説明可能 AI のための対比因子ラベル生成手法に関する研究

清野駿（筑波大学大学院 修士 2 年）

背景と目的

なぜ XAI の重要性が高まっているのか？

- AI の社会実装が進む中で、**「なぜその判断なのか？」という疑問が増加**
- 医療・法務・教育など、**説明責任**が求められる分野で利用拡大
- **LLM のようなブラックボックスモデルの普及**により、透明性の欠如が顕著に

現在の課題は何か？

- 多くの AI モデルは「正答を出す」が、**理由は出さない**
- 出力結果に対して、人が **納得・理解できないケースが多い**
- 特に LLM では、**創発的な挙動**が人間にとって予測困難
- 「説明できない AI」は、社会的信頼を失うリスク

これまでの主な取り組み

- **可視化手法**（Attention Map、SHAP、LIME など）：画像・特徴量に特化
- **ルール抽出・事後説明型 XAI**：決定木や説明生成モデル
- **生成系 XAI**（例：GPT による説明生成）：文脈に応じた自然言語説明を試みるが、一貫性・根拠の妥当性に課題
- 多くが **入力と出力の間にある“中間的特徴”の扱いが曖昧**

本研究の立場と貢献

- LLM による自然言語説明生成を活用し、**人間が読んで納得できる差分説明**を目指す
- そのために → 説明対象を「**二つのテキスト集合の差異（対比因子）**」と定義 → **抽象的な判断過程ではなく、具体的な“違い”に注目**
- 本研究は、「創発言語の意味理解」という最終目標に向けた **中間的ステップ**

背景と目的の修正したいところコメント

- 研究の目的はあくまで「検証」
    - 対比因子の生成というタスクを、GPT にやらせた例がない
    - →GPT にやらせたらできるのか、できるならどの程度できるのか、これを検証する

- 立場と貢献のところの内容をこの方向性で修正したい

研究アプローチ

本研究のアプローチ概要

- GPT にプロンプトを与え、**グループ A/B のレビュー集合**を入力
- GPT は、**グループ A にのみ特徴的な差異を自然言語で記述**
- 出力された説明が「対比因子」として妥当かを検証

使用プロンプトの構造

```
以下の2つのデータグループを比較して、グループAに特徴的で
グループBには見られない表現パターンや内容の特徴を特定してください。
{examples_section}

【グループA】
{group_a_text}

【グループB】
{group_b_text}

{output_language}で{word_count}程度で、
グループAに特徴的でグループBには見られない
主要な違いを簡潔に回答してください。
```

- `examples_section`: Few-shot 例題（0〜3 件）
- `group_a_text`, `group_b_text`: 入力テキスト群
- 出力：**自然言語 1 文**（差異の説明）

本研究における「対比因子」とは？

> 「A に含まれて B に含まれないテキスト的特徴」

- 例：「A は価格に関する言及が多い」
- 抽象的な特徴ではなく、**文中の傾向・パターン**として表現される差異
- LLM がこれを自力で抽出できるかを評価

出力評価方法

意味的類似度を使った自動評価

- **BLEU スコア**：n-gram ベースの表層一致（語彙レベル）

- **BERT スコア**：意味空間でのベクトル類似（意味レベル）

評価の流れ

1. A/B グループを「ある特徴（例：価格）」で分離しておく
2. GPT 出力（例：「A は価格に触れている」）と 正解ラベル（例：「価格に関する特徴を持つ」）を比較
3. BLEU/BERT スコアを算出（0〜1）

→ どれだけ“意味的に近い”説明が生成できたかを測定

実験

実験の方向性

- 実験はさまざまな軸を設定してその軸ごとに、変数を変えることで検証する
- → 変更する
    - 実験は各データセットに対して、変数を変えて複数パターンで bert,bleu スコアを出す。
- データセット
    - SemEval レストランレビュー
    - Steam Game Review
- 変数
    - グループのデータ数(これは 300 で固定)
    - 例題の有無(例題の数ごとに 0,1,3-shot と名付ける)
    - LLM モデル(GPT4o-mini で固定)
- コメント: 内容多いから二つのスライドに分割していきたい

