筑波大学
University of Tsukuba

2025 年 4 月 17 日

報道関係者各位

国立大学法人筑波大学

大規模言語モデルによりオンラインチャットの会話脱線予測が可能に

オンラインチャットでは、会話の内容が本来のトピックから逸脱し、個人攻撃に発展することがあります。本研究では、このような会話脱線現象を予測する手法として、大規模言語モデルとゼロショット予測を用い、プラットフォームごとに学習を行わなくても十分な精度で予測可能なことを示しました。

SNS や電子掲示板で行われるオンラインチャット上の会話には、ユーザーの有害なふるまいがしばしば観察されます。そこで、オンラインチャット上の会話が本来の会話トピックの内容から逸脱し、個人攻撃へと発展する「会話脱線現象」について、会話脱線を検出し、モデレーションするための予測アルゴリズムの開発が行われてきました。しかし、これまでの研究は、特定のプラットフォームのデータに依存して進められおり、他のプラットフォームへの応用には新しいデータセットの構築にコストがかかることが課題でした。

本研究では、大規模言語モデルを用い、学習データセットがなくてもデータを予測できる「ゼロショット予測」の手法を会話脱線現象の予測に適用しました。多様な大規模言語モデルを検証し、従来整備されたデータセットに対して、大規模言語モデルごとのゼロショット予測の精度の平均値と、学習データにより調整された深層学習アルゴリズムの精度を比較しました。

その結果、特定プラットフォームの学習を行っていない大規模言語モデルでも、従来開発されたアルゴリズムと同等以上の精度を達成することを示しました。これにより、プラットフォーム運営者が、会話脱線予測を低コストで実行できるようになり、さまざまなプラットフォームにおいて、会話モデレーションによるコミュニティの健全な発展が促されることが期待されます。

研究代表者
筑波大学ビジネスサイエンス系
吉田 光男　准教授
野中 賢也　リスク・レジリエンス工学学位プログラム 博士後期課程1年

研究の背景

SNS や電子掲示板で行われるオンラインチャット上の会話では、ユーザーの有害なふるまいが観察されます。特に、オンラインチャット上の会話が本来の会話トピックの内容から逸脱し、個人攻撃へと発展する「会話脱線現象」が問題となっていることから、会話脱線を検出し、モデレーションを行うための予測アルゴリズムの開発が行われてきました。しかし、これまでの研究は、特定のプラットフォームのデータに依存して進められおり、他のプラットフォームに応用するためには、新しいデータセットの構築にコストがかかることが課題でした。

研究内容と成果

会話脱線現象は、参考図の右側に示すように、会話が最終的に相手への明確な侮辱表現（wasting other people's time is simply rude：他人の時間を無駄にするのは失礼である）へと発展する現象です。この現象の発生を防ぐためには、最終的な発言をモデルの入力に加えずに、それ以前の発言から会話脱線を予測することが必要です。そのための手法として、本研究では、現在利用可能な主要な 5 つの大規模言語モデル（Large Language Model: LLM）[注 1)]を用い、学習データセットがなくてもデータを予測できる「ゼロショット予測[注 2)]」を適用し、その予測精度と、学習データにより調整された従来の深層学習アルゴリズムの予測精度を比較しました。

具体的には、米国の掲示板型ニュースサイト「Reddit」内のチャット機能「ChangeMyView」スレッド、および、Wikipedia の改善などについて利用者が議論する「Wikipedia Talk Page」で採集されたデータセットを対象に、会話脱線現象の予測精度の検証を行いました。

実験においては、クラウドで利用できる商用 LLM として、OpenAI 社の GPT-4o、Google 社の Gemini-1.5-flash を用いました。また、手元の計算機などローカルで動作するオープンモデル LLM として、Llama-3.1-70B-Instruct（4bit 量子化）、Gemma-2-9b-it、Gemma-2-27b-it（8bit 量子化）を検証しました。また、従来の教師あり学習モデルとして、Google 社が公開した BERT を各データセットにファインチューニングしたモデルをテストしました。その結果、利用した大規模言語モデルの予測精度の平均値が、従来の教師あり学習モデルの正解率を、Wikipedia データにおいて 4.8 ポイント、Reddit データにおいて 2.6 ポイント上回りました。また、LLM を使用する際のプロンプト（指示文）に「Reddit のスレッドである」「Wikipedia の議論である」といった事前知識を含めることで、モデルの精度自体は向上しなかったものの、議論の脱線をより早期に予測する傾向が顕著に現れることが分かり、予測タイミングに対してプロンプトが与える影響の大きさが明らかになりました。

