# 아주대 - GTA Gated Toxicity Avoidance for LM Performance Preservation 논문리뷰

논문 - 아주대(조현석교수님) · 2024. 9. 18. 18:40

GTA Gated Toxicity Avoidance for LM Performance Preservation 논문리뷰(아주대)

Comments

•전반적으로 이전 연구에 비해 연구의 수준이 높지는 않았음

•또한 설명이 부족한 부분이 매우 많았음

•옆의 예시는 GeDi라는 모델을 설명하고 있지만, 모델에 대한 설명이 (어떻게 작동하는 지 등) 이 전무함

Introduction

•언어 모델은 다양한 분야에서 놀라운 성과를 내고 있음

•하지만 때때로, 언어 모델은 공격적이거나 불쾌한 대답을 만들어내기도 함

•이는 언어 모델이 학습한 데이터 (인터넷 등)에서 나온 것들을 그대로 따라하는 것

•따라서 이러한 문제를 적절히 교정할 필요성이 있음

•예를 들자면, 마이크로소프트가 2016년에 Tay라는 챗봇을 출시하였음

•이때 Tay는 욕설과 차별적인 발언을 포함한 답변을 많이 생성하였고, 이후 서비스가 일시적으로 중단되었음•

Background

•방법 1: Ranking

•하나의 Prompt가 주어졌을 때, 여러 개의 대답을 생성한 다음 그중에서 가장 less-toxic 한 답변을 고르는 것

•방법 2: Text Style Transfer

•어떤 toxic 한 대답이 있을때, 이걸 언어 모델을 사용해 교정하는 것

-> 일반적으로 답변을 생성하는 것보다 더 많은 시간과 비용이 듬

해결책

•방법 3: Controllable Text Generation (CTG)

•언어 모델의 대답 방향성을 직접적으로 조절하는 것(특정방법은 없음. 다양한방법사용)

•Prompt를 적절히 조절하기, 원하는 방향성으로 Fine-tuning 하기, Token 예측 확률을 직접 조절하기 등이 있음

Background (PPLM)

