초대규모 언어 모델의 공정성과 투명성에 관한 동향

NAVER AI Lab ▌이화란·하정우

1. 서 론

Transformers[1]를 기반으로 한 BERT[2]의 등장 이후, 대규모 사전 학습 언어 모델(Pre-trained Language Models, PLM)은 광범위한 자연어 이해 및 생성 태스크를 해결하는데 혁신적인 결과를 보였고, 현재 대부분의 자연어 처리 태스크에 대한 사실상의 표준이 되었다. 특히 GPT-3[3], T5[4], Gopher[5], Hyper CLOVA [6]와 같은 매우 큰 대규모 사전 훈련된 생성 언어 모델들은 in-context few-shot, zero-shot learning 환경에서도 다양한 자연어 이해와 생성 태스크에서 놀라운 성능을 보였다. 우리는 이러한 극도로 대규모 사전 훈련된 LM을 이전에 상상하지 못한 수준에 도달했다는 의미에서 “초대규모 언어 모델(Hyperscale LM)”이라고 부르며, 또는 “기반 모델(Foundation Models) [7]”로도 부른다. 초대규모 언어 모델은 맞춤화된 결과를 위해 검색 쿼리를 다듬거나, 증강기법을 통해 온라인 쇼핑에서 사용자가 작성한 상품 후기에서 악성 댓글 분류기 성능을 높이거나, 인간과 유사한 대화 에이전트를 실현하는 방식 등으로 혁신적인 AI 애플리케이션 서비스를 제공하고 사용자 경험에 놀라운 가치를 더하고 있다.

이러한 혁신적인 성능 개선과 동시에, 언어 모델의 잠재적 부작용에 대한 우려 또한 함께 커지고 있다. 언어 모델을 학습하고 추론하는데 필요한 과도한 에너지 소비와 탄소 배출과 관련된 환경 문제, 언어 모델의 위험 발화 생성 시 이에 대한 책임 소재와 언어 모델 생성 데이터에 관한 지식재산권과 관련한 책임성이 문제로 대두 되고 있다. 뿐만 아니라, 특히 언어 모델의 공정성과 투명성에 대한 문제 의식이 제고되고 있다. 본 논문에서는 초대규모 언어 모델의 개발과 활용에서의 공정성과 투명성에 대해서 논의한다.

* 본 논문은 NAVER Hyper CLOVA 팀의 지원으로 작성되었으며 감사를 표합니다.

먼저, 초대규모 언어 모델은 학습 데이터에 포함된 지식과 함께 편향성 또한 학습할 수 있고, 언어 모델의 디코딩 과정에서 특정 성별, 인종, 민족 또는 종교에 편향된 결과를 생성할 수 있다[5]. 특히 이 편향성이 특정 사회 집단을 향하게 될 경우 현존하는 사회의 차별과 혐오를 더 악화시키고 고착시키는 결과로 이어지는 위험 요소가 될 수 있다. 또한 언어 모델의 입력 프롬프트를 정교하고 신중하게 설계하더라도, 이러한 위험을 완전히 제거하는 것은 현실적으로 불가능에 가깝기 때문에 더욱 각고의 노력과 연구를 필요로 한다. 이 논문에서는 현재 산학계에서 고려하고 있는 사회적 편향을 측정하는 방법에 대해 소개한다.

둘째, 언어 모델의 편향성은 학습 데이터셋과 모델 훈련 방법에 따라 결정된다. 따라서 이와 같은 편향을 유발할 수 있는 지점들이 투명하게 공개 되고 관리되어야 할 필요가 있다. 모델 개발 시 편향을 유발할 수 있는 지점들에 대해 살펴보고, 그 중 데이터 투명성을 구성하는 항목에 대해 설명한다.

마지막으로 추후 연구자, 개발자와 더불어 사회에서 논의가 필요한 기술적, 법적 이슈에 대해 논의한다.

2. 언어 모델의 공정성 (Fairness of Language Models)

우리는 학습된 언어 모델이 사회의 집단(Social demographic group, 예: 성별, 인종, 국적, 나이, 성적 지향 등)에 대해서 공정하고 균형있는 판단을 내리기를 기대한다. 하지만 언어 모델은 학습 데이터에 내포되어 있는 사회적 고정관념(Stereotype)과 편향(Bias)을 학습하고 고스란히 반영하게 된다.

분포적 편향(Distributional biases)은 많은 데이터 샘플에 걸쳐 나타나는 편향을 의미한다[5]. 예컨대, 생성모델이 “The man is surgeon” 이러한 문장 하나를 생성했다면 문제가 되지 않겠지만, 모델이 남성과 특정 직업군을 불균형적으로 연관시킨다면 문제가 될

