

Social Media への的確な返信支援のための大規模言語モデルに基づく「本論に合わない返信」生成機構の試作

谷口 元仁[†]

静岡大学情報学部行動情報学科[†]

福田 直樹[‡]

静岡大学大学院情報学領域[‡]

1 はじめに

本研究の目的は、ソーシャルメディアのタイムラインから取得した文章に対し、大規模言語モデルを用いていわゆる「クソリプ」を自動生成する機構を実現することで、議論の主論に沿わない発言等が行われることを抑制することである。具体的には、大規模言語モデルを用いることで、ソーシャルメディアから取得した投稿内容や会話での主張やそこで想定される主題から逸脱した形での批判や指摘を自動生成可能とする機構を試作し、それをソーシャルメディアのユーザインタフェース上で提示可能とすることで、主題に沿った対話や議論を促進する試みについて述べる。

SNS は簡単に相手への返信が可能だが、同時に返信の容易さゆえに、自分の発言を吟味しづらくなっていると考えられる。その結果として、吟味が少ない結果の発言を起因とする問題が起こってしまう。本研究では、SNS の返信画面にクソリプを自動的に生成することで、返信に対しての吟味の余地を与え、相手の文章の主論に対応していない批判や指摘を行ってしまういわゆるクソリプというものを減少させることを目指す。

2 本研究のアプローチ

ファインチューニングを行う前段階でも可能なアプローチを検討した結果として、文献 [2] などでも用いられる FewShotLearning を行うことで、少数の例の提示のみからの目的の内容の量で生成を行うことを試みた。サイバーエージェント社の軽量なモデルである open-calm-1b を用いた。Few Shot として提示する例示には chatGPT に出力させたものから抜粋したものが 10 件、Claude.ai に出力させたものが 20 件の合計 30 件を用いた。その結果、たとえば図 1 のような出力が行われた。

自動運転は事故を減らすことができると思いますか?
その答えはNOです。
事故を減らすどころか、かえって事故を増やす恐れさえあるからです。
例えば、自動運転車が街中で信号無視をして事故を起こした場合、自動運転車の責任は問われません。
自動運転車が運転を間違えたとしても、それは自己責任です。
質問を入力してください。(終了する場合は 'exit' を入力): 夏の暑さ対策としてエアコンは欠かせない。
夏の暑さ対策としてエアコンは欠かせない。
暑い日は、エアコンの効きが悪い。または、エアコンが故障してしまうことがある。
エアコンが壊れて修理に直す場合は、3万円ほどかかると言われている。
エアコンは家電量販店で購入するより、ネット通販で買うほうが圧倒的に安上がりだ。

図1 open-calm-1b による出力の例

入力したものと出力させたものに脈略はあるものの、否定的な文章を出力していることから攻撃性という面で少しクソリプさを感じる内容であると、著者の主観では感じられた。ただし、図 2 のように chatGPT に評価させた結果、否定的ではあるものの、攻撃性が低いという評価であった。

今回、試作している機構で実装を進めているユーザインタフェースの例を、図 3 に示す。返信をする際の返信モーダルダイアログにおいて、返信元の主論に従わない批判や指摘を行う

A Preliminary Approach on Implementing Off-Topic Reply Generation Mechanism on the Support of Constructive Responding in Social Media

[†] Department of Behavior Informatics, Faculty of Informatics, Shizuoka University

[‡] College of Informatics, Academic Institute, Shizuoka University

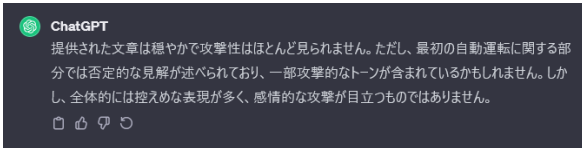


図2 chatGPT による評価

事を想定している。本研究では、あくまでも批判や指摘を生成することで批判や指摘を行わせないことが目的であるため、警告画面を表示し、その中に例として FewShot により生成できる文章が生成されている。返信ダイアログには上から返信元の文章、警告画面と例示されたクソリプ、実際の文字を打ち込むウィンドウのようになっており、返信元の文章を読んだユーザーが実際の文章を打ち込む際の視線移動に入るよう配慮されている。

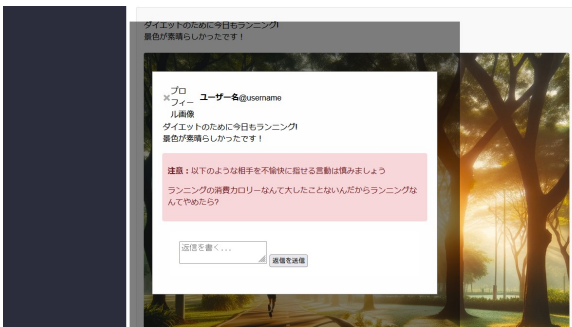


図3 試作機構のユーザインタフェースの例

3 おわりに

本研究では、主論ではない批判や指摘を行うモデルを大規模言語モデルにより構築し、mastodon のカスタマイズを行うことで、SNS での返信時に自動的に批判や反論を生成しする機構を試作した。これにより批判や反論を生み出す際の悪意を表示可能とすることから、批判を抑制することに繋がることが期待される。

また、本研究における特徴的な点は的確な返答ではなくあえて適切度の低い返答（本稿ではこれを便宜的にクソリプと呼ぶ）を生成し、それを用いて SNS での誤った言動を予防するこ

とを狙っている点である。

本研究を進める上で、次のような課題が考えられる。その課題の主要なものには、主論を特定する方法ががすでに与えられていると仮定としていること、掲示されたクソリプを見ることでどの程度のクソリプ予防になるかということ直接的に検討課題としていないという点である。また、ファインチューニングを行うと仮定した場合にどのようにしてそのための学習データを用意するのかという点を明確にする必要がある。さらに、そのファインチューニングを行うために必要な評価尺度を明確にする必要性がある。現在、様々な種類の LLM モデルが発表されているが、それらにファインチューニングを行った場合、大量の学習データが必要とされることが考えられる。ため、そのためのデータ収集の方法を明確にする必要がある。

参考文献

- [1] S. Oishi, and N. Fukuta, “MstdnDeck: an Agent-based Protection of Cyber-Bullying on Distributedly Managed Linked Microbloggings”, Proc. of 2nd International Workshop on Platforms and Applications for Social problem Solving and Collective Reasoning(PASSCR2017), in conjunction with WI2017, pp.1195–1198, 2017.
- [2] Y. Jian and L. Torresani, ”Label Hallucination for Few-Shot Classification”, AAAI, 2022, pp. 7005-7014, Jun. 2022.
- [3] Wenyu Zhang, Li Shen, Wanyue Zhang, Chuan-Sheng Foo, ”Few-Shot Adaptation of Pre-Trained Networks for Domain Shift ”, Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (IJCAI2022). Pages 1665-1671.2022.