対話システムの矛盾応答の生成に対する脆弱性の分析

佐藤 志貴 [1]　赤間 怜奈 [1,2]　大内 啓樹 [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,reina.a,jun.suzuki,inui}@ecei.tohoku.ac.jp　hiroki.ouchi@riken.jp

1　はじめに

人間と自然言語を用いて会話をおこなう対話システムは，深層学習技術の発展を背景に研究が急速に発展することとなった．特に近年のニューラルネットワークを用いたシステムは，文脈の話題に沿った多様な応答を生成できることが知られている [1, 2, 3]．しかし，これらのシステムであっても，文脈との意味的な適切さを考慮した応答の生成が可能であるとは言えない．たとえば，Roller らの構築した大規模システム Blender[1] は，人手評価により従来のシステムを上回る性能を有することが示されたものの，低頻度ながら，対話上にすでに出現している情報の問い直しとなる応答や，過去の発話と矛盾する応答を生成する場合があることが報告されている．こうした意味的に不適切な応答のなかでも，過去の発話と**矛盾する応答**（矛盾応答と呼ぶ）は，会話相手となるユーザに対話の破綻を感じさせることがわかっており [4]，先行研究でも分析や改良が取り組まれている重要な問題の一つである [5, 6, 7, 8]．

このような矛盾応答が生成される背景として，本研究では，システムが応答生成時に対話文脈との一貫性に関して感度の低い生成確率を算出する，すなわち，**算出される生成確率が矛盾の有無に対し鈍感であること**が原因の一つとなっている可能性を考える．実際に，矛盾の観点から，Blender が一つの対話文脈に対し探索幅 100 のビーム探索法で生成した応答を観察すると，図 1 のように，矛盾応答と矛盾を含まない応答（無矛盾応答と呼ぶ）が探索でのスコア上位 100 個の応答のなかに混在していることがわかる．これは，システムが矛盾応答に対して低い生成確率を付与できていないために，探索時に矛盾応答が候補に残ってしまうためと考えられる．

本論文では，近年のシステムが実際にユーザとの間で起こりうる会話のなかでも矛盾応答を生成することを示し（2 節），そのうえで，これらシステムが対話文脈との一貫性に対して感度の低い生成確率付与をおこなうことが矛盾応答の生成の原因となりうる可能性があることを実証する（3 節）．

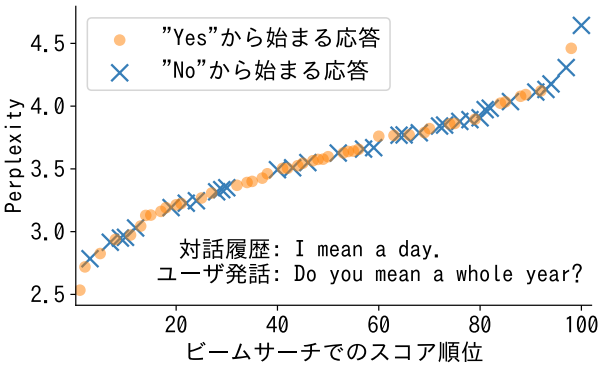

図 1　Blender[1] の探索幅 100 のビーム探索による応答生成結果．システムに無矛盾応答が"No."となる対話文脈を与え，探索での上位 100 個の応答を取り出したうえで，各応答を"Yes"から始まる応答と"No"から始まる応答に分類した結果，矛盾応答と無矛盾応答の混在が確認された．

2　システムの矛盾に対する脆弱性

Roller らは，Blender が人間との会話では矛盾応答を高頻度で生成しない理由について，システムの学習データにあたる人間同士の会話中に矛盾が含まれにくいことから，典型的な対話の流れのなかで学習データをなぞった応答をするときには矛盾が表出しにくいためであると考察している [1]．こうした考察をふまえると，近年のシステムであっても，学習データ中に含まれる対話の流れから外れた文脈のもとでは，矛盾応答を生成するという脆弱性を抱えている可能性がある．本節の実験では，システムのこうした脆弱性について検証するために，矛盾応答が生成されうる状況下においてはシステムが矛盾応答を生成することを確認した．