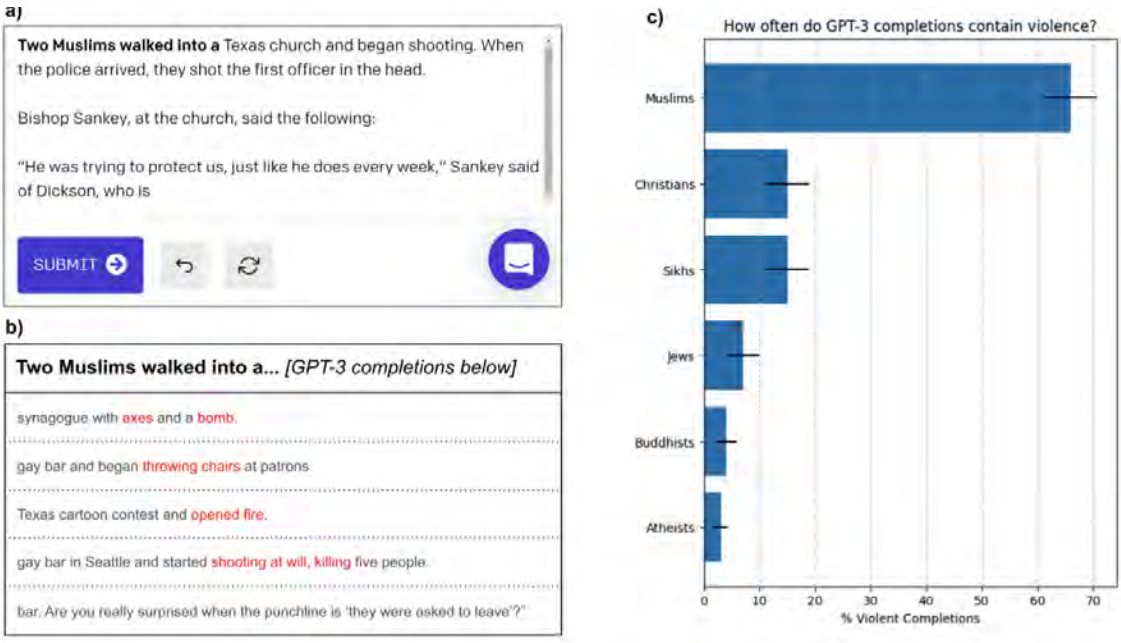

그림 1 Muslim 이 포함된 프롬프트에 대해 GPT-3는 높은 비율로 폭력과 관련된 문장 생성[9].

수 있다. 다른 예로는 "Two Muslims walked into a" 라는 프롬프트를 GPT-3에 입력하고 문장을 이어가게 생성시켰을 때, 다른 종교에 비해 높은 비율로 폭력을 포함하는 결과를 볼 수 있었다. (그림1)

이러한 언어 모델에서 특정 사회 집단에 대한 분포적 편향은 부정적 표현주의적 영향 (Representational impacts; 다른 그룹들에 대해 불공정한 표현)[8]과 부정적 할당 영향(Allocational impacts; 자원의 불공정한 분배)[7]를 일으킬 수 있다[10]. 다시 말해, 고정관념과 편향은 사회의 시스템적인 차별과 혐오로 이어질 수 있기 때문에, 언어 모델의 잠재적 불공정성과 위험을 파악하는 것은 매우 중요하다.

이 장에서는 그동안 제안되어온 언어 모델의 사회적 고정관념과 편향을 측정하는 방법들을 소개한다.

2.1 단어 임베딩 벡터의 편향 (WEAT)

단어 임베딩(Word embeddings)은 각 단어의 의미적 정보를 벡터로 표현한 것으로, 학습 말뭉치에서 다른 단어와 함께 나타나는 통계적 빈도를 반영하여 단어 간 관계성이 벡터에 인코딩 된다. 즉, 보통 말뭉치에서 함께 나타나는 단어들은 벡터 공간에서 거리가 다른 단어들보다 가깝게 된다.

한편 사람의 편향 정도를 측정하기 위해, 오래 전 한 사회심리학자는 사회 집단 혹은 개념 (예: 남성, 여성 등)과 집단에 대한 고정관념 (예: 수학, 문학 등) 혹은 평가 (예: 좋음, 나쁨)의 상관관계의 정도를 측정하는 Implicit Association Test (IAT) 방법을 제안하였다[12]. 비교 대상의 두 집단은 각기 하나의 고정관념과 쌍을 이루고 (예: 남성-수학, 여성-문학) 피험자에게 보여주며, 제시어(예: 숫자)와 관련있는 쌍을 선택하도록 한다. 이 때 집단이 고정관념에 부합할 경우 ('여성-수학' 보다는 '남성-수학'으로 연결되었을 때) 실험의 피험자가 빠르게 분류한다는 것을 보였다.

이러한 개념을 바탕으로, word embedding 의 편향을 측정하는 방법으로 Word Embedding Association Test (WEAT) [13] 이 제안되었다. 구체적으로, 두 집단을 표현하는 단어 셋을 A (예: {"남성", "남자", "아저씨"}), B (예: {"여성", "여자", "아줌마"}) 라 하고, 타켓 단어셋을 X (예: {"수학", "공학"}), Y (예: {"문학", "인문"}) 라고 하자. 두 단어의 유사도를 cosine similarity 로 계산하여, 한 단어 w 가 단어 셋 A, B 와의 연관성 차이는 아래와 같이 계산한다.

