応答の生成・評価・選択による対話システム

榮田亮真　河原大輔
早稲田大学理工学術院
s.ryoma6317@akane.waseda.jp dkw@waseda.jp

概要

本論文では、人間から良い評価が得られる応答を生成する対話システムを目指し、応答評価モデルを対話システムに取り入れることを提案する。提案するシステムは、応答生成モデルが複数生成した応答を応答評価モデルで評価し、最適な応答を選択して応答を出力するものである。提案システムによる出力をベースラインシステムの出力と比較する人手評価を行った。人手評価の結果、提案システムの応答が良いと判断されることが多く、提案手法の有効性が示された。

1　はじめに

対話システムはルールベースのシステムが利用されることが多い領域であった。近年では他の自然言語処理タスクと同様に、深層ニューラルネットワーク (以下、DNN) を利用して、自然な応答を生成することができるようになっている。

対話システムの自動評価は、BLEU [1] などの、システムの応答と正解の応答を比較する手法が一般的に用いられている。しかし、そのうえでは、対話が持つ性質で、1 つの発話に対して適切な応答が多く存在することを意味する、one-to-many の性質 [2] が障壁となっている。1 文もしくは少数の正解応答との比較では、正解応答とまったく異なる文だが適切な応答を正しく評価できない。そのため、正解応答を用いない評価手法が求められており、その 1 つが DNN による評価である。人間やシステムが行う応答に対する人手評価を正解データとして DNN の学習を行うことで、人手評価とある程度の相関がある評価ができるようになる方法が提案されている [3, 4]。

本研究では、DNN による応答評価モデルの利用法として、独立に研究されていた応答生成モデルと応答評価モデルを合わせて 1 つの対話システムとすることを提案する。応答評価モデルは正解応答を必要としないことから、応答の生成時にも利用できることが特徴である。応答評価モデルを対話システムに組み込むことで、人からより良いと判断される応答を生成できるようになることを目指す。具体的には、生成モデルによって応答を複数生成し、それらの応答を評価モデルによって評価、もっとも高い評価を得た応答を選択するという応答生成・評価・選択システムを提案する。応答生成モデルには事前学習モデルの T5 [5]、応答評価モデルには事前学習モデルの BERT [6] を利用する。それらのモデルは独立に Fine-tuning する。応答を複数生成するためには、グリーディーサーチ、ビームサーチ [7]、サンプリング [8] の 3 種類のデコーディング手法を用いる。

提案手法を評価するために、提案システムとベースラインシステムの出力を比較するクラウドソーシングで人手評価を行った。このクラウドソーシングで、提案システムがより良い応答を出力すると判断されることが多く、提案手法の有効性が示された。

なお、対話システムの既存研究として、複数ターンの対話を扱うもの、発話以外の知識 [9] や感情 [10]、個性 [11] などを扱うものも存在するが、本研究では、単一の発話のみを入力し、応答を出力する 1 ターンの対話システムを扱う。

2　関連研究

対話システムによる応答を評価する手法は、人手評価と自動評価に分けられ、さらに自動評価は正解文を必要とする評価と必要としない評価に分類できる。人手評価は、関連度、おもしろさ、流暢さなどの観点を設け、クラウドソーシング等で評価を集めるのが一般的である [12]。しかし、人手評価にはコストと時間が必要という問題がある。一方、自動評価として主に用いられるのは BLEU [1] で、正解文との N-gram の一致度で応答を評価するものである。しかし、BLEU と人手評価にはまったく相関がないことが示されている [13]。原因の 1 つは one-to-many [2] として知られる、一つの発話に対

