Learning to Compare for Better Training and Evaluation of Open Domain Natural Language Generation Models

Wangchunshu Zhou, Ke Xu
State Key Lab of Software Development Environment, Beihang University, Beijing, China
zhouwangchunshu@buaa.edu.cn, kexu@nlsde.buaa.edu.cn

Abstract

Automated evaluation of open domain natural language generation (NLG) models remains a challenge and widely used metrics such as BLEU and Perplexity can be misleading in some cases. In our paper, we propose to evaluate natural language generation models by learning to compare a pair of generated sentences by fine-tuning BERT, which has been shown to have good natural language understanding ability. We also propose to evaluate the model-level quality of NLG models with sample-level comparison results with skill rating system. While able to be trained in a fully self-supervised fashion, our model can be further fine-tuned with a little amount of human preference annotation to better imitate human judgment. In addition to evaluating trained models, we propose to apply our model as a performance indicator during training for better hyperparameter tuning and early-stopping. We evaluate our approach on both story generation and chit-chat dialogue response generation. Experimental results show that our model correlates better with human preference compared with previous automated evaluation approaches. Training with the proposed metric yields better performance in human evaluation, which further demonstrates the effectiveness of the proposed model.

1 Introduction

Recent advances in sequence-to-sequence learning architecture (Sutskever *et al.* 2014) and the transformer model (Vaswani *et al.* 2017) have raised increasing interest in natural language generation (NLG) tasks, including story generation (Fan *et al.* 2018), open-domain dialogue response generation (Sordoni *et al.* 2015) and abstractive summarization (See *et al.* 2017). Despite the fast advances of models, there remains a huge gap in the evaluation of NLG models and it is hard to measure the progress due to the lack of good evaluation metrics. While perplexity is a good measure of how well a model fits some data, it does not measure performance at the desired task. Word overlap based metrics such as BLEU (Papineni *et al.* 2002), METEOR (Banerjee and Lavie 2005) and ROUGE (Lin 2004) capture quality better than the perplexity and are useful in translation and summarization. However, they still correlate poorly with human evaluation (Liu *et al.* 2016) in open domain text generation tasks including story generation and dialogue response generation because two equally good generated texts may have no n-gram overlap. Human evaluation is generally considered to be the gold standard evaluation, however, it does not scale well as it is generally expensive and time-consuming to conduct human evaluation.

Apart from measuring relative progress between different models, automated evaluation metrics also play an important role in the training stage of NLG models. It is a common practice to tune the model hyperparameter, detect convergence, perform early-stopping, and select the best checkpoints based on the model's performance on automated evaluation metrics. While acceptable for tasks where automated metrics correlate well with human evaluations, including machine translation and text summarization, this can be erroneous and result in sub-optimal training in open domain NLG tasks because available automated metrics correlate poorly with human evaluation, as demonstrated in the experimental section of this paper.

To tackle the aforementioned problems, in this paper, we propose a self-supervised approach with transfer learning to learn to compare the quality of two samples as an automated comparative Turing test. The motivation of our approach is that we can better assess the quality of generated samples or trained NLG model by comparing it with another one. Our model is a text pair classification model trained to compare the task-specific quality of two samples, which is then used to evaluate the quality of trained NLG models. As human preference annotation is generally expensive, our model is designed to be able to perform self-supervised training using only generated samples and gold reference samples without human preference annotation. When human preference annotation is available, our model can be further fine-tuned to better imitate human judgment. To evaluate the model-level quality of NLG models based on pairwise comparison in sample-level, we adopt the skill rating system similar to ELO (Elo 1978) and Trueskill (Herbrich *et al.* 2007), which is a method for assigning a numerical skill to players in a player-vs-player game, given a win-loss record of games played. In our scenario, the players are NLG models to be evaluated and a higher rating indicates a better model. The skill rating system makes it possible to evaluate all n models without needing to run n^2 matches and is able to take