ホープスピーチ研究のための日本語データセット

和泉 悠 [1,4]　谷中 瞳 [2]　永守 伸年 [3]　荒井 ひろみ [4]
yuizumi@nanzan-u.ac.jp,　hyanaka@is.s.u-tokyo.ac.jp,
nnr12153@fc.ritsumei.ac.jp,　hiromi.arai@riken.jp
[1] 南山大学　[2] 東京大学　[3] 立命館大学　[4] 理化学研究所

概要

ヘイトスピーチをはじめ有害コンテンツの自動検出技術の重要性は言をまたないが，表現規制のハードルの高さが技術運用における課題となってきた．そのため，ヘイトスピーチを間接的に抑制する方法として，ホープスピーチの重要性が認識され始めた．本研究では，YouTube ニュース動画へのコメント約 1 万件によって構成されている，日本語ホープスピーチのデータセットを作成し，ラベルづけされたテキストの内容分析と GPT-4 を用いた評価実験を行った．

1　はじめに

オンライン上での有害コンテンツの一種である，社会的弱者への差別を煽動し攻撃するヘイトスピーチは，深刻な社会問題として学術的にも研究対象となってきた．自然言語処理分野においても，ヘイトスピーチ自動検出技術などの研究が進められてきた．

そうした技術を運用する際の大きな課題は，ヘイトスピーチやその他の有害コンテンツを検出できたとしても，どのようにそれらに介入すべきか明らかではない，というものである．市民が発信したコンテンツを削除や閲覧禁止にするためには，相応の根拠を与えなければならず，有害コンテンツ検出技術の運用は，表現の自由との間に緊張関係が存在する [1]．

そのため，ヘイトスピーチの直接的禁止ではなく，間接的抑制手段として，カウンタースピーチの重要性が認識されており，対抗言説の検出や自動生成といった研究が進められてきた．たとえば，差別的虚偽情報へのファクトチェックが自動的に行われるならば，虚偽情報の投稿を直接禁止せずとも，ユーザーに正しい情報を提供することが可能になる．

関連した研究動向として，近年ではオンライン上の「ホープスピーチ」に焦点を当てた研究が進められている．ホープスピーチとは，差別的デマの否定のような，ヘイトスピーチへの直接的対抗ではなく，何らかの意味で希望的な言説を指す．ヘイトスピーチを直接的に検出することへの補完的手法として，ホープスピーチを推奨することにより，オンライン情報空間の健全性に貢献できる可能性があるのである．

ホープスピーチ研究はいまだ端緒についた段階であり，先行文献には多くの概念的混乱が見られ，日本語での研究も存在しない．そこで本研究では，ホープスピーチ概念を整理するとともに，YouTube ニュース動画に投稿された日本語コメントにアノテーションを与えることにより，日本語ホープスピーチデータセットを構築することを目的とする．さらに，構築したデータセットを用いて，大規模言語モデルによるホープスピーチの自動検出の可能性について検討を行う．構築した日本語ホープスピーチデータセットは，今後研究利用可能な形式で公開予定である．

2　関連研究

心理的態度としての希望は，長年心理学 [2, 3] や哲学 [4] において探求されてきたが，希望的な言説が人々に与える影響については明らかではない．近年になり，自然言語意味論・言語哲学分野において，情報伝達のプロセスを超えた，言語が情動や集団行動に与える影響の重要性が認識され，そのメカニズムに関する研究が始まった [5, 6, 7]．本研究も，単なる情報伝達ではなく，他者の情動や行為に影響を与える可能性のある言説としてのホープスピーチに焦点を当てる．

ヘイトスピーチに関する研究に比べると数は少な

いものの，カウンタースピーチについては，専門家が差別煽動と対になる言説を作成したデータセット [8] や，ドイツ語のカウンタースピーチ研究 [9] などが存在する．日本語の有害表現に関連する研究としては，ネットいじめに関するものをはじめ [10, 11]，有害表現データセット [12] や Twitter に含まれる日本語のヘイトスピーチデータセット [13]，人権侵害表現データセット [14]，常識道徳データセット [15]，など様々な日本語データセットの構築が進められているが，有害表現の間接的抑止を目指した研究は管見の限り存在しない．

ホープスピーチ研究では，インド・パキスタン間の紛争に関する YouTube コメントを資料としたデータセット [16], 性的少数者の権利，Black Lives Matter 運動 (BLM)，女性の理系分野での活躍といった多様なテーマに関する YouTube コメントからなるデータセット [17, 18]，そして内容を問わずに Twitter テキストデータをもとにしたデータセット [19] が存在する．

