雑談対話における会話継続予測に基づくユーザ適応的応答評価

蔦侑磨 [1]　吉永直樹 [2]　佐藤翔悦 *　豊田正史 [2]
[1] 東京大学大学院　[2] 東京大学 生産技術研究所
{tsuta, ynaga, toyoda}@tkl.iis.u-tokyo.ac.jp

概要

雑談対話システムでは、話し相手となるユーザに即した応答を生成することが望ましいが，そのような応答を適切に評価するには、当該ユーザ自身が評価を行う必要があり、コストが大きい．本研究は雑談応答生成において，対話システムが応答する話者の視点を取り入れた自動評価手法を提案する．提案手法は，評価ユーザが会話を継続する応答が良い応答であるという発想から、評価ユーザを考慮した会話の継続性予測タスクを通じて生成応答の主観評価を行う．大規模な Twitter データセットを用いた実験から，ベースラインより高い精度で会話の継続性の予測が行えることを確認し、当該ユーザの介入なしで主観評価を行うことが可能となった。

1　はじめに

GPT-3 [1] などの大規模事前学習済み生成モデルの出現により，近年の雑談対話システムは与えられた発話に対し、ある程度自然な応答ができるようになった [2, 3, 4, 5]．今後、雑談対話システムがスマートスピーカやスマートフォンに搭載され，日常的に会話を続けるためには，過去の会話履歴やユーザの最新のプロフィールなどを参照しながら対象ユーザに即した応答を返すなど，応答生成モデルの個人適応が求められると考えられる [5, 6, 7].

では，このように特定のユーザとの会話に特化した対話システムをどのように評価するのがよいのだろうか．ユーザが好む応答は、個別ユーザに依存する [8] ため、ユーザによる対話的評価 [3, 4, 5] は理想的ではあるものの，そのようなコストが高く，再現性が担保できず，評価の信頼性が評価者に依存する [9] という問題もあり，効率的に対話システムを開発する上で課題が残る．雑談対話システムの自動評価に関する研究 [10, 11, 12, 13, 14] は様々あるが，これらの既存の自動評価手法では，対話システムの話し相手であるユーザの視点は考慮されていない．

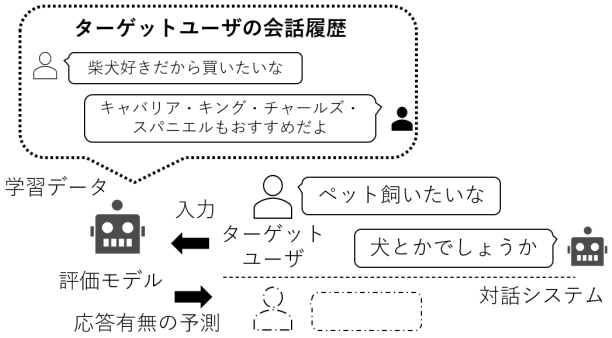

図 1　システム応答に対して話し手が応答するかどうかを予測するタスク

このような背景を受けて、我々は対話の継続性予測タスクを通じた生成応答の自動主観評価 [15]（図 1）の研究を進めている。この研究では、Twitter などから得られる大規模対話ログにおいて、ユーザが応答した発話は応答しない発話よりそのユーザにとって良い発話であるという仮定に基づき、ユーザのプロフィールを考慮して会話の継続性予測を行うモデルを用いて自動で主観評価を行うことを実現した．予備的な実験を通して、継続性の予測ができることを確認したものの、プロフィール文のあるユーザしか評価に利用できず、対話システムの生成応答の評価まで行えていなかった。

そこで本稿では、ユーザトークン [16] を用いて会話の継続性予測を個人適応する手法を提案し，これをプロフィール文を用いて個人適応する手法 [15] との比較を行う。また，実際に対話システムの生成応答に対して提案手法を適用し，評価ユーザによる人手評価との相関に基づき、提案手法の自動主観評価手法としての有効性を確認する．

実験では，Twitter 上の膨大な会話データセットを用いた実験により，本稿で提案するユーザトークンを用いた会話の継続性予測手法が既存のプロフィール文を用いた会話の継続性予測手法と同程度の精度で会話の継続性を予測できることを確認した．また，評価ユーザによる対話システムの生成応答に対

* 現在は企業に所属.

