多様な話者との自動対話に基づく雑談システムの自動評価

佐藤志貴 *,1　岸波洋介 *,1　杉山弘晃 2　赤間怜奈 1,3　徳久良子 1　鈴木潤 1,3
1 東北大学　2NTT コミュニケーション科学基礎研究所　3 理化学研究所
shiki.sato.d1@tohoku.ac.jp　yosuke.kishinami.q8@dc.tohoku.ac.jp

概要

雑談対話応答生成システムの評価方法として，システムの実際の対話を収集し性能評価を行う手法が注目されている．本研究では，評価対象システムと多様なシステムの対話を自動収集することで，対話相手の多様性を考慮しつつも人手を介さない自動評価の枠組みを提案する．評価実験では，提案手法がシステム性能の自動比較に有用であることを示す．

1　はじめに

雑談対話応答生成システム（以下，雑談システム）の自動評価は，人手評価に比べ低コストであり再現性が高いことから，日々のシステム改良の効果を検証する場合に有用である．とくに近年，評価対象システムと対話相手の間の発話交換により形成される実際の対話の質に基づいたシステム評価の自動化が注目されている [1, 2]．機械翻訳や自動要約などの他の生成タスクと異なり，対話ではシステムの運用時にユーザとのインタラクティブな発話交換が求められる．そのため，用意した入力に対する応答をもとにシステムを評価するのではなく，対話相手がいる状態でシステムを評価することが重要となる．

実際の対話の質に基づくシステム評価手法は (i) 評価対象システムの対話の収集と (ii) 収集した対話に基づくシステム評価の二段階からなり，先行研究でも (i), (ii) の自動化が試みられている．(i) の自動化手法として，評価対象システムに自分自身と対話させる方法（以下，自己対話）が一般的である [1, 2]．(ii) の自動化手法として，英語対話の評価で人手評価との強い相関が報告されている FED (fine-grained evaluation of dialog) などが挙げられる [3]．FED は，対話中の評価対象システムの発話に対する応答としてポジティブ，ネガティブ応答のどちらが妥当かを，用意した大規模雑談システムに推測させることで評価対象システムを評価する．

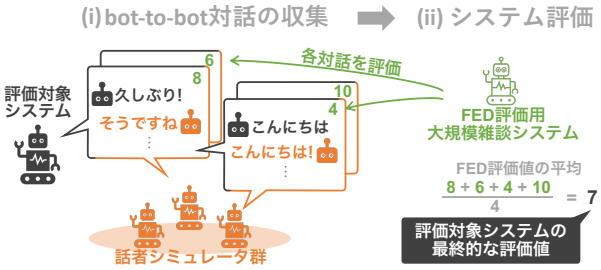


図 1: 提案する bot-to-bot 対話評価の概要．

ここで (i) の自動化について，自己対話による対話収集は対話相手となるシステム（以下，話者シミュレータ）を用意するコストがかからないという利点がある一方，対話相手の違いに対する応答の質の頑健性など，デプロイ後にシステムが直面する対話相手の多様性を考慮した評価ができない．

本研究では，多様な話者との自動対話に基づく雑談システム自動評価の枠組み，**bot-to-bot 対話評価**を提案する．(i) の自動化では，多様な話者シミュレータと評価対象システムの間の自動対話（bot-to-bot 対話）を収集することで，対話相手の多様性を考慮する．(ii) の自動化では，英語対話の評価で人手評価との相関が報告されている FED を用いる．実験では，まず FED により人手評価結果が付与された日本語対話を評価し，人手評価との相関を確認する．そのうえで雑談システムを実際に評価することで，提案手法により対話相手の多様性を考慮した高精度なシステム評価が可能であることを示す．

2　提案手法

本研究で提案する bot-to-bot 対話評価の枠組みは，(i) 評価対象システムの対話の収集と (ii) 収集した対話に基づくシステム評価の二段階からなる．

(i) 対話収集　最初に評価対象システムが参加する対話を自動収集する．このとき，多様な相手と対話することで，限られた相手と対話するよりも網羅的に評価対象システムの挙動が表出すると考えられる．本手法では，評価用に用意した n 個の多様な話

