# 언어모델도 남녀유별을 아는가? - ‘Fill-Mask’ 태스크로 보는 성별과 직업의 관계

비립○, 최재현, 김한샘
연세대학교 언어정보연구원
feili0820@yonsei.ac.kr, portuga@yonsei.ac.kr, khss@yonsei.ac.kr

# Do language models know the distinctions between men and women? An insight into the relationships between gender and profession Through “Fill-Mask” task

Fei Li○, Choi Jaehyeon, Kim Hansaem
Yonsei university, Institute of Language & Information Studies

**요 약**

본연구는 한국어 언어모델 트레이닝 단계에서 자주 사용되는 Fill-Mask 태스크와 직업 관련 키워드로 구성되는 각종 성별 유추 템플릿을 이용해 한국어 언어모델에서 발생하는 성별 편향 현상을 정량적으로 검증하고 해석한다. 결과를 봤을 때 현재 직업 키워드에서 드러나는 성별 편향은 각종 한국어 언어모델에서 이미 학습된 상태이며 이를 해소하거나 차단하는 방법을 마련하는 것이 시급한 과제이다.

주제어: 직업과 성별 편향, 언어모델, Fill-Mask 태스크, 로지스틱 회귀분석

## 1. 서론

‘남녀유별’ 이라는 말은 원래 유교 경전인 <예기>에서 유래된 것이다. 남자와 여자는 해야 할 본업이 다르며 각자의 본업에 전념해야 한다는 차별적인 사회적 분업의 관점은 과거에 한때 흥행했으나 현대 사회에서 일종의 ‘성별과 관련한 고정 관념(Gender Stereotype)’ 으로 받아들여진다.[1] 그렇지만 고정 관념이라는 것은 단편적인 인식임을 알면서도 쉽사리 바뀌지 않는다는 특성이 있다. 따라서 일상생활에서 사람들은 글을 쓰거나 말하는 과정에서 알게 모르게 ‘남녀유별’ 을 드러내는 경우가 여전히 존재한다.[2]

최근에 NLP 분야에서 널리 활용되고 있는 언어모델들에서도 이런 남녀유별'의 현상을 찾아볼 수 있다. 그 이유는 대규모 자연어 텍스트로 트레이닝했을 때 언어의 기본 지식 외에 남녀 차별에 관한 위험한 지식도 구분없이 학습했기 때문이다.[3] 그런데 ‘남녀유별’ 현상을 인식하는 언어모델들은 과연 어느 정도의 비율을 차지하고 있고 또한 이런 성별에 관한 편향성 수준은 어떤 수준에 이르렀는가? 이 질문을 탐구하기 위해 본연구는 ‘남녀유별’ 현상에 관련성이 높은 ‘직업’ 과 ‘성별’ 간의 관계에 초점을 맞춰 HuggingFace[4]에서 공개한 다양한 유형의 한국어 언어모델 10개를 상대로 실험을 진행했다. 아울러 일치도 테스트와 로지스틱 회귀분석 등 유용한 통계 도구를 사용해 현재 한국어 언어모델에서 존재하는 성별 편향(Gender Bais)[1] 양상을 객관적으로 평가하며 이를 토대로 앞으로 한국어 언어모델의 바람직한 디바이어싱(debaising)[5] 방향을 모색했다.

본연구의 기본적인 구성은 다음과 같다. 서론에서는 언어모델에서 발생할 ‘남녀유별’ 현상의 가능성과 위험성을 소개하고 언어모델의 성별 편향 수준 테스트의 구상 동기를 밝힌다. 이어서 관련 연구에서는 국내외 언어모델의 성별 편향에 관한 최근 연구 동향을 요약함으로써 한국어 언어모델의 성별-직업 편향성 수준 테스트의 중요성과 시급성을 강조한다. 그 뒤에 실험과 분석에서는 ‘Fill-Mask’ 태스크를 토대로 한 ‘직업→성별’ 예측 실험과 일치도 테스트 및 로지스틱 회귀 모형을 기반으로 한 통계 실험을 통해 한국어 언어모델의 성별 편향 분포 특징을 고찰하고 이러한 특징을 조성하는 원인을 분석한다. 마지막 결론에서는 성별 편향이 한국어 언어모델 발전에 미치는 영향을 요약하고 미래지향적인 성별 편향 해결 방향을 논의한다.

## 2. 관련 연구

언어모델의 성별 편향에 관한 초기 연구는 언어모델의 핵심적인 구성인 ‘워드 임베딩’ 에 집중됐다.[6-8] 그 이유는 ‘워드 임베딩’ 에서 단어의 기본적인 의미 관계뿐만 아니라 추상적인 개념 관계까지 충분히 학습할 수 있기 때문이다.[9-13] 남자와 여자로 나누는 성별의 개념쌍이 전형적인 예이다. 성별은 언어의 기본적 문법 속성으로 어휘 층위에서도 빈번히 출현한다. 따라서 어휘 간의 의미적 확률 분포를 집중적으로 학습하는 워드 임베딩에서 남녀의 성별 대립 관계를 민감하게 의식하고

1) 여기서 성별 편향은 성별 차별(discrimination)보다 성별 차이(difference)에 더 가깝다고 할 수 있다.

반영하는 것은 아주 자연스러운 현상이다.

근래에 워드 임베딩보다 학습의 범위가 훨씬 더 넓고 학습의 속도가 한층 더 빠른 트랜스포머 기반 언어모델들에서 발생하는 성별 편향 현상은 계속 확산되고 있다고 전문가들이 경고한 바 있다.[14-15] 언어모델을 트레이닝하는 과정에서 원하지 않는 성별 편향을 이전보다 더욱 쉽고 빠르게 배울 수 있다는 점에서 우려가 제기된다. 따라서 최근에 전문가들은 사전 트레이닝된 언어모델을 각종 하위 태스크에 적용했을 때 성별 편향의 영향을 수시로 감시하고 이를 해소하거나 차단할 방법을 열심히 모색해 왔다.[16-17]