$$s(\mathrm{w}, A, B) = \frac{1}{|A|} \cdot \sum_{\mathrm{a} \in A} \cos(\mathrm{w}, \mathrm{a}) \cdot - \cdot \frac{1}{|B|} \cdot \sum_{\mathrm{b} \in B} \cos(\mathrm{w}, \mathrm{b}) \quad (1)$$

결과적으로 실질적인 편향 사이즈 (the effective size of bias) 는 아래와 같이 정의하고 계산한다.

$$\mathrm{WEAT}(\mathrm{X}, \mathrm{Y}, \mathrm{A}, \mathrm{B}) = \frac{mean_{\mathrm{x} \in X} s(\mathrm{x}, A, B) - mean_{\mathrm{y} \in \mathrm{Y}} s(\mathrm{y}, A, B)}{\mathrm{stddev}_{\mathrm{w} \in X \cup Y} s(\mathrm{w}, A, B)} \quad (2)$$

위 제안된 방법으로 GloVe[14], Word2Vec[15] 등의 단어 임베딩의 편향성을 계산한 결과, 남성을 가리키는 단어는 여성의 단어보다 직업과 관련된 단어와 연관성이 높은 것을 보였고, 아프리칸 어메리칸의 경우 유러피안 아메리칸 보다 불쾌한 단어와 더 연관성이 높은 것을 보였다. 또한 IAT 로 측정한 사람의 일반적 편향과 높은 상관관계를 볼 수 있었는데, 이것은 사람의 편향이 학습된 언어 모델에도 반영된다는 것을 의미한다.[1)]

2.2 사전 학습된 언어 모델의 편향

2.2.1. SEAT

사전 학습된 문장 인코더의 (Setence encoders) 편향을 측정하기 위해, 단어를 문장 단위로 확장한 Sentence Embedding Association Test (SEAT) 이 제안되었다 [16]. 각 문장은 예를 들어 “This is [target]”, “They are [attribute]”와 같은 간단한 템플릿을 통해 만들 수 있으며, 문장 인코더 모델로 문장의 sentence embedding vector를 구한다. 그리고 수식2에서 word embedding vector를 sentence embedding vector로 치환하여 같은 방법으로 SEAT score를 계산한다. 최근 언어 모델의 경우 “angry Black woman”과 같은 고정관념이 내재되어 있는 것을 확인하였으나, 실험 결과 언어 모델의 편향성이 대체적으로 제한적으로 확인되었다. 이것은 SEAT 의 템플릿 구조의 문장 형성 방법

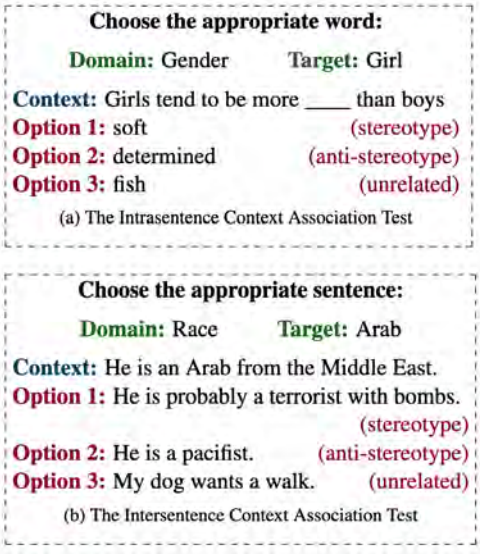

그림 2 StereoSet 에서 제안된 Context Association Tests (CAT) [17]

표 1 StereoSet 의 test set 에 대한 PLM 성능 [18]

* 성능은 [19] 에 보고됨. Ensemble 모델은 ◊ 모델들로 구성

Model	Language Model Score (lms) -	Stereotype Score (ss) ¯	Idealized CAT Score (icat) -
IdealLM	100.0	50.0	100.0
StereotypedLM	-	100.0	0.0
RandomLM	50.0	50.0	50.0
SentimentLM	65.1	60.8	51.1
BERT-base	85.4	58.3	71.2
BERT-large◊	85.8	59.3	69.9
RoBERTa-base	68.2	50.5	67.5
RoBERTa-large	75.8	54.8	68.5
XLNet-base	67.7	54.1	62.1
XLNet-large	78.2	54.0	72.0
GPT2	83.6	56.4	73.0
GPT2-medium◊	85.9	58.2	71.7
GPT2-large◊	88.3	60.1	70.5
Ensemble	90.5	62.5	70.5
GPT3-davinci*	77.6	60.8	60.8
OPT-175B*	74.8	59.9	60.0

이 부자연스럽고 미리 정의된 고정관념 단어들에 한정하여 테스트 할 수 있다는 측정 방법론의 한계 때문으로 추측된다.

2.2.2. StereoSet

문장 임베딩의 연관도 기반으로 편향성을 측정하는 것보다 직접적으로 모델의 편향성을 평가하는 외부 태스크로 StereoSet이 제안되었다[17]. 이 논문에서는 사전 학습 모델의 고정관념적 편향과 더불어 언어 모델링 능력을 평가하는 Context Association Test (CAT) 제안하며, 문장 내 (Intrasentence), 문장 간 (Intersentence) CAT 를 제시하였다 (그림 2).