---

3. 実験設計（変数）の表

軸	内容
Few-shot	0, 1, 3-shot
入力件数	300
モデル	GPT-4o-mini
データセット	SemEval, Steam

評価方法

- 評価指標：BERT（意味）／BLEU（語彙）
    - グループ A と B を、そのデータセットに元から設定されている、特徴で分けておく

    - 元からある特徴を、正解データとする
- temperature：0.7、seed：42 で、LLM とその他ランダム値を固定しています

---

4. 結果

結果（データセットごとの平均比較）

データセット	BERT	BLEU
SemEval	0.718	0.015
Steam	0.672	0.014

Steam ゲームレビューの結果概要

Steam レビュー実験：概要と傾向

- gameplay や story など語彙が明確なアスペクトは高スコア
- recommended や suggestion など抽象的な属性は苦手傾向
- few-shot により「推薦」などの語彙が明確化する例も観察

Few-shot による性能変化（例：gameplay）

Shot	BERT	BLEU	説明語彙の例
0	0.529	0.000	メカニクスへの言及
1	0.824	0.080	gameplay mechanics
3	0.824	0.080	exploration も追加

- → 例示によって差異を“特定する”語彙にシフト

アスペクト別の平均スコア（高い順）

アスペクト	BERT	BLEU
gameplay	0.726	0.054
story	0.592	0.000
audio	0.554	0.000
visual	0.535	0.000
price	0.528	0.000

アスペクト	BERT	BLEU
technical	0.520	0.000
suggestion	0.477	0.000
recommended	0.475	0.000

- gameplay/story など“ゲーム体験の内容”に関する語は高精度

考察：Steam レビューにおける特徴

- gameplay や story のような語彙的に安定した特徴は抽出しやすい
- recommendation や suggestion は文脈的・抽象的で難易度が高い
- few-shot により、LLM が抽出タスクに適した出力へ矯正される傾向

詳細結果 (PyABSA SemEval レストランレビュー)

全体統計

指標	値
総実験数	12 件
成功実験数	12 件 (100%)
平均 BERT スコア	0.681
平均 BLEU スコア	0.022
BERT スコア範囲	0.554 - 0.771
BLEU スコア範囲	0.000 - 0.080

---

few-shot による分析

Shot 設定	実験数	平均 BERT スコア	平均 BLEU スコア
0-shot	4 件	0.606	0.005
1-shot	4 件	0.730	0.022
3-shot	4 件	0.708	0.040

- **1-shot で大幅な性能向上**: 0-shot → 1-shot で BERT スコアが 0.606 → 0.730 に向上
- **例題効果**: わずか 1 つの例示で LLM の出力精度が劇的に改善
- **3-shot での安定化**: BLEU スコアも 0.005 → 0.040 と 8 倍向上

結果

以下は food アスペクトでの例：

Shot	BERT スコア	BLEU スコア	LLM 応答	データ分割
0-shot	0.554	0.000	"Group A emphasizes staff friendliness and authenticity."	613 件 vs 4115 件
1-shot	0.743	0.033	"Focus on food quality and dining experience."	613 件 vs 4115 件
3-shot	0.771	0.080	"food quality and presentation"	613 件 vs 4115 件

- **0-shot**: food 関連語彙が出現せず
- **1-shot**: "food quality" が出現開始
- **3-shot**: "presentation" など詳細化
- **結論**: 例示により LLM 出力が正解方向に強化

---

service 着目

特に service でも同様の変化が見られる：

Shot 設定	BERT スコア	BLEU スコア	LLM 応答
0-shot	0.672	0.009	service quality（全体評価）
1-shot	0.753	0.054	service quality and attentiveness
3-shot	0.758	0.080	service and dining experience

- **語彙の具体化**: "attentiveness"（注意深さ）などの詳細な表現が導入
- **アスペクトの精緻化**: 抽象的な概念から具体的な特徴へと発展
- **BERT スコア向上**: より詳細な語彙が意味類似度の改善に寄与

アスペクト別平均スコア（高い順）

抽象度の高い概念でも、十分な例示があれば抽出可能という傾向が見られる。

アスペクト	BERT スコア	BLEU スコア
service	0.728	0.048
food	0.689	0.038
atmosphere	0.663	0.000
price	0.644	0.004