•예를 들자면, 언어 모델이 toxic한 토큰을 예측할 확률을 강제로 내려버리는 것임

•반대로, non-toxic한 토큰의 확률은 증가시킴

•다른 검열 모델 (toxic discriminator)가 매 예측마다 현재 생성된 글이 toxic한지 아닌지 확인

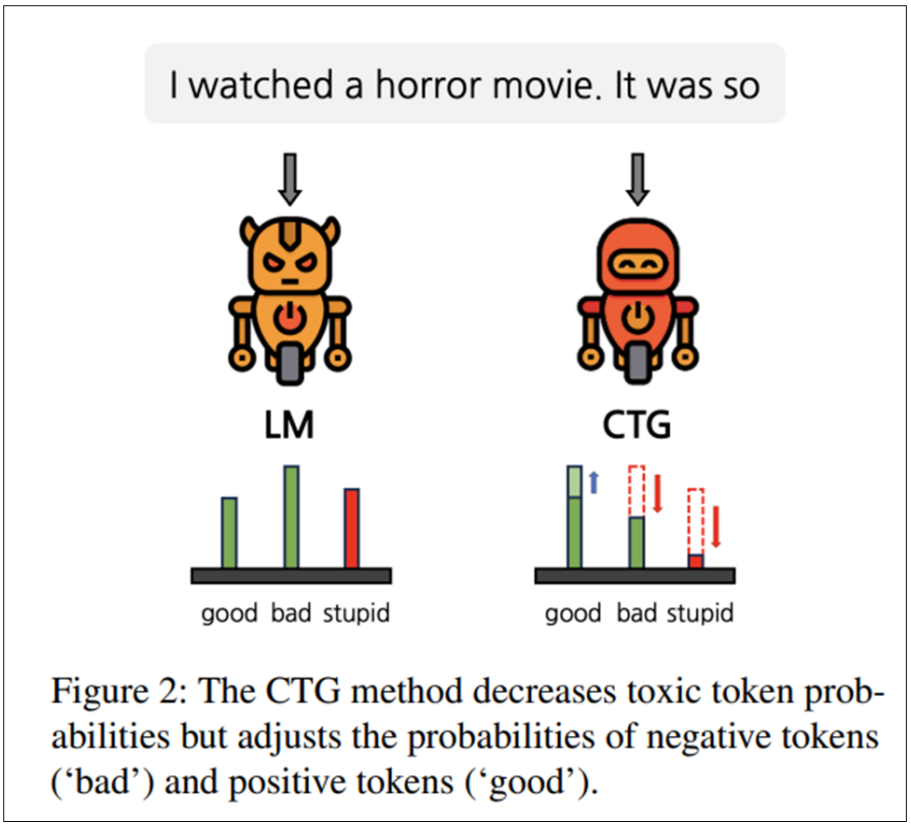



Figure 2: The CTG method decreases toxic token probabilities but adjusts the probabilities of negative tokens ('bad') and positive tokens ('good').

•Discriminator는 여러 개의 toxic / non-toxic 한 글들로부터 훈련됨

•Sentiment analysis(감정분석)와 비슷함. 어떤 글이 주어졌을 때, 이 글이 얼마나 toxic한지 점수로 매기는 것

•이때 이 toxicity 점수를 error로 생각하고, 각 token의 등장 확률을 backpropagation을 통해 조절하는 것

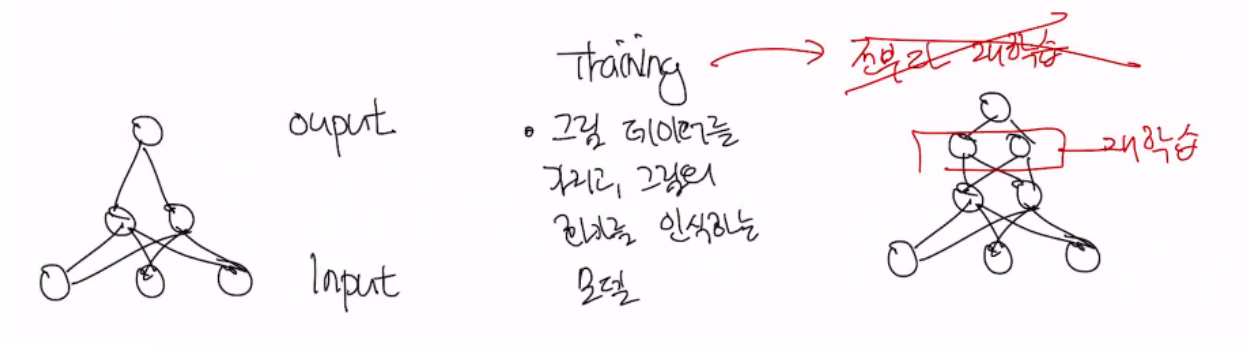
•이는 언어 모델을 Fine-tuning하는 것과는 (비슷하지만) 다름

-> (언어모델 자체를 수정하는게 아니라 주어진 맥락에서 수정)

•Discriminator와 언어 모델은 서로 독립적으로 작동

•다양한 종류의 Discriminator를 학습시켜 언어 모델의 작동 방식을 효과적으로 조절할 수 있음

아래그림) fine-tuning(모델자체를 수정하는게 아니라 1회성으로 수정)




•CTG를 통해 생성된 (검열된) 글은 원래 의도했던 내용을 담고 있지 않은 경우가 많음

•예시: The movie was (bad -> good)

•또한 이렇게 단어를 바꾸다 보니 문법적, 의미적으로 말이 안 되는 글을 생성하는 경우도 많음

and staff was not kind --> 라고
문법적으로 맞게 써야되는데 문법틀리게 수정됨

아래그림)


•다만 이러한 문제를 완화시키기 위해 CTG의 강도를 줄이면 (검열을 약하게 하면) 언어 모델이 말이 되지만 toxic한 답변을 할 확률 또한 증가함

