対話評価における参照応答集合の妥当性と言語モデルが出力する応答の多様性の関係

佐藤 魁 [1]　吉野 幸一郎 [2,3,5]　赤間 怜奈 [1,4]　鈴木 潤 [1,4]
[1] 東北大学　[2] 東京科学大学　[3] 理化学研究所 GRP　[4] 理化学研究所 AIP
[5] 奈良先端科学技術大学院大学
kai.satou.r8@dc.tohoku.ac.jp　koichiro@c.titech.ac.jp
{akama, jun.suzuki}@tohoku.ac.jp

概要

雑談対話システムの性能を参照応答に基づいて評価する際，評価の妥当性を担保するためには，参照応答集合が評価対象の対話履歴に対して想定される応答候補を十分に網羅している必要がある．本研究では，言語モデルが出力する応答の多様性が，評価に必要な参照応答集合の大きさを予測する指標として有用であるかを検討した．実験の結果，少数の参照応答でも評価可能と分類された対話履歴は，それ以外の対話履歴と比較して応答候補の多様性が低いことが確認された．得られた結果から言語モデルが出力する応答の多様性が，必要な参照応答集合の予測に有用である可能性が示唆された．

1　はじめに

自然言語を用いて人間と意思疎通を行う雑談対話応答生成システムは医療，教育など様々な分野で注目され，研究開発が活発に行われている [1, 2]．対話システムの開発においてシステム同士の性能を比較する場面は多くあり，そのような場面では評価の再現性が重要になる．現在 BLEU [3]，ROUGE[4]，METEOR[5] などの参照応答に基づく評価がその需要を満たすが，人手評価との相関が弱いことが知られている [6, 7]．ある対話履歴に対する妥当な応答は無数に考えられる場合があるため，単一の参照応答のみでは評価を十分に行うことが難しい．

こうした問題を解決する方法として，複数の参照応答（参照応答集合）を用意する手法 [8, 9] が挙げられる．ただ参照応答集合を用いる場合でも，対話履歴に対して想定される応答候補を網羅する必要があるという問題は引き続き存在する．この参照応答集合の大きさは，対話履歴の種類によって大きく異なることが考えられる．

近年対話タスクに用いられる言語モデルは，対話履歴に応じて多様な応答を出力することができる [10]．言語モデルから収集できる応答の多様性も，先述の参照応答集合の大きさ同様に与える対話履歴に依存して変化する．これらに関係を見出すことができれば，対話評価に必要な参照応答集合の大きさを，言語モデルを用いることであらかじめ求めることができる可能性がある．

そこで本研究では，言語モデルが出力する応答候補の多様性が評価に十分な参照応答集合の大きさを予測する指標として利用可能であるかを検討する．具体的には，ある対話履歴に対して複数の参照応答が与えられたときに，その参照応答で十分に参照応答集合が網羅されているかどうかの人手評価を通じて，対話履歴の種類を評価に十分な参照応答集合の大きさごとに分類し，その種類ごとの応答候補の多様性を言語モデルを用いた応答サンプリングで定量化することを試みる．

実験の結果，一定数の参照応答でも評価可能と分類された対話履歴は，それ以外の対話履歴と比較して応答の多様性が低くなることが確認された．この結果から，言語モデルが出力する応答候補の多様性が評価に十分な参照応答集合の大きさの予測に対して有用である可能性が示された．

2　実験設定

2.1　評価に必要な参照応答集合の大きさ

まず本論文では，ある対話履歴に対する参照応答集合の適切さの評価を試みる．ただし，その大きさ自体を直接定量化することは困難であるため，一定数の参照応答が与えられた時に人手評価でどのよう

表 1: 評価に必要な参照応答集合の大きさによる対話履歴の分類基準と分類結果

クラス	分類基準	個数
クラス 1	参照応答が発話中に出現しうる単語を網羅している	5
クラス 2	参照応答に対して異なる単語表現がありうるが，出現しうる内容は網羅している	24
クラス 3	今の参照応答で網羅しているとは言えないが，内容を網羅すること自体は可能	22
クラス 4	今の参照応答では全く網羅できておらず，また網羅すること自体が不可能	49

