フォローアップ発話を用いた日本語対話の自動評価

川本 稔己 [1,2] 岡野 裕紀 [1] 山崎 天 [2] 佐藤 敏紀 [2] 船越 孝太郎 [1] 奥村 学 [1]
[1] 東京工業大学　[2] LINE 株式会社
{kawamoto, okano, funakoshi, oku}@lr.pi.titech.ac.jp
{takato.yamazaki, toshinori.sato}@linecorp.com

概要

対話の人手評価には時間とコストがかかるため, 人手評価と自動評価の相関を向上し自動評価の信頼性を高めていくことは重要である. 本稿では, フォローアップ発話が生成される尤度を利用して対話を評価する FED と FULL という自動評価指標に着目し検証と改善を行う. 日本語対話データを利用した実験の結果, FED と FULL は人手評価との相関があまり見られなかったが, 我々の改善した手法の相関係数は 0.58 で最も高い値となった. また, 我々の改善した手法を用い, 対話データから対話システムの順位を予測した結果, 高い相関を示し, 特にオープンドメインの対話は正解の順位と完全に一致した.

1　はじめに

近年の事前学習済み言語モデルによる性能向上に伴い, 年々優れた対話システム [1, 2, 3, 4] が発表されているが, 対話システムの優劣を評価するためには最終的に人手評価を行うことが一般的である [5]. しかし, 人手評価には時間とコストがかかるため, 開発段階では自動評価を利用して評価できれば効率的である. 本研究では, 対話の人手評価と相関があり信頼性の高い自動評価指標を提案することを目指す.

対話の自動評価指標には正解の応答が必要な指標と不要な指標がある. 正解の応答が必要な指標としては BLEU [6], FBD [7] 等が挙げられ, 正解の応答が不要な指標としては Perplexity や USR [8] 等が挙げられる. 正解の応答が必要な指標は, 多様な応答が許容される対話の性質上適切に評価ができなかったり, 正解の応答を作成するための時間とコストがかかるという問題がある. よって本稿では, 正解の応答が不要な指標の 1 つであり, 人手評価との相関が高いことが報告されている [5], FED [9] と, FED を発展させた FULL [10] に着目する. FED と FULL が日本語対話に利用できるか検証し, 先行研究から考えられる問題点（3 節参照）を改善することを試みる. 実験の結果, 我々の改善した手法は先行研究より人手評価との相関が高いことを確認した. さらに, 対話データから対話システムの順位を予測したところ, オープンドメインの対話とシチュエーションに特化した対話それぞれで高い相関を示し, 特に, オープンドメインの対話は正解の順位と完全に一致した.

2　先行研究

本節では, 対話の自動評価指標である FED と, FED を発展させた FULL について説明する. FED は, 対話履歴が与えられたときに, それに続くフォローアップ発話の対数尤度を測ることによって評価を行う指標である. FED の評価項目とフォローアップ発話の例を図 1 に示す. 評価項目は大きく分けてターンレベルと対話レベルの 2 種類があり, ターンレベルでは, 対話において評価したい話者の発話の後にフォローアップ発話を付与し対数尤度を算出する. 一方, 対話レベルでは, 対話の最後にフォローアップ発話を付与し対数尤度を算出する. 各レベルごとに評価項目は複数あり, ターンレベルでは 8 個, 対話レベルでは 10 個, 計 18 個の評価項目が使われている. 各評価項目には, 平均 4.05 個のフォローアップ発話が設定されており, フォローアップ発話は, ポジティブとネガティブの 2 種類に分けられる. 図 1 に示すように, フォローアップ発話の例として, Interesting の項目に当てはまるポジティブの発話は”Wow that is really interesting.”といった, 前の発話が面白い場合に尤度が上がる発話を利用し, ネガティブの発話は”That’s really boring.”といった, 前の発話が面白くない場合に尤度が上がる発話が利用されている. 対数尤度の計算には, 公開されている DialoGPT [11] を使用している. 追加でデータの準備やモデルの訓練をする必要がないため, 他の対話自動評価指標 [12, 13] と比べて比較的手軽に評価を行うことができる. また, Tianbo ら [5] が行った自動評価と人手

