

実際の人々の感想を利用した雑談発話生成

成田 風香¹ 佐藤 志貴¹ 徳久 良子¹ 乾 健太郎^{2,1,3}

¹ 東北大学 ²MBZUAI ³ 理化学研究所

fuka.narita.q1@dc.tohoku.ac.jp

概要

雑談では、感想を述べる発話が対話の盛り上がり
に寄与することが知られている。我々は、文脈に対
して適切な感想を生成できる発話生成システムの実
現を目指し、対話の話題に対する実際の人々の感想
を外部情報として入力に用いることを試みる。本稿
では、我々がこれまでに構築した「感想付きニュー
ス雑談コーパス」を拡張した上で、システムの追加
学習を行うことで、人々の感想を利用して発話を生
成するシステムを構築した。生成発話の評価の結果、
文脈に対して自然であり、かつ感想を含む発話を
多く生成できることが分かった。加えて、これら
のシステムにより生成された発話は、雑談を盛り上
げるような発話であることが明らかとなった。

1 はじめに

近年、ニューラルネットワークをベースとした発
話生成システムの性能が向上している [1, 2, 3]。し
かし、ユーザと話すこと自体が目的である「非タ
スク指向型対話 (雑談)」に関する研究では、シス
テムの応答が受動的であることが指摘されており [4]、
場合によっては対話の盛り上がりへの寄与に欠けた
応答となることが考えられる。一方で、人同士の雑
談では感想を述べる発話が対話の盛り上がり
に寄与すること [5]、人はシステムと話すとき、相手
に人と同様の振る舞いを期待する傾向があること [6]
から、システムから感想を含む発話を生成すること
で、雑談をさらに盛り上げることができると考えら
れる。しかし、発話生成システムが適切でない感想
発話を生成することにより倫理面や安全性の観点で
問題が起きる危険性について指摘されている [7, 8]。

そこで我々は、倫理面や安全性の問題に配慮し
ながら自然な感想を発話するシステムを実現するた
め、実際の人々の感想を利用することを試みる。我
々はこれまでに、図 1 に示すような発話生成シス
テムの構築を目指し、対話の話題とそれに対する実

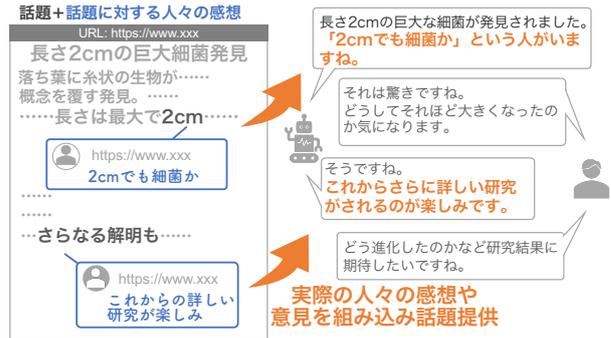


図 1 我々が構築を目指す発話生成システムの全体像。話題に対する人々の感想を外部情報として用いることで、感想を含む発話を生成する。

際の人々の感想、対話を組み合わせた「感想付き
ニュース雑談コーパス」を構築した [9]。本稿では、
構築したコーパスを大規模化した上で、学習済の発
話生成システムの追加学習を行った。そのシステム
に人々の感想や文脈を入力して生成した発話を、感
想発話の積極性や雑談の盛り上がりの観点で評価し
た。評価の結果、文脈のみから次の発話を生成する
従来法に比べて文脈に対して自然な発話を生成で
き、かつ感想を含む発話を多く生成できることが分
かった。加えて、これらのシステムにより生成され
た発話は、雑談を盛り上げるような発話となってい
ることが明らかとなった。

2 コーパスの構築

我々はこれまでに、ある話題に対する人々の感想
を用い、その話題について対話を行う図 1 のような
発話生成システムの実現を目指し、**感想付きニュー
ス雑談コーパス**を構築した。本コーパスは、人々の
感想としてツイート¹⁾を、ツイートの集まっている
話題としてニュース記事を用いることとし、「話題
(ニュース記事)」「感想 (話題に対するツイート)」「
対話」の三つ組みで構成されている²⁾。収録されて

1) 本稿では X への投稿をツイートと表現する。

2) 実際には、ニュース記事の URL、ツイート ID と対話の内
容を公開している。

いる対話は Wizard of Oz 法で行われたものである。一人はシステム役としてニュース記事の話題を投げかけ、その内容とそれに対する実際の人々の感想を提供しながら発話し、もう一人はユーザ役として自由に発話した。システム役の発話には、話題に対するツイート群の中からどのツイートが使用されたかを示す情報が付与されている。

