第67回日本学生科学賞 最終審査進出研究作品概要

HIT035IT	高校	情報技術	東京都

学校名	東京学芸大学附属国際中等教育学校
研究作品タイトル	データアクイジションにおけるAGIバイアス解決へ
研究者氏名 （共同の場合はグループ）	小谷 理人
指導教諭氏名	河野 真也

【動機】

近年GPTなどの大規模NLPのパブリック使用が増えており、毎日数百万人の方々が使用している。そこで、情報ソースとしてニュースなどを知る際に、もしAIにバイアスがあったら、でそのバイアスってどの様なものなの？というのが先行研究ではただPositive・Negativeでしかやられておらず、具体的にはわかっていない。今後人類の相棒となる上でこれの明確化は必然である。

【方法】

今までの先行研究ではMetricを利用して、グループ分けを通してバイアス度というのを見てきた。それはモデルのバイアス全体図を理解する事にはつながらない。そこでクラウドソーシングを通して哲学・社会学に基づくルーブリックで客観的にバイアスを判断する。

【結果】

２２００以上の方々のリスポンスを元にGPT3.5、GPT4、BARDがやはりバイアスというのはあるという事。バイアスが極力無いニュースのあり方についてなどを知ることができた。またこれらをSDE、CEVというアルゴリズムを通してより客観的に分析にする事でNLPモデルの「性格」の様な事を知ることができた。

【まとめ】

バイアスとはどう足掻いても生成されてしまうものである。例えばNeutralityを求めると情報にRelevanceが無くなったりしてしまうなどと。そこをどうバランスを取って、どのアスペクトを重視するかという事を考える事もできた。また、ネット上にあるデータの原因や、GPT3.5とGPT4がバイアス構成がとても似ていたり、BARDはまた違ったりなどAI学におけて良い研究をする事ができたと考える。

【展望】

この研究は、今後NLP開発を進めるのにおけてどの様にAIの「性格」というものを決めるのかという事だけでなく。やはり、AIバイアスにおける危険性というのがまだ社会に広まっていなく、

今後これらを多く使用していく若者に対しても大きく影響があると考える。