한국어 기반 pre-trained language model에 대한 bias 분석

과제 수행 학생 : 박상권
지도 교수: 윤성로
담당 조교: 박노일, 송종윤

1. Introduction

- Social Bias in Language Model

· NLP에서 social bias는 매우 중요한 문제지만, 한국어 모델에 대해서는 관련 연구가 전무함

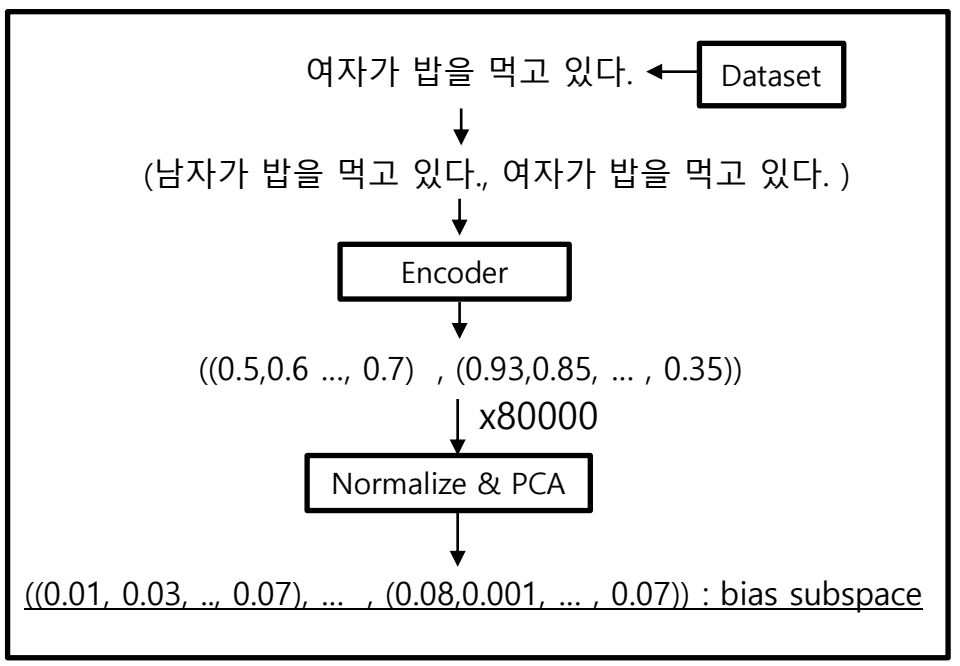
· 본 프로젝트에서는 기존의 영어 모델에 대해 발표된 debiasing 기법과 bias evaluation metric을 한국어 모델에 대해 적용해 보고, 그 결과를 관찰함

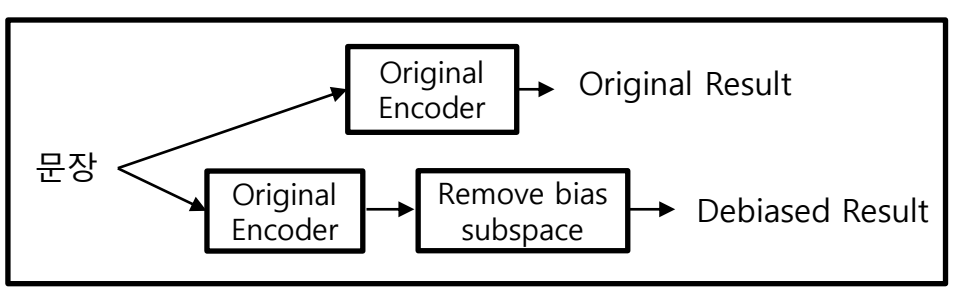
2. Debiasing & SENT-DEBIAS-KR

- Debiasing Method : SENT-DEBIAS (Liang et al., 2020)