する人手の継続性評価にとの相関については、一部のユーザに関して，人手の継続性評価と提案手法の予測結果が相関する結果が得られた．

2 関連研究

対話システムの主観評価を自動で行う研究は先行研究 [15] のみである．本節では，雑談対話生成において話し相手となるユーザや，対話システム自体のキャラクター性を考慮して応答を生成する研究を説明する (§ 2.1)．続いて，生成応答に対する自動評価に関する関連研究 (§ 2.2) を紹介する．

2.1 対話システムの個人適応

雑談対話応答生成の研究分野では，対話システムに個性を埋め込むことで，一貫した応答の生成やシステムのキャラクター付けを可能にした研究が多く行われている．例えば，Li らは，話者に固有の学習可能な埋め込み表現を利用することで話者の性別や住所などを再現することを可能にした [16]．また，PersonaChat [17] やマルチセッションチャット [5] には，特定の個性を持った人物を演じる会話が含まれており趣味や生活環境などの話者の疑似的な特徴を含む会話が利用可能である．

これらの研究では，与えられた話者性に合致する応答を生成する対話システムの開発 [17, 18, 19] だけでなく，会話相手の特徴を意識した会話を可能にしてエンゲージメント向上を目指した研究 [20, 21] も存在する．これらの研究は，キャラクター性の理解・表出により，エンゲージメントの高い回答を生成することが可能となったが，人手評価は応答対象のユーザではない第三者が行なっており，応答対象ユーザの観点からの評価はなされていない．

2.2 雑談対話システムの自動評価研究

既存の生成応答の自動評価手法は，生成応答と参照応答との一致度に基づく評価 (例えば，BLEU [22])，あるいは参照文を用いない評価モデルによる生成応答の評価 (例えば，RUBER [23]) があるが，現在のところ既存の生成応答の自動評価手法は不安定であり，機械翻訳における BLEU のような標準的な生成応答の自動評価手法は存在しない．我々は，生成応答の評価は評価者に強く依存し，対話システムのユーザと異なる評価者が行った人手評価との相関を取ることに自動評価の本質的な問題があると考えて，主観評価を自動化することを考えた．

対話システムの評価モデルの学習において，システム応答や実応答に対する人手の評価を用いた研究が存在する [11, 13]．また，SNS 上のフィードバックデータを学習データとして用いる評価手法が提案されている [12]．しかし，これらの研究は主観的な自動評価を行うことを目的としていない．

3 提案手法

3.1 会話継続性予測に基づく評価

雑談対話システムのユーザを意識した評価を可能するためには，そのユーザに好まれる文を理解すること，つまり当該ユーザが好む実応答と，好まない実応答が必要となる．そのデータはどのようにして得られるのだろうか？我々は先行研究 [15] において，この問いに対話の継続性に関する自然注釈を利用することで答えた．すなわち，与えられた応答に対し，ユーザが応答し，会話が継続しているかどうかが，ユーザの応答に対する好みを反映していると考えた．本研究でもこのアイデアを利用し，与えられた応答に対して当該ユーザが会話を継続するかを予測するモデルを学習し，その予測確率を評価値として，与えられた応答の主観評価を行う．

3.2 ユーザを考慮した会話継続性の予測

本研究では，先行研究で提案したプロフィール文を利用した会話の継続性予測手法に加えて，Li ら [16] が応答生成タスクにおいて提案したユーザトークンを用いた会話の継続性予測手法を検討する．本手法では，各ユーザの発話に対してユーザ特有の特殊トークンを発話の先頭に付加し，ユーザの会話データから与えられた応答に対する当該ユーザの会話継続性の予測モデルを学習する．プロフィールでなく、ユーザトークンを用いることで、プロフィール文が利用できない Reddit [1) 2)] のような会話データを用いた対話システムの評価を行うことができる．また，プロフィール文が利用できる Twitter 上の会話データでも，プロフィール文が存在しないユーザを対象とした評価を行うことができる．

3.3 会話継続性予測タスク

本研究の応答継続性予測タスクは，問いかけられたユーザが返答を行うかを予測するタス

1) https://www.reddit.com/
2) Reddit ユーザの話者の特徴データの抽出可方法は公開されているが [24]，抽出データが未公開のため再現が難しい．

