雑談応答生成モデルによる矛盾応答の大規模収集

佐藤志貴 [1] 赤間怜奈 [1,2] 鈴木潤 [1,2] 乾健太郎 [3,1,2]
[1] 東北大学 [2] 理化学研究所 [3] MBZUAI
{shiki.sato.d1,akama,jun.suzuki}@tohoku.ac.jp kentaro.inui@mbzuai.ac.ae

概要

雑談応答生成モデルが生成した矛盾応答データの不足は，矛盾応答の抑制に向けた取り組みの障害となっている．本研究では，多様なモデルの矛盾応答からなる大規模データセットを構築する．様々な観点からデータセットを分析し，モデルによる矛盾応答の特徴を明らかにするとともに，構築したデータセットを学習データとして用いることでデータ駆動型の矛盾抑制の有効性が向上することを確認する．

1 はじめに

雑談応答生成モデル (Response Generation Model, RGM) による誤りのなかでも，表 1 のように過去の自身の発話と矛盾する応答，いわゆる矛盾応答は，対話の円滑な進行やユーザとの信頼関係の構築を妨げる [1] ため，ユーザと信頼関係を構築可能なシステムを実現する上で解決すべき重要な課題である．

矛盾応答の抑制において，大規模かつ高品質な RGM の矛盾応答データの存在は 2 つの点で重要となる．まず，大規模矛盾応答データを分析することで RGM の矛盾応答の特徴が明らかになれば，より効果的な矛盾抑制手法を開発できる可能性がある．また，大規模矛盾応答データを学習に用いることで，近年主流となっているデータ駆動型の矛盾抑制 [1, 2, 3, 4] の性能が向上すると考えられる．しかし，Nie らが収集した DECODE データセット [1] など，人手で書かれた矛盾応答のデータセットは存在する一方，RGM が生成した矛盾応答を大規模収集した研究はこれまで存在しなかった．人手で書かれた矛盾データは，3.1 節で示すように，RGM による矛盾とは異なる特徴を有しており，分析や学習に用いるデータとしての有効性に限界がある．

そこで我々は，実際に RGM が生成した矛盾応答の大規模収集に取り組んできた [5]．しかし，これまでのデータセットに収録した矛盾応答の事例数は 2.5k と，データ駆動型システムの安定した学習に必要な事例数の目安の一つ，10k [6] を大きく下回り，学習データとしての利用に際してデータ駆動型の手法の効果を十分引き出せない可能性があった．

表 1 収集した RGM の矛盾，無矛盾応答例．A，B は話者，CG と B1 は RGM を指す．発話 u_q に対する RGM 応答の，先行発話 u_r との一貫性に関する 3 人のワーカの判断は C，N で表され，それぞれ矛盾，無矛盾を指す．発話 u_r は u_q に d_{u_r} 発話先行している．太字は矛盾箇所を示す．

対話文脈
A: **I hurt my toe doing ballet.** — u_r
B: Oh I hope you get better. Does is hurt a lot?
A: It hurts pretty bad, but it will heal. [. . .]
B: [. . .] Do you do ballet practice often? — u_q
話者 A としての u_q に対する RGM 応答
CG: **I don't do ballet myself,** I was just watching a performance. [. . .] — [C, C, C]
⋮
BL: Yes, I do ballet every day. [. . .] — [N, N, N]

本研究では，多様な RGM の矛盾応答 10503 個を収録した大規模データセットを構築する．さらに，データセットを分析し RGM の矛盾応答の特徴を明らかにする．最後に，本データセットで学習した矛盾検出器が高い精度で RGM の矛盾を検出できることを確認し，データ駆動型の矛盾抑制における有用性を示す．本データセットは公開予定である．

2 大規模データセットの構築

我々の過去の応答収集 [5] と類似した方法で，RGM の矛盾応答および無矛盾応答を大規模に収集した．構築の詳細は付録 A に示す．

2.1 各事例の構成

各事例は，表 1 のように対話文脈とそれに対する RGM の応答，応答の一貫性を示すラベルからなる．

対話文脈. 過去の我々の調査 [5] により，先行する発話 (u_r とする) に関連するような質問である Follow-up 質問 (FQ) が RGM から矛盾応答を誘発す

