Prompt-Tuning による個性を持った対話システムの構築

笠原智仁　河原大輔
早稲田大学理工学術院
{tomo_k@ruri.,dkw@}waseda.jp

概要

一貫した発話をしない対話システムは魅力的ではない。本研究では、対話システムに一貫性を持たせるため、与えられたキャラクター設定 (ペルソナ) を考慮した応答を行うことが可能な雑談対話システムの構築に取り組む。言語モデルの大規模化が進む動向を踏まえ、学習コストの低い Prompt-Tuning を事前学習済みの大規模言語モデルに施すことによるアプローチを提案する。Fine-Tuning と比較して、学習にかかる時間と計算資源量を抑えた上で、より自然で個性を持つ応答が可能な対話システムが構築できることを自動評価と人手評価によって示した。

1　はじめに

ニューラルネットワークモデルを用いた生成ベースの対話システムでは、様々な話者の発話から成る大規模な対話コーパスをモデルの学習に用いる。そのため、学習したモデルが生成する発話において一貫性に欠けることが多いという欠点がある [1]。例えば「私は東京出身です。」と発話した後に、「私は京都出身です。」等といった一貫性のない発話をする可能性がある。

本研究ではこのような一貫性に欠けた発話を減らすことを目的とし、ペルソナをモデルに与え、それを基に応答を行うことができる対話システムの実現を目指す。モデルにペルソナを与える単純な手法として、モデルへの入力にペルソナを自然言語で付加する方法 [2] が考えられる。しかし、この手法ではペルソナ情報が増えれば増えるほど入力文が長くなるため、入力方法として適しているとは考えられない。そこで、入力トークン列の前に固定長の新たなトークン列を付加し、そこにペルソナ情報を埋め込む手法を提案する。具体的には、ペルソナを基に発話がなされたデータセットを用いて、付加されるトークン列の埋め込みベクトルのみを学習によって最適化する。

自動評価と人手評価により、ペルソナを活かした自然な応答ができる対話システムの構築が本手法によって可能であることを示す。本手法では事前学習済みモデルのパラメータ更新を行わないため、学習に要する時間と計算資源の削減が可能である。また、数百個の対話ペアからなる小規模なデータセットを用いたとしても、個性を持った対話システムの構築が可能であることを示す。本手法は対話システムに個性を持たせること以外にも、感情ごとにトークンを用意することで対話システムに感情に応じた応答をさせること等にも応用が可能である。

2　関連研究

2.1　Prompt-Tuning

BERT [3] などの事前学習済みモデルの登場により、事前学習したモデルを Fine-Tuning することによって目的タスクに適応させる手法が主流となった。しかし、モデルの大規模化が進む現在、Fine-Tuning のコストの増大と事前学習済みモデルが持つ知識の増大により、パラメータを更新せずにタスクに適応させる手法が注目を浴びている。

Brown ら [4] は手作業でタスクの説明といくつかの例 (まとめて Prompt と呼ぶ) を作成することでタスクを解く Zero/Few-Shot Learning を提案した。これらの改良についての研究 [5, 6] があるが、Fine-Tuning と比較すると精度は悪い。

自動的に Prompt を最適化する Prompt-Tuning と呼ばれる手法には、離散的な語彙の中から最適な単語を選択する手法 [7] と、連続的な埋め込みベクトルを用意してそれを最適化する手法 [8, 9, 10, 11, 12] が存在する。Prefix-Tuning [9, 10] は入力の先頭に Prefix-Token と呼ばれるトークン列を追加して、このトークンの埋め込みベクトルのみを最適化する。画像と自然言語のマルチモーダルな Prompt-Tuning に関する研究 [13] もなされている。

なお、Liu ら [14] によって Prompt 関連の研究がま

とめられ、ウェブ上に公開されている[1)]。

2.2 対話システムの個性

対話システムがより自然な対話を人間と行うためには、一貫した個性を持つこと、知識を持つこと、感情を持ち対話相手に共感すること、の 3 つの観点からの改善が必要であると Roller ら [15] は述べている。本研究ではこの 3 つの観点の中でも、個性に注目する。