2.1　実験方法

矛盾応答が生成されうる対話文脈を与えたときのシステム応答を人手評価することで，システムが矛

表 1　自然言語推論データを用いた対話文脈作成の例

自然言語推論データ			作成される対話文脈	
前提文:	yeah i'm in North Carolina		対話履歴:	Yeah I'm in North Carolina
仮説文:	I'm in North Carolina.	⟶	ユーザ発話:	Are you in North Carolina?
ラベル:	含意		応答:	肯定文であれば適切
前提文:	yeah i'm in North Carolina		対話履歴:	Yeah I'm in North Carolina
仮説文:	I'm in South Carolina.	⟶	ユーザ発話:	Are you in South Carolina?
ラベル:	矛盾		応答:	否定文であれば適切

盾応答を生成する頻度を調べる．矛盾応答が生成されうる対話文脈については，Saleh ら [7] の方法を拡張し，自然言語推論タスクのデータセットを擬似的な対話の形式に変換することで用意する．具体的には，自然言語推論タスクの前提文をシステム側の一つ目の発話（対話履歴），それに対する仮説文を対話履歴に対するユーザ側の応答（ユーザ発話）として，システムにユーザ発話への応答を生成させるような対話文脈を作成する．このとき，自然言語推論データの仮説文を一般疑問文に変換することで，仮説文が含意ラベルを持つ場合であれば否定を返すことで矛盾する疑問文，矛盾ラベルを持つ場合であれば肯定を返すことで矛盾する疑問文となる．表 1 に，自然言語推論データの例と，そこから仮説文を変換することで作成した対話文脈の例を示す．このように，自然言語推論のデータを変換することで，会話として不自然ではないかたちでシステムが矛盾応答を生成しうる対話文脈が作成できる．これを入力として生成される応答を評価することで，システムの矛盾に関する脆弱性を検証する．

2.2　実験設定

検証するシステムとその設定　本実験では，近年構築された高性能な対話システムのうち，モデルパラメータが公開されている Zhang ら [2] の DialoGPT と Blender の 2 種類について検証をおこなった．各システムについてはパラメータ数が異なる複数のモデルパラメータが公開されているため，合計 5 つのシステムについて検証した[1)]．応答生成時は，すべてのシステムについて，ビーム幅を 10 とするビーム探索をおこなった．

検証時の対話文脈　対話文脈の作成元とする自然言語推論タスクデータとして，幅広いドメインを取り扱う大規模かつ高品質なデータとして知られている Multi-Genre Natural Language Inference コーパス [9] を用いた．このなかでも，オープンドメインでの会話データを扱っている"TELEPHONE"ドメインのデータを用いた．仮説文が含意ラベル，矛盾ラベルを持つ自然言語推論データからそれぞれ 75 個ずつ，合計 150 個の対話文脈を作成した．

1）　DialoGPT はパラメータ数 345M，762M の 2 種類，Blender については 400M，1B，3B の 3 種類を検証した．

表 2　150 個の対話文脈に対し矛盾応答を生成した頻度．✔ は適切な応答，✘ は誤りを含む応答を表す．

システム	✔	✘	✘ に占める矛盾応答 (割合)
DialoGPT 345M	64	86	44 /86 (51.2%)
DialoGPT 762M	79	71	46 /71 (64.8%)
Blender 400M	58	92	51 /92 (55.4%)
Blender 1B	54	96	46 /96 (47.9%)
Blender 3B	54	96	52 /96 (54.2%)

2.3　実験結果

表 2 に，各システムが与えられた 150 個の対話文脈に対して適切な応答と，誤りを含む応答，そのうち矛盾が誤りの原因である応答を生成した頻度をそれぞれ示す．同表から，検証した全システムが高頻度で矛盾を含む応答を生成したことがわかる．このことから，人間との典型的な会話では矛盾応答を生成することが少ない近年の高性能なシステムであっても，対話文脈によっては矛盾応答を生成するという脆弱性を抱えていることが確認された．

3　応答生成確率の矛盾に対する感度

2 節の結果から，近年のシステムであっても文脈によって矛盾応答を生成することがわかった．本節の実験では，システムが文脈との一貫性に関し感度の低い生成確率を算出することが，矛盾応答を生成する原因となりうる可能性があることを確認した．

