リフレーミング可能な対話モデル構築に向けた基礎実験

Basic experiments for Constructing a Reframing Dialogue Model

船迫 龍之介[*1]，當間 愛晃[*2]
Ryunosuke FUNASAKO[*1], Naruaki TOMA[*2]
[*1] 工学部工学科知能情報コース
[*1]Presently with Computer Science and Intelligent Systems Program, School of Engineering, Faculty of Engineering
[*2] 琉球大学
[*2]University of the Ryukyus
Email: e205732@ie.u-ryukyu.ac.jp, tnal@ie.u-ryukyu.ac.jp

あらまし：リフレーミングとは，ある枠組みで捉えられている物事に対して，違う枠組みで捉え直すことを指す．心の不調者の増加への一助として対話システムを利用した認知行動療法が注目されている．リフレーミングを対話システムに適用し，ユーザーのネガティブ発話をポジティブな表現に言い換えることで，認知行動療法における効果の向上が期待できる．本研究では，ChatGPT（GPT4-Turbo）を利用してネガティブ発話とポジティブ返答をペアとしたリフレーミングコーパスを構築し，その分析を行なった．また，構築したコーパスを利用して言い換え生成モデルの試作を行い，その言い換え性能の評価を行なった．
キーワード：リフレーミング，言い換え

1. はじめに

昨今，心の不調を訴える人が増加している[(1)]．こうした状況に対して，認知行動療法を体験できる対話システムの開発が行われている[(2)]．心の不調者がネガティブな発話を多く行う傾向があることを考慮すると，ユーザーのネガティブ発話を対話システムがポジティブな表現に言い換えて返答することは，認知行動療法の効果を高める重要なアプローチとなり得る．こうしたネガティブ発話をポジティブな表現に言い換える対話技術をリフレーミングと呼ぶ．リフレーミングは，ある枠組みで捉えられている物事に対して，違う枠組みで捉え直すことを指す[(3)]．

本研究では，リフレーミング可能な対話モデルを構築するための最初のステップとして，ChatGPT（GPT4-Turbo）を利用してリフレーミングコーパスを構築した．また，構築したコーパスの分析，言い換え生成モデルの試作を行なった．さらに，試作したモデルの言い換え性能の評価を行なった．

2. リフレーミングコーパスの構築と分析

ChatGPT を利用してリフレーミング事例を収集し，リフレーミングに基づいた言い換え生成モデルを試作するための小規模なリフレーミングコーパスを構築する．

2.1 リフレーミングコーパスの構築

リフレーミングコーパスは相談者役 A のネガティブ発話とカウンセラー役 B のポジティブ返答のペアで設計する．まず，ChatGPT に相談者役 A のペルソナ 250 人を生成してもらう．次に，ペルソナ情報を入力として与え，1 人に対して 4 つのトピックに関するネガティブ発話を生成してもらう．トピックは生活全般に関する悩み，職場または学業に関する悩み，家族または恋愛に関する悩み，自己評価に関する悩みの 4 つである．次に，ネガティブ発話を入力として与え，ポジティブ返答を生成してもらう．最後に，ネガティブ発話とポジティブ返答を校閲する．

2.2 リフレーミングコーパスの分析

リフレーミングコーパスの統計情報を調査した．リフレーミングコーパスは 1,000 ペア全てがユニークであり，それはネガティブ発話が全てユニークであることに起因していることがわかった．しかし，実際のところは主語や言い回しの違いによってユニークなものになっているだけで，内容的には同じものが多くあった．また，いくつかのポジティブ返答に関しては複数の異なるネガティブ発話のペアとして使われていることがわかった．平均文字数，平均単語数，平均文節数に関しては，ネガティブ発話に対してポジティブ返答の方が少なく，比較的簡潔にポジティブな言い換えを行えていることがわかった．

3. 言い換え生成モデルの構築と評価

本研究で構築したリフレーミングコーパスを学習データとして利用し，リフレーミングに基づいた言い換え生成モデルの試作を行う．また，構築した言い換え生成モデルがリフレーミングに基づいた言い換えをどれだけ効果的に生成できるかについての評価を行う．

3.1 言い換え生成モデルの構築

事前学習済みモデルを使用し，構築したリフレーミングコーパスでのファインチューニングを通じて，リフレーミングに基づいた言い換え生成モデルの試作を行う．まず，コーパスをモデルが処理できる形に整形する．この過程で，<NEG_START>と<POS_START>という特殊トークンを追加し，モデルがリフレーミングに基づいた言い換えタスクを効果