•Toxic한 답변을 피하는 것과 언어 모델의 성능을 유지하는 것 사이에는 Trade-off 관계가 있음

•이전 CTG의 문제점은, 모든 Token 예측에 Discriminator를 적용하는 것임
•만약 예측된 Token이 문제될게 없다면, 굳이 적용할 필요가 없음
•따라서 연구진은 주어진 글이 Toxic한지 아닌지 구분하는 모델 (Gate) 을 사용해, 꼭 필요할 때만 CTG를 사용하기로 함

I dislike you --> toxic -> 검열받고 -> 공격적 답변 낮추고, 긍정적 답변으로 높임

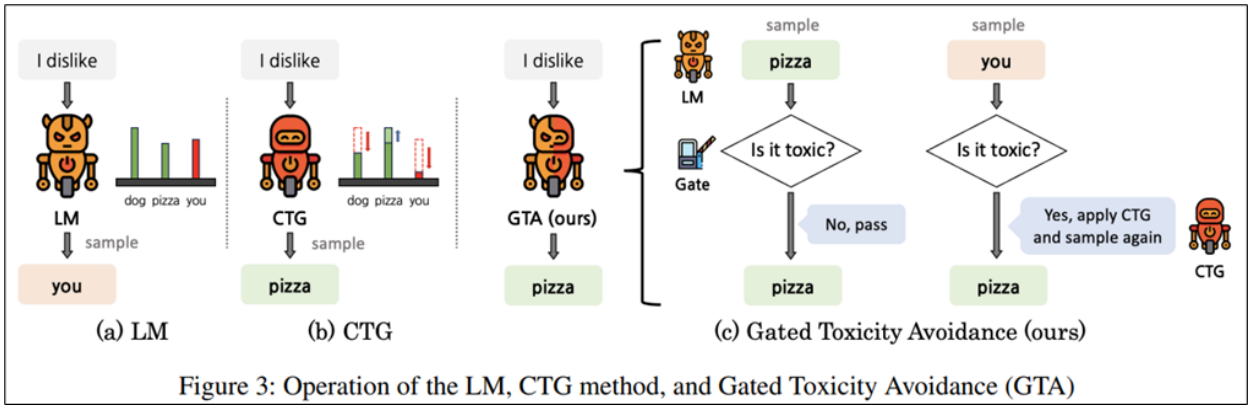


Figure 3: Operation of the LM, CTG method, and Gated Toxicity Avoidance (GTA)

•이때 Gate 모델은 기존에 존재하는 Toxicity detector 를 사용함

• Roberta 기반의 모델 (https://huggingface.co/s-nlp/roberta_toxicity_classifier)

•이때 주어진 글의 점수가 0.5 이상이면 (0~1사이) toxic 한걸로 간주하고, CTG를 사용함

•다만 의문이 드는 것은, 어짜피 기존 CTG 중에서도 내부적으로 주어진 글이 toxic한지 아닌지 예측하는 것들이 있음

•PPLM에서도 일반적으로는 주어진 글이 toxic한걸로 예측될 때만 Token 확률을 조정함

•또한 이 연구에서는 컴퓨팅 성능의 한계로 GPT-2를 사용하였는데 이는 2019년 2월에 공개된 모델임

•그 당시에는 매우 충격적인 성능을 보여주었지만, 현재는 매우 많이 뒤쳐진 모델 (사실상 아무도 사용하지 않음)

Evaluation

•CTG를 적용하기 전과 후를 비교했을 때

•Accuracy: 얼마나 원래 말하려던 주제에 충실한지
•Grammar: 얼마나 문법적으로 정확한지
•Toxicity: 얼마나 공격적이고 부적절한 표현이 있는지
•Perplexity: 얼마나 모델이 예측할 때 확신을 가지고 하는지(확신을 갖고 풀때 정답확률높아짐, 얼마나 당황을 하는지)
•Perplexity를 제외한 세가지는 현재 공개되어있는 모델을 사용해서 점수를 매김

•세가지 주제를 사용

•Sentiment (Positive / Negative)

•Emotion (Anger / Sadness / Fear / Surprise / Joy / Love)

•News (Tech / Sport / Entertainment / Politics / Business)

•언어 모델 (GPT-2) 에게 위 주제에 대한 글을 생성하도록 시키고 (CTG를 적용한 것과 안한 것),

생성된 글이 얼마나 주어진 주제와 일치하는지 확인 (Accuracy)