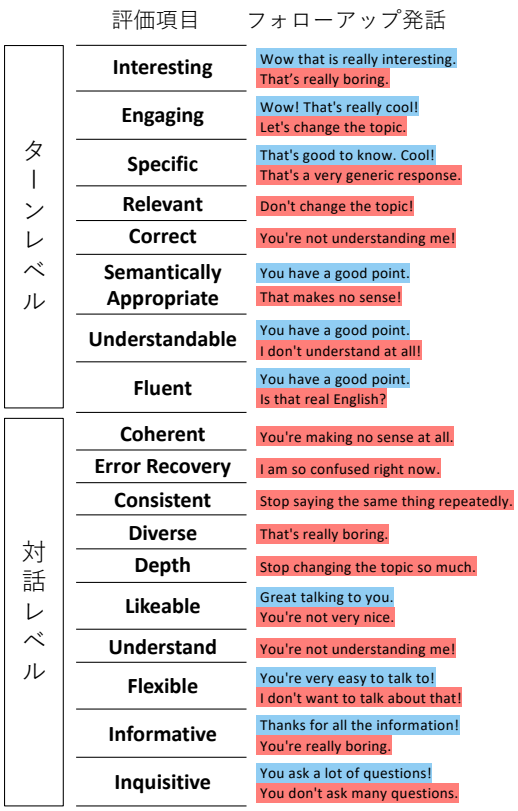

図 1 FED の評価項目と各項目に設定されているフォローアップ発話の例. フォローアップ発話のうち, ポジティブを青色, ネガティブを赤色で示している. ネガティブ発話しか表示されていない評価項目はポジティブ発話が定義されていないことを示す.

評価の相関を測る実験で FED と人手評価の相関が 0.59 と報告されており, 彼らの論文中で使われた自動評価指標の中では最も高い相関を得ている.

FULL は FED を発展させた指標であり, FED との差分としては, (1) FED では対話履歴とフォローアップ発話を含めた対数尤度を算出していたが, 対話履歴が与えられた場合のフォローアップ発話の条件付き対数尤度に変更したこと, (2) 各評価項目を撤廃し, 5 つのフォローアップ発話のみに限定したこと, (3) 複数の言語モデルで比較を行い, DialoGPT から BlenderBot [14] に変更したことである. FED データセットで人手評価との相関係数を計った結果, state-of-the-art となる 0.69 が報告されている.

3 先行研究の問題点と調査

本節では, FED・FULL から考えられる問題点と, 問題点を改善するための手法を説明する. 2 節で述べたように FED は, 18 個の評価項目を用いているが, Semantically Appropriate, Understandable, Fluent の 3 つの項目のポジティブのフォローアップ発話は全く同じ発話が使用されている. それ以外にもフォローアップ発話が同一である評価項目が存在し, 全ての評価項目が独立した項目として測れているか疑わしい. そのため, 我々は対話評価を行う際に, 18 項目全てを用いるよりも, 一部の評価項目を用いたほうが優れているのではないかと仮説を立てた. 具体的には, 検証データで最も高かった評価項目の組み合わせのみを選択し, その組み合わせを使ってテストデータで相関係数の評価を行った.

一方, FULL で限定された 5 個のフォローアップ発話は, FED データセットで相関が高かった発話が選ばれており, 別のデータセットでも適切なフォローアップ発話なのか不明であるため, FULL で 5 発話に限定しない手法も比較する. さらに, 限定された 5 発話は全てネガティブのフォローアップ発話なので, ポジティブの発話のみや, ネガティブの発話のみを使って評価を行う. それ以外にも, GPT [15] のような基盤モデル [16] を用いて対数尤度を算出するため, フォローアップのテキストが発話である必要はない可能性があり, 評価項目名をそのままフォローアップとした手法の評価も行う.

4 実験

4.1 実験設定

本実験では, 対話ごとにつけられた人手評価と各手法によって得られる自動評価の相関係数を比較することで自動評価指標を評価する. 評価データには, 対話システムライブコンペティション 3 [17] (以降, 対話コンペ 3) で行われた予選のデータ[1)]を使用する. 対話コンペ 3 にはオープントラックとシチュエーショントラックの 2 種類のトラックがあり, オープントラックではオープンドメインの対話, シチュエーショントラックではシチュエーションに沿った対話が行われている. どちらのトラックもワーカとシステムがチャットベースで対話を行っており, 人手評価は対話を行ったワーカが行う. オープントラックの評価指標は以下の 3 つでそれぞれ 5 段階で評価している.