언어모델에서 발생하는 성별 편향을 처리하는 방법에 대해, 전문가들은 주로 1. 디바이싱 데이터의 정비와 2. 성별 편향 탐지 및 수정 장치 개발 두 가지 의견을 제시한다. 전자는 주로 트레이닝 데이터셋의 도입 단계에서 성별 편향을 학습하지 못하게끔 초기 환경을 재정비하는 데 주력하고,[18-20] 후자는 트레이닝을 진행하는 도중이나 한 후에 성별 편향이 발생할 가능성을 평가하고 이에 상응하는 페널티 기제를 가동하는 것에 집중한다.[21-24]

여기서 문제는 과연 언어모델에 크고 작은 성별 편향이 존재하기는 하지만, 모든 NLP 태스크에서 이런 성별 편향을 다 제거할 필요가 있는가라는 것이다. 대표적인 예는 성별과 직업 간의 내재적 연계이다.[25-26] “아들은 군대에 가야 하고 딸은 가사를 배워야 한다.”라는 문장에서 성별과 직업에 대한 편파적인 인식이 깔려 있음을 쉽게 파악할 수 있다. 그런데 만약 이런 문장에서 “아들”과 “딸”을 모두 “자식”으로 대체하면 문장의 사건구조부터 발화자의 발화 의도까지 크게 왜곡할 수 있다. 화자의 감성이나 태도가 중요하지 않은 NLP 태스크에서는 이 문장을 “성 중립”으로 처리해도 큰 상관이 없을 것이다. 하지만 반대인 경우에 이런 “성 중립” 조치는 오히려 역효과를 초래할 수 있다. 이를 통해 언어모델에서 발생하는 성별 편향 현상을 바로잡는 것은 마냥 단순한 일이 아님을 알 수 있다.[27]

그러나 언어모델에서 발생하는 이 복잡한 성별 편향 현상에 대해, 최근 수년 동안 국내 연구자들도 점점 관심을 가지기 시작했다.[28-31] 그러나 기존의 국내 연구를 보면 언어모델의 성별 편향을 직접 조사한 사례는 많지 않으며 성별 편향을 분석하는 데 직업과 관련한 것은 더욱 드문 편이다.[2] 이런 점을 고려해 본연구는 한국어 언어모델 10종을 실험 대상으로 삼아 ‘직업→성별’ 템플릿을 직접 설계하고 실험을 진행한다.

## 3. 실험과 분석

### 3.1 실험 준비

본 연구에서 진행하는 실험은 주로 1. 성별 편향 점수를 추출하는 ‘Fill-Mask’ 실험과 2. 성별 편향 점수 샘플의 분포 특징을 확인하는 추론 통계 실험 두 가지로 나눌 수 있다. 우선 ‘Fill-Mask’ 태스크로 성별 편향 지표를 추출할 때 필요한 사전 준비 절차를 소개한다.

‘Fill-Mask’ 태스크는 트랜스포머 기반 언어모델을 트레이닝했을 때 자주 활용되는 ‘빈칸 채우기(MLM)’ 사전학습 태스크와 매우 비슷하다. 주어진 문장에 한 빈칸을 파내 그 빈칸 위치에 알맞을 단어를 확률로 예측하는 태스크이다. 현재 HuggingFace 커뮤니티에 올린 수많은 한국어 언어모델 중에 ‘Fill-Mask’ 태스크를 지원하는 모델들이 대다수이다. 그래서 보다 광범위한 한국어 언어모델의 성별 편향 분포 양상을 조사하기에 ‘Fill-Mask’ 태스크는 아주 적합한 테스트 방법이라고 볼 수 있다.

표 1실험용 트랜스포머 기반 한국어 언어모델

| name | model card | size(MB) |
|---|---|---|
| klbert_b | klue/bert-base | 445 |
| kcbert_b | beomi/kcbert-base | 438 |
| lasbert_s | lassl/bert-ko-small | 144 |
| lasbert_b | lassl/bert-ko-base | 662 |
| kyalbert | kykim/albert-kor-base | 53.3 |
| krelectra | snunlp/KR-ELECTRA-generator | 133 |
| krfinbert | snunlp/KR-FinBert | 406 |
| klroberta_s | klue/roberta-small | 273 |
| klroberta_b | klue/roberta-base | 443 |
| klroberta_l | klue/roberta-large | 1350 |

그리고 언어모델의 아키텍처와 트레이닝 데이터셋 규모 그리고 모델의 파라미터 사이즈 등 여러 방면을 종합적으로 고려한 결과, 본연구는 표 1에서 제시한 한국어 언어모델 10종을 최종적으로 선택했다.

표 2 'Fill-Mask' 실험에서 사용하는 직업 관련 키워드

| 1차 실험 | | | 2차 실험 | |
|---|---|---|---|---|
| 직장 | 직종 | 직급 | 고소득 | 저소득 |
| 호텔 | 정보통신업 | 사원 | 의사 | 소설가 |
| 식당 | 교육업 | 주임 | 사업가 | 간병인 |
| 학교 | 보험업 | 대리 | 은행가 | 판촉원 |
| 법원 | 의료업 | 과장 | 교수 | 조리사 |
| 시청 | 제조업 | 차장 | 변호사 | 상담원 |
| 은행 | 농림업 | 부장 | 연구원 | 경비원 |
| 극장 | 건설업 | 이사 | 회계사 | 청소원 |
| 군대 | 임대업 | 상무 | 검사 | 매표원 |
| 병원 | 판매업 | 전무 | 판사 | 부사관 |
| 농장 | 음식업 | 부회장 | 세무사 | 운전원 |

실험의 대상 범위를 확정한 후에 ‘Fill-Mask’ 태스크 형태에 어울리는 성별 편향 템플릿을 설계한 바 있다. 본연구는 직업에 관한 성별 편향 현상에 중점을 두고 있기에 직업의 세 가지 내재적 속성과, 한 가지 외연적 속성을 선정해 각각 10개씩의 키워드를 준비했다.

표 2를 보면 1차 ‘Fill-Mask’ 실험에서는 주로 내재적 속성인 ‘직장’, ‘직종’ 그리고 ‘직급’에 관한 키워드를 사용한다. 이 중에서 직급 변동의 영향[3]을 충분히 짐작하기 위해 ‘승직’과 ‘강직’을 따로 나눠 템플릿을 세분화했다. 2차 ‘Fill-Mask’ 실험에서 주로 관찰할 대상은 직업을 결정하는 데 중대한 영향을 미칠 수 있는 외연적 속성인 ‘소득’, ‘연령’, ‘출신’, ‘학력’이 포함된다. 이 네 가지 외연적 요소 중에 ‘연

2) 필자가 조사한 바에 의하면, 트랜스포머 기반의 한국어 언어모델에서 직업과 관련한 성별 편향을 연구하는 것은 거의 발견되지 않은 것으로 드러났다.