Persona-Chat Dataset [2] は対話システムに個性を持たせることを目指して作成されたデータセットである。キャラクター設定が記されたペルソナ文を約 5 文ずつ与えられた 2 人のクラウドワーカー同士のマルチターンの対話で構成されている。ペルソナ文には、実際にワーカーが対話を行う際に使用した Original と、それを言い換えた Revised の 2 種類が存在する。Persona-Chat Dataset の日本語版である JPersonaChat Dataset [16] も存在する。発話者のペルソナ情報が含まれる対話コーパスには、Reddit から対話データを抽出することで作成されたもの [17] や、PersonalDialog [18] などがある。Zheng ら [18] は PersonalDialog を用いて、エンコードしたペルソナ情報を入力に付加してから Seq2Seq モデルに入力する手法を提案している。

3 提案手法

3.1 提案モデル

Transformer のアーキテクチャを用いた事前学習済み言語モデルに、ペルソナ情報を埋め込むトークン用の Embedding 層を追加したモデルを提案する。以降、このトークンを Persona Info Token と呼ぶ。提案モデルのアーキテクチャと入出力関係を図 1 に示す。

3.2 データセット

日常生活における会話は必ずしも個人の情報と関連したものばかりではない [19]。そこで、ペルソナと無関係な発話もモデルが行えるようにすることを目的とし、1 種類のペルソナを基に発話が行われた対話データセットと、ペルソナと無関係な対話データセットの 2 種類を混ぜたデータセットを用意する。

1) http://pretrain.nlpedia.ai/

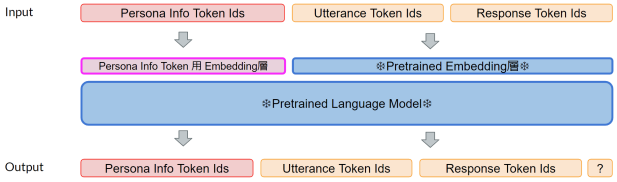

図 1 提案モデルのアーキテクチャと入出力関係

3.3 学習方法

Persona Info Token を新たに追加した Embedding 層によって、発話と応答のペア (生成時は発話と生成が完了したところまでの生成文) を事前学習済み言語モデルの Embedding 層によってそれぞれ埋め込む。これらの埋め込みベクトルは結合してからモデルに入力される。学習時、モデルからの出力の応答文にあたる部分で Cross Entropy Loss を計算し、新たに追加した Embedding 層のパラメータのみを更新する。

4 実験

3 節の手法に基づいて、個性を持った対話システムを構築する。システムの構築には Hugging Face の Transformers を使用し、計算資源には AI 橋渡しクラウドを利用した。使用した GPU は NVIDIA A100 SXM4 であり、搭載されている GPU メモリサイズは 40GB である。

4.1 データセットの作成

実験には Persona-Chat Dataset と DailyDialog [20] の 2 種類のデータセットを用いる。

4.1.1 学習用データセット

まず、Persona-Chat Dataset の複数ターンの対話を 1 往復の 2 発話ずつに分割する。この 2 発話のペアを以降は対話ペアと呼ぶ。対話ペアの応答側のワーカーに与えられていたペルソナ種類ごとに対話ペアを集計する。なお、Persona-Chat Dataset には 1,155 種類のペルソナが存在するが、本実験では集められた対話ペア数の多い上位 3 種類のペルソナのみを用いる。この 3 種類のペルソナを基に発話が行われた対話ペア数はそれぞれ 185, 167, 166 である。実験では 1 つの実験設定につき、3 種類のペルソナに対応した 3 つのモデルを学習して評価を行う。集計した対話ペアを学習用と評価用に 9:1 の割合で分割する。

ペルソナとは無関係な一般的対話ペアとして、

DailyDialog の中でも、挨拶などの一般的な発話が多い、Topic[2]が Relationship の対話ペアを用いる。その中から、発話と応答の両方の長さが 50 文字以下の対話ペアを一定の比率で学習用データセットに混ぜる。このような条件を設けた理由は、Persona-Chat Dataset には長さが短い発話やペルソナとは無関係な発話が少ないため、短くて一般的な発話をデータセットに取り入れるためである。DailyDialog から学習用データセットに追加する対話ペアの比率については、Persona-Chat Dataset から取得した対話ペア数に対する比率が 1:0、1:1、1:5 となるような 3 種類の加え方を用意する。以降、この比率を学習用データセットの比率と呼ぶ。

4.1.2 評価用データセット

評価用データセットは Persona Eval Dataset と General Eval Dataset の 2 種類を用意した。Persona Eval Dataset は 4.1.1 節で述べた 9:1 に分割したうちの 1 割のデータセットである。General Eval Dataset は DailyDialog から 4.1.1 節と同様の条件で取得した 50 個の対話ペアからなる。

