TITLE		Rep-No.
先行研究と今後の計画		1
AFFILIATION		
九州大学大学院 システム情報科学府 鈴木研究室		
AUTHOR	POSITION	DATE
Ichinsoe Haruki	B4	2025/05/14

1 先行研究：データの偏見的説明の判定

本節では, Rosling らの『Factfulness』で提唱された 10 の人間の本能（ギャップ本能、恐怖本能、単一視点本能など）[1] が悪用されることで, 統計的に非倫理的であり偏りがあるが, 一定数の人が信じてしまうような偏見的説明になっているか否かということを判定する Zhang らの 2 つの研究を先行研究として紹介する [2, 3].

1.1 フレーズ間の意味的類似度に基づくデータの偏見的説明判定

Zhang ら（2022）は, このような非倫理的であり偏りがあるが, 一定数の人が信じてしまう偏見的説明を 18 個（Type I〜XVIII）定義し, さらに, その説明が偏見的説明か否かを判定する 3 つの手法（α）,（β）,（γ）を提案している [2].

(1) 手法（α）

法 α は, 習慣 X と病気 Y に関する解説文を判定する手法であり, 以下の 3 つの指標によって判定する。

- 習慣と病気の関連度：$\theta_{\text{relevance}}$
- 恐怖度：$\theta_{\text{bad habit}}$
- 悪い習慣度：θ_{fear}

それぞれの指標は次のように定義される.

$$\theta_{\text{relevance}} = \frac{\text{Sim}(X, Y)}{\text{Sim}(\text{base word}, Y)}$$

$$\theta_{\text{bad habit}} = \frac{\text{Sim}(X, \text{“bad habit”})}{\text{Sim}(X, \text{“good habit”})}$$

$$\theta_{\text{fear}} \Rightarrow \text{DALYs（障害調整生存年数）に基づくスコア}$$

これら 3 つの値がそれぞれ 1 を超えた場合, その説明は偏見的説明であると判定される. この条件を, 以下のように表す.

$$\begin{aligned}&\text{IF} \quad \theta_{\text{relevance}} > 1 \quad \wedge \quad \theta_{\text{bad habit}} > 1 \quad \wedge \quad \theta_{\text{fear}} > 1\\ &\text{THEN} \quad 1\\ &\text{ELSE} \quad 0\end{aligned}$$

(2) 手法（β）

手法 β は、「主語 X が他のいくつかの主語 X' と比較して性質 Y を持つ」ことを説明する解説文を判定する手法であり, 次の 2 つの条件がともに成立する場合にのみ, その説明を偏見的説明と判定する.

(1) 主語 X が, 逆の性質 $\bar{Y}$ よりも性質 Y に関連していること.
(2) 性質 Y が, 他クラスに属するすべての主語 X' よりも主語 X に強く関連していること.

これを X と Y の関連度 θ^{β}_{XY}, X と $\bar{Y}$ の関連度 $\theta^{\beta}_{X\bar{Y}}$, X' と Y の関連度 $\theta^{\beta}_{X'Y}$ を用いて以下のように表す.

$$\text{IF} \quad \theta_{XY} > \theta_{X\bar{Y}} \quad \wedge \quad \forall X'\,(\theta_{XY} > \theta_{X'Y}) \tag{1}$$
$$\text{THEN} \quad 1 \tag{2}$$
$$\text{ELSE} \quad 0 \tag{2.2}$$

(3) 手法（γ）

手法 γ は,「主語 X が他の種類の傾向 Y' とは異なる傾向 Y をもつ」ことを説明する解説文を判定する手法であり, 主語 X が, 異なるどの傾向 Y' よりも傾向 Y と強く関連する場合のみ, 偏見的説明と判定する. この条件を, X と Y の関連度 θ^{γ}_{XY}, X と Y の関連度 $\theta^{\gamma}_{XY'}$ を用いて以下で表す.

$$\text{IF} \quad \forall Y'\,(\theta_{XY} > \theta_{XY'}) \tag{3}$$
$$\text{THEN} \quad 1 \tag{4}$$
$$\text{ELSE} \quad 0 \tag{5}$$

これらの手法を用いて, 定義した解説文とそのフレーズを一部変えた変種に対して実験を行った結果, （α）と（γ）は正答率 1, 0.914 と比較的高い結果を残したが, （β）は, 正答率 0.510 という結果となった. これは,（β）が限られてフレーズ間の意味的類似性のみ測定を行っているためであると考えられた.