クである．具体的には，N 個の発話を含む会話 $U = [u_0, u_1, ..., u_{N-1}]$ が 2 人の話者 s_i と s_j によって交互に行われている[3] と仮定し（なお，u_{N-1} は s_j による発言とする），予測モデルは会話における s_i の応答確率 $P(u_N = exists \mid U, s_i)$ を予測する．そのため，モデルは会話文 U と，返答の有無を予測するユーザ s_i が入力され，返答予測確率を出力する．ユーザを考慮するために固有のトークンを利用する際には，対象となるユーザ s_i に対して固有トークン（[USER i]）を用意し，ユーザ s_i の全ての発言の前にそのトークンを結合しモデルに入力する．考慮されないユーザは，代わりに汎用のユーザトークンが用いられる．プロフィールでユーザを考慮する場合は，ユーザ固有トークン [USER i] の代わりに二人の話者を区別する目的で汎用トークン（例えば，[SPK A] と [SPK B]）を用い，ユーザ s_i のプロフィール文（p_i）をモデルに入力する会話文の先頭に結合する ([SPK A] p_i { 会話文 }）．なお，モデルの最大入力長を超える場合，古い文脈が優先的に無視される．

4 実験

本研究は，話者が識別可能な Twitter から収集した膨大な実会話ログを用いて応答評価モデル（会話の継続性予測モデル）の学習，評価を行う．まず，会話データ中の対象ユーザに関して評価モデルを学習し，人同士の会話について対象ユーザの応答予測が可能かどうかを確認する（以下，内的評価と呼ぶ）．次に，雑談対話システムによる生成応答に対してアノテータによる評価を行い，これを正解ラベルとして応答評価モデルを評価した（以下，メタ評価と呼ぶ）．

4.1 会話データ

本研究では，Twitter 公式 API を用いて収集した Twitter 上の日本語の投稿（ツイート）を用いた．会話データとして構築するために，2 人のユーザ間の投稿とその返信からなる一連の投稿のみを利用した．2017 年 1 月から 2018 年 3 月の間に 30 回以上会話したユーザ（以下，ターゲットユーザ）をランダムに 1 万人抽出し，1 人あたり最大 400 回の会話，合計約 125 万件の会話データを収集した．収集した会話を，モデルの訓練データおよび検証データとして使用した[4]．このうち，各ユーザの 5%の会話を検証用データとして確保し，残りを訓練データとして利用した．訓練データの統計量などは付録 § A.1 に記載した．また，内的評価のためのテストデータとして，2018 年 3 月から 12 月の間のターゲットユーザの会話約 200 万件を別途収集した．

雑談対話システムの生成応答を利用して行うメタ評価には，評価者と対話システムとの会話，および正解ラベルとして対話システムの生成応答に対する評価者の評価が必要である．そこで，と研究室のメンバー 2 名がそれぞれ Twitter 上で行った会話を収集した．これら会話データ，それぞれ（訓練・検証・テスト）の順で（165・43・27）・（19・6・10）の会話を上記のデータセットに追加した．

4.2 評価モデルとハイパーパラメタ

実験では，事前学習済みの日本語 BERT [25][5] を評価モデルのベースモデルとして採用し，分類モデルとして微調整 (fine-tuning) した．ユーザを考慮して会話の継続性予測を行うことで，予測精度が上がるか，すなわち，より信頼性のある応答評価が可能かを確認するため，ユーザを考慮したモデルと考慮しない（すなわち，ターゲットユーザ独自のトークンやプロフィールを用いない）応答評価モデルを学習した．また，メタ評価にあたって，Gao らの研究 [12] を参考に妥当性検証用のモデルを利用しているが，すべての学習可能なモデルで同様の利用を行っているため，モデル間の比較に影響はない（§ A.4）．その他の詳細は，付録 § A.2 に記載する．

4.3 メタ評価のための人手評価の収集

対話システムの生成応答での性能を評価するために，正解ラベルとなる人手評価データが必要である．なお，生成応答までの会話文脈は § 4.1 で述べたアノテータのテストデータから作成し，生成応答によってアノテータに問いかける形式にした．Ji らの研究 [26] を参考に，アノテータ（共著者 1 名，学生 1 名）は，7 種の対話モデル (§ A.3) による生成応答と 1 つ実応答を 0 から 100 のスコアで評価した．結果，400 件と 192 件の会話データおよび正解ラベルを得た．

3） 本研究では 簡単のため二人の間での会話を扱う (§ 4.1)

4） なお，自分への返信，送信元の URL や名前に「bot」が含まれる bot による発言，3 語以下の発言などの会話文として不適切な文および会話は除外した．また，Twitter API により 2021 年 12 月時点で，追加の返信がないことを再確認した．

