신뢰할 수 있는 언어 기초 모델의 연구 동향

포항공과대학교 ▌박상돈

1. 서 론

Google에서 제안한 새로운 인공 신경망인 transformer [1]로 시작해 이를 기반으로 한 대화형 서비스인 ChatGPT로 인해 생성 모델은 급격한 발전을 이루었다. 특히, ChatGPT의 인간 수준의 성능으로 언어 기초 모델(language foundation model)[2]에 대한 관심이 급격히 고조되었고 언어 기초 모델과 인간의 상호 작용도 마찬가지로 급증하게 되었다. 이에 따라 모델 개발 시 의도하지 않았던 환각 효과 등의 신뢰성 문제로 인해 언어 기초 모델이 사회에 미치는 악영향도 마찬가지로 큰 문제가 되고 있다. 이에 더해 2023년 10월에 공표된 미국의 신뢰할 수 있는 인공 지능(trustworthy AI) 개발을 촉구하는 행정 명령으로 인공 지능의 신뢰성 문제에 대한 관심이 급증하고 있다[3].

본고에서는 생성 모델 중 인간과 상호 작용을 손쉽게 유발하여 신뢰성 문제가 두드러지는 언어 기초 모델의 신뢰성 문제에 관한 연구 동향을 소개하려 한다. 구체적으로, 먼저 언어 기초 모델과 이를 학습하기 위한 필요 요소에 관해서 설명하고 신뢰할 수 있는 인공 지능의 속성을 안전성과 밀접한 관련이 있는 성능 보장성, 제어 가능성, 견고성, 보호성 및 사회성으로 분류 및 정의하고 각 안전성 속성을 언어 모델에서 부여하기 위해 진행되고 있는 연구를 소개하겠다. 그리고 나아갈 연구 방향을 제시한 후 마치려고 한다.

2. 언어 기초 모델과 신뢰성

해당 섹션에서는 언어 기초 모델 및 그 학습 환경과 본고에서 고려하는 신뢰성 기준에 대해서 안전성을 기준으로 설명하려 한다.

2.1 언어 기초 모델 및 학습 환경의 정의

기초 모델(foundation model)은 다양한 데이터에 대해서 학습되고 폭 넓은 다운스트림 작업(downstream task)에 적응할 수 있는 모델을 말한다[2]. 기초 모델로 언어, 시각, 로봇 제어 등의 다양한 모델이 존재하지만 본고에서는 상업적 활용성이 크고 인간과 손쉬운 상호 작용을 유발할 수 있어서 신뢰성이 유독 강조되는 언어 기초 모델(language foundation model)에 대해서 다루려 한다.

언어 기초 모델은 입력과 출력이 텍스트인 모델을 말한다. 구체적으로, 입력 텍스트 x와 출력 텍스트 y는 텍스트 처리 최소 단위인 토큰(token)의 연속으로 이루어져 있고, 언어 기초 모델 f는 입력에 대해 출력을 대응한다(즉, $y = f(x)$).

언어 기초 모델은 데이터 분포(data distribution)가 주어지고 데이터 분포에서 수집한 데이터(data)를 가지고 자기 지도 학습(self-supervised learning)으로 대표되는 학습 알고리즘(learning algorithm)을 통해서 학습되게 된다. 신뢰할 수 있는 언어 기초 모델은 데이터 분포, 데이터, 학습 알고리즘, 및 학습된 언어 기초 모델을 포함한 모든 부분이 신뢰성 조건을 만족해야 얻을 수 있다. 다음은 본고에서 고려하는 신뢰할 수 있는 인공 지능의 속성에 대해서 언급하겠다.

2.2 신뢰할 수 있는 인공 지능의 속성

신뢰할 수 있는 인공 지능 속성으로 안전성(safety), 공평성(fairness), 투명성(transparency), 책임성(responsibility) 등 다양한 속성이 알려져 있다[4]. 하지만 이런 속성은 서로 연관이 되어 있어서 정확히 구분하기 힘든 경향이 있다. 이에 본고에서는 안전성의 입장에서 관련된 세부 속성을 분류하고 설명하도록 하겠다[그림 1].

