ペルソナ対話システムにおけるペルソナ選択と応答生成

吉田快 [1,2]　吉野幸一郎 [2,1]　品川政太朗 [1]　須藤克仁 [1]　中村哲 [1]
[1] 奈良先端科学技術大学院大学　[2] 理化学研究所 ガーディアンロボットプロジェクト
{yoshida.kai.yf1, koichiro, sei.shinagawa, sudoh, s-nakamura}@is.naist.jp

概要

ペルソナ対話システムでは対話が進むのに従い，対話の文脈上生成された新しいペルソナに相当する情報が現れる場合がある．この新しいペルソナを考慮しない場合，対話文脈と矛盾した応答を生成することがあり得る．そこで本研究では，こうした新しいペルソナに相当する情報をペルソナプールに保持し，その中から対話の文脈に合わせて必要なペルソナを選択し利用する枠組みを提案する．このため，人手によるペルソナ選択データセットの構築を行い，より良いペルソナ選択手法について分析を行った．構築したデータセットを用いたペルソナ文選択の評価では，名詞ベース選択の手法が既存の文ベース選択の手法より精度が高いことを確認した．

1　はじめに

近年大規模言語モデルやその周辺技術の発展により，対話システムは与えられた文脈に対して自然な応答生成が可能となった．この結果，次のステージの対話システム研究，例えばシステム自身の発話との整合やシステムに与えられた役割の保持などを目的とするもの，が増えている．例えばペルソナ対話 [1] は，システムにあらかじめ与えられたペルソナに従った発話を行わせることにより，応答の不整合問題を解決しようとする研究のアプローチである [2, 3, 4]．ペルソナの表現方法は大きく分けて 2 種類あり，「私は犬を飼っています」のように文形式で明示的に与える場合 [1, 3, 4] や，特定のユーザの対話データを基に学習されるユーザ表現ベクトル [2, 5] のような，文によらない暗黙的な表現を用いる場合がある．ペルソナ対話システムの多くの先行研究では，Transformer [6] による事前学習済み言語モデルをペルソナ対話のデータセットで微調整する方法が採用されてきた．ペルソナ対話用のデータセットには PersonaChat [1] と呼ばれるデータセットがよく用いられており，PersonaChat の日本語版データセットである JPersonaChat [7] も公開されている．

既存の多くのペルソナ対話システムは，大規模言語モデルをはじめとする大量の学習データで事前訓練された学習済み言語モデルを活用する．このような場合，対話システムが生成する内容に，与えたペルソナに含まれない事実が含まれる幻覚（Hallucination [8]）という現象が発生することがある．例えば，「私は北陸に住んでいます」というペルソナのみを事前に与えられたシステムが「私は現在休職中です」のような発話を生成した場合，この新しい「休職中」という情報もその後の応答生成で考慮しなければ応答の一貫性が保たれない．

ペルソナ対話システムに関する先行研究では，こうした不整合に対応するため，応答生成モデルが生成した内容に応じてシステムのペルソナを逐次的に更新する手法が提案されている [9, 10]．これらの手法では，システムが生成した応答文からペルソナとみなすことができる部分を抽出し，外部メモリに記憶する．この際問題となるのは，入力となるペルソナが逐次的に増える点である．対話が進むにしたがい，利用しうる新しいペルソナは単調に増加する．そのため，訓練時に想定されていない数のペルソナを扱うことは考慮されておらず，応答生成に悪影響を及ぼしうる．

この問題に対し，入力されたユーザ発話毎に適切なペルソナのみを選択し，選択したペルソナを入力して応答生成を行うことで，システムの応答精度を高めることが期待できる．対話の文脈に応じて新しいペルソナから必要なペルソナを選択して利用する場合，応答に使うべきペルソナの選択は再現率を下げないようにしつつ，適合率を上げることが必要である．こうした目的では，類似する可能性が高いペルソナ候補をできるだけ網羅できるような基準を用いることが重要である．この観点から本研究では，ペルソナ対話システムにおけるペルソナ選択の有効性を確認すると共に，どのような基準で選択することがどのような場面で有効か明らかにする．

