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

蔦侑磨 [1]　吉永直樹 [2]　佐藤翔悦 [2]　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, shoetsu, ynaga, toyoda}@tkl.iis.u-tokyo.ac.jp

概要

雑談対話を対象とした自動評価手法に関する既存研究の多くは，実在する会話の応答全てを等しく妥当な応答として想定し，評価手法を構築しているため，ユーザが興味を持たない応答であっても高い評価を行う．そこで本研究では，生成応答がユーザの返答を促し会話が継続するかを評価するために，ユーザの返答の有無を予測するタスクから学習された自動評価手法を提案する．この際に同一の発話内容でも返答可能性は人によって変化しうるため，会話相手であるユーザのパーソナリティを考慮する．実験では内的評価により会話初期の継続可能性の判定において提案手法の有効性が確認された．

1　はじめに

人は一日の約 15%以上の時間を雑談会話に費やすと報告 [1] されるように，日常的に会話を行い雑談欲求を満たしている．しかし，個人ドライバーや独り身の高齢者など適切な会話相手がいない状況や近年の新型コロナウイルスの影響など，会話自体が困難で雑談欲求があっても満たせない状況がある．加えて，人の話し相手となりうるスマートスピーカーが爆発的に普及したこともあり，雑談対話システムへの関心が急速に高まりつつある．

ユーザの雑談欲求を満たせる雑談対話システムの構築のためには，その生成応答が応答として妥当か，そして生成応答により会話が継続するかなどの多様な観点を考慮して，応答を生成・評価する必要がある．近年の雑談対話研究では，対話システムが頻繁に「そうですか」のような汎用的で無難な応答 (dull response) [2] を繰り返すことが問題視されているが，与えられた発話に対する実応答を評価における参照応答に用いる通常の評価手法では，このような dull response に高い評価を与えてしまう．この解決のため，Ghazarian らは応答継続可能性に関する人手評価を教師データに利用した評価手法により応答の妥当性評価以外の評価を行っている [3]．しかし，少数の人手評価を学習データとした自動評価器は，その汎用性の低さが問題となる．

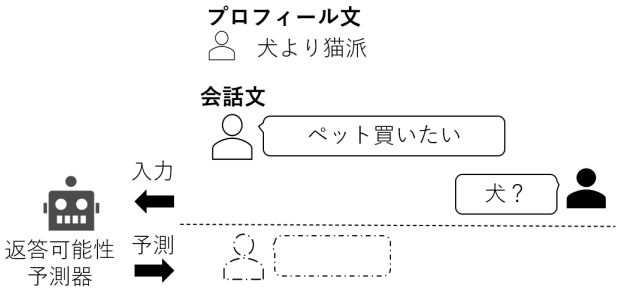

図 1　ユーザの返答の有無を予測するタスク

本研究ではユーザのパーソナリティを考慮して，ユーザとの継続的な会話を促す生成応答を高く評価する自動評価手法を提案する．提案手法では，人手の評価データを用いることなく評価モデルを学習することで，多様なドメインの応答を評価できる汎用性の高い手法を目指す．具体的には，ある会話において次に返答が行われたかどうかを予測する会話継続可能性評価タスク（図 1）による学習を行う．さらに，同一の発話内容でも返答可能性が人によって変化しうるため，提案手法では会話相手のパーソナリティを考慮したモデルを構築する．

本研究で目的とする雑談対話システムのユーザのパーソナリティを考慮した応答継続可能性評価を行うには，パーソナリティの明示されたユーザに関する対話システムとの長期的な会話データが必要である．しかし対話システムとの長期的な会話データを得ることはコストが高く，さらに対話相手であるユーザのパーソナリティを得るのはプライバシの観点から非常に困難である．そこで本研究では，Twitter 上の人同士の会話データについて，そのプロフィール文をパーソナリティとみなして，実験用の会話データを構築した．実験ではこの会話データに関する内的評価から会話初期の継続可能性の判定においてパーソナリティを考慮することにより予測精度が向上することが確認された．