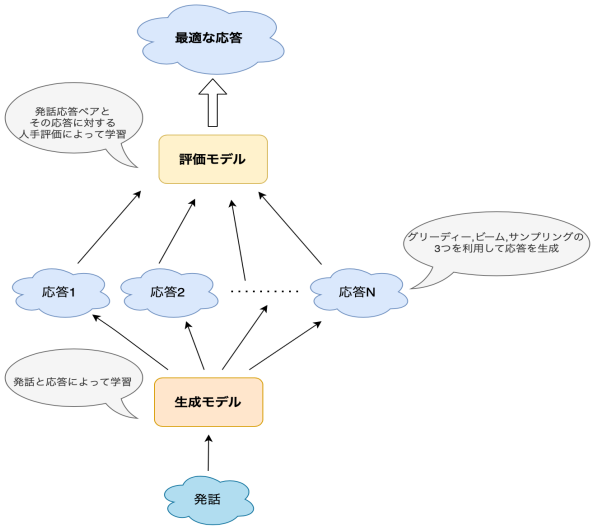

図 1　提案システムの概要図

し、複数の適切な応答が存在する対話の性質にある。この性質を考慮したとき、正解文との一致度を計るような手法は応答の適切な評価手法とはいえない。そこで、自動かつ、正解文を必要としない評価手法が研究されており、DNN を利用した手法がある。Zhao ら [3] や Ghazarian ら [4] は、BERT [6] を人手評価のデータセットで学習させることで、人手評価と相関がある自動評価ができるシステムを構築している。

3　提案手法: 応答生成・評価・選択システム

対話システムとして、応答を生成するモデルと評価するモデルの 2 つを利用し、応答を複数生成、それらの応答を評価、もっとも高い評価を得た応答を選択して出力するシステムを提案する。提案システムの概要図を図 1 に示す。応答を複数生成するために、グリーディーサーチ、ビームサーチ、サンプリングの 3 種類のデコーディング手法を用いる。特にサンプリングを繰り返すことで、多様な応答を得ることができる [14]。サンプリングは top-50 サンプリングを用いる。

応答生成モデルとしては Huggingface Transformers に登録されている sonoisa/t5-base-japanese を Fine-tuning したものを利用する。sonoisa/t5-base-japanese は、Raffel ら [5] にならって、日本語で事前学習が行われたモデルである。Fine-tuning 用の対話コーパスとして、Twitter から Twitter API を用いて我々が収集したツイート-リプライ対を利用する。学習に用いる対話ペア数は 800,000 ペアである。

表 1　応答評価コーパスのデータ数

評価観点	対話コーパス	ペア数
関連度	Twitter/Model	4,000/4,000
おもしろさ	Twitter	2,000
魅力	Twitter/Model	4,000/4,000
感情	Twitter	2,000

応答評価モデルについては 4 節で詳しく述べる。

4　応答評価モデルの構築

応答評価モデルの学習に用いるのに適切な日本語のコーパスは公開されていないため、まず応答評価コーパスを構築し、そのコーパスを利用して応答評価モデルを作る。

4.1　応答評価コーパス

クラウドソーシングを利用して、発話応答ペアに 1 から 5 の評価を付与する。評価観点は、対話システムの評価において一般的に用いられている関連度、おもしろさ、会話相手としての魅力の 3 つに、会話相手の感情に寄り添っているかを加えた 4 つである。会話相手の感情に寄り添っているか、については、対話システムが生成する応答はユーザの感情に寄り添うべきもので、それを評価したいため追加した。各観点は以下のようにして尋ねた。質問に含まれる A、B はそれぞれ、ある発言をした話者と、それに対する応答をした話者を表す。

- 関連度　「B の応答は A の発言に関係したものですか」
- おもしろみ　「B の応答はおもしろいですか」
- 魅力　「あなたが A だとして、B は会話相手として魅力的であり、B ともっと会話を続けたいですか」
- 感情　「B の応答は A の感情に寄り添っていますか」

発話応答ペアとしては、Twitter のツイート-リプライ対 (Twitter データセット) と、Twitter のツイート-応答生成モデルの出力対 (モデルデータセット) の 2 種類を利用する。モデルデータセットについては、1 つの発話に対して応答生成モデルが出力した、複数の応答に関して評価を集めることで、応答が異なれば評価が異なる様子を表現したデータセットとする。複数の応答は 3 種類のデコーディング手法を利用して得た。各観点のデータ数を表 1 に示す。

表 2　応答評価モデルの評価例

発話	応答	評価
聴いて良かったと心から思った。	聴いてるだけで心が落ち着く。	3.4
	はよ寝ろ w	2.8
	ありがとうございます。心からそう思えるのが一番嬉しいですね。	4.0
素足で踏んじゃった。	お腹がすいた。	2.9
	素足で踏んでるのが可愛い。	3.1
	素足で踏むのはダメですよー。	3.4