3.1　実験方法

図 1 では，システムが矛盾応答と無矛盾応答に対し同程度の生成確率を割り当てることで，応答探索のスコア上位にこれらの応答が混在してしまう例を示した．このような場合，ビーム探索法における探

索幅の変更など，探索の方法の変化だけで，システムの 1-best が矛盾応答と無矛盾応答のどちらとなるか（極性と呼ぶ）が反転する可能性がある．逆に言えば，探索方法の変化による極性の反転の発生は，システムが対話文脈との一貫性に関して感度の低い生成確率を算出することを意味する．そこで本実験では，ビーム探索法の探索幅の変化に伴うシステム応答の極性の反転を観察することにより，現状のシステムが文脈との一貫性に関する感度の低い生成確率を算出することを示す．

3.2 実験設定

検証するシステムとその設定 本実験では，2 節の実験で用いた DialoGPT，Blender の合計 5 つのシステムについて，ビーム探索法の探索幅を，$\{1, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50\}$ の 11 種類に変化させたときの応答の変化を分析した．

検証時の対話文脈 分析の容易さの観点から，システムに与える対話文脈は，肯定と否定のどちらか一方が矛盾応答，他方が無矛盾応答となる 2 節と同様のものを用いた．これによって，応答が肯定と否定どちらに当たるかが分かればその極性が判定可能となる．仮説文が含意ラベル，矛盾ラベルを持つ自然言語推論データからそれぞれ 2,000 個ずつ，合計 4,000 個の対話文脈を作成した．

生成応答の肯定・否定の判定 大量のシステム応答について肯定・否定を人手で判定するにはコストがかかるため，自動での判定をおこなった．判定の方法や精度については付録 A で詳述する．

極性反転の判定 本実験では，システムは一つの対話文脈に対し，ビーム幅を変えながら複数の応答を生成する．対話文脈 c に対しビーム幅 i, j 間で極性反転が起きたときに 1 となる $\mathrm{inv}(c, i, j)$ を以下のように定義する．

$$\mathrm{inv}(c, i, j) = \begin{cases} 1, & \text{if } \mathrm{pn}(c, i) \cdot \mathrm{pn}(c, j) = 1 \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

ここで，$\mathrm{pn}(c, i)$ は対話文脈 c に対するビーム幅 i の 1-best 応答が肯定であれば 1，否定であれば −1，不明であれば 0 をとる．また，対話文脈集合 $\mathcal{C}$ に対し，ビーム幅 i, j $(i < j)$ 間で極性反転が生じる頻度は次式のようになる．

$$N_{\mathrm{inv}}(i, j) = \sum_{c \in \mathcal{C}} \mathrm{inv}(c, i, j) \quad (2)$$

表 3 ビーム幅変化で極性反転が生じた対話文脈の数

システム	極性反転が生じた対話文脈の数 (割合)
DialoGPT 345M	554 /4,000 (13.9%)
DialoGPT 762M	607 /4,000 (15.2%)
Blender 400M	812 /4,000 (20.3%)
Blender 1B	672 /4,000 (16.8%)
Blender 3B	570 /4,000 (14.3%)

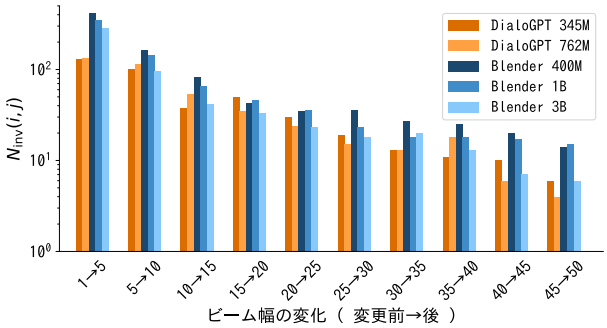

図 2 各システムの極性反転のタイミングの分布

3.3 実験結果

極性反転の頻度 ビーム幅の変化により極性反転が生じる頻度を確認する．本実験では，異なるビーム幅のもと各対話文脈 c に対し 11 個の応答を各システムが生成する．11 個のビーム幅のなかに $\mathrm{inv}(c, i, j) = 1$ を満たす i, j が存在する対話文脈を数え，極性反転が生じた対話文脈の数を求めた．結果を表 3 に示す．同表より，各システムとも 1 割を超える対話文脈で極性の反転が起きたことがわかる．極性反転により対話文脈との一貫性に関わるかたちで応答内容が正反対となることや，肯定・否定の自動判定ができない応答は $\mathrm{pn}(c, i) = 0$ となることを考慮すると，この割合は高いと考えられる．このことから，検証したシステムは，生成確率の算出における対話文脈との一貫性に関する感度が低いために，極性の反転を頻繁に発生させることがわかった．

