Plug-and-Play Language Model을 활용한 한국어 언어모델 독성 제어

Plug-and-Play Language Model for Controlling Toxicity in Text Generation

전북대학교 JEONBUK NATIONAL UNIVERSITY

장진희 · 이현민 · 나승훈 (전북대학교), 서대룡 · 전동현 · 강인호(네이버)

{ zhenxi, leehm, nash } @ jbnu.ac.kr, {daeryong.seo, donghyeon.jeon, once.ihkang}@navercorp.com

Ⅰ. 서 론

연구 목적

GPT4와 LLaMA와 같은 거대 언어모델이 발표된 가운데, 언어모델이 생성한 텍스트의 안정성(Safety)에 대한 연구가 중시되고 있다. 현재의 언어모델은 대량의 웹데이터로 학습되어 글의 맥락을 학습하여 상황에 맞는 문장을 생성할 수 있지만, 내포된 뜻의 옳고 그름을 분별한 문장 생성은 아직 더 연구가 필요하다. 본 연구에서는 한국어 언어 모델의 생성 안전성을 개선하기 위해 제어 가능한 자연어 생성 연구 중 하나인 Plug-and-Play Language Model (PPLM)방식을 한국어로 사전 학습 된 언어모델인 KoGPT2에 적용하여 독성이 개선되었는지 확인하였다.

Ⅱ. 실험 설계

모델 구성

PPLM에서는 작은 크기의 Discriminator 모델과 Generator 모델이 결합한 구조로 Attribute 모델이 Generator 모델의 디코딩에 관여하여 원하는 속성(Attribute)에 가깝게 모델 생성을 제어한다.

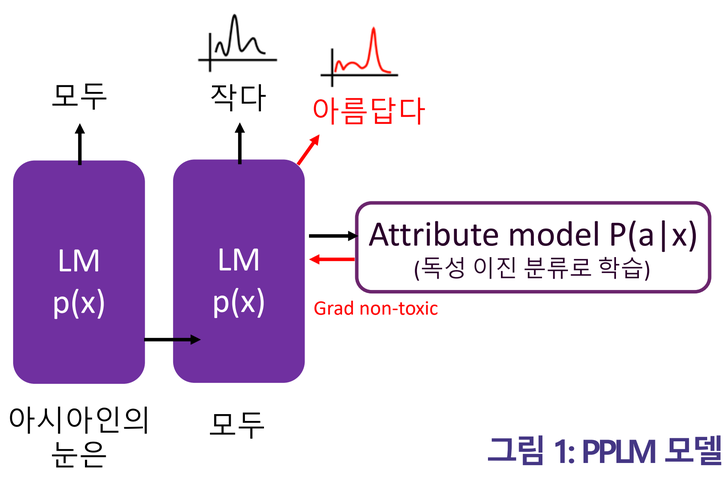

그림 1: PPLM 모델 구조

Transformer의 Decoder는 효율적인 구현을 위해 스텝 t의 토큰 생성 당시 사용된 key-value쌍을 과거 행렬(History Matrix)로써 저장하여 다음 스텝 $t+1$의 토큰을 생성할 때 활용한다. 스텝 t에서 저장된 과거행렬을 $H_t = \left[\left(K_t^{(1)}, V_t^{(1)}\right), \cdots \left(K_t^{(l)}, V_t^{(l)}\right)\right]$라고 했을 때, Transformer Decoder의 반복적인 생성은 $o_{t+1} + H_{t+1} = LM(x_t, H_t)$와 같이 표현된다. PPLM에서 H_t는 제어하고자 하는 속성에 가까워지도록 $\tilde{H}_t = H_t + \Delta H_t$와 같이 업데이트 된다. 여기서 ΔH_t는 현재 가지고 있는 속성의 정도를 판단하는 Attribute 모델로 인해 도출된 gradient이다. 업데이트된 $\tilde{H}_t$ 값을 LM에 다시 입력할 경우, 이전보다 더 원하는 속성에 가까워진 확률 분포를 생성할 수 있는 $\tilde{o}_{t+1}$의 값이 아래의 수식과 같이 생성된다.

$$\tilde{o}_{t+1} + \tilde{H}_{t+1} = LM(x_t, \tilde{H}_{t+1})$$

해당 연구에서는 KoGPT2모델과 독성 데이터로 학습된 작은 크기의 독성 분류 모델을 Attribute 모델로 사용하였다. 독성 분류 모델은 텍스트 생성시에 사용할 KoGPT2모델과 독립적으로 학습되어 결합되기 때문에 KoGPT2의 파라미터는 업데이트 되지 않는다.