안전한 인공 지능이란 사회에 해를 미치지 않는 인공 지능을 말한다. 이때, 사회에 미치는 해는 여러 가지 성능 지표로 표현할 수 있고 이는 세부 속성에 따라 달라진다. 가령, 인공 지능 모델이 환각 효과(hallucination)로 인해 기대한 성능을 만족하지 못하거나 모델의 출력이 편향(bias)된 결과를 도출해 사회 규범에 따르지 않아 직간접적으로 사회 전반에 해를 미치는 경우를 들 수 있다. 아래에서는 안전성 측면에

안전성

인공 지능이 사회에 해를 미치지 않는 경우 안전하다고 말한다.

성능 보장성(reliability)	인공 지능 모델의 추론 성능을 보장하는 속성으로 사용자에게 모델 성능 기대치를 제시
제어 가능성(controllability)	인공 지능 학습 알고리즘이 희망 성능을 보장하는 모델을 찾을 수 있는 속성으로 사용자에게 학습 알고리즘이 찾는 모델의 성능 기대치를 제시
견고성(robustness)	인공 지능 모델이 적대적 사례에 대해서 성능을 유지하는 속성으로 적대적인 상황에서 사용자에게 모델 성능 기대치를 제시
보호성(protection)	인공 지능 모델이 민감한 정보를 드러내지 않을 속성으로 모델 이해관계자의 정보 및 지적재산을 보호
사회성(social)	인공 지능 모델이 사회적인 규범을 지킬 속성으로 사용자에게 모델 출력의 사회적인 기대치를 제시

그림 1 안전한 인공 지능이 지녀야 할 주요 속성

서 흔히 언급되는 5가지의 인공 지능의 세부 안전성 속성에 관해서 설명하도록 하겠다.

2.1.1 성능 보장성(reliability)

인공 지능 모델의 추론 성능을 보장하는 속성이다. 성능 평가 기준은 작업(task)에 따라 다르게 정해진다. 가령, 항등 분포 가정하에 분류 모델의 경우 평가 데이터 세트(test set)에서 추론의 정확성과 이의 신뢰 구간을 이용하여 성능을 평가한다. 언어 기초 모델의 경우 주어진 텍스트에서 다음 토큰 또는 텍스트의 추론 정확성을 사용한다. 사용자는 성능 보장성을 만족하는 모델에 대해 특정 수준의 성능을 기대하게 되고 이런 성능 보장성이 만족하지 않게 되면 사용자는 모델이 기대 또는 가치에 부합하지 않는다고 판단한다. 가령, 우수한 성능이 기대되는 ChatGPT가 환각 효과를 유발하면 사용자는 해당 모델이 거짓 정보를 생성하는 안전하지 않은 모델이라고 판단한다.

2.1.2 제어 가능성 (controllability)

제어 가능성은 성능 보장성과 밀접한 관계를 가지고 있다. 즉, 학습 알고리즘이 주어진 모델 집합에서 사용자가 희망하는 성능을 보장하는 모델을 찾을 때, 학습 알고리즘은 해당 모델 집합에서 제어 가능성을 가지고 있다고 말한다. 가령, 얼추 거의 맞는(probably approximately correct, PAC) 알고리즘[5]은 제어 가능한 알고리즘의 고전적인 예이다. 제어 가능한 알고리즘은 앞서 말한 성능 보장이 되는 모델을 찾으므로 사용자의 안전성에 대한 기대에 부합하게 된다.

2.1.3 견고성(robustness)

인공 지능 모델의 입력으로 적대적 예제(adversarial example)가 주어지면 모델이 잘못된 추론을 하고 기대하는 성능을 만족하지 못한다[6]. 하지만 이런 적대적인 상황에서도 기대하는 성능을 보장하는 경우 모델은 견고성을 가지고 있다고 말하고 성능 보장하는 모델과 마찬가지로 사용자의 안전성에 대한 기대에 부합하게 된다.

2.1.4 보호성(protection)

모델의 학습 데이터에 사용자의 민감한 정보가 들어있을 수 있고, 모델의 파라미터를 지적 재산으로 보아 민감한 정보로 고려할 수 있다. 모델이 이런 민감한 정보를 드러내지 않으면 보호성을 가지고 있다고 말하고 이는 안전성과 직접적으로 관련이 되겠다.

