2W-07

ユーザの発話を取り込むペルソナ対話エージェントの対話性能評価

近藤一希† 佐久間拓人‡ 加藤昇平‡

†名古屋工業大学 工学部情報工学科
‡名古屋工業大学大学院工学研究科工学専攻

1 はじめに

近年，ストレス緩和や娯楽などの観点からエージェントに非タスク対話（雑談）を求めることが増加している．しかし，既存の雑談チャットエージェントは長い対話に一貫性を持たないことや，個性を持たず画一的である欠点が存在する．

Zhang ら [1] はその解決策として，話者の情報をペルソナとして組み込んだ対話データセットの PERSONA-CHAT dataset と，それを用いた単純な対話モデルを発表した．Song ら [2] は PERSONA-CHAT dataset に基づいて作られた Convai2 dataset を用いて，より優れた対話モデルである Bert-Over-Bert(BoB) モデルを発表した．このモデルは既存のモデルと比較して高い語彙能力や一貫性を持っており，事前に与えたペルソナに応じた発話を可能としたが，対話の中で生まれた情報を取り入れず，一貫性を保たない．吉田ら [3] は対話の中で生まれた情報を取り入れるため，エージェント発話を新たなペルソナとして取り入れる手法を考案し，応答に与える影響について分析した．

本研究ではエージェント自身の発話だけでなくユーザ発話もエージェントのペルソナとして取り入れるエージェントを提案する．これにより，ユーザーの発言を取り入れることによる語彙能力の上昇や，ユーザに類似した親しみやすい個性を獲得することを試みる．

2 提案手法

図 1 に提案モデルの概観を示す．提案モデルでは，会話が行われるとまずユーザの発話がペルソナ追加機構を経てエージェントのペルソナに追加される．その後発話文とエージェントのペルソナは共に発話文生成モデルに入力され，エージェントの発話文を出力する．その後生成された発話文はユーザの発話文と同様にペルソナ追加機構への入力となり，エージェントのペルソナとして追加される．

2.1 発話文生成モデル

発話文生成モデルは Song らの事前学習済みモデルである BoB モデルを用いる．学習に使用したデータセットは Conv AI2 Persona Chat dataset[4] であり，これには 1,155 人のペルソナと，131,438 の対話例が含まれている．

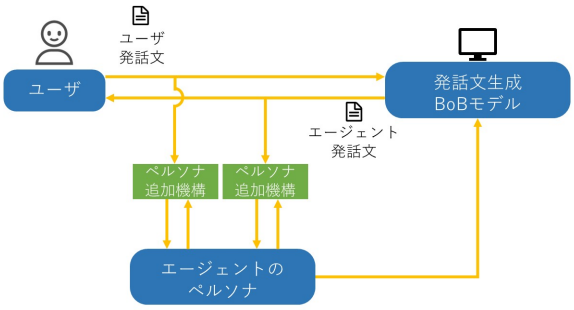

図 1: 提案モデルの概観

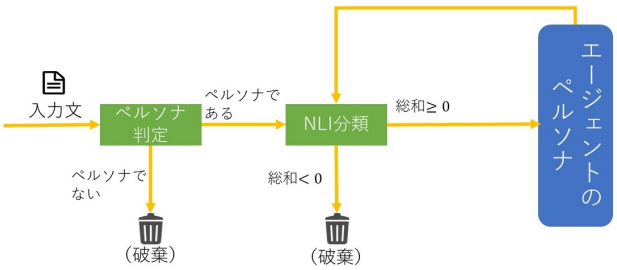

図 2: ペルソナ追加機構図

2.2 ペルソナ追加機構

ペルソナ追加機構では吉田らの手法と同一の条件で発話からペルソナを抽出し，抽出されたペルソナと，エージェントが持つペルソナ全てを対象に NLI 分類を行い，総和が 0 以上の場合新たなペルソナとして追加する．ペルソナ条件は以下の 3 つである．

1. 4 から 20 の単語（句読点含む）で構成される
2. ”I” もしくは”my” を含む
3. 名詞・代名詞・形容詞のいずれかを含む

NLI 分類器は入力に二つの文を取り，それらを「矛盾」「中立」「含意」に分類して下式に従い値を出力する．本研究では PyTorch 用ライブラリ fairseq に組み込まれた RoBERTa モデル（GLUE MNLI Score 90.2）を用いる．

$$NLI\,(r, p_i) = \begin{cases} -1 & r \text{ が } p_i \text{ と矛盾} \\ 0 & r \text{ が } p_i \text{ と中立} \\ 1 & r \text{ が } p_i \text{ と含意} \end{cases} \quad (1)$$

ただし，r はエージェントの発話，p_i はエージェントの i 番目のペルソナを表す

3 実験

本研究では，人に対話させる有人対話実験と，有人対話において試行ごとに対話内容が変化することを防

Dialogue performance evaluation of persona dialogue agents that incorporate user utterances
Kazuki KONDO† Takuto SAKUMA‡ Shohei KATO ‡
†Dept. of Computer Science, Nagoya Institute of Technology
‡Dept. of Computer Science and Engineering, Graduate School of Engineering, Nagoya Institute of Technology
†‡Gokiso-cho, Showa-ku, Nagoya 466-8555, Japan
{kkondo, sakuma, shohey}@katolab.nitech.ac.jp

