適応的対話システムのための終盤の会話を予測する埋め込みモデルの構築

飯塚慎也　　東中竜一郎
名古屋大学大学院情報学研究科
iizuka.shinya.a8@s.mail.nagoya-u.ac.jp
higashinaka@i.nagoya-u.ac.jp

概要

本研究では，対話相手に適応した応答生成を目指し，序盤の対話から，その話者との終盤の対話を予測するための埋め込みモデルを構築する．具体的には，対照学習を利用し，話し方に基づいて序盤と終盤とを紐づける埋め込みモデルを学習する．さらに，最近傍探索ライブラリである Faiss を用いて序盤の対話から，類似した話し方を持つユーザの終盤の対話を獲得する仕組みを構築する．本手法の評価では，旅行代理店タスク対話コーパス（Tabidachi）および RealPersonaChat（RPC）コーパスを使用し，提案モデルが序盤の対話から終盤の対話を予測するための埋め込み表現の獲得に有効であることを確認した．

1　はじめに

大規模言語モデル（Large Language Model; LLM）の登場により，対話システムの性能は飛躍的に向上している [1, 2, 3, 4]．そのような対話システムが普及していく中で，人間が対話において行う，相手の性別や年代，性格に合わせた話し方の調整は，タスクの達成や満足度向上において，より重要となっている．これまでに，相手に適応する対話システムを構築する取り組みはいくつか提案されている [5, 6, 7]．しかし，従来の対話システムでは，適応のルールや話し方の調整方法を事前に設計する必要があるため，未知のユーザに柔軟に適応することが難しい．

我々は，対話相手の話し方に着目し，序盤の対話から対話相手との終盤の対話を予測することで，対話相手に適応した応答生成を実現する手法を提案する．人間の対話では，進行に伴い互いの話し方が調整され，親密さや信頼感が築かれる [8]．また，会話

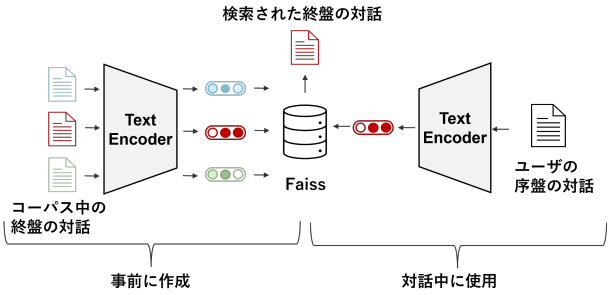

図 1　対話相手に適応した応答生成を行う手法の概要．

における言語スタイルの適応が，相手への関与や相互理解を高めるうえで重要な役割を果たすことも示唆されている [9]．本研究では，序盤の対話から終盤の話し方を予測し，その結果を応答生成に活用することで，通常対話を重ねる中で得られる適応を短時間で実現することを目指す．

図 1 に手法の概要を示す．本手法では，対照学習 [10] を利用し，話し方に基づいて序盤と終盤とを紐づける埋め込みモデルを学習する．さらに，最近傍探索ライブラリである Faiss[1] [11] を用いて序盤の対話から類似した話し方を持つユーザの終盤の対話を獲得する仕組みを構築する．実際の対話時には，蓄積された対話履歴から，対話相手との終盤の対話となるものを検索し，それを利用して対話相手に適応した発話を実現する．例えば，得られた対話をプロンプトにショットとして含めるなどが想定される．

本稿では，序盤の対話から，その話者との終盤の対話を予測するための埋め込みモデルの構築と評価を行った．評価の結果，対照学習を導入することで，終盤の対話に対する検索の精度が向上することを確認した．特に，一つの学習データに含める発話数を増やした場合に大幅な精度向上が見られ，長い文脈情報が話し方の特徴を捉える上で有効であることが示された．

1）https://github.com/facebookresearch/faiss

2　関連研究

2.1　ユーザに適応する対話システム

Komatani らは，ユーザの「システムに対する習熟度」，「ドメインに関する知識レベル」，「性急度」を推定してユーザに応じてシステムの振る舞いを変化させるシステムを構築した [5]．また，Ohashi らは，強化学習を活用して環境ノイズやユーザの語彙レベルに適応する発話生成を提案した [6]．さらに，Yamamoto らは，ユーザの性格に応じたシステムのキャラクタ表現を設計し，対話の印象向上を目指した [7]．しかし，従来の対話システムでは，適応のルールや話し方の調整方法をあらかじめ設計する必要があるため，未知のユーザに対して柔軟に適応することが難しい．本研究では，序盤の対話から，その話者との終盤の対話を予測するための埋め込みモデルを構築する．これにより，未知のユーザにも適応可能な対話システムの実現を目指す．

