Text Generation and Evaluation for Human-Machine Collaborative Writing

Elizabeth Clark

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington
2021

Reading Committee:
Noah A. Smith, Chair
Yejin Choi
Katharina Reinecke

Program Authorized to Offer Degree:
Computer Science and Engineering

University of Washington

Abstract

Text Generation and Evaluation for Human-Machine Collaborative Writing

Elizabeth Clark

Chair of the Supervisory Committee:

Professor Noah A. Smith
Computer Science and Engineering

Natural language generation (NLG) models' ability to generate long, fluent texts has enabled progress across many NLG subfields and increased the types of contributions models can make to human-machine collaborative writing tasks. However, the improved quality of generated text also poses challenges for NLG model evaluation. In this dissertation, we develop and evaluate methods for using NLG models in a collaborative setting to offer suggestions to people as they write. We identify new modeling directions for this setting and build one such model and demonstrate its effectiveness. Finally, we improve automatic and human evaluations for long, fluent generated text, both by developing and testing new automatic metrics and by evaluating the effectiveness of human evaluations for state-of-the-art language generation models.

First, we explore the possibility of machine-in-the-loop creative writing. We performed two case studies using two system prototypes, one for short story writing and one for slogan writing. Participants found the process fun and helpful and could envision use cases for future systems. At the same time, machine suggestions do not necessarily lead to better written artifacts, and we suggest modeling and design choices that may better support collaborative writing. We explore one such direction (adding character representations as additional context for the model) and find it achieves improved generation results according to human and automatic metrics. We then consider the challenge of evaluating NLG models for collaborative writing and demonstrate how a collaborative writing platform can be used to collect pairwise, utterance-level human evaluations.

For evaluating long machine-generated texts, automatic methods avoid the collection of human judgments, which can be expensive and time-consuming. We introduce methods based on *sentence mover's similarity*; our automatic metrics evaluate text in a continuous space using word and sentence embeddings. We find that sentence-based metrics correlate with human judgments significantly better than ROUGE and can be used as a reward when learning a generation model via reinforcement learning.

Finally, we examine human evaluations of text generated by state-of-the-art models and find non-expert evaluators are unable to distinguish between human- and machine-generated text from three text domains. We explore various evaluator training methods, but find none is able to significantly improve evaluators' performance. We also find that evaluators focus on the form of the text more often than the text's content and often underestimate the capabilities of current NLG models. Based on these findings, we discuss future directions for collecting human evaluations for NLG models.

Acknowledgments

It is because of great mentors, colleagues, and friends that I have completed the work in this dissertation. First and foremost among these is my advisor, Noah Smith. I thank him for his guidance, support, and encouragement over the last six years. His advice, whether about research, professional development, or grammatical pet peeves, has been invaluable, and I will depend on it for years to come. Many thanks to my committee, Yejin Choi, Katharina Reinecke, and Gina-Anne Levow. Their perspectives and feedback strengthened this work, not only in the final dissertation but also over the entire course of my studies. I also thank Asli Çelikyilmaz for her mentorship that extended beyond my internship and continues to this day.

This dissertation is based on work that was written with the help of many others. I'd like to thank my coauthors: Noah Smith, Yangfeng Ji, Annie Ross, Chenhao Tan, Hao Fang, Hao Cheng, Maarten Sap, Ari Holtzman, Mari Ostendorf, Yejin Choi, Asli Çelikyilmaz, Karen Qin, Antoine Bosselut, Chandra Bhagavatula, Rowan Zellers, Ali Farhadi, Jianfeng Gao, Tal August, Katharina Reinecke, Sofia Serrano, Nikita Haduong, and Suchin Gururangan.