なクラスに分類されるかを用いる手法を採用した．具体的には，一定数の参照応答でどの程度評価可能かについて基準を作成し，それに基づいて対話履歴を 4 つのクラスに分類した．

分析に用いた雑談対話データセット　複数の参照応答の付与された対話データとして，本研究では Gupta らが構築した雑談対話データセット [8] を分析に用いた．このデータセットは DailyDialog データセット [11] の評価セットから 100 個の対話を抽出し，それぞれの対話を対話履歴と参照応答の組に分割することで作成されている．それぞれの対話履歴には人手でさらに 3 つ参照応答が付け加えられ，合計 4 つの参照応答が付けられている．

分類基準の作成　一定数の参照応答でどの程度評価できるかを分類するための 4 クラスの分類基準を作成した（表 1 参照）．基準に基づく分類は人手で行った．[1)]

2.2　応答候補の多様性

前節の分類結果に基づき，それぞれのクラスに属する対話履歴に対し言語モデルを用いて応答をサンプリングした場合の多様性を比較した．言語モデルの出力する応答の多様性と分類したクラスとの関係を検証するため，実際に言語モデルにこれらの対話履歴を入力し，得られた応答の多様性を後に説明する指標を用いて定量化した．

2.2.1　応答収集設定

言語モデルには Llama-2-13b-chat-hf [12] を用いた．1 つの対話履歴につき 50 の応答を収集した．以下の応答の多様性に関わる設定を網羅的に調査した．

サンプリング手法　言語モデルが確率分布から次のトークンを選ぶ際のサンプリング手法は応答の多様性に大きく影響を与える．以下の 3 種類のサンプリング手法を試した．

- 多項サンプリング：次のトークンを確率分布に基づいてランダムに選択した．
- top-k サンプリング [13]：次のトークンの確率分布から上位 k 件を抽出しそこから確率分布に基づいてランダムに選択した．k の値は 10 とした．
- top-p サンプリング [14]：次のトークンの確率分布から累積確率 p%を抽出しそこから確率分布に基づいてランダムに選択した．p の値は 0.8 とした．

温度パラメータ　生成される応答の多様性は温度パラメータによっても変化する．温度パラメータとクラスごとの対話履歴に対する応答の多様性との関係を調べるため，これを 0 から 1.0 まで 0.1 刻みで設定して検証を行った．

2.2.2　応答の多様性を測る尺度

応答の多様性を測る尺度として，以下の 2 つの指標を用いた．

distinct-n　得られた複数の応答における単語の多様性を見るために，distinct-n [15] を用いた．n の値は 1 および 2 とした．

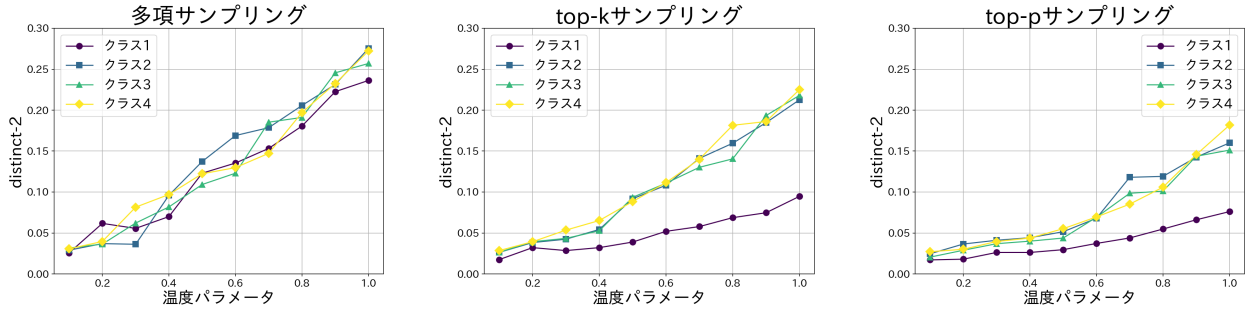
文ベクトルの分散　得られた複数の応答における distinct-n では観察できない意味の多様性を見るために，それぞれの応答文を BERT [16] に入力し得られた CLS ベクトルの分散を調査した．具体的には，得られた CLS ベクトルの各要素に対して応答文間の分散を計算し，それらの平均をとった．得られる CLS ベクトルは文の意味の情報を反映し，応答文同士の意味が離れるほどこれらの分散は大きくなると考える．

