対数尤度の周波数成分への変換に基づく自然言語文内のバイアス検知

石戸谷 由梨　(指導教員：小林 一郎)

1　はじめに

近年，情報の発信源の信頼性を担保するために，人か言語モデルかによるテキストを区別する手法が研究されている [1][2][3]．言語モデルは，生成する文の尤度が高くなるように語彙を選ぶため生成文は尤度が高くなるが，人によるものはその限りではないという傾向がある．しかし，言語モデルが人間に近い文章を生成する能力が向上するにつれ，この区別は困難になりつつある．このような現状に対して，Xu ら [3] は言語の相対尤度スペクトルビューという手法を用いて，尤度を周波数領域で分析することが，識別に優位であることを示した．また，人間が生成するテキストには，しばしば無意識にバイアスが含まれており，そのようなデータで学習された言語モデルも，同様のバイアスを引き継ぐ問題が指摘されている．この問題に対処するため，様々な文章のバイアスを検出する技術が研究されている [4, 5]．これらを踏まえ，本研究では，言語の尤度を元にバイアスの有無を検知すること，また尤度を周波数領域へ展開して得られた周波数特性値を用いてバイアスの検知を行う．

2　研究概要

実験 1 では，尤度を用いてバイアスの有無を識別する．単調な文章に関してバイアスのあるものとないものを用意し，尤度スコアに基づき両者を比較分析する．実験 2 では，実験 1 でのデータに加え，より複雑な文章を使用し，Xu らが考案した文章の尤度を離散フーリエ変換によって周波数成分に変換した値に基づき尤度の微細な傾向を捉える手法を用いてバイアスの有無を識別する．識別には 3 種類の教師あり学習手法を用い，バイアス有無の二値分類を行った．

3　実験 1

3.1　実験設定

尤度の算出方法　尤度を算出したい文章をトークン化し数値ベクトルに変換する．得られたベクトルを，尤度を計算するモデルに入力し，各トークンの損失を計算した．ここで算出される損失が負の対数尤度となる．今回，尤度の算出には GPT-2 [6] を使用した．

使用データ　実験には StereoSet [4] と Bias Evaluation Across Domains(BEADs) [5] を使用した．StereoSet は，性別，職業，人種，宗教の 4 つの領域における典型的なバイアスを測定するためのデータセットである．データは，文章においてバイアスのある位置が文章間と文章内の 2 つに区分されており，文章内の特定の単語が stereotype（バイアスのあるもの）と anti-stereotype（バイアスのないもの）に分類されているデータセットを使用した．バイアスの種類は性別に関するもののみで統一し，総計 255 個のバイアスの有無が対となったデータを作成した．その例を表 1 に示す．

BEADs はバイアスの検出において，テキスト分類，トークン分類，バイアスの定量化，良性言語生成など，幅広い NLP タスクを対象とするように設計されたデータセットである．今回は，良性な言語生成に関するデータセット中の，バイアスのある文章とない文章が対になった全部で 8,000 ペアのデータを使用した．

表 1: StereoSet のデータ例．太字の単語はバイアスが現れている / 取り除かれている箇所を示す．

Type	Sentence
stereotype	School girls are so **innocent** in every movie that I watch.
anti-stereotype	School girls are so **strong** in every movie that I watch.

データの選定　尤度の計算過程のうち，トークン化を行う際に未知語に対してサブワードによるトークン化が行われる．これらのデータは，トークンごとの純粋な尤度を算出すること難しいと考え，分析対象から除外した．操作としては，トークン化して得られる数値ベクトルの配列の長さが一致しているもののみをデータとして使用した．最終的に 162 個のデータを使用して分析を行った．

3.2　実験課題

バイアスがある文章とない文章では尤度に違いが現れるのかを調査するため，上記の方法で各文に対して尤度を算出した．尤度はトークン化した際の次元数と同じ個数算出される．そのため，定量的に調査するために，尤度の平均値を算出し，バイアスの有無で大小関係にどのような特徴が現れるのかを調査した．

3.3　実験結果

StereoSet は，162 個のデータ中 118 個のデータにおいて，バイアスのある文章の方が尤度の平均値が高くなっていることがわかった．BEADs は 8145 個のデータ中 7798 個のデータにおいて，バイアスのない文章の尤度の平均値が高くなった．明確な傾向としては確立されていないが，尤度によってバイアスのある文章とない文章の判定が可能であることがわかった．

3.4　考察

StereoSet と BEADs のバイアスの有無による尤度の傾向が異なった要因として，StereoSet の文章は単調かつ短文であるため，相対的なトークンごとの尤度に傾向が現れにくいことが考えられる．また，BEADs の方が尤度での分類結果が優れていた要因として，BEADs のデータセットにおいて，バイアスがある文章と比較を行った結果，バイアスがあるとされている箇所のトークンが，意味的に不自然な言葉選びになっているものなどがあり，バイアスの有無よりも尤もらしさに起因していると考えられる．