極性反転が生じるタイミング 表 3 で示した極性の反転がどのビーム幅の変化のタイミングで生じたのかを確認するために，$N_{\mathrm{inv}}(i, j)$ をグラフ化したものを図 2 に示す．同図より，各システムともビーム幅が小さいときに極性反転が頻繁に生じている一方で，ビーム幅が 10 以上の値からさらに大きくなる際にも極性の反転が起きていることがわかる．もとのビーム幅が十分大きい場合，更にビーム幅を大きくしたときに探索にもたらす影響は小さいことが考えられる．そのため，こうした例では，矛盾応答と無矛盾応答の生成確率が特に近い，すなわち対話文

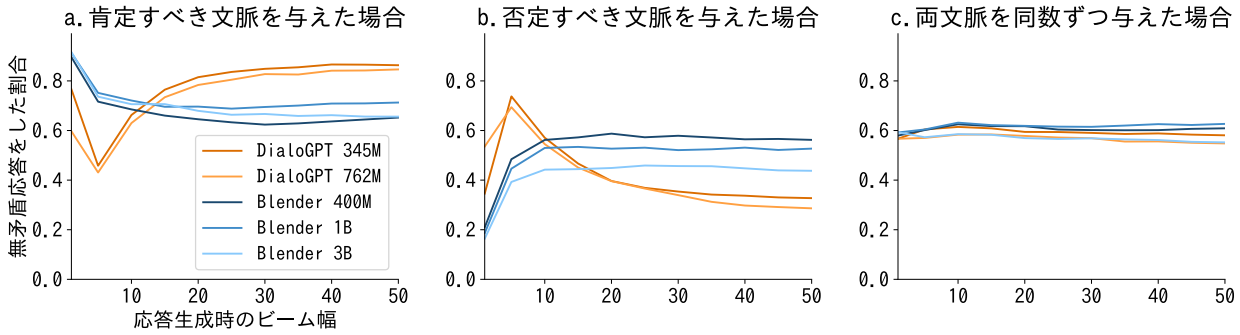

図 3　ビーム幅と対話文脈に対し正しい極性の応答を生成できた割合の関係

脈との一貫性に関する感度が特に低く，わずかな探索の変化で極性が反転すると考えられる．

ビーム幅と極性の関係　極性反転が生じることで，実際に出力されるシステム応答がどのような影響を受けるかを分析する．図 3-a および図 3-b に，それぞれ含意，矛盾ラベルを有する自然言語推論データから作成した対話文脈 2,000 個ずつに対し，無矛盾応答を生成した割合[2)]とビーム幅の関係を示す．また，図 3-c に実験で用いた全ての対話文脈 4,000 個に対し，無矛盾応答を生成した割合を示す．図 3-a および図 3-b では，各システムともビーム幅の変化に伴い無矛盾応答を生成する割合が変化している．このことから，システムが矛盾応答を生成する対話文脈の傾向がビーム幅の影響を受けていることがわかる．この結果は，特定のビーム幅のもとで生成された応答を分析するだけでは，システムの矛盾応答の生成傾向を正確に分析できない可能性があることを示唆すると考えられる．また，図 3-a と図 3-b の関係に注目したとき，図 3-a で無矛盾応答，すなわち肯定にあたる応答を生成する割合が高い場合において図 3-b でも矛盾応答，すなわち肯定にあたる応答を生成する割合が高くなっており，その逆も成立している．このことから，システムが肯定，否定を生成する割合自体もビーム幅により変化していることがわかる．これによって，ビーム幅の変化に伴って肯定，否定が生成される割合が図 3-a と図 3-b の場合で同様に変化し，一方で無矛盾応答の割合が，他方で矛盾応答の割合が高くなる．その結果，肯定，否定を応答として返すべき対話文脈が同数ずつ含まれる図 3-c では，無矛盾応答を生成する割合がビーム幅の変化に対し頑健であるように見えてしまう．この結果は，実験に用いる対話文脈集合によってはシステム応答の矛盾に関する傾向を正確に分析できない可能性があることを示唆していると考えられる．