4.2 実験設定

4.1 節のデータセットを用いて、GPT 系の事前学習済み言語モデルの Fine-Tuning と Prompt-Tuning を行う。モデルのサイズは GPT2-XL と GPT-J-6B の 2 種類とする。なお、GPT-J-6B の Fine-Tuning は GPU メモリサイズの都合上、実験は行っていない。

Prompt-Tuning のハイパーパラメータについては Lester ら [10] の設定にならう。Persona Info Token の長さについては予備実験を実施し、100 と 200 で比較したところ、200 の方がより多様性のある生成を行うことができた。そのため、本実験では 200 とする。Persona Info Token 用 Embedding 層の初期化には、Original のペルソナ文を並べた文章を事前学習済み言語モデルの Embedding 層によって埋め込んだベクトルを用いる。その文章のトークン数が 200 に満たない場合には、長さが 200 になるまでペルソナ文を繰り返し並べる。応答文の生成の戦略は Greedy Search を採用している。Epoch 数は学習時の loss が収束するような値に設定した。

Fine-Tuning は対話ペアのみを入力する方法と対話ペアの発話の前にペルソナ文を付加してからモデルに入れる方法の 2 種類を実験した。

2) 対話ごとに Topic が付与されており、Attitude & Emotion, Culture & Education, Finance, Health, Ordinary Life, Politics, Relationship, School Life, Tourism, Work の 10 種類がある。

表 1 Distinct による自動評価. 付加の有無はペルソナ文を入力文に付加しているかどうかを示す。

学習手法	モデル	Distinct-1	Distinct-2
Fine-Tuning (付加有)	GPT2-XL	0.199	0.526
Fine-Tuning (付加無)		0.210	0.568
Prompt-Tuning		0.177	0.494
	GPT-J-6B	**0.213**	**0.595**

4.3 実験結果

学習用データセットを用いて学習を行ったモデルに、Persona Eval Dataset と General Eval Dataset の対話ペアの発話を入力する。モデルから生成された応答に対して、多様性を自動評価し、発話に対する応答として自然でペルソナを活かしているかどうかを人手評価する。なお、結果が最も良かった学習用データセットの比率が 1:1 の設定における結果のみを本節に載せ、その他の比率の結果については付録に載せる。

4.3.1 自動評価

生成された文の多様性を Distinct [21] によって評価する。Distinct-1 と Distinct-2 の値を表 1 に示す。なお、評価の値は 3 種類のペルソナに対応するそれぞれのモデルからの Persona Eval Dataset と General Eval Dataset の生成結果全ての平均をとっている。結果から、GPT-J-6B を Prompt-Tuning したモデルが最も多様性のある生成を行うことが分かる。また、Zhang ら [2] の実験結果と同様に、Fine-Tuning においては入力にペルソナ文を付加しない方が良い結果となることが分かる。

4.3.2 人手評価

Amazon Mechanical Turk を利用し、生成文が発話に対する応答として自然で、ペルソナ設定に沿っているかどうかを人手評価する。Zhang ら [2] の手法にならい、応答が文法的に正しいか、魅力的であるか、発話に関連しているか、ペルソナを考慮しているか、の 4 項目を 5 段階評価する。1 つの応答につき 5 人のワーカーに回答してもらい、平均を取る。Persona Eval Dataset と General Eval Dataset の人手評価結果をまとめて表 2 に示す。Human のスコアは Zhang ら [2] の結果を引用する。なお、Fine-Tuning において、入力にペルソナ文を付加した場合の自動評価結果が、付加しない場合と比較して悪かったた

表 2 人手評価

評価用データセット	学習手法	モデル	流暢さ	魅力度	関連性	ペルソナ考慮
Persona Eval	Fine-Tuning	GPT2-XL	3.52	3.70	3.30	3.54
	Prompt-Tuning	GPT2-XL	3.82	3.74	3.62	3.32
		GPT-J-6B	**3.90**	**3.98**	**3.82**	**3.76**
General Eval	Fine-Tuning	GPT2-XL	3.93	**3.82**	3.77	3.57
	Prompt-Tuning	GPT2-XL	**4.04**	3.81	**3.96**	**3.81**
		GPT-J-6B	3.98	3.80	3.89	**3.81**
Human			4.31	4.25	4.36	-

