Follow-up 質問による矛盾応答収集の提案

佐藤志貴 [1] 赤間怜奈 [1,2] 鈴木潤 [1,2] 乾健太郎 [1,2]
[1] 東北大学 [2] 理化学研究所
{shiki.sato.d1,akama,jun.suzuki,kentaro.inui}@tohoku.ac.jp

概要

対話応答生成システムによる矛盾応答の抑制を考えるうえで，矛盾応答データの不足がボトルネックとなっている．データ不足の原因に，矛盾の発生が入力に依存した低頻度なものであるために効率的な収集が困難であることが挙げられる．本研究では，矛盾応答を誘発する発話として Follow-up 質問に注目し，自動収集した Follow-up 質問を用いた矛盾応答収集の効率化を試みる．提案法により矛盾応答を効率的に収集できることを確認したうえで，実際に矛盾応答を大規模収集する．収集した矛盾応答データの有用性を，矛盾検出器の精度改善を例に示す．

1 はじめに

近年のニューラル応答生成システム [1, 2, 3]（システムと呼ぶ）は流暢性は高いが，意味的に不適切な応答を生成する場合があることが広く知られている [1]．特に図 1 の r_2 のように過去の自身の発話と**矛盾**する応答（矛盾応答と呼ぶ）の生成は，一度でも発生すればシステムが内容を理解した対話をしていないという致命的な悪印象をユーザに与える深刻な問題である [4]．矛盾応答の抑制は，ユーザと信頼関係を構築し共生していく対話システムの実現を目指すうえで中心的な課題の一つといえる．

先行研究で一定の効果が報告されている矛盾応答の抑制手法として，矛盾検出器による矛盾応答候補の除去 [5, 4]，対照学習 [6, 7] がある．これらの手法では，品質の高い矛盾応答データをいかに多く獲得できるかが性能の良し悪しを決める大きな要因となっている．しかし，現状は品質の高い矛盾応答データの効率的な収集方法が確立されているとは言い難い．特に，矛盾の発生が直前の相手発話（入力と呼ぶ）に依存することが示唆されている [4, 8] ものの，どのような入力に対し矛盾が発生する傾向にあるか十分研究されていない．実際に，これまでの矛盾応答データの収集は，人手により作成した矛盾応答に限られている [4]．人手による矛盾応答は，実際のシステムの矛盾応答とは性質が異なる可能性が高く，大規模な作成が難しいなどの弱点もある．

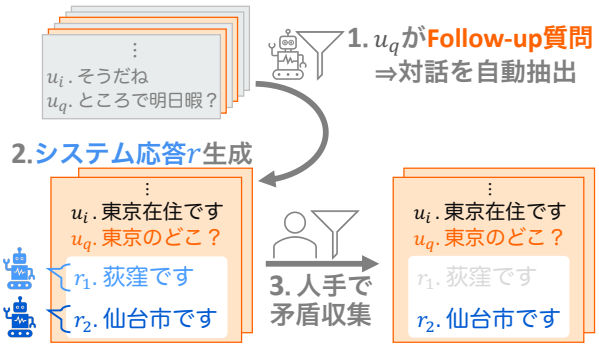

図 1 FQ を用いた矛盾応答収集方法の概要．

本研究では，システムがどのような入力に対し矛盾応答を生成しやすいかを考慮したシステムの矛盾応答の効率的な収集を考える．具体的な取り組みとして，(1) 先行研究と本研究の分析に基づき，過去の発話に含まれる情報に関連する質問，**Follow-up 質問**（**Follow-up Question** (**FQ**) と呼ぶ）が矛盾応答を誘発しやすいという仮説を立てたうえで，FQ の自動収集方法を提案する．(2) 自動収集した FQ が矛盾を誘発する傾向にあることを確認した後，(3) FQ を用いて図 1 の方法によりシステムの矛盾応答を大規模収集する．[1)] (4) さらに，構築したデータセットの有用性を，矛盾検出器の精度改善を例に示す．