＊ 両者の貢献は同等である．

者シミュレータそれぞれと m 回対話させることで，$n \times m$ 個の対話を自動収集する．

(ii) システム評価　収集した対話をもとに任意の観点 v について評価対象システムの性能を評価する．具体的には，各対話での評価対象システムの発話 l 個を v について評価したときの評価値の平均を v における各対話の評価値とする．さらに，nm 個の対話の v における評価値の平均値を最終的な評価対象システムの v における評価値とする．システム発話の評価には FED を用いる．FED は，評価対象システムの発話に対する応答として v に関するポジティブ，ネガティブ応答のどちらが言語モデル的に妥当かを推測することで，v についてシステム発話を評価する．各ポジティブ，ネガティブ応答の妥当性は，学習済み大規模雑談システムを用いて自動評価する．文脈 c に対する評価対象システム発話 r の v に関する評価値は，以下のように算出される．

$$\sum_{p \in \mathcal{P}_v} w_p D(c+r, p; \theta) - \sum_{n \in \mathcal{N}_v} w_n D(c+r, n; \theta) \quad (1)$$

ここで，$\mathcal{P}_v, \mathcal{N}_v$ はそれぞれ v に関するポジティブ，ネガティブ応答の集合，w は各応答の重み，$D(c, \cdot; \theta)$ は c に対する応答文の生成確率を雑談システム（パラメータ θ）により算出する関数である．なお，Mehri らは $w = 1$ とする手法を提案した [3] が，本研究では少量の人手評価スコア付き対話データで w を学習させ，より高精度な評価を目指す．

3　FED による対話評価の精度検証

FED を用いて人手評価結果が付与されたシステムと人間の対話を評価し，w の学習の効果に注目しながら評価結果が人手評価と相関するか確かめる．

3.1　実験設定

評価する対話データセット　Sugiyama らが収集したシステムと人間の対話 1,459 件を使用した [4]．本データセットの各対話には，13 の観点についてシステムを 11 段階（0 から 10）で人手評価した結果が付与されている．データセットは学習用データと評価用データに 8 : 2 で分割した．

評価観点　上述したデータセットに人手評価スコアが付与されている 13 の観点について評価した．

評価に用いる応答文と雑談システム　FED 評価に用いる日本語のポジティブ，ネガティブ応答はクラウドソーシングを用いて収集した．13 観点それぞれに対して，ポジティブ，ネガティブ応答を 50 個ずつ収集し重複するものを除いたうえで，人手評価により各観点との関連性が低いものを除外した．また，FED スコアを算出する大規模雑談システムとして，Sugiyama らの事前学習済みのパラメータ数 1.6B の Transformer [4] を使用した[1)]．

3.2　結果

FED による評価結果と人手評価結果の順位相関係数について，各応答の重みを学習しない場合，観点平均は弱い相関が認められるとされる [5] 0.2 をわずかに上回る 0.232 であった[2)]．このことから，重みを学習しない場合，FED による対話評価の精度は高いものとはいえない．一方，重みを学習することで，順位相関係数の平均は 0.420 と，一定の相関が認められるとされる [5] 0.4 を上回った．以上より，重みを学習することで FED により一定の精度を持つ対話自動評価が可能であることがわかった．

4　提案手法による評価実験

3 節で FED が対話の自動評価方法として有効であることを確認した．そこで，FED を利用した提案手法による雑談システムの自動評価を実施する．

1 節で述べたように，自動評価の枠組みの主な用途は日々のシステム改良の効果の検証である．そのため，あるシステムと，そのシステムを改良したシステムなどの性能差を識別可能であることが望ましい．本実験では，次の三つの設定において提案手法がシステムの性能差を識別可能か検証する．

(a)アーキテクチャや学習設定が異なり，評価対象システム同士の性能も大きく異なる場合
(b)評価対象システム同士の性能が比較的近く，かつ平均的に性能が低い場合
(c)評価対象システム同士の性能が比較的近く，かつ平均的に性能が高い場合

4.1　実験設定

用意した雑談システム　評価対象もしくは話者シミュレータとして，合計 31 個の雑談システムを用意した．内訳は，杉山ら [4]，藤原ら [6]，岸波ら [7] が構築した Transformer 雑談システムがそれぞれ 25 個，2 個，1 個と，これらと比べて明らかに性能が劣ることを想定した LSTM 雑談システム 3 個である．以下，各システムは「製作者代表を示す記号 - アー