2.1.5 사회성(social)

모델의 추론 결과가 사회적인 규범을 지킬 때 사회성을 지닌다고 말한다. 여기서 사회적인 규범은 광범위하게 정의 가능하고, 일 예로 편향적인 대답이나 혐오 발언을 피하는 모델을 사회적인 모델이라고 말하며 안전한 모델이라고 고려하겠다.

3. 언어 기초 모델의 안전성 연구 동향

해당 섹션에서는 언어 기초 모델의 안전성 연구 동향을 안전한 인공 지능이 가지는 주요 속성별로 구분하여 살펴보고자 한다. 분류기(classifier)와 회기자(regressor)로 대표되는 고전적인 모델에서의 안전성 연구와는 달리 언어의 다중적, 복합적, 사회적인 속성 때문에 언어 모델에서 안전성 연구는 구분이 된다.

3.1 성능 보장성 및 제어 가능성 연구 동향

일반적으로 인공 지능 모델의 성능 보장성은 독립항등 분포(independent and identical distribution) 가정하에 평가 데이터(test dataset)에서 성능에 대한 평균과 오차 범위 등의 통계적인 방법을 통해서 만족할 수 있다. 하지만 언어 기초 모델 및 다운스트림 작업에서 미세 조정(finetuing)된 언어 모델의 경우 무수히 많은 출력이 정답이 되는 정답의 다양성으로 인해서

정확한 일치(exact match) 또는 F1 수치(F1 score) 등 상용자 가치와 일치하지 않는 성능에 대해 보장한다. 이런 성능 평가 기준 불일치[7]로 인해서 사용자 요구에 맞는 성능 보장성이 만족하기가 힘들다. 그래서 차선책으로 언어 모델의 평가는 보수적인 EM을 많이 사용하고 있다[8,9].

언어 기반 모델의 성능 보장성에서 더 나아가 성능 보장이 되는 모델을 찾는 학습 알고리즘의 제어 가능성에 관한 연구가 환각 효과(hallucination effect)의 현재 해법으로 언급되고 있다. 특히, 언어 기초 모델의 추론에 대한 불확실성(uncertainty)을 집합의 크기로 모델링하는 정합 추론(conformal prediction)[10]이 주류를 이루고 있다.

구체적으로 정합 추론은 언어 기초 모델을 기반으로 정합 집합 모델(conformal set model) $\hat{C}$을 학습하여 입력 x가 주어지면 정합 집합(conformal set) $\hat{C}(x)$을 출력하는 방식으로 추론 하는 방식을 말한다. 이때, 정합 집합 모델의 학습 알고리즘은 언어 모델의 입력에 대해 참인 출력 y를 정합 집합이 특정 확률로 포함하도록, 즉 특정 커버리지(coverage)를 만족하는, 정합 집합 모델을 찾게 된다. 구체적으로 커버리지는 $\Pr(y \in \hat{C}(x))$로 정의 된다. 여기서 사용자는 학습 알고리즘에 희망하는 커버리지 수준 α를 정하면 알고리즘은 이 수준을 지키는 정합 모델을 찾게 된다. 즉, 이는 제어 가능성을 보장하는 알고리즘이다.

학습된 정합 집합은 참인 출력을 가질 가능성이 보장되므로 사용자는 이 보장을 고려하여 정합 집합의 원소 중 정답을 최종적으로 선택할 수 있고 이는 사용자가 최종적으로 올바른 결정을 하는 데 도움이 된다[11]. 여기서 정합 집합이 크다는 말은 참일 가능성이 높은 출력이 많다는 말이므로 직관적으로 불확실성이 높다고 생각할 수 있다.

3.1.1 언어 모델

일반적으로 정합 집합이 주어지면 사용자는 최종적으로 집합의 원소 중 답을 하나 고르게 된다. 즉, 정합 집합의 모든 원소를 나열할 수 있어야 한다는 말이다. 이는 유한한 출력 집합을 고려하는 고전적인 분류 문제에는 적용할 수 있지만 언어 관련 작업의 경우 대답과 동일한 무수히 많은 대답이 존재하므로 정합 집합의 모든 원소를 나열하는 것은 불가능에 가깝다.