2 なぜ FQ に注目するか

本研究では，矛盾を誘発する発話として Follow-up 質問に注目する．FQ は「対話相手が過去の発話で提示した任意の情報 i に関連する質問」とする．例えば，図 1 の橙色の質問は直前の相手発話の情報に対する FQ である．FQ に注目する理由は二つある．

第一に，システムの矛盾応答は対話中の既出情報を再度言及する際に生じやすいことが知られている [4, 8]．こうした状況は，既出情報に関連する質問

1) 構築したデータセットは公開予定である．

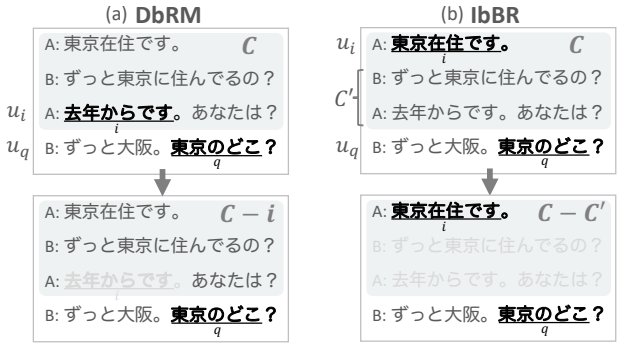

図 2 DbRM と IbBR での C に対する処理．A，B は話者．

である FQ への応答時に生じやすいと考えられる．

第二に，Nie ら [4] が小規模に収集したシステムの矛盾応答 382 個から無作為抽出した 50 個を分析したところ，42% の矛盾応答が相手の FQ に対する応答だった．矛盾応答を誘発した相手発話の種類の分布は FQ のみに偏ったロングテールとなったため，FQ は矛盾を引き出す代表的な発話といえる．

システムは FQ のみに対し矛盾応答を生成するわけではないため，FQ への注目で全ての矛盾を考慮できるわけではない．しかし，代表的な入力に注目し知見を蓄えることは第一歩として重要と考える．

3 FQ の自動収集

本研究では，FQ に対するシステム応答から矛盾応答を効率的に収集することを考える．先行研究 [9] では 教師あり学習により FQ を検出する方法が提案されたが，教師あり学習では検出可能な FQ が学習データに依存するという問題がある．本研究では，学習を要さない FQ の自動収集方法を考える．

3.1 アイデア：応答生成システムの利用

FQ は過去の発話の情報 i に関する質問であり，i に対し質問 q が FQ か判定するには，i と q の関連性を適切に捉える必要がある．ここで，近年の大規模ニューラル応答生成システムは入力との関連性が高い応答を生成可能であり [2]，発話間の関連性を適切に捉えることができると考えられる．以上から，大規模ニューラル応答生成システムにより i と q の関連性が強いと推測された場合，q が i に対する FQ である可能性は高いと考えられる．

3.2 FQ 収集のための指標

ニューラル応答生成システムを用いて q と i の関連性を自動評価する指標を提案し，その値が高い q を FQ として収集する．q と i の関連性を評価する指標として，異なる考え方に基づき，**D**ecrease in probability **b**y **ReM**oving information (**DbRM**) と **I**ncrease in probability **b**y **BR**inging information closer (**IbBR**) の 2 種類を提案する．以下，u_i のうち情報は疑問文以外に含まれると考え，i は u_i の疑問文以外の文，q は u_q の疑問文とする．また，隣接発話間距離を 1 とした u_i と u_q の発話間距離を n とし，n 発話前の i に対する FQ を FQ$_{-n}$，n 発話前の i に対する q の DbRM と IbBR をそれぞれ DbRM$_{-n}$，IbBR$_{-n}$ とする．

