

ペルソナ更新型対話システムにおける効果的なペルソナ選択手法の分析

吉田 快 品川 政太朗 須藤 克仁 中村 哲

奈良先端科学技術大学院大学

{yoshida.kai.yf1, sei.shinagawa, sudoh, s-nakamura}@is.naist.jp

概要

ペルソナ対話システムにペルソナの更新を導入する場合、蓄積されたペルソナを応答生成に活用するためにペルソナの選択方法が重要となる。従来の文ベースの類似度によるペルソナ選択手法は、表層的に似た表現の文にも高い類似度を与えるため、問題がある。本研究では、ペルソナ選択手法の改善のために、ペルソナ選択について分析を行い、入力文とペルソナ文に含まれる名詞語の類似度を用いた選択手法が有望であることを示す。また、名詞ベースと文ベースの選択手法による応答生成文をユーザ評価により比較することで、両者の比較を行った。その結果、ペルソナを常に用いた応答をすると、応答の自然性が低下することが分かった。

1 はじめに

ペルソナ対話システム [1, 2, 3, 4] は**ペルソナ**と呼ばれるプロフィール情報を持つ対話システム的一种である。ペルソナ対話システムは近年の機械学習モデルの発達により増加している、生成ベースの雑談対話システム [5, 6] における応答の不整合問題 [3] を解決することを目的としている。本論文では、ペルソナ文の更新が可能なペルソナ対話システムに焦点を当てる。ペルソナ更新は、ペルソナ対話システムにとって重要である。なぜなら、ペルソナ対話システムの生成文は、対話の後で生成されるシステムプロフィールとも整合していなければならないからである。たとえば、図 1 のように、ユーザがシステムプロフィールを引き出すために「Hello what are you doing today?」という質問をすると、「I am student」というペルソナ文が存在しないにもかかわらず、「I am student」を含む応答が生成されることがある。この現象は、いわゆる **Hallucination** としてよく知られており、大規模な事前学習済みのニューラル言語モ

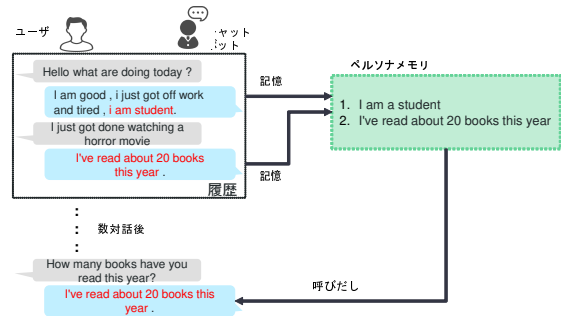


図 1 ペルソナ更新可能な対話システムの例

デルに基づく対話システムでしばしば発生する。

このような Hallucination に対応するために、ペルソナ対話システム分野ではペルソナ更新という概念が新しいタスクとして導入されている [7, 8]. ペルソナ更新とは Hallucination によって得られた、ペルソナ文を含む文を新しいペルソナ文として保存し、システムが応答生成時に利用できることを指す。システムが応答生成時に利用できるようにすることで、後の対話で生成された情報と一致する応答を生成するようにシステムを制約することができる。ペルソナ対話システムにペルソナ更新を導入した場合、ペルソナ文は徐々に増加することになる。システムの応答を継続的に過去のシステム応答におけるペルソナ（ペルソナ更新によって蓄積されたペルソナ）に対して一貫させるには、**与えられたユーザ入力に対して、システムが応答を生成するために蓄積された大量のペルソナ文から適切なペルソナ文をどのように選択するか**という問題が生じる。

最近の研究 [7, 8] では、ユーザ入力との文同士の類似度を用いて上位 N 個のペルソナ文を選択することが提案されている。これにより、ユーザ入力に関連した応答を生成するようにシステムが制約を受けるが、表面的な類似性で無関係なペルソナ文を選択することが考えられる。しかし、ペルソナの選択方法に着目した研究は少なく、選択戦略も確立されていない。そこで本研究では、Integrated gradient [9]