表 2 構築したデータセットの基本統計情報．括弧内の値は，対話文脈の種類数を表す．

d_{u_r}	矛盾応答数	無矛盾応答数
1	8108 (2703)	12471 (2920)
3	2175 (739)	4378 (953)
5	220 (74)	422 (94)
Total	10503 (3516)	17271 (3967)

る主要な発話であることが判明している．そこで今回のデータセット構築でも，過去の収集と同様，先行発話 u_r に対する FQ を含む発話 u_q が末尾に来る対話文脈を Multi-session Chat (MSC) データセット [7] から収集した．なお，u_r と u_q は d_{u_r} 発話離れていることとする．表 1 は $d_{u_r} = 3$ の例となる．

RGM の応答． 各対話文脈に対し，既に社会で広く活用されている ChatGPT (CG，https://openai.com/chatgpt) に加え，高性能な RGM である Plato-2 (P2)，Plato-XL (PX) [8, 9]，Blenderbot1-3B (B1) [10]，Blenderbot2-3B (B2) [7, 11]，Blenderbot3-3B (B3) [12]，Blenderbot3-30B (BL)，Opt-66B (O6) [13] に応答を生成させることで，多様な RGM の応答を収集した．

一貫性に対するラベル． 各応答が先行発話 u_r と矛盾するかを，3 人ずつのクラウドワーカに評価させ，2 人以上が矛盾と判断した場合は矛盾応答，全員が無矛盾と判断した場合は無矛盾応答とした．

2.2 統計情報

表 2 にデータセットの統計情報を示す．事前調査で d_{u_r} が 3 より大きいと矛盾発生頻度が低くなると判明しているため，d_{u_r} の最大値を 5 と設定した．

3 分析データとしての有用性の検証

RGM による矛盾応答の特徴を分析することは，より有効性の高い矛盾抑制手法を開発する上で不可欠である．本研究で大規模データセットを構築したことにより，こうした分析が可能となる．3.1 節では RGM 矛盾応答自体の特徴に関する分析結果を，3.2 節では RGM が特に矛盾応答を生成する傾向にある対話文脈の特徴に関する分析結果を報告する．

3.1 矛盾応答自体の分析

我々の分析により，RGM に特有の矛盾応答として，発話内矛盾に起因する矛盾応答と，曖昧性のある表現を含む矛盾応答が特定された．

発話内矛盾に起因する矛盾応答． RGM によって生成された矛盾応答を観察したところ “I like tennis, but I dislike tennis.” のように，応答内に相反する情報を組み込むことで，そのうちのどちらかの情報が対話文脈と矛盾するという事例が散見された．具体的には，本データセットの各 RGM の矛盾応答を無作為に 50 個ずつ取り出し確認したところ，ChatGPT を除く 7 個の RGM で発話内矛盾に起因する矛盾応答が 1 個以上観測された．一方，人手で書かれた矛盾応答を DECODE データセットから 50 個無作為抽出し調べたが，発話内矛盾に由来する矛盾応答は確認されなかった．以上から，発話内矛盾に起因する矛盾は RGM 応答で特に頻出するといえる．

表 3 曖昧性のある表現を含む Plato-2 の矛盾応答の例．太字の “interview” が今日のものを指していると解釈するかによって一貫性に関する判断が異なると考えられる．

対話文脈
A: I had a promising interview today!
B: Oh excellent! How did it go, what made it so excellent?
話者 A としての RGM 応答
P2: i think i did well because they called me back to set up **an interview**.