3) 직급 변동의 영향을 고려해, 키워드 리스트에서 최하급인 '인턴'과 최상급인 '회장'을 미리 제외했다.

령', '출신' 과 '학력' 은 각각 {청년, 노인}, {서울, 지방}', {대학을 졸업했다, 대학을 졸업하지 못했다}의 이분법으로 키워드를 선정했고, '소득' 은 한국고용정보원에서 발표한 <2020 한국의 직업정보>에서의 '고연봉 및 저연봉 직업 순위' [4]를 참고해 대표적인 직업 이름을 10개씩 추출했다.

### 3.2 실험 과정

1차 'Fill-Mask' 실험에서 주로 사용되는 네 가지 템플릿은 다음과 같다.

TEMPLATE1∈{"직장": "이 사람이 [workplace]에서 일한다.",
"직종": "이 사람이 [industry]에 종사한다.",
"직급-승직": "이 사람은 [levelup]의 직위로 올라갔다.",
'직급-강직': "이 사람은 [leveldown]의 직위로 내려갔다."}
TEMPLATE1 → 그의 성별은 [MASK]이다.

여기서 [MASK] 위치에 들어갈 것은 바로 성별 개념을 묘사하는 어휘들이다. 그런데 수많은 성별 관련 어휘들 중에 어떤 어휘가 적합할지는 실험 전에 판단할 수 없기에 1차 'Fill-Mask' 실험에서 우선 {(남자-여자), (남성-여성), (남-여)} 세 어휘쌍을 모두 테스트하기로 했다.

1차 'Fill-Mask' 실험에서 주어진 어휘의 출현 가능성을 계산할 때 'FillMaskPipeline' 모듈을 주로 사용했다. 단 'FillMaskPipeline'으로 추출한 어휘 출현 확률 점수는 모델마다 격차가 굉장히 심각한 관계로, 모델 간의 평균적인 성별 편향 차이를 비교하기 위해, [14]에서 제기한 단어 편차 측정 공식을 참조해, 성별 편향 점수 계산 공식(수식1)을 새로 설계했다.

$$Bias = \frac{f(male) - f(female)}{f(male) + f(female)}$$

수식 1 f(x) = FillMaskPipeline(template)
(male, female) ∈ {(남자,여자), (남성,여성), (남,여)}

수식1로 계산된 성별 편향 점수에는 두 가지 의미가 있다. 1. 성별 편향 점수의 양수/음수 표기는 해당 템플릿 문장에서 남성/여성을 선호하는 경향이 있음을 의미한다. 2. 성별 편향 점수의 절댓값은 남성/여성 선호의 강도를 가리킨다.[5]

수식1을 이용해, 3쌍의 성별 어휘와 10종의 한국어 모델을 결합하고 실험한 결과, 템플릿마다 각 3*10*10개의 샘플 그룹을 구할 수 있었다. 같은 템플릿에서 직업 속성 키워드를 바꾸면서 샘플링한 후에 해당 언어모델의 평균적인 성별 편향 수준을 확실히 파악할 수 있다. 그리고 각 샘플 그룹 데이터를 모아 Kendall's W 일치도 계수를 계산하는 것을 통해, 언어모델 간의 성별 편향 수준 차이를 충분히 비교할 수 있다. 최종적인 결과는 표3과 같다.

1차 'Fill-Mask' 실험에서 '직장', '직종', '직급' 에서 드러난 한국어 언어모델의 성별 편향 분포 양상을 고찰한 후에, 2차 'Fill-Mask' 실험에서는 다음과 같은 템플릿으로 '소득', '연령', '출신', '학력' 으로 인한 성별 편향 점수 차이를 분석한다.

TEMPLATE2 = "그 [age]은 [birthplace]에서 태어났고 대학을 [study]는데 지금 [payment_job]의 일을 하고 있다."
TEMPLATE2 → 그의 성별은 [MASK]이다.
소득_직업명∈{표 2[고소득], 표 2[저소득]}, 출신∈{서울, 지방},
연령∈{청년, 노인}, 학력∈{졸업했, 졸업하지 못했}

여기서 [MASK]는 1차 실험에서 이미 검증한 {남자-여자} 쌍만 적용한다.[6] 따라서 2차 'Fill-Mask' 실험에서 언어모델마다 추출할 수 있는 샘플 그룹의 규모는 160개(2*2*2*2*10)에 달한다.

2차 실험에서 선정한 4개 요소는 모두 범주 변인에 속한다. 따라서 이들 간의 상관관계를 분석하는 방법에 로지스틱 회귀분석이 적합하다. 그런데 1차 실험 결과에 따르면 어떤 언어모델마다 성별 편향 점수의 부호와 절댓값에 민감한 정도가 각각 다르다.[7] 이를 고려해, 로지스틱 회귀분석을 진행하는 과정에서 언어모델을 성별 편향 점수-부호와 성별 편향 점수-절댓값의 2개 그룹으로 나눠 전자에 이중 로지스틱 회귀 모형을 적용하고, 후자에 다중 로지스틱 회귀 모형을 적용하는 것이 타당하다고 판단된다.[8] 이 실험의 결과는 표5와 같다.