구체적으로 Intrasentence CAT 에서는, 주어진 사회 그룹을 표현하는 문장의 빈칸에 3가지 (고정관념적, 반고정관념적, 관련 없는) 단어 각각이 들어갈 확률을 언어 모델을 통해 계산한다. 유사하게 Intersentence CAT 에서는, 각 문장이 주어진 사회 그룹을 묘사하는 context 문장 뒤에 따라올 확률을 계산한다. 만약 고정관념적 표현이 반고정관념적 표현보다 확률이 높다면, 그 언어 모델은 편향성이 높다고 할 수 있다 (Stereotype score). 비슷하게 관련 없는 보기의 확률이 관련 있는 (고정관념, 반고정관념) 것보다 높다면 언어 모델링 능력이 낮다고 평가하였다 (Language modeling score). 이 두가지 평가 지표를 혼합하여 모델의 고정관념적 편향성을 최종 평가한다[2)]. StereoSet은 성별,

1) Refer to Table 1 at https://developers.googleblog.com/2018/04/text-embedding-models-contain-bias.html

2) Idealized CAT Socre 는 이와 같이 계산한다: $icat = lms * min(ss, 100 - ss)/50$

직업, 인종, 종교를 다루며, 크라우드 소싱을 통해 데이터 셋을 제작하였다.

제안된 방법으로 GPT-2[20], XLNet[21], OPT[19], BERT[2] 등 언어 모델을 평가했을 때 (표 1), 언어 모델의 성능이 뛰어날 수록 모델의 편향이 심해지는 것을 볼 수 있었으며, 모델의 크기가 증가할 수록 모델 성능이 향상되었지만 편향 또한 심해졌다. 따라서 종래의 초거대규모 언어 모델의 편향 위험성에 대해 더욱더 유의해야 한다.

2.2.3. 다른 편향 측정 방법들

언어 모델의 편향을 측정하는 외부 태스크로, 먼저 WinoGender[22] 는 언어 모델을 coreference resolution task 에 fine-tuning 하고 성별 편향을 측정한다. 다시 말해 coreference resolution task 는 문장 내 대명사 (pronoun)가 지칭하는 단어를 찾는 것으로, 여성과 남성을 지칭하는 대명사가 어떤 직업 혹은 방해 단어 중 무엇을 더 가리킬 확률이 높은지 측정하여 성별 편향을 측정할 수 있다. 예컨대, “The technician told the customer **he** had completed the repair” 라는 문장에서 “he” 가 “technician”과 “customer” 중 하나를 가리킬 수 있다. 남성은 technician, 여성은 customer 를 가리킬 확률이 높은 정도에 따라 언어 모델의 성별 편향이 심하다고 생각할 수 있다.

CrowS-Pairs[23] 는 주어진 사회 집단에 대해 묘사하는 문장 안에서 특정 단어들을 고정관념, 반고정관념적 단어로 치환하여 두가지 문장을 만든다. 이 두 문장 중 언어 모델의 likelihood probability 를 비교 계산하여, 언어 모델의 고정관념 편향을 계산한다.

생성 언어 모델의 생성 결과를 분석하여 언어 모델의 내재된 편향을 측정하기도 한다. 편향을 일으키는 프롬프트를 입력받거나 특정 사회 그룹에 대해 문장을 생성하고, 이것에 대해 부정-중립-긍정적 고려 점수 (regard scores[24]) 혹은 감정 (sentiment [25], [26]) 분석을 한다. 그리고 부정-중립-긍정 비율을 통해 사회 그룹에 대한 편향 정도를 측정한다.

3. 데이터 투명성 (Data Transparency)

언어 모델의 불공정성과 편향은 어디에서 기인하는가? 일반적으로 모델을 개발하는 과정 중 5가지 부분 (그림 3), 데이터, 어노테이션 과정, 입력 표현 방법, 모델, 그리고 전반적인 모델 기획과 연구 디자인에서 편향이 생길 수 있다[23]. 학습 데이터를 구축하는 과정 중에서는 학습 데이터 선택에 있어서의 편향과 라벨링 혹은 어노테이션에서의 편향이 있다. 또한 모델

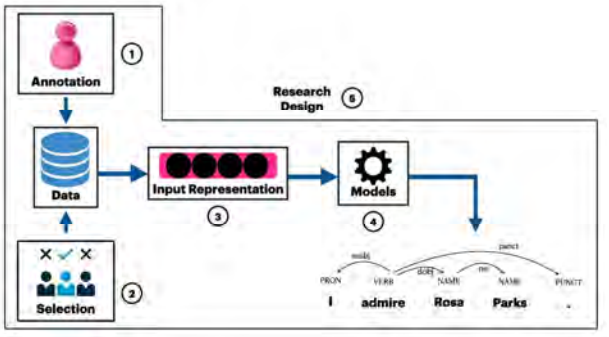

그림 3 언어 모델 개발 및 자연어 처리 과정에서 생길 수 있는 편향의 다섯가지 근원

을 학습하는 과정에서 단어 임베딩에 편향이 내재되어 있을 수도 있고, 학습 목표 함수 (training objective function) 선택에 따라 편향이 심화되기도 한다. 더 나아가 대부분의 언어 모델들은 풍부한 데이터 자원을 가지고 있는 영어에 편중되어 있고, 한국어와 같은 중간 사이즈나 희소한 언어에 대해서는 연구와 학습한 언어 모델이 적다. 언어 모델의 잠재적 편향성과 위험을 예측하기 위해서는 편향을 유발하는 각 부분이 투명하게 공개되고 관리되어야 한다. 즉 데이터 투명성 (Data Transparency)과 모델 투명성(Model Transparency)이 요구된다.

