大規模言語モデルを報道のバイアス検出に利用する!

LLM ファミリによるバイアスの検出

マーブ イファット (山岸研究室) エディソン・M・テイラー（東京大学）
セバスチャン・パド（ドイツ・シュトゥットガルト大学） 松尾豊（東京大学）

どんな研究？

- 報道のバイアスや偏向をAIにより自動認識し分析する方法の研究
- 大規模言語モデル（LLM）で採用されているプロンプト技術がメディアのバイアスや偏向を分析するのに有効か調査

何を検討している？

- LLMをどう利用するのがバイアス検出に適しているかを実験し調査
- プロンプトをどう書くか？数例のサンプルは役に立つか？ファインチューニングは？

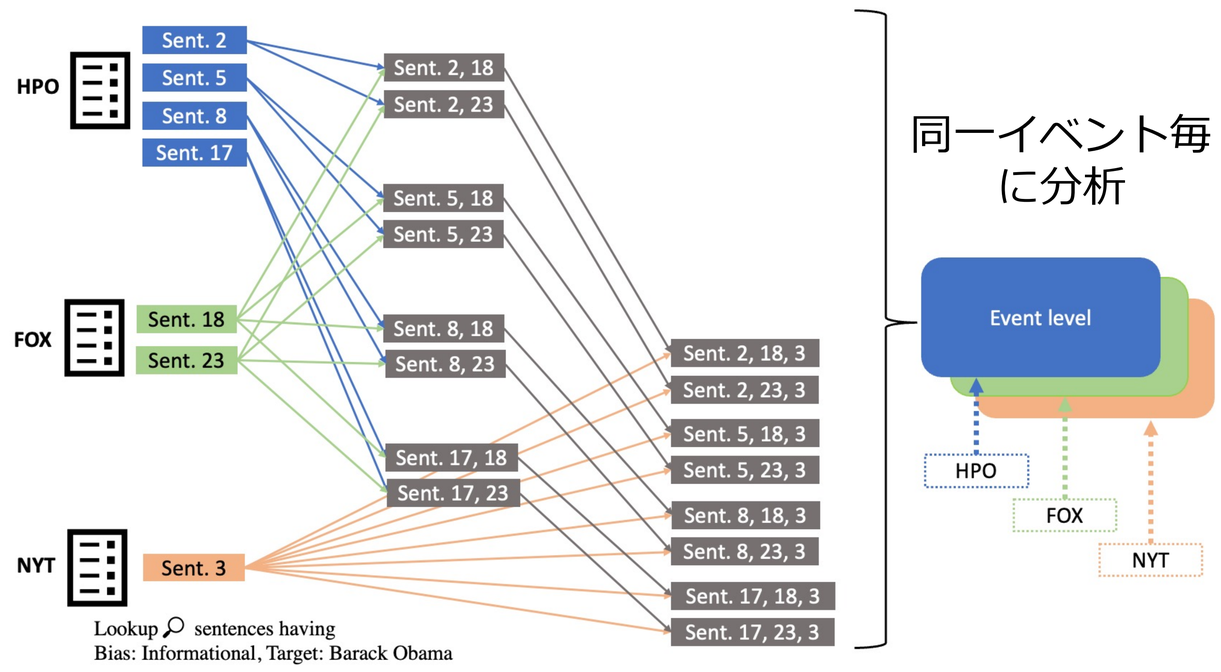
３社により報道された同一事象のバイアスをLLMにより推測させる

- 放送・メディアのバイアス：リベラル派・保守派があると言われる。アメリカの放送で特に顕著。
- LLMのバイアス：LLMにも学習データに起因するバイアスが存在
- LLMが報道記事のバイアスを正しく検知できるか、どのLLMなら可能か厳密に分析することが重要
- Fox News (FOX)とNew York Times (NYT)とHuffington Post (HPO)により報道された同一事象の300記事を分析し、専門家がアノテーションしたBASILデータセットを利用

BASILデータベースに含まれるNew York Timesのサンプル

文章毎に情報バイアス（INF）と語句バイアス（LEX) が付与

バイアスタイプ	文章	文章番号
INF	**The president again suggested that he should win the Nobel Peace Prize, and he reviewed which conservative commentators had been supportive of him, while dismissing Ann Coulter, who has not.**	10
LEX	Sounding alternately **defensive and aggrieved,** Mr. Trump explained his failure to secure wall funding during his first two years in office when Republicans controlled both houses of Congress by saying, I was a little new to the job.	11



実験結果

Prompt Template	INF/OTH	INF/LEX
CONCISE: Given the following text, classify it as either c_a or c_b, Text: s, Classification:	c_a: “bias” c_b: “no bias”	c_a: “inf” c_b: “lex”
DETAILED: I want you to act as a bias detector. The definition - of “Informational bias” is it uses speculative and tangential clauses in a factual and neutral way to sway readers. In contrast, the definition of “Lexical bias” is it depends on linguistic attributes like word choice and syntax. Given the definitions, classify if the following text exhibits c_a or c_b. Text: “s”, Classification:	c_a: “inf” c_b: “lex”	

バイアスの有無の予測精度　バイアスタイプの予測

	INF/ OTH			INF/ LEX		
			F1 Score			F1 Score
Model	**Acc.**	**INF**	**OTH**	**Acc.**	**INF**	**LEX**
FLAN-T5-XXL	76.67%	60.71	77.34	72.88%	56.74	38.73
FLAN-UL2	78.87%	61.08	69.40	81.89%	62.91	41.36
GPT-3.5-turbo	75.98%	65.86	75.83	76.08%	60.89	40.58
GPT-4 1.78T	78.88%	**69.07**	73.59	82.99%	64.55	42.76
OPT 2.7B	76.24%	46.11	53.98	69.33%	45.25	33.15
BERT (Chen et al., 2020)	-	41.46	-	-	-	-
RoBERTa (Lei et al., 2022)	-	46.47	-	-	-	-
ArtCIM (van den Berg and Markert, 2020)	-	42.80	-	-	-	-
EvCIM (Guo and Zhu, 2022)	-	45.81	-	-	-	-
MultiCTX (Guo and Zhu, 2022)	-	46.08	-	-	-	-
BERT+ctx (Maab et al., 2023)	84.90%	56.88	-	83.36%	74.01	66.97
BERT+ctx (**) (Maab et al., 2023)	**86.40%**	58.15	-	**84.77%**	**75.46**	**71.93**

(**) 追加の学習データを利用した結果

結論

- プロンプトによるLLMまたは、微調整されたBERTのいずれかを使用してバイアス検出可能
- 特にLLMは訓練されたデータがなくてもバイアスを検出可能
- BASILデータセットでの精度は約70～80%程度

連絡先：山岸研究室／ 国立情報学研究所　コンテンツ科学研究系
TEL：03-4212-2576　Email：maab@nii.ac.jp