大規模汎用言語モデルによるペルソナを考慮した応答生成

川本 稔己 [1,2]　山崎 天 [1]　佐藤 敏紀 [1]　奥村 学 [2]
[1]LINE 株式会社　[2] 東京工業大学
toshiki.kawamoto@linecorp.com

概要

対話における応答生成では, 自分のこれまでの発話内容や話者のペルソナとの一貫性を保つことが重要である. 本稿では長期間行われる対話を想定し, 従来のペルソナ対話システムでは十分考慮されなかった対話中のペルソナの変動性に着目する. 変動するペルソナを追跡する過程を「抽出」・「選択」・「更新」の 3 つのタスクに分割し, それぞれを解くことで応答を生成する対話システムを提案する. 提案手法は大規模汎用言語モデルである HyperCLOVA を用いて複数タスクを解く. 評価実験の結果から, 提案手法は従来のペルソナ対話システムより高い一貫性スコアを得られ, 長期間行われる対話の一貫性を保つために有効であることを確認した.

1　はじめに

従来の対話システムは, 対話履歴や話者のペルソナと矛盾した発話を生成してしまうことがある. ペルソナとは好みやプロフィールなど個性を表す短文のことであり, 従来の研究においては対話履歴やペルソナと一貫性のある応答を生成することが課題であった. 本稿では, 大規模汎用言語モデル HyperCLOVA [1] を用いてその課題に取り組む.

関連研究として, Zhang ら [2] は Persona Chat データセットを提供した. Persona Chat データセットには, 複数話者のペルソナとそのペルソナに沿った対話が含まれている. そのデータを利用することでペルソナに沿った応答生成を学習, 評価することが可能となり, 対話と自分のペルソナを用いて応答を生成する手法が提案されている [3, 4]. しかし, 長期間行われる対話では対話履歴が増加し続けるため, 全てをシステムへの入力として扱うことが難しい. そのため, 対話履歴から自分のペルソナを抽出する取り組みが行われている [5, 6]. しかし, これらの取り組みには, (Ⅰ) 相手のペルソナには注目できていない, (Ⅱ) ペルソナの数が多くなるとペルソナを選択する必要がある, (Ⅲ) ペルソナの変化には対応していない, といった欠点がある. そこで, 本稿では従来のペルソナ対話システムで十分考慮できていない対話中のペルソナの変動性に着目し, 変動するペルソナを追跡する過程を「抽出」・「選択」・「更新」の 3 つのタスクに分割する. そして, これらのタスクから得られたペルソナを使用して応答を生成するシステムを提案する. 本システムは複数のタスクを同時に解く必要があるが, それぞれにモデルを用意するのは多くの時間とコストがかかる. そのため, これらのタスクを大規模汎用言語モデルである HyperCLOVA を用いて解く. HyperCLOVA とは, GPT-3 [7] と同様な性能・性質をもつ日本語に特化した言語モデルである. HyperCLOVA はショットと呼ばれる特定の言語タスクのサンプルを複数個埋め込んだプロンプトを作成し, その形式を学習することで特定の言語タスク処理能力の獲得が期待できる.

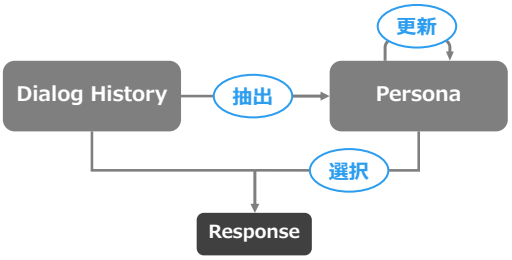

図 1　提案手法. 対話履歴からペルソナを抽出する. 抽出したペルソナに既存のペルソナからの変化が見られた場合更新する. ペルソナ集合から応答生成に必要なペルソナを選択し, 選択されたペルソナと対話履歴を用いて応答を生成する.