- 自然性：対話が自然かどうか
- 話題追随：システムはユーザが選択した話題に関して適切に応答できたかどうか
- 話題提供：システムはユーザが選択した話題に

1) https://dialog-system-live-competition.github.io/dslc3/data.html

表 1 対話コンペ 3 データセットのサイズ.

	オープン	シチュエーション
対話数	239	296
チーム数	5	6
1 チームの平均対話数	47.80	49.33
1 対話の平均発話数	30.00	30.00

関して新たな情報を提供できたかどうか

これら 3 指標の平均点を各対話における人手評価のスコアとした. シチュエーショントラックの評価指標は以下の 1 項目で, 同様に 5 段階で評価している.

- どれくらいシチュエーションに適しており、かつ、人らしい会話か

データセットのサイズを表 1 に示す. 対話コンペ 3 に参加した一部のチームのデータは公開されておらず, 公開されているデータのみを実験の対象とした.

尤度を計算するための言語モデルは日本語で学習された GPT[2)] を利用した. フォローアップ発話は FED・FULL のフォローアップ発話を第一著者が人手で日本語に翻訳した. 対話ごとの人手評価と自動評価の相関は Spearman の順位相関係数と Pearson の相関係数を利用し算出する. 2 分割交差検証を行い 2 つの相関係数の平均を最終的な結果とする.

4.2 手法

実験では以下の手法を比較する.

FED FED を日本語に適応した手法.

FULL FULL を日本語に適応した手法.

FED-Cond FED に 2 節で説明した FULL との差分, (1) 条件付き対数尤度を適用した手法.

FED-Cond-Pos FED-Cond のポジティブのフォローアップ発話のみを用いた手法.

FED-Cond-Neg FED-Cond のネガティブのフォローアップ発話のみを用いた手法.

FED-Cond-Tag FED-Cond のフォローアップとして, 発話の代わりに評価項目名を用いた手法.

FED-Selected FED の評価項目から検証データで選ばれた項目のみを用いた手法.

FED-Cond-Selected FED-Cond の評価項目から検証データで選ばれた項目のみを用いた手法.

4.3 実験結果

実験結果を表 2 に示す. 先行研究である FED と FULL の相関係数は, オープントラックではマイナスで相関がないこと, シチュエーショントラックでは, Spearman の順位相関係数が 0.27 と多少の相関はあることがわかったが, 英語対話の評価結果とは大きく乖離があることを確認した. 先行研究をそのまま日本語対話の評価に適用することは難しいことがわかった.

表 2 実験結果. Spear. は Spearman の順位相関係数, Pear. は Pearson の相関係数を示す.

	オープン		シチュエーション	
	Spear.	Pear.	Spear.	Pear.
FED [9]	-0.283	-0.244	0.278	0.339
FULL [10]	-0.018	-0.008	0.279	0.311
FED-Cond	0.279	0.277	0.297	0.333
FED-Cond-Pos	0.302	0.257	-0.010	0.005
FED-Cond-Neg	-0.296	-0.251	0.258	0.262
FED-Cond-Tag	0.040	0.038	0.001	-0.041
FED-Selected	0.485	0.476	0.249	0.312
FED-Cond-Selected	**0.585**	**0.576**	**0.315**	**0.371**

