AI が生成する不採択コメントを用いた研究指導

宮下芳明* 大友千宙*

概要．本稿では，AI の文章生成言語モデルによって論文の概要から「不採択コメント」を生成し，それに学生と教員が一緒に向き合うことによる研究指導を提案する．実装したシステムは Slack Bot であり，文末に「なーんてな」といった言葉を追加することによって表現の威圧感を緩和して，実際に学生が読んだときの精神的なダメージを低減する．動作テストでは，10 の不採択コメントで役に立ったと思える点があるものは 7 つで，中には新しいアイデアまで提供するものがあった．これを実際に当研究室での研究指導に活用する計画である．

1 はじめに

1.1 不採択コメントの生成

OpenAI 提供の文章生成言語モデル GPT-3[1]を用いて，「不採択コメントの文章生成」を行う．

以下の概要をもつ論文に対する不採択の査読コメントを生成せよ

WISS2021 で採択された本稿第一著者の論文[2]の概要を上記文章に付して生成させたところ，以下の不採択コメントが得られた．(各種パラメータは後述．和英・英和翻訳 DeepL API[3]を使用)

この論文にはいくつかの懸念がある．まず，味覚ディスプレイのプロトタイプが実際にどの程度機能するのかが不明である．著者らは，プロトタイプをテストしたデータや結果を提供していない．第二に，味覚ディスプレイの潜在的な応用が十分に検討されていない．著者らは，味覚ディスプレイがさまざまな種類の味を再現するために使用できることに簡単に触れているだけである．味覚ディスプレイが実世界でどの程度実用的で有用なものであるかは，明らかでない．最後に，味覚ディスプレイのコストと実現可能性についても言及されていない．著者らは，味覚ディスプレイの製造コストや大量生産が可能かどうかについては言及していない．全体として，著者らは味覚ディスプレイプロトタイプに関する主張を支持する十分な証拠を提供していない．著者が上記の懸念を払拭するまでは，本論文の掲載を認めるべきではない．

本論文の査読結果はウェブで公開されている[4]ので，人間（専門家）による（論文本文の）査読結果と比較することができる．偶然にも，有用性を示す評価実験がない点など，指摘事項が似ていることがわかる．もしこの不採択コメントを当時受け取っていたら，本稿第一著者はその評価を真摯に受け入れ，それどころか，コメントに沿った追実験や加筆を行っていたのではないかとすら考えている．

* 明治大学

1.2 指導する院生の感想

さらに，本稿第一著者が指導する院生が投稿予定の論文概要を，本人同意の上で入力し，不採択コメントを生成して本人と一緒に読むこととした．それを読んだときの院生の最初の感想は「これは傷つきますね」であった．たしかに，GPT-3 の優秀さゆえ，査読コメント特有の「威圧感」も再現されてしまっているように感じられる．ただしその院生からは，

① 人間でなく AI による生成である点
② 採否判定をさせたわけではなく意図的に不採択コメントを生成させている点
③ 論文全体ではなく概要だけの情報をもとに生成されている点

から「実際の不採択コメントに比べれば精神的ダメージははるかに少ない」という感想を得た．

GPT-3 では，同じ入力文でも毎回生成結果が異なる．精神的ダメージが少ないことを確認できたため，本人の同意のもとでさらに 10 種類の不採択コメントを生成し，一緒に読んでディスカッションを行うことにした．それらの生成結果には，論文を要約しただけのものや全く的外れなものも混じり，「これは査読にもなってない」「この査読者は全く論文を読んでない」「AI なのにこれは意外に良い指摘をしている」など，まるでこちらが査読コメントを審査しているような気持ちになり，笑顔や冗談も出るほどポジティブに研究の議論をすることができた．

本稿第一著者はこれまで 15 年間，学生・院生達と共著で査読付きの学会や論文誌に投稿してきたが，不採択コメントをこのような気分で見られたことはない．怒りや落胆に満ちた学生達をうまく慰めたりなだめたりし，軌道修正・再投稿に向けてエンカレッジしていくことに，とても苦労してきた．それに失敗して学生が研究活動をやめたことすらある．

不採択コメントというのは，査読者が労力と時間

をかけて論文を客観的に審査した末に，不備を列挙して，なぜよくないのか，どうすればよくなるのか，といったことをわかりやすくまとめたものであり，貴重な教育的価値があるはずである．採択コメントの方が概して文章量が少ないので，むしろ不採択時こそ，そうした報酬が多く得られていると言っても過言ではない．しかし，そういうポジティブな気持ちで不採択コメントを読める学生はなかなかいない．

では，AI が概要のみに基づいて生成した不採択コメントに，そういった教育効果があるのかを見てみる．まず，わずかながら，指導教員として指摘したいが学生本人には言いにくい，研究の本質的な弱点を言い当てているものがあった．これを学生本人に伝える場合には，指導教員という立場上，言い方やタイミングにものすごく気を遣うであろう．