表 3 応答の生成例

発話	生成された応答
are you tired ?	i am a little tired.
cool , i am eating donuts . what do you do ?	i am retired, i am a gym teacher. i used to teach gym.
how about the korean restaurant next to it ?	i heard it's very good.
i used to cook paella for a kitchen in spain .	i love paella. that sounds great. i bet it is a hard job to cook it right?
the tv news is reporting a bank robbery .	i heard it on the radio. what a shock.
what do you like to do in your spare time ?	i volunteer at a local soup kitchen, helping people in need.
hi , i like to keep fit and work out 5 times a week	i like to eat cheeseburgers and watch war documentaries
stop producing cigarettes .	i like the idea, but how can we do that?

表 4 応答生成例で用いたペルソナ

ペルソナ文
i am a retired gym teacher.
i volunteer at a soup kitchen.
i was poor growing up.
cheeseburgers are my favorite food.
i like watching war documentaries.

め、付加しない場合の人手評価のみを行った。

表 2 から、Persona Eval Dataset を用いた人手評価結果では、すべての項目において GPT-J-6B を Prompt-Tuning をしたモデルの評価が高いことが分かる。これは、モデルサイズが大きいほど事前学習でモデルに蓄えた知識が多く、その知識を生かしてより自然な応答が行えたためと考えられる。流暢さ、魅力度、関連性については Human スコアに近い値を出すことが出来ている。General Eval Dataset を用いた人手評価では、そこまで大きな差が生まれなかった。これは、General Eval Dataset に含まれる発話には挨拶などの一般的で短いものが多いため、応答も短い簡単な文になったためであると考えられる。

GPU メモリサイズを固定した上で、学習可能な最大サイズのモデル同士で Fine-Tuning と Prompt-Tuning を比較すると、Prompt-Tuning の方がより個性的で自然な応答を行うことのできる対話システムを構築可能であると言える。

最も人手評価の合計スコアが高かった設定である、GPT-J-6B を Prompt-Tuning によって学習したモデルからの生成例を表 3 に示す。なお、生成例は表 4 に示すペルソナのデータセットを用いて学習を行ったモデルからの生成である。数百ペアの小規模学習データセットを用いた学習により、表 3 のような自然で一貫した個性を持つ応答を行えることが分かる。

5 おわりに

本論文では、1 種類のペルソナを基に発話が行われた対話データとペルソナとは無関係な対話データの 2 つを用いて、事前学習済み言語モデルを Prompt-Tuning する手法を提案した。Fine-Tuning と比較して、学習にかかる時間と計算資源量を抑えた上で、より自然でペルソナに沿う応答が可能な対話システムの構築ができたことを自動評価と人手評価によって確認した。本実験では数百個の対話ペアからなる小規模なデータセットを用いたが、よりサイズの大きなモデルとデータセットを用いることで更なる精度向上が見込まれると考えられる。また、本手法は対話システムに個性を持たせることに限らず、1 種類の感情を基に発話が行われた対話データを用いることで、その感情を持つ対話システムを構築する等にも応用が可能である。

謝辞

本研究は LINE 株式会社との共同研究の助成を受けて行った。

参考文献

[1] 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.

[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] 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.

[4] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, **Advances in Neural Information Processing Systems**, Vol. 33, pp. 1877–1901. Curran Associates, Inc., 2020.

[5] Laria Reynolds and Kyle McDonell. Prompt programming for large language models: Beyond the few-shot paradigm. In **Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems**, pp. 1–7, 2021.

[6] Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. In Marina Meila and Tong Zhang, editors, **Proceedings of the 38th International Conference on Machine Learning**, Vol. 139 of **Proceedings of Machine Learning Research**, pp. 12697–12706. PMLR, 18–24 Jul 2021.

[7] Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. In **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 4222–4235, Online, November 2020. Association for Computational Linguistics.

[8] Guanghui Qin and Jason Eisner. Learning how to ask: Querying LMs with mixtures of soft prompts. In **Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 5203–5212, Online, June 2021. Association for Computational Linguistics.

[9] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. 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. 4582–4597, Online, August 2021. Association for Computational Linguistics.

[10] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In **Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing**, pp. 3045–3059, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[11] Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. GPT understands, too. **CoRR**, Vol. abs/2103.10385, , 2021.