이 문제는 무수히 많은 대답 중 표본 추출을 통해서 대답을 고르면서 정답을 포함하게 정합 집합을 구성하는 방법으로 해결할 수 있다[12]. 구체적으로 주어진 입력에 대해서 언어 기초 모델에서 출력을 생성하고 생성된 대답의 신뢰성이 높고, 기존에 생성된 대답이랑 너무 유사하지 않으면 정합 집합에 점진적으로 추가하는 방법을 사용한다. 그리고 최종적으로 정합 집합이 충분히 좋은 대답을 많이 가지고 있으면 점진적 생성을 정지하는 방식으로 최종적인 정합 집합을 형성하게 되는 방법이다. 이렇게 정합 집합을 구성하는 파라미터를 학습하는 알고리즘은 기존의 정합 추론과 마찬가지로 최종 정합 집합이 사용자가 원하는 확률로 정답을 갖는 제어 가능성을 가지고 있다.

3.1.2 검색 증강 언어 모델

언어 기초 모델은 최신 정보를 반영하기 힘들어서 최신성 정보랑 관련된 환각 효과를 가지고 있다. 이를 해결하는 유효한 방법으로 검색 증강 생성 기법(retrieval-augmented generation, RAG)이 있다[13]. 이는 언어 모델이 검색 모델을 이용하여 질문과 관련된 최신 정보를 검색하여 최종 대답을 생성하는 구조를 따른다. 하지만 최신성 문제랑 별개로 검색 모델과 언어 모델이 가지고 있는 대답의 불확실성에 기인한 환각 효과는 여전히 제어하지 못한다.

이를 해결하기 위해서 검색 모델과 언어 모델에 정합 모델을 적용하여 최종적으로 합성된 정합 모델이 정답을 포함하는 확률을 사용자가 원하는 확률로 제어하는 방법이 제시되었다[14]. 구체적으로 검색 모델은 질문을 입력으로 받으면 출력으로 관련 문맥(passage)의 집합을 갖는다. 이 집합을 구성하는 방법은 다양하지만, 집합을 학습하는 방법을 정합 집합을 학습하는 방법을 사용하게 되면 사용자가 원하는 확률로 검색 모델의 정합 집합은 올바른 관련 문맥을 갖게 된다. 이와 비슷하게 언어 모델의 출력도 일반적인 정합 추론이나 이전 섹션에서 제시한 방법으로 정합 집합을 출력하도록 한다. 이렇게 두 개의 정합 집합이 주어지면 union bound를 이용하여 최종적으로 생성되는 두 정합 집합의 곱집합은 최종적으로 사용자가 원하는 확률로 정답을 포함하도록 제어할 수 있는 제어 가능한 알고리즘이 됨이 증명 되어 있다.

이렇게 정합 추론 기법을 이용하여 환각 효과를 줄이는 방법이 제시되었다. 하지만 앞의 기법은 독립 항등 분포 가정을 따르기 때문에 적대적인 입력이 들어오면 제어 가능성이 깨지고 만다. 다음은 적대적인 환경에서 언어 모델의 문제점을 드러내는 연구를 소개하도록 하겠다.

3.2 견고성 연구 동향

ChatGPT로 대표되는 언어 기초 모델의 우수성을 보여주는 상용 제품이 출시됨에 따라 해당 언어 모델

의 취약점을 찾으려는 화이트 해커의 시도가 많이 있었다. 특히, 프롬프트 인젝션(prompt injection)으로 대표되는 공격 방법은 언어 기초 모델이 의도하지 않은 정보를 출력하게 하고, 언어 모델의 다양한 사례와 연계하여 공격을 구현할 수 있다.

이렇게 대상 언어 모델의 성능을 저하하기 위한 적대적인 입력에 대해서 성능을 유지하는 속성을 견고성(robustness)이라고 하고, 본 섹션에서는 언어 모델에 특화된 적대적 입력 생성 방법과 알려진 방어 방법에 대해서 알아보도록 하겠다.

3.2.1 언어 모델

GPT-3 같은 언어 기초 모델은 초거대 데이터를 이용하여 다음 토큰을 올바르게 추측하도록 학습된다. 이런 방법은 대상 작업에 대한 몇 개의 입력-출력 쌍을 프롬프트로 제공함으로써 해당 작업에 손쉽게 적응할 수 있는, 즉 in-context 학습을 쉽게 할 수 있다는 사실이 실험적으로 검증된 바 있다[8]. In-context 학습이 가능한 점은 더 나아가 언어 기초 모델이 주어진 프롬프트에 쉽게 조건화(conditioning)가 가능하다는 점을 시사하고 이를 이용하여 언어 모델을 공격하는 다양한 방법이 알려져 있다.

특히, 프롬프트 인젝션(prompt injection) 공격은 언어 모델의 쉬운 조건화 성질을 이용하여 언어 모델이 공격자가 원하는 대답을 하게 만든다. 가령, GPT-3와 지시 프롬프트를 이용하여 대학 지원 서류를 평가하는 작업에 특화된 언어 모델을 생각할 수 있다. 즉, 지원 서류가 주어지만 언어 모델은 합격 또는 불합격을 결정한다. 하지만 공격자는 이 언어 모델에게 지시 프롬프트가 무엇인지 알아내거나, 지시 프롬프트를 무시하고 항상 합격을 결정하게 만들거나[15], 역시 지시 프롬프트를 무시하고 원래 GPT-3의 기능을 하게 만드는 공격[15]을 수행할 수 있다.

앞의 프롬프트 인젝션 공격을 방어하기 위한 다양한 휴리스틱 방법이 제시됐다. 입력된 데이터 프롬프트를 의역(paraphrase)[16]하거나, 지시 프롬프트를 다시 상기[17]시키거나 공격을 고려해 지시 프롬프트를 설계[18]하는 등의 공격의 영향을 줄이는 공격 방지 방법이 있다. 또한 데이터 프롬프트의 perplexity를 이용하여 비정상적으로 perplexity가 높은 데이터 프롬프트를 탐지[19]하거나, 주어진 데이터 프롬프트에 더해 추가로 답을 알고 있는 질문을 요청함으로써 데이터 프롬프트의 올바름을 탐지[20]하는 방법이 알려져 있다. 하지만 앞에서 소개된 방법은 휴리스틱 하여 그 효과가 보장되지 못한다.

3.2.2 언어 모델이 통합된 애플리케이션

언어 기초 모델의 최신성을 유지하기 위해서 검색 증강 생성(RAG)을 포함하여 외부 애플리케이션을 이용하는 사례가 늘고 있다. 여기서 외부 데이터베이스 또는 애플리케이션을 공격하여 언어 모델이 오답을 생성하게 만드는 공격이 알려져 있다.

앞의 프롬프트 인잭션 공격을 실현하기 위해서 간접적인 프롬프트 인잭션(indirect prompt injection)[21]을 수행할 수 있다. 이는 웹페이지의 코맨트에 프롬프트를 숨겨두어 이를 입력으로 이용하는 언어 모델에 프롬프트 인젝션을 수행하게 하는 실용적인 공격 방법이다. 또한 PoisonedRAG[22]는 공격자가 검색 대상 문서를 수정하여 검색 모델이 잘못된 정보가 들어 있는 문서를 언어 모델에 제공하게 하여 오답을 유발하는 공격을 선보였다.

3.3 보호성 연구 동향

인공 지능 모델의 보호성은 모델이 민감한 정보를 드러내지 않을 속성이다. 언어 기초 모델의 사용 관점에서는 학습 데이터에 개인 정보 및 저작권 보호가 필요한 코드 등의 데이터가 개인 정보와 관련된 민감한 정보에 해당하고, 언어 기초 모델의 작업 적응을 위한 지시 프롬프트와 언어 모델의 파라미터 정보가 서비스 제공자의 지적 재산과 관련된 민감한 정보에 해당한다. 이랑 관련된 연구 동향을 설명하겠다.

