**졸업프로젝트 최종 보고서**

**한국어 기반 pre-trained language model에 대한 bias 분석**

전기정보공학부

2019-12684 박상권

**1. 소개**

최근 자연어 처리에서 social bias 문제의 중요성이 강조되고 있다. social bias 문제는 인공지능 전 분야에서 나타나는 문제로, 인공지능이 기존 사회의 사회적 편향을 반영하는 문제를 뜻한다. 자연어 처리에서의 social bias의 대표적인 예시는, 번역기에서의 social bias로, 아래 그림과 같이 “He is a nurse”, “She is a doctor”를 성별 대명사가 없는 터키어로 번역 후, 다시 영어로 번역하였을 때, “She is a nurse”, “He is a doctor”로 원본과 반대로 번역이 된다. 이는 의사는 남자가, 간호사는 여자가 한다는 사회적 편향이 반영이 된 것으로 해석할 수 있다.



이러한 자연어 처리에서의 social bias 연구는 매우 활발히 이루어지지만, 대부분이 영어 및 영어 모델에 대해서만 이루어지며, 이외의 언어들에 대해서는 연구가 거의 진행되지 않는다. 특히 한국어 및 한국어 모델에 대해서는 이러한 연구가 전무하다. 따라서 본 프로젝트에서는 기존의 영어 모델에 대해 제안되었던 debiasing method와, bias evaluation metric을 한국어 모델에 맞게 수정하여 성별 bias에 대해 적용하여 보고, 그 결과를 관찰한다.