を用いて、PERSONA-CHAT データセットのいくつかの対話を分析し、システムの応答のペルソナ文とユーザ発話への依存性を検討する。この分析の結果から名詞の類似度を用いた選択法を提案し、実際に文の類似度と名詞の類似度に基づく応答生成の結果をユーザ評価により評価を行う。

2 応答生成における入力の寄与度の分析

良いペルソナ選択とは、多くの候補の中から応答生成に適したペルソナやそのトークンを選択できることである。ここでは、システムがペルソナを用いた応答をするか否かを予測する分類器を構築し、入力であるユーザ発話とペルソナ文において分類に寄与する入力トークンを分析することで、優れたペルソナ選択器を検討する。同様に、システムがペルソナを用いて Hallucination を起こすか否かを予測する分類器も構築し、寄与する入力トークンを分析することで、優れたペルソナ選択器を検討する。

入力の寄与度を計算するために、Integrated Gradients [9] を使用する。また、寄与度の計算のために必要な分類器の構築のためのデータセットの前処理については、付録 A に記載する。

2.1 ペルソナを用いた応答生成におけるユーザ発話とペルソナの寄与度

ペルソナを用いた応答生成に対するユーザ発話文とペルソナ文の寄与度を計算するには、ペルソナ使用予測器が十分な精度を持つ必要がある。このペルソナ使用予測器を学習データの PERSONA-CHAT(付録 A) によって評価したところ Accuracy は 0.985, F1 は 0.979 (Recall: 0.969, Precision: 0.989) となり、十分な精度を持つことが確認できた。ここで、既知の入力に対する寄与度を求めるのが目的であるため、学習データで評価を行った。このペルソナ使用予測器を用いて、Integrated Gradients を適用し、各応答に対するペルソナ文とユーザ発話の寄与度を算出した。

分析には、応答への寄与度が最も高い単語を上位 3 つまでカウントした。図 2 は、上位 3 語がどの品詞タグに属するかと、応答がペルソナを含んでいるか (Label 1) と含んでいないか (Label 0) を表している。なお、All tokens は、ユーザ発話またはペルソナ文に出現した単語数を表し、可視化のために 10 で割っている。その結果、入力トークン全体では代名詞 (PRON)、名詞 (NOUN)、特殊トークン

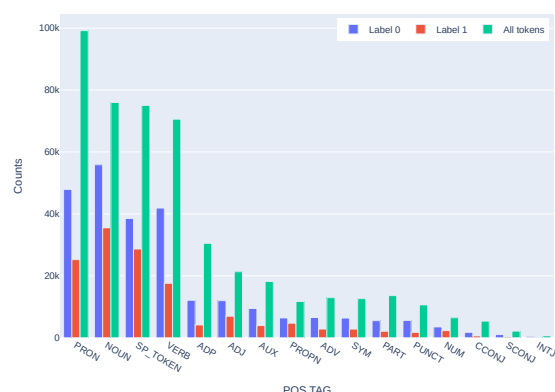


図 2 ペルソナを使用した応答生成に寄与した token の品詞上位 3 つ

(SP_TOKEN) が多いものの、ペルソナを使った応答生成では、名詞、特殊トークン、代名詞の寄与度が高いことが分かった。そこでさらに、名詞語の固有表現に着目して、ユーザ発話とペルソナ文の関係を調べた。寄与度の高い名詞語に注目し、図 3 に示すように、ユーザ発話とペルソナ文の固有表現の出現頻度の相関を可視化した。ここでは、spaCy ¹⁾ を固有表現抽出のために用いた。その結果、ペルソナ文と