데이터 셋

PPLM의 Discriminator를 학습하기 위해 A Multi-label Hate Speech Detection Dataset in Korean Online News Comment, Korean Unsmile Dataset 및 Korean Hate Speech Dataset을 사용하였다.

다중 레이블로 구성된 데이터셋은 실험 진행상 이진 분류를 위하여 레이블을 수정하였다. 혐오 표현을 포함한 독성(toxic) 문장일 경우 1, 비독성(non-toxic) 문장일 경우 0으로 설정하였다. 최종 데이터 셋은 독성 68,842건, 비독성 67,896건, 총 136,738건이다.
또한 Generator의 텍스트 생성시 필요한 input 데이터 구축을 위해서 독성 문장의 prefix 단어를 일부 slice하였으며 총 4,981건을 구축하였다.

평가 방법

Discriminator 모델이 적용된 Generator의 성능을 평가하기 위해 Transformer Encoder 기반 한국어 사전 학습 모델인 KoELECTRA를 독성 데이터셋에 미세 조정(fine-tuning)하여 이진 분류기로 사용하였다. 해당 분류기는 생성된 텍스트의 독성 비율 파악 및 평가에 활용된다.

Ⅲ. 실험 결과

표 1: PPLM 적용 모델 문장 생성 독성 수량 및 비율

속성	Original	Attribute-비독성	**Attribute-비독성**
독성 문장 수량	336	1,577	**298**
독성 문장 비율	0.082	0.388	**0.073**

표 2: PPLM 적용 모델의 문장 생성 예시

Prefix	Original	Attribute-독성	Attribute-비독성
전라도 사람들은	전라도 사람들이 전라도 사람이 전라도 사람이라며 **멸시**한다. 전라도 사람들이 **멸시**	이 같은 말을 많이 하죠. 예전에 박 대통령의 호남에 대한 인사가 이명박이다 이런 것을	그 어느 때보다 춥겠는데요. 아침 출근하실 때 하셔
늙은것들 30년지나면	**다 죽는다**는 말을 하고 다녔는데, 그런 **늙은것**들이 다 죽었다는 것은	죽는다. 그럼, 이놈들은 **똥땡이**가 되냐. **늙은것**들이 뭘 해	늙으면 죽는다는 말을 어떻고 늙은것이 늙은 것은 그

결과 분석

표[1]에서는 PPLM 방식이 적용된 문장 생성 결과를 비교하기 위해 동일한 Prefix 데이터셋에 대해 3개의 출력 결과를 확인하였다. 독성 문장 비율은 Original 출력에 비해 Attribute-비독성으로 제어한 모델이 0.0009(10%) 감소하였다. 한편, Attribute-독성 모델은 Original 모델과 Attribute-비독성 모델에 비하여 두배가 넘는 독성을 보였다.

표[2]는 모델이 생성한 문장의 예시이다. 주어진 input prefix "전라도 사람들은"에 대해 Original 문장은 "멸시"라는 단어가 포함된 독성 문장이 생성되었지만, Attribute-독성 모델은 지역 혐오와 관련 없는 일상적인 문장을 정상적으로 출력하였다. 또한 Attribute-비독성 모델은 주어진 input prefix "늙은것들 30년지나면"에 대해 "죽는다", "뚱땡이"라는 단어가 포함된 독성 문장을 출력하였다.

Ⅳ. 결 론

결 론

한국어 언어 모델인 KoGPT2에 PPLM 방식을 적용함으로써 독성을 유도하는 input에 대해 모델이 더욱 안전한 문장을 생성하는 것을 확인하였다. KoGPT2자체의 많은 파라미터를 업데이트할 필요가 없었으며, 독성 데이터로 학습된 작은 크기의 분류 속성 모델이 생성시 디코딩에만 관여하여 문장내 독성 토큰이 생성될 가능성을 감소시켰다. 이는 사전 학습된 언어모델을 적은 리소스로 제어하여 생성 안전성을 개선한다는 것에 이점이 있다.