5） https://huggingface.co/cl-tohoku/bert-base-japanese-v2

4.4 評価モデル

ベースラインとしては，ユーザごとに多数ラベルを予測する手法（以下，多数クラス）を用いる．評価モデルとしては，ユーザを考慮しない手法 (BERT) と，ユーザを考慮する提案手法を追加したものを利用する．§ 3.2 で説明したように，ユーザの考慮の方法は 2 種類（ユーザトークン，プロフィール) 検討し，それぞれを排他的または同時に考慮する手法（以下,「全て」と呼称）も比較する．

4.5 評価指標

評価モデルの内的評価における評価指標には，実際の返答の有無を正解ラベルとする分類精度 (acc.) を用いる．また，Twitter 上の実際の会話ログでは，一般に問いかけが応答されやすく会話が継続する傾向なため (図 2)，この偏りを是正するために会話の有無の両方のラベルに対する F_1 を平均化したマクロ F_1 も計測した．なおメタ評価では，§ 4.3 で述べたように，正解ラベルが [0,100] の範囲のため，分類問題ではなくモデル出力との相関値 (Pearson's r) を記載した[6]．また，内的評価の結果との比較のために，50 以上の評価値の応答を正例とみなした分類問題としての結果も記載した．なお，メタ評価ではユーザごとの会話継続性の割合が異なることを考慮し，ユーザごとに結果を表記した．

4.6 実験結果

表 1 は，Twitter ユーザ間の会話での内的評価での結果である．この結果から，ユーザを考慮しない BERT による応答継続性予測は，（ユーザ性を考慮した）ユーザごとの多数クラスを出力する単純な方法と差がないことが確認された．一方で，ユーザを考慮した手法は，いずれもより高い予測精度となることが確認できた．特に，その方法としてプロフィールよりもユーザ独自のトークンを用いる方が有効であることが確認された．

表 2 は，対話システムによる生成応答（および実応答）に対する予測を人手の評価ラベルに基づいて評価した結果である．アノテータ 1 に関しては，正解率・マクロ F_1 ではユーザ独自のトークンによりユーザを考慮したモデルの予測が最も人手評価と一致しており，人手評価との相関ではプロフィールも考慮すること (+all) で最も人手評価と一致することが確認できた．一方で，アノテータ 2 ではユーザを考慮しない方法がすべて人手評価に近いことが確認できた．内的評価とは異なり，メタ評価では会話相手が対話システムであるため，アノテータ 1 では対話システムが会話相手でも問題なく受け入れられたが，アノテータ 2 では対話システムに対する返答の傾向が変化した可能性などが考えられる．

6） このため，メタ評価では常に多数クラスを出力する方法は記載しない．

表 1 Twitter の会話ログから対象ユーザーが反応したかどうかを予測することに関する内的評価.

評価モデル	acc.	F_1
多数クラス	0.683	0.659
BERT	0.668	0.653
+ ユーザトークン	**0.751**	**0.744**
+ プロフィール	0.746	0.738
+ 全て	**0.751**	**0.744**

表 2 雑談対話システムに対する生成応答への応答の有無の予測におけるメタ評価.

評価モデル	アノテータ 1			アノテータ 2		
	acc.	F_1	r	acc.	F_1	r
多数クラス	0.555	0.335	-	0.293	0.227	-
BERT	0.680	0.680	0.516	**0.654**	**0.638**	**0.429**
+ ユーザトークン	**0.719**	**0.718**	0.520	0.634	0.621	0.409
+ プロフィール	0.711	0.710	0.517	0.647	0.632	0.405
+ 全て	0.711	0.711	**0.524**	0.639	0.626	0.409

5 おわりに

本論文では，対話システムの評価において、システムの話し相手であるユーザの視点を取り入れた自動主観評価手法を提案した．提案手法では，与えられた応答に対してユーザが返答を行う確率を予測し，これをユーザによる応答評価に転用する．ユーザの視点を取り入れるために，ユーザトークンを用いてユーザをモデル化した．Twitter から収集した人同士の膨大な会話ログを用いた評価では，ユーザを考慮することで会話の継続性予測の性能が改善すること，すなわち，より信頼性の高い評価が行えることを確認した．さらに，本手法を複数の雑談対話システムによる生成応答の評価に適用した場合，提案手法が一部のユーザでは有効であることを確認した．

謝辞

この研究は国立情報学研究所（NII）CRIS と LINE 株式会社が推進する NII CRIS 共同研究，および JSPS 科研費 JP21H03494 の助成を受けています．人手評価にご協力くださった研究室の方々に深く感謝致します．

参考文献

[1] 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.

[2] Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. DIALOGPT : Large-scale generative pre-training for conversational response generation. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations**, pp. 270–278, Online, July 2020.