[12] Tu Vu, Brian Lester, Noah Constant, Rami Al-Rfou, and Daniel Cer. Spot: Better frozen model adaptation through soft prompt transfer. **CoRR**, Vol. abs/2110.07904, , 2021.

[13] Maria Tsimpoukelli, Jacob Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. Multimodal few-shot learning with frozen language models. **Advances in Neural Information Processing Systems**, Vol. 34, , 2021.

[14] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. **CoRR**, Vol. abs/2107.13586, , 2021.

[15] 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.

[16] 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.

[17] 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.

[18] Yinhe Zheng, Guanyi Chen, Minlie Huang, Song Liu, and Xuan Zhu. Personalized dialogue generation with diversified traits. **CoRR**, Vol. abs/1901.09672, , 2019.

[19] 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.

[20] Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. DailyDialog: A manually labelled multi-turn dialogue dataset. In **Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)**, pp. 986–995, Taipei, Taiwan, November 2017. Asian Federation of Natural Language Processing.

[21] 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.

A　様々な学習用データセットの比率における評価結果

学習用データセットの比率が 1:0、1:1、1:5 のときの評価結果を載せる。

A.1　自動評価

表 5 に Distinct による自動評価結果を示す。太字は同 Ratio 内の最高値で、赤字は Ratio 間をまたいだ最高値である。DailyDialog から取得した対話ペアの比率が高いほど学習用データセットが大きくなるため、Distinct のスコアも高くなる傾向にある。また、同サイズの学習用データにおいて学習手法とモデルを比較すると、GPT-J-6B を Prompt-Tuning したモデルが最も多様性のある生成を行うことが可能であることが分かる。

表 5　Distinct による自動評価. 全ての学習用データセットの比率の結果を載せる。

学習手法	モデル	データセットの比率	Distinct-1	Distinct-2
Fine-Tuning (ペルソナ文付加有)	GPT2-XL	1:0	0.125	0.319
Fine-Tuning (ペルソナ文付加無)			0.105	0.266
Prompt-Tuning			0.153	0.377
	GPT-J-6B		**0.183**	**0.546**
Fine-Tuning (ペルソナ文付加有)	GPT2-XL	1:1	0.199	0.526
Fine-Tuning (ペルソナ文付加無)			0.210	0.568
Prompt-Tuning			0.177	0.494
	GPT-J-6B		**0.213**	**0.595**
Fine-Tuning (ペルソナ文付加有)	GPT2-XL	1:5	0.220	0.502
Fine-Tuning (ペルソナ文付加無)			**0.241**	0.643
Prompt-Tuning			0.183	0.463
	GPT-J-6B		0.240	**0.678**

A.2　人手評価

人手評価結果を表 6 に示す。

表 6　Persona Eval Dataset の人手評価. 全ての学習用データセットの比率の結果を載せる。

評価用データセット	学習手法	モデル	データセットの比率	流暢さ	魅力度	関連性	ペルソナ考慮
Persona Eval	Fine-Tuning	GPT2-XL	1:0	3.35	**3.80**	3.52	**3.78**
	Prompt-Tuning			3.49	3.54	**3.85**	3.75
		GPT-J-6B		**3.64**	3.76	3.64	3.70
	Fine-Tuning	GPT2-XL	1:1	3.52	3.70	3.30	3.54
	Prompt-Tuning			3.82	3.74	3.62	3.32
		GPT-J-6B		**3.90**	**3.98**	**3.82**	**3.76**
	Fine-Tuning	GPT2-XL	1:5	3.44	3.62	3.68	**3.82**
	Prompt-Tuning			**3.89**	**3.78**	**3.77**	3.81
		GPT-J-6B		**3.89**	3.70	3.68	3.69
General Eval	Fine-Tuning	GPT2-XL	1:0	3.33	3.22	3.19	3.60
	Prompt-Tuning			3.49	3.64	3.65	**3.78**
		GPT-J-6B		**3.77**	**3.67**	**4.01**	3.61
	Fine-Tuning	GPT2-XL	1:1	3.93	**3.82**	3.77	3.57
	Prompt-Tuning			**4.04**	3.81	**3.96**	**3.81**
		GPT-J-6B		3.98	3.80	3.89	**3.81**
	Fine-Tuning	GPT2-XL	1:5	3.52	**3.80**	**3.84**	**3.83**
	Prompt-Tuning			4.10	3.60	3.66	3.62
		GPT-J-6B		**4.11**	3.71	3.81	3.77
Human				4.31	4.25	4.36	-