2 関連研究

本章の 2.1 節では自動評価手法に関する既存研究を，2.2 節ではパーソナリティ（個人属性）を考慮した既存の雑談対話研究の紹介を行う．

2.1 自動評価手法

雑談対話システムは基本的に人手評価によって評価されるが，効率的な開発のために人手評価と相関する自動評価手法の確立が強く期待されている．実応答を再現することを目的とした雑談対話システムの単純な評価方法として，実応答とのトークンの重複により計算する BLEU [4] や ROUGE [5] が利用可能だが，Liu らにより人手評価との相関が低いことが指摘された [6]．これらの単純な評価手法からの改善として，評価時に利用できる唯一の実応答だけでなく雑談対話で許容される多様な応答を利用した，応答多様性を考慮した手法 [7, 8, 9] が提案されている．一方で，生成応答と入力発話の関連性に着目した評価手法に関して，人手評価を利用した教師あり学習手法や負例サンプリング [10] を用いた教師なし学習手法なども多数提案されている [11, 12, 13]．しかし既存研究の多くは実在する会話の応答全てを等しく妥当な応答として想定するため，dull response のような応答であっても高い評価を与えてしまう．ユーザの興味を引かない応答では，会話が継続せずユーザの雑談欲求を満たすことができないため，これらの評価手法では会話継続可能性や会話満足度などの評価には適さない．

本研究と同様の動機に基づく研究としては，Ghazarian らが応答継続可能性に関する人手のアノテーションデータを教師データに利用した自動評価手法を提案している [3]．しかし人手評価を教師データとして利用するため，そのコストや評価器が学習データのドメインに対して過学習する可能性があることが問題点として挙げられる．提案手法では，人手のアノテーションデータを利用しない学習手法により，多様なドメイン上の会話に利用可能な汎用性の高い手法を目指す．

2.2 個人属性を考慮した雑談対話研究

雑談対話における個人属性（パーソナリティ）を考慮した研究として，デモグラフィック属性などのパーソナリティについて会話上で再現する対話システムが提案されている [14, 15, 16] が，これらの研究はパーソナリティの再現や理解，あるいは会話の一貫性向上が主体となっている．一方で本研究は個人ごとの応答の嗜好性が異なるというパーソナリティを意識した，会話継続性の評価に関する研究となっており，従来のパーソナリティを考慮した研究とは利用目的が異なる．また雑談応答生成では与えられたパーソナリティを再現する研究 [14, 15] は多くの研究で古くから行われているが，会話相手のパーソナリティを理解することを目的とした研究は新しく少数である [16]．本研究は自動評価手法として，特に会話相手のパーソナリティを考慮するという点で新規的な研究といえる．

3 提案手法

本研究ではユーザとの継続的な会話を促す生成応答を高く評価する自動評価手法を提案する．この際に同一の発話内容でも返答可能性が人によって変化しうるため，提案手法では会話相手のパーソナリティも考慮する．先行研究 [3] とは異なり，多様なドメインにおいて汎用性の高い手法としての構築を目指すため，人手のアノテーションを利用しない教師なし学習により分類器を学習する．そこで提案手法では，会話の特定のターンにおいて次に返答が行われたかどうかを予測する会話継続可能性評価タスク（図 1）として学習を行う．

3.1 会話継続可能性評価モデル

本研究では図 1 のように，会話履歴や応答文と同時に返答予測を行う対象のパーソナリティを入力し，実際に返答を行うか予測するタスクにより評価モデルを学習する．提案手法での会話継続可能性評価を行うモデルには BERT [17] を利用する．

