, tunable neural conversational response generation model, DIALOGPT (dialogue generative pre-trained transformer). Trained on 147M conversation-like exchanges extracted from Reddit comment chains over a period spanning from 2005 through 2017, DialoGPT extends the Hugging Face PyTorch transformer to attain a performance close to human both in terms of automatic and human evaluation in single-turn dialogue settings. We show that conversational systems that leverage DialoGPT generate more relevant, contentful and context-consistent responses than strong baseline systems. The pre-trained model and training pipeline are publicly released to facilitate research into neural response generation and the development of more intelligent open-domain dialogue systems.

1 Introduction

We introduce DIALOGPT, a tunable gigaword-scale neural network model for generation of conversational reponses, trained on Reddit data.

Recent advances in large-scale pre-training using transformer-based architectures (Radford et al., 2018; Devlin et al., 2019; Raffel et al., 2019) have achieved great empirical success. OpenAI's GPT-2 (Radford et al., 2018), for example, has demonstrated that transformer models trained on very large datasets can capture long-term dependencies in textual data and generate text that is fluent, lexically diverse, and rich in content. Such models have the capacity to capture textual data with fine granularity and produce output with a high-resolution that closely emulates real-world text written by humans.

DIALOGPT extends GPT-2 to address the challenges of conversational neural response generation. Neural response generation is a subcategory of text-generation that shares the objective of generating natural-looking text (distinct from any training instance) that is *relevant* to the prompt. Modelling conversations, however, presents distinct challenges in that human dialogue, which encapsulates the possibly competing goals of two participants, is intrinsically more diverse in the range of potential responses (Li et al., 2016a; Zhang et al., 2018; Gao et al., 2019a,b). It thus poses a greater *one-to-many* problem than is typical in other text generation tasks such as neural machine translation, text summarization and paraphrasing. Human conversations are also generally more informal, noisy, and, when in the form of textual chat, often contain informal abbreviations or syntactic/lexical errors.

Most open-domain neural response generation systems suffer from content or style inconsistency (Li et al., 2016b; Zhang et al., 2019; Gao et al., 2019c), lack of long-term contextual information (Serban et al., 2017), and blandness (Li et al., 2016a; Zhang et al., 2018; Qin et al., 2019). While these issues can be alleviated by modelling strategies specifically designed to boost information content, a transformer-based architecture like GPT-2 (Radford et al., 2018), which uses a multi-layer self-attentive mechanism to allow fully-connected cross-attention to the full context in a computationally efficient manner, seems like a natural choice for exploring a more general solution. Transformer models, for example, allow long-term dependency information to be better be preserved across time (Radford et al., 2018), thereby improving content consistency. They also have higher model capacity due to their deep structure (up to 48 layers in GPT-2) and are more effective in leveraging large-scale datasets (more than 100 million training instances) than RNN-based approaches (Vaswani et al., 2017).

* A collaboration between Microsoft Research and Microsoft Dynamics 365 AI Research.