話者情報に条件づけられた対話モデルにおける話者情報を抽出する方法の比較分析

A Comparative Analysis of Methods on Extracting Speaker Information in Dialogue Models Conditioned on Speaker Information

安川浩貴 *1 Hiroki Yasukawa　品川政太朗 *1 Seitaro Shinagawa　水上雅博 *2 Masahiro Mizukami　杉山弘晃 *2 Hiroaki Sugiyama　須藤克仁 *1 Katsuhito Sudoh
中村哲 *1 Satoshi Nakamura

*1奈良先端科学技術大学院大学 Nara Institute of Science and Technology　*2NTT コミュニケーション科学基礎研究所 NTT Communication Science Laboratories

In recent dialogue research, many models have been proposed in which sentences describing the speaker's information are input into a encoder with dialogue sentences to generate a response sentence that reflects the speaker's information. However, such models have a problem that they generate responses that are strongly conditioned on the words appearing in the persona description. In this study, we aimed to alleviate this conditioning by inputting distributed representations obtained by encoding Persona Descriptions with VAE. As a result, we were able to generate a response that is not contradict the persona description while relaxing the conditioning.

1. はじめに

近年の対話モデルの研究では，ニューラルネットワークを用いた返答文生成手法が数多く提案されている．特に SNS から収集した大量のデータを用いて事前学習した大規模な Transformer [Vaswani 17] ベースのモデルを，高品質な対話データセットを用いて finetune することで，人間と同程度に自然な対話や魅力的な対話を実現している [Roller 20][杉山 20]．このような手法では，対話の入力文を Encoder へ入力し，Decoder で入力文に対する返答文を生成する．また話者の情報を反映した返答生成を行うようにモデルを学習させる際には，Encoder に「話者の情報を記述した文（以下，ペルソナ記述文）」及び「対話の入力文」を連結して入力する．これによりペルソナ記述文に則った返答文を生成するように学習させる．実際に杉山らが公開している日本語 Transformer Encoder-decoder 対話モデル [Sugiyama 21] を，JPersonaChat [Sugiyama 21] を使用して finetune した場合には話者情報に準拠する応答を生成するようにモデルを学習できた．

その一方で，ペルソナ記述文に登場する単語に強く条件づけられている返答文が散見された．表 1 に示すように，ペルソナ記述文の一部である「香水の匂いが苦手」といった部分が，前後の文脈とは関係なしにそのまま反映された返答文が生成されてしまう現象も見られた．このような返答文はペルソナ記述文の内容を反映している一方で，日本語として不自然なものになってしまっている．また入力文に対する一貫性が低く，対話モデルとして適切な生成文とは言い難い．また一般に，与えられるペルソナ記述文の数は限られているため，含まれる単語に強く条件づけられた返答文を生成するモデルでは，返答文の多様性が低下することが危惧される．

このようにペルソナ記述文に強く条件づけられた返答文の生成してしまう問題は，文を直接入力してコピーするようにモデルが学習してしまっているからであると考えている．よって本研究では対話生成モデルがペルソナ記述文の単語を直接参照できない形で話者の情報を入力することで，この問題を緩和することを目指す．具体的には，対話文の Encoder とは別にペルソナ記述文をエンコードするモデルとして Variational Autoencoder（VAE）[Kingma 14] を導入する．VAE を使用してペルソナ記述文を話者情報に圧縮してから利用する事で，与えられたペルソナ記述文からの条件づけを緩和し，多様な応答を生成する事を目指す．

本研究では Distinct-N [Li 16] を用いた生成文の多様性の評価及び，BLEU を用いたペルソナ記述文と生成文の被覆率の評価を行った．またベースラインモデルにおいて主観評価実験を行い，ペルソナ記述文と生成文の矛盾度合いについて調べた．

連絡先: 安川浩貴，奈良先端科学技術大学院大学 知能コミュニケーション研究室，yasukawa.hiroki.ye6@is.naist.jp

2. 関連研究

2.1 大量の発話を用いた話者埋め込み表現の獲得とその反映

ニューラル対話モデルにおいて初めて返答文に現れるペルソナの制御を目指した Li らの研究 [Li 16] では，ある話者の大量の発話から話者の情報を分散表現として獲得し，それをデコーダでの返答文生成の際に利用する Speaker Model を提案している．このモデルでは話者 ID を付与した対話データセットを使用して，単語の分散表現と同様の形で話者の分散表現を獲得する．そして返答文生成デコーダへ話者の分散表現と入力文と連結して入力する事で，明示的に表現できない話者らしさを反映した返答文を生成している．しかしこのモデルにおいて話者に関する情報として与えられる話者 ID には，話者の性質についての情報が含まれておらず，話者の分散表現の学習は対話文からしか行えない．本研究では VAE を用いてペルソナ記述文をエンコードしたものを話者の分散表現として用いることで，ペルソナ記述文及び対話文から話者の分散表現を学習し，デコーダでの返答文生成の際に利用する．