This work has also been shaped and supported by the broader UW NLP and CSE communities. Many thanks to the professors and TAs of my classes, to the students I've TAed and mentored, to the CSE staff who has made my life easier in countless ways, and to my friends and classmates. In particular, I have continually been impressed by how smart and how kind the UW NLP community is, and their academic and personal support has made me feel at home at UW and in the broader NLP community. I also owe many thanks to Noah's ARK for all their revisions of paper drafts, talk feedback, and pilot study participation. Finally, I need to thank my fellow 6th-years in my office and in ARK: Mandar Joshi, Julian Michael, Kelvin Luu, Maarten Sap, Lucy Lin, Phoebe Mulcaire, and Rahul Nadkarni. Our many conversations have been a highlight of my time at UW, and I look forward to continuing them over the rest of our careers.

DEDICATION

To my family

with all my love and gratitude.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Automatic tools have changed the way we write and collaborate, supporting writers with contributions like spelling and grammar error detection (e.g., spellcheckers like Grammarly) and enabling real-time collaboration and version control (e.g., Google Docs, Overleaf). Most writing tools, though, are focused on the writing process or a text's style or grammar; they do not involve the machine contributing directly to a text's content.

As natural language generation (NLG) models have rapidly improved, researchers are increasingly questioning the role that automatic tools can play in the writing process [Yang et al., 2019; Swanson and Gordon, 2012; Ghazvininejad et al., 2017]. Large neural models are more flexible and can handle more context than their rule-based or retrieval-based counterparts. Pretrained large models provide high levels of fluency, and finetuning and few-shot learning methods allow them to adapt to specific writing tasks and styles. Given these improvements, NLG models show potential for contributing more than surface-level suggestions to a writing task. Can they play the role of a collaborator instead, contributing content and ideas to writers?

Although the rapid development of NLG models is promising for human-machine collaboration, it has also outpaced our ability to develop effective evaluation methods for open-ended text generation. Traditional word-overlap metrics do not work well for many NLG tasks [Novikova et al., 2017] and are especially poorly suited to creative and collaborative text generation tasks. Human evaluations are typically used to evaluate generated text in these domains, but human evaluations face difficulties of their own. Current NLG models can produce long, fluent text passages, for which human evaluations are expensive and difficult to collect.

Human evaluations are also often disconnected from the end tasks and users who would be interacting with the NLG models [van der Lee et al., 2021].

In this dissertation, we explore the potential of neural models to collaborate with people as they write creative text. We examine different types of human-machine interactions and discuss directions for improving NLG models for human-machine collaboration. We also demonstrate how human-machine collaborative systems can be effectively used as evaluation platforms for NLG researchers to compare and analyze models.

We then discuss the challenge of evaluating state-of-the-art NLG models, looking at both automatic and human evaluations in NLG. We propose a new automatic evaluation metric for multi-sentence text and demonstrate how it can be used both as a text quality metric and as a reward when generating long texts. We finally turn to human evaluations in NLG and investigate how well non-expert evaluators can detect machine-generated text. Given current models' ability to generate fluent and stylistically-consistent text, we discuss the role and the future directions of human evaluations in NLG.

1.1 Background

Human-Machine Collaborative Writing

Creative applications of computing have been proposed for decades [Meehan, 1977; Ryan, 2017; Riedl et al., 2021], and past human-machine collaborative tasks include improvisational music [Hoffman and Weinberg, 2011; Quick and Thomas, 2019] and dance [Jacob and Magerko, 2015]. In natural language generation, past work has proposed collaborative writing tasks ranging from collaborative poetry writing [Ghazvininejad et al., 2017] to headline writing [Gatti et al., 2016] to story writing [Swanson and Gordon, 2012]. Because previous collaborative writing systems have suffered due to limited abilities to handle long contexts and generate fluent text, as NLG models improve, human-machine collaborative writing becomes increasingly viable. Collaborative writing systems for creative writing tasks, particularly for story writing, have risen in popularity and are currently an active area of research.

While early story generation models were based on rules and templates [Ryan, 2017; Meehan, 1977], neural models' ability to generate open-domain text and to adapt to new text styles through finetuning or few-shot training has resulted in their prevalence in recent work in both standalone and collaborative story

generation. Large neural language models can be used directly to generate stories [See et al., 2019], but past work has also incorporated story elements into the generation models, such as the story's structure [Fan et al., 2018], characters [Clark et al., 2018a], and events or plot points [Rashkin et al., 2020; Martin et al., 2018]. Collaborative story writing systems have also leveraged the flexibility of neural models, allowing writers to take turns with a neural model to write a story [Clark et al., 2018b] or request on-demand story suggestions [Roemmele and Gordon, 2018; Akoury et al., 2020].