以前の我々の研究 [9] でコーパスに収録された対話数は 425 対話であったが、より自然な発話生成の学習を行うため対話数を増やし、47 件のニュース記事に対して、11,256 発話から成る 1,005 対話とした。1 対話あたりの発話数は 11.2 発話、1 対話あたり使用されたツイート数は 2.8 個である。

3 実験設定

我々が構築したコーパスを用いることで、人々の実際の感想を用いて発話するよう学習されたシステムが、倫理的な問題を含まない自然な感想を積極的に生成できるかを検証した。さらにそれらの発話が雑談を盛り上げることにに対し効果的かを検証した。

3.1 コーパスの分割

評価実験にあたり、構築したコーパスを学習データ 837 対話、開発データ 72 対話、評価データ 96 対話に分割した。それぞれのデータセット間の独立性を保てるよう、各データセットに含まれるニュース記事には重なりがないようにした。

3.2 発話生成システムの学習と推論

本評価実験では、評価データ中の対話に含まれるシステム役の特定の発話を**参照応答**、参照応答に先行する発話の系列を**対話文脈**、またコーパス構築において同対話がなされたときにシステム役に提示されていたニュース記事全文およびツイート群を**外部情報候補**と呼ぶ。発話生成システムに参照応答直前の対話文脈 3 発話や外部情報候補の 1 つを入力し、対話文脈直後の発話として生成された発話の質を評価した。生成発話の評価には、以下の 4 個の発話生成システムを用いた。下記の (1) はベースラインシステムで、(2)~(4) は感想付きニュース雑談コーパスを用いて追加学習したシステムである。

(1) baseline-model. 対話文脈のみの入力ですべて自然な発話生成を行うベースラインシステムとして、Sugiyama らが公開している `japanese-dialog-transformers` [10] を用いた。本システムには評価デー

タの対話文脈のうち直近 3 発話のみを入力として与え、続く発話を生成させた。

(2) oracle-w/NT. 人手で選ばれた文脈に適切な外部情報を入力して生成される発話を評価するため、本発話生成システムを用いた。外部情報候補の中から、コーパス構築時にシステム役が参照応答に用いた外部情報を対話文脈と連結して入力し、文脈に続く発話を生成するよう学習して本システムを構築した。具体的には、参照応答がいずれかのツイートを使用して発話されていた場合、そのツイートと対話文脈を連結させたものを入力とした。いずれのツイートも使用されていない場合、参照応答がニュース記事の一文を参考していると推測される場合はその文³⁾を対話文脈と連結して入力とし、それ以外の場合は対話文脈のみを入力とした。このようにして決定した入力に対し、参照応答を出力として `baseline-model` を追加学習することで本システムを構築した。生成時は、学習時と同様に、参照応答に応じて外部情報を決定し発話を生成した。

(3) oracle-w/N. 本システムでは、入力にツイートを使用せず、発話を生成した。本システムの生成発話を `oracle-w/NT` の生成発話と比較することでツイートを入力することの有効性を確認した。ツイートを使用した発話を参照応答とするデータを学習データから除いた点以外は `oracle-w/NT` と同様に本システムを構築した。生成時もツイートを使用せず、参照応答に応じて、外部情報として用いるニュース記事の一文を決定した。

(4) select-w/NT. (2), (3) のシステムでは、参照応答を利用して入力する外部情報を決定し発話を生成したが、実際の対話の場面では参照応答は存在しない。そこで本システムでは、システム自身が外部情報候補から文脈に適切なものを選択して発話を生成し、`oracle-w/NT` による生成発話と比較して、参照応答がある場合と参照応答がない実用時を想定した場合の差を検証した。本システムは `BlenderBot 3` [11] に倣い、特殊トークンを付与することで、外部情報の選択と発話の生成をどちらも生成問題として同一の系列変換器に学習させた。外部情報の選択の学習時には、(2) で述べた手法で選択される外部情報と対話文脈を入力した場合の出力を '1'、選択されない外部情報と対話文脈を入力した場合の出

3) 具体的には、対話の話題となっているニュース記事本文を `ja-sentence-segmenter` (<https://pypi.org/project/ja-sentence-segmenter/>) で分割し、参照応答と類似度の高い 1 文を発話に用いた外部情報として決定した。