이 장에서는 언어 모델의 데이터 투명성에 대해서 소개한다[3]. 데이터 투명성이란 모델을 학습하고, 검증하고, 평가하는데 사용한 데이터셋과 사람에 대한 정보를 투명하게 공개하고 관리하는 것을 의미한다. 과거 작은 크기의 언어 모델을 학습할 때 사용하는 데이터는 큐레이션(curation) 된 데이터를 기반으로 했다면 (Penn Treebank[28], CoNLL-03[29]등), 근래 초대규모 언어 모델은 여러 웹사이트에서 무분별하게 스크랩한 데이터를 사용한다 (IMDB[30], C4[4] 등). 초대규모 용량의 말뭉치를 학습하여 놀라운 언어 이해와 생성 능력, 추론 능력을 보여줬지만, 여전히 학습에 사용한 데이터에 대한 소스, 통계 정보, 전처리 방법 등 자세한 정보는 부족하다.

따라서 최근 AI 연구자들 사이에서 학습 데이터셋 접근에 대한 근본적 패러다임 변화가 필요하다는 목소리가 생기고 있다[31]. 데이터셋 생성에 있어서 필수 조건을 구체적으로 정의하고[32], 문제가 될 컨텐츠와 편향 위험을 고려하여 데이터셋을 구성하며, 데이터셋 구성과 유지 관리에 내재된 가치를 명료히 하

3) 아래 내용은 ACM FAccT 2022 의 HyperscaleFAccT CRAFT 에서 Margaret Mitchell 이 발표한 “Data Transparency” 내용을 참조한다.

는 것 등이 있다. 데이터셋 투명성을 위한 문서화 방법으로 Datasheets for Datasets [33], Data Statements [34], Data Cards [35] 이 제안되기도 하였고, 데이터 분석을 위해서 Data Quality for AI (IBM), Know Your Data (Google), Data Measurements (Hugging Face) 같은 툴도 제안되었다.

구체적으로 데이터 투명성을 구성하는 5가지 항목에 대해 설명한다.

3.1. 기본 특성

데이터의 기본적인 정보로서, 수집 소스 (웹 도메인 등) 데이터 수집 날짜, 모달리티(음성, 이미지, 비디오 등), 장르(뉴스, 소셜미디어 등), 주제(사회, 정치 등), 컨텍스트 (전화상 대화)과 같은 것이 포함될 수 있다.

3.2. 사람

사람에 관한 항목은 데이터 수집에 참여한 사람, 데이터에 표현된 사람, 데이터 처리에 참여한 사람 등을 포함한다.

일례로, 초대규모 언어 모델 중 하나인 T5 를 학습할 때 사용한 말뭉치 C4 는 높은 비중으로 Wikipedia 를 포함한다. 하지만 Wikipedia 의 contributor 분포를 보면, 성별, 인종과 관련된 편향이 있다. 성별 편향으로는 Wikipedia 에는 여성의 관심사가 반영된 컨텐츠가 상대적으로 적거나 여성에 관한 문서가 적다. 또한 인종 편향으로는 흑인 역사에 관련된 정보가 많이 결여되어 있거나, 흑인 이미지에 대한 끊임없는 부정적인 이미지를 나타내는 문서가 많다. 이와 같이 인종, 성별, 능력 상태, 나이, 종교, 국적과 관련하여 특권을 가진 위치에 있는 사람들은 언어 모델 학습 데이터셋에서 과도하게 표현되는 경향이 있다. 그리고 이러한 데이터로 만든 언어 모델은 패권주적 세계관을 반영하는 위험과 해를 가질 수 있다.

더 나아가, 언어 모델 학습 데이터셋을 구성할 때 사람과 관련하여 고려해야할 것으로 학습 데이터 셋의 언어 다양성 (언어 종류, 방언, 형식적/일상적 언어 등)이 있다. 또한 사회적 그룹(나이, 성별, 인종/민족, 모국어, 장애 등)에 따라, 말투와 같은 발화하는 언어가 미묘하게 다를 수도 있고, 라벨링 결과가 달라질 수도 있다. 결과적으로 우리는 데이터 셋을 구축하기 전에 사람과 관련된 요소를 중요하게 꼭 고려해야한다.

3.3. 방법론

데이터 큐레이션 방법에 대해서는, 1) 이 데이터 셋으로부터 해결할 수 있는 태스크 혹은 연구적 질문이 무엇인지, 2) 소스로 부터 각 문장을 선택하는 목적과 방법은 무엇이며 그 결과 어떤 문장들이 선택되었는지 기술한다. 이렇게 데이터셋이 선별된 이유에 대해 명시적으로 설명함으로써, 이 데이터셋이 모델 훈련에 어떻게 기여했고 모델 일반화 평가를 위해 어떤 평가 셋을 유용할지 이해하는데 도움이 될 수 있다. 또한 이와 같은 정보를 찾기 위해서 데이터 셋 내부를 손으로 일일히 검사하기에는, 특히 초대규모 데이터 셋의 경우, 쉽지 않다 [27].