[16, 17, 18] における「ホープスピーチ」の基準は，一般的ないし価値中立的な心理的態度としての希望ではなく，何らかの理念を推奨，推進するという観点から見た「希望」概念に依拠する．たとえば [17] は，「平等性・多様性・包摂性」のスローガンに現れる理念を推進する言説を「ホープスピーチ」とみなす．

情報空間に好影響を与える可能性があるものとして，理念的ホープスピーチの研究は重要であるが，これらの研究における「ホープスピーチ」基準は明らかな定義の循環を含み（例「希望とは . . . 希望をうながすもの」「. . . 希望的未来について語っているもの」），判断を困難にするあいまいな語句が多数含まれる（例「インスピレーション」「洞察」）．実際，[17] が提供するテキストデータを細かく見たところ，“All lives matter” といった, 文脈を踏まえると, BLM への保守的反動としてむしろ包摂性を否定する言説なども Hope とラベルづけされている．

一方 [19] は、先行するホープスピーチ研究を批判し，より価値中立的な，単なる心理的態度の表明として分析される「ホープスピーチ」基準を提案する．これは，哲学文献 [4] において標準的な，希望がある種の信念と欲求から構成されるとする分析的定義と非常に近く，基準適用が容易になると見込まれ，実際，アノテーション一致率が改善されたと報告されている．

しかしながら，[19] の基準は，哲学的文献で議論されてきた希望と絶望を区分する課題 [20] を克服できていない．そこで本研究では，哲学文献における「希望」概念を参考としつつ，[19] の基準を改変したものを採用する．

3 データセット構築手法

3.1 “Hope” 基準とアノテーション

本研究で使用する「希望」概念は以下のように特徴づけられる．そのような意味での希望を表すようなコメントを “Hope” そうでないものを “Non-Hope” とラベルづけする，というのがアノテーション基準となる．

- **何らかの結果や状態についての希望**
  1. その結果や状態が実現することは保証されていないが，（可能性が小さくとも）実現しうると考えている．
  2. その結果や状態の実現を望んでいる．
  3. その結果や状態が実現すると考えることに，気分を明るくさせたり，実現に向けて何かをしようと思わせたりといった前向きな態度がともなわれる．

(1) と (2) は [19] における基準と類比的で，内容中立的に希望的言説を取り出している．たとえば「早く戦闘が終わるように心から願います」も「早く敵が殲滅されますように心から願います」もどちらも Hope と分類される．また，(3) により，「戦争が早く終わればいいがどうせ何をしても無駄だ」といった，同一の選好についてのものではあるが，単なる諦念や絶望の表明となっているものや，「がんばれ」，「負けるな」といった常套句や挨拶の定型も除外される．

研究グループのうちの哲学者・倫理学者の 2 名が，以下で述べる投稿データすべてを確認し，基準に従いラベルを付与した．2 人の判断が一致したものをその投稿の最終的なラベルとした．

3.2 対象データ

本研究では [16] と同じくホープスピーチのトピックを絞り，戦争や紛争・テロ行為に関する YouTube ニュース動画に対して投稿された日本語コメントを収集した．

事前調査においては，Yahoo ニュースにおける

ユーザーコメントを分類する可能性を探ったが，それらのコメントは評論的作文，意見の開陳が多く，情報伝達を超えて他者に影響を与えうるホープスピーチとみなせるものがわずかであったため，より他者への呼びかけに近い YouTube コメントを素材とした．

YouTube 上で，日本語で視聴できるニュース動画配信媒体（BBC News Japan，ANNnewsCH，テレ東 BIZ，TBS News DIG Powered by JNN，日テレ NEWS，FNN プライム）の投稿動画から関連する動画を選択し，ブラウザ上から直接コメントを収集した．収集期間は 2023 年 11 月 22 日から 12 月 8 日であり，もっとも古い動画が 2021 年 5 月 12 日，もっとも新しい動画が 2023 年 11 月 21 日に投稿されたものである．多くがイスラエルによるパレスチナ自治区ガザ地区への侵攻，あるいはロシアによるウクライナ侵攻に関するニュース動画であった．なお，動画に直接向けられた投稿のみをデータとし，投稿への投稿は文脈の把握が複雑になるため除外した．