3.2 評価器への会話データの入力形式

BERT を入力の分類タスクに用いる場合には入力の先頭に [CLS] トークンを，入力中の各文末に [SEP] トークンを入力するため，図 2 のような形式でデータを入力する．また通常の BERT の利用では最大 2 文のみを入力するが，提案手法では 3 文以上が入力される可能性がある．このため本研究では BERT に入力する Segment Embeddings についてプロフィール文と会話履歴を同一の分散表現とした．また BERT では入力長に制限があるため，この制限を超える場合，古い会話履歴を優先的に省略する．

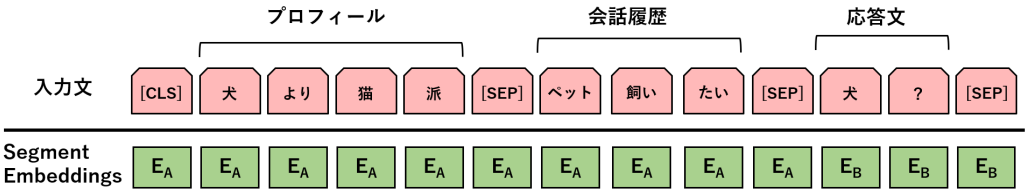

図 2　評価器（BERT）への入力データの構造

3.3　学習データ

提案手法では，会話の特定ターンにおいて次に返答が行われるかどうかを予測する会話継続可能性評価タスクとして学習を行う．本研究で目的とする雑談対話システムのユーザのパーソナリティを考慮した応答継続可能性評価を行うには，パーソナリティの明示されたユーザに関する対話システムとの長期的な会話データが必要である．しかし対話システムとの長期的な会話データを得ることがコストが高く，さらに対話相手であるユーザのパーソナリティを得るのはプライバシの観点から非常に困難である．

そこで本研究では，Twitter 上の人同士の会話データについて，そのプロフィール文をパーソナリティとみなして，実験用のデータを構築した．具体的には，大規模会話ログにおける特定の 2 話者間の会話について，片方を対話システム，他方をユーザと想定する．そして対話システムが発言した後にユーザが返答するかどうかを予測するタスクとしてデータを構築する．

3.4　会話データにおける正例と負例

本研究で負例として扱われる応答文には返答が行われない会話データを利用する．このため会話が終了した時点での最後の応答文を負例として扱う．一方で，全ての正例には実際に返答が行われない会話データを利用するため，負例より前のすべての応答文を利用する．

4　実験

本章では，実験に利用したデータやパラメタ，実験結果について説明を行う．

4.1　大規模日本語対話コーパス

実験で利用する大規模日本語対話コーパスは，著者らの研究室で 2011 年 3 月から継続的に収集している Twitter アーカイブから構築した．本研究では会話が実際に終了したかの判断が重要なため，Twitter API を利用して 2021 年の時点で実際に応答が存在しないかを確認した．Twitter 上のツイート文のやりとりを会話データとみなして，メンションもしくはリツイート以外の投稿を発話，それに連なるメンションを応答とした対話ツリーを抽出した．但し，対話を行うユーザは常に 2 ユーザのみ存在し，同一のユーザが自身の投稿に続けてメンションを行っていない対話ツリーのみを利用した．