曖昧性のある表現を含む矛盾応答． RGM の応答と，DECODE データセット中の人手応答の間には，矛盾の有無に対する人手評価の傾向において差異があることが確認された．本研究と Nie らのデータセット構築 [1] では，矛盾ラベルを付与するワーカの選定方法やラベルの付与方法は同様である．しかし，3 人中 1 人以上のワーカが矛盾と判断した事例からなる部分集合の中で，他の 2 人も矛盾であると同意した割合に有意な差が観察された．この割合は，人手応答では 78.4%，RGM の応答では 30.4% だった．この異なりは，RGM が一貫性に関して曖昧な表現を含む応答を生成する傾向があることに起因している可能性がある．表 3.1 に曖昧な表現を含む応答の例を示す．曖昧性を含む応答に対する矛盾の有無の判断は，個々のワーカが一貫性に関してどのように解釈するかによって異なる．こうした応答は，たとえ一部のワーカには矛盾を感じさせないとしても，ひとたび実際のユーザに矛盾と認識されてしまうと，対話の質を著しく損なう可能性があり，これらの矛盾に対処することも極めて重要である．

3.2 矛盾応答を誘発する対話文脈の分析

多様な RGM から矛盾応答を誘発する対話文脈の特徴の特定は，より効果的な矛盾抑制のための技術

表 4 相手発話中の情報 (太字部分) を自身の情報として応答に取り入れたことによる Plato-2 の矛盾応答の例.

対話文脈
A: Have you taken any new pictures?
B: I managed to get out at the weekend and get loads of shots in the snow we had. [. . .]
A: Oh wow you had snow!? **We just had rain all weekend :)** [. . .] Did you have a nice chilled weekend? [. . .]
話者 B としての RGM 応答
P2: it was a good weekend here, **we got to enjoy the cold rain**!

の開発において有用である可能性がある．我々のデータセットでは，各対話文脈 8 個もの多様な RGM の応答が収集されており，この調査に適している．

解釈性が高い分析観点として対話行為ラベルと使用語彙に着目しつつ，統計的手法[1)]を用いて対話文脈を分析したところ，RGM の矛盾応答を誘発しやすい対話文脈の特徴が 3 つ特定された．

対話行為分類に基づく特徴． データセット中の各対話文脈の発話 u_q に対話行為ラベルを付与したところ，[2)]“Declarative Yes-No-Questions” および “Statement-non-opinion” が u_q に付与された対話文脈はより多くの RGM から矛盾応答を誘発する傾向にあることが確認された．前者のラベルについて，該当する 193 個の対話文脈において矛盾応答を生成した RGM の個数の平均が 2.77 だった一方，該当しない 4084 個の対話文脈における平均は 2.41 だった．この差は，過去の発話で自身が提供した情報の繰り返しを相手から求められていることを理解した上で一貫した情報を繰り返す能力が RGM に欠けていることが原因で発生している可能性がある．後者のラベルについて，該当する 2118 個の対話文脈における平均は 2.49 だった一方，該当しない 2159 個の対話文脈における平均は 2.36 だった．この差の原因として，表 4 の例のように，RGM が対話文脈中の自身の発話と相手発話の区別に失敗し，相手発話 u_q 中の情報を自身に関するものとして認識し応答することで矛盾する場合があることが考えられる．

1) 本データセットには，対話文脈ごとに 8 個の RGM による計 8 個の応答が収録されており，各対話文脈について 8 個のうち矛盾応答を生成した RGM の個数が算出できる．データセット中の対話文脈をある特徴の有無により 2 つの集合に分け，集合ごとに矛盾応答を生成した RGM の個数の平均を求めた．有意水準 1%の片側 t 検定に基づき 2 つの集合間で矛盾応答を生成した RGM の個数の平均が有意に異なる場合，その特徴を，矛盾応答の生成を誘発する特徴とみなした．

2) RoBERTa を Switchboard コーパス [14] で追加学習し対話行為分類器を構築した．同コーパスにおける評価データセットでの正解率は 80.4% だった．

語彙に基づく特徴． また，疑問詞 “how” を u_q に含む対話文脈も矛盾応答を誘発しやすいことを確認した．“how” を u_q に含む対話文脈 3513 個において矛盾応答を生成した RGM の個数の平均は 2.60 だった一方，“how” を含まない場合の平均は 2.39 にとどまった．このことから，方法や度合いに関する質問に対し一貫性を保ちながら説明することは現在の RGM でも比較的難しいことがわかった．