1.2 類義語フレーズ類似度に基づくデータの偏見的説明判定

その後の研究で Zhang らは、特に手法 (β) における判定精度の低さを改善するため、Phrase Similarity Graph を導入し, 新たな判定手法 (β^2) を提案した [3].

これは, 主語, 比較主語, 目的語, 性質, 比較性質を表す X, X', Y_{base}, Y, $\bar{Y}$ のそれぞれに対して, ChatGPT を基に類義語を含んだフレーズ群を生成し, 幅広いフレーズ間で意味的関係を探索するものである. 生成の過程では, 元々の語と類似語群の間に 3 層のフレーズ類似度グラフを作成し, それを図 1 のような部分グラフに分解を行った. その後, グラフエントロピーを求めて, 偏りのある部分グラフに重み付けを行い, 信頼生成スコアを測定した.

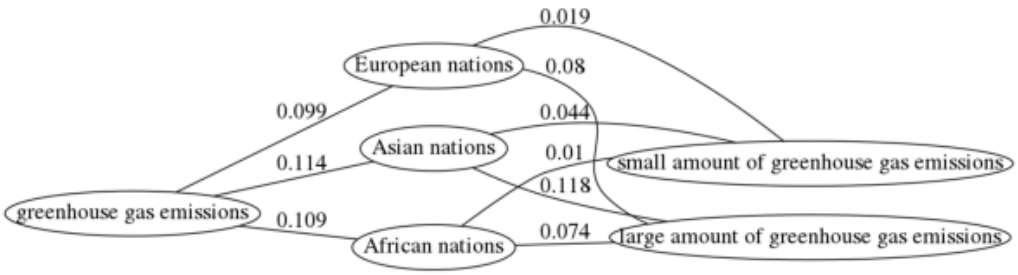

図 1: サブグラフの例

この改良により、従来の β 手法の精度が 81.1% まで向上し, 直感に反する類似度の問題を緩和できることが示された.

2 先行研究：データの偏見的説明に関する系統的分類と LLM 生成

樋口氏の修士論文では, Zhang らの分類を基に図 2 に示すようにデータの偏見的説明を対象 S（1 つか 2 つ以上）, 時間的拡張の有無 T（ありかなし）に基づき 4 つのパターンに系統的に分類する手法を提案している [4].

T \ S	1つ	2つ以上
なし	I, II&III, XIV, XX, XXI, XXII, XXIV	IV, V, VII, XII, XV, XVI, XVII, XVIII, XIX, XXIII, XXV
あり	VIII, X, XI	VI, IX, XIII

図 2: データの偏見的説明の分類

また, この系統的分類を基に図 3 に示すような本能決定の手続きを定め, LLM を用いてデータの偏見的説明の生成を行った.

図 3: 本能の決定手続き

LLM を用いて生成を行う過程では, ペルソナ設計と Chain-of-Thought によるプロンプトエンジニアリングを実施して出力精度の向上を図っている．その結果, 提案した分類パターンと本能決定手続きをプロンプト構造に明示的に組み込むことで, 意図した偏見的説明パターンに沿った説明文が高い一貫性で生成されることが確認された．以下に結果の一例として, ”gpt-4o-mini”を用いた際の結果を判定方法毎に示す.

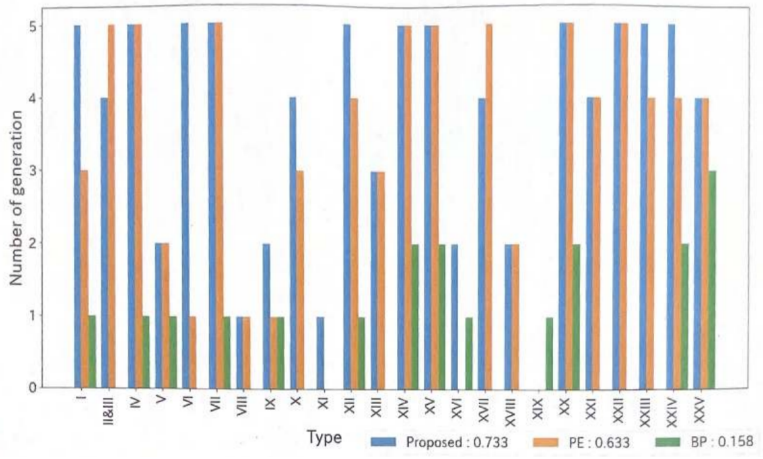

図 4: LLM:gpt-4o-mini, 判定方法：自身

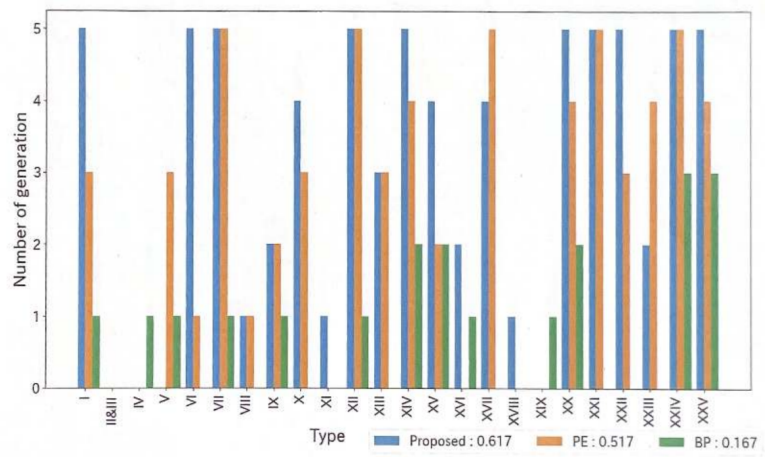

図 5: LLM:gpt-4o-mini, 判定方法：LLM

以上より, 本研究で提案された分類フレームワークとプロンプト設計手法は, 偏見的説明の多様なパターンを体系的に再現・制御可能であることを実証し, 今後の説明文生成タスクや偏見評価タスクにおいて有用な基盤となりうると結論づけられる.

3　今後の計画：一般時事ニュースからの偏見的説明の生成

これまでの先行研究を踏まえて, 私は統計データに基づく一般時事ニュースの中から, 信憑性が高く見える一方で, 読者に偏った印象を与えるような説明を生成することを今後のテーマに据えようと考えている. 実現を目指す流れは以下の 3 段階から成る.

1. **統計データからの要素抽出と本能決定**：時事データから, 主語・性質・対象群などの要素を構造的に抽出し, それに対応する人間の本能を決定する。
2. **偏見的説明文の生成**：抽出された要素と選定された本能に基づき, データの一部のみを強調・誇張するような説明文を生成する.
3. **生成された説明の判定**：Zhang らの手法（α, β, γ, β^2）や LLM を用いて, 生成された説明文が偏見的説明であるかを判定する.

この一連のフローにより, 単なる誤情報や虚偽ではないが, 受け手の直感や判断を不適切に誘導する説明文を系統的に生成・分析・判定する枠組みの実現を目指している.

今後の課題としては, 以下の点が挙げられる.

- 統計データからどのように有効な構成要素（主語・性質・比較対象など）を抽出し, それに対して適切な本能ラベルを自動的に割り当てるか
- 抽出要素と本能に基づいて, どのようなモデル（LLM, VLM, Deep Learning 等）を用いれば, 効果的かつ制御可能に偏見的説明文を生成できるか
- 具体的にどのような一般時事ニュースのデータを利用していくか（米の価格推移、外国人労働者数と犯罪件数の変化など）

なお, 本研究における偏見的説明文の生成は, あくまで「そのような説明がどのように構成され, どのような要素が人間の認知に影響を与えるか」を明らかにし, ひいてはそのような説明文の生成の防止や検知につながることを期待しており, 不適切な説明文の自動生成自体を目的とするものではない.

参考文献

[1] H. Rosling, A. R. Rönnlund, and O. Rosling. *Factfulness: Ten Reasons We ' re Wrong About the World—and Why Things Are Better Than You Think*. Sceptre, London, United Kingdom, 2018.

[2] K. Zhang, H. Shinden, T. Mutsuro, and E. Suzuki. Judging Instinct Exploitation in Statistical Data Explanations Based on Word Embedding. In *Proc. AIES ' 22*, pp. 867–879, 2022.

[3] K. Zhang and E. Suzuki. Judging Credible and Unethical Statistical Data Explanations via Phrase Similarity Graph. In *Proc. 2023 Pacific Asia Conference on Information Systems (PACIS)*, 2023.

[4] T. Higuchi. データの偏見的説明に関する系統的分類と *LLM* 生成. Master's Thesis, Kyushu University, Graduate School of Information Science and Electrical Engineering, 2024.