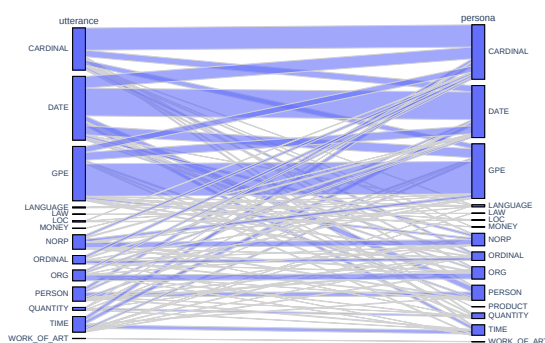


図 3 ペルソナを使用した応答において最も寄与度が高かった token の固有表現の出現頻度の相関

ユーザ発話が同じ固有表現を持つ場合、固有表現の出現頻度の相関が高いことが分かった。

2.2 Hallucination 発生時のユーザ発話とペルソナの寄与度

2.1 節と同様に、生成された応答の Hallucination に対するユーザ発話文とペルソナ文の寄与度を計算するには、Hallucination 予測器が十分な精度を持つ必要がある。この Hallucination 予測器を学習データの PERSONA-CHAT(付録 A) で評価したところ Accuracy 0.988, F1 スコア 0.980 (Recall: 0.989,

1) <https://spacy.io/>

Precision: 0.971) となり、十分な精度を持つことが確認できた。ここでも、2.1 節と同様に既知の入力に対する寄与度を計算するのが目的のため、学習データで評価を行った。この Hallucination 予測器を用いて、システムが生成した応答に対して 2.1 節と同様に Integrated Gradients を適用し、各応答に対するペルソナ文とユーザ発話の寄与度を算出した。分析には、応答への寄与度が最も高い単語を上位 3 つまでカウントした。図 4 は、上位 3 語がどの品詞タグに属するかと、Hallucination が発生した (Label 1) か、していない (Label 0) かを示している。なお、All tokens は、ユーザ発話またはペルソナ文に出現した単語数を表し、可視化のために 10 で割っている。図 4 から分かることは、ラベル 0 の予測に

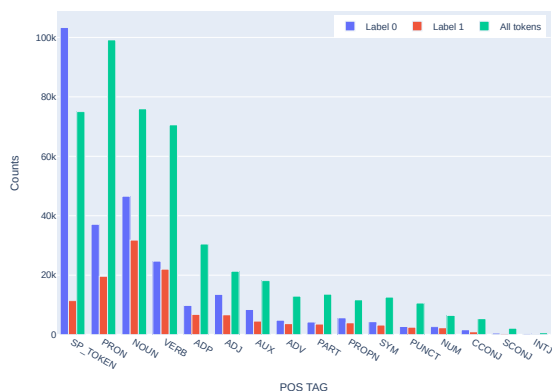


図 4 Hallucination が起こる応答生成に寄与した token の品詞上位 3 つ

は特殊トークンの寄与が大きく、ラベル 1 の予測には名詞、動詞、代名詞の寄与が大きいことである。特に、特殊トークンは 13 万文中 11 万文とかなり高い寄与度であった。