4 学習データとしての有用性の検証

RGM の矛盾応答からなる大規模データセットを学習データとして用いることで，データ駆動型の矛盾抑制の有効性が向上することを示す．一例として，対話文脈中の先行発話 u_r と RGM 応答の間の一貫性を自動判定する矛盾検出器の性能向上を試みる．矛盾検出器は，矛盾抑制のための後処理 [1, 2] や RGM の矛盾生成頻度の自動評価に用いられ，特に重要な矛盾抑制のための手法の一つである．

既存の検出器は，実際に RGM が生成する矛盾応答の代わりに，自動合成された矛盾応答や人手で書かれた矛盾応答を学習資源として用いることによって開発されてきた [1, 4]．我々は，実際に RGM が生成した矛盾データを学習に利用することで，有効性の高い検出器を構築できることを確認する．特に本実験では，我々のデータセットで学習した検出器の性能を，Nie らが人手矛盾データを用いて構築した既存の最高性能の検出器 [1] と同様の方法で構築した検出器の性能と比較する．

4.1 実験設定

矛盾検出器． Nie らは，人手で書かれた矛盾応答からなる DECODE データセットで RoBERTa [15] を追加学習し矛盾検出器を構築した．本実験では，Nie らの設定に従い，我々のデータセットで RoBERTa を追加学習し検出器 $\mathrm{CD}_{\mathrm{RGM}}$ を構築した．比較対象として，$\mathrm{CD}_{\mathrm{RGM}}$ の学習に用いた事例数と同数の事例を DECODE データセットから無作為抽出したものを学習データセットとする $\mathrm{CD}_{\mathrm{HUM}}$ を用意した．

評価データセット． 本研究で収集した RGM の矛盾応答は，MSC という特定の対話コーパスの FQ に対して生成されたものに限られている．しかし，矛盾検出器を実際に矛盾抑制に用いる場合，異なるドメインの対話における FQ 以外に対する RGM の応答の矛盾を検出できることも重要となる．そこで，異なる対話コーパスの，FQ に限定しない対話文

表 5 矛盾検出器の正解率. Human-Bot は B1 の応答を含むため CD_{RGM} の Human-Bot の値は B1 が評価対象の時のもの.

		Topical-test / Daily-test						
Detector	Human-Bot	P2	PX	B1	B2	B3	BL	O6
CD_{HUM}	.749	.55/.52	.58/.60	.61/.55	.60/.59	.68/.61	.67/.55	.59/.53
CD_{RGM}	**.787**	**.77/.77**	**.72/.67**	**.74/.68**	**.70/.72**	**.73/.76**	**.82/.64**	**.81/.75**

脈に対する RGM の応答からなる評価データセットを 2 種類構築した．一方は Topical-Chat データセット [16]，他方は DailyDialog データセット [17] に収録されている対話文脈からなり，前者を Topical-test データセット，後者を Daily-test データセットと呼ぶ．各評価データセットは 7 個のサブセットで構成されており，各サブセットは P2，PX，B1，B2，B3，BL，O6 のいずれかの RGM の矛盾応答と無矛盾応答 50 個ずつからなる．[3] 対話文脈は，**FQ に限らない**質問発話が末尾に来るものをコーパスから無作為抽出し用意した．[4] 付録 B に構築の詳細を示す．また，これらのデータセットに加え，Nie ら [1] により構築された Human-Bot データセットも評価データセットとして用いた．Human-Bot データセットは，人間と RGM の間の対話において発生した RGM による矛盾応答と無矛盾応答 382 個ずつからなる．

CD_{RGM} の学習． 本研究で構築した大規模データセットを用いて CD_{RGM} を学習した．同データセットの応答を生成した 8 個の RGM は，Topical-test，Daily-test データセットの応答を生成した 7 個の RGM と重複する．未知の RGM 応答に対する矛盾検出精度を評価するため，Topical-test，Daily-test データセットの応答を生成した 7 個の RGM のうち 1 個の RGM を評価対象 RGM とし，Topical-test，Daily-test データセットのうち評価対象 RGM の応答からなるサブセットを評価データセット，我々の大規模データセットから評価対象 RGM の応答を除いたものを学習データセットとした．評価対象 RGM を変えながら CD_{RGM} の学習と評価を 7 回実施した．B2 を評価対象とした際，学習データセットの事例数は矛盾応答と無矛盾応答それぞれ 8023 個で最小となったため，他の RGM を評価対象とした場合もこの事例数に揃えた．付録 C に学習の詳細を示す．