2.2　話し方に着目した埋め込みモデル

Akama らは，「同一発話内に含まれる単語は同一のスタイルを持つ」という仮定のもと，スタイルに敏感な単語埋め込みを学習することで，発話全体のスタイルを反映した埋め込み表現を生成する手法を提案した [12]．Zenimoto らは，対照学習を用いて日本語の多様な話し方を表現するスタイル埋め込みモデルを構築した [13]．これらの研究は，ユーザの話し方に焦点を当てることで，従来よりも適応性の高い対話システムの構築に貢献する可能性がある．本研究では，これらの研究と同様にユーザの話し方に着目し，序盤の対話と終盤の対話を紐づけるような埋め込みモデルを構築する．

3　ユーザの話し方に基づく対照学習

本研究では，対照学習を用いて，序盤の対話から，その話者との終盤の対話を予測するための埋め込みモデルの構築を目指す．本研究における対照学習の概要を図 2 に示す．あるユーザの対話データを 1/3 ずつ分割し，それぞれを序盤対話（D_{start}），中盤対話（D_{middle}），終盤対話（D_{end}）とする．同じ話者のみからデータを作成することで，話し方だけに着目し，内容に依存しない話し方の変化の特徴を捉える．アンカーとなる発話集合を D_{start} から窓幅 n で取り出し，それらの発話を特殊トークン $[SEP]$ で

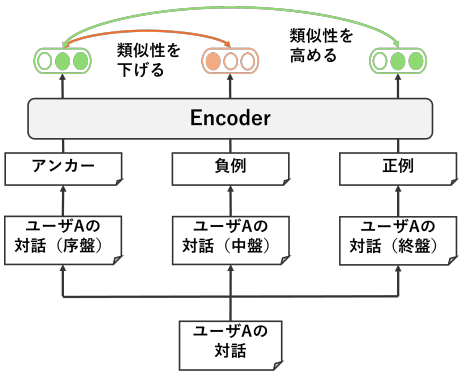

図 2　ユーザの話し方に基づく対照学習の概要．序盤の対話（アンカー）を基に，終盤（正例）との類似性を高め，中盤（負例）との類似性を下げるように学習を行う．

連結することで作成する．同様に正例となる発話集合，負例となる発話集合をそれぞれ D_{end}，D_{middle} から作成する．これらの文を Encoder に入力し，それぞれ埋め込みベクトル (E_a, E_p, E_n) を生成する．対照学習では，Triplet Margin Loss [2] [14] を用いて，以下の損失関数を最小化する．

$$L = \max\left\{\|E_a - E_p\|_2 - \|E_a - E_n\|_2 + \alpha, 0\right\}$$

ここで α はマージン（ハイパーパラメータ）である．

この学習によって，同一対話内のアンカーと正例の埋め込みが近くなる一方，負例の埋め込みは遠ざかるように調整される．同一対話内で正例と負例を生成することで，対話内容の影響を抑え，話し方の特徴のみを捉える埋め込みモデルを構築することが可能となる．また，この埋め込みモデルを利用しデータベースを作成することで，序盤の対話から類似した話し方を持つユーザの終盤の対話を獲得できると考える．

4　実験

4.1　学習データの作成

初対面の対話から始まり，徐々に相手に適応していく様子が記録された対話データを対象に実験を行った．具体的には，旅行代理店タスク対話コーパス（Tabidachi） [15] と RealPersonaChat コーパス（RPC） [16] の 2 種類を利用した．

Tabidachi は，観光プランに関する相談の対話を記録した音声対話の書き起こしコーパスである．このコーパスには，カスタマとオペレータの間で交わされた対話が含まれており，オペレータはカスタマの

2）https://pytorch.org/docs/stable/generated/torch.nn.TripletMarginLoss

発話内容に応じて適応的に対話を進める．カスタマ役は，一般 25 名，高齢者 10 名，子供 20 名で構成され，合計で 55 ペアの対話が収録されている．本研究では，コーパス中のペアを train，validation，test に 45:5:5 の割合で分割して使用した．