結果 62 個の動画から，10,137 投稿を取得した．ユーザー名数は 8322 で，10 回以上投稿をしていた同一ユーザーは 17 名しかなく，最大投稿数は 22 であった．17 ユーザーのうちひとつは，固定投稿を行う動画投稿企業であり，他に明らかなボットは含まれていない．ただし，17 のうちの 3 つは，ほぼ重複した内容でそれぞれイスラエル（1 名）とウクライナ（2 名）を応援する投稿のみを行っていた．また，2 回以上重複するコメントが 123 コメント（上記企業投稿を除く）含まれている．これらの多くは，収集の際の単純ミスと思われるが，少なくとも一部は，同一ユーザーが異なる動画にまったく同じコメントを投稿したものも含まれる．なお，ユーザー名は，プライバシーへの配慮からデータセット自体からは削除し，上記アノテーション作業においても使用していない．

4　データセットの分析と考察

4.1　データセットの概要

アノテーター 2 人の判断が一致したコメントは 9956/10137（約 98.2 %），判断が一致したラベルにおける Hope の割合は 2.67%であり（Hope/Non-Hope: 271/9685），YouTube コメントに自然と含まれるホープスピーチの割合としては，[16] における数値 2.45%と類似的であった．

「希望」概念を上述のように構成要素に分析した結果，高い一致率が達成されたと考えられる．先行研究ではアノテーターが研究者でないため単純な比較はできないが，[17] における一致率（Krippendorff's alpha) は 0.63（英語データセット），0.76（タミル語），0.85（マラーヤラム語）であり，[19] では 0.85（英語データセット）とされている．

4.2　内容の分析と考察

本研究では信念と欲求の複合物としての「希望」概念を採用したため，Hope と分類されたコメントには，「〜してほしい」「〜を望みます」といった選好にまつわる類似的な構文が多く見られた．KH Coder 3 による形態素解析によると，Hope コメントでは「平和」と「戦争」が頻出語 1 位と 2 位であり，それぞれ平和の実現，戦争の終結を希望するコメントに関連していた．一方で，内容中立的に選んでいるため，平和を希望するコメントだけでなく，戦争の継続や，それぞれの国や組織の勝利を希望するコメントも含まれていた．

以上のような Hope コメントは出現率が非常に小さいだけでなく，印象論ではあるが，コメント全体の有害性 (toxicity) が非常に高かった．Non-Hope の中には，「勉強になった」といった動画に対する感想，事実関係への論評的コメント，また，「日本は平和でよかった」「中東の争いは永久に終わらないでしょうね」といった「他人事」の意思表示をするコメントも数多く含まれており，これらは有害とはみなせない．しかし，それ以外にも，ネットミーム的やりとりを含む冗談／不謹慎なコメント，女性出演者の容姿に関する侮蔑的コメント，「自業自得」「何の意味もない」といった冷笑や揶揄のコメント，また数は多くないが韓国などに対する差別的発言まで含まれていた．

以下に Hope/Non-Hope それぞれの事例をいくつか示す．

Hope ラベルの事例

- どうかこれ以上犠牲者が出ない事を祈ります．記者の方々も気を付けて下さい． 無事に帰国される事を祈ってます
- 国連が，責任を持ってエジプトへの避難民の受け入れを推して欲しい
- １日も早く停戦すること願うばかりです

Non-Hope ラベルの事例

表 1 GPT-4 の予測とアノテータの判断との一致率 (%).

設定	全体	Hope	Non-hope
zero-shot	97.5	36.9	99.2
few-shot	97.5	49.1	98.9

- こりゃ大変だ！！ でも私は本日のお昼ごはんの方が優先課題である ごめんなさい
- もっとやれい (*・ω・) ノウクライナにかけてんだよ！
- 鳩山，宗男「・・・・・・・・・・・」
- ハマスを選んだ民衆の自業自得で草 www
- 強さが正義ですからね
- 他国の戦争をネタにして注目されたいだけの日本人
- じゃあアンタらが現地に行って止めに行けばいいでしょ！戦争は簡単にはなくなりませんよ！平和な国で抗議した所で何の意味もないです！

当然ながら，Non-Hope の具体例には差別的コメントや過激な表現を含めていない．公共性の高いニュース動画のコメント欄でありながら，実際の有害性はより高いケースが見られる．少なくとも公共性の高いオンライン空間において，有害性を調整する手段として，ホープスピーチの出現を何らかの形でコントロールする手法が今後検討されるべきである．