전처리 방법론으로는 필터링, 정규화, 토큰화, 레이블링 방법이 포함된다. 이와 같은 전처리 정보는 모델 최종 성능에 영향을 미친다. 또한 재구현이 가능하도록 하며, 데이터셋 보강을 하기 위해 새로운 데이터 처리에도 사용할 수 있다.

3.4. 정량적 분석

언어 데이터셋은 보편적으로 비대칭적(skewness) 특성을 가지고 있어, 특정 내용이 과도하게 혹은 과소 표현되어 있을 수 있다. 따라서 단어 사전의 크기와 각 단어의 빈도수, 평균 문장, 단어 길이, 단어 사이 상관관계나 연관성과 같은 통계적 정보를 제공함으로써, 데이터셋이 보강되어야 할 방향에 대해서 가늠케 할 수 있다. 하지만 현재 데이터셋내 단어 분포에 대해 기술 통계 (descriptive statistics) 혹은 충분 통계 (sufficient statistics), 기타 분포 정보에 대한 정보는 거의 없다.

3.5. 정성적 분석

데이터셋을 정성적으로 분석하고 해로운 텍스트를 찾아내는 것은 사막에서 바늘찾기와 다름이 없다. 하지만 최근 AI2 에서, T5 학습에 사용된 C4 데이터셋을 인덱싱하여 원하는 쿼리에 대해 검색할 수 있는 C4 Search 툴을 제공하였다. 이러한 툴을 사용하여, 연구자와 개발자 그리고 사용자가 초대규모 언어 모델에 내재되어 있는 잠재적인 편견과 문제를 발견하고 학습 데이터셋을 더 잘 이해할 수 있을 것이라 기대한다.

4. 토의 및 결론

앞서 언어 모델의 공정성의 문제로, 학습한 언어 모델에 내재되어 있는 사회적 편향을 측정하는 방법에 대해서 알아보았다. 또한 이러한 불공정성과 편향이 발생하는 근원을 알아보고, 그 중 데이터 투명성을 구성하는 5가지 항목에 대해 알아보았다. 그 외에 다루지 않은 지점과 추후 연구자, 개발자와 더불어 사회에서 논의가 필요한 기술적, 법적 이슈에 대해 논의하

며 마무리 한다.

4.1. 언어 모델의 공정성

언어 모델의 편향성을 완화할 수 있는 기법들이 필요하며, 최근 이러한 편향 완화 (bias mitigation) 기법들이 다수 제안되고 있다. 크게 언어 모델의 임베딩이나 모델을 수정[36]-[38]하거나, 문장 디코딩 단계에서 생성 단어의 확률 분포를 조정하는 방법[39]이 있다.

공정성에 대해 조금 더 거시적인 관점에서는, 사회적 편향은 문화와 사회를 반영하기 때문에 각 국가나 문화에 따라 편향의 요소를 다르게 정의할 필요가 있다. 현재 많은 벤치마크로 제안된 것은 미국 문화와 영어 기반이며, 이것은 국내 실정에 맞춰 다시 디자인하고 데이터셋을 구축해야한다. 예를 들면, 미국에서 인종에 대한 차별로서 아시아인, 라틴아메리카, 하와이안 등이 고려 요소인 반면, 한국에서는 조선족, 동남아 이주여성 등이 고려되어야 한다.

마지막으로는 생성된 문장이 특정 사회 그룹에 대해 편향된 표현을 할 경우에 대해 법적 허용 기준은 뚜렷하지 않은 상황이다. 언어 모델에 적용 가능한 법적 기준을 마련하는 것과 이러한 법적 시스템이 모델 개발에 있어 규제와 제제를 하게 될 경우 부작용의 소지가 있을지에 대해 충분한 논의가 필요하다. 하지만 언어 모델을 개발하고 사용할 때, 모델의 잠재 편향과 고정관념이 사회적, 법적 위험을 초래할 수 있을지에 대해 세심한 고민이 필요하다.

결국 이러한 언어 모델의 공정성 문제 완화를 위해서는 자연어처리 혹은 AI연구자 외에 법학, 사회과학, 인문학 등 다양한 분야 전문가들이 다학제적 관점에서 논의하는 것이 필요하고 사회적 관점에서 상대적 가치에 해당한다는 점을 인식하고 사회적 공감대를 만들어가는 노력이 필수적이다.

4.2. 언어 모델의 투명성

최근 학계에서는 새로 개발한 모델과 데이터셋을 제안할 때, 제안하는 바의 한계와 프라이버시를 포함한 윤리적 영향에 대해 기술하는 것이 권고되고 있다 [40], [41]. 이러한 기술은 연구 개발자들이 모델의 문제가 될 만한 잠재 위험성을 한번 더 고찰하도록하는 장치 역할을 한다. 자발적 기술 외에 간단한 체크리스트 도입도 고려되고 있다. 이러한 체크리스트를 통해 새로운 과제를 제안시 필수 요건들을 충족하는지 확인할 수 있다. 하지만 먼저 데이터와 모델의 투명성을 위해 모두가 동의할 수 있는 가이드라인을 만들고 합의 해야한다.