1）4.1 節で導入する表記では *s*-Tfm-1.6B-t2.1B に該当する．
2）観点ごとの順位相関係数は付録 A に示す．

キテクチャ名 - システムパラメータ数 - 学習データ[3]とその数」の形式で示す．

各雑談システムの割当　全システムを評価対象と話者シミュレータに分けた．設定 (a) ではアーキテクチャや学習データが異なる 4 個のシステム *s*-Tfm-1.6B-m300K, *f*-Tfm-480M-t300M+, *k*-Tfm-220M-t250M, *k*-LSTM-110M-t100M を評価対象とした．設定 (b) では 3 個の LSTM システム *k*-LSTM-{5M, 30M}-t10M, *k*-LSTM-110M-t100M を評価対象とした．設定 (c) では 4 個の Transformer システム *s*-Tfm-{0.3B, 0.7B, 1.1B, 1.6B}-f50K を評価対象とした．各設定で評価対象以外は話者シミュレータとした．

Bot-to-bot 対話収集の設定　評価対象システム，話者シミュレータの各組み合わせで 500 対話を収集した．会話冒頭の発話として，Sugiyama らによって公開されている JEmpatheticDialogues [4] の各対話の最初の 2 発話を使用した．評価対象システムが第一話者となる形式で各対話 5 ターンを収集した．対話評価は 3 節と同様に 13 の観点について行った．

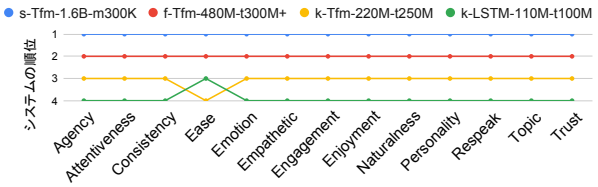
検証方法　観点ごとに，評価値に基づき作成した比較対象システムのランキングの妥当性を確認することで，提案手法の有効性を検証する．

4.2　結果

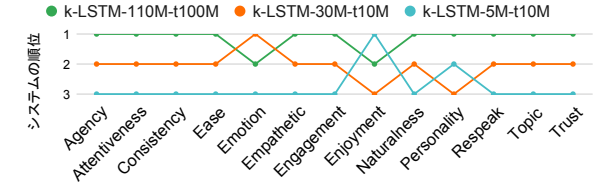
各設定での比較実験において，提案手法による自動評価結果をもとに作成した評価対象システムのランキングを図 2 に示す．

設定 (a)　評価対象システムは，アーキテクチャ，学習設定の差から，*s*-Tfm-1.6B-m300K, *f*-Tfm-480M-t300M+, *k*-Tfm-220M-t250M, *k*-LSTM-110M-t100M の順に性能が高いと考えられる．図 2 から，自動評価結果に基づくランキングは，1 観点を除き事前の予想に従うものとなった．したがって提案手法は，アーキテクチャや学習設定が大きく異なるシステムの性能差を識別可能であることが確認できた．

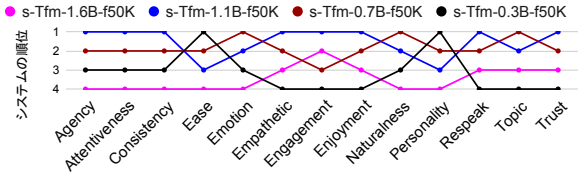
設定 (b)　評価対象システムは，モデルパラメータ数，学習データ数の差から，*k*-LSTM-110M-t100M, *k*-LSTM-30M-t10M, *k*-LSTM-5M-t10M の順に性能が高いと考えられる．図 2 から，自動評価結果に基づくランキングは，3 観点を除き事前の予想に従うものとなった．したがって提案手法は，比較的性能が低いとされるシステムの性能差を識別可能であることが確認できた．



(a) アーキテクチャや学習設定が異なるシステムの比較．



(b) LSTM の比較．



(c) Transformer の比較．

図 2: 評価対象システムのランキング．

設定 (c)　評価対象システムは，モデルパラメータ数の差から，*s*-Tfm-1.6B-f50K, *s*-Tfm-1.1B-f50K, *s*-Tfm-0.7B-f50K, *s*-Tfm-0.3B-f50K の順に性能が高いことが予想できる．しかし図 2 から，自動評価結果に基づくランキングは事前の予想とは異なっていた．このことから提案手法は，近年の高性能システムの性能差の識別には評価の解像度において改善の余地があることがわかった．