3) 大規模矛盾応答収集において CG の矛盾応答の生成頻度は比較的低かったことから，本研究ではコストを鑑みて CG の応答からなる評価データセットについては構築しなかった．

4) 非質問発話が末尾となる対話文脈は，話題転換を促す応答など，対話文脈と無関係な応答も妥当な応答とする可能性があるため，矛盾応答が発生する頻度はより低くなると考えられる．本研究では，コストを鑑みて質問以外の発話に対する RGM の応答からなる評価データは収集しなかった．

CD_{HUM} の学習． DECODE データセットから矛盾応答と無矛盾応答を 8023 個ずつ無作為抽出し学習に用いた．他の設定は CD_{RGM} と同様である．

4.2 実験結果

表 5 に，RGM 応答の文脈中発話 u_r との一貫性を 2 値分類した時の正解率を示す．CD_{HUM} について，Human-Bot を除く全評価データセットにおける正解率が 0.7 を下回った．特に P2 が評価対象だった時の Daily-test での正解率は 0.52 であり，チャンスレートと同程度となった．以上から，RGM 矛盾応答の検出における CD_{HUM} の精度には実用上問題があることがわかった．一方，我々のデータセットで学習した CD_{RGM} は，学習事例数が CD_{HUM} と同じであるにも関わらず全評価データセットで CD_{HUM} の正解率を上回った．以上から CD_{RGM} は，未知のドメインの FQ とは限らない質問に対する，未知の RGM の応答の矛盾検出に有効であることがわかった．更に，対話文脈の種類を一切制限しない Human-Bot データセットでも CD_{RGM} の検出精度は高く，任意の対話文脈に対する矛盾応答も検出できると考えられる．なお，CD_{HUM} の学習に DECODE データセット中の全事例 (矛盾応答無矛盾応答それぞれ 15605 個) を用いた場合でも，Human-Bot を除く全評価データセットで CD_{RGM} が CD_{HUM} の正解率を上回った．

5 おわりに

矛盾応答の特徴に関する分析やデータ駆動型の手法の学習に利用可能な RGM の矛盾応答データはこれまで大規模収集されておらず，矛盾応答の抑制に向けた取り組みの障害となっていた．

本研究では，多様な RGM による矛盾応答からなる大規模データセットを構築した．構築したデータセットの分析を通し，RGM の矛盾応答や RGM の矛盾応答を誘発する対話文脈の特徴など，より有効性の高い矛盾応答の抑制手法の開発に有用な知見が得られた．さらに，矛盾検出器の学習を例として，本データセットがデータ駆動型の矛盾抑制のための学習データとして有効であることがわかった．

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 JPMJMS2011-35 (fundamental research) の助成を受けて実施されたものです．研究遂行にあたり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] Yixin Nie, Mary Williamson, Mohit Bansal, Douwe Kiela, and Jason Weston. I like fish, especially dolphins: Addressing Contradictions in Dialogue Modeling. In **Proc. of ACL2021**, pp. 1699–1713, 2021.

[2] Sean Welleck, Jason Weston, Arthur Szlam, and Kyunghyun Cho. Dialogue natural language inference. In **Proc. of ACL2019**, pp. 3731–3741, 2019.

[3] 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 **Proc. of ACL2020**, pp. 4715–4728, 2020.

[4] Weizhao Li, Junsheng Kong, Ben Liao, and Yi Cai. Mitigating Contradictions in Dialogue Based on Contrastive Learning. In **Findings of ACL2022**, pp. 2781–2788, 2022.