4.3 自動検出の可能性と課題

近年では，大規模言語モデルが自然言語処理の様々なタスクで優れた性能を示している．そこで，OpenAI が提供する GPT-4 (gpt-4-1106-preview) の API[1)] を用いて，大規模言語モデルによるホープスピーチ自動検出の可能性を検討した．

ここでは与えられたコメントに対して，Hope の基準に則して Hope であるか Non-Hope であるかの 2 値を予測する分類タスクとして，ホープスピーチ検知タスクを定義する．本研究では構築したデータセットのうちアノテーターの判断が一致している 9956 件のコメントに対して，タスクの指示のみをプロンプトとして与える zero-shot の設定と，プロンプトに加えてデータセットに含まれない 2 例（Hope/Non-Hope の例を 1 例ずつ）を正解例として与える few-shot の設定という 2 種類の設定で，GPT-4 の予測性能を比較した．実験に使用したプロンプトと正解例は付録に示す．

表 1 に GPT-4 の予測とアノテータの判断との一致率を示す．結果を見ると，一般的には zero-shot よりも few-shot の設定の方がモデルの予測性能が上がることが期待されるが，全体の一致率は zero-shot と few-shot とで差が見られない結果となった．

しかし，ラベルごとの一致率を見ると，few-shot の設定で Hope の予測性能は向上している．これは今回構築されたデータの大部分が Non-hope であるため，全体の一致率では差が見られなかったが，少数の訓練事例を与えることでホープスピーチの検知性能が上がる可能性を示している．

一方で，Non-Hope の予測性能が 9 割を超えているのに対して，Hope の予測性能が 3-4 割程度であるという結果から，Hope の基準を正確に理解してホープスピーチの検知を行うことは大規模言語モデルにおいて挑戦的なタスクであることが示唆される．

5 おわりに

本研究では，YouTube ニュース動画へのコメントにもとづいた，日本語ホープスピーチのデータセットを作成し，分類されたテキストの内容分析と GPT-4 を用いた評価実験を行った．

本データセットはおよそ 1 万件のコメントにより構成されているが，自然に出現する 300 件弱のホープスピーチのみを含み，先行研究と比較するとホープスピーチの数が少ない．また，トピックを戦争や紛争にまつわるものに絞ったため，数と内容ともに今後拡張させることが望ましい．また，ホープスピーチの優先的表示ないし自動生成が，他のユーザーにどのような影響を与えるのか実証的な研究が必要である．

ホープスピーチの自動検出や生成技術は，有害コンテンツ規制に比べ，その運用の法的・倫理的ハードルは低くなると予想される．公開動画の公開コメントのみから構築された本データベースが，それらの技術開発のきっかけになること，また，近い将来において，ホープスピーチの研究が進み，オンライン情報空間の有害性が減少することを強く希望する．

1) https://platform.openai.com/docs/guides/text-generation (2024 年 1 月 12 日 参照)

謝辞

GPT-4 を用いた分析について，杉本智紀氏（東大）のサポートに感謝します．本研究は，JSPS 科研費 JP22K00020 の助成を受けたものである．

参考文献

[1] 和泉悠, 仲宗根勝仁, 朱喜哲, 谷中瞳, 荒井ひろみ. Ai はレイシズムと戦えるのか–自然言語処理分野におけるヘイトスピーチ自動検出研究の現状と課題. 思想, Vol. 1169, pp. 88–105, 2021.

[2] C. R. Snyder. Hope theory: Rainbows in the mind. **Psychological Inquiry**, Vol. 13, No. 4, pp. 249–275, 2002.

[3] C. R. Snyder, Kevin L. Rand, and David R. Sigmon. Hope Theory: A Member of the Positive Psychology Family. In Matthew W. Gallagher and Shane J. Lopez, editors, **The Oxford Handbook of Hope**, pp. 27–44. Oxford University Press, 2018.

[4] Claudia Bloeser and Titus Stahl. Hope. In Edward N. Zalta, editor, **The Stanford Encyclopedia of Philosophy**. Metaphysics Research Lab, Stanford University, Summer 2022 edition, 2022.

[5] Herman Cappelen and Josh Dever. **Bad Language**. Oxford University Press, Oxford, 2019.

[6] Jason Stanley. **How Propaganda Works**. Princeton University Press, New Jersey, 2015.

[7] David Beaver and Jason Stanley. **The Politics of Language**. Princeton University Press, Princeton, 2023.