表 3　応答評価モデルの性能

評価観点	ピアソン	スピアマン
関連度	0.54	0.54
おもしろさ	0.19	0.19
魅力	0.47	0.49
感情	0.47	0.48

表 4　魅力の評価における応答評価モデルの出力と応答の表層的性質の相関

応答の表層的性質	相関係数
応答の長さ	0.34
発話との単語重複割合	0.17

データセットの例を付録 A に示す。

1 つの発話応答ペアに対して 5 人のクラウドワーカから評価を集め、その平均を評価値とする。

本研究においてクラウドソーシングはすべて、Yahoo!クラウドソーシングを利用して行った。

4.2　応答評価モデル

2 節の Zhao ら [3] にならい、BERT を応答評価コーパスで Fine-tuning することで、応答評価モデルを構築する。モデルは Huggingface Transformers の cl-tohoku/bert-base-japanese-whole-word-masking を利用する。評価値は BERT の出力の [CLS] に対応するベクトルを線型層に通してから、Sigmoid 関数に入力することで、0 から 1 の値を得て、それをスケール変換して 1 から 5 の連続値として得ている。

なお、対話システムによる応答の人手評価は複数の観点を設けるのが一般的であるが、本研究の応答生成・選択・評価システムにおいては、それら複数の観点の総合といえる、会話相手としての魅力を評価観点とする応答評価モデルを利用する。

4.3　応答評価モデルの評価

応答評価モデルの評価には、5 分割交差検証を行い、評価指標として、ピアソンの相関係数とスピアマンの順位相関係数を用いた。評価結果を表 3 に示す。

応答評価モデルが応答の長さ、発話との単語の重複割合のような、表層的な性質に基づいて評価をしていることが懸念されるため、それら表層的性質とモデルの出力の相関を調べる。テストコーパスは、学習に使ったものとは別に Twitter から収集したツイートと、そのツイートを応答生成モデルに入力したときの出力 14,000 ペアであり、評価観点は魅力である。ここでは相関係数はピアソンの相関係数である。結果を表 4 に示す。相関係数がそれほど高くないことから、応答評価モデルの評価が、それら表層的性質のみを手がかりにしていないことがわかる。

表 5　魅力の評価における応答評価モデルと応答の表層的性質による評価の性能比較

応答の評価手法	相関係数
応答の長さ	0.15
発話との単語重複割合	0.19
応答評価モデル	**0.47**

また、表層的な性質による評価をベースラインととらえ、人手評価との相関を調べ、応答評価モデルと比較する。テストコーパスは学習に用いたコーパスと同じ、4.1 節の対話評価コーパス 8,000 ペアで、評価観点は魅力である。ここでは相関係数はピアソンの相関係数である。応答評価モデルの相関係数は、5 分割交差検証によって得たものである。結果を表 5 に示す。結果から、表層的な性質による単純なベースラインに比べて、応答評価モデルによる評価がより正確に人間の評価を表現できることがわかる。

応答評価モデルによる評価の例を表 2 に示す。発話に関連していて、丁寧な応答がより高い評価を得ており、人間に近い評価ができていることがわかる。

表 6　発話「美味しいですよね。台湾茶。大好き。」に対する提案システムの応答例

グリーディー/ビーム	ありがとうございます。台湾茶は美味しいですよね。	3.6
サンプリング 1	美味しいですよね。台湾茶が食べたくなったので、今度試してみます。	3.8
サンプリング 2	美味しいですよね。味も飲みやすくて、 お値段の割にちょっぴり高級感があって良いですよね	**3.9**

表 7　提案システムの評価結果

システム比較	勝ち	負け	引き分け
提案システム vs グリーディーシステム	49%	27%	24%
提案システム vs ランダムシステム	50%	24%	26%

5　応答生成・評価・選択システムの評価

5.1　実験設定

提案手法の有効性を確かめるため、クラウドソーシングによる人手評価を行う。ある発話に対する、提案システムの応答とベースラインの対話システムの応答の 2 つをクラウドワーカに見せ、どちらの応答をする相手と会話を続けたいかを尋ねる。1 つの発話応答ペアについて 3 人に尋ね、多数決をとって結果とする。テストコーパスは応答生成モデルの学習に使ったものと同様の Twitter コーパスで、2,000 文である。ベースラインとしては以下の 2 種類を用いる。

- グリーディーシステム
  応答生成モデルがグリーディーサーチによる応答だけを生成するシステム
- ランダムシステム
  応答生成モデルが生成した複数の応答を評価せず、それらからランダムに選択した応答を出力とするシステム

ランダムシステムと提案システムの応答生成モデルでは、グリーディーサーチによる応答 1 つ、ビームサーチによる応答 1 つ、サンプリングによる応答 5 つの計 7 つの応答を生成して、そこから 1 つを選択して出力とする。