DbRM． q が i に関連する場合，図 2 (a) のように u_q 以前の対話の全発話 $C \ni u_i$ から i を除去[2] した $C - i$ に対する u_q の生成確率は，$P(u_q|C)$ より低いと考えられる．そのため，ニューラル応答生成システムで $P(u_q|C)$ と $P(u_q|C-i)$ を計算したとき，次の値は高くなると考えられる:

$$\mathrm{DbRM}(i, q, C) = \log \frac{P(u_q|C)}{P(u_q|C-i)} \tag{1}$$

IbBR． q が i に関連する場合，図 2 (b) のように u_q と u_i が挟む発話系列 C' を除去した $C - C'$ に対する u_q の生成確率は，$P(u_q|C)$ より高いと考えられる．そのため，次の値は高くなると考えられる:

$$\mathrm{IbBR}(i, q, C) = \log \frac{P(u_q|C-C')}{P(u_q|C)} \tag{2}$$

DbRM と IbBR による FQ 収集． DbRM と IbBR は算出においてそれぞれ制限がある．DbRM について，$|C|$ をニューラル応答生成システムに入力可能なトークン数に抑えるために C の冒頭部分を省略するとき，i も省略されるほど n が大きいと，$C - i = C$ となり値が計算できない．IbBR について，$C' \neq \emptyset$ を前提とした指標であるため $n = 1$ の場合は計算できない．そこで本研究では，FQ$_{-1}$ は DbRM が，FQ$_{-n}(n > 1)$ は IbBR が高い質問を収集する．

3.3 DbRM と IbBR の算出結果の定性分析

表 1 に，4 節の検証にて前述の方法により DbRM が計算された質問の例を示す．[3] DbRM および IbBR の上位下位の例を観察したところ，同表のように指標の値が高い / 低い例では u_i 中の情報との関連性が高い / 低い質問が u_q に含まれる傾向を確認した．以上から，前述の方法により FQ を一定の精度で自動収集できると考えられる．ここで，本研究の目的は

2) 疑問文以外を除去し，空なら u_i を “__SILENCE__” に置換．
3) 両指標の分布を付録 A に，IbBR の例を付録 B に示す．

表 1 DbRM$_{-1}$ 計算例．I の u_q 太字部は u_i 太字部に関係．

I. DbRM$_{-1}$ が高い質問の例 (DbRM$_{-1}$ = 53.6)

話者	発話
A:	What do you attribute your longevity to?
B:	Well, I think it has to do in part with not smoking. Also, I think the red wine has provided me with lasting health benefits.
A:	Doctors do say that red wine is great for health. Wish I could get into red wine more, I do prefer the white. Had a great **gewurztraminer** this month. ..... u_i
B:	**What in tarnation is a gewurztraminer?** I've never heard of that wine. ..... u_q

II. DbRM$_{-1}$ が低い質問の例 (DbRM$_{-1}$ = −32.4)

話者	発話
A:	I think Japanese food is not for me! I prefer French food. I had some very good French food in Paris once.
B:	I heard that there are other foods in japan. just like here in the states we have mexican, italian, jamaican food, that they have mexican, italian etc. food there...is that true?
A:	yes, that's true, but the quality of the food can vary! japan specializes in sushi ..... u_i
B:	so what part of japan were you in where you got your fried octopus and okonomiyaki? ..... u_q

矛盾応答の効率的な収集であり，収集した質問がシステムの矛盾を誘発するかが重要となる．そのため，FQ 自動収集自体の精度算出は行わない．

4 矛盾応答収集方法の効果検証

自動収集した FQ に対し，実際にシステムが矛盾応答を生成しやすい傾向にあるかを検証する．具体的には，まず DbRM や IbBR が高い $(i, q, C \ni u_i)$ の三つ組（入力対話と呼ぶ）を収集する．次に，入力対話に対するシステムの応答を収集する．最後に，システム応答が入力対話との矛盾を含むかを確認していき，矛盾応答が生成される頻度を求める．無作為抽出した入力対話についても同様の頻度を求め比較することで，DbRM や IbBR が高い入力対話において矛盾応答の発生頻度が高くなるかを確認する．