対話システムの一貫性を評価するため, JPersona Chat データセット[1)]を用いた実験を行った. 自動評価の結果から大規模汎用言語モデルはベースラインモデルより一貫性がある応答生成を行うこと, 提案手法はペルソナをより考慮した応答を行うことを確認した. また, 長期間行われる対話を想定した人手評

[1)] https://github.com/nttcslab/japanese-dialog-transformers

```
===  # ショット
ユーザ：最近血液型を知ったんですよ。 # 対話例
クローバ：思った通りの血液型でしたか？
ユーザ：いえ、私の予想とは違いました。
クローバ：そうなんですね、何型だったんですか？
ユーザ：AB 型だったんです。 # 対象となる発話例

[ユーザの属性情報] 血液型は AB 型。 # 抽出ペルソナ例
===  # ショット数は 5
...
===  # 進行中の対話
ユーザ：...  # 対話履歴
クローバ：... # 最長で 6 発話
ユーザ：  ... # 最後の発話がペルソナの抽出の対象

[ユーザの属性情報]# ここから生成開始
```

図 2　ペルソナの抽出を行うためのプロンプトのテンプレート. #以降はコメントである.

```
===  # ショット
既存ペルソナ：煮込み料理を作る。
新規ペルソナ：煮込み料理を作った。
更新後ペルソナ：煮込み料理を作った。
===
既存ペルソナ：りんごが好き。
新規ペルソナ：みかんが好き。
更新後ペルソナ：りんごが好き。みかんが好き。
===  # ショット数は 5
...
===  # 更新の対象とするペルソナ対
既存ペルソナ：... # 関連のあるペルソナを記述
新規ペルソナ：... # 抽出したペルソナを記述
更新後ペルソナ：# ここから生成開始
```

図 3　ペルソナの更新を行うためのプロンプトのテンプレート. #以降はコメントである.

価の結果, 提案手法は従来のペルソナ対話システムより高い一貫性スコアを得られ, ペルソナの変動性を考慮した提案手法の有効性を確認した.

2　提案手法

従来の対話システムでは, 対話と初めに与えられた自分のペルソナを用いて応答を生成する. しかし, 一貫した応答生成のためには対話中に新たに自分のペルソナが発生することや, 相手のペルソナを考慮する必要がある. よって, 本稿では対話から自分と相手のペルソナを抽出することを試みる (ペルソナの抽出). しかし, ペルソナの抽出を行えばペルソナの数が増加する. その結果, 対話の話題と関連性が低いペルソナが応答生成時のノイズになる可能性や, 言語モデルの入力長に収まらない可能性がある. そのため, 最適なペルソナを選択する必要がある (ペルソナの選択). また, 対話が長くなるとペルソナに変化が起こることも想定される. よって, ペルソナの変化を検知し, 必要に応じて更新を行う (ペルソナの更新). まとめると, 本システムはペルソナの「抽出」・「選択」・「更新」を行った上で応答を生成する. 概要を図 1 に示す.

2.1　ペルソナの抽出

対話から話者のペルソナを抽出するのには HyperCLOVA を用いる. HyperCLOVA に与えるプロンプトのテンプレートを図 2 に示す. 1 つのショットには対話の最終発話からペルソナを抜き出し,"[ユーザの属性情報]" 以降に記述する. ショットには, 対話の最終発話にペルソナが含まれていないパターンも記述し, その場合"[ユーザの属性情報]"以降には何も記述しない. このプロンプトを用いて, 対話履歴の発話それぞれが抽出の対象となるように HyperCLOVA を実行し, 自分と相手のペルソナを抽出する.

2.2　ペルソナの選択

ペルソナの抽出を行ったことで保持するペルソナの数が増加する. そのため, 応答生成の際には適切なペルソナを選択する必要がある. ペルソナの選択には直前の対話を用いる. 具体的には, 直前に行われた自分と相手の発話を時系列順に繋ぎ合わせ, それを検索文字列とする. その検索文字列の埋め込み表現と, 保持しているペルソナの埋め込み表現とのコサイン類似度を計算し, 自分と相手のペルソナからそれぞれコサイン類似度の上位 5 つを使用する.