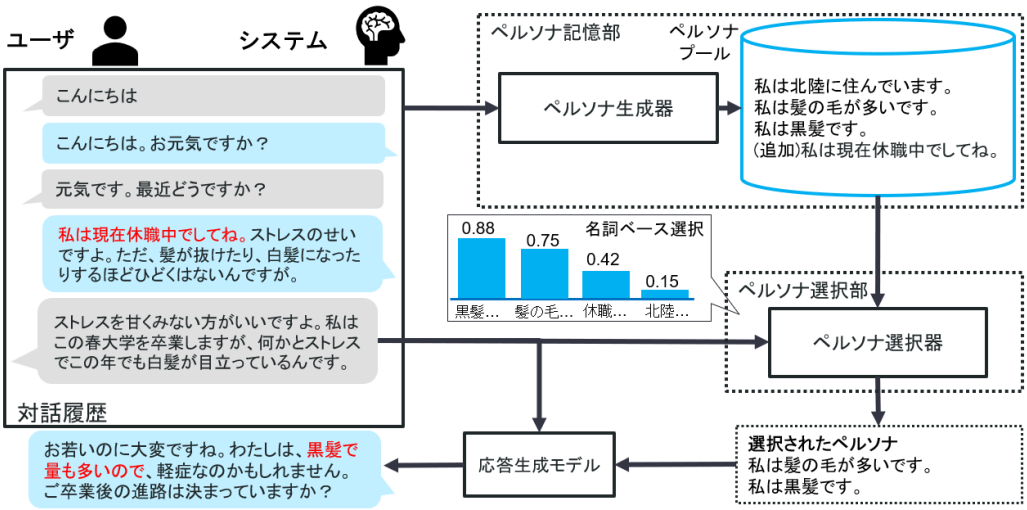

図 1　ペルソナ更新可能な対話システムの例

2　ペルソナ更新とペルソナ選択

本節ではまずペルソナ更新システム [5, 9] について説明する．さらに，更新によってプールに蓄積されたペルソナを選択する機構について説明する．また，この選択における選択基準について説明する．

2.1　ペルソナ更新とペルソナ選択機構を持つペルソナ対話システム

ペルソナ更新機構を持つペルソナ対話システムの全体像を図 1 に示す．まず，ペルソナ対話システムは，与えられたペルソナと対話履歴を入力として応答を生成する．この対話の流れの中で，「私は現在休職中でしてね」のように，新しいペルソナに相当する情報が応答生成モデルから生成される場合がある．このとき，ペルソナ生成器がペルソナ更新として，対応するペルソナ文をペルソナプールに保存する（追加されたペルソナ: 私は現在休職中でしてね）．ペルソナプールへの追加は先行研究 [3] に従い，以下のルールを満たすものを抽出・記憶する．

1. 4 から 20 の単語 (句読点を含む) で構成される
2. ”私”か”わたし”という単語を含む
3. 名詞，代名詞，形容詞のうち少なくともどれか 1 つを含む

それ以降の応答生成を行う際に，このペルソナプールから現在の対話文脈に合わせて利用するべきペルソナを選択する．このペルソナ選択によって選択されたペルソナのみが応答生成モデルに与えられ，次の応答生成に用いられる．

2.2　ペルソナ選択手法

ペルソナ選択においては，文の表層における類似として Transformer ベースの構造を持つ ERNIE [11] を用いる手法が提案されている（文ベース選択 [9]）．しかしこの方法では，文構造や表現など表層における類似が重視され，内容に関する考慮が軽視されがちという問題があった．そこで本研究では，特に内容語に注目し，発話中に生じた名詞を抽出して用いる手法を提案する．文脈となるユーザ発話とペルソナプール中のペルソナ文からそれぞれ抽出された名詞を全て word2vec によってベクトル化し，総当たりでコサイン類似度の計算を行う．このうち各ペルソナ文が持つ最大値をそのペルソナ文が持つユーザ発話に対する類似度として扱う（名詞ベース選択）．