5.2　実験結果

実験結果を表 7 に示す。グリーディーシステム、ランダムシステムどちらのシステムとの比較においても、提案システムが勝利する結果となり、提案手法の有効性が示された。

提案システムによって出力される応答の例を表 6 に示す。発話「美味しいですよね。台湾茶。大好き。」に対して応答生成モデルから 7 個の応答を得たが、グリーディーサーチとビームサーチによる応答は同一であった。またサンプリングによる応答は 2 つだけを示している。複数の応答の中で、人間からの評価が高いと考えられる応答「美味しいですよね。味も飲みやすくて、お値段の割にちょっぴり高級感があって良いですよね」を選択できており、提案手法の有効性がわかる。

6　おわりに

本研究では、人手評価を学習データとして学習した応答評価モデルを対話システム内に組み込むことで、人間からより良いと判断される応答を出力できる対話システムを構築することを目指した。応答生成モデルが生成した複数の応答を応答評価モデルで評価し、最適な応答を選択して出力することで、グリーディーサーチのみを利用するシステムや、複数の応答からランダムに応答を選択するシステムより良い応答を出力できることがクラウドソーシングによる人手評価によって示された。

DNN による応答評価モデルは応答の評価のみならず、本研究のように応答の生成に応用することや、対話コーパスのフィルタリングに利用することなどもでき、今後もさらなる研究が求められる。そのために、評価に関するデータセットが大規模に収集され、公開されることが期待される。

また、本研究では応答生成モデルと応答評価モデルは完全に独立して学習しており、本研究の発展として、応答生成モデルと応答評価モデルの相互作用を利用して学習するような構造が考えられる。

謝辞

本研究は LINE 株式会社との共同研究の助成を受けて行った。

参考文献

[1] 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.

[2] Tiancheng Zhao, Ran Zhao, and Maxine Eskenazi. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. In **Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 654–664, Vancouver, Canada, July 2017. Association for Computational Linguistics.

[3] Tianyu Zhao, Divesh Lala, and Tatsuya Kawahara. Designing precise and robust dialogue response evaluators. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 26–33, Online, July 2020. Association for Computational Linguistics.

[4] Sarik Ghazarian, Johnny Wei, Aram Galstyan, and Nanyun Peng. Better automatic evaluation of open-domain dialogue systems with contextualized embeddings. In **Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation**, pp. 82–89, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[5] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. **Journal of Machine Learning Research**, Vol. 21, No. 140, pp. 1–67, 2020.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)**, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[7] Yuanlong Shao, Stephan Gouws, Denny Britz, Anna Goldie, Brian Strope, and Ray Kurzweil. Generating high-quality and informative conversation responses with sequence-to-sequence models. In **Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing**, pp. 2210–2219, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[8] Angela Fan, Mike Lewis, and Yann Dauphin. Hierarchical neural story generation. In **Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 889–898, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[9] Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. Wizard of wikipedia: Knowledge-powered conversational agents. In **International Conference on Learning Representations**, 2019.

[10] Tatsuya Ide and Daisuke Kawahara. Multi-task learning of generation and classification for emotion-aware dialogue response generation. In **Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop**, pp. 119–125, Online, June 2021. Association for Computational Linguistics.

[11] 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.

[12] Sarah E. Finch and Jinho D. Choi. Towards unified dialogue system evaluation: A comprehensive analysis of current evaluation protocols. In **Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue**, pp. 236–245, 1st virtual meeting, July 2020. Association for Computational Linguistics.

[13] Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In **Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing**, pp. 2122–2132, Austin, Texas, November 2016. Association for Computational Linguistics.

[14] Hannah Rashkin, David Reitter, Gaurav Singh Tomar, and Dipanjan Das. Increasing faithfulness in knowledge-grounded dialogue with controllable features. 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. 704–718, Online, August 2021. Association for Computational Linguistics.

表 8　応答評価コーパスの例

種類	発話	応答	評価
Twitter	Twitter に電話機能があったら良いのに・・・	ツイコールみたいなやつあった気が	4.2
Twitter	今日は地元	。さんもきっと気持ちがわかるんですね。	2.4
モデル	いいなあホテル暮らし	ホテル暮らしが一番いいよ。	4.2
モデル	いいなあホテル暮らし	いいなあ、私もホテル暮らしです。	2.2
モデル	いいなあホテル暮らし	ホテル暮らしはいいぞー。	4.4

A　応答評価コーパスの例

応答評価コーパスの例を表 8 に示す。