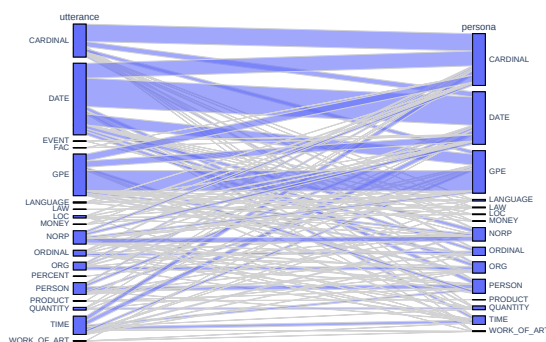


図 5 Hallucination が起こる応答において最も寄与度が高かった token の固有表現

さらに、名詞語の固有表現に着目して、ユーザ発話とペルソナ文の関係を調べた。寄与度の高い名詞

語に注目し、ユーザ発話とペルソナ文の名前付き固有表現の共起の相関を図 5 のように可視化した。その結果、図 3 が示すペルソナ使用の結果と同様に、ペルソナ文とユーザ発話が同じ固有表現を持つ場合、固有表現の出現頻度の相関が高いことが分かった。また、Hallucination と非 Hallucination に分類されたサンプルを別々に集計したところ、Hallucination に分類されたサンプルでは、最も寄与度の高いトークンは発話中の 38,868 トークンであり、ペルソナ文中は 2,958 トークンのみであった。また、非 Hallucination に分類されたサンプルでは、最も寄与度の高いトークンは、発話中の 36,662 トークンであり、ペルソナ文中の 52,943 トークンであることがわかった。これらの結果は、Hallucination がユーザ発話のみに依存していることを示唆している。

3 文ベースと名詞ベースによるペルソナ選択手法の比較

2 章で得られた知見を基に、名詞ベースのペルソナ選択手法が文ベースのペルソナ選択方法よりも優れているのか比較を行う。具体的には、PERSONA-CHAT で fine-tuning した BERT2BERT を用いて構築した応答生成モデルを基に、文同士の類似度と提案する名詞ベースの選択法の 2 通りに分けて比較を行う。文ベースの選択では Sentence BERT を用いてペルソナ文と入力文をベクトル化し、それらのコサイン類似度の一番高いものを選択する。名詞による選択では、ペルソナ文と入力文に含まれる名詞をそれぞれ word2vec によってエンコードし、それらのコサイン類似度の一番高い名詞を含むペルソナを選択する。それぞれの選択手法を用いて一つのペルソナ文を決定し、このペルソナ文に条件づけられた応答生成を行い、得られたそれぞれの応答文に対してユーザ評価を実施する。評価者は 5 人で各 40 サンプル、TOEIC650 点以上という条件で実験を行った (内一人は英語母語話者)。評価データはテストデータ中から 2 通りの選択方法によって異なるペルソナが選択されたデータ²⁾からランダムに 40 件選択した。評価者には入力となるペルソナ文、ユーザ発話、生成された応答の 3 種類の文を提示し、以下の 2 つの項目について 5 段階で評価を行った。

一貫性 生成した応答に含まれるペルソナが与えられたペルソナに一貫しているか

2) テストデータ 7,793 件中、4,822 件で選択されたペルソナが異なった

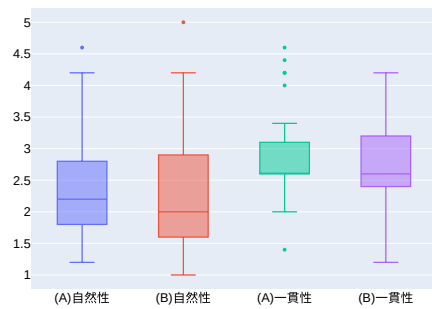


図 6 (A) 文ベースと (B) 名詞ベースの各サンプルの平均値の分布

自然性 文脈に沿ったペルソナを提示できているか

3.1 ユーザ評価結果

ユーザ評価の全体の評価項目ごとの平均値は表 1 に示すように、提案手法である名詞ベースの選択手法の方が低い結果となった。

表 1 ユーザ評価の各サンプルの平均値

	自然性	一貫性
文ベースの選択	2.405	2.845
名詞ベースの選択	2.25	2.805

各サンプルにおける平均値の分布を図 6 に示す。自然性に関しては値にばらつきが見られなかったが、一貫性は提案手法である名詞ベースの選択手法の方が最大値が大きいことが確認できた。

4 考察

2.1 節の実験結果から名詞がペルソナを用いた応答生成に大きな影響を与えることが分かった。PERSONA-CHAT の会話には、他の対話者の発話内容に関連した応答が含まれるため、これは妥当な結果である。そのため、図 3 に示すように、ユーザ発話の名詞語が応答生成において同じ名前の固有表現を持つ名詞語を誘発させると考えられる。また、2.2 節の実験結果から名詞トークンが Hallucination の誘発に最も影響力があることがわかった。この結果は、2.1 節で述べたように、他の対話者の発話内容に関連した応答が含まれる PERSONA-CHAT データセットの特徴も反映していると思われる。