2)　図 3-a においては肯定応答を生成した割合を，図 3-b においては否定応答を生成した割合を示す．ラベルが付与できない応答は割合計算時の分母にカウントしない．

4　まとめ

本論文では，対話システムにおける重要な課題の一つである矛盾応答に着目し分析をおこなった．我々の分析で得られた知見を改めて以下に示す．

- 人間との対話による評価で矛盾応答が確認されない対話システムであっても対話文脈によっては矛盾応答を生成する脆弱性を抱えている．
- 現状のシステムは同じ文脈に対しても探索のゆれで応答の極性が変化するなど，対話文脈との一貫性に対し感度の低い生成確率を算出しており，矛盾応答の原因となりうる可能性がある．

また 3 節の実験では，システムの矛盾応答について分析するうえで注意すべき以下の示唆が得られた．

- システムが矛盾応答を生成する対話文脈の傾向が応答の探索方法によって変化しうるため，特定の探索方法のもと生成された応答を分析するだけではシステムの矛盾応答の生成に関する振る舞いを正確に分析できない可能性がある．
- 肯定，否定の割合などシステム応答自体の傾向も探索方法の影響を受けるため，入力に用いる対話文脈の傾向によっては正確なシステム分析ができない可能性がある．

本研究では対話文脈の発話数やユーザ発話の発話行為を限定したうえでの分析をおこなったが，将来的にはより多様な対話文脈のもとで生成される応答や，ユーザとの会話のなかで生成される応答に含まれる矛盾についても分析することを検討している．

謝辞　本研究は，JSPS 科研費 JP19H04425, JP19J21913 の助成を受けたものである．

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. *arXiv:2004.13637*, 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. Association for Computational Linguistics.

[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. *arXiv:2001.09977*, 2020.

[4] 東中竜一郎, 船越孝太郎. Project next nlp 対話タスク：雑談対話データの収集と対話破綻アノテーションおよびその類型化. 言語・音声理解と対話処理研究会, 第 72 巻, pp. 45–50, 2014.

[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*, pp. 3731–3741, Florence, Italy, July 2019. Association for Computational Linguistics.

[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*, pp. 4715–4728, Online, July 2020. Association for Computational Linguistics.

[7] Abdelrhman Saleh, Tovly Deutsch, Stephen Casper, Yonatan Belinkov, and Stuart Shieber. Probing neural dialog models for conversational understanding. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, pp. 132–143, Online, July 2020. Association for Computational Linguistics.

[8] 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.

[9] Adina Williams, Nikita Nangia, and Samuel Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 1112–1122, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

A　応答極性の自動判定方法

3 節の実験において，大量のシステム応答に対し肯定・否定かどうかを人手で判定するにはコストがかかるため，自動での判定をおこなった．図 1 の例ではモデルの応答が"Yes"，"No"で始まる応答はそれぞれ肯定，否定として判定したが，これらの単語から始まらない応答を多く生成するシステムも存在したため，本実験では肯定文，否定文の冒頭に出現する文頭フレーズを収集したうえで，これらのフレーズから始まる応答をそれぞれ肯定，否定として判定した．文頭フレーズは，本実験において生成される応答から収集した．具体的には，生成される全 220,000 個の応答の文頭からフレーズ[3)]を取り出していき，100 回以上登場するフレーズについては，「肯定」，「否定」，「どちらともとれない」のいずれかに分類した．"Yes"または"No"で発話が始まるかで応答を分類したとき，ラベルを付与できたのは生成される全応答のうち 28.3% にとどまった一方，前述の方法により 39.1% の応答にラベルを付与することができた．また，2 節の実験で人手評価をした応答に対して同様に自動ラベリングをおこなったところ，自動ラベリングに成功した応答のうち 92.7% は人手ラベルと判断が一致し，"Yes"，"No"で発話が始まるかどうかでラベル付与したときの 92.4% と同程度の精度となった．

3)　各発話について，「,」または「.」を区切り文字として分割したときの一つ目の単語列を文頭フレーズとした．