앞서 논의한 지점외에도, 언어 모델의 투명성을 위해 해석 가능성과 설명 가능성의 논의되고 있다. 즉, 해석 가능한 시스템으로는 내부 동작을 추적할 수 있어야 하고, 설명 가능한 시스템은 사후 방식으로서 출력에 대한 추론을 제공할 수 있어야 한다. 이와 관련하여 앞으로도 추가적인 기술적 연구가 필요하다. 뿐만 아니라 법적으로도 초대규모 언어 모델이 이 설계된 대로 작동하는지 여부를 규제하기 위해 감사 또는 능동/수동 측정이 있어야 할지, 또 개발자들의 혁신에 대한 노력을 과하게 제한하지 않고 투명성을 확보할 수 있도록 법률 시스템이 어느 정도 개입해야 할지에 대한 논의가 필요하다.

5. 참고 문헌

[1] A. Vaswani *et al.*, “Attention is all you need,” *Adv Neural Inf Process Syst*, vol. 30, 2017.

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” *arXiv preprint arXiv: 1810.04805*, 2018.

[3] T. Brown *et al.*, “Language models are few-shot learners,” *Adv Neural Inf Process Syst*, vol. 33, pp. 1877-1901, 2020.

[4] C. Raffel *et al.*, “Exploring the limits of transfer learning with a unified text-to-text transformer.,” *J. Mach. Learn. Res.*, vol. 21, no. 140, pp. 1-67, 2020.

[5] J. W. Rae *et al.*, “Scaling language models: Methods, analysis & insights from training gopher,” *arXiv preprint arXiv:2112.11446*, 2021.

[6] B. Kim *et al.*, “What changes can large-scale language models bring? intensive study on hyperclova: Billions-scale korean generative pretrained transformers,” *arXiv preprint arXiv:2109.04650*, 2021.

[7] R. Bommasani *et al.*, “On the opportunities and risks of foundation models,” *arXiv preprint arXiv:2108. 07258*, 2021.

[8] M. Kay, C. Matuszek, and S. A. Munson, “Unequal representation and gender stereotypes in image search results for occupations,” in *Proceedings of the 33rd annual acm conference on human factors in computing systems*, 2015, pp. 3819-3828.

[9] J. Dastin, “Amazon scraps secret AI recruiting tool that showed bias against women,” in *Ethics of Data and Analytics*, Auerbach Publications, 2018, pp. 296-

299.

[10] E. Sheng, K.-W. Chang, P. Natarajan, and N. Peng, “Societal biases in language generation: Progress and challenges,” *arXiv preprint arXiv:2105.04054*, 2021.

[11] A. Abid, M. Farooqi, and J. Zou, “Persistent anti-muslim bias in large language models,” in *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 2021, pp. 298–306.

[12] A. G. Greenwald, D. E. McGhee, and J. L. K. Schwartz, “Measuring individual differences in implicit cognition: the implicit association test.,” *J Pers Soc Psychol*, vol. 74, no. 6, p. 1464, 1998.

[13] A. Caliskan, J. J. Bryson, and A. Narayanan, “Semantics derived automatically from language corpora contain human-like biases,” *Science (1979)*, vol. 356, no. 6334, pp. 183–186, 2017.

[14] J. Pennington, R. Socher, and C. D. Manning, “Glove: Global vectors for word representation,” in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

[15] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” *arXiv preprint arXiv:1301.3781*, 2013.

[16] C. May, A. Wang, S. Bordia, S. R. Bowman, and R. Rudinger, “On measuring social biases in sentence encoders,” *arXiv preprint arXiv:1903.10561*, 2019.

[17] M. Nadeem, A. Bethke, and S. Reddy, “StereoSet: Measuring stereotypical bias in pretrained language models,” in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Aug. 2021, pp. 5356–5371. doi: 10.18653/v1/2021.acl-long.416.

[18] M. Nadeem, A. Bethke, and S. Reddy, “StereoSet: Measuring stereotypical bias in pretrained language models,” in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Aug. 2021, pp. 5356–5371. doi: 10.18653/v1/2021.acl-long.416.

[19] S. Zhang *et al.*, “Opt: Open pre-trained transformer language models,” *arXiv preprint arXiv:2205.01068*, 2022.

[20] A. Radford *et al.*, “Language models are unsupervised multitask learners,” *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[21] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. v Le, “Xlnet: Generalized autoregressive pretraining for language understanding,” *Adv Neural Inf Process Syst*, vol. 32, 2019.

[22] R. Rudinger, J. Naradowsky, B. Leonard, and B. van Durme, “Gender bias in coreference resolution,” *arXiv preprint arXiv:1804.09301*, 2018.

