TITLE		Rep-No.
関連研究の調査と今後の計画		3
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
Ichinose Haruki	B4	2025/06/24

1 前回の発表について

前回の発表では, 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の提案を行った. この手法は, 以下の 2 つのフェーズから構成されていた.

- 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 紹介論文：Towards Knowledge-Grounded Counter Narrative Generation for Hate Speech [1]

本研究は, ヘイトスピーチ（HS）に対抗するカウンターナラティブ（CN）の自動生成において, 従来の生成モデルが抱える具体性・根拠の欠如やハルシネーションといった課題を解決することを目的としている. そのため, Wikipedia やニュース記事などの外部知識に基づいて, より情報量が多く説得力のある CN を生成する, 初の包括的パイプラインを提案した.

2.1 論文情報

- **タイトル**: Towards Knowledge-Grounded Counter Narrative Generation for Hate Speech
- **著者**: Yi-Ling Chung, Serra Sinem Tekiroğlu, Marco Guerini
- **発表年**: 2021 年
- **公開**: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021
- **被引用数**: 78

2.2 背景と課題

- **HS 対策と CN の現状**
  - 現在, コンテンツ削除やシャドウバンなどが主な対策として行われているが, 言論の自由や多様性の観点から懸念も存在する.
  - 事実に基づいた議論や異なる視点による反論として, 非否定的な応答である CN が注目されている.
- **効果的な CN の例**

HS: *The world would be a better place without Muslims. They are only killing and raping our children.*
CN1: *Evidence for this? This is not true. How can you say this about an entire faith?*
CN2: *On the contrary, most children abuse is operated by people they know: a relative, family friends, sports coach, someone in a position of trust and authority. Besides, Muslims help people - A Muslim woman rushed to help the victims of a triple stabbing in Manchester on New Year's Eve.*
CN3: *You are truly one stupid backwards thinking idiot to comment on Muslims like that.*

図 1: HS に対する適切な CN（CN1, CN2）および不適切な CN（CN3）の例 [1]

  - CN1 と CN2 はいずれも適切だが, CN2 は事実や知識, 論理的推論に基づいており, より説得力が高い.
  - CN3 は攻撃的で不適切な応答である.

2.3 HS-CN データセット

- **データセット**: CONAN
- **特徴**: NGO の専門家によって作成されており, 高品質
- **データ数**: 6,645 組（英語）
- **内訳**:
  - オリジナルの HS-CN ペア: 1,288 組
  - パラフレーズ付きペア: 2,576 組
  - 翻訳ペア（仏・伊語から英語）: 2,781 組
- **分割**:
  - 訓練: 4,069 組
  - 開発: 1,288 組
  - テスト: 1,288 組

この研究では, リバースエンジニアリングを用いて, 既存データセットに知識を検索・付与する方法で新たなデータセットを構築している.

2.4 提案手法

本研究は, 外部知識に基づいた CN を生成するため, 以下に示す 2 段階の構成のアーキテクチャを提案している.

2.4.1 知識検索モジュール（Knowledge Retrieval）

与えられた HS に対して, CN の根拠となる知識を外部リポジトリから検索・抽出する.

- **知識ソース**:
  - Newsroom（130 万件以上のニュース記事）
  - WikiText-103（28,595 件の Wikipedia 記事）
- **クエリ構築法**: 関連知識を検索するためのキーワード（クエリ）を構築する.
  1. **抽出型（Extraction）**: 訓練時には, HS と正解 CN の両方からキーフレーズを抽出し, 検索精度を高める.
  2. **生成型（Generation）**: テスト時には正解 CN が利用できないため, HS を入力として, 反論に有効なキーフレーズを予測・生成する AI モデルを別途用意する.
- **知識選定**: クエリを用いて関連文書を検索後, 各文とクエリの関連度（ROUGE-L スコア）を計算し, 最も関連性の高い上位 5 文を知識文（KN）として選定し, 生成モジュールに入力する.

2.4.2 CN 生成モジュール（Counter Narrative Generation）

- **入力**: 元の HS および検索された知識文（KN）
- **モデル**: GPT-2 などの大規模言語モデルをファインチューニングし, 与えられた知識を反映させながら CN を生成する. 本研究では特に GPT-2 モデルを重点的に評価している.

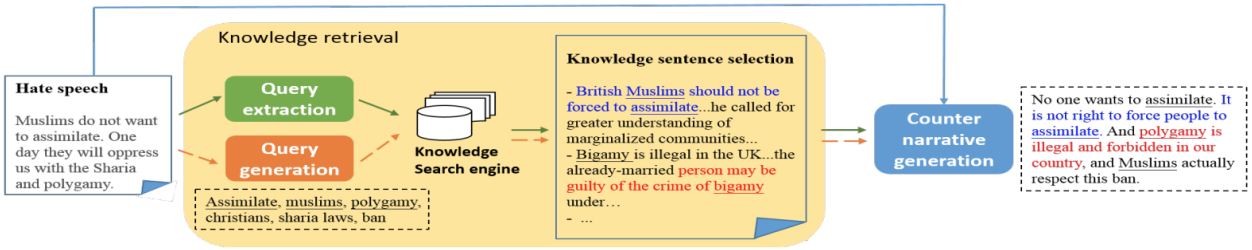

図 2: 抽出型（緑の実線矢印）と生成型（点線矢印）のクエリを用いて関連知識を検索する, 知識文に基づいた生成アーキテクチャ [1]

2.5 実験と結果

BLEU-2 や ROUGE-L を用いた自動評価と, NGO の専門家による手動評価の 2 軸で性能を検証を行った.

- **自動評価**: 知識を用いたモデルは, ベースライン（知識なし）と比較して, 生成文の新規性が著しく向上した. これは, 知識を導入することで, ありきたりでない, より情報量の多い反論が生成されたことを示唆する.
- **手動評価**: 専門家による評価では, 提案手法である GPT-2 が, 他のどのモデルよりも HS に対して適切であると最も高く評価された.
- **追加検証**: 高品質な知識を与えた場合や, 未知のトピックでテストした場合でも, 知識を注入するアプローチの有効性が確認された.

2.6 結論と貢献

- 本研究は, 外部知識に基づいた CN 生成の初の包括的な手法を提案し, その有効性を実証した.
- 外部知識の注入, 特にそのためのキーフレーズ生成が, より具体的で説得力のある CN 生成に寄与することを示した.
- 提案手法は未知のトピック（ゼロショット）に対しても一定の効果を発揮し, CN 研究に新たな方向性を示した.
- ただし, AI による自動生成にはリスクも伴うため, 最終的な判断は人間が行う「提案ツール」としての活用を推奨している.

3 自らの研究の展望

3.1 SelectInstinct による本能特定アルゴリズムの改善

- サブルーチン PessimismInstincts: より精緻な感情分析の導入
- サブルーチン TrendInstincts: 変化点検出アルゴリズムの導入

3.2 LLM による対抗文生成アルゴリズムの検討

- データセットを用いたファインチューニング

- プロンプトエンジニアリングによる最適化
- マルチモーダル LLM の導入検討
- 生成した対抗文の評価方法の検討

参考文献

[1] Yi-Ling Chung, Serra Sinem Tekiroğlu, and Marco Guerini. 2021. Towards Knowledge-Grounded Counter Narrative Generation for Hate Speech. In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 899–914, Online. Association for Computational Linguistics.