As researchers develop new models for human-machine collaborative writing, evaluation remains a challenge. As in other areas of NLG, human evaluation is considered the gold standard evaluation in story generation and other creative generation domains, though the use of automatic measures to evaluate story quality has also been explored [Roemmele et al., 2017; Purdy et al., 2018; See et al., 2019; Guan and Huang, 2020]. Most evaluations of collaborative writing systems depend on writers' individual or system-level ratings using Likert scales, though some also include additional analyses, e.g., the writers' edits to the generated text [Roemmele and Gordon, 2015; Akoury et al., 2020].

Automatic Evaluation of Generated Text

Although n-gram overlap metrics like BLEU [Papineni et al., 2002] and ROUGE [Lin, 2004] have become commonplace in NLG evaluation beyond machine translation and summarization tasks, they only capture similarity when information is repeated verbatim between a source and reference text. One way to capture a more nuanced notion of similarity is to represent each text as a collection of word embeddings and consider the collective distance between the word embeddings in a reference text versus a generated text, as in word mover's distance [Kusner et al., 2015]. Due to the improvements in language representation enabled by large pretrained models like BERT [Devlin et al., 2019], this evaluation approach has become increasingly popular, resulting in evaluation metrics like the BERT-based BLEURT [Sellam et al., 2020], BERTScore [Zhang et al., 2020b], and MoverScore [Zhao et al., 2019].

In cases where there is no reference text to compare the generated text against (e.g., a story generated in a collaborative setting) or where a reference text may be just one of many suitable responses (e.g., a dialogue agent's answer to a chitchat question), automatic analysis is still possible, often focusing on different dimensions of the text's quality rather than its similarity to a gold reference text. For example, metrics like

self-BLEU [Zhu et al., 2018] and distinct-n [Li et al., 2016] measure the diversity of generated text, and sentence length and the distribution of the generated words' parts-of-speech have been used as proxies for the complexity and style of a text [Roemmele et al., 2017; See et al., 2019].

Human Evaluation of Generated Text

Human evaluations are considered the gold-standard evaluation method for many NLG tasks [van der Lee et al., 2021]. Though extrinsic human evaluations (i.e., downstream evaluations performed by end-users [Belz and Reiter, 2006]) are encouraged, they are rare in practice, with most evaluations consisting of ratings or rankings of a text's intrinsic qualities [van der Lee et al., 2021]. In most cases, evaluators assign generated text a rating along a Likert scale; common evaluation dimensions include "overall quality," "fluency," and "grammaticality" [van der Lee et al., 2021].

Despite the importance of human evaluations in NLG, there is little consensus in how human evaluations are run in NLG, even within a single task or domain. There is little consistency in the evaluation format, the types of evaluators, and the text quality dimensions that are evaluated (and even when evaluating for the same text dimensions, they can be defining these terms differently) [van der Lee et al., 2021; Howcroft et al., 2020]. These problems are exacerbated by a lack of reporting; descriptions of human evaluation procedures are often underspecified or missing altogther. For example, van der Lee et al. [2019] found only 55% of papers in their survey included the number of evaluators. The inconsistencies in human evaluations, obstacles to transparent and replicable research, have led to recent efforts to better document and clarify human evaluation procedures [Howcroft et al., 2020] and to build human evaluation platforms to standardize human evaluation results [Khashabi et al., 2021; Gehrmann et al., 2021].

1.2 Challenges

The high-level goal of many NLG tasks is to interact with and support people as they complete tasks, but in practice NLG problems are often simplified to static input-output tasks. While this allows rapid model development and evaluation, models trained to complete simplified tasks and evaluated with automatic metrics approximating desired behavior may be mismatched with the goals and expectations of an end user. One challenge is understanding what these user goals are and how current NLG models are or are not aligned

with them.

