TITLE 提案手法とその具体例		Rep-No. 2
AFFILIATION 九州大学大学院 システム情報科学府 鈴木研究室		
AUTHOR Ichinose Haruki	POSITION B4	DATE 2025/06/04

1　前回の発表について

前回の発表では, データの偏見的説明に関する 3 つの先行研究の概要の説明を行った. またそれらを踏まえ, 当初は新たな統計データから偏見的説明の生成を行う研究を提案していたが, その後の検討により, 偏見的説明に対する対抗的な説明文を生成する研究へと方針を変更した.

2　提案手法

先行研究で示された偏見的説明の分類や, それらが悪用する人間的な本能の知見を活用し, LLM を用いて, 統計データとその偏見的説明文から, それに対する効果的な対抗説明文を生成する手法を提案する. 本提案手法は, 大きく分けて以下の 2 つのフェーズで構成する.

- フェーズ 1: SelectInstinct による本能の特定
- フェーズ 2: LLM による対抗文生成

2.1　フェーズ 1: SelectInstinct による本能の特定

このフェーズでは, 偏見的説明文と関連する統計データから, その説明文が悪用している可能性の高い人間の本能を特定する. その手法として, 先行研究で提案された偏見的説明文生成時の本能選択手続きである SelectInstinct の考え方を使用する [2].

Algorithm 2: SelectInstinct: selecting human instincts

Input: S: subject number; T: temporal change; $\boldsymbol{V}$: subject names; $\boldsymbol{D}$: statistical data
Output: $\boldsymbol{I}$: set of human instincts
1: Set $\boldsymbol{I}$ based on Table 2 with S and T
2: $\boldsymbol{I} \leftarrow \boldsymbol{I} \cup \text{PessimismInstincts}(\boldsymbol{V})$
3: **if** $T =$ "True" **then**
4: $\quad \boldsymbol{I} \leftarrow \boldsymbol{I} \cup \text{TrendInstincts}(\boldsymbol{D})$
5: **end if**

図 1: SelectInstinct

SelectInstinct は, 本来, 統計データの対象数 S, 時間変化 T, 対象名 V, そして統計データ D を分析し, LLM が特定の偏見的説明文を生成するために, どの本能 (例: 恐怖本能, ギャップ本能など) を悪用すべきかを自動で選択するアルゴリズム群として設計されている.

本フェーズでは, この SelectInstinct の「本能を選択する」というロジックを逆の視点から活用する. つまり, 偏見的説明文が, どの本能を悪用することでその偏見を生み出しているのかを分析・特定するために, SelectInstinct の判断基準や手続きを指針として用いる.

- 入力
  - S：対象数
  - T：時間拡張の有無
  - V：変数名
  - D：統計データ

w3

- 出力
  - I : 偏見的説明文が悪用している可能性のある本能の集合

2.2 フェーズ 2: LLM による戦略的対抗文生成

このフェーズでは, フェーズ 1 で特定した情報に基づき, LLM を用いて対抗的な説明文を生成する. プロンプトの主要構成要素: LLM に与えるプロンプトは, 主に以下の要素で構成する。

- 入力
  - 統計データ
  - 偏見的説明文
  - 悪用している可能性のある本能の集合 **I**
  - 対抗文生成のためのプロンプト:
    - ペルソナ設計
    - Chain-of-Thought (CoT) を促す指示
    - 出力形式の指定
- 出力
  - 対抗文: 偏見的説明文の誤りを指摘し, 統計データに基づいたより正確でバランスの取れた解釈を提示する説明文.

3 具体例

ここでは, 先行研究 [1, 2] で挙げられている「Type V: 女性は男性よりも数学の点数が低い (Women have lower math scores than men.)」を例に, 提案手法の適用イメージを示す.

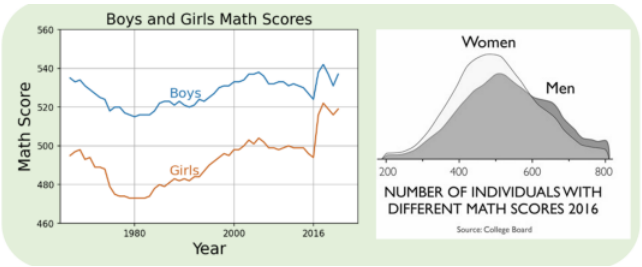

図 2: Type V の統計データ (zhang2022[1] より)

- 提示される情報:
  - 偏見的説明文: 「女性は男性よりも数学の点数が低い.」
  - 統計データ: 男女別の数学の平均点の比較と, それぞれのスコア分布図.
- フェーズ 1: **SelectInstinct** に基づく悪用されている本能の特定
  - 入力:
    - S : 2 つ (男性と女性)
    - T : なし
    - V :「男性の数学の点数」「女性の数学の点数」
    - D : 男女の数学の平均点とスコア分布図 (図 2 を参照)
  - 出力: { 分断本能, パターン化本能, ネガティブ本能, 単純化本能 }
- フェーズ 2: LLM による対抗文生成

– プロンプトの構成要素:
  * **統計データ:** 男女の数学の平均点とスコア分布図.
  * **偏見的説明文:** 「女性は男性よりも数学の点数が低い.」
  * **特定された本能:** { 分断本能, パターン化本能, ネガティブ本能, 単純化本能 }
  * **ペルソナ設計:**
    「あなたはデータリテラシーの専門家です. 統計データの誤解を解き, より正確でバランスの取れた理解を促すことを目指しています.」
  * **Chain-of-Thought (CoT) を促す指示:**
    1. 偏見的説明文が悪用している特定の本能を分析する.
    2. 偏見を正し, 客観的な事実を伝える対抗説明文を生成する.
  * **出力形式の指定:** 簡潔で分かりやすい説明.

– **期待される出力:**
「平均点だけを見ると女性は男性よりも数学の点数が低いと感じるかもしれませんが, スコア全体の分布図に注目すると, 実際には男女のスコアは広範囲に重なり合っています. 高い点数の女性も, 低い点数の男性も多く存在するため, 性別だけで数学の能力を一般化することは, データの全体像を見誤る可能性があります.」

参考文献

[1] 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.

[2] Anonymous. Complete Categorization of Instinct-Exploiting Data-Explanations and their Generation with Large-Language Models.