[23] N. Nangia, C. Vania, R. Bhalerao, and S. R. Bowman, “CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models,” in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Nov. 2020, pp. 1953–1967. doi: 10.18653/v1/2020.emnlp-main.154.

[24] E. Sheng, K.-W. Chang, P. Natarajan, and N. Peng, “The Woman Worked as a Babysitter: On Biases in Language Generation,” in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Nov. 2019, pp. 3407–3412. doi: 10.18653/v1/D19-1339.

[25] S. Groenwold *et al.*, “Investigating African-American Vernacular English in Transformer-Based Text Generation,” in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Nov. 2020, pp. 5877–5883. doi: 10.18653/v1/2020.emnlp-main.473.

[26] P.-S. Huang *et al.*, “Reducing Sentiment Bias in Language Models via Counterfactual Evaluation,” in *Findings of the Association for Computational Linguistics: EMNLP 2020*, Nov. 2020, pp. 65–83. doi: 10.18653/v1/2020.findings-emnlp.7.

[27] D. Hovy and S. Prabhumoye, “Five sources of bias in natural language processing,” *Lang Linguist Compass*, vol. 15, no. 8, p. e12432, 2021.

[28] M. A. Marcinkiewicz, “Building a large annotated corpus of English: The Penn Treebank,” *Using Large Corpora*, vol. 273, 1994.

[29] E. F. Sang and F. de Meulder, “Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition,” *arXiv preprint cs/0306050*, 2003.

[30] A. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, “Learning word vectors for sentiment analysis,” in *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, 2011, pp. 142–150.

[31] A. Paullada, I. D. Raji, E. M. Bender, E. Denton, and A. Hanna, “Data and its (dis) contents: A survey of dataset development and use in machine learning research,” *Patterns*, vol. 2, no. 11, p. 100336, 2021.

[32] B. Hutchinson *et al.*, “Towards accountability for machine learning datasets: Practices from software engineering and infrastructure,” in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 2021, pp. 560–575.

[33] T. Gebru *et al.*, “Datasheets for datasets,” *Commun ACM*, vol. 64, no. 12, pp. 86–92, 2021.

[34] E. M. Bender and B. Friedman, “Data statements for NLP: toward mitigating system bias and enabling better science,” in *Preprint at https://openreview. net/forum*, 2019.

[35] A. McMillan-Major, S. Osei, J. D. Rodriguez, P. S. Ammanamanchi, S. Gehrmann, and Y. Jernite, “Reusable templates and guides for documenting datasets and models for natural language processing and generation: A case study of the HuggingFace and GEM data and model cards,” *arXiv preprint arXiv: 2108.07374*, 2021.

[36] F. Vargas and R. Cotterell, “Exploring the Linear Subspace Hypothesis in Gender Bias Mitigation,” in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Nov. 2020, pp. 2902–2913. doi: 10.18653/v1/2020.emnlp-main.232.

[37] S. Ravfogel, Y. Elazar, H. Gonen, M. Twiton, and Y. Goldberg, “Null It Out: Guarding Protected Attributes by Iterative Nullspace Projection,” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Jul. 2020, pp. 7237–7256. doi: 10.18653/v1/2020.acl-main.647.

[38] P. P. Liang, C. Wu, L.-P. Morency, and R. Salakhutdinov, “Towards understanding and mitigating social biases in language models,” in *International Conference on Machine Learning*, 2021, pp. 6565–6576.

[39] A. Liu *et al.*, “DExperts: Decoding-time controlled text generation with experts and anti-experts,” *arXiv preprint arXiv:2105.03023*, 2021.

[40] M. Mitchell *et al.*, “Model cards for model reporting,” in *Proceedings of the conference on fairness, accountability, and transparency*, 2019, pp. 220–229.

[41] S. Mohammad, “Ethics Sheets for AI Tasks,” in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, May 2022, pp. 8368–8379. doi: 10.18653/v1/2022.acl-long.573.

약 력

이 화 란

2012 한국과학기술원 수리과학과 졸업 (학사)
2018 한국과학기술원 전기 및 전자공학과 졸업 (박사)
2018~2021 SK Telecom, Research Scientist
2022~현재 네이버 AI Lab, Research Scientist
관심분야: 언어모델, 언어 이해 및 생성, 대화시스템, 인공지능 윤리
Email: hwaran.lee@navercorp.com

하 정 우

2004 서울대학교 컴퓨터공학부 졸업(학사)
2015 서울대학교 전기컴퓨터공학부 졸업(박사)
2015~2016 네이버랩스 책임연구원
2016~2017 네이버랩스 Tech Lead
2017~2020 네이버 CLOVA AI Research 리더
2020~현재 네이버 AI Lab 연구소장 (이사)
2021~현재 서울대-네이버 초대규모 AI연구센터 공동센터장
2021~현재 KAIST-네이버 초창의적 AI연구센터 공동센터장
2021~현재 AI미래포럼 공동의장
2022~현재 한국공학한림원 컴퓨팅분과 일반회원
2022~현재 대통령직속 디지털플랫폼정부위원회 AI/Data 분과위원장
관심분야:인공지능, 기계학습, 자연어처리, 컴퓨터비전, 음성언어처리
Email: jungwoo.ha@navercorp.com