名詞語によるペルソナ選択の優位性を示すために、ユーザの発話とペルソナとのコサイン類似度を用いた文ベースの類似度（赤）と名詞の単語ベースの類似度（青）の比較を図 7 に可視化した。図 7 より、名詞ベースの類似度は文ベースの類似度より広く分布していることが確認できた。分布の広さは 2

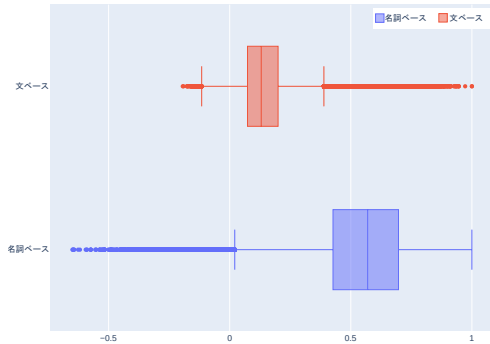


図 7 発話とペルソナのコサイン類似度の文ベースの類似度（赤）と名詞の単語ベースの類似度（青）の分布

文間の違いを捉えられていると考えることができ、名詞ベースの類似度によって効果的な選択が可能であることが示唆される。このことから判断して、ユーザ発話とペルソナ文の名詞ベースの類似度は、選択手法として有効であると考えられる。

ユーザ評価において実際に選択されたペルソナを確認すると、どちらの手法のスコアも低い場合の例として、入力に対応したペルソナが無い場合があった（付録 A）。両方の手法においてペルソナを開示するべきでない場面にもペルソナを開示して自然性が下がるケースが多々見られたため、そのケースへの対応も必要であると考えられる。一方で、そのような場合にペルソナが選択されたとしても、モデルは常にペルソナを含む応答を生成してはいなかった。

以上のことから、PERSONA-CHAT の対話では、開示できる適切なペルソナが無い場合にペルソナを用いると応答の自然性が低下することが確認できた。そのため、入力すべきペルソナが無い時は、ペルソナを選択しない工夫が選択には必要である。図 7 では、名詞による選択の方が優れていることが示唆されたが、ユーザによる評価では文ベースの選択より劣っている結果となった。これは評価データがペルソナを開示できるような入力を多く含んでいなかったため、ペルソナの選択が必要な事例が少なく、評価が正しくできなかった可能性がある。

5 まとめ

本研究では、ペルソナ更新により適したペルソナ選択のために、PERSONA-CHAT データセットを分析した。分析の結果、名詞ベースの選択手法が良い選択であることが示唆され、適切でない場面でペルソナの開示を行うと、応答の自然性が下がるということが確認された。

謝辞

本研究は JSPS 科研費 JP21K17806 の助成を受けたものである。

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 **ACL (1)**, 2018.
- [2] 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.
- [3] 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.
- [4] 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.
- [5] Iulian Vlad Serban, Ryan Lowe, Laurent Charlin, and Joelle Pineau. Generative deep neural networks for dialogue: A short review. **arXiv preprint arXiv:1611.06216**, 2016.
- [6] Oriol Vinyals and Quoc Le. A neural conversational model. **arXiv preprint arXiv:1506.05869**, 2015.
- [7] 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.
- [8] Hongjin Qian, Zhicheng Dou, Yutao Zhu, Yueyuan Ma, and Ji-Rong Wen. 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.
- [9] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Ax-omatic attribution for deep networks. In Doina Precup and Yee Whye Teh, editors, **Proceedings of the 34th International Conference on Machine Learning**, Vol. 70 of **Proceedings of Machine Learning Research**, pp. 3319–3328. PMLR, 06–11 Aug 2017.