4 実験 2

4.1 実験設定

尤度の周波数領域への変換方法 尤度の周波数成分への変換するには，Xu らの手法と同様，3.1 の手法で求めた負の対数尤度に対して，離散フーリエ変換を行うことで値を得た．

識別器の作成 StereoSet と BEADs のそれぞれのペア文の負の対数尤度の系列データに対して，フーリエ変換により求めた周波数成分の値を一般化加法モデルを用いて離散的な周波数成分から連続の関数を求め，特定の周波数からの値をサンプリングすることにより，同一の次元数の周波数成分を用意した．最終的に周波数成分を説明変数とし，バイアスの有無を目的変数として，ロジスティック回帰と多層パーセプトロン，ナイーブベイズ分類器の 3 種類の教師あり識別器を作成した．

使用データ 実験 1 と同様のデータをセットを使用した．また，ベースラインとの条件を揃えるために，サンプリングの際の次元数はそれぞれ 384 次元，768 次元，1536 次元にした．

4.2 実験課題

バイアスがある文章とない文章の尤度に対して，上記の方法で各文に対して周波数成分の値に変換した．尤度から変換した周波数領域の値は，低周波成分には尤度が高い成分が出現し，高周波成分には尤度が低い成分が出現する傾向があることから尤度よりもより微細な違いや傾向を捉えることができる．これにより，バイアスの有無を判別する識別器を作成し，既存の手法との精度比較を行った．

4.3 ベースライン

性能を比較する際のベースラインとして，文章の埋込みベクトルを算出し，それを元にバイアスの有無で二値分類をした．埋込みベクトルの算出には OpenAI の API[1]，BERT[7]，Sentence-BERT[8] の 3 つを使用した．上述した 2 種類の識別器を作成し，その結果を表 2 に示す．

4.4 実験結果

実験の結果を表 3 に示す．StereoSet のようなバイアスがあるとされる特定の箇所の単語のみが異なる意味的に単調な文章のデータセットに関しては，ベースラインよりもわずかに本手法の精度が低くなった．BEADs のような口語のように複雑なデータセットにおいては，ベースラインの結果の方が優位であった．

4.5 考察

使用データセットのバイアスがあるものとないものを用意する過程にあたり，BEADs はバイアスがある文から言語モデルを通じてバイアスがない文を生成しペア文を作成している．バイアスがある文章とない文章のペアにおいて，作成元が異なることが，StereoSet と BEADs の精度の違いの要因となったと考えられる．また，周波数成分を用いた手法においては，入力ベクトルを作成する際の周波数成分のサンプリング位置などにおいて工夫が必要であることが考えられる．

[1]https://platform.openai.com/docs/guides/embeddings

表 2: 埋込みベクトルによる実験結果．2 種類の識別器のうち，精度の良いものだけを表に示した．

Dataset	Model	Best Acc.	Classifier
StereoSet	OpenAI	0.59	NB
	BERT	0.56	NB
	Sen-BERT	0.53	LR
BEADs	OpenAI	0.99	MLP
	BERT	0.90	LR,MLP
	Sen-BERT	0.80	MLP

表 3: 埋込みベクトルによる実験結果．2 種類の識別器のうち，精度の良いものだけを表に示した．

Dataset	Dimension	Best Acc.	Classifier
StereoSet	384	0.53	NB
	768	0.53	NB
	1536	0.53	NB
BEADs	384	0.61	LR
	768	0.62	LR
	1536	0.61	LR

5 まとめと今後の課題

本研究では，文章におけるバイアスの検出において，尤度とその値を周波数成分に変換した値を用いてバイアスの有無の識別を行い，ベースラインとなる埋込みベクトルによる識別精度と比較を行った．その結果，尤度や周波数領域に変換した値による識別にはそれぞの利点があると言えるが，総じて埋込みベクトルを入力情報として採用した際の結果が良いという結果となった．今後の課題，尤度の性質によるノイズを減らすために，バイアスのある文章とない文章，どちらもモデルで作成し今回と同様の実験を実施することや周波数成分のサンプリング位置の工夫などが考えられる．

参考文献

[1] Mitchell, et al. Detectgpt: Zero-shot machine-generated text detection using probability curvature. Vol. 202, 2023.

[2] Bao, et al. Fast-detectgpt: Efficient zero-shot detection of machine-generated text via conditional probability curvature. *arXiv preprint arXiv:2310.05130*, 2024.

[3] Xu, et al. Detecting subtle differences between human and model languages using spectrum of relative likelihood. *arXiv preprint arXiv:2406.19874*, 2024.

[4] Moin Nadeem, et al. Stereoset: Measuring stereotypical bias in pretrained language models, 2020.

[5] Raza, et al. Beads: Bias evaluation across domains. *arXiv preprint arXiv:2406.04220*, 2024.

[6] Radford, et al. Language models are unsupervised multitask learners. *OpenAI blog*, Vol. 1, No. 8, 2019.

[7] Devlin, et al. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[8] Reimers, et al. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.