RPC は，話者のペルソナや性格特性を含む雑談対話を記録したテキスト対話コーパスである．このコーパスには，233 人の参加者による合計 1572 ペアの対話が含まれている．話者の年齢構成は，20 代，30 代，40 代がそれぞれ約 30%を占め，幅広い話者層の多様な対話が収録されている．RPC の収集には対話相手に適応的に進めるという意図は含まれていないが，収録された対話が初対面の話者同士で行われている点に着目して採用した．初対面から徐々に相手に合わせていく話し方の変化を捉えるため，本研究において適したデータセットであると判断した．同様に，コーパス中のペアを train，validation，test に 1272:150:150 の割合で分割して使用した．

実験では，各対話からアンカー 5 個，正例 5 個，負例 5 個を窓幅 n をずらしながら抽出した．具体的には，D_{start}，D_{end}，D_{middle} からそれぞれ発話集合を取り出し，それらの発話を特殊トークン $[SEP]$ で連結した．窓幅 n は Tabidachi では 5 と 10 の 2 種類を用い，RPC では窓幅 5 のみを用いた．この設定は，RPC の 1 対話あたりの発話数が約 30 発話程度であり，窓幅を広げると多様なデータを生成することが困難なためである．一方，Tabidachi は 1 対話あたりの発話数が約 360 発話と多くの発話が収録されているため，窓幅 10 を適用することで多様性のある学習データを生成可能であった．結果として，1 対話あたり合計 125 個の学習データを生成した．

4.2 モデルの学習

本研究では，ベースモデルとして東北大学が公開している日本語 $BERT_{base}$[3] [17] を使用した．このモデルは，日本語に特化した事前学習済みの BERT モデルであり，日本語の文脈情報を効果的に捉えることができる．また，このモデルは提案モデルの基盤となるだけでなく，比較モデルとしても使用した．

埋め込みモデルの構築では，このベースモデルのパラメータを初期値として利用し，提案手法に基づいてデータセットに対する対照学習を行った．具体的には，対話データからアンカー，正例，負例を抽出し，Triplet Margin Loss を使用して学習を進めた．学習時のハイパーパラメータとしては，バッチサイズを 64，オプティマイザには AdamW を選択し，学習率は 1e-05 に設定した．また，Triplet Margin Loss のマージンは 1 とした．

さらに，比較モデルとして ファインチューニングモデル（FT モデル）を構築した．このモデルは，ベースモデルをデータ量が比較的多い RPC データセットで事前学習した後，Tabidachi データセットで追加学習を行ったものである．この二段階学習プロセスにより，学習データ量が増加し，埋め込みモデルの表現力や性能の向上が期待される．

4.3 評価

生成された埋め込みの有効性を評価するために Cosine Similarity を用いた評価を設計した．具体的には，埋め込みベクトル間のコサイン類似度 (cos) を用いて，以下の条件を満たすテストデータの割合 P を精度として計算した．

$$P = \frac{\text{個数}\left(\cos(E_a, E_p) > \cos(E_a, E_n)\right)}{\text{総テストデータ数}}$$

ここで，E_a は対話序盤のアンカーの埋め込み，E_p は対話終盤の正例の埋め込み，E_n は対話中盤の負例の埋め込みを表す．

この評価指標によって，埋め込みモデルが対話内容ではなく，終盤の話し方や発話スタイルに基づいた特徴を捉えているかどうかを検証する．

仮にモデルが対話の内容に依存して埋め込みを作成している場合，正例 (E_p) と負例 (E_n) の間で類似度がほぼ同じになると想定される．これは，学習データを同じ対話の中から作成しているためである．一方，話し方や発話スタイルに着目した埋め込みモデルが適切に機能している場合，正例 (E_p) はアンカー (E_a) と類似性が高く，負例 (E_n) とは類似性が低くなると期待される．特に，提案モデルでは，対話内容に依存せずに話し方や発話スタイルを埋め込み空間上で反映することで，正例とアンカーのコサイン類似度 ($\cos(E_a, E_p)$) が負例との類似度 ($\cos(E_a, E_n)$) を大きく上回る結果が得られると予想される．

3） https://huggingface.co/tohoku-nlp/bert-base-japanese-v3

表 1　Tabidachi における Cosine Similarity 評価の結果．それぞれの窓幅 (n) において最も高いスコアを太字で表す．

窓幅 (n)	モデル	精度 (%)
5	比較モデル	53.4
	提案モデル	**55.5**
	FT モデル	51.4
10	比較モデル	58.7
	提案モデル	**71.2**
	FT モデル	63.0