文ベース選択では，入力発話文とメモリに存在する全てのペルソナ文について特徴抽出を行い，入力発話文とペルソナ文間のコサイン類似度を計算し，閾値以上の類似度となったペルソナ文をすべて選択する．文からの特徴抽出は bert-large-japanesev2 の最終層の埋め込みの average pooling によって行った．名詞ベース選択では，まず，入力発話文と各ペルソナ文を形態素解析器である MeCab [12] と辞書の ipadic によって名詞のみを抽出し，word2vec により単語ベクトルとした．続いて，入力発話文とペルソナ文に含まれる全ての名詞の組の間で単語ベクトルのコサイン類似度を計算し，最も高い類似度となる組が持つ類似度を当該文対の類似度とした．名詞ベース選択でも文ベース選択と同様に閾値を設定し，閾値を超えたペルソナ文をすべて選択する．

3 実験

本研究では，ペルソナ選択器として文ベース選択と名詞ベース選択の 2 種類の選択器について検証を行う．まず，文ベース選択と名詞ベース選択のペルソナ選択器について，それぞれの選択がどの程度適切に行われているかを人手で構築されたテストデータによって評価する．また，それぞれのペルソナ選択手法に基づいてペルソナ更新・選択を行うペルソナ対話システムを構築し，それらを用いて対話実験を用いた場合の人手評価を行う．以降ではそれぞれの実験設定について説明する．

3.1 ペルソナ選択の評価

JPersonaChat の一部を抽出し，その対話の文脈において利用可能なペルソナ候補文の中からどの文が対話応答に利用されたかのアノテーションを付与したテストデータを構築した [1)]．具体的には，以下のような手順によってアノテーションを行った．

1. JPersonaChat から同じ名詞を共有していない発話 u，ペルソナ文 p_n を事前に選別
2. u と p_n を評価者に提示
3. 評価者は各 u と p_n を比較し，各 p_n に「反映されていない」と「やや反映されている」，「反映されている」3 通りのラベルを付ける

事前に 1000 件選別したペルソナに対し，1 件あたり 3 名がラベル付けを行った．「反映されていない」を 0，「やや反映されている」を 1，「反映されている」を 2 としてラベル付けを行い，3 名のラベルの総和が 3 を超えるもののみを「応答生成に用いられたペルソナ文」として付与した．

3.2 ペルソナ対話システムでの評価

3.2.1 自動評価指標

応答生成モデルの正確さを測るために，テストデータ中の正解応答に対する応答生成モデルの尤度の代わりに用いられているテストセットパープレキシティ (PPL) を用いる．PPL は低いほど，応答生成モデルがテストデータに近い応答生成を行っていることを示す．

3.2.2 人手評価

ペルソナ選択による応答への影響を確認するため，複数ターンの文脈が存在する状態での応答生成を行い生成結果を評価する．ここでは，JPersonaChat に含まれる会話文を用いて，疑似的に複数ターンが与えられた場合の応答生成を行った．具体的には，JPersonaChat であらかじめ定義されている 5 つのペルソナ，対話履歴として 12 ターン（6 応答ペア）の履歴を提示し，それらを入力として生成されたシステム応答をあわせて評価者に提示した．この際，1 手法あたり 4 件，計 16 件を 1 セットとして，各 1,000 件のデータに対して 250 名の評価者をクラウドソーシングで以下の 3 件を付与した．

自然性　自然な応答をしているか

会話履歴の反映度　生成した応答に与えた対話履歴の内容が反映されているか

ペルソナ文の反映度　生成した応答に対話履歴の内容が反映されているか

対話実験は次の 4 通りのモデルで実験を行った．

1. **ベースライン**　Japenese Transformer [7]
2. **更新のみ**　Japenese Transformer + ペルソナ更新
3. **文ベース選択**　Japenese Transformer + ペルソナ更新 + 文ベース選択
4. **名詞ベース選択**　Japenese Transformer + ペルソナ更新 + 名詞ベース選択