[5] 志貴佐藤, 怜奈赤間, 潤鈴木, 健太郎乾. Follow-up 質問による矛盾応答収集の提案. 言語処理学会第 29 回年次大会発表論文集, pp. 387–392, 2023.

[6] Tianyi Zhang, Felix Wu, Arzoo Katiyar, Kilian Q. Weinberger, and Yoav Artzi. Revisiting Few-sample BERT Fine-tuning. In **Proc. of ICLR2021**, 2021.

[7] Jing Xu, Arthur Szlam, and Jason Weston. Beyond Goldfish Memory: Long-Term Open-Domain Conversation. In **Proc. of ACL2022**, pp. 5180–5197, 2022.

[8] 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 ACL-IJCNLP2021**, pp. 2513–2525, 2021.

[9] 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 AACL-IJCNLP2022**, pp. 107–118, 2022.

[10] 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 **Proc. of EACL2021**, pp. 300–325, 2021.

[11] Mojtaba Komeili, Kurt Shuster, and Jason Weston. Internet-Augmented Dialogue Generation. In **Proc. of ACL2022**, pp. 8460–8478, 2022.

[12] Kurt Shuster, Jing Xu, Mojtaba Komeili, Da Ju, Eric Michael Smith, Stephen Roller, Megan Ung, Moya Chen, Kushal Arora, Joshua Lane, Morteza Behrooz, William Ngan, Spencer Poff, Naman Goyal, Arthur Szlam, Y.-Lan Boureau, Melanie Kambadur, and Jason Weston. BlenderBot 3: a deployed conversational agent that continually learns to responsibly engage. In **arXiv preprint arXiv:2208.03188**, 2022.

[13] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. OPT: Open Pre-trained Transformer Language Models. In **arXiv preprint arXiv:2205.01068**, 2022.

[14] Dan Jurafsky, Liz Shriberg, and Debra Biasca. Switchboard SWBD-DAMSL shallow-discourse-function annotation coders manual. **Institute of Cognitive Science Technical Report**, 1997.

[15] 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.

[16] Karthik Gopalakrishnan, Behnam Hedayatnia, Qinlang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, and Dilek Hakkani-Tür. Topical-chat: Towards knowledge-grounded open-domain conversations. In **Proc. of INTERSPEECH2019**, pp. 1891–1895, 2019.

[17] Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset. In Greg Kondrak and Taro Watanabe, editors, **Proc. of IJCNLP2017**, pp. 986–995, 2017.

[18] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The Curious Case of Neural Text Degeneration. In **Proc. of ICLR2020**, 2020.

[19] Alexander H. Miller, Will Feng, Adam Fisch, Jiasen Lu, Dhruv Batra, Antoine Bordes, Devi Parikh, and Jason Weston. ParlAI: A dialog research software platform. In **Proc. of EMNLP2017: System demonstrations**, pp. 79–84, 2017.

[20] 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 **Proc. of EMNLP2020: System Demonstrations**, pp. 38–45, 2020.

A　大規模データセット構築の詳細

A.1　構築方法

最終発話が FQ を含む対話文脈を用意し，それらに対する応答を RGM に生成させる．生成された各 RGM 応答について，人手で矛盾の有無を判定することで矛盾応答と無矛盾応答を収集する．

FQ 収集のアイデア． 質問文 q を含む発話 u_q を末尾とする対話文脈を C，$u_q \in C$ より d_{u_r} 個前の非質問文 r を含む発話を $u_r \in C$ とする．質問文 q が r を参照する質問である場合，q を r に対する FQ とみなす．質問文 q が r の FQ であるかを判定するには，q が r に関連する文であるかを確認する必要がある．近年の RGM は対話文脈に関連する応答を生成できることを踏まえ，本研究では RGM を用いて q と r の関連性を自動評価することを考える．