2.2 明示的なプロフィールに基づく話者情報の反映

話者の埋め込み表現を用いる研究では，大量のデータセットを用いて話し方のような明示的に表現できない情報を学習する事に重きを置いていた．一方で Zhang らの研究 [Zhang 18] では話者の情報を明示的に記述したペルソナ記述文を含んだ対話データセットである PersonaChat を用いて話者の情報を

表 1: ペルソナ記述文に登場する単語に強く条件づけられた返答文の例（ベースラインモデル）と条件づけを緩和した返答文の例（提案モデル）.

ペルソナ記述文	私は鹿児島から引っ越してきました。私は専業主婦です。私は手が小さいです。 私は焼肉が大好物です。**私は香水の匂いが苦手**です。
対話入力文	もう、すっごいかわいいんですよ。人懐っこいのでモフモフ出来ます！
ベースラインモデル	モフモフいいですね 。**私は香水の匂いが苦手**なので、モフモフできないのが残念です。
提案手法	モフモフしたいです!私の実家の猫も人懐っこいですよ。

返答文へ反映させるモデルを提案している．このデータセットは，5 文程度のペルソナ記述文で構成されたセットを作成し，それに則ってクラウドワーカーに対話させる事で収集したものである．このように話者の情報を記述した文章とそれに関係する内容を含んだ対話データを用いて，ペルソナ記述文と入力文を与えられた際に話者の情報に基づいた返答文を生成可能なモデルを提案している．また近年の事前学習済み対話モデルを Finetune する研究 [Roller 20][Song 21] では，対話の入力文と連結して同じ Encoder へ入力することでペルソナ記述文に基づく返答を生成するように学習するモデルが提案されている．本研究では，VAE によってペルソナ記述文をエンコードすることで獲得した話者の分散表現を用いて話者情報を反映した返答文を生成するモデルを提案する．このようにペルソナ記述文の内容を直接参照できない構造にすることによって，ペルソナ記述文に登場する単語からの条件づけを緩和することを目指す．

3. モデル

3.1 ベースラインモデル

本研究では，ベースラインモデルとしてシンプルな Transformer Encoder-Decoder モデル [Vaswani 17] を使用した．モデルの構造を図 1 に示す．

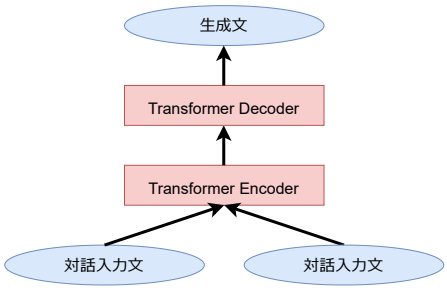

図 1: ベースラインモデルの図．ペルソナ記述文と対話入力文を同じ Encoder へ入力し，Decoder で返答文を生成する

JPersonaChat から対話の入力文，返答文，返答者の性質を記述したペルソナ記述文のセットを作成してモデルの訓練を行った．ペルソナ記述文と対話の入力文を [SEP] トークンで連結して Transformer Encoder の入力とし，返答文を正解データとした．

3.2 提案モデル

話者の分散表現 z を獲得する VAE と，Transformer Encoder Decoder を用いて返答文を生成する対話モデルの二つを用いて提案モデルを構築した．モデルの構造を図 2 に示す．

VAE（図 2 左側）は図 3 のようにして構成した．

まず Transformer Encoder を使用してペルソナ記述文 P_{in} をエンコードして出力 E^P_{out} を獲得する．次にランダムに初期化したベクトル c を利用して次の形で Multi Head Attention[Vaswani 17] の出力 h を得る．ここで Query を c, Key 及び Value を E^P_{out} とした．