[3] Daniel Adiwardana, Minh-Thang Luong, David R. So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, and Quoc V. Le. Towards a human-like open-domain chatbot, 2020.

[4] Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Kurt Shuster, Eric M. Smith, Y-Lan Boureau, and Jason Weston. Recipes for building an open-domain chatbot, 2020.

[5] Jing Xu, Arthur Szlam, and Jason Weston. Beyond goldfish memory: Long-term open-domain conversation. In **Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 5180–5197, May 2022.

[6] 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, May 2022.

[7] Sanghwan Bae, Donghyun Kwak, Soyoung Kang, Min Young Lee, Sungdong Kim, Yuin Jeong, Hyeri Kim, Sang-Woo Lee, Woomyoung Park, and Nako Sung. Keep me updated! memory management in long-term conversations. **arXiv preprint arXiv:2210.08750**, 2022.

[8] Shikib Mehri and Maxine Eskenazi. USR: An unsupervised and reference free evaluation metric for dialog generation. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 681–707, 2020.

[9] Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A. Smith. All that's 'human' is not gold: Evaluating human evaluation of generated text. 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. 7282–7296, 2021.

[10] Michel Galley, Chris Brockett, Alessandro Sordoni, Yangfeng Ji, Michael Auli, Chris Quirk, Margaret Mitchell, Jianfeng Gao, and Bill Dolan. deltaBLEU: A discriminative metric for generation tasks with intrinsically diverse targets. In **Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)**, pp. 445–450, 2015.

[11] Ryan Lowe, Michael Noseworthy, Iulian Vlad Serban, Nicolas Angelard-Gontier, Yoshua Bengio, and Joelle Pineau. Towards an automatic Turing test: Learning to evaluate dialogue responses. In **Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 1116–1126, 2017.

[12] Xiang Gao, Yizhe Zhang, Michel Galley, Chris Brockett, and Bill Dolan. Dialogue response ranking training with large-scale human feedback data. In **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 386–395, 2020.

[13] Sarik Ghazarian, Ralph Weischedel, Aram Galstyan, and Nanyun Peng. Predictive engagement: An efficient metric for automatic evaluation of open-domain dialogue systems. **Proceedings of the AAAI Conference on Artificial Intelligence**, Vol. 34, No. 05, pp. 7789–7796, Apr. 2020.

[14] Yuma Tsuta, Naoki Yoshinaga, and Masashi Toyoda. uBLEU: Uncertainty-aware automatic evaluation method for open-domain dialogue systems. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop**, pp. 199–206, 2020.

[15] 蔦侑磨, 吉永直樹, 佐藤翔悦, 豊田正史. パーソナリティを考慮した雑談対話の会話継続可能性評価. 言語処理学会第 28 回年次大会発表論文集, pp. 583–587, 2022.

[16] 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, 2016.

[17] 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, 2018.

[18] Hyunwoo Kim, Byeongchang Kim, and Gunhee Kim. Will I sound like me? improving persona consistency in dialogues through pragmatic self-consciousness. In **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 904–916, 2020.

[19] Haoyu Song, Yan Wang, Wei-Nan Zhang, Xiaojiang Liu, and Ting Liu. Generate, delete and rewrite: A three-stage framework for improving persona consistency of dialogue generation. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 5821–5831, 2020.

[20] Qian Liu, Yihong Chen, Bei Chen, Jian-Guang Lou, Zixuan Chen, Bin Zhou, and Dongmei Zhang. You impress me: Dialogue generation via mutual persona perception. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 1417–1427, 2020.

[21] Itsugun Cho, Dongyang Wang, Ryota Takahashi, and Hiroaki Saito. A personalized dialogue generator with implicit user persona detection. In **Proceedings of the 29th International Conference on Computational Linguistics**, pp. 367–377, 2022.

[22] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: A Method for Automatic Evaluation of Machine Translation. In **Proceedings of 40th Annual Meeting of the Association for Computational Linguistics**, pp. 311–318, 2002.

[23] Chongyang Tao, Lili Mou, Dongyan Zhao, and Rui Yan. RUBER: An Unsupervised Method for Automatic Evaluation of Open-Domain Dialog Systems. In **AAAI Conference on Artificial Intelligence**, pp. 722–729, 2018.

[24] 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, 2018.

[25] 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, 2019.

[26] Tianbo Ji, Yvette Graham, Gareth Jones, Chenyang Lyu, and Qun Liu. Achieving reliable human assessment of open-domain dialogue systems. In **Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 6416–6437, May 2022.