아래그림)

예시: I feel so angry -> happy
Accuracy :낮아짐
Tocxity: 낮아짐
Grammar: 크게 떨어지지않음. 다른 문맥에서는 떨어짐
Perplexity: 당황성 높아짐

•전반적으로 CTG를 적용했을 때 주어진 글들이 문법적으로 부정확해지고 주어진 주제를 잘 따라가지 못하는 경향을 보임

•대신 Toxicity는 확실히 낮아짐

•다만 원래 GPT-2가 매우 오래된 모델이라 이러한 결과가 요즘 언어 모델에도 적용될지는 미지수

| Topic Group | Method | Accuracy (↑) | Toxicity (↓) | Grammar (↑) | PPL (↓) |
|---|---|---|---|---|---|
| Sentiment | GPT-2 | 76.25 | 0.3 | 88.22 | 4.72 |
| | PPLM | 75.05 | 0.5 | 50.05 | 38.21 |
| | GeDi | 74.35 | 0.05 | 82.09 | 6.39 |
| | DExperts | 73.55 | 0.05 | 69.65 | 12.41 |
| Emotion | GPT-2 | 62.98 | 3.8 | 88.31 | 8.27 |
| | PPLM | 47.19 | 1.05 | 45.91 | 24.44 |
| | GeDi | 56.3 | 0.03 | 91.06 | 7.25 |
| | DExperts | 51.8 | 0.07 | 28.18 | 55.88 |
| News | GPT-2 | 80.26 | 0.12 | 78.63 | 9.15 |
| | PPLM | 84.58 | 0.04 | 38.93 | 34.95 |
| | GeDi | 80.04 | 0.02 | 78.8 | 9.5 |
| | DExperts | 72.04 | 0.02 | 46.34 | 54.48 |
| Average | GPT-2 | **71.67** | 1.85 | 84.57 | 7.89 |
| | PPLM | 65.86 | 0.58 | 43.86 | 30.04 |
| | GeDi | 68.21 | **0.03** | **84.97** | **7.89** |
| | DExperts | 62.93 | 0.05 | 41.54 | 43.9 |

아래그림

PPLM (파란색) - 말이 안되는 언어

GeDi - 문맥에안맞는말함

Table 2: These are cherry-picked sample texts that LM and CTG methods generated for an anger topic. Red text indicates a toxic text. Blue text indicates that the quality of the sentence has dropped sharply. Green text indicates text which is not toxic.

| Method | Cherry-picked samples | Anger probability (%) |
|---|---|---|
| GPT-2 | i have no idea what the fuck is going on when you feel the need to be so stubborn | 99.2 |
| PPLM | im feeling kind– Ic [o._: (I, I (u to n: a a you o l to a to b to d f u b o e n s r i s t t u l a h u s t e s u l u i t s s i s t s u | 0.2 |
| GeDi | i feel i am not a saint | 0.2 |
| DExperts | i don t feel angry at the plants in yenell and i get mad at the birds and mosquitos and spiders and the bees and bugs and bugs and the stuff i do the most in these fields gets the large ones away and i can see it kind of pitted back together in it out and i try to get | 99.4 |

아래그림)

GTA(Gated Toxicity Avoidance) -

•이에 반해 Gate를 사용했을 경우, Accuracy, Grammar, Perplexity 등이 좋아지는 경향을 보임(선택적 검열)

•또한 답변 생성 속도도 큰 폭으로 감소하였음

Table 3: Overall performance of gated toxicity avoidance

| Method | Accuracy (↑) | Toxicity (↓) | Grammar (↑) | PPL (↓) |
|---|---|---|---|---|
| GPT-2 | 71.67 | 1.85 | 84.57 | 7.89 |
| PPLM | 65.86 | 0.58 | 43.86 | 30.04 |
| $\text{PPLM}_{GTA}$ | 72.65 | 0.62 | 85.26 | 8.73 |
| GeDi | 68.21 | 0.03 | 84.97 | 7.89 |
| $\text{GeDi}_{GTA}$ | 71.12 | 0.03 | 84.22 | 7.9 |
| DExperts | 62.93 | 0.05 | 41.54 | 43.9 |
| $\text{DExperts}_{GTA}$ | 70.96 | 0.05 | 84.33 | 7.91 |