4 実験結果

4.1 ペルソナ選択の精度

最初に，文ベース選択，名詞ベース選択それぞれの手法で選択を行った場合の ROC 曲線および Precision-Recall 曲線を図 2 および図 3 に示す．これらの図から，全体の傾向として名詞ベース選択が文ベース選択よりも Recall/FPR に対して高い Precision/TPR を達成していることが確認でき，より良いペルソナ選択ができていることがみて取れる．また，このテストデータによって F1 値が最大となるような値を選択することで最適な閾値を設定し，文ベース選択が 0.72，名詞ベース選択が 0.39 となった．

最適な閾値を設定した際の，文ベース選択と名詞ベース選択の選択精度を表 1 に示す．結果から，提

表 1　選択の評価

手法	Accuracy	Recall	Precision	F1
文ベース選択	0.184	**0.971**	0.158	0.272
名詞ベース選択	**0.692**	0.458	**0.244**	**0.319**

1) https://github.com/riken-grp/PersonaSelection

案する名詞ベース選択は 7 割近い Accuracy を達成した．また，文ベース選択は Recall は高いものの，Accuracy や，F1 が低い結果となった．この結果から，全体の傾向として名詞ベース選択は文ベース選択よりも良い選択を実現可能であることが示せた．

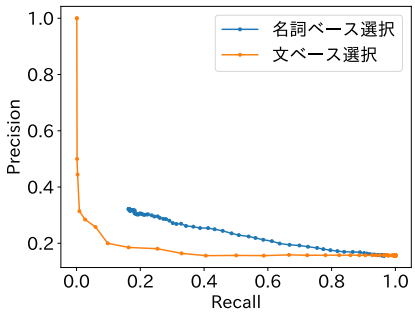

図 2 名詞ベース選択 (青) と文ベース選択 (橙) の Precision-Recall 曲線

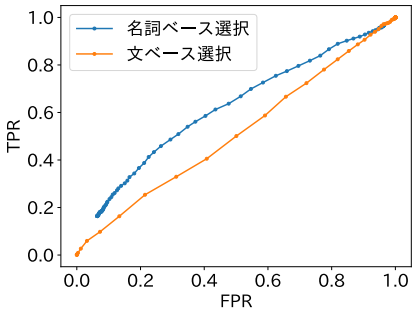

図 3 名詞ベース選択 (青) と文ベース選択 (橙) の ROC 曲線

4.2 生成された応答の評価

4.2.1 自動評価

4.1 節で得られた閾値を用いて実際にユーザ発話によって使用するペルソナ文を選択し，評価した結果を説明する．まず，自動評価におけるテストセットパープレキシティ（PPL）を表 2 に示す．

表 2 応答生成における PPL．p_{avg} は応答生成に使用されたペルソナの数の平均を表す．

	PPL	p_{avg}
ベースライン	13.3	5
更新あり	13.63	6.57
文ベース選択	12.93	5.78
名詞ベース選択	**12.01**	**1.6**

文ベースと名詞ベース選択によってベースラインよりも PPL の低下が確認できた．また，更新のみの手法では応答に使用するペルソナ数が増加しているが，更新と選択を導入した手法では更新のみの手法よりも応答に使用するペルソナ数を削減できた．特に名詞ベース選択は，入力するペルソナ数を大幅に削減しつつ，最も低い PPL を達成できた．

4.2.2 人手評価

人手評価の結果を表 3 に示す．

表 3 人手評価

手法	自然性	対話履歴の反映度	ペルソナの反映度
ベースライン	3.856	3.831	3.483
更新あり	**3.960**	**3.941**	3.455
文ベース選択	3.929	3.860	**3.484**
名詞ベース選択	3.921	3.866	3.477