2.3　ペルソナの更新

本節では, 既存のペルソナを抽出したペルソナに置き換えるかどうか決定する手法について述べる. まず, 抽出したペルソナと関連のあるペルソナが既存のペルソナの中に存在するか調べる. 抽出したペルソナの埋め込み表現と既存のペルソナそれぞれの埋め込み表現の中からコサイン類似度が最も高いペルソナを選択し, コサイン類似度が閾値 (0.7) 以上ならその選択されたペルソナを関連のあるペルソナとする. 一方, コサイン類似度が閾値以下の場合抽出したペルソナは既存のペルソナと中立な関係であるとし, ペルソナの更新は行わず新しいペルソナとして追加する. 次に, 関連のあるペルソナを抽出したペルソナに置き換えるか HyperCLOVA を用いて決定する. プロンプトのテンプレートを図 3 に示す. ここで, 関連のあるペルソナと抽出したペルソナには以下の 3 つの関係が考えられる: (a) 等しい, (b) 変更が起きている, (c) 中立. (a), (b) の場合は関連のあるペルソナを抽出したペルソナに置き換える必要があるので, 抽出したペルソナを出力する. (c) の場合は関連のあるペルソナを置き換える必要がないので, ど

```
=== # ショット
クローバのペルソナ：私は早起きが苦手です。私は高校生です。私はピーマンが嫌いです。私はおばあちゃん子です。私は陸上部に入っています。
ユーザのペルソナ：私は以前、沖縄に住んでいたことがあります。私は海のそばに住んでいます。私はダンスが得意です。私は怒りっぽいです。私が尊敬する人は、母親です。

ユーザ：そうなんだ！近いね！わたし、沖縄に住んでたことがあるんだけど、この辺も似てるね。
クローバ：そうなんだね。じゃあここでは沖縄気分が味わえるわけだ。
ユーザ：うん。ちょっとだけだけど、そんな感じがするよ！ところで、わたし、ダンスが得意なんだけど、あなたは何が得意？
クローバ：私は陸上部だから、走るのが得意かな。短距離走よりも長距離走が自信ある。
=== # ショット数は 5
...
=== # 進行中の対話
クローバのペルソナ：...
ユーザのペルソナ：...

ユーザ：... # 直近の対話を最大 3 発話
クローバ：...
ユーザ：...
クローバ：# ここから生成開始
```

図 4 ペルソナを考慮した応答生成のためのプロンプトのテンプレート. #以降はコメントである.

ちらのペルソナも出力する. (a), (b), (c) 全てのパターンをショットに記載し, 得られた出力を新しいペルソナとして追加する.

2.4 ペルソナを利用した応答生成

ペルソナの抽出 (§2.1)・更新 (§2.3) を行い, ペルソナの選択 (§2.2) で選択されたペルソナを用いて応答を生成する. 応答の生成には HyperCLOVA を用いる. プロンプトのテンプレートを図 4 に示す. ショットの前半がペルソナ, 後半が対話である. ショットに与えるペルソナと対話は JPersona Chat の訓練データから 5 つ使用した. また, 情報量が少なくつまらない応答となることを避けるため, 生成した応答の長さが短い場合にはその応答の続きとなる文を生成する. 具体的には, 生成した応答が 12 文字 (事前の観察に基づく) 以下の場合, 図 4 のショットを図 5 のように変更し, 追加の文を生成する. そして, その文を元の応答に繋げて最終的な応答とする.

3 実験

3.1 実験設定

本稿で使用した HyperCLOVA のパラメータ数は 39B で, 1.8TB の日本語データで学習を行っている. 埋め込み表現を入手するためには Universal Sentence Encoder [8] を用い, コサイン類似度を計るためには FAISS [9] を用いた. データセットは JPersona Chat を用いる. このデータは Persona Chat を日本語に変換

```
クローバのペルソナ：私は早起きが苦手です。私は高校生です。私はピーマンが嫌いです。私はおばあちゃん子です。私は陸上部に入っています。
ユーザのペルソナ：私は以前、沖縄に住んでいたことがあります。私は海のそばに住んでいます。私はダンスが得意です。私は怒りっぽいです。私が尊敬する人は、母親です。

ユーザ：そうなんだ！近いね！わたし、沖縄に住んでたことがあるんだけど、この辺も似てるね。
クローバ：そうなんだね。
[続き] じゃあここでは沖縄気分が味わえるわけだ。
ユーザ：うん。ちょっとだけだけど、そんな感じがするよ！ところで、わたし、ダンスが得意なんだけど、あなたは何が得意？
クローバ：私は陸上部だから、走るのが得意かな。
[続き] 短距離走よりも長距離走が自信ある。
```

図 5 複数文を生成するプロンプトに与えるショット例.

したデータセットで, 話者同士のペルソナと最大 6 ターンの対話が含まれている. データの分割は元の分割方法[1)]に従い, テストデータを用いて応答を生成し実験を行った. 実験は以下のモデルで比較を行う.

BlenderBot ベースラインの Transformer ベースのペルソナ対話システム [3]. 事前学習に加えて JPersona Chat の訓練データで fine-tuning を行っている [10].

w/o Persona 対話履歴のみを利用して HyperCLOVA で応答を生成するモデル.

w/ Persona 対話履歴と初めに与えられたペルソナを利用して HyperCLOVA で応答を生成するモデル.