아래그림)

회색:검열에 따른시간변화

파란색: 선택적 검열에 따른 시간변화  -> 시간이 많이 줄어듬. 불필요한 시간이 줄어듬. 효율적

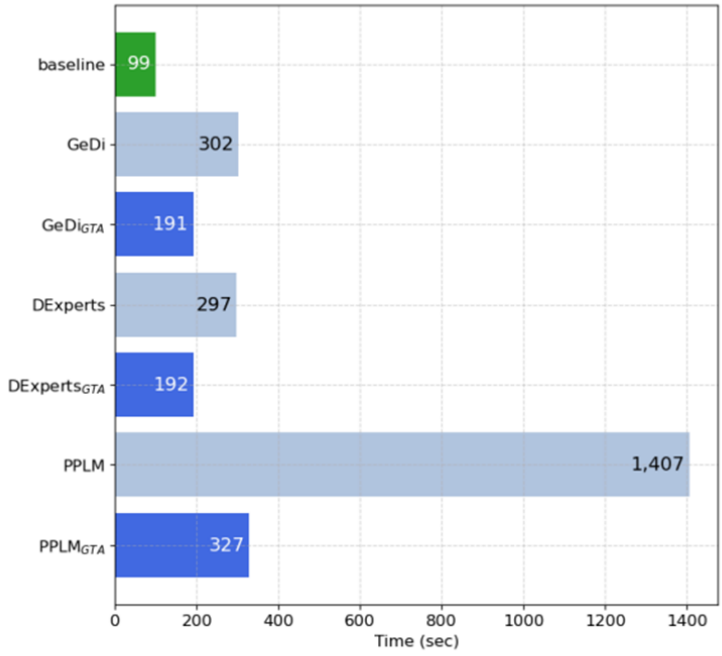



Figure 5: Time to generate 100 texts 100 tokens in length

아래그림)

•동일한 Evaluation을 사람들을 통해 진행함

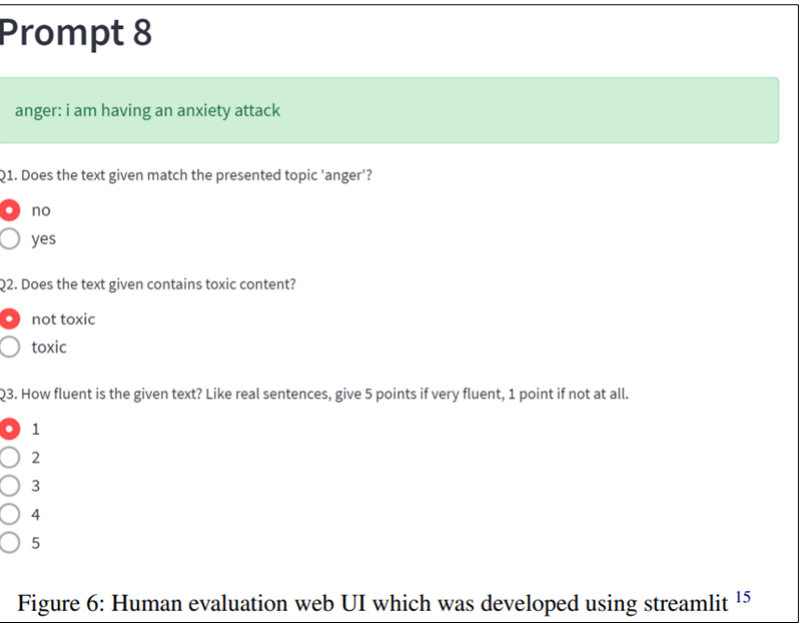



Figure 6: Human evaluation web UI which was developed using streamlit [15]

아래그림)