今後の展開

本研究成果は、プラットフォーム運営者が、会話脱線予測を低コストで実行することを可能にし、さまざまなプラットフォームにおいて、会話モデレーションによるコミュニティの健全な発展が促されることが期待されます。今後さらに、会話脱線の予測のみならず会話脱線発生のメカニズムの解明が求められます。例えば、なぜ会話が個人攻撃へ陥ってしまうのかを、会話のテキストスタイルなどから説明することができれば、会話モデレーション時の明確な方針を作ることができます。

参考図

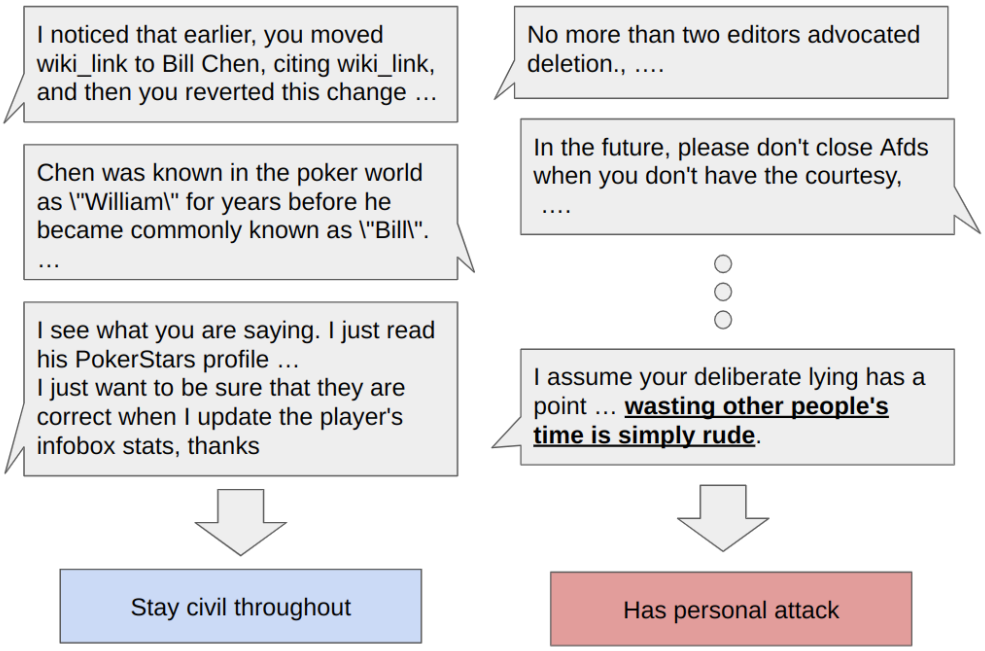

図　会話脱線現象の概要図

左側は、会話が最後まで本来のトピックに集中しており、他者に対する攻撃表現は現れていない。一方、右側は会話が脱線している状態で、最後のコメントにおいて、明確に侮辱的な表現（wasting other people's time is simply rude：他人の時間を無駄にするのは失礼である）が現れている。

用語解説

注１）　大規模言語モデル（Large Language Model; LLM）

非常に大きなパラメータ数を持つニューラルネットワークを大規模なデータセットを用いて学習したモデル。与えられたコンテキストに対して次の文字列の生成確率を高い精度で予測することができるため、テキストの分類をはじめ、要約や翻訳などさまざまな応用が期待されている。

注２）　ゼロショット予測（Zero-Shot Prediction）

機械学習における問題設定の一つ。典型的な機械学習タスクでは、あらかじめ定義されたさまざまなカテゴリ（ラベル）に属する訓練データを使ってモデルを学習し、同じカテゴリのテストデータに対する精度を評価する。一方、ゼロショット予測では、訓練時に一度も見たことのないカテゴリ（未知のラベル）に属するデータがテスト時に与えられても、それを正しく分類・予測することを目指す。

研究資金

（該当なし）

掲載論文

【題　名】 Zero-Shot Prediction of Conversational Derailment With Large Language Models
（大規模言語モデルによる会話脱線のゼロショット予測）

【著者名】 Kenya Nonaka and Mitsuo Yoshida

【掲載誌】 *IEEE Access*

【掲載日】 2025 年 3 月 25 日

【DOI】　10.1109/ACCESS.2025.3554548

問合わせ先

【研究に関すること】
吉田　光男（よしだ　みつお）
筑波大学 ビジネスサイエンス系　准教授
URL: https://www.gssm.otsuka.tsukuba.ac.jp/~mitsuo/

【取材・報道に関すること】
筑波大学広報局
TEL: 029-853-2040
E-mail: kohositu@un.tsukuba.ac.jp