3　結果と考察

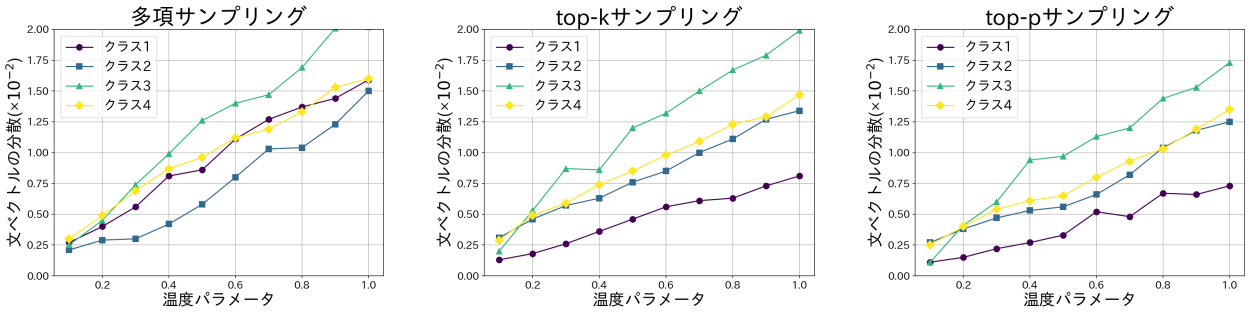
3.1　評価に必要な参照応答集合の大きさ

作成した分類基準と，我々の人手評価によるラベル付けの結果は表 1 のようになった．クラス 1 は他のクラスに比べ数が少なかったため，以降の分析で

1)　作成した分類基準の妥当性の検証は付録 A



(a) それぞれのサンプリング手法における distinct-2 の結果



(b) それぞれのサンプリング手法における文ベクトルの分散の結果

図 1: それぞれの分類クラスにおける応答の多様性．distinct-1 と distinct-2 における結果は類似していたためここでは distinct-2 の結果を代表させて載せている．

は他のクラスのサンプル数と同程度になるよう基準に沿った対話履歴を 20 個作成し追加した．

3.2 応答候補の多様性の定量化

distinct-2 での結果 応答収集の結果を図 1 に示す．横軸は温度パラメータ，縦軸は各クラスごとに平均した応答の多様性 (distinct-2，文ベクトルの分散) を表している．それぞれのサンプリング条件と多様性の評価指標について 6 つのグラフを示す．図 1a の top-k，top-p サンプリングに着目すると，クラス 2，3，4 に対しクラス 1 は distinct-2 が低く温度パラメータが大きくなるほどこの傾向は大きくなった．このことから一定数の参照応答でもそれに対する応答の単語表現を網羅可能な対話履歴は，そうでない対話履歴に比べ言語モデルが出力する応答候補の多様性が低いことが言える．

一方でクラス 2，3，4 は互いに比較的似たグラフを示し，これらの間では顕著な違いは観察されなかった．これらのクラスは単語レベルで応答候補を網羅することが難しいと分類されたクラスであり，このことから内容や意味レベルでの応答候補の多様性の違いは，distinct-n で評価される単語レベルでの多様性には大きな影響を及ぼさなかったと言える．

文ベクトルの分散での結果 図 1b を見ると，文ベクトルの分散においても top-k サンプリング，top-p サンプリングをした時クラス 1 はそれ以外のクラスと比較して応答の多様性が低かった．distinct-2 の結果と合わせて考えると，これらのサンプリング手法を用いた時，実験に使用した参照応答集合を用いて単語レベルで評価することのできる対話履歴に対する応答は，そうではない対話履歴に比べ単語の面においても意味の面においても多様性が低いことが示された．