**2. 선행 연구 및 한국어에서의 구현**

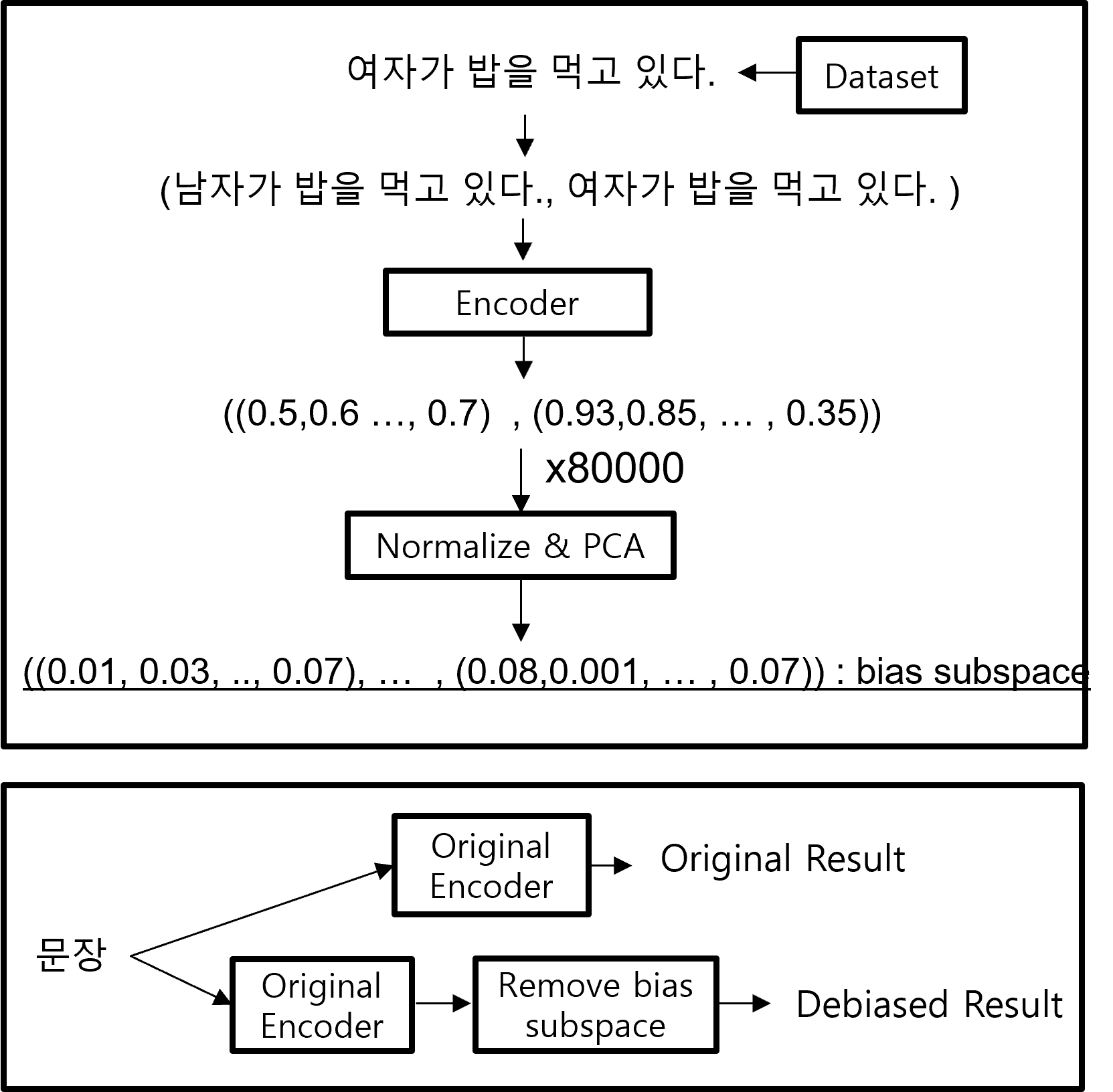
(1) SENT-DEBIAS

SENT-DEBIAS[1]는 2020년 발표된 debiasing method로, pretrain된 언어 모델의 출력을 후처리하는 post-hoc debias method이다. 문장을 하나의 숫자 벡터로 인코딩하는 언어 모델을 sentence encoder라고 하는데, SENT-DEBIAS는 이런 sentence encoder 모델에 대해 debiasing 하는 방법이다.

SENT-DEBIAS 알고리즘은 데이터셋을 활용하여 bias의 성분을 뽑아낸 후, 기존 sentence encoder의 결과에서 해당 bias 성분을 제거하는 두 과정으로 이루어진다.

데이터셋에서 여자, 남자, 아빠, 엄마 등과 같이 성별을 나타내는 단어를 포함하는 문장들을 모두 선별한다. 그 후, 해당 문장의 성별을 나타내는 단어로 반대 성별로 치환한 문장을 만들어, (여자 문장, 남자 문장)의 순서쌍을 만든다. 그 후, debias하고자 하는 encoder에 두 문장을 대입하여 각각의 출력 v1, v2를 얻는다. 그리고, 둘의 평균 성분을 각각에서 뺀 벡터(v2-(v1+v2)/2)를 정규화하여 벡터 v3를 얻는다. 이러한 과정을 반복하여, PCA 알고리즘으로 주성분을 분석한다. 그러면 bias 성분을 나타내는 bias subspace의 기저 벡터들을 얻게 된다.

그 후, debias된 결과를 얻기 위하여, 문장을 기존 encoder에 대입한 출력을 bias subspace에 대해 사영시킨 성분을 빼서, 해당 bias subspace에 orthogonal한 벡터를 얻는다. 그러면 이 값이 입력으로 넣은 문장의 debias된 출력이 된다.



본 연구에서 제안하는 방법론인 SENT-DEBIAS-KR은 기존의 SENT-DEBIAS[1]를 대부분 그대로 한국어에 적용하여 구현한 것이다. 성별 단어쌍으로는 6쌍을 사용하였으며, bias subspace를 구할 때 사용한 데이터셋은 국립국어원 모두의 말뭉치(https://corpus.korean.go.kr/)의 문어 말뭉치를 사용하였다. 총 79,980개의 문장 순서쌍에서 bias 성분을 구하여 10개의 기저 벡터를 얻었다.

debias에 사용한 언어모델은 세 종류로, KOR-BERT[2], KR-BERT[3], KLUE-BERT[4]를 사용하였다. 이들은 모두 BERT model[5]로, BERT model은 마지막 hidden layer에서 각 토큰 별로 768 차원의 벡터를 매핑한다. 이때 문장의 시작을 뜻하는 [CLS] 토큰의 벡터를 문장의 representation 벡터로 쓸 수 있고, 이들은 sentence encoder의 역할을 한다.

(2) SEAT TEST

SEAT TEST[6]는 2019년에 발표된 metric으로, sentence encoder 형태의 언어 모델이 얼마나 bias되어있는지를 정량화하여 얻는 방법이다.

