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(説明可能 AI) 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？

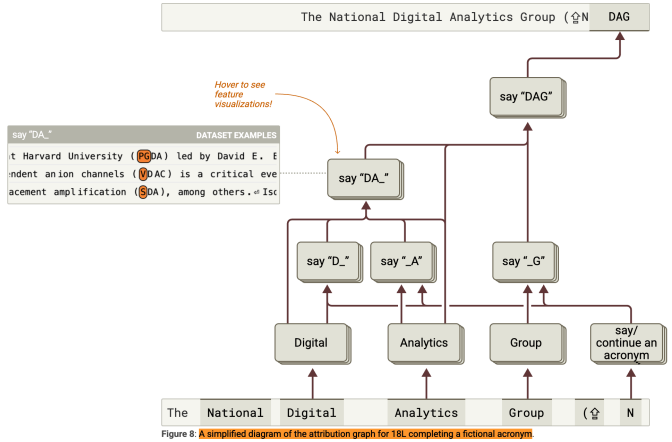
- 多くの AI モデルは「正答を出す」が、**理由は出さない**
- 出力結果に対して、人が **納得・理解できないケースが多い**
- 特に LLM では、**創発的な挙動**が人間にとって予測困難
- 「説明できない AI」は、社会的信頼を失うリスク

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

可視化手法の例

- Circuit Tracing: Revealing Computational Graphs in Language Models
- 「ここでニューロンが発火している」という、LLMの判断基準部分が可視化されている。
- しかし発火に至る判断過程は不明　→ 人間が目で見て、ラベリングしている



ニューロン発火パターンの可視化(anthropic)

引用元: https://transformer-circuits.pub/2025/attribution-graphs/methods.html

説明可能 AI（XAI）の究極のゴール

XAI が目指すもの

- **説明可能 AI（XAI）の究極のゴール**は、
  → AI が内部で使っている"AI 言語"を人間が理解できる形で翻訳すること

現在の課題

- **可視化手法**：ニューロンの発火は見えるが、**人手でのラベリングに依存**
- **既存の説明手法**：未知のデータセットに対して、**毎回人手でラベリング**が必要
- **根本的問題**：AI 内部の「意味」を人間が理解できる形で自動抽出する仕組みが不十分

→ **人手に依存しない自動説明生成**の仕組みが求められている

本研究の位置づけ

中間ステップとしてのアプローチ

- **「AI 言語の翻訳」に向けた中間ステップ**として、
  ２つのテキスト集合の意味的な差（対比因子）を説明させるタスクを設計
- **差異に含まれる意味**を自然言語で取り出せれば、
  → AI 内部で発生している**意味のラベル化・可視化**が可能

本研究の貢献

- LLM による「**対比因子生成**」というタスクの実現可能性
- 二つのテキスト集合の差異を **自然言語で説明させるプロンプト設計**
- LLM による出力の**正確さ・具体性・再現性**を定量的に評価

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」として妥当かを評価

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

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。

具体例：ゲームレビューでの対比因子抽出

グループ A（ストーリー重視）

- "Amazing storyline"（素晴らしいストーリー）
- "The plot is engaging"（プロットが魅力的）

グループ B（技術面重視）

- "Good graphics"（良いグラフィック）
- "Nice sound effects"（良い音響効果）

期待される対比因子
→ "story and narrative"（ストーリーと物語性）

LLM に与えられる課題：この対比因子を自動で発見・言語化できるか？

評価の流れ

1. A/B グループを「特徴（例：価格）」で分離しておく
   この**特徴を正解ラベル**とする
2. A/B グループを入力 → LLM 出力（例：「**グループ A は価格に言及している**」）
3. LLM 出力と正解ラベルを比較して、二つの類似度を 0~1 のスコアとして算出

→ どれだけ“意味的に近い”説明が生成できたかを測定

出力評価方法

意味的類似度を使った自動評価

- **BERT（Bidirectional Encoder Representations from Transformers ※1）スコア**
  - 意味空間でのベクトル類似（意味レベル）
- **BLEU（Bilingual Evaluation Understudy ※2）スコア**
  - n-gram ベースの表層一致（語彙レベル）

※1: Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv preprint arXiv:1810.04805. https://arxiv.org/abs/1810.04805

※2: Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002). BLEU: a Method for Automatic Evaluation of Machine Translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), (pp. 311-318). https://aclanthology.org/P02-1040/

実験

実験の目的とアプローチ

- **目的**：LLM が「対比因子」をどの程度適切に生成できるかを **定量的に評価**
- **手法**：
  - **2 種類のデータセット**に対して、**異なる検証軸**で GPT に説明生成を行わせる
  - 出力を **BERT/BLEU スコア**で評価
- **検証軸**：
  - Few-shot の効果
  - アスペクトの難易度
  - データセットごとの差異