表 2 入力に対応したペルソナが無いと考えられる例

入力文：
“that sounds good , i bet you’d be a good teacher .”
ペルソナ文：
1. hey there i’m 23 and i love food
2. i also like to cook but i’m not very good at it
3. i have been trying all types of food everywhere i go
4. i own a yacht and i rent it out when i’m not using it
5. i’ve been traveling the world for a years
入力文：
“lol . i thought maybe my tv watching had fried my brain for a minute .”
ペルソナ文：
1.i am single and with two dogs
2.i do not drink alcohol
3.i like to play chess
4.i love taking bubble baths
入力文：
“i don’t like it either . love music”
ペルソナ文：
1.i hate math class
2.i ride the bus to school
3.i’m 13 years old
4.i’m on the soccer team
5.my brother is older than me

A 付録

Integrated Gradients

応答生成モデルがペルソナを強調した応答を生成する場合や Hallucination を引き起こす場合に、入力のどの部分が応答生成に強く影響するかを Integrated Gradients を用いて調べる。 i^{th} 次元の入力 x とベースライン x' に対する Integrated Gradients は、式 (1) のように定義される。ここで、 $\frac{\partial F(x)}{\partial x_i}$ は i^{th} 番目の $F(x)$ の勾配である。

Integrated Gradients は、デフォルトで 0 に設定されているベースライン x' から入力 x までの勾配を積分し、入力とベースラインの差と積を取ることで寄与度を計算する。文ベクトル x とその各トークン x_i に Integrated Gradients を適用することで、各トークンの寄与度を求めることができる。

$$IG_i(x) ::= (x_i - x'_i) \int_{\alpha=0}^1 \frac{\partial F(x' + \alpha \times (x - x'))}{\partial x_i} d\alpha \quad (1)$$

寄与度分析のための分類器構築

Sect. 2 で述べた Integrated Gradient を用いた分析には分類モデルの構築が必要となる。2つの分析には、それぞれ2つの2値分類モデルを用いた。最初の分析では、ペルソナ使用予測モデルを用い、システムが生成した応答がペルソナ文に強く影響されているかを予測した。2つ目の分析では、Hallucination 予測モデルを用い、システムが生成した応答が、ユーザの発話やペルソナ文から判断して、Hallucination を含んでいる可能性があるかどうかを予測した。

これらの2つのモデルはユーザの発話と複数のペルソナ文からなる2種類の入力で学習させた。学習のために、入力は分割トークン $\langle SEP \rangle$ を用いて次のように結合される。

$$\langle CLS \rangle \text{ utterance } \langle SEP \rangle P_1$$

$$\langle SEP \rangle \dots \langle SEP \rangle P_N \langle SEP \rangle$$

ここで P_N は N 番目のペルソナを表す。また、分類器の学習のために、事前学習された BERT を PERSONA-CHAT を用いてファインチューニングした。2つの分類モデルを学習するためのターゲットラベルは、自動でアノテーションを行った。発話文とペルソナ文のどれかが同じ名詞を含んでいる場合に 1、それ以外に 0 をラベル付けした。例えば、ペルソナ「i like shoot a bow」と発話「i have bow」は「bow」という名詞を共有しているため、「1」とラベル付けした。結果として、131,431 件中ラベル 0 は 84,440、ラベル 1 は 46,991 となった。

Hallucination 分類モデルでは、応答と発話文が次の4つのルールを満たす場合、1 とラベル付けした。

1. 応答文が 4-20 語で構成される
2. 応答文が “I” か “my” を含む
3. 応答文に動詞、名詞、代名詞、形容詞のうち少なくとも 1 つの単語が含まれている
4. 発話文と応答文が同じ名詞を含んでいない

それ以外の場合は 0 とした。結果として、ラベル 0 が 89,605 件、ラベル 1 が 41,826 となった。

入力に対応したペルソナが無いと考えられる例

入力文に対して、応答する文に含められそうなペルソナが無かった例を表 2 に示す。

多くの場合、これらは聞かれたことに対するペルソナが無い場合 (表 2 一つ目) や、ユーザに関する話題が中心となっているものであった。