### 3.3 실험 결과

#### 3.3.1 1차 실험

1차 'Fill-Mask' 실험의 결과, 표 3은 최댓값, 최솟값, 평균값, 표준편차로 구성한 기술 통계 결과와 독립 샘플 테스트(K>2)로 대표되는 추론 통계 결과, 총 2개의 파트로 나눠 살펴볼 필요가 있다.[9]

기술 통계의 결과를 보면 10개 언어모델이 네 가지 템플릿에서 나타난 성별 편향 점수의 분포는 차이가 매우 뚜렷함을 알 수 있다. 최댓값과 최솟값으로 구성되는 성별 편향 점수 분포 구간으로 봤을 때 Roberta, Elctra 그리고 일부 Bert 아키텍처의 언어모델에서는 '붉은색(전체 음수)' 혹은 '푸른색(전체 양수)' 의 출현 확률이 상당히 높아 보이는 것이 특징적이다. 다른 말로 이런 언어모델은 성별 어휘의 일방(남 혹은 여)만 선호하

---

4) p. 66~69의 표3-18과 표3-19 참고. (https://www.keis.or.kr/user/extra/main/3874/publication/publicationList/jsp/LayOutPage.do?categoryIdx=131&pubIdx=8332&onlyList=N)

5) 본연구에서 바이어스 점수의 절댓값은 0~0.1, 0.1~0.3, 0.3~0.5, 0.5~0.7, 0.7 이상 총 5개 등급[0, 4]으로 나눴다. 등급 수가 높을수록 편향성의 심각도가 계속 올라간다.

6) 이유는 3.3. 결과 분석을 참고.

7) 자세한 설명은 3.3 결과 분석을 참고.

8) 다중 로지스틱 회귀 모형을 쓰는 이유는 각주 5 참고.

9) 본연구에서 사용된 모든 실험 데이터는 https://github.com/feili0820/Gender-Occupation_Bais_in_KorPLM 에서 공개한다.

는 경향이 확실한 것이다. 반대로 Albert 아키텍처의 언어모델인 경우, 한쪽에 기우는 경향도 뚜렷하지 않고 표준편차의 기복도 심하지 않은 것으로 나타났다. 즉 사이즈가 작은 언어모델에서는 성별 편향 수준도 상대적으로 낮을 수 있다는 셈이다.

그리고 'klroberta' 계열과 'lasbert' 계열의 언어모델들은 계열 내부에서 학습 방식에 비슷한 면이 많다는 점과 함께 성별 편향 수준도 아주 선명하다는 공통점을 공유하고 있다. 이점은 성별 편향 형상을 조성하는데 언어모델의 파라미터 규모 외에 학습 방식의 영향도 있음을 보여준다.

아울러 아키텍처가 다르지만 트레이닝 데이터셋의 구성이 고도로 유사한 'klbert_b'와 'klroberta_b' 언어모델은 성별 편향 점수의 분포도 서로 많이 비슷한 편이다. 이점을 통해 언어모델의 성별 편향 분포를 결정하는 것에 트레이닝 데이터 소스가 중요한 역할을 발휘함을 엿볼 수 있다.