When modifying NLG models, we want to make sure the changes are improving our models in meaningful ways and in directions that benefit end users. To do this, we need good evaluation methods. Generated text can be evaluated by both automatic metrics and human evaluations, but human evaluations can be expensive and time-consuming, particularly when dealing with long texts.

Most popular automatic metrics require reference texts, "correct" answers against which they compare the generated text. However, in open-ended and creative tasks, there are rarely reference texts, and even when there are, they do not cover the space of all possible "correct" answers. In collaborative generation settings, this is especially true as the models are generating text for a brand-new and dynamic context; any reference texts would need to be collected post-hoc. For these reasons, human evaluations are most often used to evaluate NLG models for open-ended text generation tasks.

Human evaluations are typically treated as the "gold-standard" for evaluating generated text, but there is relatively little discussion of these methods within NLG and how to improve them. In fact, papers frequently omit details of how human evaluations altogether [van der Lee et al., 2019]. While our NLG models have greatly improved, our methods for evaluating them via crowdsourcing have not. The increasing fluency and length of text that current models are able to generate pose challenges to traditional approaches to collecting human evaluations of generated text.

1.3 Approach

To address these challenges, we first consider NLG models and evaluation for human-machine collaboration (Chapters 2-4), before discussing evaluation in NLG more generally (Chapters 5-6).

To see how NLG models can contribute to creative writing, we discuss a framework for human-machine collaboration and use it to design two human-machine collaborative writing systems. We collect participants' feedback about their experience writing with the help of generated text to better understand the strengths and weaknesses of current models in collaborative settings. Based on this feedback, we identify several directions for improving NLG models for collaborative writing, one of which we implement and find improves the generated text.

To address the challenge of NLG model evaluation, particularly in collaborative settings, we demonstrate

how the human-machine collaborative writing platform can also be used for pairwise model evaluation, providing utterance-level paired human evaluations of model quality. We then consider the broader challenges in evaluating state-of-the-art generation models and their ability to generate long, fluent texts. We propose an automatic evaluation metric designed for multi-sentence text passages by measuring similarity at the sentence level. To address the challenge that high-quality generated text poses for human evaluations, we analyze the effectiveness and the focus of non-expert human evaluators when identifying text generated by state-of-the-art NLG models. We explore different evaluator training methods to overcome this challenge, but find them unsuccessful, leading us to recommend future directions for human evaluations in NLG.

1.4 Outline

In Chapter 2, we first consider the role NLG models can play in the creative writing process. We present a "machine-in-the-loop" framework for human-machine collaboration on a creative writing project and run user studies with two different creative writing systems. Based on participants' feedback, we discuss the challenges that are specific to collaborative generation models that are not currently addressed by general-purpose language models and recommendations for improving these models. We explore one such improvement in Chapter 3, using character information to improve generated text. We show that incorporating representations of a story's entities as additional context can improve performance on several generation-related tasks. In Chapter 4 we see how human-machine collaborative writing systems can also be used as an evaluation platform for generation models.

We then discuss current evaluation methods in NLG and how they perform on state-of-the-art NLG models and tasks. In Chapter 5, we introduce an automatic evaluation method for handling long, generated texts. "Sentence Mover's Similarity" is a metric based on word- and sentence-embeddings, rather than word overlap metrics like BLEU and ROUGE, allowing it to capture a more nuanced sense of similarity. We find that the sentence-based metrics perform well at automatically evaluating both machine- and human-authored text, correlating with human judgments better than ROUGE. Furthermore, we show how the metrics can also be used directly in the generation process by incorporating them into a summarization model's loss function.

Finally, in Chapter 6, we discuss human evaluation practices in NLG. We find that non-expert human evaluators struggle to distinguish between human-authored text and text generated by state-of-the-art models

and that they often focus on the form of the text rather than its content. We explore three methods for training evaluators at this task, but their limited success points to the need for new human evaluation methods in NLG.