FQ 収集の方法． 発話 u_r と u_q の間にある発話系列を C_{mid}，C から u_q および C_{mid} を除いた発話系列を $C_{\text{w/o}\{u_q,C_{\text{mid}}\}}$ とする．同様に，u_q と r を除いた C を $C_{\text{w/o}\{u_q,r\}}$ とする．文 q が r に対する FQ である場合，u_q の元の条件付き確率に比べ，$P(u_q|C_{\text{w/o}\{u_q,C_{\text{mid}}\}})$ は低くならず，$P(u_q|C_{\text{w/o}\{u_q,r\}})$ は低くなると考えられるため，$\text{FQness} = P(u_q|C_{\text{w/o}\{u_q,C_{\text{mid}}\}})/P(u_q|C_{\text{w/o}\{u_q,r\}})$ は高くなると考えられる．Blenderbot1-3B を用いて各確率を求めることで FQness を計算し，対話文脈集合から FQness が最も高くなる対話文脈を取り出す．

A.2　構築設定

対話文脈の用意． 対話コーパスから，$d_{u_r}+1$ 個以上の発話からなり最終発話が質問文を含む発話系列を取り出すことで，対話文脈集合を作成した．MSC データセットを収集元とし，最終的に 59k 個の発話系列からなる対話文脈集合を得た．ここから，$d_{u_r}=1$，3，5 と設定したときに FQness が最も高い対話文脈を 3250，1000，100 個ずつ取り出した．

RGM 応答の収集． 各対話文脈に対し，8 個の RGM から応答を一つずつ集め計 8 個の応答を得た．このとき，矛盾応答を効率的に集めるため，対話文脈それぞれに対して各 RGM に 100 個の応答候補を生成させ，Nie らが構築した矛盾検出器により u_r と矛盾する可能性が最も高いと予想された候補をその RGM の最終的な応答とした．応答候補の生成には，top-p サンプリング [18] を用いた．サンプリングにおける p の値を，ParlAI [19] などの 0.9 と比べて大幅に小さい 0.5 に設定し，生成確率の低い候補の生成を回避した．ChatGPT の応答生成には OpenAI の API，Plato-2 と Plato-XL の応答生成には Knover, [5] 他の RGM の応答生成には ParlAI を用いた．

B　評価データセット構築の詳細

Topical-Chat, DailyDialog データセットから，$d_{u_r}+1$ 個以上の発話を有し最終発話が質問文を含むような発話系列を無作為抽出し，評価データセット構築のための対話文脈とした．Topical-test について，$d_{u_r}=1$ の評価セット構築のために 300 個，$d_{u_r}=3$ の評価セット構築のために 100 個対話文脈を収集した．Daily-test について，$d_{u_r}=1$ の評価セット構築のために 200 個，$d_{u_r}=3$ の評価セット構築のために 100 個対話文脈を収集した．FQness を用いた対話文脈の抽出を行わない点と ChatGPT の応答を収集しない点を除き，その他の手順および設定は大規模データセット構築と同様である．

C　矛盾検出器学習の詳細

学習データ． 負例は，我々のデータセットに収録されている RGM の矛盾応答とその先行発話 u_r の組とした．一方正例は，データセット中の無矛盾応答および無作為に選択された同話者の対話文脈中の先行発話の組とした．これは，u_r と矛盾しない応答であれば他の発話とも矛盾しない可能性が高いと考えられること，また u_r と RGM 応答の組と比べ関連性が低いと考えられる発話の組を正例とすることで，関連性の低い発話の組は正例であることを検出器が学習できると考えられるためである．

ハイパーパラメータ． 一部を除き，Hugging Face [20] の初期値を用いて検出器を構築した. [6]

各学習率におけるモデル学習時，パラメータ更新 200 回ごとに検証データセットにおける正解率を計算し，一回前に計算した正解率を下回った段階で，一回前に正解率を計算した際のパラメータをその学習率におけるモデルの最終パラメータとした．学習率ごとにモデルを学習していき，検証データセットにおける正解率が最も高くなった学習率の最終パラメータを評価に用いた．なお，検証データセットは学習データセットから 10% のデータを無作為に取り出したサブセットであり，モデルパラメータの学習には使わなかった．

5）www.github.com/PaddlePaddle/Knover.

6）train_batch_size: 128, learning_rate: {1e-6, 5e-6, 1e-5, 5e-5}, weight_decay: 0.01, and eval_steps: 200.