4.1 検証における設定

入力対話の収集． 対話コーパスから最終発話が質問（q とする）を含むような n 発話以上の発話系列を切り出し作成した集合（入力対話集合と呼ぶ）のうち，DbRM か IbBR が高いものを収集する．$n = 1$，$n > 1$ 両方の場合の検証をするため，DbRM$_{-1}$，IbBR$_{-3}$ それぞれ上位 100 個の入力対話を収集する．また，入力対話集合から無作為に 100 個取り出したものを比較対象とする．入力対話集合は，Multi-Session Chat [10]（MSC と呼ぶ）から作成する．FQ$_{-1}$，FQ$_{-3}$ 収集のための入力対話集合の対話数はぞれぞれ 59,234，44,041 となった．DbRM と IbBR は，高性能であり応答生成システムの性能評価 [11] にも使われる Blender-1B [1] により算出する．

システム応答の収集． u_q に対する応答生成には，近年の高性能な大規模ニューラル応答生成システムである DialoGPT-Medium [2] (DG)，PLATO-2 [12] (P2)，Blender-3B (B3)，PLATO-XL [13] (PX) を用いる．各入力対話に対し，4 個の応答生成用システムにより 2 個ずつ[4] 計 8 個の応答を生成させる．

矛盾頻度の計算． 各入力対話に対する 8 個の応答が入力対話と矛盾するかを人手で確認する．システム応答 1 個あたり 5 人のクラウドワーカー[5] が矛盾の有無を 5 段階評価し，3 人以上が矛盾を含むと判断した場合[6] 矛盾応答とする．8 個のシステム応答のうち 1 個以上が矛盾応答と判定された場合，その入力対話は矛盾を誘発したとみなす．

表 2 入力対話と矛盾応答の生成頻度の関係．

入力対話	FQ$_{-1}$ 矛盾頻度	FQ$_{-3}$ 矛盾頻度
無作為抽出	29 / 100	26 / 100
FQ 自動収集	**46** / 100	**38** / 100

4.2 検証結果

表 2 に，収集した入力対話のうち，矛盾応答を誘発したものの数を示す．同表より，FQ$_{-1}$，FQ$_{-3}$ いずれについても，無作為抽出した入力対話に比べ，DbRM や IbBR が高い入力対話に対して応答を生成させたときに矛盾応答の生成頻度が高くなった．入力対話を無作為抽出した場合も一定の割合の質問が矛盾応答を誘発した理由として，入力対話の収集元の MSC では，一つの話題に対し深堀りした議論をする傾向にあるため，コーパス中の質問に占める FQ の割合が比較的高いことが考えられる．

5 データセットの構築

前節より，DbRM や IbBR が高い入力対話に対しシステムが高頻度で矛盾応答を生成することを確認した．これら入力対話に対しシステムに生成させた

4) p = 0.5 の top-p sampling を 100 回実施し，尤度上位 2 候補を抽出する．

5) https://www.mturk.com/

6) 評価基準は 1 が「明らかに矛盾」，5 が「明らかに無矛盾」で，1 と 2 は矛盾，4 と 5 は無矛盾とした．コストを考慮して，u_i 以降の発話と応答の矛盾についてのみ評価する．

表 3　構築したデータセットの統計情報．

	矛盾応答数	無矛盾応答数
FQ_{-1} に対する応答	1,620	1,085
FQ_{-3} に対する応答	897	1,325

応答から，人手評価により矛盾応答を検出することで，システムの矛盾応答を大規模収集する．

5.1　収集設定

入力対話集合の作成元となる対話コーパスは 4 節と同様に MSC を用いる．収集に使う入力対話は FQ_{-1} および FQ_{-3} とし，入力対話集合から $DbRM_{-1}$，$IbBR_{-3}$ が高い上位 500 個ずつの入力対話をデータセット構築に用いる．DbRM，IbBR の計算とシステム応答の収集は 4 節と同じ設定を用いる．[7] 各システム応答に対し 3 人のクラウドワーカーによって 4 節と同様の人手評価を行い，1 人以上が矛盾，かつ 1 人以下が無矛盾と判断した応答を矛盾応答として収集する．また，3 人のクラウドワーカー全員が無矛盾と評価した応答は無矛盾応答として収集する．