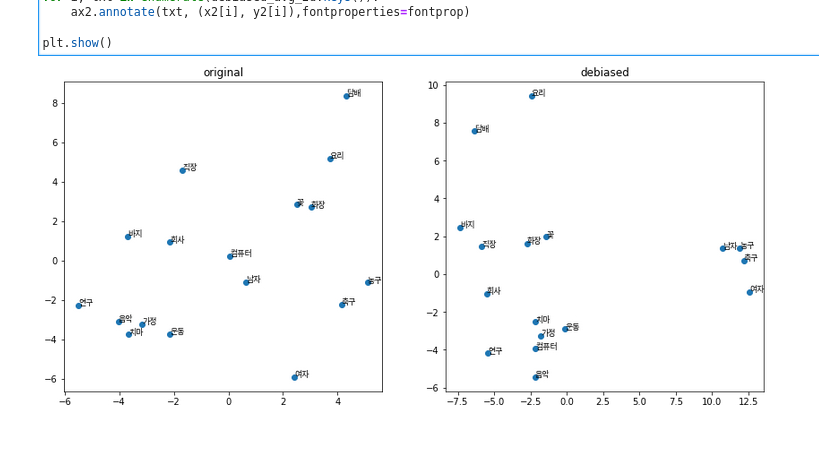
텍스트이(가) 표시된 사진

자동 생성된 설명

SEAT score를 구할 때 sentence encoder에 대해 위 수식을 통하여 하나의 실수 값을 도출한다. 절댓값이 클수록 모델이 많이 편향됨을 뜻한다. 수식의 A, B, X, Y는 저용량 텍스트 집합으로, 각각 10개 남짓의 간단한 문장들로 구성된다. A, B, X, Y를 설정함에 따라 metric이 대변하는 bias의 종류가 바뀌며 이들을 바꾸어 가며 여러 개의 test를 만들 수 있다.

본 프로젝트의 SEAT TEST-KR에서는 X, Y로 각각 가장 많이 사용하는 남자와 여자의 한국 이름들을 사용하는 것으로 고정하였다. 그리고 A, B에 각각 직업/취미/특기와 관련된 stereotype set을 넣어주어 Job Test/Hobby Test/Specialty Test의 3개의 test를 구현하였다.

**3. 정성적 분석**

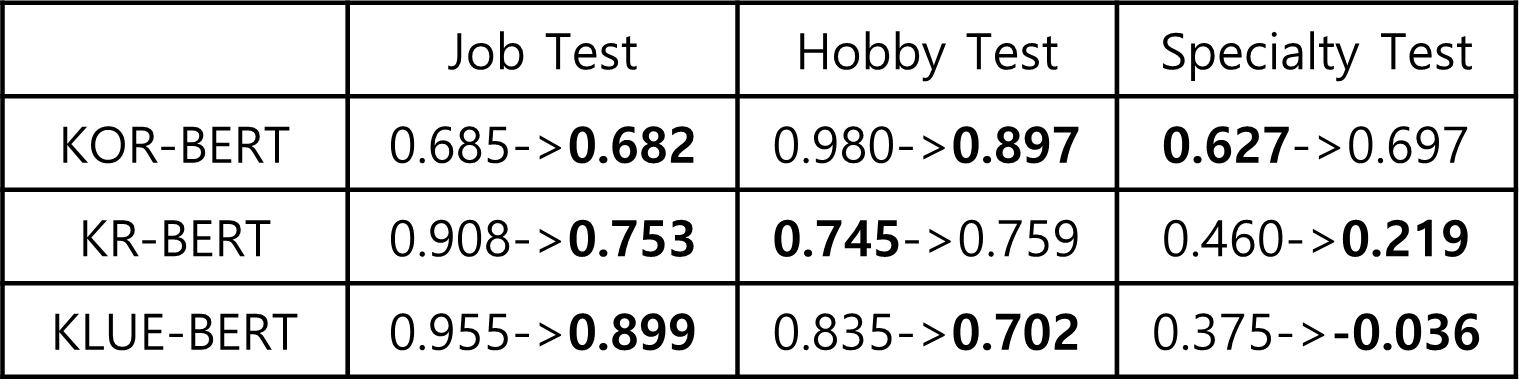


위 그림은 단어들 간의 관계를 나타낸 그림으로, debias 전후 결과를 비교한 것이다. 그림은 기존의 word2vec과 같은 형태로 단어들을 이 차원 벡터로 나타내어 평면에 도시한 것으로, 단어가 가까울수록 이들이 많이 연관되어 있음을 뜻한다. Sentence encoder는 단어 차원에서 벡터로 매핑하지 않으므로, 별개의 대용량 데이터셋에서, 벡터를 얻고자 하는 단어가 들어간 모든 문장들에 대해 encoder로 인코딩 후 평균을 구하여 단어 벡터로 사용한다. 또한 문장 벡터는 768 차원 벡터이므로 t-SNE[7] 알고리즘을 통하여 2차원으로 축소한다.

대용량 데이터셋으로는 aihub의 문서요약 텍스트(https://aihub.or.kr/aidata/8054)를 사용하였다.

결과를 보면, 기존의 그림에서는 “여자”와 “남자”의 단어 벡터가 다른 단어들과 매우 가까운 반면, debias 이후에는 이 벡터들이 다른 단어들과 상대적으로 멀리 떨어지게 된다. 다만 “농구”와 “축구” 또한 함께 멀어지는데, 이는 학습 데이터 양이 부족하여 발생하는 것으로 추측된다. 이는 원 논문[1]과 매우 흡사한 결과로, “남자”와 “여자” 단어를 멀리 떨어뜨려 다른 단어들과의 상대적인 위치를 비슷하게 만든 것으로 해석할 수 있다.