的に学習できるようにする．次に，5 分割交差検証を用いて epoch 数の最適化を目的としたハイパーパラメータチューニングを行う．最後に，最適な epoch 数で事前学習済みモデルをファインチューニングする．本研究で使用したモデルは，japanese-gpt2-small と t5-base-japanese であり，以降 GPT-2 及び T5 として言及する．

3.2 評価実験

構築した言い換え生成モデルの性能について自動評価と人による評価を行う．自動評価の指標としては，ポジティブ返答候補上位 100 件に正解のポジティブ返答が完全一致で含まれている割合 Accuracy，ポジティブ返答候補上位 100 件を用いたランキング評価値 MRR，生成されたポジティブ返答と正解のポジティブ返答の bigram までの一致を考慮した BLEU の平均，生成されたポジティブ返答と正解のポジティブ返答の分散表現ベクトルのコサイン類似度 Similarity の平均を用いた．人による評価の指標としては，ネガティブ発話の返答として適切かといった対話の自然さ，ネガティブ発話をポジティブに言い換えることができているかといったポジティブ度，ネガティブ発話の文意を保持しているかといった文意の保持を用いた．上記 3 つの評価指標を用いて，生成されたポジティブ返答に対して 5 段階評価を行なってもらい，その平均値をスコアとする．

4. 評価結果

4.1 自動評価

表 1 に，自動評価の結果を示す．T5 が GPT-2 と比較して性能が高いことが示された．特に MRR と Similarity に関しては顕著な差が見られる．正解のポジティブ返答が完全一致で含まれていたものの中でランキング 1 位だった件数は GPT-2 が 4 件，T5 が 17 件だった．MRR と Similarity における顕著な差は，正解のポジティブ返答が完全一致でポジティブ返答候補のランキング 1 位として含まれていた数に起因すると考える．

表 1 自動評価の結果

Model	Acc	MRR	BLEU	Sim
GPT-2	20.0	0.07	33.1	0.78
T5	39.0	0.22	39.5	0.94

4.2 人による評価

表 2 に，10 人の評価者による評価の結果を示す．自然さと文意の保持に関しては T5 が，ポジティブ度に関しては GPT-2 が高いスコアを得た．しかしその差は若干で，自動評価ほど顕著な差は見られなかった．この結果は，人による評価指標や実験設計が両モデルの違いを明確に区別するのに十分でないことを示唆している可能性がある．また，標準偏差に基づいた評価結果のばらつきを調査したところ，両モデルとも文意の保持に関する評価にばらつきが見られた．この結果は，人による評価において一貫性の向上が必要であることを示唆している．

表 2 人による評価の結果

Model	自然さ	ポジティブ度	文意の保持
GPT-2	3.38	3.88	3.05
T5	3.44	3.80	3.08

4.3 生成されたポジティブ返答

人による評価の対象とならなかった生成されたポジティブ返答に関しても，その傾向を確認した．GPT-2 は文脈に合わせて言葉を組み合わせる能力を活かし，学習した知識を基に新しい表現を創出できていた．例えば，正解のポジティブ返答と類似した内容かつ「趣味の時間」を「趣味に没頭する時間」というような強調した表現に変化させて生成できていた．しかし，返答が一文でなかったり，文としてまたは返答として自然でなかったりする返答を生成している事例も散見された．T5 はテキスト変換能力を活かし，リフレーミングコーパスに存在する表現を使って，文として自然な返答を生成できていた．しかし，入力のネガティブ発話に存在しない情報を含んでいたり，返答として自然でなかったりする返答を生成している事例も散見された．

5. おわりに

本研究では，ユーザーのネガティブ発話をポジティブな表現に言い換えることができる対話モデルを構築するために，ChatGPT を利用してリフレーミングコーパスを構築した．また，構築したコーパスの分析と，構築したコーパスを学習データとして利用して試作したモデルの評価を行なった．収集したリフレーミング事例は比較的簡潔にポジティブな言い換えを行えているものの，内容の多様性や出現頻度のバランスが課題となった．今後はリフレーミングコーパスの評価と拡張，言い換え生成モデルの改善，及びリフレーミング以外の対話技術の導入を目指す．

参考文献

(1) 株式会社インテージリサーチ:“新型コロナウイルス感染症に係るメンタルヘルスとその影響に関する調査報告書”，https://www.mhlw.go.jp/content/12200000/001097924.pdf（参照 2024-01-11）

(2) 中込和幸:“COVID-19 等による社会変動下に即した応急的遠隔対応型メンタルヘルスケアの基盤システム構築と実用化促進に向けた効果検証”，https://researcher.jp/projects/view/1125242（参照 2024-01-11）

(3) 竹田葉留美:“出来事の視点を変えてポジティブに考える～リフレーミングを活用したストレスマネジメント～”，情報の科学と技術, Vol. 67, No. 3, pp. 121-122 （2017）