過去の対話セッションを考慮した雑談対話システム

髙﨑 環 [1]　佐藤 翔悦 [2]　吉永 直樹 [2]　豊田 正史 [2]
[1] 東京大学大学院 情報理工学系研究科　[2] 東京大学 生産技術研究所
{takasa-m,shoetsu,ynaga,toyoda}@tkl.iis.u-tokyo.ac.jp

概要

スマートスピーカーの普及が進み，対話システムとユーザとの雑談が継続的に行われるようになりつつある．そのような環境下では，これまでに行った全ての対話の内容を踏まえて，ユーザの知識や情報を把握した上で応答を行うことが望ましい．しかし，現状の深層学習に基づく雑談応答生成では，モデルの系列長に制限があるため，限られた対話履歴しか入力することができない．本研究では，膨大な過去の対話履歴の中から直前の対話履歴と関連度の高い対話や単語を離散的に選択し，入力に加えることの有効性を検証する．実験では，Twitter 上の対話ログを用いて GPT-2 をファインチューニングする際に，選択された過去の対話履歴を入力に追加して訓練を行い，未知ユーザ同士の対話において提案システムの有効性を評価する．

1　はじめに

スマートスピーカーの普及が進み，雑談対話システムとの会話を継続的に行う機会が増えつつある．その際に，システムは同一ユーザとの過去の対話の内容を踏まえ，過去の類似した話題に関する情報や，個人特有の知識などを考慮した応答を生成することが望ましい．具体的には，過去の対話内容を踏まえて話題を展開したり（図 1），過去に話した内容を繰り返さないような応答を生成することが重要であると考えられる．また，言葉遣いや特有の話題など，ユーザの嗜好に即した対話を展開することが期待される．

雑談対話システムにおいて，過去の対話履歴を活用する研究は広く行われてきた．SNS 上の大規模対話ログを模倣するように学習させる深層学習モデルが登場した当初は，直前の 1 文のみを発話履歴として入力する手法 [1] が提案された．その後，複数文の発話履歴を入力する研究 [2] が登場し，より効果的な学習をするために，進行中の対話セッションの履歴について，複数の発話をトークン単位・発話単位の 2 階層に分けて処理を行う研究 [3] が提案された．しかし，アーキテクチャの階層化にも限度はあり，対話履歴が膨大な場合は全てを処理することは難しい．また，これまでの研究では，現在の対話セッションを超えるような長い対話履歴をモデルに入力した際に，応答生成に及ぼされる効果について，十分に検証されていない．

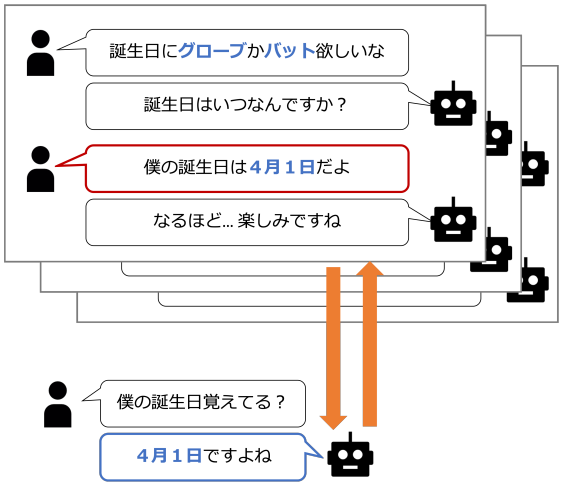

図 1　過去の対話セッションに手がかりを得た応答生成

本研究では，雑談対話システムの応答生成において，直前までの対話セッションの履歴に加え，過去にそのユーザと話した対話セッションから得られる情報を付与することの有用性を検証する（図 1）．モデルの入力長制限の問題を解決するために，直前の会話履歴をクエリとして関連度の高い過去の対話の選択や，ユーザ特有の単語の抽出を行うことで，大規模な対話セッションの情報から重要な情報を離散的に選択した．

実験では，Twitter から収集された対話ログを使用し，GPT-2 [4] のファインチューニングを行なった．訓練データに登場していない未知のユーザ同士の対話を用い，自動評価と人手評価によって提案手法を検証した．

2 関連研究

大規模な対話ログからの学習に基づく雑談対話システムでは，運用を通じて獲得できる対話履歴を発話の改善に使用したり [5, 6]，対話履歴から知識を獲得したり [7, 8] する研究が行われてきた．

Hancock ら [5] は，雑談対話モデルの運用中に収集した対話ログを用いてモデルを再訓練し，応答性能を継続的に改善する研究を行なった．この研究では複数人のユーザからの対話ログを収集するが．対話ログを共有する際は，プライバシーの観点で，使用可能な情報に制限がかかる．一方，本研究では個人の対話の中で動的に生の対話ログを参照するため，より多くの情報を使用できると考えられる．