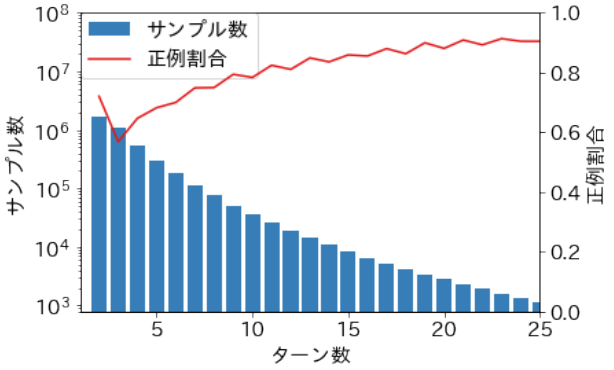

図 3　訓練データにおける各ターンの会話数と正例割合

4.2　実験に利用する会話データ

実験では 4.1 節で説明した会話データから最大 25 ターンまでの会話データを利用した．分類器の学習データには 2017 年 1 月中の全データを，評価データには 2018 年 1 月中のデータを，プロフィール文は 2020 年 11 月時点のデータを利用した．プロフィール文がこれらの中で最新のデータとなっているが，API などを利用して会話データを収集する場合，最新の情報のみしか利用できないことが多く，過去のデータを利用しない本実験の実験設定は実験の再現性の観点から問題ないと考えられる．検証用のデータは学習データ中の無作為に抽出した 1 割のデータを利用し，残りを訓練データに利用した．詳細は表 1 に記載する．また訓練データのターン数ごとの詳細を図 3 に示す．

表 1　実験に利用したデータの詳細（概算）

種別	会話件数	正例割合
訓練	4M	0.67
検証	450K	0.67
テスト	25K	0.65

表 2　テストデータにおける分類精度

モデル	Acc.	Prec.	Recall	F_1
提案手法	**0.73**	**0.74**	0.91	0.81
ベースライン	0.70	0.71	0.92	0.80

4.3　実験設定

提案手法で利用する BERT には事前学習済みモデルを利用した[1)]．提案手法の比較手法としてのパーソナリティを考慮しない（プロフィール文を入力しない）方法をベースラインとして比較する．全てのモデルは，学習率 $1e-5$，最適化手法 Adam [18]，損失関数は交差エントロピー，バッチサイズ 8，エポック数 1 で学習した．

4.4　実験結果

表 2 にテストデータにおける分類精度を示す．この結果から，パーソナリティを考慮した提案手法はパーソナリティを考慮しないベースラインと比べて，F_1 スコアに関しては同程度だったものの，若干高い精度で予測できたことが確認できる．

次に実験結果に関する詳細な解析として，訓練データに出現したユーザを既知ユーザ，その他のユーザを未知ユーザに分類した分類精度を示す．同時にターン[2)]（会話履歴に応答文を加えた発言回数）ごとに分類した分類精度を示す．但し 3 ターン以上の会話については結果が全て同様であったため，2 ターン目と 3 ターン以上で分類した結果を示す．ユーザ・ターン別の分類精度を表 4 に示す．なお各分類時のデータ詳細を表 3 に示す．この結果から，特に 2 ターン目に関して，パーソナリティを考慮することによって予測精度と F_1 スコアが共に向上することが確認できる．一方で 3 ターン以上のデータに関しては提案手法とベースラインでほぼ同等であることが確認できる．

4.5　考察

未知ユーザ・既知ユーザ問わず 2 ターン目の会話ではプロフィールを考慮することで分類精度が向上することが確認できた．一方で，3 ターン以上の会話では分類精度の向上が確認できなかった．これは Twitter 上のプロフィール文は主にユーザの興味のあるトピックやコンテンツに関して記載されているため，会話初期において評価対象のユーザの興味のある話題かといった判断を行うのに役立ったのだと推測される．一方で，長期的な会話における相手の話し方や会話の進め方などに関してはユーザごとの嗜好性がプロフィール文からは読み取りづらいため，3 ターン以上の会話では精度差がなかったのではないかと予想される．

1）　https://github.com/cl-tohoku/bert-japanese
2）　例えば図 1 のターン数は 2 である．

表 3　テストデータ詳細（ユーザ・ターン別）

ユーザ	ターン	データ数	正例割合
既知	2	7K	0.68
未知	2	4K	0.73
全て	2	11K	0.70
既知	≥ 3	9K	0.60
未知	≥ 3	5K	0.64
全て	≥ 3	14K	0.61