[27] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In **7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019**. OpenReview.net, 2019.

[28] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language Models Are Unsupervised Multitask Learners. 2019.

[29] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, L ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, **Advances in Neural Information Processing Systems**, Vol. 30. Curran Associates, Inc., 2017.

[30] 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.

表 3 訓練会話におけるターン・文字数に関する統計．

種別	最小	平均	最大
1 会話当たりの発言数	2	3.43	117
1 発言当たりの文字数	4	31.05	191
1 会話当たりの文字数	8	106.49	3451

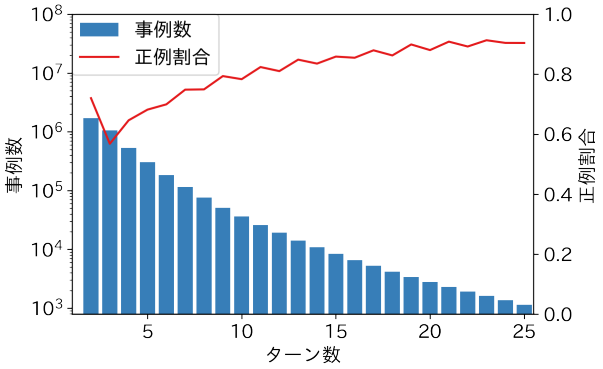

図 2 訓練データのターンごとの応答返答率の分布．ターンは，問いかけを含んでそれまでに何回発言が行われたかを意味し，正例割合はその中でどの程度問いかけに返答されたか（つまり，訓練データでの正例の割合）を示す．

A 付録 (Appendix)

A.1 訓練データの情報

表 3 に訓練データのターン・文字数に関する統計を，図 2 に訓練データにおけるターンごとの事例数と，正例の割合を示す．

A.2 モデルの学習パラメータ

すべてのモデルのハイパーパラメタは，学習率 $3e-5$，バッチサイズ 64，エポック数 5 で学習し，optimizer として AdamW [27]，損失関数には交差エントロピー損失を使用した．なおモデルの実装には全て Hugging Face Transformers[7] を用いた．実験に用いた全てのライブラリのライセンスは，学術目的での使用を許可している．

A.3 雑談対話システム

メタ評価のためのデータセットには，対話システムによる生成応答が必要であるため，§ 4.1 で述べた訓練データを用いて事前学習済み GPT-2[8] [28] の転移学習を行った．対話システムは，再学習時に事前学習パラメタを継承するかランダム初期化するか，またユーザトークンを利用するかしないか（§ 3.3 と同様の方法) を組み合わせて、4 種のモデルを学習した．さらに，事前学習済み Transformer[9] [29, 30] の公開済みの 3 種のモデルも使用した．上記の合計 7 種の応答生成モデルの生成応答をメタ評価に利用した．

A.4 妥当性評価モデル

実験における評価モデルは，Twitter 上の人同士のデータで学習される．一方で，メタ評価では対話システムの応答の評価がメインであり，対話システムの性能が低い場合，生成された応答が理解できない非文になっている可能性が考えられる．理解可能な人同士の文章のみで学習されたモデルの，非文を含むデータに対する脆弱性を考慮し，補強する必要がある．

Gao らの研究 [12] では，実会話となる問いかけと応答のペアを正例，ランダムな問いかけと応答のペアなどを不例として学習した（従来の評価モデルと同様の）妥当性評価モデルにより，応答として不適格な文にペナルティを与えることで，生成応答に対するエンゲージメント評価の性能向上を行った．そこで，本研究でも同様に，BERT の Next Sentece Prediction タスクにより学習したモデルを妥当性評価モデルとして採用し，生成応答にペナルティを与えた．なお，このペナルティはすべての生成応答に同一の条件で与えられ，その後の評価モデルにも依存しないため，本研究における実験での学習可能なモデルの比較には影響がない．

応答妥当性評価モデルの学習のためのデータセットは，§ 4.1 の学習データと同じ会話データから作成した．正例は，会話データのうち正例（会話の後に応答があるもの）約 1M の対話データを用いた．負例には，会話中のすべての応答について，会話文脈をランダムにサンプリングして組み合わせることで作成した約 2M の会話データを利用した．学習時には，正例と負例が同数となるように学習した．

7） https://github.com/huggingface/transformers
8） https://huggingface.co/rinna/japanese-gpt2-medium
9） https://github.com/nttcslab/japanese-dialog-transformers