5.2　収集結果

表 3 に，構築したデータセットの統計情報を示す．実際にシステムが生成した矛盾応答を収集したものとしては最大規模のデータセットとなる．付録 C にデータセットの例を示す．

6　実験：データセットの活用例

矛盾応答の抑制における本データセットの有用性を示す．例として，本データセットを用いた追加学習による矛盾検出器の精度の改善を試みる．

6.1　実験設定

学習する矛盾検出器．　Nie ら [4] が構築した矛盾検出器を，本データセットにより追加学習する．[8] Nie らは既存の対話コーパス中の対話に続く矛盾応答を人間に書かせ収集したうえで，与えられた 2 つの発話同士が矛盾するかの 2 値分類を行うよう RoBERTa [14] をファインチューニングした．実際にシステムが生成する矛盾応答を検出する際には，本データセットのようにシステムの矛盾応答も学習に用いることで性能が向上すると考えられる．

7）　ただし，効率的に矛盾応答を収集するため，各システム 100 個の応答候補を Nie ら [4] の矛盾検出器によって予測される矛盾確率でリランキングし上位 2 個を取り出す．

8）　学習の詳細は付録 D に示す．

表 4　ターゲットシステムごとの矛盾応答検出の正解率．括弧内の数字は評価におけるシステム応答の数．

矛盾検出器	DG (186)	P2 (238)	B3 (156)	PX (194)
追加学習前	54.8	61.3	67.9	67.0
追加学習後	**64.0**	**70.2**	**74.4**	**68.0**

データセット分割．　本データセットを入力対話によって学習データセット，テストデータセットに 8 : 2 で分割する．本データセットには各入力対話に対し最大 4 種類のシステムの応答が存在する．そこで，4 個のうち 3 個のシステムの学習データセット中の応答を学習データとしたとき，残りの 1 個のシステム（ターゲットシステムと呼ぶ）のテストデータセット中の各応答[9] が矛盾を含むかどうかの 2 値分類を矛盾検出器に行わせる．ターゲットシステムを変えて，これを 4 回実施する．そのため，矛盾分類器は未知の入力対話に対する未知のシステムの応答に含まれる矛盾を検出する必要がある．

6.2　実験結果

表 4 に，矛盾検出器による 2 値分類の正解率を示す．追加学習前の矛盾検出器について，Nie ら [4] は人間が作成した矛盾応答を 2 値分類させた際の正解率が 93.2% と報告したが，同表より実際にシステムが生成した矛盾応答については正解率が 60% に満たない場合があることがわかる．一方，本研究で構築したデータセットを用いた追加学習によって，未知の入力対話や未知のシステムの応答についても 2 値分類の正解率が向上した．

7　おわりに

本研究では，矛盾応答生成の抑制において必要な矛盾応答の効率的な収集に取り組んだ．特に，FQ がシステムの矛盾を引き出す代表的な発話であるという分析から，自動収集した FQ を用いた矛盾応答の収集を提案した．提案法で効率的に矛盾応答を収集できることを確認したうえで，FQ に対するシステムの矛盾応答を大規模収集し，矛盾検出器の性能向上を例にデータセットの有用性を示した．

今後の課題として，データセットのさらなる大規模化が挙げられる．また，本研究の分析により FQ が矛盾応答を引き出す代表的な発話であることがわかったものの，FQ 以外の多様な発話に対するシステムの矛盾応答の収集にも取り組む予定である．

9）　テストデータセット中のターゲットシステムの矛盾，無矛盾応答数が一致するようアンダーサンプリングする．

謝辞

本研究は，JSPS 科研費 JP22K17943, JP21J22383, JST ムーンショット型研究開発事業 JPMJMS2011 (fundamental research) の助成を受けて実施されたものです．また，本研究の遂行にあたり多大なご助言，ご協力を賜りました Tohoku NLP グループの皆様に感謝申し上げます．

参考文献

[1] 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. In **Proceedings of the 16th conference of the european chapter of the association for computational linguistics: Main volume (EACL)**, pp. 300–325, 2021.