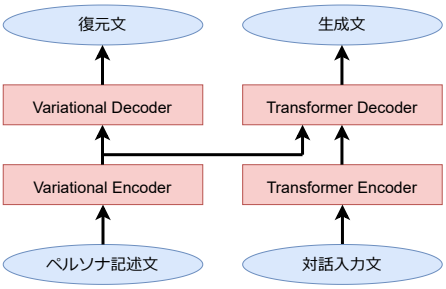

図 2: 提案モデルの全体図．復元文は，話者表現 z から VAE によって復元されたペルソナ記述文を表す．

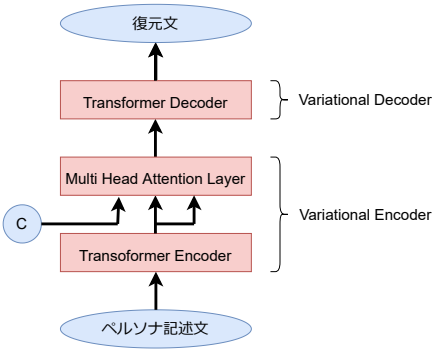

図 3: VAE モデルの図．c はランダムに初期化されたベクトルを示す．

$$h = \mathrm{MultiHeadAttetion}(\mathrm{c}, \mathrm{E}^{\mathrm{P}}_{\mathrm{out}}, \mathrm{E}^{\mathrm{P}}_{\mathrm{out}})$$

この h に基づき，話者を表現する潜在変数 z が従う母平均 μ と分散 σ を以下の式で求める．

$$\begin{pmatrix} \mu \\ \log(\sigma^2) \end{pmatrix} = hW_q + b_q$$

ここで reparameterization trick [Kingma 14] を用いて h から話者を表現する潜在変数 z を求める．これは VAE において End to End な学習を行うために提案された手法である．z を $\mathcal{N}(\mu, \sigma^2 I)$ からランダムサンプリングする代わりに，ランダムノイズ $\epsilon \sim \mathcal{N}(0, I)$ を用いて，潜在変数 z を $z = \mu + \sigma \otimes \epsilon$ で求め，VAE の Transformer Deocder へ入力することで復元文 P_{out} を生成する．

対話モデル（図 2 右側）では，対話の入力文 D_{in} を事前学習済み対話モデルの Transformer Encoder でエンコードして獲得した出力 E^D_{out} を Transformer Decoder の各層へ入力する．また VAE を用いて獲得した話者の分散表現 z を，各トークン

及び位置情報の埋め込み表現と加算して Transofrmer Decoder へ入力し，話者の情報を反映した返答文 D_{out} を生成する．

4. 実験設定

本研究では提案モデルと従来モデルが生成する文章について比較分析を行う．両モデルが生成する文章の多様性及びペルソナ記述文からの条件づけの度合いを検証するために，生成文の多様性とペルソナ記述文と生成文の被覆率について自動評価を行った．また提案モデルにおいてペルソナ記述文と矛盾した返答文を生成していないかの主観評価実験も行った．本章ではモデルの学習に関わる事項や評価手法について記述する．

4.1 使用データセット

本研究では，JPersonaChat [Sugiyama 21] を使用した．このデータセットは Zhang ら [Zhang 18] と同様の形で構成されており，ペルソナ記述文とそれに則って行われた対話がセットになっている．その例を表 2 に示す．

データセットにはペルソナ記述文のセットが 100 セット含まれており，それぞれを与えられたクラウドワーカーが対話を行っている．またワーカーは与えられたペルソナ記述文の内容について触れるようにしながら交互に雑談を行う．1 発話は最大で 100 文字，全体で最低 12 発話，最高 15 発話となるように収集されている．本研究ではこの内の 8 割の対話を用いて訓練を行い，残りの 1 割を検証データ，残りの 1 割をテストデータとした．

4.2 損失関数

提案手法の訓練には，VAE についての損失関数と，対話モデルについての損失関数を用いる．

まず VAE における損失関数を説明する．VAE における潜在変数の生成に関する損失関数 $\mathcal{L}_{KL}$ と入力文の復元に関する損失関数 $\mathcal{L}_R$ をそれぞれ次の式で求める．

$$\mathcal{L}_{KL} = -\frac{1}{2}(1 + \log(\sigma^2) - (\mu^2) - (\sigma^2))$$

$$\mathcal{L}_R = -\log(\mathrm{p}_\theta(\mathrm{P}_{\mathrm{out}}|\mathrm{P}_{\mathrm{in}}))$$

VAE の学習では $\mathcal{L}_{kl}$ の項が消失しやすいことが問題となるため，cyclical annealing[Fu 19] を用いる．これは $\mathcal{L}_{kl}$ の重みを一定の間隔で上下させることで，$\mathcal{L}_{kl}$ の項の減衰を抑えることができる．この重み W_A は，サイクルの間隔 T とその間隔内での第 N ステップによって，次のように定義する．

$$\mathcal{W}_A = \min(\frac{2N \bmod T}{T}, 1.0)$$

また返答文生成に関する損失関数 $\mathcal{L}_D$ を次の式で求める．

$$\mathcal{L}_D = -\log(\mathrm{p}_\psi(\mathrm{D}_{\mathrm{out}}|\mathrm{D}_{\mathrm{in}}))$$