現在の対話セッションと異なる対話の情報を利用して応答を生成する研究として，Examplar を用いたものがある，Pandey ら [6] は，学習データから類似した対話を TF-IDF をもとに検索し，応答生成に使用することで，より情報量が多く多様な応答が生成できたと主張した．この研究では，ユーザを制限せず追加する対話セッションを選択するが，本研究では特定ユーザ間での対話セッションに制限する．

Mazumder ら [7] は，知識ベースを参照した質問応答において，知識ベースにない情報をユーザに問い，未知の知識を更新し，以降の応答生成に使用する手法を提案した．この手法では限られた状況での質疑応答を対象としており，応答はトリプレットを基に作成される簡潔なものである一方で，本研究で扱うのは雑談応答生成モデルである．

Wu ら [8] の研究では，対話履歴から個人特有の情報をトリプレットで獲得する．この研究の知識獲得の対象はユーザ属性のみであるが，継続的にユーザ特有の知識を蓄積し，動的に参照する対話システムの研究に有用であると主張している．本研究においても，継続的に獲得可能な対話履歴から個人に有用な情報を入力する手法をとっており，関連性が高い．一方で，本研究の場合は個人情報の使用にとどまらず，話題の繰り返しを避けるなど，より広範な問題をターゲットにしている．

3 提案手法

3.1 課題

本研究では，雑談対話システムを大規模コーパスで訓練する際に，進行中の対話セッション内の履歴だけでなく，過去の対話セッションの履歴も使用する手法を提案する．この手法では，特定の 2 ユーザ間の対話をユーザと雑談対話システムと見做し，その過去の対話を継続的に行われた対話履歴として使用する．

膨大な対話履歴を扱う必要がある一方で，高性能なモデルアーキテクチャには，通常は系列長の制限がある．入力される系列長に対して線形以上のオーダの計算量が必要である以上，長い系列長は短縮して入力しなければならない．また，全ての過去の対話履歴が，現在の対話セッションにおける発話生成に有効ではないため．モデルの最適化が難しくなる可能性がある．そこで，既定の検索手法を用いるなどの帰納バイアスによって，入力長の圧縮を行う手法が考えられる．

3.2 対話セッションの選択・圧縮

入力する過去の対話セッションを選択または圧縮するために，本研究では二つの手法を提案する．一つ目は直前の対話履歴をクエリとして，関連する過去の対話セッションを選択する手法（図 2）で，話題の繰り返しを避けるなど，過去の対話を踏まえた対話を行うことを期待する．二つ目は話者特有の情報を抽出する手法（図 3）で，話者の口癖や頻繁に出す話題など，個人特有の情報を考慮することを狙いとしている．

3.2.1 現在の対話と関連する過去の対話履歴の選択

継続的な会話を行う際，対話セッションを跨ぐような文脈も考慮し，同じ話題を繰り返さないなど，長距離文脈を適切に考慮した応答を生成することが望ましい．しかし，膨大な履歴を全て入力するのは困難なため，過去の対話セッションのうち，進行中の対話と関連性の高いものを選択する必要がある．これを実現するために，同じ 2 ユーザ同士の過去の対話セッションをベクトルで表現し，進行中の対話セッションとのコサイン類似度が最も高いものを選択する．対話セッションのベクトルは，出現単語について，その対話セッション中の出現回数 TF と，同ユーザ間の全対話セッション中での単語の希少性 IDF を計算し掛け合わせることで算出され，語彙毎のスコアを用いて表現される．

過去の対話セッションを選択後，現在の対話セッションの履歴とともにモデルに入力する．，現在の対話セッションの履歴のうち i ターン目の発話を

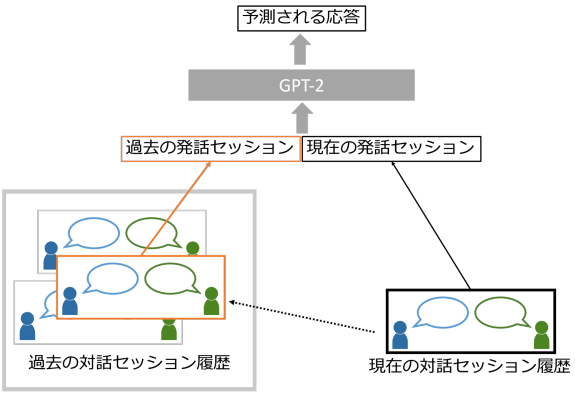

図 2　過去の対話セッションを選択・入力するモデル

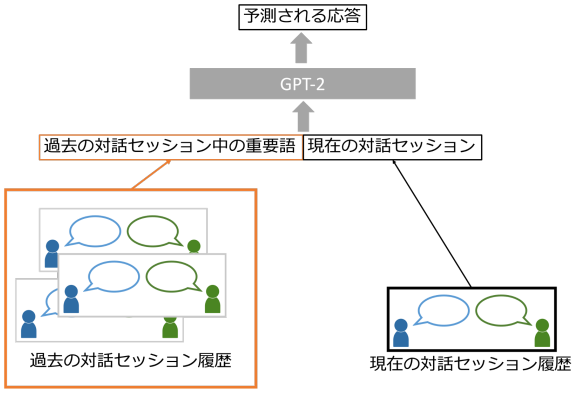

図 3　ユーザ特有の重要語を入力するモデル

$\boldsymbol{u}_i = (u_{i,1}, u_{i,2}, ...)$，選択された対話セッションのうち i ターン目の発話を $\boldsymbol{h}_i = (h_{i,1}, h_{i,2}, ...)$ としたとき，モデルへの入力は，

$$\texttt{[HEAD]}h_{1,1}...h_{m,l_m}\texttt{[EOH]}u_{1,1}...u_{n,l_n}\texttt{[EOS]}$$

