P1

영어와 한국어에서의 gender bias 비교를 주제로 발표

P4

Tay에 대한 소개 ‘Tay는 트위터로 대화하면서 사용자들로부터 실시간으로 학습하는 방식의 Ai chat bot이었는데, 사용자들로부터 인종차별과 욕설을 배워 결국 하루도 안되 서비스가 중단되었다’ 학습에 기반이 된 빅 데이터가 편향적이기 때문이다.

P5

또한

Assignment 1중에서 bias를 analysis 하는 과정이 있었음

Gender bias를 정량화?

P6

관련 논문 3개 소개

P7

**Word Embeddings and Gender Stereotypes in Swedish and English**

한국어와 영어에서의 gender bias를 비교하자는 아이디어를 직접적으로 얻게 된 논문이다.

이 논문에서는 영어에서 word embedding을 만드는 과정이 스웨덴어에서도 동일하게 적용될 수 있음을 확인하고, bias를 파악하는 방법도 동일하게 적용시킨다.

P8

**Man is to Computer Programmer as Woman is to Homemaker?**

**Debiasing Word Embeddings**

이 논문은 gender stereotype들을 word embedding에서 제거하는 방법을 제시한다.

제거되는 gender stereotype을 수치화하여 비교하면 편향된 정도를 알 수 있지 않을까? 아이디어

.

**P9**

**On Measuring Gender bias in Translation of Gender-neural Pronouns**

자동 변역기에서 번역하는 과정 중 드러나는 gender bias를 연구, 한국어를 직접 연구하기 보다는 번역 과정중 드러나는 bias에 집중

P11

Gender Subspace는 [man is to](2번논문)의 구현을 그대로 따라갔는데, 표에 주어진 단어에 해당하는 벡터를 빼서 gender 벡터쌍 10개를 구한 뒤, 이를 각각의 축으로 하는 공간으로 구현

P12

SVD 이용 차원을 축소

P13

10개의 벡터를 평균을 내서 타겟 벡터가 어디에 더 가까운지 측정

P14

[man is to](2번논문)에서 bias를 제거시킬 떄 그 제거된 값을 비교하여서 편향성 파악

P16

한국어, 영어 위키를 corpus로 이용

P17

이전의 논문들에서는 직업 단어들에만 집중하는 경향, 하지만 예쁘다, 사납다 처럼 남성성/여성성을 나타낼 수 있는 다른 단어들도 사용해야 한다고 생각함(형용사) 추가. 명사40개, 형용사 40개를 사용

P18

결과 분석은 뒤의 파트에서 자세히, 단 scale의 차이가 있을 수 있어, 벡터를 normalize한 형태로 계산함

P20

한국어 Gender bias가 영어에 비해 낮을 것이라 예상

한국에서는 직업의 성 구분 명확x, 유럽쪽은 성 구분이 명확

P21

예상과는 다르게 대체로 대체로 한국어가 Gender bias가 크게 나옴. 특히 approach2에서 형용사 파트가 큰 차이가 나타남

P22

원인 추론

한국어는 영어에 비해 서순이 자유로움, 형용사가 서술어 수식어 기능을 모두 수행-> 수치를 증폭 가설

다만 어느정도 뒤틀렸는지 알 수는 없다

P24

한국어는 영어에 비해 성 편향이 심하다 결론

결론을 긍정한다면 : 페미,pc 의견

결론을 부정한다면 : 수식구조에서 에러가 난 결과이다

P25

형용사 측정 결과가 뒤틀렸을 것이라는 가정 검정 필요 : 다른 여러 모델에 적용시켜볼 필요가 있음

실험 결과를 이용하여 추가적인 인문 사회적 연구가 가능