3.3.1 개인 정보 추출

언어 기초 모델은 초거대 데이터를 이용하여 자기지도 학습을 통해 학습되므로 데이터에 전화번호, 이메일 주소, 저작권 보호 코드 등의 다양한 민감한 정보가 데이터에 포함될 수 있다. 이렇게 학습 데이터로 사용한 개인 정보는 언어 모델이 overfitting으로 데이터를 기억할 때 유출이 된다. 이 점을 이용하여 언어 기초 모델의 학습 데이터에서 사용한 민감한 정보를 추출할 수 있음을 GPT-2를 이용하여 보인 바 있다[23]. 구체적으로 GPT-2를 이용하여 임의로 생성된 문장을 perplexity를 포함한 6가지 민감정보 검출 기준을 이용해 정렬한 뒤 수작업을 통한 검색으로 개인 정보가 유출됨을 확인하였다. 결과적으로 개인 이름이 담긴 46의 예시와 알려진 코드와 거의 동일한 31개의 예시 등 민감한 정보가 GPT-2를 통해 추출됨을 알 수 있었다. 모델 추출을 방어할 수 있다고 알려진 방어 방법으로는 차분 프라이버시(differential privacy)를 이용한 모델 학습, 데이터 세트 필터링, 생성된 문장의 민감정보 탐지 후 후처리를 통한 방어가 있지만 보통

성능을 저해한다고 알려져 있다.

3.3.2 모델 추출 공격 및 방어

언어 기초 모델이 상용 서비스로 이용되는 상황에서 모델의 작업 적응을 위한 지시 프롬프트와 모델 파라미터는 지적 재산으로 민감한 정보에 포함이 된다. 지시 프롬프트의 추출은 앞선 프롬프트 인젝션 공격으로 지시 프롬프트를 출력하라는 프롬프트로 추출(탈옥 공격, jailbreaking attacks)이 가능함을 보인바 있다[24].

언어 모델에서는 탈옥 공격뿐만 아니라 고전적인 모델 추출 공격(model stealing attack)도 가능함이 최근에 알려졌다. 구체적으로, [25]에서는 상용 모델인 ChatGPT 또는 구글의 PaLM-2의 임베딩 투사층(embedding projection layer) 모델의 매개변수를 추출할 수 있음을 보였다. 일반적으로 서비스로 제공되는 언어 모델의 임베딩 공간의 크기는 알 수 없지만, 임의 프롬프트에 대한 대답의 logit을 알 수 있으면 선형대수적 성질을 이용하여 추출할 수 있음을 보였다. 하지만 이는 상위 k개의 logit만 알려주는 상용 언어 모델 API에는 적용 불가능하다. 이 문제를 해결하기 위해서 사용자 지정한 bias term에 대한 logit을 알 수 있는 API의 특성과 logsoftmax의 수학적 특성을 이용하여 logit을 추출하는 알고리즘이 제시되었다.

위 모델 추출 공격은 최근에 소개가 되어서 이에 대한 방어 기법은 잘 알려진 바 없다. 해당 논문에서는 사용자 지정 bias term을 사용하지 못하게 API를 수정하는 공격 의존적이거나 logit에 noise를 넣는 공격 비의존적인 방어 기법이 소개되어 있다.

3.4 사회성 연구 동향

언어에는 사회 및 문화적인 정보가 담겨있고, 이런 언어 기반의 데이터로 언어 기초 모델은 학습되므로 언어 모델에서 편향성 및 다양성 문제는 주로 사회적인 편향(social bias) 문제와 연결되어 있다. 가령, 특정 사회적 그룹에 대한 혐오 발언 또는 고정관념 표출이나 사회적 그룹 간의 성능 편차를 들 수 있다[26]. 이런 사회적 편향 문제 해결 방안에 대해서 언어 모델의 작업 흐름을 기준으로 알아보려 한다. 구체적으로 언어 모델의 편향성은 학습에 사용되는 데이터, 학습에서 고려되는 학습 알고리즘, 추론 알고리즘 및 추론 결과의 후 처리를 통해서 완화 가능하다. 아래는 대표적인 방법을 소개하려 하고, 좀 더 자세한 연구 리뷰는 [26]를 참고하길 바란다.

3.4.1 데이터의 편향성 완화

언어 기초 모델은 학습에 쓰이는 언어 데이터에 대해서 편향이 일어날 수밖에 없다. 데이터에 존재하는 편향성을 제거하기 위해 데이터 증강, 데이터 reweighting, 데이터 생성 방법을 사용할 수 있다.

구체적으로, [27]는 자기 주도 학습 과정에서 성별과 관련된 단어를 임의로 다른 성별에 대응하는 단어로 수정하는 데이터 증강 방법을 이용해 학습을 진행한다. 데이터 reweitghing 방법으로, 성별 분류기(gender classifier)를 지도 학습으로 학습할 수 있으면 성별 분류기를 사용하여 성별 분류가 용의한 데이터 편향이 있다고 생각하여 중요도를 낮춰서 언어 모델의 학습에 사용한다 [28]. 만약 성별 분류기를 지도 학습으로 학습 불가능하면 언어 모델이 downsteam 작업을 쉽게 풀 수 있는지 판별하는 분류기를 학습하여 분류가 쉬운 데이터는 성별 정보를 이용한다는 경향성을 이용하여 풀기 쉬운 데이터는 중요도를 낮춰서 학습한다. 마지막으로 사회적인 규범을 지키는 대화형 데이터 세트를 새롭게 생성 후 언어 모델을 학습하면 사회적 편향성을 근본적으로 줄일 수 있다[29].

3.4.2 학습 알고리즘에서 편향성 완화

학습 알고리즘에서 편향성 완화는 학습 목적 함수를 수정하여 언어 모델의 파라미터를 학습하는 방법이 주류를 이룬다. 구체적으로, 입력이 남성 또는 여성 등 여러 그룹으로 나뉠 때 언어 모델이 생성하는 서로 다른 그룹의 임베딩 벡터를 유사하게 만드는 방식으로 언어 모델을 학습하면 그룹 간의 성능 편향성이 완화된다[30].

임베딩 벡터를 고려하는 편향성 완화 기법 외에 생성되는 토큰의 분포를 고려해 완화하는 방법도 있다. 구체적으로, 각 그룹에 속하는 토큰이 생성되는 확률을 같도록 유지하는 학습 목표를 기존 학습 목표에 넣어서 모델을 학습하는 방법이 있다[31].

3.4.3 추론 및 후처리 과정에서 편향성 완화

언어 모델 학습 알고리즘의 목적 함수를 수정하여 학습하는 경우 학습 비용이 발생한다. 이를 피하고자 기존에 학습된 언어 모델의 디코딩 과정에서 편향성을 완화하는 방법이 제시됐다.

[32]에서는 빔서치(beam search)를 이용하여 토큰을 생성하되 성별 관련 편향성에 대한 제한 조건이 있어서 편향이 적은 문장을 생성하게 한다. 이와 다르게 생성 과정에서 점진적으로 생성되는 토큰이 속성 분류기가 편향성을 완화하는 방향으로 잠재 공간을 업데이트하여 최종적으로 편향이 덜한 문장을 추론 과

정에서 생성하는 방법도 있다[33].

마지막으로 추론이 끝난 후 후처리를 하는 방법도 있다. 특히, 생성된 문장을 편향이 완화된 문장으로 변환하는 기계 번역[34] 또는 스타일 변형[35] 문제로 보아 편향성을 완화하기도 한다.

4. 결 론

본고에서는 언어 기초 모델에 관한 안전성 연구 동향에 대해서 알아보았다. 견고성 및 보안성 문제는 이제 문제가 제기되어 해결책을 생각해야 하는 단계이고 사회성 연구는 기존의 공평성 연구와 많은 부분을 공유하되 언어 모델의 특성을 이용하는 연구가 선호되고 있다. 성능 보장성 및 제어 가능성 연구는 안전성을 확보하기 위한 이론적 보장을 주는 연구들로 초기 단계에 머물러 있다. 필자는 다른 안전성 속성에서도 제어 가능성 연구가 필요하다고 본다.

참고문헌

[1] Vaswani, A. et al.. “Attention is all you need.” Advances in Neural Information Processing Systems, 2017.

[2] Bommasani, R. et al., “On the opportunities and risks of foundation models,” arXiv preprint arXiv:2108.07258, 2021.

[3] The White House, “Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence,” 2023.

[4] 한국정보통신기술협회, “신뢰할 수 있는 인공지능 개발 안내서,” 2024.

[5] Valiant, Leslie G, “A theory of the learnable,” Communications of the ACM 27, no. 11 1134-1142, 1984.

[6] Szegedy, C. et al, “Intriguing properties of neural networks.” International Conference on Learning Representations, 2014.

[7] Lee, M. et a.l. “Semi-Supervised Selective Generation for Trustworthy Language Models.” arXiv preprint arXiv: 2307.09254, 2024.

[8] Brown, T. et al, “Language models are few-shot learners.” Advances in Neural Information Processing Systems, 2020.