- **主観的概念が高スコア**: service・food は頻出語彙で安定した抽出が可能
- **price は低スコア**: 定量的だが表現が抽象的（「高い」「コスパ悪い」）で抽出困難

まとめ

Shot 設定	実験数	平均 BERT スコア	平均 BLEU スコア
0-shot	4 件	0.606	0.005
1-shot	4 件	0.730	0.022
3-shot	4 件	0.708	0.040

- **1-shot 効果**: 対比因子語彙（"food quality", "staff attentiveness"）が出現し、BLEU スコア上昇
- **3-shot 限界**: 内容充実する一方、焦点散漫化で BERT スコア低下のケースあり

Steam 実験との比較

Steam 実験との比較

指標	PyABSA 平均	Steam 平均
BERT 類似度	0.681	0.725
BLEU スコア	0.022	0.027

PyABSA の特徴：

- Steam より低スコアだが、抽象語彙・文体揺らぎが大きくタスク難易度が高い
- 情報密度・多様性では PyABSA が上回る
- Steam は英語として単純明快で対比構造が分かりやすい

---

5. 考察

考察 ①：Few-shot の効果と限界

Few-shot の効果

- 0-shot では曖昧な説明が多く、差異の特定が困難
- 1-shot 以上で、語彙の精度や焦点の明確化が顕著
    - 例：「food」「service」「recommendation」などのアスペクト語が登場
- 3-shot によって語彙がさらに詳細化される傾向
    - 例：「attentiveness」「presentation」などの具体的表現
- → Few-shot により、説明が“伝える”から“特定する”へと最適化

Few-shot の限界

- スコアの伸びが 1-shot で頭打ちになるケースがある
    - 3-shot では焦点が分散し、BERT スコアが低下する場合も

- 抽象的・主観的なアスペクト（suggestion など）は依然困難
- BLEU スコアは低水準で停滞
    - → 語彙の一致よりも、表現の多様性が影響している可能性

考察 ②：対比因子抽出の難しさと LLM の特性

アスペクトごとの難易度

- 高スコア（gameplay, food, service）
    - 語彙が安定・頻出しやすく、差異が言語的に現れやすい
- 低スコア（recommended, suggestion）
    - 概念的・メタ的であり、テキストから直接抽出が困難
- → LLM の生成傾向とズレやすい

LLM の応答スタイルの影響

- LLM は「共感的・抽象的」な表現を好む傾向
- → 0-shot では説明が曖昧になりがち
- Few-shot によって「比較的で説明的」なスタイルに矯正可能
- → 例示は LLM にとっての出力スタイルの“教師”となる

---

6. 結論と貢献

研究の結論

- GPT を用いた 対比因子生成の可能性と限界を定量評価した
- Few-shot プロンプティングにより最大 20% 程度のスコア向上
- 特に 1-shot 設定での効果が顕著
- 一部アスペクトでは、人間にとっても納得感のある出力が得られた

本研究の貢献

- LLM を活用した 対比的説明生成タスクの枠組みを初提案・検証
- 評価手法（BERT/BLEU）による定量的分析フレームワークを構築
- Steam / SemEval という異なるジャンルでの 再現可能な比較実験
- 創発言語の意味理解という XAI 研究への中間的貢献を達成

---

7. 今後の展望

実用性・汎用性の拡張

- 多言語レビュー・ノイズ含むテキストへの対応
- → より現実的な応用場面への展開
- ストリーミング処理でのリアルタイム説明生成
- → チャットボットなど対話型 AI への応用

評価と改善の方向性

- 人手による説明の主観評価（納得性・簡潔性・明瞭さなど）の導入
- 説明の「引用妥当性」「情報源の明示性」などの質的指標
- → GPT の“ハルシネーション傾向”の抑制に活用

長期的なゴール

- 本研究で得られた知見を踏まえ、
- 創発言語の意味理解フレームワークの一部として統合
- “言語を創る AI”と“その意味を理解する人間”の接続点の創出

ご清聴ありがとうございました