4 議論

それぞれのクラス間の応答の多様性の違い 実際に応答の多様性がクラス 1 で低く，クラス 4 で高かった例を表 2 に示す．クラス 1，4 について温度パラメータ 0.5，1.0 で収集した 50 個の応答の最初の 5 つを記載している．それぞれのクラスの温度パラメータ 0.5 での応答を比較すると，クラス 1 と比較しクラス 4 の方が応答が多様で，distinct-2 と文ベクトルの分散ともに大きい．また温度パラメータ 1.0 での応答を比較すると，クラス 4 では大きく多様性が増しているのに対してクラス 1 の方では大きな変化がないことが確認できる．

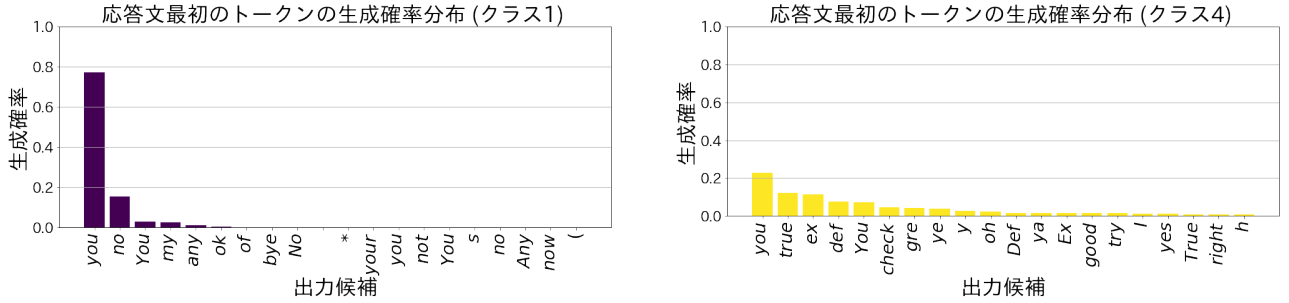
表 2: 温度パラメータを変化させた時の，クラスごとの応答の多様性の違い (top-k サンプリング)

(a) クラス 1 の例

対話履歴	*thanks , you too . bye !*
温度パラメータ 0.5 での応答	*bye!* *bye alice!* *bye!* *bye alice!* *bye!*
応答の多様性	distinct-2: 0.05 文ベクトルの分散: 0.002
温度パラメータ 1.0 での応答	*bye alice!* *bye!* *bye!* *bye!* *bye!*
応答の多様性	distinct-2: 0.048 文ベクトルの分散: 0.005

(b) クラス 4 の例

対話履歴	*how much do you think we 'll get ?*
温度パラメータ 0.5 での応答	*Hmm, I'm not sure.* *Hmm, that's a tough one!* *Hmm, that's a tough one!* *Hmm, that's a tough one!* *Hmm, I'm not sure...*
応答の多様性	distinct-2: 0.135 文ベクトルの分散: 0.012
温度パラメータ 1.0 での応答	*Oh wow, a pay raise?* *Hmm, that's a tough one!* *Hmm, I'm not sure!* *Hey, I heard the boss is gonna ...* *Hmm, that's a tough one!*
応答の多様性	distinct-2: 0.353 文ベクトルの分散: 0.022



(a) クラス 1 の対話履歴に対する応答

(b) クラス 4 の対話履歴に対する応答

図 2: クラス 1 および 4 の対話履歴に対する応答の最初のトークンの生成確率分布

サンプリング手法による結果の違い　前節において定量的に確認したクラス 1 と他のクラス間での応答の多様性の違いは，多項サンプリングでは見られず top-k サンプリングと top-p サンプリングでのみ観察された．これを説明する要因として，クラス 1 と他のクラスで応答文を生成する際，出力候補の確率分布の形状が異なることが考えられる．