[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 (ACL): System demonstrations**, pp. 270–278, 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. In **arXiv preprint arXiv:2001.09977**, 2020.

[4] Yixin Nie, Mary Williamson, Mohit Bansal, Douwe Kiela, and Jason Weston. I like fish, especially dolphins: Addressing Contradictions in Dialogue Modeling. In **Proceedings of the 59th annual meeting of the association for computational linguistics (ACL)**, pp. 1699–1713, 2020.

[5] Sean Welleck, Jason Weston, Arthur Szlam, and Kyunghyun Cho. Dialogue natural language inference. In **Proceedings of the 57th annual meeting of the association for computational linguistics (ACL)**, pp. 3731–3741, 2019.

[6] Margaret Li, Stephen Roller, Ilia Kulikov, Sean Welleck, Y-Lan Boureau, Kyunghyun Cho, and Jason Weston. Don't say that! Making inconsistent dialogue unlikely with unlikelihood training. In **Proceedings of the 58th annual meeting of the association for computational linguistics (ACL)**, pp. 4715–4728, 2020.

[7] Weizhao Li, Junsheng Kong, Ben Liao, and Yi Cai. Mitigating Contradictions in Dialogue Based on Contrastive Learning. In **Findings of the Association for Computational Linguistics: ACL 2022**, pp. 2781–2788, 2022.

[8] Zekang Li, Jinchao Zhang, Zhengcong Fei, Yang Feng, and Jie Zhou. Addressing Inquiries about History: An Efficient and Practical Framework for Evaluating Open-domain Chatbot Consistency. In **Findings of the joint conference of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (ACL-IJCNLP)**, pp. 1057–1067, 2021.

[9] Souvik Kundu, Qian Lin, and Hwee Tou Ng. Learning to Identify Follow-Up Questions in Conversational Question Answering. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 959–968, 2020.

[10] 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, 2022.

[11] Maxime De Bruyn, Ehsan Lotfi, Jeska Buhmann, and Walter Daelemans. Open-Domain Dialog Evaluation Using Follow-Ups Likelihood. In **Proceedings of the 29th International Conference on Computational Linguistics**, pp. 496–504, 2022.

[12] Siqi Bao, Huang He, Fan Wang, Hua Wu, Haifeng Wang, Wenquan Wu, Zhen Guo, Zhibin Liu, and Xinchao Xu. PLATO-2: Towards Building an Open-Domain Chatbot via Curriculum Learning. In **Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021**, pp. 2513–2525, 2021.

[13] Siqi Bao, Huang He, Fan Wang, Hua Wu, Haifeng Wang, Wenquan Wu, Zhihua Wu, Zhen Guo, Hua Lu, Xinxian Huang, Xin Tian, Xinchao Xu, Yingzhan Lin, and Zheng-Yu Niu. PLATO-XL: Exploring the Large-scale Pre-training of Dialogue Generation. In **Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022**, pp. 107–118, 2022.

[14] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach. In **arXiv preprint arXiv:1907.11692**, 2019.

[15] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-Art Natural Language Processing. In **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations**, pp. 38–45, 2020.