以上より，モデル全体の損失関数 $\mathcal{L}_{Model}$ は，各損失関数の重みを λ_{KL}，λ_R，λ_D として，次の式で求める．

$$\mathcal{L}_{model} = W_A\lambda_{KL}\mathcal{L}_{KL} + \lambda_R\mathcal{L}_R + \lambda_D\mathcal{L}_D$$

4.3 モデル学習の設定

杉山らが公開している日本語 Transformer Encoder-decoder 対話モデル [Sugiyama 21] を使用する．このモデルは 2 層の Encoder と 24 層の Decoder を持つ．また隠れ層の次元数が 1920 次元，Attention head が 32 である．

提案手法では，対話モデルに組み合わせる VAE のモデルパラメータとして，潜在変数の次元数を 1920 次元とした．また学習の Optimizer には Adafactor[Shazeer 18] を使用し，学習率は 5e-6 とした．損失関数の重みは $\lambda_{KL} = 0.001$，$\lambda_R = 1.0$，$\lambda_D = 1.0$ とし，cyclical annealing のサイクル間隔を 10000 ステップとした．

モデルの実装，学習には Pytorch をベースとしたニューラル系列モデリングライブラリである Fairseq[*1] を利用した．学習データには同様に公開されている JPersonaChat を用いた．

4.4 評価手法

本研究では三つの観点から評価を行う．

提案モデルとベースラインモデルが生成する文章の多様性を評価するために Distinct-N [Li 16] を用いた．また両モデルにおけるペルソナ記述文からの条件づけの度合いを比較するために BLEU を用いた．更に提案モデルの生成する文章がペルソナ記述文と矛盾していないかを評価するために，ペルソナ記述文と生成文の間の矛盾度を測る主観評価実験を行った．

5. 実験結果と考察

4.4 節で述べた手法を用いて，提案手法と従来手法の評価を行った．本章ではその結果を示す．

5.1 生成文の多様性の評価

生成文の多様性について比較するために，Distinct-N（$N = 1, 2$）を用いた．本研究では，辞書に mecab-ipadic-NEologd を用いた MeCab[*2][Sato 17] を使用して生成文を分かち書きした後に Distinct-N の計算を行った．結果を表 3 に示す．

Distinct-1，Distinct-2 の双方で提案手法がベースラインモデルよりも高い値を記録し，生成する返答文の多様性が提案手法を上回る事が示された．

5.2 ペルソナ記述文からの条件づけ度合いの評価

生成文がペルソナ記述文にどの程度条件づけられているかを評価するために，ペルソナ記述文と生成文の BLEU を計算した．また実際のデータセットにおける被覆率とも比較するため，テストデータにおけるペルソナ記述文と返答文の被覆率も計算した．本研究では，BLEU3 及び BLEU4 を用いた．結果を表 4 に示す．

ベースラインモデルでは，BLEU3 及び BLEU4 でテストデータを大きく上回る値を記録した．これは，ベースラインモデルが生成する返答文にはペルソナ記述文に含まれている N-gram が過剰に登場している事を意味している．その一方で提案手法はテストデータを下回る値を記録した．ペルソナ記述文からの条件づけを弱める事はできた一方で，ペルソナ記述文の内容を反映する能力も低下していると考えられる．

5.3 ペルソナ記述文と生成文の矛盾度合いの評価

提案手法の生成文において，VAE によって強い条件づけを緩和していることは確認できた．しかし条件づけの緩和によりペルソナ記述文と矛盾した文章を生成していると大きな問題がある．そこで提案モデルにおけるペルソナ記述文と生成文の矛盾度合いについて検証するために主観評価を行った．実験参加者に「ペルソナ記述文」「対話の入力文」「生成文」のセットを渡し，生成文がペルソナ記述文として与えられた五つの文章全てと矛盾していないかについて評価してもらった．実験参加者は五名であり，上述のセット 100 セットを渡した．

∗1 https://github.com/pytorch/fairseq
∗2 http://taku910.github.io/mecab/

表 2: Persona Chat の例

話者 1 に与えられたペルソナ記述文	話者 2 に与えられたペルソナ記述文
私は北海道で生まれました。	私は青森で生まれました。
私は東京に住んでみたいです。	私は山のそばに住んでいます。
私は一軒家に住んでいます。	私は昭和生まれです。
私は介護福祉士です。	私は広告代理店勤務です。
私は彼氏いない歴が長いです。	私の趣味は写真撮影です。
対話	
話者 1 こんにちは。お元気ですか？	
話者 2 はい、元気です。広告代理店での仕事が忙しいですが。	
話者 1 お疲れさまです。私は介護福祉士をしています。	
話者 2 介護福祉士として働いていらっしゃるんですね。大変なお仕事ですよね。	