FED を条件付き対数尤度に変更した FED-Cond は, FED より高い相関を得ていることから, 2 節で説明した FULL との差分である, (1) の条件付き対数尤度は効果的であることを示しており, FULL より高い相関を得ていることから, (2) の 5 発話に絞ることは日本語対話の評価に悪影響を与えていることを示している. FED-Cond-Pos と FED-Cond-Neg を比較すると, オープントラックにおいては, ポジティブのフォローアップ発話のほうが効果的であるが, シチュエーショントラックにおいては, ネガティブのフォローアップ発話のほうが効果的であった. 効果的な発話が 2 つのトラックで異なったのは, シチュエーショントラックの設定が ”先輩からの「同窓会の幹事の依頼」を断りたい状況” であるため, システムの発話がネガティブに当てはまる場合が多かったことが理由として考えられる. その上で, FED-Cond の方が高い相関係数を得ていることから, ポジティブとネガティブどちらの発話も使ったほうが効果があると考えられる. FED-Cond-Tag は, どちらのトラックでも相関は見られず, 基盤モデルで実験を行う場合も評価項目名をフォローアップとして与えるより, 評価項目の意図を反映するような発話を用いたほうが良いことがわかる. その上で, 検証データを用いて評価項目を選ぶ FED-Cond-Selected は他の手法と比較してどちらのトラックでも最も高い相関係数を得た. よって, 18 個の評価項目から一部の項目を使用することで相関係数が向上することを確認した.

2) https://huggingface.co/rinna/japanese-gpt-1b

表 3 オープントラックのランキング結果. 結果は”チーム名 (スコア)” で表記する. 赤が正しく順位を予測できたチーム, 青が誤って順位を予測したチームを示す.

順位	正解	FED	FULL	FED-Cond-Selected
1.	A (3.83)	C (1.55)	E (3.55)	A (9.86)
2.	B (3.11)	E (1.34)	B (3.45)	B (9.25)
3.	C (2.64)	B (1.19)	C (3.44)	C (9.14)
4.	D (2.10)	D (1.18)	A (3.43)	D (8.92)
5.	E (1.45)	A (1.09)	D (3.38)	E (8.71)
Spear.	-	-0.50	-0.30	1.00

表 4 シチュエーショントラックのランキング結果.

順位	正解	FED	FULL	FED-Cond-Selected
1.	A (4.26)	D (1.49)	A (3.94)	A (8.09)
2.	B (3.92)	A (1.40)	C (3.65)	C (7.48)
3.	C (3.76)	E (1.39)	B (3.61)	B (7.40)
4.	D (3.76)	B (1.33)	E (3.54)	F (7.20)
5.	E (3.63)	C (1.24)	D (3.52)	D (7.07)
6.	F (3.28)	F (1.20)	F (3.47)	E (7.03)
Spear.	-	0.37	0.89	0.77

4.4 ランキング

本稿で提案した手法 FED-Cond-Selected は最も高い相関があることを確認したが, 今回使用したデータセットである対話コンペ 3 予選の本来の目的は, チームのランキングを決めることである. よって, 対話それぞれの自動評価スコアをチームごとに平均することでチームのランキングを試みる.

ランキング結果とその Spearman の順位相関係数を, オープントラックは表 3 に, シチュエーショントラックは表 4 に示す. チーム名はわかりやすさのために順位の高い方からアルファベット順で表記した. オープントラックで, FED・FULL が正しく順位を予測したのはそれぞれ 1, 2 チームだけであったが, FED-Cond-Selected は正解の順位と完全に一致した. 一方で, シチュエーショントラックでは, どの手法も正しく順位を予測したのは数チームであり, 相関係数は FULL が最も高かったが, FED-Cond-Selected も 0.77 と高い相関を示した.

4.5 考察

実験結果から, オープントラックで行われたオープンドメイン対話に関しては比較的高い相関が見られ, ランキングが正解と一致したことから, 本稿で提案した手法は対話の自動評価を行える可能性があることを示唆している. 一方, シチュエーショントラックに関してはオープントラックと比べて相関係数が低く, ランキングが一致するチーム数も少なかった. これは, シチュエーショントラックの指標である ”どれくらいシチュエーションに適しており、かつ、人らしい会話か” の前半部分であるシチュエーションに適しているかを今回使用したフォローアップ発話では測れなかった可能性があり, シチュエーションに特化した評価項目とそのフォローアップ発話を新たに作成することで対応できる可能性がある.