表 2　RPC における Cosine Similarity 評価の結果．最も高いスコアを太字で表す．

窓幅 (n)	モデル	精度 (%)
5	比較モデル	49.4
	提案モデル	**61.5**

4.4　結果

Tabidachi および RPC を用いて提案モデルの性能を評価した．表 1 に，Tabidachi を用いた実験結果を示す．窓幅 5 では 55.5%，窓幅 10 では 71.2%の精度を達成し，ベースモデルに比べてそれぞれ 2.1 ポイント，12.5 ポイントの改善を確認できた．窓幅を大きくすることで精度が向上する傾向が見られ，長い文脈を取り込むことで対照学習の効果がより顕著に現れることを示している．

一方で FT モデルは RPC での事前学習を行ったにもかかわらず，提案モデルよりも低い精度となった．これは，RPC が主にテキスト対話で長い応答が特徴であるのに対し，Tabidachi は音声対話で相槌など短い応答を含むというデータセット間の特性の違いが，適応を難しくしたことの原因と考えられる．

表 2 に，RPC を用いた実験結果を示す．提案モデルでは 61.5%の精度を達成し，比較モデルに比べて約 12 ポイントの改善を確認できた．この結果は，提案モデルが発話の内容ではなく，話し方に着目した埋め込みを生成できていることを示唆している．

4.5　検索の事例

本研究では，構築した埋め込みモデルを用いて，序盤の対話から類似した話し方を持つユーザの終盤の対話を検索する仕組みを構築した．検索には，Faiss [11] を利用した近似最近傍探索を採用し，対話データベースから類似した話し方を持つ対話を検索できるようにした．Tabidachi コーパスに含まれる子供（8 歳）の序盤の対話を検索クエリとして使用した検索事例を表 3 に示す．この結果は，検索クエリとして与えた 8 歳の子供の話し方に近い 10 歳の子供の対話が検索されていることを示している．検索結果に含まれる発話では，「釣りをしますっていうのと、パンケーキ食べますって決めたよね。」のような子供に合わせた言葉遣いが観察された．このことから，提案モデルは話し方に基づいた埋め込みを適切に学習しており，特定の話し方に対応した類似対話の検索が可能であることが確認できた．

表 3　検索の事例

クエリ：子供の対話（8 歳，対話 ID: 306_1_1）
<オペレータ>こんにちは。 [SEP] <カスタマ>こんにちはー。 [SEP] <オペレータ>今日はご利用いただきましてありがとうございます。 [SEP] <カスタマ>あ、こちらこそ。 [SEP] <オペレータ>はい、で、今日は旅行の相談でよろしいですか？ [SEP] <カスタマ>はい。 [SEP] <オペレータ>はい、ありがとうございます。どこに旅行に行くんですか？ [SEP] <カスタマ>北海道です。 [SEP] <オペレータ>北海道ですね。いいですね、北海道。北海道って行ったことありますか？ [SEP] <カスタマ>あ、はい。
検索結果：子供の対話（10 歳，対話 ID: 310_1_1）
<オペレータ>そっか、じゃあ、これ、今ね、スキーをしますっていうのと、釣りをしますっていうのと、パンケーキ食べますって決めたよね。 [SEP] <カスタマ>はい。 [SEP] <オペレータ>他、何か、お父さんとかお母さん、こんなことしたいって言ってなかった？ [SEP] <カスタマ>言っていません。 [SEP] <オペレータ>言っていません。 [SEP] <オペレータ>じゃあ他に、お客様自身が何か食べてみたいのとか、そういうの、決めていこっか。 [SEP] <カスタマ>はい。 [SEP] <オペレータ>どうしよう、じゃあ、他、何か、食べ物決めるのがいい？それとも行く場所がいい？ [SEP] <オペレータ>スキーとか釣り以外、何かしてみたい？ [SEP] <カスタマ>お寿司でも食べたい。

5　おわりに

本稿では，序盤の対話から，その話者との終盤の対話を予測するための埋め込みモデルを構築した．その結果，対照学習を導入することで，終盤の対話に対する検索の精度の向上を確認した．特に，一つの学習データに含める発話数を増やした場合に大幅な精度の向上が見られ，長い文脈情報が話し方の特徴を捉える上で有効であることが示された．また，対話データの特徴に応じて埋め込みモデルの性能が変化することが明らかとなった．

今後は，構築した埋め込みモデルをさらに発展させ，実際の対話システムに応用することを目指す．具体的には，Retrieval-Augmented Generation（RAG）のアプローチ [18] を参考に，検索された対話をショットとして応答生成に組み込むことで，対話相手の話し方に適応した発話の実現を目指す．このアプローチにより，従来の対話システムが抱えていた，未知のユーザへの適応性の課題を克服することが期待される．

謝辞

本研究は，JST ムーンショット型研究開発事業，JPMJMS2011 の支援を受けた．また，本研究は，株式会社アイシンとの共同研究の成果を含む．モデルの構築には，名古屋大学のスーパーコンピュータ「不老」を利用した．

参考文献

[1] OpenAI. GPT-4 technical report. **arXiv preprint arXiv:2303.08774**, 2023.