ベースラインと比較して，更新を取り入れた手法はすべて自然性と対話履歴の反映度が向上していることが確認できた．また，更新あり手法に選択を加えた場合は自然性と対話履歴の反映度が低下する反面，ペルソナの反映度が向上する結果となった．しかし, 手法間の有意差を示すためにマンホイットニーの U 検定により，名詞ベース選択とその他の手法での評価指標ごとの有意差の確認を有意水準 5% で行った．その結果，各手法間の有意差は確認できなかった．

5 まとめ

本研究では，ペルソナ対話システムが自身の過去の応答内容に対して一貫性を持った応答を行うことを目的として，ペルソナの更新と選択を取り入れた対話システムを提案しその評価を行った．応答文と利用されたペルソナの対応について人手で紐づけたテストデータを構築し評価した結果，提案する名詞ベース選択手法はより良いペルソナ選択を実現できていることがわかった．さらに，ペルソナ選択を導入した応答生成ではペルソナ選択によりペルソナ数を削減した上で PPL を改善できることが確認された．

対話実験では、ペルソナ選択を行う場合と行わない場合（更新のみ）で人手評価値に大きな差がつかなかった．今回，実験に用いた対話履歴はそれほど長くなく, 結果として既存の対話モデルでもうまく応答生成が行える設定になっていた可能性が考えられる．こうした問題について議論するためには，さらに長い対話履歴や特定の個人との長期間にわたる対話において，提案したようなペルソナ選択に基づく手法がうまく動作するかを議論する必要がある．

参考文献

[1] Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. Personalizing dialogue agents: I have a dog, do you have pets too? In **Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 2204–2213, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[2] Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and Bill Dolan. A persona-based neural conversation model. In **Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 994–1003, Berlin, Germany, August 2016. Association for Computational Linguistics.

[3] Pierre-Emmanuel Mazaré, Samuel Humeau, Martin Raison, and Antoine Bordes. Training millions of personalized dialogue agents. In **Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing**, pp. 2775–2779, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[4] Haoyu Song, Yan Wang, Kaiyan Zhang, Wei-Nan Zhang, and Ting Liu. BoB: BERT over BERT for training persona-based dialogue models from limited personalized data. In **Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)**, pp. 167–177, Online, August 2021. Association for Computational Linguistics.

[5] Qian, et al. Learning implicit user profile for personalized retrieval-based chatbot. In **Proceedings of the 30th ACM International Conference on Information & Knowledge Management**, pp. 1467–1477, 2021.

[6] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, L ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, **Advances in Neural Information Processing Systems**, Vol. 30. Curran Associates, Inc., 2017.

[7] Hiroaki Sugiyama, Masahiro Mizukami, et al. Empirical analysis of training strategies of transformer-based japanese chit-chat systems, 2021.

[8] Ziwei Ji, et al. Survey of hallucination in natural language generation. **ACM Comput. Surv.**, Vol. 55, No. 12, mar 2023.

[9] Xinchao Xu, Zhibin Gou, Wenquan Wu, Zheng-Yu Niu, Hua Wu, Haifeng Wang, and Shihang Wang. Long time no see! open-domain conversation with long-term persona memory. In **Findings of the Association for Computational Linguistics: ACL 2022**, pp. 2639–2650, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[10] Zhengyi Ma, et al. One chatbot per person: Creating personalized chatbots based on implicit user profiles. In **Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval**, SIGIR '21, p. 555–564, New York, NY, USA, 2021. Association for Computing Machinery.

[11] Sun, et al. Ernie 2.0: A continual pre-training framework for language understanding. **Proceedings of the AAAI Conference on Artificial Intelligence**, Vol. 34, No. 05, pp. 8968–8975, Apr. 2020.

[12] Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. Applying conditional random fields to Japanese morphological analysis. In **Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing**, pp. 230–237, Barcelona, Spain, July 2004. Association for Computational Linguistics.