5　議論: 提案手法の特長と課題

5.1　対話相手の多様化の重要性

提案手法の特徴は，多様な話者シミュレータを用意することで，対話相手の多様性を考慮した評価を行う点である．4 節の実験では，多様な対話相手を用意する重要性を示唆する例が確認されていた．表 1 に，*s*-Tfm-1.6B-m300K の評価時に収集した対話の例を示す．高性能な *s*-Tfm-1.6B-p50K が対話相手であった同表 I の対話では，両話者が適切な発話の生成に成功している．一方で 低性能な *k*-LSTM-5M-t10M が対話相手であった同表 II の対話では，対話相手が過去発話の繰り返しなど，低品質な応答を生成している．このような低品質な応答の直後に，評価対象システムもつられるように不自

3)　学習データの e は JEmpatheticDialogues，f は趣味雑談コーパス，p は JPersonaChat [4]，m は e, f, p の混合コーパス，t は Twitter のリプライチェーンを示す．Sugiyama らのシステムにおいて学習データが t 以外のシステムは，*s*-Tfm-{0.3B, 0.7B, 1.1B, 1.6B}-t2.1B を追加学習したものである．

表 1: 評価対象システム（話者 A）が *s*-Tfm-1.6B-m300K である場合の対話例.

I. 対話相手（話者 B）は *s*-Tfm-1.6B-p50K

話者	発話
A:	寒い日はしっかり湯舟に浸かりたいと思いますよね。
B:	日本人はやっぱりお風呂が好きですからね。
A:	そうですね。 嫌いな人はいないですよね。 お肌もきれいになりますし。
B:	はい、エステティシャンをしているのですが、肌の手入れにはついつい時間をかけちゃいます。
A:	お肌の手入れにはこだわるんですね。 毎日されているんですか?
B:	はい、20 代のころから肌の手入れは欠かしません。
A:	そうなんですね。 お手入れを欠かさないって素敵ですね。
B:	ありがとうございます。あなたは何かご趣味はお持ちですか?
A:	最近テレビを観ることと音楽を聴くことにはまっています。
B:	そうなんですね。どんなジャンルの音楽がお好きですか?

II. 対話相手（話者 B）は *k*-LSTM-5M-t10M

話者	発話
A:	寒い日はしっかり湯舟に浸かりたいと思いますよね。
B:	日本人はやっぱりお風呂が好きですからね。
A:	はい、肌もつるつるになりますからね。
B:	はい、そうします。
A:	旅行は国内ですか?国外ですか? いい温泉があれば教えてほしいです。
B:	京都は国外ですか?
A:	京都も国内です。 京阪で行けますので。
B:	京都は国内ですか?
A:	**京都も国外です。 関西です。**
B:	京都は国内ですか?

然な発話を生成している（表太字部分）．本例から，評価対象システムは対話相手の高品質な応答には適切な発話を生成できるが，低品質な応答には不適切な発話を生成する可能性があることがわかった．

運用時では，操作ミスをするユーザやシステムに非協力的なユーザとの対話なども想定されるため，低品質な発話に対し適切な応答を生成する能力もシステムにとって必要となる．そのため，対話 I だけでなく対話 II のような状況におけるシステム発話を評価することも重要となる．

5.2 観点ごとの評価に関する今後の課題

3 節では，FED により各観点について人手評価と一定の相関を持つ自動評価が可能であることを確認した．一方で，観点 v の FED による評価結果が，v 以外の観点の人手評価と強く相関する場合があることがわかった．たとえば，観点 Naturalness について FED により算出した自動評価スコアは，Naturalness の人手評価スコアと順位相関係数 0.463 の相関を持つ（付録 A）が，同時に観点 Trust の人手評価スコアと順位相関係数 0.493 の相関を持つ．このことから，FED を用いた自動評価は現状，各観点を隔てる細かい違いの評価まではできていないと考えられる．