[8] Yi-Ling Chung, Elizaveta Kuzmenko, Serra Sinem Tekiroglu, and Marco Guerini. CONAN - COunter NArratives through nichesourcing: a multilingual dataset of responses to fight online hate speech. In **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics**, pp. 2819–2829, Florence, Italy, 2019. Association for Computational Linguistics.

[9] Joshua Garland, Keyan Ghazi-Zahedi, Jean-Gabriel Young, Laurent Hébert-Dufresne, and Mirta Galesic. Countering hate on social media: Large scale classification of hate and counter speech. In **Proceedings of the Fourth Workshop on Online Abuse and Harms**, pp. 102–112. Association for Computational Linguistics, 2020.

[10] 松葉達明, 桝井文人, 河合敦夫, 井須尚紀. カテゴリ別関連度最大化手法に基づく学校非公式サイトの有害書込み検出. 言語処理学会第 17 回年次大会発表論文集, pp. 388–391, 2011.

[11] 新田大征, 桝井文人, Ptaszynski Michal, 木村泰知, Rzepka Rafal, 荒木健治. カテゴリ別関連度最大化手法に基づく学校非公式サイトの有害書込み検出. 人工知能学会全国大会論文集, Vol. JSAI2013, pp. 2039–2039, 2013.

[12] 小林滉河, 山崎天, 吉川克正, 牧田光晴, 中町礼文, 佐藤京也, 浅原正幸, 佐藤敏紀. 日本語有害表現スキーマの提案と評価. 言語処理学会第 29 回年次大会, pp. 933–938, 2023.

[13] 荒井ひろみ, 和泉悠, 朱喜哲, 仲宗根勝仁, 谷中瞳. ソーシャルメディアにおけるヘイトスピーチ検出に向けた日本語データセット構築の試案. 言語処理学会第 27 回年次大会, pp. 466–470, 2021.

[14] 久田祥平, 若宮翔子, 荒牧英治. 権利侵害と不快さの間：日本語人権侵害表現データセット. 言語処理学会第 29 回年次大会, pp. 363–368, 2023.

[15] 竹下昌志, ジェプカラファウ, 荒木健治. Jcommonsensemorality: 常識道徳の理解度評価用日本語データセット. 言語処理学会第 29 回年次大会, pp. 357–263, 2023.

[16] Shriphani Palakodety, Ashiqur R. KhudaBukhsh, and Jaime G. Carbonell. Hope speech detection: A computational analysis of the voice of peace, 2020.

[17] Bharathi Raja Chakravarthi. HopeEDI: A multilingual hope speech detection dataset for equality, diversity, and inclusion. In **Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in Social Media**, pp. 41–53, Barcelona, Spain (Online), December 2020. Association for Computational Linguistics.

[18] Bharathi Raja Chakravarthi. Hope speech detection in youtube comments. **Social Network Analysis and Mining**, Vol. 12, No. 1, p. 75, 2022.

[19] Fazlourrahman Balouchzahi, Grigori Sidorov, and Alexander Gelbukh. Polyhope: Two-level hope speech detection from tweets. **Expert Systems with Applications**, Vol. 225, p. 120078, 2023.

[20] Ariel Meirav. The nature of hope. **Ratio**, Vol. 22, No. 2, pp. 216–233, 2009.

付録 GPT-4 を用いた実験のプロンプト

プロンプトは以下を使用した.

次のコメントが Hope か Non-hope か回答してください。それ以外には何も含めないことを厳守してください。
制約：何らかの結果や状態に対して 1, 2, 3 の全てに合致するコメントであれば Hope、そうでなければ Non-hope と回答してください。
1 その結果や状態が実現することは保証されていないが、（可能性が小さくとも）実現しうると考えている。
2 その結果や状態の実現を望んでいる。
3 その結果や状態が実現すると考えることに、気分を明るくさせたり、実現に向けて何かをしようと思わせたりといった前向きな態度がともなわれる。

few-shot の設定では，上記のプロンプトに以下の Hope/Non-hope の例を 1 例ずつ追加した．どちらの例も評価データには含まれない例であり，研究グループの哲学者・倫理学者によって Hope の基準に基づいて作成されたものである.

コメント：貴重な取材ありがとうございます。 どうかご無事で帰ってきてください。
回答：Hope
コメント：怪我などしていないことを祈りますが、難しいでしょう。
回答：Non-hope