表 4　テストデータにおける分類精度（ユーザ・ターン別）

ユーザ	ターン	モデル	Acc.	Prec.	Recall	F_1
既知	2	提案手法	**0.80**	**0.80**	0.94	**0.86**
	2	ベースライン	0.73	0.73	0.96	0.83
未知	2	提案手法	**0.79**	**0.80**	0.94	**0.87**
	2	ベースライン	0.75	0.77	0.94	0.85
全て	2	提案手法	**0.79**	**0.80**	0.94	**0.87**
	2	ベースライン	0.74	0.74	0.95	0.84
既知	≥ 3	提案手法	0.67	0.67	0.89	0.77
	≥ 3	ベースライン	0.66	0.66	0.89	0.76
未知	≥ 3	提案手法	0.70	0.71	0.89	0.79
	≥ 3	ベースライン	0.70	0.71	0.90	0.79
全て	≥ 3	提案手法	0.68	0.68	0.89	0.77
	≥ 3	ベースライン	0.67	0.68	0.89	0.77

5　おわりに

本研究ではパーソナリティを考慮した応答継続可能性評価器を構築し，パーソナリティを考慮しない手法との比較を内的評価により行った．結果として，会話初期についてはパーソナリティの考慮により応答継続可能性の予測精度が向上したが，長期的な会話については精度差が確認できなかった．

今後の課題として，評価対象の長期的な会話におけるユーザの応答継続可能性を理解するために，過去の会話履歴などをパーソナリティとして導入することを検討したい．また本研究では，パーソナリティを考慮した手法とそうでない手法を内的評価による比較のみを行った．今後は実践的な環境を想定するために，実際に雑談対話システムが行った会話についての評価や，人手評価との比較を検討する．

謝辞

この研究は国立情報学研究所（NII）CRIS と LINE 株式会社とが推進する NII CRIS 共同研究の助成を受けています。

参考文献

[1] 花絵小磯, 智行土屋, 涼子渡部, 大輔横森, 正夫相澤, 康晴伝, KOISO Hanae, TSUCHIYA Tomoyuki, WATANABE Ryoko, YOKOMORI Daisuke, AIZAWA Masao, DEN Yasuharu. 均衡会話コーパス設計のための一日の会話行動に関する基礎調査. 国立国語研究所論集 = NINJAL research papers, No. 10, pp. 85–106, jan 2016.

[2] 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.

[3] Sarik Ghazarian, Ralph Weischedel, Aram Galstyan, and Nanyun Peng. Predictive Engagement: An Efficient Metric for Automatic Evaluation of Open-Domain Dialogue Systems. In **AAAI Conference on Artificial Intelligence**, p. 8, 2020.

[4] 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, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics.

[5] Chin-Yew Lin. ROUGE: A Package for Automatic Evaluation of Summaries. In **Text Summarization Branches Out: Proceedings of the ACL-04 Workshop**, pp. 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[6] Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation. In **Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing**, pp. 2122–2132, Austin, Texas, November 2016. Association for Computational Linguistics.

[7] 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, Beijing, China, July 2015. Association for Computational Linguistics.

[8] Prakhar Gupta, Shikib Mehri, Tiancheng Zhao, Amy Pavel, Maxine Eskenazi, and Jeffrey Bigham. Investigating Evaluation of Open-Domain Dialogue Systems With Human Generated Multiple References. In **Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue**, pp. 379–391, Stockholm, Sweden, September 2019. Association for Computational Linguistics.

[9] 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, Online, July 2020. Association for Computational Linguistics.

[10] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed Representations of Words and Phrases and their Compositionality. In **Advances in Neural Information Processing Systems**, Vol. 26. Curran Associates, Inc., 2013.

[11] Yi-Ting Yeh, Maxine Eskenazi, and Shikib Mehri. A Comprehensive Assessment of Dialog Evaluation Metrics. In **The First Workshop on Evaluations and Assessments of Neural Conversation Systems**, pp. 15–33, Online, November 2021. Association for Computational Linguistics.

[12] 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, Vancouver, Canada, 2017. Association for Computational Linguistics.

[13] 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.

[14] 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.

[15] 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.

[16] 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, Online, July 2020. Association for Computational Linguistics.

[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 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.

[18] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In **International Conference for Learning Representations**, 2015.