[9] Touvron, H., et al, “Llama: Open and efficient foundation language models.” arXiv preprint arXiv:2302.13971., 2023.

[10] Vovk, V, et al., “Algorithmic learning in a random world,” Vol. 29. New York: Springer, 2005.

[11] Straitouri, E. et al, “Improving expert predictions with conformal prediction,” In International Conference on Machine Learning, 2023.

[12] Quach, V. et al., “Conformal language modeling,” International Conference on Learning Representations, 2024.

[13] Lewis, P. et al., “Retrieval-augmented generation for knowledge-intensive nlp tasks.” Advances in Neural Information Processing Systems, 2020.

[14] Li, S. et al, “TRAQ: Trustworthy Retrieval Augmented Question Answering via Conformal Prediction,” Annual Conference of the North American Chapter of the Association for Computational Linguistics, 2024.

[15] Liu, Y. et al., “Prompt injection attacks and defenses in llm-integrated applications,” arXiv preprint arXiv: 2310.12815, 2023.

[16] Jain, N. et al., “Baseline defenses for adversarial attacks against aligned language models.” arXiv preprint arXiv:2309.00614, 2023.

[17] Learn Prompting, “Sandwich Defense,“ 2023.

[18] Learn Prompting, “Instruction Defense,“ 2023.

[19] Alon, G. and Michael K., “Detecting language model attacks with perplexity,” arXiv preprint arXiv:2308. 14132, 2023.

[20] https://twitter.com/yoheinakajima/status/ 1582844144640471040, 2022.

[21] Greshake, K. et al.,. “Not what you’ve signed up for: Compromising real-world llm-integrated applications with indirect prompt injection,” In Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security, 2023.

[22] Zou, W. et al, “Poisonedrag: Knowledge poisoning attacks to retrieval-augmented generation of large language models,” USENIX Security Symposium, 2025.

[23] Carlini, N. et al., “Extracting training data from large language models.” In 30th USENIX Security Symposium, 2021.

[24] Liu, Y. et al, “Prompt Injection attack against LLM-integrated Applications.” arXiv preprint arXiv: 2306. 05499, 2023.

[25] Carlini, N. et al., “Stealing part of a production language model,” International Conference on Machine Learning, 2024.

[26] Gallegos, I. et al, “Bias and fairness in large language models: A survey.” Computational Linguistics, 2024.

[27] Ghanbarzadeh, S. et al, “Gender-tuning: Empowering

fine-tuning for debiasing pre-trained language models.” Findings of the Association for Computational Linguistics, 2023.
[28] Orgad, H. and Yonatan B., “BLIND: Bias removal with no demographics,” Annual Meeting of the Association for Computational Linguistics, 2023..
[29] Kim, H. et al., “Prosocialdialog: A prosocial backbone for conversational agents,” Empirical Methods in Natural Language Processing, 2022.
[30] Huang, P. et al., “Reducing sentiment bias in language models via counterfactual evaluation,” Empirical Methods in Natural Language Processing, 2020.
[31] Qian, Y. et al., “Reducing gender bias in word-level language models with a gender-equalizing loss function.” arXiv preprint arXiv:1905.12801, 2019.
[32] Saunders, D. et al., “First the worst: Finding better gender translations during beam search,” Findings of the Association for Computational Linguistics, 2022.
[33] Dathathri, S. et al., “Plug and play language models: A simple approach to controlled text generation,” International Conference on Learning Representations, 2020.
[34] Vanmassenhove, E. et al., “Neutral rewriter: A rule-based and neural approach to automatic rewriting into gender-neutral alternatives.” Empirical Methods in Natural Language Processing, 2021.
[35] Tokpo, E. K. and Calders, T., “Text style transfer for bias mitigation using masked language modeling,” arXiv preprint arXiv:2201.08643, 2022.

약 력

박 상 돈

2010 서울대학교 컴퓨터공학과 졸업(학사)
2012 서울대학교 전기컴퓨터공학부 졸업(석사)
2021 미국 University of Pennsylvania(박사)
2023 미국 Georgia Institute of Technology(포닥)
2023~현재 POSTECH 인공지능대학원 조교수
관심분야 : 신뢰가능한 인공지능, 기계 학습, 컴퓨터 보안
Email : sangdon@postech.ac.kr