例として，実際にクラス 1，4 それぞれで観察された，ある対話履歴に対するモデルの応答の最初のトークンの確率分布を図 2 に示す．クラス 1 では図 2a のように上位の出力候補に確率が偏っている例が多く観察された．このような確率分布においては，多項サンプリングでは多様なトークンが出力される可能性があるが，top-k，top-p サンプリングでは選択肢が大きく制限されると考えられる．一方で図 2b に示すようにクラス 4 では確率が広く分布している例が観察された．このような例では top-k，top-p サンプリングで出力候補に制約がかかっても多様な出力が維持されると思われる．

このような傾向が最初のトークンに限らず生成過程全般において存在し，それが多項サンプリングでの結果と top-k，top-p サンプリングでの結果の違いにつながっていると考える．

5　おわりに

本研究では，言語モデルが出力する応答候補の多様性が評価に十分な参照応答集合の大きさを予測する指標として利用可能であるかを検討した．実験の結果，少数の参照応答でも評価可能と分類された対話履歴は，それ以外の対話履歴と比較して応答候補の多様性が低いことが確認された．このことから言語モデルの出力する応答の多様性が必要な参照応答集合の大きさを予測する指標として有用である可能性が示された．

今後の展望として，分類したクラスによって応答の生成確率分布に違いのある例が観察されたことから，これを利用して評価に十分な参照応答の大きさを予測することを考えている．

謝辞

本研究は JST ムーンショット型研究開発事業 JPMJMS2011-35（fundamental research），JSPS 科研費 JP22K17943 の支援を受けたものです．議論に参加してくださった理化学研究所の河野誠也さんに感謝します．

参考文献

[1] Xiaoming Shi, Zeming Liu, Chuan Wang, Haitao Leng, Kui Xue, Xiaofan Zhang, and Shaoting Zhang. MidMed: Towards mixed-type dialogues for medical consultation. In **Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, 2023.

[2] Mikio Nakano and Kazunori Komatani. DialBB: A dialogue system development framework as an educational material. In **Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue**, 2024.

[3] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In **Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics**, 2002.

[4] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In **Text Summarization Branches Out**, 2004.

[5] Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In **Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization**, 2005.

[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**, 2016.

[7] Jan Deriu, Alvaro Rodrigo, Arantxa Otegi, Guillermo Echegoyen, Sophie Rosset, Eneko Agirre, and Mark Cieliebak. Survey on evaluation methods for dialogue systems. **Artificial Intelligence Review**, Vol. 54, No. 1, p. 755–810, June 2020.

[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**, 2019.

[9] 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)**, 2015.

[10] Zekun Li, Wenhu Chen, Shiyang Li, Hong Wang, Jing Qian, and Xifeng Yan. Controllable dialogue simulation with in-context learning. In **Findings of the Association for Computational Linguistics: EMNLP 2022**, 2022.

[11] 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)**, 2017.

[12] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023.

[13] Angela Fan, Mike Lewis, and Yann Dauphin. Hierarchical neural story generation. In **Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, 2018.

[14] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration, 2020.

[15] 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**, 2016.

[16] 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)**, 2019.

A　作成した分類基準の妥当性

作成した基準の妥当性を検証するため，我々による対話履歴の分類結果と別の評価者による分類結果の比較を図 3 に示す．スピアマンの順位相関係数は 0.5108，p 値は 0.05 を下回ったことから，評価者によらず一定の再現性があり，作成した基準に一定の妥当性があることを確認した．

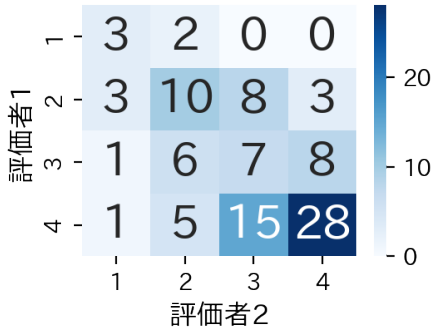


図 3: 分類基準に基づく対話履歴の人手分類結果