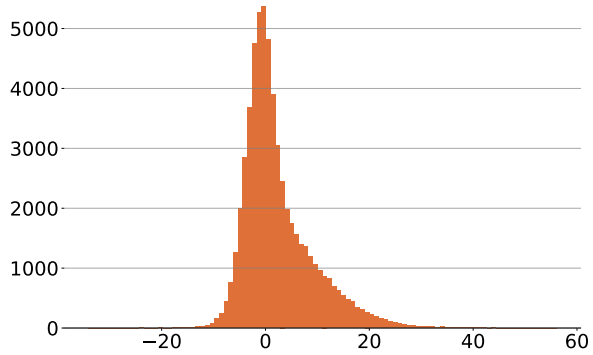

図 3　MSC における DbRM_{-1} の分布．

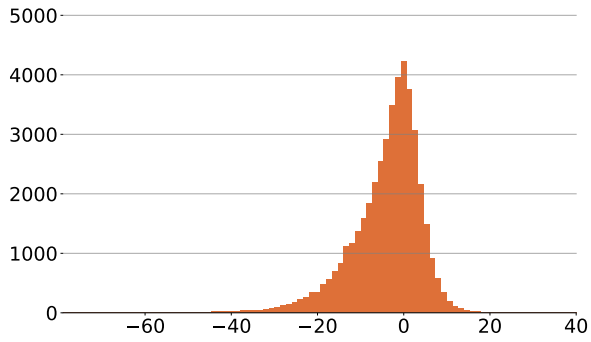

図 4　MSC における IbBR_{-3} の分布．

表 5　IbBR_{-3} が高い質問の例 (IbBR_{-3} = 43.2)．u_i 太字部に関する質問が u_q 太字部でされている．

話者	発話
A:	I meant to tell you that I like cycling too. Where did you end up meeting Sufjan Stevens anyways?
B:	I think it was New York City. Where do you like to go cycling at?
A:	That's super cool. **I'm kind of jealous! You met my favorite artist in the greatest city in the world!** Usually I ride around town and on some mountain bike trails. Sometimes, I'll ride to work when it's nice out. ..... u_i
B:	Mountain bike trails sound fun and challenging! I think I should look on my map and see if I can find any. Would you recommend any? Do you have a favorite?
A:	Well, when I visit my friend in Seattle, Washington, I love riding the trails at Mount Rainier National Park and Mount Baker Snoqualmie National Forest. I know you like to stay indoors reading, and you're busy with your tech job, but you should give mountain biking a go!
B:	I'll have to make a playlist for biking haha. I do enjoy reading indoors, but that's only when it's raining. I prefer to read under a tree or maybe a hammock. **Do you plan on going to one of Sufjan Stevens' concerts?** ..... u_q

表 6　データセット中の FQ_{-1}，FQ_{-3} およびそれに対するシステム矛盾応答の例．u_i 太字部への u_q 太字部の質問に対し，システムが応答 r 太字部にて矛盾を生成している．

I. FQ_{-1} に対する矛盾応答の例．

話者	発話
A:	**My allergies have been acting up like crazy the past couple days. I think there's a cat nearby.** ..... u_i
B:	**How do you know its a cat and not a different animal you are allergic to?**..... u_q
A:	I'm not sure. **I've never had allergies before.** ..... r

II. FQ_{-3} に対する矛盾応答の例．

話者	発話
A:	**I read an article yesterday on the history of the Premier League that was very interesting.** Did you know that it has been around for almost 30 years? ..... u_i
B:	I did not. I had no idea it had such a history. I guess if you think about it, 30 years ago is only the 1990's now!
A:	Yeah, that's hard to believe! I wonder what the oldest soccer team is.
B:	I would have to guess it would be one in Europe, but I'm curious as well. **Did the history show get you interested in watching some games?** ..... u_q
A:	Not really. I just like watching soccer. I like the game. **I don't care about the history.** ..... r

A　DbRM と IbBR の分布

4 節にて算出した，MSC の入力対話集合に含まれる u_q の DbRM_{-1} および IbBR_{-3} の分布を図 3 および図 4 にそれぞれ示す．

B　自動収集された FQ の例

表 5 に，4 節の検証にて収集された，IbBR_{-3} が高い質問の例を示す．

C　データセットの例

表 6 に，本研究で構築したデータセットの入力対話およびシステムが生成した矛盾応答の例を示す．

D　矛盾検出器の学習詳細

6 節の矛盾検出器の学習において，エポック数は 10，学習率は 1.0×10^{-5}，バッチサイズは 64，重み減衰係数は 0.01，他の値は Hugging Face[15] で設定されている初期値を用いた．