以下のモデルが提案手法を利用したモデルである.

Extract w/ Persona に加えて, 2.1 節のペルソナの抽出を行ったモデル.

Extract+Select Extract に加えて, 2.2 節のペルソナの選択を行ったモデル.

Extract+Select+Update Extract+Select に加えて, 2.3 節のペルソナの更新を行ったモデル.

応答生成時には, 自分のペルソナとそれまでの対話履歴を用い, 相手のペルソナは使用していない. 対話履歴はベースライン手法の BlenderBot と合わせるために直前の 3 発話を入力に加えた.

3.2 自動評価

評価指標は応答と正解文との一致度を測るために BLEU [11], 応答の多様性を測るために Dist-1, Dist-2 [12] を使用する. 一貫性を測るための指標としては, 生成した応答とペルソナの間で自然言語推論（NLI）を行うことによって導出される Consistency Score (C.Score) [13] がある. しかし, 手法によって推論の対象となるペルソナの数が大きく異なることから, 従来の方法で比較することはできない. よって本

表 1　自動評価結果

	BLEU	Dist-1	Dist-2	含意 (%)	中立 (%)	矛盾 (%)
BlenderBot	**3.34**	2.03	5.75	21.47	59.97	18.56
w/o Persona	1.32	2.52	7.16	22.88	55.55	21.57
w/ Persona	1.71	2.84	7.82	29.90	52.10	**18.00**
Extract	1.59	2.83	7.88	**31.71**	48.33	19.95
Extract+Select	1.56	2.87	7.92	30.82	49.82	19.36
Extract+Select+Update	1.60	**2.97**	**8.19**	31.38	48.67	19.96

表 2　人手評価結果

	関連性	一貫性
BlenderBot	2.24	1.84
w/ Persona	2.45	2.20
Extract+Select+Update	**2.56**	**2.36**

稿では, 生成した応答が生成のために用いた自分のペルソナを含意/中立/矛盾している割合を NLI 分類器で評価する. NLI 分類器には, 日本語で学習された BERT [2] [14] を NLI データセットで fine-tuning したモデルを用いた. NLI データセットとしては発話とペルソナのペアで構成される DNLI [15] を用いるのが本稿の設定には適しているが, 日本語ではないため, 本稿では日本語で記述された SNLI [16] である JSNLI [17] を用いた. JSNLI はテストデータを提供していないため, 元の訓練データを訓練データと検証データに分割し, 元の検証データをテストデータとした. 正解率は 92.31%であった.

結果を表 1 に示す. 従来のペルソナ対話システムの手法を大規模汎用言語モデルに適応させた w/ Persona はベースラインである BlenderBot より含意の割合が高く, 矛盾の割合が低いことから一貫性がある応答生成を行うことを確認した. 抽出などの提案手法を利用したモデルは w/ Persona と比較して Dist-1, Dist-2 の値が高いことから, Dull Response と呼ばれるつまらない応答ではなく, ペルソナを考慮した応答を生成している割合が高いと考えられる. それにより, 含意の割合が増加している一方で矛盾の割合も増加している.

3.3　人手評価

長期間行われる対話の一貫性は自動評価だけでは十分に測ることができないため, データセットの続きとなる対話を行い, その応答に対して人手評価を行った. 具体的には, テストデータからランダムに 10 対話を選び, その対話から 1 週間後の対話を想定して 10 ターンの対話を行った. 3 つのモデル (BlenderBot, w/ Persona, Extract+Select+Update) で対話のそれぞれの応答を評価した. 評価指標は関連性と一貫性の 2 つで, 関連性は前の発話またはペルソナと関連性のある応答ができているかを示す. 一貫性は生成した応答が対話履歴とペルソナに矛盾していないかを示す. それぞれ 3 段階で 1 点 (悪い) から 3 点 (良い) を付与した. その平均点を最終的な結果とし表 2 に示す.

評価結果から, 提案手法は関連性, 一貫性ともに最も高いスコアを得られ, 提案手法が有効であることを確認した. 今回の実験では, 提案手法以外のモデルは新しいペルソナの発生や, 初めに与えられたペルソナからの変化を考慮できなかった可能性がある.