**4. 정량적 분석**



위 표는 앞서 언급하였듯이 3개의 모델에 대해 3개의 테스트를 만들어 측정한 결과이다. debias 적용 전후의 측정값을 화살표로 구분하였고, 굵은 것으로 절대값이 작은, 즉 더 편향이 적은 모델을 표시하였다. 9개의 경우 중 두 개를 제외한 대부분의 경우에서 모두 debias 이후 편향이 더 적어짐을 확인할 수 있다.

**5. 결론**

본 프로젝트를 통하여 한국어 모델에도 영어 모델과 같이 bias가 존재함과, 기존의 영어 모델에서 제안된 biasing method가 한국어 모델에도 적용됨을 최초로 확인하였다. 또한 한국어 모델에 대한 bias evaluation metric을 최초로 구현하고 적용하였다.

하지만 여러 한계점이 있었는데, 우선 debias의 효과가 뚜렷하지만 원본 논문보다는 적으며, 세밀한구현 방식과 기타 hyperparameter에 대해 매우 sensitive 하였다. 또한 seat test 같은 경우 저용량 텍스트(A, B, X, Y)의 구성에 대해 매우 sensitive하였다.

또한 한국어와 영어의 차이 때문에 구현에 어려움이 있었다. 예를 들어, 한국어에는 영어에는 없는 조사가 존재하며 언어 모델들마다 이들을 처리하는 방식들이 다르기 때문에 데이터를 처리하는 데에 어려움이 있었다. SEAT TEST의 text sensitivity 또한 한국어의 문장 특성 때문에 더 커진 것으로 생각된다.

SENT-DEBIAS 기법[2]은 여러 대용량 텍스트를 활용하는데, 한국어 데이터가 부족하여 이 데이터들의 다양성을 포기하였고, 연산 속도 및 메모리의 한계로 인해 실험 데이터의 크기에 있어 어느 정도 타협이 불가피하였다. 따라서 향후에는 더 다양한 한국어 데이터를 수집 및 정제하여 적용하였을 때 제안한 방법의 효과를 검증하는 것이 필요하다.

또한 원 논문[1]에서는 bias와 관련된 fine-tuning task에 대해 debiasing method 적용 전후 결과를 비교하였는데, 한국어에서는 이러한 bias를 평가할 수 있는 fine-tuning task 또는 sentence encoder 형태로 활용할 수 있는 fine-tuning task가 없어 불가능하였다. 다만 최근 발표된 KLUE 데이터셋[4]을 통해 한국어 언어 모델에 대해서도 다양한 fine-tuning task를 수행할 수 있게 되었다. 따라서 추후에는 이와 같은 데이터셋을 활용하여 bias를 평가할 수 있는 fine-tuning task를 만들고, SENT-DEBIAS-KR 방법론의 효과를 검증하고자 한다.

마지막으로, 본 프로젝트의 구현은 영어권의 debias 기법과 metric을 한국어로 단순히 옮긴 것인데, 한국어의 특수성을 고려한 요소들을 적용시켜 보다 더 발전된 debias 기법과 metric을 개발할 수 있을 것으로 생각된다.

**6. 참고문헌**

[1] Paul Pu Liang, Irene Li, Emily Zheng, Yao Chong Lim, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2020. “Towards Debiasing Sentence Representations.” In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.*

[2] Pretrained Language Models for Korean, <https://github.com/kiyoungkim1/LMkor>

[3] [Sangah Lee](https://arxiv.org/search/cs?searchtype=author&query=Lee%2C+S), [Hansol Jang](https://arxiv.org/search/cs?searchtype=author&query=Jang%2C+H), [Yunmee Baik](https://arxiv.org/search/cs?searchtype=author&query=Baik%2C+Y), [Suzi Park](https://arxiv.org/search/cs?searchtype=author&query=Park%2C+S), [Hyopil Shin](https://arxiv.org/search/cs?searchtype=author&query=Shin%2C+H). 2020. “KR-BERT: A Small-Scale Korean-Specific Language Model” *arXiv preprint arXiv:2008.03979*

[4] Sungjoon Park, Jihyung Moon, Sungdong Kim, WonIk Cho, Jiyoon Han, Jangwon Park, Chisung Song, Junseong Kim, Yongsook Song, Taehwan Oh, et al. 2021. “KLUE: Korean Language Understanding Evaluation.” *arXiv preprint arXiv:2105.09680.*

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova. 2019. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.” *NAACL-HLT (1) 2019: 4171-4186*

[6] Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019. “On Measuring Social Biases in Sentence Encoders.” In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers).*

[7] van der Maaten, Laurens and Hinton, Geoffrey. 2008. "Visualizing Data using t-SNE." *Journal of Machine Learning Research 9 (2008): 2579--2605.*