表 1: 有人対話実験結果

Agent	ペルソナ数	Dist-1	Dist-2	C.Score.AVG
Both	20	**0.39**	**0.75**	-0.32
Response	**31**	0.34	0.67	-0.29
Nothing	5	0.36	0.68	**-0.24**

表 2: 機械対話実験結果

Agent	ペルソナ数	Dist-1	Dist-2	C.Score.AVG
Both	**116**	0.39	0.74	-0.24
Response	68	**0.42**	**0.76**	-0.34
Nothing	5	0.34	0.63	**-0.22**

ぎつつ対話回数を増加させるための機械対話実験の 2 実験を実施し，語彙能力や一貫性を評価する．両実験の共通設定として同じペルソナを持った以下の 3 つのエージェントを用いる．

- Both: 今回の提案手法であり，ユーザ・エージェント両方の発話文をペルソナに追加する
- Response: エージェントの発話文のみをペルソナに追加する
- Nothing: ペルソナを追加しない

3.1 有人対話実験設定

20 代男性一人を対象に対話実験を実施した．被験者には Both と Response の 2 つのエージェントに 100 往復の対話を事前に行わせ，対話の中でペルソナを追加させた．その後，Nothing を含めた 3 エージェントを相手に 50 往復の対話を 2 試行実施した．

3.2 機械対話実験設定

定型文を 1500 文用意し，有人対話実験と同様に 2 つのエージェントに対話させ，ペルソナを追加させる．その後，Nothing を含めた 3 エージェントに 50 往復の対話を機械的に行った．

3.3 評価指標

語彙能力の評価には式 (2) で算出される Distinct-n（Dist.n）を，一貫性の評価には式 (3) で算出される Consistency Score Average（C.Score.AVG）を用いる．一貫性評価で一般的に用いられる指標は，対話文と全てのペルソナのペアに NLI 分類することで算出される Consistency Score であり，式（4）で算出される．しかし，本研究ではペルソナの数を増やす影響を受けるため判断に適さない. よって，全ての対話の C.Score の和を取り，累計判定回数 Judge で除した C.Score.AVG を利用する．

$$distinct\text{-}n = \frac{\mid unique(n\text{-}gram) \mid}{\mid n\text{-}gram \mid} \tag{2}$$

$$C.Score.AVG = \sum_{j=1}^{M_r} \frac{C.Score(r_j)}{Judge} \tag{3}$$

$$C.Score(r) = \sum_{i=1}^{M_p} \mathbf{NLI}\,(r, p_i) \tag{4}$$

ただし，M_r はエージェント発話の総数，M_p はエージェントのペルソナ数，r_j は j 番目のエージェントの発話，p_i はエージェントの i 番目のペルソナを表す．

4 結果

有人対話実験の結果を表 1 に，機械対話実験の結果を表 2 に示す．太字はその列での最大値を表し，有人対話実験では 2 試行の平均値を表記している．

表 1 より，Both の語彙能力が Nothing だけでなく Response と比較しても上昇する傾向にあることが判明した．しかし，新たにペルソナを加えることで，一貫性が低下する傾向にあることも読み取れる．

表 2 より，長期的な会話の後では，Response の語彙能力が最大値をとった．一方，有人実験の際には Response と同程度であった Both の C.Score.AVG は，最大をとる Nothing の C.Score.AVG の値に近付いた．

5 考察

ペルソナを追加することで語彙能力が向上する傾向にあることは先行研究のとおりであることが確認出来たが，一貫性の向上については確認できなかった．語彙能力に関して，吉田らの研究ではペルソナ数が大きいほど語彙能力が上がるとされていたが，表 2 より，単純にペルソナ数を増加しても語彙能力が向上するとは限らないことがわかる．

また，先行研究の結果に反し，ペルソナを追加すると一貫性は低下した．これについて，先行研究ではペルソナ数を考慮しない総和をとる手法で評価したため向上したと考えられる．今回の指標を用いた場合，ペルソナ数が増え複雑性が増している以上一貫性の低下は妥当であると考えられる．

6 今後の展望

今後の展望として，一貫性や語彙力などを対話を通して評価させる評価実験と，事前に対話したエージェントを複数のエージェントの中から判別させる判別実験の 2 つの実験を実施する．評価実験では 3 節の実験で用いたものと同種の 3 種類のエージェントと対話させ，それを感性評価する．評価内容は「一貫性」「語彙性」「流暢性」「親密性」の 4 因子 5 段階を予定している．判別実験では評価実験と同じ被験者に，直前の評価実験で対話したエージェントと，事前に用意した別のペルソナを持つエージェントを相手にブラインドで対話させ，事前に対話したエージェントを判別させる．これを Both と Response でそれぞれ実施し，結果を比較する．これにより対話による個性の獲得を評価する．

参考文献

[1] Saizheng Zhang et al: Personalizing Dialogue Agents: I have a dog, do you have pets too?, arXiv:1801.07243 (2018).

[2] Haoyu Song et al: Generate, Delete and Rewrite: A Three-StageFramework for Improving Persona Consistency of Dialogue Generation,arXiv:2004.07672,(2020).

[3] 吉田 快 他：応答履歴に応じたペルソナの更新が対話システムの応答生成へ与える影響の分析, 第 93 回言語・音声理解と対話処理研究会会議録, pp.32-37 (2021).

[4] Dinan et al. The second conversational intelligence challenge (convai2)arXiv:1902.00098, (2019).