3.4　ペルソナの抽出の評価

表 1 の結果から, ペルソナの抽出を利用したモデルは w/ Persona と比較して矛盾の割合が増加したので, ペルソナの抽出の性能を人手で評価した. テストデータからランダムに 100 件の発話を選び, 抽出したペルソナが正しいかを二値で判定した. その結果, 正しく抽出されたのが 71 件, 誤って抽出されたのが 29 件となった. ペルソナの抽出は対話履歴の全発話に対して行うので, 対話の長さが長くなれば全抽出結果が正しい確率が低くなることから, 抽出性能が十分であるとは言えない. 抽出の性能を改善することで提案手法の性能も向上することが予測される.

4　おわりに

本稿では, 一貫性がある対話を行うにはペルソナを動的に「抽出」・「選択」・「更新」する必要があるという仮説の元, 大規模汎用言語モデルを利用してシステムを設計した. 評価実験の結果, 提案手法は初めに与えられたペルソナのみを利用したモデルより高い一貫性スコアを得られ, 提案手法の有効性を確認できた. しかし, 提案手法は大規模汎用言語モデルに依存した実装やタスク設計になっており, 特に精度面で不十分な部分があったので, 今後はその改善に取り組んでいきたい.

[2] https://huggingface.co/cl-tohoku/bert-base-japanese-v2

参考文献

[1] Boseop Kim, HyoungSeok Kim, Sang-Woo Lee, Gichang Lee, Donghyun Kwak, Jeon Dong Hyeon, Sunghyun Park, Sungju Kim, Seonhoon Kim, Dongpil Seo, Heungsub Lee, Minyoung Jeong, Sungjae Lee, Minsub Kim, Suk Hyun Ko, Seokhun Kim, Taeyong Park, Jinuk Kim, Soyoung Kang, Na-Hyeon Ryu, Kang Min Yoo, Minsuk Chang, Soobin Suh, Sookyo In, Jinseong Park, Kyungduk Kim, Hiun Kim, Jisu Jeong, Yong Goo Yeo, Donghoon Ham, Dongju Park, Min Young Lee, Jaewook Kang, Inho Kang, Jung-Woo Ha, Woomyoung Park, and Nako Sung. What changes can large-scale language models bring? intensive study on HyperCLOVA: Billions-scale Korean generative pretrained transformers. In **Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing**, pp. 3405–3424, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[2] 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.

[3] Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. Recipes for building an open-domain chatbot. In **Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume**, pp. 300–325, Online, April 2021. 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] Jing Xu, Arthur Szlam, and Jason Weston. Beyond goldfish memory: Long-term open-domain conversation. **arXiv preprint arXiv:2107.07567**, 2021.

[6] 吉田快, 品川政太朗, 須藤克仁, 中村哲. 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. 人工知能学会研究会資料 言語・音声理解と対話処理研究会, Vol. 93, pp. 32–37, 2021.

[7] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. **arXiv preprint arXiv:2005.14165**, 2020.

[8] Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Céspedes, Steve Yuan, Chris Tar, et al. Universal sentence encoder. **arXiv preprint arXiv:1803.11175**, 2018.

[9] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with gpus. **arXiv preprint arXiv:1702.08734**, 2017.

[10] Hiroaki Sugiyama, Masahiro Mizukami, Tsunehiro Arimoto, Hiromi Narimatsu, Yuya Chiba, Hideharu Nakajima, and Toyomi Meguro. Empirical analysis of training strategies of transformer-based japanese chit-chat systems, 2021.

[11] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In **Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics**, pp. 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics.

[12] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. A diversity-promoting objective function for neural conversation models. In **Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 110–119, San Diego, California, June 2016. Association for Computational Linguistics.

[13] Andrea Madotto, Zhaojiang Lin, Chien-Sheng Wu, and Pascale Fung. Personalizing dialogue agents via meta-learning. In **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics**, pp. 5454–5459, Florence, Italy, July 2019. Association for Computational Linguistics.

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)**, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[15] Sean Welleck, Jason Weston, Arthur Szlam, and Kyunghyun Cho. Dialogue natural language inference. In **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics**, pp. 3731–3741, Florence, Italy, July 2019. Association for Computational Linguistics.

[16] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. In **Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing**, pp. 632–642, Lisbon, Portugal, September 2015. Association for Computational Linguistics.

[17] 卓見吉越, 大輔河原, 禎夫黒橋. 機械翻訳を用いた自然言語推論データセットの多言語化. Technical Report 6, 京都大学, 京都大学／現在，早稲田大学, 京都大学, jun 2020.