GPT-2가 Vanilla Accouracy 76.15%, Toxicity 4.23, Fluency: 3.62

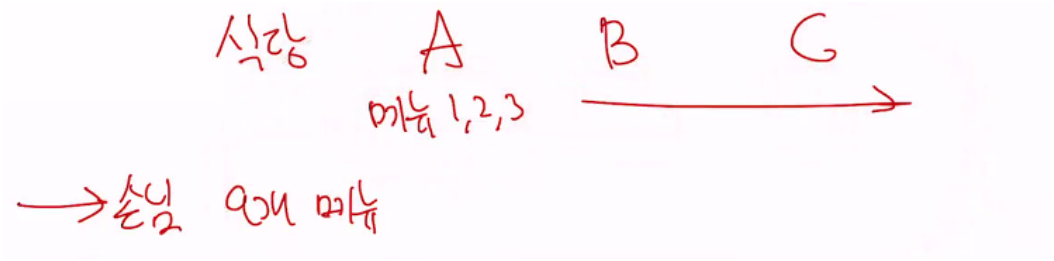
GTA 적용되면 정확도는 유지되고, 공격성은 유의미하게 낮아짐(원래 문제는 정확도는 떨어지고 공격성이 떨어짐)

Table 6: Human evaluation result on large-scale LM

| Method | Accuracy (↑) | Toxicity (↓) | Fluency (↑) |
|---|---|---|---|
| GPT-2 | 76.15 | 4.23 | 3.62 |
| GeDi | 65.77 | **1.15** | **3.73** |
| $\text{GeDi}_{GTA}$ | 75.77 | **1.15** | 3.57 |
| DisCup | 70.77 | 1.54 | 3.56 |
| $\text{DisCup}_{GTA}$ | **76.54** | 2.69 | 3.69 |

예를들어

A,B,C식당의 각 메뉴1,2,3 씩 만들어서 손님들한테 평가를 받아서 높은 메뉴를 찾는식




## Conclusion (Summary)

•방법 3: Controllable Text Generation (CTG)

•언어 모델은 때때로 공격적이거나 불쾌한 답변을 생성하는 경우가 있음 (toxic한 답변)
•이를 제어하기 위해 CTG Controllable Text Generation (CTG)가 흔히 사용됨 (Prompt 조절, Token 확률 조절 등)
•다만 CTG는 생성된 답변이 의미적 / 문법적으로 적절하지 않게 만들 가능성이 있음. 또한 추론 시간도 증가함

•따라서 주어진 글이 toxic 할때만 선택적으로 CTG를 적용하면 이러한 문제를 (일부) 극복할 수 있음

•GPT-2를 사용해 실험했지만, 사실상 이 모델은 현재 아무도 사용하지 않는 모델임. 따라서 이 연구가 최근 언어 모델에서도 유효할꺼라는 보장이 없음

•또한 Evaluation에서 사용된 평가 기준들 (Accuracy, Toxicity, Fluency) 외에도 다양한 기준들이 있을 수 있음. 예를 들자면 얼마나 글이 예의바른지, 주어진 요청사항을 잘 따르는지 등

•또한 일부 경우에서 CTG를 사용했음에도 toxic한 대답을 만들어내는 경우도 있음

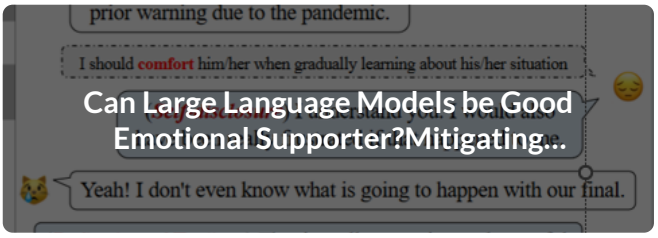
GTA(Gated Toxicity Avoidance) 를 이용해서 CTG의 문제점을 해결함



'논문 - 아주대(조현석교수님)' 카테고리의 다른 글

Can Large Language Models be Good Emotional Supporter?Mitigating Preference Bias ... (0) 2024.09.15

관련글 관련글 더보기






댓글 2

익명

비밀댓글입니다.

2024. 9. 18. 18:45

**OneTop's**

글이 너무 좋아요! 💯

2024. 9. 19. 09:06 · 답글

이름

비밀번호

내용을 입력하세요.

등록