表 3: Distinct-N を用いた評価

	Distinct-1	Distinct-2
ベースラインモデル	0.029	0.1197
提案モデル	0.0336	0.1415

表 4: テストデータの応答文及び生成モデルの生成した文と，ペルソナ文の重複度の比較

応答文	BLEU3	BLEU4
テストデータの応答文（参考値）	0.035	0.017
ベースラインモデルの生成文	0.135	0.094
提案モデルの生成文	0.017	0.005

結果，矛盾していると判断されたセットは五名の平均で 4.8% に留まった．また全員から矛盾があると評価されたセットは存在しなかった．このことから，提案モデルはペルソナ記述文と矛盾するような返答文を生成することは少ないと言える．

5.4 まとめと考察

Distinct-N，BLEU，主観評価の結果をまとめ，考察を行う．ベースラインモデルでは，ペルソナ記述文の単語に強く条件づけられた文章を生成しており，多様性が低下していた．一方で提案モデルでは多様な文章を生成できるようになったが，ペルソナ記述文の内容の反映度合いが低くなっているように見受けられた．これは，ペルソナ記述文を直接モデルに与えずに抽象化した分散表現を用いてモデルに与えていることで，ペルソナ記述文に含まれる固有名詞などの具体的なコンテンツの反映が難しくなっている事が原因であると考えられる．

一方で主観評価の結果，ペルソナ記述文の内容と矛盾した返答文の生成はほとんど行っていないことがわかった．よってペルソナ記述文から無関係に返答文を生成している訳ではないと言える．

6. おわりに

本研究では，文章で与えられた話者の情報に条件づけられた対話モデルにおいて話者情報を抽出する方法について比較分析を行った．具体的には，ペルソナ記述文を対話の入力文と同じエンコーダへと入力する手法と，対話のエンコーダとは異なる VAE を用いてペルソナ記述文をエンコードする手法を使用した．BLEU を用いた評価により，従来手法ではペルソナ記述文に強く条件づけられた返答文を生成してしまっている事が示された．また提案手法では，ベースラインよりも多様であり，ペルソナ記述文からの条件づけが緩和された返答文を生成できる事もわかった．

その一方で生成結果を見ると，具体的なコンテンツの反映が難化しているといった難点があるように感じた．今後はペルソナ記述文に登場する単語からの条件づけを緩和しながら，言及する必要がある時には言及できるようなモデルの構築に取り組みたいと考えている．

参考文献

[Fu 19] Hao Fu et al, Cyclical Annealing Schedule: A Simple Approach to Mitigating KL Vanishing. NAACL. (2019)

[Kingma 14] Diederik P Kingma et al, Auto-Encoding Variational Bayes. arXiv:1312.6114. (2014)

[Li 16] Li Jiwei et al, A Diversity-Promoting Objective Function for Neural Conversation Models ACL. pp.110–119 (2016)

[Li 16] Li Jiwei et al, A persona-based neural conversation model. arXiv preprint arXiv:1603.06155. (2016)

[Roller 20] Roller Stephen et al, Recipes for building an open-domain chatbot. arXiv preprint arXiv:2004.13637. (2020)

[Sato 17] Toshinori Sato et al, Implementation of a word segmentation dictionary called mecab-ipadic-NEologd and study on how to use it effectively for information retrieval (in Japanese). ANLP (2017)

[Shazeer 18] Shazeer Noam et al, Adafactor: Adaptive learning rates with sublinear memory cost. ICML. pp.4596–4604. (2018)

[Song 21] Song Haoyu et al, BoB: BERT Over BERT for Training Persona-based Dialogue Models from Limited Personalized Data. arXiv preprint arXiv:2106.06169. (2921)

[杉山 20] 杉山弘晃 et al, Transformer encoder-decoder モデルによる趣味雑談システムの構築. 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 90 回, p.24. 一般社団法人 人工知能学会. (2020)

[Sugiyama 21] Hiroaki Sugiyama et al, Empirical Analysis of Training Strategies of Transformer-based Japanese Chitchat Systems. arXiv:2109.05217. (2021)

[Vaswani 17] Vaswani Ashish et al, Attention is all you need. NeurIPS. pp.5998–6008. (2017)

[Zhang 18] Zhang Saizheng et al, Personalizing dialogue agents: I have a dog, do you have pets too?. arXiv preprint arXiv:1801.07243. (2018)