また，そういう「鋭い」不採択コメント以外は無用かというと，そうでもないと感じた．それらのコメントは，いわば「ろくに論文を読まず，概要を読んだ程度の印象に基づいて，ありがちな批判を書き連ねる査読者」のような趣きがあった．そういう査読者に誤解されることのないように，論文の第一印象をブラッシュアップしようという気持ちにはなる．たとえば，このときの院生の論文は，綿密な評価実験を行って強固な実証を行ったものであったが，生成された不採択コメントには「評価実験が行われていない」というものがあった．たしかに概要文には評価実験の実施や結果まで詳細に書いていなかったので，加筆することにした．

このようなことから本稿では，「AI が生成する不採択コメントを用いた研究指導」に可能性を感じ，システムを試作した．対象ユーザは研究室の学生と教員であり，ＡＩによる不採択コメントに対する反論を一緒に考えることで研究指導につなげることを意図している．生成される不採択コメントが，何かしらその研究に有益な指摘をもたらすこと，そしてそれが学生に精神的ダメージを与えないことを狙ってシステムをデザインした．

2　システム

本稿で試作したシステムは，論文の概要を入力するとそれに対する不採択コメントを返す Slack bot である．システムの動作例を図 1 に示す．ユーザは本ボットに対して論文の概要をメンションと共に送信する．すると，入力した概要に対する不採択コメントをメッセージ形式で受け取ることができる．

本システムは AI であることを明示するよう名前にも AI と付す．アイコンも，画像生成 AI の Stable Diffusion [5]で「コミカルでダメそうなかわいいロボットのイラスト」と指示して生成した画像を用いた．意図的に不採択コメントを生成していること，概要だけの情報をもとにしていることを明示した．複数の査読結果が出ている方が精神的ダメージが少ないという内観から 10 種類の査読コメントを生成した．また，文章の不快感レベルを制御する手法[6]から腰砕け手法（文末に「なーんてな」といった言葉を追加する手法）を用いることとし，威圧感の低減を狙った．翻訳に対しては DeepL の翻訳がそれなりに正確に訳されており，その一方で，ですます調とである調が混在するなど，文章が少しカタコトになる点も，精神的ダメージを軽減させるということがわかったので，このまま採用した．

図 1 システムの動作例

OpenAI API の Usage guidelines[7]によると，ユーザに AI とやり取りしていることを明示する必要がある．そのため本ボットは Slack 上において通常のユーザと区別されて表示される．また本ボットを使用するユーザには事前に AI によって生成されたコメントであると周知する．GPT-3 への入力は英語のため，DeepL API で翻訳を行った後に GPT-3 に入力している．翻訳された入力の冒頭に，「以下の概要をもつ論文に対する不採択の査読コメントを生成せよ」という一文を英訳したものを付与し入力とする．パラメータは以下の通り．Model: text-davinci-002, Temperature: 1.0, Maximum length: 150, Top P: 1, Frequency penalty: 0, Presence penalty: 0, Best of: 1, Inject start text: ON, Inject restart text: ON, Show probabilities: OFF.

ユーザからの入力と生成されたコメントは，処理前に適切なものか否かを確認すべく，Moderation endpoint[8]を用い，暴力的，性的，増悪的，自傷的，または OpenAI のコンテンツポリシーに違反していないかを判定する．論文[9]をもとに不採択コメントを生成したところ，総じて，指摘として正しい，役に立ったと思えるコメントは 7 つで，中には新しいアイデアまで提供するものがあった．一方で 3 つ，全く役に立たない査読コメントがあった．

「腰砕け手法」がとても効果的であり，文末に「なんつって」「たぶん」と書かれているだけで，重く受け止めずに軽く読めるようになっている．

有用そうなコメントが得られ，精神的ダメージも少なそうだということから，これを実際の当研究室の研究指導に用いる計画である．

参考文献

[1] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P.,Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., M.Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., & Amodei, D. (2020). Language models are few-shot learners. Advances in neural information processing systems, 33, 1877-1901.

[2] 宮下芳明. 液体噴霧混合式の味ディスプレイの試作, 第 29 回インタラクティブシステムとソフトウェアに関するワークショップ(WISS2021)論文集, pp.121-127, 2021.

[3] DeepL API. https://www.deepl.com/ja/pro-api?cta=header-pro-api（2022/09/24 確認）

[4] 液体噴霧混合式の味ディスプレイの試作 査読結果. https://www.wiss.org/WISS2021Proceedings/data/18.html（2022/09/24 確認）

[5] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 10684-10695).

[6] 大家眸美, 宮下芳明. 文章の不快感レベルを制御する手法群とその実装，インタラクション 2013 論文集, pp.550-555，2013.

[7] Usage guidelines (responsible use), Open AI. https://beta.openai.com/docs/usage-guidelines (2022/09/24 確認)

[8] Markov, T., Zhang, C., Agarwal, S., Eloundou, T., Lee, T., Adler, S., A, Jiang., & Weng, L. (2022). A holistic approach to undesired content detection in the real world. arXiv preprint arXiv:2208.03274.

