GRACE: Gradient-guided Controllable Retrieval for Augmenting Attribute-based Text Generation

Zhihua Wen, Zhiliang Tian[*], Zhen Huang, Yuxin Yang, Zexin Jian, Changjian Wang, Dongsheng Li[*]
College of Computer, National University of Defense Technology, Hunan, China
{zhwen, tianzhiliang, huangzhen,
yangyuxin21a, jianzexin21, wangcj, dsli}@nudt.edu.cn

Abstract

Attribute-based generation methods are of growing significance in controlling the generation of large pre-trained language models (PLMs). Existing studies control the generation by (1) finetuning the model with attributes or (2) guiding the inference processing toward control signals while freezing the PLM. However, finetuning approaches infuse domain bias into generation, making it hard to generate out-of-domain texts. Besides, many methods guide the inference in its word-by-word generation, pushing the word probability to the target attributes, resulting in less fluent sentences. We argue that distilling controlling information from natural texts can produce fluent sentences while maintaining high controllability. In this paper, we propose **GRA**dient-guided **C**ontrollable r**E**trieval (GRACE), a retrieval-augmented generation framework to facilitate the generation of fluent sentences with high attribute relevance. GRACE memorizes the semantic and attribute information from unlabeled corpora and applies a controllable retrieval to obtain desired information. For the generation, we design techniques to eliminate the domain bias from the retrieval results and integrate it into the generation model. Additionally, we propose a gradient-guided generation scheme that iteratively steers generation toward higher attribute relevance. Experimental results and quantities of examples verify the effectiveness of our method.

1 Introduction

Controlling the text generation model toward a specific direction remains an active research area, covering many tasks, including storytelling, text debiasing, and attribute-based generation (Xu et al., 2020; Liu et al., 2021; Dathathri et al., 2019). Attribute-based text generation requires generating text that satisfies the given attribute, which is a control code for a specific topic, sentiment, or style (Prabhumoye et al., 2020; Zhang et al., 2022). Pre-trained language model (Radford et al., 2019) (PLM) can generate fluent texts by learning on large corpora but is difficult to control because it does not learn to adapt controlling signals.

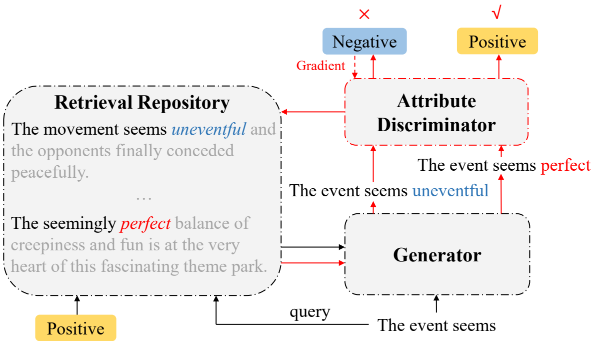
×
√
Negative
Positive
Gradient
Retrieval Repository
The movement seems *uneventful* and the opponents finally conceded peacefully.
…
The seemingly *perfect* balance of creepiness and fun is at the very heart of this fascinating theme park.
Attribute Discriminator
The event seems perfect
The event seems uneventful
Generator
Positive
query
The event seems

Figure 1: The idea of GRACE. Black lines indicate the first phase (i.e. attribute-based text generation augmented by the retrieval). Red lines indicate the gradient-guided generation to revise the previous generation.

Some researchers re-train a PLM supervised with control signals (Keskar et al., 2019; Zhang et al., 2020) or fine-tuning on domain-specific data (Bakker et al., 2022). CTRL (Keskar et al., 2019) pre-trains with texts from the Internet and extracts control code from URLs. PPVAE (Duan et al., 2020) fine-tunes part of the parameters for the target condition to bridge the conditional latent space and the global latent space. These methods bring high controllability and fluency to the generated text by modeling the relationship between the attribute and its contexts from supervised data. However, attribute-based supervised datasets usually derive from some specific domains (see App. F). Fine-tuning on those datasets brings in not only attribute information but also domain bias. The generated texts, without eliminating the domain bias, likely fall into the specific domain and lack the generalization ability across domains. Besides, the computational overhead of re-training a large PLM is becoming increasingly expensive (Liu

[*] Corresponding Authors.

Findings of the Association for Computational Linguistics: ACL 2023, pages 8377–8398
July 9-14, 2023 ©2023 Association for Computational Linguistics