評価における各観点の違いの理解は，とくに設定 (c) など評価対象の性能が高い場合に重要となると考えられる．たとえば，低性能システムが生成しうる対話破綻をもたらす発話[4]は，多数の観点で低く評価されることが予想される．そのため低性能システムの自動評価では，各観点で着目すべき点が違うことを考慮できなくとも，各観点に対する人手評価との相関は高くなりやすい．一方で，設定 (c) で評価対象システムとなっているような Transformer ベースの大規模システムは，文脈に沿った自然な応答が生成可能であることが知られている．これらのシステムの観点ごとの性能比較では，各システムが自然な応答を生成するなかで，観点ごとに挙動の違いを的確に見出し優劣を判定する必要がある．

各観点を隔てる細かい違いを理解した高精度な自動対話評価を実現し，設定 (c) などの高難度なシステム比較を自動化することが今後の課題の一つである．自動対話評価手法の改良案として，FED 評価に用いるポジティブ・ネガティブ文の増強や，各文の重みの学習方法の改良を検討している．

6 おわりに

本研究では，評価対象システムと多様なシステムの対話を自動収集することで，対話相手の多様性を考慮しつつも人手を一切介さない bot-to-bot 対話評価の枠組みを提案した．評価実験では，評価対象システムが高性能な場合において課題が残るものの，評価対象システム同士のアーキテクチャや学習データが異なる場合や，評価対象システム性能が低性能である場合，提案手法によりシステムの性能差を自動で識別できることがわかった．また，本手法の特徴である対話相手の多様化により，網羅的な評価対象システムの挙動を考慮した評価が可能であることを確認した．今回は提案手法が評価対象システムを予想される順位に並び替えられるかにより評価の有効性を検証したが，今後は各観点で評価対象システムを人手評価し得られるシステムのランキングとの一致をもとに評価の有効性を検証する予定である．

4） 文脈に無関係な発話や同じ発話の過度な繰り返しなど.

謝辞

本研究の一部は JSPS 科研費 JP19H04162, JP19H05693，JP21J22383 の助成を受けたものです．

参考文献

[1] Asma Ghandeharioun, Judy Hanwen Shen, Natasha Jaques, Craig Ferguson, Noah Jones, Agata Lapedriza, and Rosalind Picard. Approximating Interactive Human Evaluation with Self-Play for Open-Domain Dialog Systems. In **Proceedings of the 2019 Conference on Neural Information Processing Systems (NeurIPS)**, 2019.

[2] Jan Deriu and Mark Cieliebak. Towards a Metric for Automated Conversational Dialogue System Evaluation and Improvement. In **Proceedings of the 12th International Conference on Natural Language Generation (INLG)**, pp. 432–437, 2019.

[3] Shikib Mehri and Maxine Eskenazi. Unsupervised Evaluation of Interactive Dialog with DialoGPT. In **Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)**, pp. 225–235, 2020.

[4] Hiroaki Sugiyama, Masahiro Mizukami, Tsunehiro Arimoto, Hiromi Narimatsu, Yuya Chiba, Hideharu Nakajima, and Toyomi Meguro. Empirical Analysis of Training Strategies of Transformer-based Japanese Chit-chat Systems. **arXiv:2109.05217**, 2021.

[5] Haldun Akoglu. User's guide to correlation coefficients. **Turkish journal of emergency medicine**, Vol. 18, No. 3, pp. 91–93, 2018.

[6] 藤原吏生, 岸波洋介, 今野颯人, 佐藤志貴, 佐藤汰亮, 宮脇峻平, 加藤拓真, 鈴木潤, 乾健太郎. ILYS aoba bot: 大規模ニューラル応答生成モデルとルールベースを統合した雑談対話システム. 第 90 回 言語・音声理解と対話処理研究会, pp. 110–115, 2020.

[7] 岸波洋介, 赤間怜奈, 佐藤志貴, 鈴木潤, 徳久良子, 乾健太郎. 対話システムの先読み能力実現に向けた未来の展開まで生成する学習戦略の提案と分析. 2021 年度人工知能学会全国大会, 2021.

A FED による評価結果と人手評価結果の順位相関係数

3 節において算出した FED による評価結果と人手評価結果の順位相関係数について，観点ごとに算出した結果を表 2 に示す．同表から，各応答文の重みを学習しない場合は観点によって人手評価との相関にばらつきがあり，相関が全く認められないものも存在する．一方，重みを学習することでどの観点も人手評価との順位相関係数が 0.4 前後まで向上した．

表 2: 人手評価結果との順位相関係数.