표 3 1차 'Fill-Mask' 실험 결과

| | | statistics | kcbert_b | klbert_b | klroberta_b | klroberta_l | klroberta_s | kyalbert | lasbert_b | lasbert_s | krelectra | krfinbert |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 직종 | **남자-여자:** | Min. | -0.476 | -0.250 | -0.293 | 0.382 | 0.018 | -0.066 | -0.013 | 0.104 | -0.395 | -0.649 |
| | N=10, | Max. | 0.606 | -0.085 | -0.061 | 0.640 | 0.314 | 0.014 | 0.243 | 0.181 | -0.059 | -0.398 |
| | Chi-S.=25.44 | Mean | 0.084 | -0.145 | -0.134 | 0.511 | 0.215 | -0.020 | 0.154 | 0.138 | -0.214 | -0.536 |
| | Kendall-W=0.283*** | StDev. | 0.326 | 0.051 | 0.072 | 0.072 | 0.090 | 0.034 | 0.084 | 0.031 | 0.102 | 0.076 |
| | **남성-여성:** | Min. | -0.476 | -0.250 | -0.293 | 0.382 | 0.018 | -0.066 | -0.013 | 0.104 | -0.395 | -0.649 |
| | N=10, | Max. | 0.606 | -0.085 | -0.061 | 0.640 | 0.314 | 0.014 | 0.243 | 0.181 | -0.059 | -0.398 |
| | Chi-S.=24.44 | Mean | 0.084 | -0.145 | -0.134 | 0.511 | 0.215 | -0.020 | 0.154 | 0.138 | -0.214 | -0.536 |
| | Kendall-W=0.272*** | StDev. | 0.326 | 0.051 | 0.072 | 0.072 | 0.090 | 0.034 | 0.084 | 0.031 | 0.102 | 0.076 |
| | **남-여:** | Min. | -0.476 | -0.250 | -0.293 | 0.382 | 0.018 | -0.066 | -0.013 | 0.104 | -0.395 | -0.649 |
| | N=10, | Max. | 0.606 | -0.085 | -0.061 | 0.640 | 0.314 | 0.014 | 0.243 | 0.181 | -0.059 | -0.398 |
| | Chi-S.=10.36 | Mean | 0.084 | -0.145 | -0.134 | 0.511 | 0.215 | -0.020 | 0.154 | 0.138 | -0.214 | -0.536 |
| | Kendall-W=0.115 | StDev. | 0.326 | 0.051 | 0.072 | 0.072 | 0.090 | 0.034 | 0.084 | 0.031 | 0.102 | 0.076 |
| 직장 | **남자-여자:** | Min. | 0.009 | -0.311 | -0.206 | 0.450 | 0.009 | -0.082 | 0.096 | 0.070 | -0.340 | -0.745 |
| | N=10, | Max. | 0.649 | 0.219 | 0.044 | 0.711 | 0.315 | 0.010 | 0.358 | 0.185 | -0.148 | -0.410 |
| | Chi-S.=27.03 | Mean | 0.431 | -0.096 | -0.130 | 0.530 | 0.126 | -0.035 | 0.206 | 0.127 | -0.210 | -0.574 |
| | Kendall-W=0.300*** | StDev. | 0.181 | 0.171 | 0.086 | 0.072 | 0.092 | 0.028 | 0.071 | 0.034 | 0.053 | 0.103 |
| | **남성-여성:** | Min. | -0.504 | -0.430 | -0.294 | 0.128 | -0.519 | -0.078 | -0.034 | 0.042 | -0.363 | -0.900 |
| | N=10, | Max. | -0.067 | 0.123 | -0.012 | 0.525 | -0.179 | 0.079 | 0.110 | 0.254 | -0.028 | -0.617 |
| | Chi-S.=27.45 | Mean | -0.319 | -0.249 | -0.167 | 0.220 | -0.455 | 0.001 | 0.041 | 0.170 | -0.135 | -0.794 |
| | Kendall-W=0.305*** | StDev. | 0.124 | 0.171 | 0.094 | 0.117 | 0.100 | 0.060 | 0.049 | 0.057 | 0.093 | 0.093 |
| | **남-여:** | Min. | -0.504 | -0.430 | -0.294 | 0.128 | -0.519 | -0.078 | -0.034 | 0.042 | -0.363 | -0.900 |
| | N=10, | Max. | -0.067 | 0.123 | -0.012 | 0.525 | -0.179 | 0.079 | 0.110 | 0.254 | -0.028 | -0.617 |
| | Chi-S.=9.58 | Mean | -0.319 | -0.249 | -0.167 | 0.220 | -0.455 | 0.001 | 0.041 | 0.170 | -0.135 | -0.794 |
| | Kendall-W=0.106 | StDev. | 0.124 | 0.171 | 0.094 | 0.117 | 0.100 | 0.060 | 0.049 | 0.057 | 0.093 | 0.093 |
| 직급 (승직) | **남자-여자:** | Min. | -0.222 | -0.215 | -0.217 | 0.429 | -0.030 | -0.075 | 0.027 | 0.168 | -0.398 | -0.552 |
| | N=10, | Max. | 0.279 | -0.046 | -0.040 | 0.626 | 0.270 | 0.025 | 0.315 | 0.303 | 0.104 | -0.293 |
| | Chi-S.=18.55 | Mean | 0.102 | -0.121 | -0.126 | 0.545 | 0.080 | -0.030 | 0.192 | 0.248 | -0.154 | -0.393 |
| | Kendall-W=0.206** | StDev. | 0.166 | 0.048 | 0.056 | 0.063 | 0.095 | 0.043 | 0.100 | 0.046 | 0.173 | 0.069 |
| | **남성-여성:** | Min. | -0.674 | -0.086 | -0.159 | 0.039 | -0.555 | -0.029 | -0.329 | 0.363 | -0.611 | -0.902 |
| | N=10, | Max. | -0.396 | 0.187 | 0.084 | 0.211 | -0.366 | 0.135 | -0.031 | 0.492 | 0.017 | -0.762 |
| | Chi-S.=16.36 | Mean | -0.518 | 0.033 | -0.023 | 0.149 | -0.477 | 0.062 | -0.158 | 0.421 | -0.342 | -0.839 |
| | Kendall-W=0.182* | StDev. | 0.089 | 0.092 | 0.071 | 0.058 | 0.060 | 0.064 | 0.090 | 0.044 | 0.214 | 0.046 |
| | **남-여:** | Min. | 0.823 | 0.091 | -0.489 | 0.033 | -0.338 | -0.008 | 0.765 | 0.551 | -0.895 | 0.837 |
| | N=10, | Max. | 0.942 | 0.427 | -0.221 | 0.319 | -0.137 | 0.128 | 0.924 | 0.760 | -0.253 | 0.915 |
| | Chi-S.=9.95 | Mean | 0.901 | 0.240 | -0.313 | 0.188 | -0.222 | 0.075 | 0.889 | 0.690 | -0.593 | 0.868 |
| | Kendall-W=0.111 | StDev | 0.034 | 0.090 | 0.081 | 0.082 | 0.075 | 0.052 | 0.046 | 0.070 | 0.221 | 0.028 |
| 직급 (강직) | **남자-여자:** | Min. | -0.181 | -0.161 | -0.167 | 0.501 | -0.016 | -0.056 | 0.064 | 0.100 | -0.373 | -0.507 |
| | N=10, | Max. | 0.302 | -0.026 | -0.020 | 0.683 | 0.282 | 0.055 | 0.335 | 0.256 | 0.129 | -0.252 |
| | Chi-S.=16.47 | Mean | 0.144 | -0.074 | -0.097 | 0.613 | 0.091 | -0.005 | 0.216 | 0.176 | -0.126 | -0.345 |
| | Kendall-W=0.183* | StDev. | 0.155 | 0.039 | 0.047 | 0.057 | 0.095 | 0.044 | 0.094 | 0.054 | 0.170 | 0.068 |
| | **남성-여성:** | Min. | -0.676 | -0.087 | -0.043 | 0.153 | -0.531 | -0.074 | -0.256 | 0.066 | -0.577 | -0.899 |
| | N=10, | Max. | -0.400 | 0.144 | 0.136 | 0.338 | -0.334 | 0.096 | 0.041 | 0.352 | 0.054 | -0.747 |
| | Chi-S.=17.48 | Mean | -0.524 | 0.010 | 0.055 | 0.264 | -0.453 | 0.022 | -0.092 | 0.203 | -0.299 | -0.830 |
| | Kendall-W=0.194* | StDev. | 0.088 | 0.083 | 0.063 | 0.060 | 0.061 | 0.071 | 0.088 | 0.082 | 0.216 | 0.050 |
| | **남-여:** | Min. | 0.636 | 0.620 | 0.629 | 0.622 | 0.627 | 0.637 | 0.651 | 0.627 | 0.620 | 0.628 |
| | N=10, | Max. | -0.941 | -0.954 | -0.926 | -0.894 | -0.959 | -0.928 | -0.960 | -0.938 | -0.931 | -0.908 |
| | Chi-S.=9.58 | Mean | 0.950 | 0.927 | 0.886 | 0.906 | 0.913 | 0.927 | 0.920 | 0.942 | 0.918 | 0.933 |
| | Kendall-W=0.106 | StDev. | 0.087 | -0.015 | 0.112 | 0.117 | -0.048 | 0.066 | -0.006 | 0.173 | 0.032 | 0.054 |
| | N:샘플수, Chi-S.:카이제곱, Kendall-W:일치도 계수; Min.:최소값, Max.:최대값, Mean:평균값, StDev.:표준편차; *:P<0.1, **:P<0.05, ***:P<0.01 | | | | | | | | | | | |