[2] Shinya Iizuka, Shota Mochizuki, Atsumoto Ohashi, Sanae Yamashita, Ao Guo, and Ryuichiro Higashinaka. Clarifying the dialogue-level performance of GPT-3.5 and GPT-4 in task-oriented and non-task-oriented dialogue systems. In **Proceedings of the AAAI Symposium Series**, Vol. 2, pp. 182–186, 2023.

[3] Chuyi Kong, Yaxin Fan, Xiang Wan, Feng Jiang, and Benyou Wang. PlatoLM: Teaching LLMs in multi-round dialogue via a user simulator. In **Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 7841–7863, 2024.

[4] Kurt Shuster, Jing Xu, Mojtaba Komeili, Da Ju, Eric Michael Smith, Stephen Roller, Megan Ung, Moya Chen, Kushal Arora, Joshua Lane, et al. Blenderbot 3: A deployed conversational agent that continually learns to responsibly engage. **arXiv preprint arXiv:2208.03188**, 2022.

[5] Kazunori Komatani, Shinichi Ueno, Tatsuya Kawahara, and Hiroshi G. Okuno. Flexible guidance generation using user model in spoken dialogue systems. In **Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics**, pp. 256–263, 2003.

[6] Atsumoto Ohashi and Ryuichiro Higashinaka. Adaptive natural language generation for task-oriented dialogue via reinforcement learning. In **Proceedings of the 29th International Conference on Computational Linguistics**, pp. 242–252, 2022.

[7] Kenta Yamamoto, Koji Inoue, and Tatsuya Kawahara. Character adaptation of spoken dialogue systems based on user personalities. In **Proceedings of the International Workshop on Spoken Dialogue System Technology**, 2023.

[8] Howard Giles, Tania Ogay, et al. Communication accommodation theory. **Explaining communication: Contemporary theories and exemplars**, pp. 293–310, 2007.

[9] Kate G Niederhoffer and James W Pennebaker. Linguistic style matching in social interaction. **Language and Social Psychology**, Vol. 21, No. 4, pp. 337–360, 2002.

[10] Ashish Jaiswal, Ashwin Ramesh Babu, Mohammad Zaki Zadeh, Debapriya Banerjee, and Fillia Makedon. A survey on contrastive self-supervised learning. **Technologies**, Vol. 9, No. 1, p. 2, 2020.

[11] Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. The Faiss library. **arXiv preprint arXiv:2401.08281**, 2024.

[12] Reina Akama, Kento Watanabe, Sho Yokoi, Sosuke Kobayashi, and Kentaro Inui. Unsupervised learning of style-sensitive word vectors. In **Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)**, pp. 572–578, 2018.

[13] Yuki Zenimoto, Shinzan Komata, and Takehito Utsuro. Style-sensitive sentence embeddings for evaluating similarity in speech style of Japanese sentences by contrastive learning. In **Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics: Student Research Workshop**, pp. 32–39, 2023.

[14] Vassileios Balntas, Edgar Riba, Daniel Ponsa, and Krystian Mikolajczyk. Learning local feature descriptors with triplets and shallow convolutional neural networks. In **Proceedings of the British Machine Vision Conference**, Vol. 1, p. 3, 2016.

[15] Michimasa Inaba, Yuya Chiba, Zhiyang Qi, Ryuichiro Higashinaka, Kazunori Komatani, Yusuke Miyao, and Takayuki Nagai. Travel agency task dialogue corpus: A multimodal dataset with age-diverse speakers. **ACM Transactions on Asian and Low-Resource Language Information Processing**, Vol. 23, No. 9, pp. 1–23, 2024.

[16] Sanae Yamashita, Koji Inoue, Ao Guo, Shota Mochizuki, Tatsuya Kawahara, and Ryuichiro Higashinaka. RealPersonaChat: A realistic persona chat corpus with interlocutors' own personalities. In **Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation**, pp. 852–861, 2023.

[17] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In **Proceedings of the North American Chapter of the Association for Computational Linguistics**, pp. 4171–4186, 2019.

[18] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-Augmented Generation for knowledge-intensive NLP tasks. **arXiv preprint arXiv:2005.11401**, 2021.