重み学習	Agency	Attentiveness	Consistency	Ease	Emotion	Empathetic	Engagement	Enjoyment	Naturalness	Personality	Respeak	Topic	Trust	平均
なし	0.148	0.312	0.266	0.170	0.235	0.314	0.259	0.313	0.204	0.173	0.277	0.115	0.226	0.232
あり	0.373	0.446	0.423	0.470	0.426	0.453	0.420	0.350	0.463	0.400	0.375	0.341	0.518	0.420

B 評価実験の詳細

表 3 に，4 節の評価実験で用いた雑談システムの一覧を示す．システム名は「製作者代表を示す記号 - アーキテクチャ名 - システムパラメータ数 - 学習データ[5)] とその数」の形式となっている．

また，算出した各システムの FED による評価値を表 4 に示す．

表 3: 使用した雑談システムの一覧.

製作者代表	システム名
Sugiyama	*s*-Tfm-{0.3B, 0.7B, 1.1B}-t2.1B, *s*-Tfm-{0.3B, 0.7B, 1.1B, 1.6B}-{e, f, m, p}50K, *s*-Tfm-{0.3B, 0.7B, 1.1B, 1.6B}-m150K, *s*-Tfm-1.6B-f100K, *s*-Tfm-1.6B-m300K
藤原	*f*-Tfm-480M-t{300M, 300M+}
岸波	*k*-Tfm-220M-t250M, *k*-LSTM-{5M, 30M}-t10M, *k*-LSTM-110M-t100M

表 4: 各設定におけるシステムの評価値.

設定 (a)

評価対象システム	Agency	Attentiveness	Consistency	Ease	Emotion	Empathetic	Engagement	Enjoyment	Naturalness	Personality	Respeak	Topic	Trust
s-Tfm-1.6B-m300K	7.701	6.665	7.485	7.536	8.239	6.619	7.703	8.080	7.913	7.407	8.326	7.765	7.189
f-Tfm-480M-t300M+	7.246	5.740	6.743	6.919	7.925	6.072	7.352	7.056	7.663	7.201	8.118	7.650	6.627
k-Tfm-220M-t250M	6.819	4.743	6.554	6.531	7.179	5.731	6.426	6.497	7.279	6.965	7.611	6.577	6.093
k-LSTM-110M-t100M	6.502	4.672	6.194	6.560	6.825	5.567	6.293	6.297	6.992	6.871	7.190	6.400	5.301

設定 (b)

評価対象システム	Agency	Attentiveness	Consistency	Ease	Emotion	Empathetic	Engagement	Enjoyment	Naturalness	Personality	Respeak	Topic	Trust
k-LSTM-110M-t100M	6.547	4.690	6.226	6.533	6.846	5.679	6.297	6.345	7.015	6.888	7.250	6.406	5.356
k-LSTM-30M-t10M	6.065	4.636	5.918	6.049	6.850	5.322	6.235	6.228	6.555	6.705	7.159	6.321	4.736
k-LSTM-5M-t10M	5.858	4.363	5.694	5.234	6.844	4.840	5.971	6.393	5.909	6.722	6.847	5.958	3.066

設定 (c)

評価対象システム	Agency	Attentiveness	Consistency	Ease	Emotion	Empathetic	Engagement	Enjoyment	Naturalness	Personality	Respeak	Topic	Trust
s-Tfm-1.6B-f50K	7.792	6.691	7.404	7.677	8.066	6.406	7.875	7.979	7.860	7.596	8.597	7.924	6.749
s-Tfm-1.1B-f50K	7.828	6.777	7.524	7.740	8.122	6.423	7.971	8.049	7.972	7.622	8.729	7.973	6.870
s-Tfm-0.7B-f50K	7.795	6.715	7.495	7.777	8.180	6.422	7.874	7.995	7.973	7.629	8.717	7.979	6.797
s-Tfm-0.3B-f50K	7.794	6.714	7.437	7.812	8.085	6.312	7.789	7.966	7.927	7.665	8.590	7.901	6.657

5） 学習データの e は JEmpatheticDialogues，f は趣味雑談コーパス，p は JPersonaChat [4]，m は e, f, p の混合コーパス，t は Twitter のリプライチェーンを示す．