[9] 宮下芳明，青木秀憲． Ha & Fu：スマートフォンに息を吹きかけるポインティングインタフェース，第 29 回インタラクティブシステムとソフトウェアに関するワークショップ(WISS2021)論文集，2021.

[10] 永瀬翔，栗原一貴，宮下芳明. チームプレゼン！，第 19 回インタラクティブシステムとソフトウェアに関するワークショップ（WISS2011）論文集，pp.221-223，2011.

未来ビジョン：AI に悪役を担ってもらう価値

当研究室では十年以上前から，「汎用的いじわる質問」という文書が共有されており，学会発表や学位審査前に活用されている．「〜なんだけど，…？」というスタイルの文章で，〜で傷つくことを言って前置きしたのちに，…で本質的な質問をするという，学生を泣かせるほど厳しいかつての理工系文化にのっとった質問である．一部を紹介すると，「僕は全く面白くないと思うんだけど，この研究で何が嬉しいの？」「誰にも役立たない提案のような気がするけど，ターゲットユーザーは誰のつもりなわけ？」「はっきり言ってどこにでもありそうな内容だけど，一体何が新規なの？」といった質問群である．

このような文書を用意した経緯は，当研究室の学生がかつてこのような質問をされて答えられないことがあったためである．もちろん，そのような威圧的な質問がされるような発表会や学会はなくなってほしいし，現にかなり減っている．が，万一そのような質問がきてもさらっと笑顔で答えられるようになってほしい，それが学生たちを守ることにもなると考えて，研究室内で共有している．なぜ文書にしたかというと，そもそもこんな嫌味な質問は，たとえ発表練習であっても自分の口からは言いたくないからである．あえて「汎用的」にし，同じ文書を更新せずに毎年使いまわしているのもポイントで，特定の誰かの特定の研究を指して言っているわけではないことの証明にもなると考えた．研究テーマによっては，これらの質問が当てはまらないときもあるので，その「ずれ」もちょうど良いと思っている．かつては質問群をまとめてコピー&ペーストして Slack に貼っていたが，それすら行うのも気が引けるので，「Google ドライブで汎用的いじわる質問を検索して，返答を用意しておいてください」の一言で済ますようになっている．

指導教員と学生という関係で共に研究するのは，とても気を遣う．研究を始めたばかりの学生は，研究内容について指摘されただけで，自分自身が否定されたと誤解することすらある．プロの研究者はそこまでではないだろうが，それでも，対面での議論で何かを指摘するのはやはり気を遣う．WISS のようなワークショップで活発に議論を行うと言っても，根本的・致命的な質問をするのには躊躇するし，査読者でなければそれを指摘する義務があるわけでもないので，あえて触れないこともある．

一方で，共著で研究している場合には，最後は共にその研究の責任を負うし，研究者としての信用問題にもつながりかねないので，おかしいと思うところは，しっかり共著者間で指摘しあわなければならない．最終的にその研究の価値を外に訴えていく「戦友」となるためにも，投稿前までは互いに意見をぶつけ合えるぐらいの関係が望ましいだろう．もちろん，ときには最後まで平行線となる主義主張もある．本稿第一筆者らもかつて，WISS の未来ビジョン欄を利用し，共著者間で見解が異なるところをあえて分筆したことがある[10].

ただ，こうした共同研究者同士の指摘や議論に，学生-教員という立場の差がある場合には，気を遣わざるをえない．対面で発話する場合には，優しく笑顔で，穏やかに和やかに話すように心がける．Slack やメールの場合は，きつい表現がないか確認し，文章表現を何度も推敲し，エンカレッジする書き出しで始めたり，文末に（笑）などの表現を加えて和ませたりする．指摘のタイミングも重要である．先輩など他のメンバーによるフォローや指摘分担を頼むときもある．どの研究室でも，研究指導の裏には常にこのような努力があると思われる．人間関係・協力関係を維持しながら，それでも指摘すべきところは指摘しなければならない．…こうしたジレンマで苦しんでいるときに，厳しい指摘を他人がしてくれたらどんなに嬉しいかと思う．悪役を誰かが背負ってくれたら，まず自分が嫌悪されずに済む．誰かが悪役になってくれれば，それに対して共に戦う同志として，結束すらできる．本稿は AI にその悪役を担ってもらう研究だといえる．学生-教員間の人間関係・協力関係の維持にも貢献できるのではないかと考えている．

なお，本稿をきっかけに，研究の効率化・自動化，研究指導の効率化・自動化，査読の効率化・自動化といった応用が他にも考えられるかもしれないが，著者らはそれに対してはあまり興味がない．本研究のモチベーションが効率化や自動化と異なることを，本章で紙面を割いて説明した．同一の文脈で評価されることが万一あれば，それこそ本文を読んでいないずれた評価だと考える．