A　ペルソナ選択の評価者への教示内容

この作業では，発話文 1 文とプロフィール文 9 文を比較して，発話文に使われていると考えられるプロフィール文を選択していただきます.

【選択の条件】

- 発話文の根拠となっていそうなプロフィール文を選択してください
- 発話文にプロフィール文が含まれない場合は選択しないでください
- 発話文にプロフィールが複数含まれる場合は複数選択してください
- 発話文とプロフィール文に部分一致がある場合はそれを選択してください
- 発話文から容易に連想できる場合はそのプロフィール文を選択してください

B　発話評価者への教授内容

1. 自然な応答ができているか

・会話履歴の最後の発話に対して自然な応答ができている場合は「5」を選択してください

・明らかに不自然な応答の場合は「1」を選択してください

・それ以外の場合は，応答の自然さに応じて「2〜4」を選択してください

2. 会話履歴を反映した応答ができているか

・ここでは，会話履歴にたいして自然な応答ができてるかは関係なく，応答文に会話履歴の内容が反映されているかで選択を行ってください

・応答文の中に会話履歴の中に登場するキーワード，もしくはそれに関連する情報が含まれている場合は「5」を選択してください

・応答文に会話履歴の内容が含まれない場合は「3」を選択してください

・過去の応答と矛盾した応答をしている際は「1」を選択してください

3. プロフィール文を反映した応答ができているか

・ここでは，応答文にプロフィール文が反映されているかで選択を行ってください

・応答文の根拠となっていそうなプロフィール文が 1 つでも応答文に反映されていれば「5」を選択してください

・応答文にプロフィール文が含まれない場合は「3」を選択してください

・プロフィールと矛盾した発話をしている際は「1」を選択してください

C　評価の分布

各評価値の分布を図 4 と図 5，図 6 に示す.

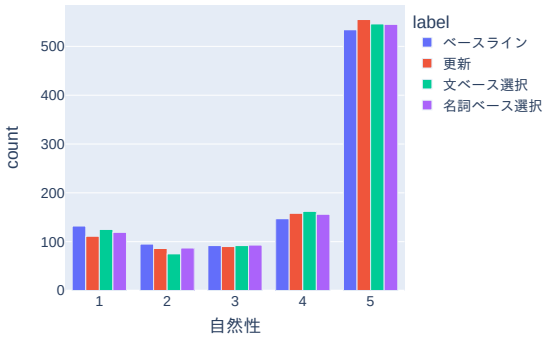

図 4　自然性の分布

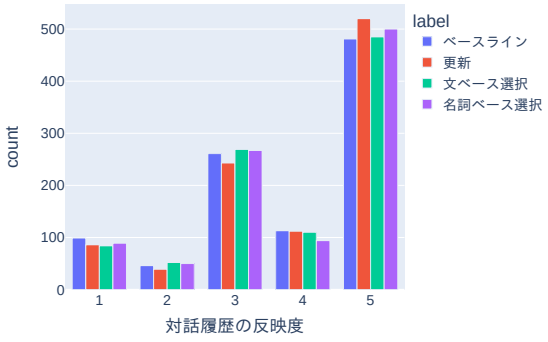

図 5　対話履歴の反映度反映度の分布

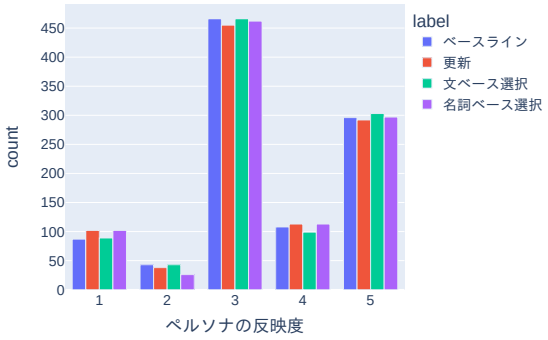

図 6　ペルソナの反映度の分布