**Automatic evaluation metrics for open-domain text generation**

**Motivation**: Open-domain text generation is a fundamental challenge in NLP that covers many challenging tasks including: social conversational systems (i.e., chitchat), data-to-text generation (e.g., commentary generation for sports games), generating document-long text from a prompt (e.g., GPT-3), generative long-from question answering (i.e, retrieving documents relevant to a given question and using them to generate an elaborate paragraph-length answer), multi-document summarization (i.e, extracting salient facts that are scattered across unstructured documents relevant to the question, then composing them into a faithful, and coherent text that reflects complex relations connecting input facts), structured data-to-text generation (e.g., summarizing or long-question answering from tables), generating personal stories from structured inputs (i.e., composing a creative, coherent, and plausible story that connects past events or relations), to name a few. Despite encouraging recent advances, there lacks reliable automatic [evaluation metrics](https://arxiv.org/abs/2006.14799) for conveniently benchmarking the systems and tuning hyper-parameters (e.g., GPT-3 text outputs were only evaluated with demos), which hinders fast development of the models. Popular automatic metrics such as BLEU, ROUGE, BERTScore, and BLEURT showed [low correlation with human judgement](https://arxiv.org/abs/2105.08920) for such open-ended tasks. This is because these metrics follow a reference-based paradigm that relies on similarities between systems’ outputs and human written “reference outputs'' to evaluate the generation, which fundamentally deviate from the open-ended nature of these tasks. In addition, reference-based metrics only consider semantic similarities, which ignore many other aspects of generation qualities such as fluency, coherence, fairness, faithfulness, factuality, grammatically, plausibility, etc. Developing reliable automatic metrics that can assess multiple aspects of the generation results become imperative to facilitate the progress of natural language generation (NLG).

In this proposal, we describe approaches for building reference-free automatic evaluation metrics for open domain NLG tasks. Our evaluator no longer requires human written references to assess the quality of the generation. Rather, it will solely base on the generation results to evaluate the systems, just as humans do. This does not only remove the reliance on expensive human annotations for generation evaluation, but also is fundamentally more suitable for evaluating open-ended text generation. Only a handful of work has recently investigated reference-free evaluation methods, including experimental works on summarization tasks using [contrastive learning](https://aclanthology.org/2020.emnlp-main.294.pdf) to improve linguistic quality of generated text, [classification-based approach](https://ojs.aaai.org/index.php/AAAI/article/view/6283) to predict the engagement of a conversation based on the utterances for social conversational systems, [context matching](https://arxiv.org/pdf/1907.13337v1.pdf) and [question answering (QA) models](https://aclanthology.org/2020.acl-main.454.pdf) to improve factuality and [pre-trained language models](https://aclanthology.org/2020.acl-main.460.pdf) to maximize coverage and fluency. PI Celikyilmaz’s [recent work](https://aclanthology.org/2021.findings-acl.42.pdf) that takes a closer look at some of these metrics has yielded that QA metrics generally improve over standard metrics that measure factuality, however, performance is highly dependent on the way in which questions are generated. To this end, we will extend recent work for evaluating open-domain text generation tasks and investigate various approaches including discriminator based models with adversaries, information theoretical frameworks (e.g., pointwise mutual information), to name a few.

**Impact:**

If successful, the proposed research will revolutionize the general practice of NLG model development. It will enable easy, quick, and proper comparisons among different model architectures and hyper-parameters; it will also remove the reliance on laborious human annotation. Both will significantly speed up the development cycle of NLG models for open-domain tasks. Within facebook, there are many applications that rely on open-domain text generation (e.g., BlenderBot, etc.). Our research can provide reliable tools for developers to conveniently evaluate their models and reduce costs on human annotations.

**Outcomes:**

Formally, we build evaluation metrics that take a piece of text (the output from an NLG system) as input and produce a score for it. We can similarly collect human judgements (scores) for the same piece of text. The goal is to construct automatic evaluation metrics that have high correlation with human judgements.

One critical step in training the evaluator is to compose good negative examples. Specifically, ideal negative examples should mimic common problems exhibited in generative models such that the discriminators trained on such negative examples can better distinguish human written texts from implausible machine-generated texts, and serve as an automatic evaluator. To achieve this goal, we propose a controllable generation framework that generates adversarial negative examples from strategically manipulated plots that mimic typical problems that exhibit in state-of-the-art generative models. Specifically, PI Peng has a [recent work](https://arxiv.org/abs/2104.05801) that pioneered the direction of plot-guided adversarial example generation for evaluating long-term coherence of open-domain story generation. In the course of the proposed research, we will expand this framework to study a rich set of evaluation dimensions of the open-domain generation, including factuality, safety (including fairness and toxicity), and interestingness/creativity. To We will also extract supporting related content from the structured factual, commonsense knowledge and reasoning graphs to compare against the generated content based on these metrics.

**Stretch Goal**: Time permitting, we also propose to set up the first evaluation platform to benchmark the evaluation metrics for open-domain generation. The GEM benchmark pioneered the direction of benchmarking *NLG models* with a list of tasks and evaluation metrics. In contrast, our platform focuses on pushing the development of more reliable *automatic evaluation metrics* for open-domain NLG tasks by establishing the correlation between evaluation metrics and human judgements. The platform will help benchmark the evaluation metrics we develop under this proposed research, as well as encouraging community involvement to build better automatic evaluation metrics that have higher correlations with human judgements. The successful development of reliable automatic evaluation metrics can speedup the development cycle of NLG models, identify early problems, and facilitate research in open-doman NLG. To this end, we propose to collect human judgements on several important aspects of generation outputs (e.g., factuality, safety, plausibility (based on commonsense models), content diversity, and interestingness/creativity), and compute the correlation between our proposed automatic metrics as well as the existing automatic metrics with human judgements to understand their effectiveness and limitations.