力を‘0’として baseline-model の追加学習を行った。発話生成の学習時は、oracle-w/NT で述べた手法で選択される外部情報と対話文脈を連結して入力し、参照応答を出力とする学習を行った。評価データの発話生成は、1. 入力する外部情報の選択、2. 選択された外部情報を利用した発話生成の2段階からなる。外部情報選択時には、外部情報候補を文脈と合わせて1つずつ入力し、‘1’が出力された外部情報の中から1つを外部情報として選択した⁴⁾。評価データの発話生成時は、選択された外部情報を対話文脈と連結して入力し、生成した。

3.3 人手による発話の評価

評価データ中のシステム役の発話 100 個を無作為抽出して参照応答とし、それらの対話文脈と、文脈に続く4つのシステムからの生成発話をアノテータに提示し評価させた⁵⁾。評価の観点と集計方法を以下に示す。

(1) 発話の自然性。 感想の発話のために外部情報を用いることで自然性が損なわれないか確認するため、文脈に続く発話としての自然性を1(不自然)、2(捉え方によっては問題ない)、3(自然)の3段階で絶対評価をさせ、平均値を算出した。

(2) 倫理性。 感想を含む発話には倫理的な問題を含みやすいと考えられる。それぞれの発話に問題があるかどうかを評価させ、少なくとも1人が問題ありとした発話を倫理的な問題のある発話とした。

(3) 盛り上がり。 外部情報を入力に加えるシステムが雑談の盛り上がりにも効果的かどうかを調べるため、4つのシステムの生成発話をランク付けにより評価させた⁶⁾。そのランクを一対比較に変換し、それぞれのシステムの勝率を計算した⁷⁾。

さらに、生成された発話が感想を含んでいるかを著者らにより評価した。

4) 優先順位をツイート、ニュース記事とし、複数の候補に対して‘1’が出力された場合はその生成尤度が最も高いものを選択した。外部情報候補のいずれに対しても‘1’が出力されなかった場合は、外部情報を入力に加えないこととした。

5) アノテータは研究室の学生3人であり、4つのシステムの特徴は公表せず、無作為な順序で生成発話を提示した。

6) 後の雑談が盛り上がりそうな順にランク付けするよう指示し、2つ以上の発話の順位を同率にすることを許容した。

7) 具体的には、あるシステムAとBの発話を比較するとき、まず各ワーカーのランキングから、システムAとBの発話の優劣関係を取り出し、上位のシステムを+1点、下位のシステムを-1点とした。この点数を3人のアノテータで合計し、0点を上回る場合は勝ち、0点の場合は引き分けとした。以上を全組み合わせ、全100個の文脈に対する発話を行うことで、それぞれのシステムの勝率を算出した。

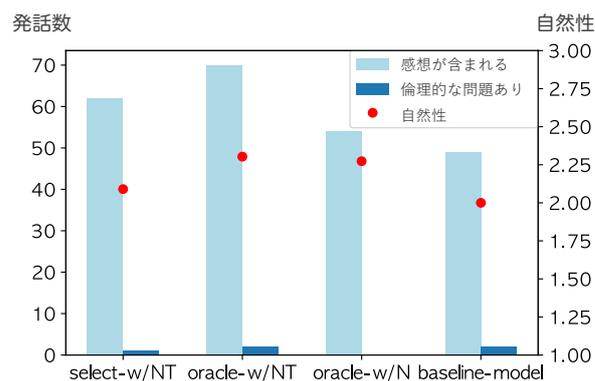


図2 感想発話数と倫理的に問題がある発話の数および各システムの自然性の評価値の平均。

4 実験結果

4.1 感想を生成する能力に関する評価結果

本研究では、雑談を盛り上げるために積極的に感想を発話する雑談発話生成システムの構築を目標としているが、その感想は自然かつ倫理的な問題がないことが望まれる。各システムの生成発話のうち感想が含まれていた数、生成発話の自然性に対する3段階評価の平均値、倫理的に問題があるとされた発話の数を図2に示す。

感想発話の生成に対する積極性。 発話生成時、外部情報としてツイートを使用した select-w/NT と oracle-w/NT は、ツイートを_usingしていない oracle-w/N や baseline-model と比較して感想を含む発話が生成された数が多かった。特に、oracle-w/NT と、oracle-w/N, baseline-model における頻度を比較したところ、統計的に有意な差が得られた⁸⁾。以上から、他者の感想を入力することで感想を含む発話が積極的に生成でき、特にその感想の選択において参照応答を利用した場合にはより積極的に感想を含む発話を生成できた。

自然性。 発話の自然性に関する3段階評価の平均値は、baseline-model は1.99、oracle-w/NT は2.30、oracle-w/N は2.28、select-w/NT は2.07であった。参照応答をもとに入力する外部情報を定めた oracle-w/NT, oracle-w/N の生成発話は、baseline-model の生成発話と比べて自然であった。また、入力する外部情報をシステム自身が選択した select-w/NT の生成発話の自然性に関する評価値はベースラインと

8) カイ二乗検定を行った結果、oracle-w/N との比較において、 $\chi^2(1) = 8.3, p = 0.0040$ 、baseline-model との比較において、 $\chi^2(1) = 4.8, p = 0.029$ となった。

比べて同程度であった。以上から、他者の感想を外部情報として発話生成に用いてもシステム発話の自然性は損なわれず、特に、参照応答をもとに、入力する外部情報を定めて発話生成を行った場合には自然性が向上した。

倫理性. 倫理的に問題があるとされた発話の数は、100 発話中、baseline-model は 2 発話、oracle-w/NT は 2 発話、oracle-w/N は 0 発話、select-w/NT は 1 発話であった。感想の積極的な発話により、倫理的問題を含む発話の生成頻度が高くなると考えられるが、感想の生成にツイートを用いる select-w/NT や oracle-w/NT では、倫理的に問題があるとされた発話の数はベースラインと同程度に低かった。問題があるとされた発話の例とその詳細を付録 A に示す。

以上の 3 点から、参照応答をもとに外部情報を入力して発話生成を行う場合、自然性と倫理性を損なわずに、積極的に感想を含む発話を生成できた。このことから、適切な外部情報を使用するシステムは対話を盛り上げることができると考えられる。select-w/NT はベースラインと同程度の自然性であったが、自然性や感想発話の生成に対する積極性は oracle-w/NT を下回ったことから、適切な外部情報の選択手法には改善の余地があると考えられる。

4.2 雑談の盛り上がりへの寄与

盛り上がりの評価に関する結果を表 1 に示す。ニュース記事とツイートを入力して発話を生成した oracle-w/NT、select-w/NT は、ツイートを利用しない oracle-w/N、baseline-model 両方との比較で勝率が 0.5 を超えた。特に、参照応答をもとに外部情報を選択して入力し、生成した oracle-w/NT は、勝率の平均値が最も高くなった。このことから、他者の感想を外部情報として適切に選択し、積極的に感想を生成可能な発話生成システムを実現することで、雑談を盛り上げる発話が生成できると考えられる。

4.3 発話生成例

表 2 に、各システムの発話生成例を示す。select-w/NT、oracle-w/NT では、入力されたツイートを利用した具体性の高い感想の生成が見られた。この対話において、3 人のアノテータによる盛り上がりに関する評価結果は、高い順に select-w/NT、oracle-w/NT、oracle-w/N、baseline-model となった。以上の例からも、実際の人々の感想を外部情報として選択し、発話生成に利用することが、雑談の盛り上がり寄与

表 1 発話生成システムの盛り上がり評価の勝率

	baseline-model	oracle-w/NT	oracle-w/N	select-w/NT	平均値
baseline-model	-	0.24	0.30	0.29	0.28
oracle-w/NT	0.76	-	0.59	0.57	0.64
oracle-w/N	0.70	0.41	-	0.46	0.52
select-w/NT	0.71	0.43	0.54	-	0.56

表 2 感想を含む生成発話の例.

	入力された外部情報	生成発話
文脈		
S:	はじめまして宜しくお願いします	
U:	こちらこそよろしく宜しくお願いします	
S:	ハイブリッドの記事読みましたでしょうか?	
U:	はい。読みました。どう感じましたか?	
S:	日本のハイブリッド車の売上げに影響するのかなと思います販売できないとなると販売戦略もなにも無いですね	
U:	そうですね。もう少し考えて欲しいですね。	
select-w/NT	[TWEET] さぞクリーンな発電方法を推進するんやろなあ	記事に付されたコメントでは「ぞクリーンな発電方法を推進しとるんやろなあ」という意見もありました。
oracle-w/NT	[TWEET] あと 13 年。他の州はどうするのかな。	あと 13 年。他の州はどうするかうするのかな。
oracle-w/N	[NEWS] 他の州がカリフォルニア州の規制に制にならうことも「多ならうことも許されてく」の州で新たな規制あり、carb の担当者は案を採用する動きがある」と話している	カリフォルニア州の規制にならうことも「多ならうことも許されてく」の州で新たな規制案を採用する動きがいです
baseline-model	なし	販売戦略もなにも無いですね。

する発話の生成につながると考えられる。

5 おわりに

我々の構築した「感想付きニュース雑談コーパス」を用いて人々の感想を使って感想を含む発話を生成するシステムを構築し、生成された発話を評価した。その結果、構築したシステムは文脈に対して自然な発話が生成でき、かつ感想を含む発話を多く生成できること、それらの発話は、雑談を盛り上げるような発話となっていることが明らかとなった。

今後の展望として、コーパス構築時は人手で行った、対話に使用する感想の絞り込みの自動化や、ニュース以外の話題の使用により、より広いドメインに対応した発話生成システムの構築を目指す。

謝辞

本研究のコーパス構築には京都大学・児玉貴志氏らが開発された Chat Collection Framework を改変したものを用いました。本システムの開発者の皆様に深謝いたします。また、Yahoo!クラウドソーシングにクラウドワーカーとしてご協力いただいた皆様へ、深く感謝を申し上げます。本研究は CREST, JSPS 科研費 JP21J22383 の助成を受けたものです。

参考文献

- [1] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020.
- [2] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022.
- [3] Amelia Glaese, Nat McAleese, Maja Trębacz, John Aslanides, Vlad Firoiu, Timo Ewalds, Maribeth Rauh, Laura Weidinger, Martin Chadwick, Phoebe Thacker, Lucy Campbell-Gillingham, Jonathan Uesato, Po-Sen Huang, Ramona Comanescu, Fan Yang, Abigail See, Sumanth Dathathri, Rory Greig, Charlie Chen, Doug Fritz, Jaume Sanchez Elias, Richard Green, Soňa Mokrá, Nicholas Fernando, Boxi Wu, Rachel Foley, Susannah Young, Iason Gabriel, William Isaac, John Mellor, Demis Hassabis, Koray Kavukcuoglu, Lisa Anne Hendricks, and Geoffrey Irving. Improving alignment of dialogue agents via targeted human judgements, 2022.
- [4] Satoshi Akasaki and Nobuhiro Kaji. Conversation initiation by diverse news contents introduction. In **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), Volume 1 (Long and Short Papers)**, pp. 3988–3998, 2019.
- [5] 徳久良子, 寺寫立太. 雑談における発話のやりとりと盛り上がりの関連. 人工知能学会論文誌, Vol. 21, No. 2, pp. 133–142, 2006.
- [6] Clifford Nass and Youngme Moon. Machines and mindlessness: Social responses to computers. **Journal of Social Issues**, Vol. 56, pp. 81–103, 2000.
- [7] Dian Yu and Kenji Sagae. Automatically exposing problems with neural dialog models. **CoRR**, Vol. abs/2109.06950, , 2021.
- [8] Jing Xu, Da Ju, Margaret Li, Y-Lan Boureau, Jason Weston, and Emily Dinan. Recipes for safety in open-domain chatbots. **CoRR**, Vol. abs/2010.07079, , 2020.
- [9] 成田風香, 佐藤志貴, 徳久良子, 乾健太郎. 感想付きニュース雑談対話コーパスの構築. 言語処理学会第29回年次大会, pp. 131–135, 2023.
- [10] 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, 2021.
- [11] 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, 2022.

A 倫理的な問題のある発話の具体例

ニュース記事やツイートを入力して発話を生成する oracle-w/NT, select-w/NT による生成発話 100 発話の中で、それぞれ 2 発話、1 発話が倫理的に問題があると評価された。表 3 に実際の生成発話を示す。これらの具体例について、oracle-w/NT による生成発話は、ツイートを利用した発話ではあるが、ツイートがもとなった部分に対して倫理的に問題があると判断されたと推測される。select-w/NT による発話の大部分はツイートに基づくものだったが、最後の部分にツイートの記述と異なる内容が生成され、その部分に対して問題を感じられたと推測される。外部情報を用いて発話を生成する場合、外部情報候補の適切な選択や、外部情報に基づかない情報の出力の抑制が重要であると考えられる。

表 3 倫理的に問題があるとされた発話の例。

	入力された情報	生成発話
oracle-w/NT	[TWEET] 航続距離 10km ってピョンと飛んですぐ落ちる、バツタみたいなもんですな	この技術が軍事力に使われないことを願うばかりです。航続距離 10 キロってピョソと飛んですぐ落ちる、バツタみたいなもんですよ。
select-w/NT	[TWEET] 開発費の 96 億は、それくらいかかるかなとは思う。ドローンでもそうだが、関連法規の整備が遅れ、事業の立ち上げが難しい中で、とりあえず万博というイベントを口実に既成事実を作って、ビジネスの道筋を作りたいってことではないか。	開発費の 96 億は、それくらいかかるかなとは思う。ドローンでもそうだが、関連法規の整備が遅れ、事業の立ち上げが遅れる中で、とりあえず万博というイベントを口実に既成事実を作って、不倫関係に発展させるってことだな。