· SENT-DEBIAS는 pretrain된 LM의 결과를 후처리하는 post-hoc debias method로, 문장을 숫자 벡터로 인코딩하는 sentence encoder 형태의 LM을 debias하는데 사용될 수 있음

- SENT-DEBIAS-KR



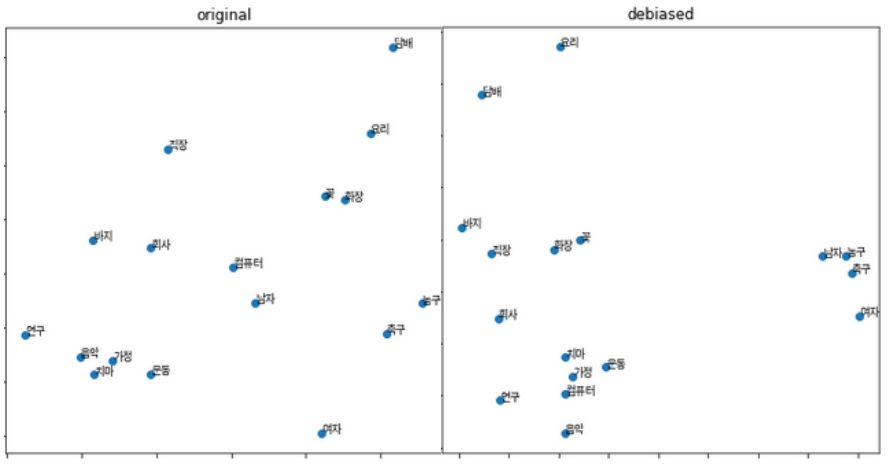


3. Qualitative Approach

· 대용량 text에서 특정 단어들이 포함된 문장 representation의 평균을 계산하여 word vector를 생성

· 해당 벡터들을 t-SNE 알고리즘으로 2D로 압축하여 시각화

- Visualization



· 원본 논문과 동일한 visualization 결과로, 남자와 여자의 단어 벡터가 debias 후 타 단어들에서 멀리 떨어져 상대적 거리가 동일해짐

4. Quantitative Approach

- SEAT Test (May et al., 2019)

$$\frac{\text{mean}_{x \in X} s(x, A, B) - \text{mean}_{y \in Y} s(y, A, B)}{\text{std\_dev}_{w \in X \cup Y} s(w, A, B)},$$

· sentence encoder가 bias되어있는 정도를 측정하는 metric

· 저용량 텍스트 기반이며 영어 implementation만 존재

- SEAT Test Result

	Job Test	Hobby Test	Specialty Test
KOR-BERT	0.685->**0.682**	0.980->**0.897**	**0.627**->0.697
KR-BERT	0.908->**0.753**	**0.745**->0.759	0.460->**0.219**
KLUE-BERT	0.955->**0.899**	0.835->**0.702**	0.375->**-0.036**

· 한국어로 SEAT Test 중 3종류 Test를 implement하여 3개의 한국어 LM(sentence encoder)를 평가

· 대부분의 경우 debias 후 bias 정도가 작아짐을 확인 가능

5. Conclusion & Future Work

- Overall Conclusion

· 한국어 모델에도 bias가 존재함과 기존 biasing method가 한국어 모델에도 적용됨을 최초로 확인

· 효과가 뚜렷하지만 원본 논문보다는 적으며 hyperparameter와 저용량 텍스트에 대해 매우 sensitive함

- Difficulties in Implementation

· 한국어와 영어의 차이

· 한국어 데이터의 부족

· 명확하지 않은 기존 implementation

- Future Work

· 한국어의 특수성을 고려한 발전된 debias 기법 및 metric 개발

· Fine-Tuning Task에 대해 적용 후 비교

6. References

Paul Pu Liang and Irene Mengze Li and Emily Zheng and Yao Chong Lim and Ruslan Salakhutdinov and Louis-Philippe Morency. 2020. Towards Debiasing Sentence Representations

Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders