Gender Bias in Korean Compared with English

컴퓨터학과 2016150145 이성주

컴퓨터학과 2018320105 구재승

Abstract

Concisely describe aims

Index

1. Introduction
2. Related work
3. Approach
4. Experiments
   1. Data
   2. Evaluation Method
   3. Experimental Details
   4. Results
5. Analysis
6. Conclusion
7. References
8. Appendix: Team contributions

1. Introduction

인공지능은 아이와 같다. 처음 학습을 시작할 때에 입력 받는 데이터들을, 스스로 판단하지 않고 그저 받아들인다는 점에서 그렇다. 그 데이터에 오류가 있거나 편견, 혐오가 있다고 하더라도. 2016년 마이크로 소프트사의 chat bot ‘Tay’를 알고 있는가? Tay는 트위터로 대화하면서 사용자들로부터 실시간으로 학습하는 방식의 Ai chat bot이었는데, 사용자들로부터 인종차별과 욕설을 배워 결국 하루도 안되 서비스가 중단되었다.[[1]](#footnote-1) 이는 학습에 기반이 된 빅 데이터가 편향적이기 때문이다.

이 사건을 관점을 바꿔서 바라보자. 어떤 데이터를 학습시켰을 때 인공지능 모델이 편향된다면, 그 결론을 통해 우리는 그 데이터가 얼마나 편향되어 있는지 알 수 있지 않을까? 언어는 사람의 인식을 나타내는 도구이기에, 언어에는 사회적인 통념이 녹아 들 수밖에 없고, 그 속에 내재되어 있는 우리 사회의 편향을 인공지능을 이용해서 측정할 수 있다. 이러한 연구는 자연어 처리 분야의 주된 연구 부분은 아니다. 그러나 자연어 처리를 할 때 이용되는 여러 모델을 이용하여 영어의 편향성을 측정하기 위한 연구는 어느정도 존재했다. 이를 한국어에도 적용시킬 수 있지 않을까?

즉, 자연어 처리 분야에서 인공지능은 공학적으로 활용되는 측면이 강하지만, 이를 이용하여 사회의 일면을 해석하는 작업을 진행하려고 한다. 일반적인 사회 과학의 연구는 현상을 해석하는데 있어서 연구자의 생각과 관점이 어느정도 반영될 수밖에 없지만, 인공지능은 주관이 없이 단순히 계산과 현상의 해석을 할 뿐이기에 보다 객관적인 연구를 진행할 수 있을 것이다.

사회적으로 드러나는 편견은 여러가지 종류가 존재한다. 선행된 영어 연구에서는 인종에 관한 편향 연구, 건강에 대한 편향 연구, responsible innovation에 대한 연구 등 여러 편견들을 다루었다. 우리 팀은 예전부터 널리 다루어져 왔고, 최근에 들어서 더 크게 부각되고 있는 성적 편향을 다룰 예정이다. 한국어에서 편향과 편견이 얼마나 존재하는지 알아보는 것은 분명 의미 있는 연구가 될 것이다.

이번 연구에서 사용할 corpus는 영어와 한국어 모두 위키피디아로 통일한다. 두 corpus를 Word2vec을 이용하여 구성하고. 이를 동일한 차원의 Glove vector로 구현한다.

2. Related work

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

**Rasmus Précenth**

이 논문은 초반부에 Word embedding이 어떻게 구성되었는지, 그리고 그것이 단어 간의 유추를 만들기 위해 수학적으로 무엇을 의미하는지 설명한다. 그 후 영어에서 word embedding을 만드는 과정이 스웨덴어에서도 동일하게 적용될 수 있음을 확인하고, bias를 파악하는 방법도 동일하게 적용시킨다. 그리고 스웨덴어와 영어에서의 bias를 논문에서 제시하는 새로운 방법을 통해 비교한다.

한국어와 영어에서의 gender bias를 비교하자는 아이디어를 직접적으로 얻게 된 논문이다. 다만 스웨덴어와 영어는 서로 유사한 알파벳을 사용하는데 반해 한국어와 영어는 한글과 알파벳이라는 전혀 다른 문자 체계를 사용하고, 문법적인 부분에서도 많은 차이가 존재하여 그 간극을 dataset의 전 처리 과정에서 해결해야 하는 문제가 있었다.

**2) Men Are Elected, Women Are Married: Events Gender Bias on Wikipedia**

**Jiao Sun and Nanyun Peng**

이 논문은 직접적으로 자연어 처리를 연구에 활용하지는 않았다. 위키피디아에서 한 사람에 일생을 담은 페이지들을 모아서, 동일한 사건이 벌어졌을 때 여자에 해당하는 페이지와 남자에 해당하는 페이지가 어떻게 다르게 설명하는지를 비교한 연구이다. 이 연구에서 gender bias를 측정하기 위한 수식을 제시했고, 이를 참고했다.

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

**Debiasing Word Embeddings**

**Tolga Bolukbasi, Kai-Wei Chang, James Zou, Adam Kalai**

이 논문은 gender stereotype들을 word embedding에서 제거하는 방법을 제시한다. 다만 receptionist(접수 담당자)와 같이 단어가 직접적으로 남성과 여성중 한쪽에 치우치지는 않았으나 word embedding이 bias된 경우는 제거하고, queen(여왕)과 같이 명확하게 한쪽으로 치우친 결과가 나와야 하는 경우는 제거하지 않는다.

이 논문에서 사용된 수식을 이용하여, 제거된 수치의 크기가 bias의 정도를 나타낸다는 가정 하에서 특정 단어의 bias의 정도를 측정했다.

1. 한국일보, 전창배 한국인공지능윤리협회 이사장, m.hankookilbo.com/News/Read/A2021091609460000420 [↑](#footnote-ref-1)