표 4 2차 'Fill-Mask' 실험 결과(이중 로지스틱 회귀)

| | factor | β | S.E. | Wald | Df | Sig. | OR |
|---|---|---|---|---|---|---|---|
| **klbert_b**: | **연령(노인)** | **-1.350** | **0.630** | **4.592** | **1** | **0.032** | 0.259 |
| Hosmer-Lemeshow | 출신(지방) | 0.000 | 0.577 | 0.000 | 1 | 1.000 | 1.000 |
| Sig. = 1.0 | 학력(비대졸) | 0.000 | 0.577 | 0.000 | 1 | 1.000 | 1.000 |
| -2 log likelihood = 74.876 | 소득(저) | -19.843 | 4382.521 | 0.000 | 1 | 0.996 | 0.000 |
| | 상수 | 22.040 | 4382.521 | 0.000 | 1 | 0.996 | 3.732E+09 |
| **kcbert_b**: | 연령(노인) | -21.678 | 4385.029 | 0.000 | 1 | 0.996 | 0.000 |
| Hosmer-Lemeshow | **출신(지방)** | **1.076** | **0.477** | **5.093** | **1** | **0.024** | 2.932 |
| Sig. = 0.999 | 학력(비대졸) | 0.000 | 0.473 | 0.000 | 1 | 1.000 | 1.000 |
| -2 log likelihood = 101.942 | 소득(저) | -0.663 | 0.476 | 1.945 | 1 | 0.163 | 0.515 |
| | 상수 | 0.125 | 0.461 | 0.074 | 1 | 0.786 | 1.134 |

factor: 독립변수; β: 회귀계수; S.E.:표준 오차; Wald:월드테스트 점수; Df:자유도; Sig.:유의수준; OR:승산 비율
Hosmer-Lemeshow: HLGOF 적합도; -2log likelihood: -2로그우도값

표 5 2차 'Fill-Mask' 실험 결과(다중 로지스틱 회귀)

| | factor | Est. | S. E. | Wald | Df | Sig. |
|---|---|---|---|---|---|---|
| **lasbert_b** | **연령(노인)** | **2.602** | **0.771** | **11.401** | **1** | **0.001** |
| parallel-lines test | **출신(지방)** | **-1.783** | **0.594** | **9.009** | **1** | **0.003** |
| Sig. = 0.602 | 학력(비대졸) | 0.000 | 0.490 | 0.000 | 1 | 1.000 |
| -2 log likelihood = 32.33 | 소득(저) | 0.720 | 0.503 | 2.043 | 1 | 0.153 |
| **krfinbert** | **연령(노인)** | **-3.262** | **0.474** | **47.306** | **1** | **0.000** |
| parallel-lines test | 출신(지방) | 0.810 | 0.418 | 3.760 | 1 | 0.052 |
| Sig. = 0.638 | 학력(비대졸) | 0.000 | 0.407 | 0.000 | 1 | 1.000 |
| -2 log likelihood = 42.74 | **소득(저)** | **-1.484** | **0.451** | **10.831** | **1** | **0.001** |

이 외에 템플릿 별로 평균값 분포를 봤을 때, 직장과 직종 템플릿에서는 각 언어모델이 남성 어휘와 여성 어휘를 선택할 확률은 균형적인 편인데 직급에 관련한 2개 템플릿에서는 많은 언어모델이 여성 어휘보다 남성 어휘를 더 자주 선택하는 경향이 있다. 이 점은 한국 사회에서 존재하는 '권력'과 '남자' 간의 강력한 연계성이 언어모델에 그대로 학습되었음을 암시한다.

이상의 기술 통계 분석 결과를 다시 요약하자면,

1. 성별 편향 점수의 분포로 봤을 때, 현재 트랜스포머 기반 한국어 언어모델 중에 성별-극성(성별 편향 점수의 부호)에 민감한 유형과 성별-강도(성별 편향 점수의 절댓값)에 민감한 유형 두 가지가 존재한다.

2. 언어모델의 내부적 속성이 성별 편향 조성에 크고 작은 영향을 미칠 수 있는데 그중에 트레이닝 데이터 소스의 영향력이 상대적으로 명확하고 파라미터 사이즈와 학습 방식도 중요한 영향 요인으로 추정된다.

3. 언어모델이 성별 편향을 학습할 때 주어진 텍스트의 내용, 구체적으로는 포함된 어휘의 부류에 따라 학습 효과가 달라진다. 예컨대 많은 언어모델은 직장이나 직종에 관한 어휘보다 직급에 관한 어휘에 '남녀유별'을 의식하는 경향이 더욱 확실한 편이다. 즉 한국어 텍스트에서 직급 관련 어휘는 언어모델의 성별 편향(극성)을 더욱 쉽게 유발할 수 있는 셈이다.

이어서 추론 통계의 결과를 보면 Kendall-W값은 (남자-여자)와 (남성-여성)에서 대부분 유의미한 결과를 보인 것과 달리, (남-여)에서 대부분 무의미한 결과를 보였다. 이 현상은 '남'과 '여' 두 어휘에 존재하는 의미적 중의성과 무관해 보이지 않는다. 구체적으로 '남'과 '여'는 성별을 구분하는 대립적인 개념쌍으로 사용되기도 하지만 성별과 무관한 다른 의미로 쓰일 경우도 있기 때문이다. 따라서 현실 세계에서 사람들이 자주 '남' 혹은 '여'로 성별을 명시하는 습관과 달리, 언어모델의 성별 구분 기준에 오히려 (남자-여자) 혹은 (남성-여성)의 어휘쌍이 더욱 적합할 수 있다.

그리고 각 언어모델의 성별 예측 결과 간의 일치도 테스트에서 재미있는 현상이 더 있었다. (남자-여자)와 (남성-여성) 조합의 Kendall-W값은 템플릿을 막론하고 유의미한 수준으로 나왔는데 그 점수는 대부분 0.3보다 낮은 것으로 드러났다.[10] 이점은 직업의 내재적 속성에 관한 모든 템플릿에서 언어모델들의 성별 구분 기준은 서로 유사하지 않음을 알 수 있다. 다른 말로 언어모델마다 성별 편향은 보편적으로 존재하나 언어모델마다 개별적인 특징도 없지 않다.

이상의 추론 통계 분석 결과를 요약하자면,

1. 한국어 언어모델에서 성별 편향을 측정할 때 일상에서 자주 쓰이는 (남-여) 척도보다 (남자-여자)와 (남성-여성) 척도가 훨씬 더 적합할 수 있다.

2. 한국어 언어모델의 성별 편향 문제는 거시적인 보편성과 미시적인 복잡성을 동시에 지닌다. 이를 해결하기 위해 이 두 가지 측면을 모두 고려할 필요가 있다.

### 3.3.2 2차 실험

1차 실험 결과를 통해 현재 한국어 언어모델은 거시적으로 '극성 민감 유형'과 '강도 민감 유형' 두 가지로 분류할 수 있으나 미시적으로 같은 유형에 속한 언어모델 간에 내집단 차이도 있음을 알 수 있었다. 2차 실험은 이런 언어모델의 개별적인 특징에 대해 집중적으로 분석하고자 한다.

전에 이미 소개한 2차 실험의 기본 템플릿(TEMPLATE2)을 도식화하면 다음 그림 1과 같다. 그림 1을 보면 모든 변수에 소수의 하위 범주가 더 있음을 쉽게 확인할 수 있다. 따라서 그림 1의 모형에 적합할 회귀분석 방법은 로지스틱 회귀 모형이다. 최종적으로 도출한 종속변수의 카테고리 개수를 보면 극성에는 2개, 강도에는 5개가 있다. 그래서 극성에 민감한 언어

[10] 보통 Kendall-W값은 0.6보다 높으면 일치도가 높은 수준, 0.2보다 낮으면 일치도가 낮은 수준이라고 본다.

모델은 이중 로지스틱 회귀 분석 모형을, 강도에 민감한 언어모델은 다중 로지스틱 회귀분석 모형을 각각 적용해 분석할 필요가 있다.

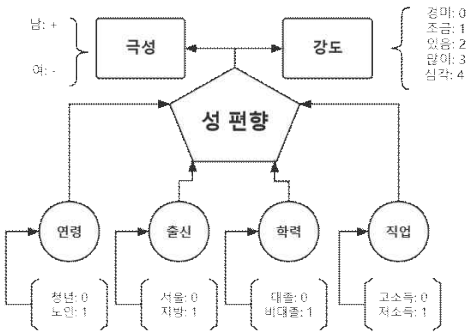


그림 1 성별 편향 평가 로지스틱 회귀 모형

그림 1의 모형으로 각 언어모델의 샘플을 테스트하기 전에 로지스틱 회귀분석의 기본 적용 조건[11]에 충족할지를 먼저 확인해야 한다. 실제로 검증한 결과, 로지스틱 회귀분석에 적합한 언어모델은 많지 않은 편이며, 사전 검증을 통과한 4개 언어모델의 회귀분석 결과는 표 4와 표 5에 기록했다.

로지스틱 회귀분석 모형을 적용할 수 있는 언어모델의 총수가 적지만 언어모델들의 개별적인 특징을 확실히 파악할 수 있다. 우선 모델의 사이즈나 아키텍처 그리고 사전훈련 태스크 유형이 유사한 'klbert_b'와 'kcbert_b'는 성별 편향의 극성 속성에 민감하다는 공통점이 있지만, 전자는 연령 정보가 담긴 어휘를 더욱 중요시하고 후자는 출신 정보가 담긴 어휘에 관심을 더 기울이는 편이다. 이런 차이를 보이는 이유는 한국사회에서 공유된 일반적인 가치관[12]과 연관성이 있다고 추정된다.

한편, 다중 회귀 분석 모형을 도입한 'lasbert_b와 'krfinbert' 언어모델은 공통적으로 강조하는 연령 정보 이외에, 전자는 출신 정보에, 후자는 직업(소득) 정보에 각각 치중하는 경향이 있다.[13] 이런 차이는 두 언어모델의 트레이닝 데이터 소스가 다르다는 점과 연계해 고찰할 필요가 있다. 'lasbert_b'는 주로 <위키백과>와 <모두의 말뭉치>의 텍스트 자료를 위주로 사용했지만, 'krfinbert'는 일반적인 텍스트 자료로 트레이닝한 'Krbert'[14]를 기반으로 금융 도메인 코퍼스 자료를 추가로 투입해 새로 fine-tuning한 모델이다. 따라서 금융 도메인의 텍스트에 더욱 익숙한 'krfinbert'는 소득 개념에 관한 어휘에 대해, 이해도가 한층 더 높을 수 있다고 짐작된다. 이런 관계로 lasbert_b'의 회귀 모형에서 잡히지 않은 직업(소득) 변수는 'krfinbert'의 회귀 모형에서 오히려 유의미한 결과가 나타날 수 있는 것이다.

로지스틱 회귀분석의 결과를 다시 정리하자면,

1. 거시적인 면에서 한국어 언어모델은 성별 편향의 극성과 강도로 나눌 수 있는데, 미시적인 면에서는 같은 분류에 속한 언어모델 간의 차이를 토대로 세분되는 유형을 더 귀납할 수 있다.

2. 한국어 언어모델은 직업에 관한 네 가지 외연적 요소 중에 어떤 변수를 더욱 중요시하는지에 따라, 성별 편향의 극성/강도를 표출하는 방식이 달라질 수 있다. 위 네 가지 요소의 중요도를 결정하는 데 트레이닝 데이터 소스의 구성 및 규모가 중대한 영향을 미칠 수 있다. 특히 특정 도메인의 데이터를 추가로 학습시킬 경우, 뜻밖의 성별 편향 요인을 초래할 가능성이 있다.

## 4. 결론

본연구는 언어모델 트레이닝 과정에서 많이 활용해온 'Fill-Mask' 태스크와 직업 속성에 기반한 성별 편향 탐지용 템플릿을 결합해 한국어 언어모델에서 이미 널리 퍼진 성별 편향 현상의 거시·미시적 분포 특징을 자세히 살펴봤다. 이를 토대로 앞으로 한국어 언어모델의 성별 편향 문제를 해소하는 방향에 대해, 다음과 같은 세 가지 의견을 제시한다.

1. 트레이닝 데이터 소스의 재정비: 남녀 속성을 구별하지 못하는 성 중립적인 텍스트 환경보다 자발적 데이터 검증으로 남녀 어휘 비율이 맞춰져 있는 성 균형적 텍스트 환경을 추구한다.

2. 성별 편향 평가 기준의 확장적 해석: 언어모델의 성별 편향을 측정할 때 극성과 강도는 두 가지 기본 척도로 선정하되, 구체적인 NLP 태스크 환경에 따라, 성별 편향 제거의 기준과 방식을 유연하게 적용한다.

3. 성별 편향 현상에 대한 포괄적 이해 및 접근: 성별 편향 현상은 언어모델의 공통적인 문제이며 이를 단순히 성차별로 취급할 수 없다. 이 문제를 해결하기 위해 NLP 분야뿐만 아니라 언어학, 심리학, 사회학, 의학, 생물학 등 다양한 접근 시각이 필요하다.

## 참고문헌

[1] J. Y. Mun, "The impact of Confucianism on gender equality in Asia," The Georgetown Journal of Gender and the Law, vol. 16, (3), pp. 633, 2015.

[2] F. T. Asr et al, "The gender gap tracker: Using natural language processing to measure gender bias in media," PloS One, vol. 16, (1), pp. e0245533-e0245533, 2021.

[3] R. Bansal, "A Survey on Bias and Fairness in Natural Language Processing," arXiv preprint arXiv:2204.09591, 2022.

[4] T. Wolf et al, "HuggingFace's Transformers:

11) 이중 로지스틱 회귀에서는 HLGOF 값이 0.05보다 높아야 하고, 다중 로지스틱 회귀에서는 parallel-lines test 값이 0.5를 초과해야 한다.

12) 여기서 주로 한국사회의 '장유유서' 특징을 가리킨다.

13) 'krfinbert'의 회귀 모형에서 출신 변수의 유의수준은 0.052에 달했기에 넓은 기준으로 봤을 때 'krfinbert'도 출신 정보에 관심이 많다고 볼 수 있다.

14) https://huggingface.co/snunlp/KR-BERT-char16424

State-of-the-art Natural Language Processing," arXiv preprint arXiv:1910.03771, 2019.
[5] R. P. Larrick, "Debiasing," Blackwell handbook of judgment and decision making, pp.316-338, 2004.
[6] T. Bolukbasi et al, "Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings," Advances in neural information processing systems, 2016.
[7] J. Zhao et al, "Gender Bias in Contextualized Word Embeddings," arXiv preprint arXiv:1904.03310, 2019.
[8] M. Brunet et al, "Understanding the Origins of Bias in Word Embeddings," International conference on machine learning, 2018.
[9] A. C. Kozlowski, M. Taddy and J. A. Evans, "The Geometry of Culture: Analyzing the Meanings of Class through Word Embeddings," American Sociological Review, vol. 84, (5), pp. 905-949, 2019.
[10] B. Ghai, M. N. Hoque and K. Mueller, "WordBias: An Interactive Visual Tool for Discovering Intersectional Biases Encoded in Word Embeddings," 2021.
[11] M. El-Assady et al, "Semantic Concept Spaces: Guided Topic Model Refinement using Word-Embedding Projections," IEEE Transactions on Visualization and Computer Graphics, vol. 26, (1), pp. 1001-1011, 2020;2019;.
[12] G. Grand et al, "Semantic projection recovers rich human knowledge of multiple object features from word embeddings," Nature Human Behaviour, 2022.
[13] A. H. Bailey, A. Williams and A. Cimpian, "Based on billions of words on the internet, people = men," Science Advances, vol. 8, (13), pp. eabm2463-eabm2463, 2022.
[14] S. Bordia and S. R. Bowman, "Identifying and Reducing Gender Bias in Word-Level Language Models," arXiv preprint arXiv:1904.03035, 2019.
[15] R. Bhardwaj, N. Majumder and S. Poria, "Investigating Gender Bias in BERT," Cognitive Computation, vol. 13, (4), pp. 1008-1018, 2021.
[16] J. Zhou et al, "Towards Identifying Social Bias in Dialog Systems: Frame, Datasets, and Benchmarks," arXiv:2202.08011, 2022.
[17] J. Zhao et al, "Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods," arXiv preprint arXiv:1804.06876, 2018.
[18] K. Webster et al, "Mind the GAP: A Balanced Corpus of Gendered Ambiguous Pronouns," Transactions of the Association for Computational Linguistics, vol. 6, pp. 605-617, 2018.
[19] N. Sevim, F. Sahinuc and A. Koc, "Gender bias in legal corpora and debiasing it," Natural Language Engineering, pp. 1-34, 2022.
[20] A. V. Nadimpalli and A. Rattani, "GBDF: Gender Balanced DeepFake Dataset Towards Fair DeepFake Detection," preprint arXiv:2207.10246, 2022.
[21] M. Nadeem, A. Bethke and S. Reddy, "StereoSet: Measuring stereotypical bias in pretrained language models," 2020.
[22] Dhamala, Jwala, et al. "Bold: Dataset and metrics for measuring biases in open-ended language generation." Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. 2021.
[23] P. Delobelle et al, "Measuring Fairness with Biased Rulers: A Survey on Quantifying Biases in Pretrained Language Models," 2021.
[24] P. P. Liang et al, "Towards Understanding and Mitigating Social Biases in Language Models," International Conference on Machine Learning, 2021.
[25] K. A. Lassonde and E. J. O'Brien, "Occupational stereotypes: activation of male bias in a gender-neutral world," Journal of Applied Social Psychology, vol. 43, (2), pp. 387-396, 2013.
[26] H. Kirk et al, "Bias Out-of-the-Box: An Empirical Analysis of Intersectional Occupational Biases in Popular Generative Language Models," Advances in neural information processing systems, 2021.
[27] A. Lauscher, A. Crowley and D. Hovy, "Welcome to the Modern World of Pronouns: Identity-Inclusive Natural Language Processing beyond Gender," 2022.
[28] 정원섭, "인공지능 알고리즘의 편향성과 공정성," 인간.환경.미래, (25), pp. 55-73, 2020.
[29] 정현규 et al, "인공지능 분야의 편향 연구에 관한 사회적 요인의 영향," 한국정보과학회 학술발표논문집, vol. 2020, (12), pp. 1460-1462, 2020.
[30] 김효은, "인공지능 편향식별의 공정성 기준과 완화," 한국심리학회지:일반, 40(4), pp. 459-485, 2021.
[31] 이수연, 조성준, "Transformer 기반의 한국어 언어 모델의 사회적 편향 분석," 한국경영과학회 학술대회논문집, vol. 2022, (6), pp. 3623-3631, 2022.