また, FED-Cond-Selected において, 2 分割交差検証で最も高い相関係数を得た評価項目の組み合わせは, オープントラックでは Specific, Relevant, Fluent の 3 項目と Interesting, Specific, Correct, Fluent の 4 項目, シチュエーショントラックでは Interesting, Engaging, Specific, Fluent, Depth の 5 項目と Specific, Relevant, Semantically Appropriate, Depth, Understand の 5 項目であった. 評価する対話の性質や目的, 状況により必要な項目は変化すると考えられるが, 今回使われた項目数は最大で 5 項目であることから, 18 項目全てを利用する必要はないことを確認した. 選ばれた評価項目で互いのトラックを評価すると, オープントラックの組み合わせを使用したシチュエーショントラックの Spearman の順位相関係数は 0.320, シチュエーショントラックの組み合わせを使用したオープントラックの Spearman の順位相関係数は 0.529 となり, 性能の劣化はあまり見られなかった.

5 おわりに

本稿では, 対話の自動評価指標である FED・FULL が日本語対話の評価に利用できるか検証した結果, 人手評価とはほとんど相関がないことを確認したが, 一方で 18 個の評価項目からいくつかの項目だけを選んで利用することで相関係数が向上することを示した. その上で, 選んだ項目のみを用いる手法でシステムの順位を予測したところ, オープンドメインの対話とシチュエーションに特化した対話のどちらも高い相関を示した. 今後の展望としては, 今回提案した手法をさらに発展させて, 使用するフォローアップ発話を対話ごとに動的に選んで自動評価を行う枠組みを構築したい. また, 今回選ばれた評価項目を他の対話データに適用したときに, 同様の結果を得られるか検証を行っていきたい.

参考文献

[1] 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, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[2] Kurt Shuster, Jing Xu, Mojtaba Komeili, Da Ju, Eric Michael Smith, Stephen Roller, Megan Ung, Moya Chen, Kushal Arora, Joshua Lane, et al. Blenderbot 3: a deployed conversational agent that continually learns to responsibly engage. **arXiv preprint arXiv:2208.03188**, 2022.

[3] Daniel Adiwardana, Minh-Thang Luong, David R So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, et al. Towards a human-like open-domain chatbot. **arXiv preprint arXiv:2001.09977**, 2020.

[4] Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. Lamda: Language models for dialog applications. **arXiv preprint arXiv:2201.08239**, 2022.

[5] Tianbo Ji, Yvette Graham, Gareth Jones, Chenyang Lyu, and Qun Liu. Achieving reliable human assessment of open-domain dialogue systems. In **Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 6416–6437, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[6] 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**, pp. 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics.

[7] Jiannan Xiang, Yahui Liu, Deng Cai, Huayang Li, Defu Lian, and Lemao Liu. Assessing dialogue systems with distribution distances. In **Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021**, pp. 2192–2198, Online, August 2021. Association for Computational Linguistics.

[8] Shikib Mehri and Maxine Eskenazi. USR: An unsupervised and reference free evaluation metric for dialog generation. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 681–707, Online, July 2020. Association for Computational Linguistics.

[9] 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**, pp. 225–235, 1st virtual meeting, July 2020. Association for Computational Linguistics.

[10] 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, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics.

[11] 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.

[12] Zekang Li, Jinchao Zhang, Zhengcong Fei, Yang Feng, and Jie Zhou. Conversations are not flat: Modeling the dynamic information flow across dialogue utterances. In **Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)**, pp. 128–138, Online, August 2021. Association for Computational Linguistics.

[13] Chen Zhang, Yiming Chen, Luis Fernando D'Haro, Yan Zhang, Thomas Friedrichs, Grandee Lee, and Haizhou Li. DynaEval: Unifying turn and dialogue level evaluation. In **Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)**, pp. 5676–5689, Online, August 2021. Association for Computational Linguistics.

[14] Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael 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**, pp. 300–325, Online, April 2021. Association for Computational Linguistics.

[15] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

[16] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. **arXiv preprint arXiv:2108.07258**, 2021.

[17] 東中竜一郎, 船越孝太郎, 高橋哲朗, 稲葉通将, 角森唯子, 赤間怜奈, 宇佐美まゆみ, 川端良子, 水上雅博, 小室允人, ドルサテヨルス. 対話システムライブコンペティション 3. 人工知能学会研究会資料 言語・音声理解と対話処理研究会, Vol. 90, p. 23, 2020.