実験条件の詳細（変数と設定）

使用データセット

1. SemEval レストランレビュー
    - 例："good food, great price, gread!"
    - 含有アスペクト："food","price", "service"
2. Steam ゲームレビュー
    - 例："great sounds and story I have ever played"
    - 含有アスペクト："audio","story"

※1: Pontiki, M., Galanis, D., Papageorgiou, H., Androutsopoulos, I., Manandhar, S., AL-Smadi, M., ... & Peev, V. (2014). SemEval-2014 Task 4: Aspect Based Sentiment Analysis. In Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), (pp. 27-35). https://aclanthology.org/S14-2004/
※2: ilos-vigil. (2024). Steam Review Aspect Dataset. GitHub. https://github.com/ilos-vigil/steam-review-aspect-dataset Kaggle. https://www.kaggle.com/datasets/ilosvigil/steam-review-aspect-dataset Licensed under CC BY 4.0.

変数と固定要素

項目	設定内容
グループの件数	各 300 件（A/B）で固定
Few-shot 設定	0-shot / 1-shot / 3-shot
使用 LLM	GPT-4o-mini（固定）
評価指標	BERT / BLEU スコア

実験設計

- Few-Shot 例題数を変えて出力を生成し、**説明精度の傾向を比較・分析**

評価方法の詳細

評価指標

- **BERT スコア**（意味類似度）
- **BLEU スコア**（語彙一致度）

評価の流れ

1. 正解ラベルと LLM 生成テキストを比較
2. 類似度を 0~1 で数値化
3. 高いスコア = より正確な対比因子の抽出

評価例と実験設定

具体的な評価例

- **正解ラベル**：「価格」
- **GPT 出力**：「グループ A は価格に言及」
- **結果**：BERT スコア: 0.76、BLEU スコア: 0.00

実験設定

- temperature: 0.7
- seed: 42
- ランダム値を固定して再現性を確保

結果

データセットごとの平均値

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

- どちらも BERT が高く BLEU が低い。
- 単語レベルでの一致などが少ない
- → 意味的には正解ラベルと LLM の出力が一致している

Steam ゲームレビューの結果概要

全体統計

指標	値
総実験数	24 件
平均 BERT スコア	0.5506
平均 BLEU スコア	0.0067
BERT スコア範囲	0.4400 - 0.8240
BLEU スコア範囲	0.0000 - 0.0800

- **BERT スコアは中程度**：意味レベルでの一致は確認できるが、抽出の精度はアスペクトに依存
- **BLEU スコアは極めて低い**：語彙レベルの一致がほとんど発生していない（自由生成傾向）

Few-shot ごとの平均値

Shot 設定	BERT	BLEU
0-shot	0.5114	0.0000
1-shot	0.5742	0.0100
3-shot	0.5662	0.0100

- **例題でスコアが向上**：0 → 1-shot でスコアが向上している
- **例題数の影響は小**：1-shot と 3-shot の差は小さい

Few-shot による性能変化の具体例（例：gameplay）

Shot	BERT	BLEU	説明語彙の例
0	0.529	0.000	グループ A はゲームの具体的なメカニクスや比較に焦点を当てている。
1	0.824	0.080	gameplay mechanics and exploration
3	0.824	0.080	gameplay mechanics and exploration

- **Few-shot による出力の変化**：例示によって、レビューグループ A と B の差異を**説明する → 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
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

PyABSA SemEval レストランレビューの結果概要

全体統計

指標	値
総実験数	12 件
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

アスペクトごとの比較

以下は food アスペクトでの例：

Shot0,1,3	BERT スコア	BLEU スコア	LLM 応答
0-shot	0.554	0.000	"Group A emphasizes staff friendliness and authenticity."
1-shot	0.743	0.033	"Focus on food quality and dining experience."
3-shot	0.771	0.080	"food quality and presentation"

- **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

SemEval レストランデータセット結果まとめ

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

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 対比因子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

評価と改善の方向性１

- 評価指標が BERT,BLEU のみで、文脈的な判断ができない
    - 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

評価と改善の方向性２

- データセットがレビュー系のみ
  - 感情分類など、より抽象度の高いデータセットでの検証
  - 例：GoEmotion データセットなど

評価と改善の方向性 3

- ベースライン手法
  - 基準となる手法が検討できていない（データセットのアスペクト同士のスコアを現在基準としている）
  - TF-IDF 等、LLM 以外によるアプローチでのベースラインを考える

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