とした．ここでの [EOH] トークンは，過去の対話セッション履歴の終了を表すトークン，[EOS] トークンは，現在の対話セッション履歴の終了を表すトークンとする．また，[HEAD] トークンは二種類存在し，最初の話者がユーザ側かシステム側かによって使い分けることで，過去の対話セッション上の話者情報を提供する．

3.2.2　ユーザ特有の重要語の抽出

この手法では，特定ユーザ同士の過去の対話セッションの全てから，話者特有の重要語を抽出し．モデルの入力に加える（図 3）．Fikri ら [9] は，生成される応答の多様性を高めるために，ユーザのツイートから抽出される重要語を入力に加えているが，本手法では，過去の対話セッションから抽出された重要語を入力に加えることで，個人特有の話題や語彙を考慮した応答が生成されることを期待する．

そこで，あるユーザの対話履歴の各単語について，重要度を算出する．具体的には，特定の 2 ユーザ間での対話セッション履歴を全て結合し，一つの文章と考え，訓練データ中での TF-IDF 値を計算する．そして，2 ユーザ間での対話において，TF-IDF 値が高い単語を重要語とみなし，特定個数選択する．この個数については，対話履歴の単語の 25%以下かつ，50 個以下となるようにした．選択された重要語を $\boldsymbol{w}_i = (w_{i,1}, w_{i,2}, ...)$ とした時，モデルの入力は，$h_{1,1}...h_{m,l_m}\texttt{[EOW]}u_{1,1}...u_{n,l_n}\texttt{[EOS]}$ とした．ここでの [EOW] トークンは，重要語の終端を表すトークンである．

4　実験設定

本研究の提案手法とベースラインに対して自動評価と人手評価を行い，応答生成の性能を比較した．

4.1　データセット

本研究では，我々の研究室で Twitter API[1] を用いて収集した Twitter データセットから対話データセットを構築した．Twitter 上でのリプライツリーを一連の対話セッションとみなし，2 人のユーザが交互に発話する対話のみを使用した．このうち，bot との対話や，画像データや URL を含むものを除外し，同一ツイートから枝分かれしている対話は最長のもののみを使用した．また，対象ユーザは過去の対話セッションが 2〜9 件存在するものに限定した．今回は，2018 年のデータを訓練・検証データ，2019 年のデータを評価データとし，訓練データに出現したユーザを評価データから除外した．最終的に，訓練・検証・評価データの特定のユーザペアの数はそれぞれ 957,799 件，50,411 件，10,901 件であった．最も新しい対話セッションを進行中の対話，それ以外を過去の対話セッションとし，1 組のユーザペアあたり 1 つの発話応答のサンプルを作成した．人手評価の際には，アノテータが現在の対話セッションから必要な情報を推測しやすいように，未知語を含む対話や，ツイートへのリプライで開始されるものを除外した．また，コストの問題から，評価データのうち対話セッションを抽出した際の類似度が 0.20 以上のサンプルから 100 件，それ未満のサンプルから 100 件選択したデータセットを使用した．抽出時の類似度別に評価を行い，発話履歴に類似した対話履歴が応答生成により効果的であるかを検証する．

1）　https://developer.twitter.com/en/docs

	BLEU-2	dist-1	dist-2	平均長
GPT-2	2.77	6.07	**29.52**	12.16
+ 対話セッション	**3.41**	**6.10**	29.49	**12.55**
+ 重要語のみ	3.14	5.99	29.04	12.31
参照応答	-	8.86	41.98	16.01

表 1　自動評価の結果

	ゲイン (%)	95%信頼区間
全サンプル	12.83	[1.579, 24.08]
類似度が低いサンプル	11.66	[-16.04, 39.37]
類似度が高いサンプル	14.31	[7.674, 20.94]

表 2　人手評価の結果

4.2　モデル

先述した提案手法をもとに，GPT-2 [4] の日本語版事前学習済みモデルである，rinna/japanese-gpt2-small [2] を最大 3 エポック分ファインチューニングした．入力を直前までの対話履歴のみとし，同様にファインチューニングを行なったモデルをベースラインとした．モデルの実装には，huggingface/transformers [10] ライブラリを用いた．評価時の生成応答を多様にするため，累積確率を基にした top-p [11] サンプリングを採用した．この時の p の値は 0.9 とした．また，対話セッションの選択や自動評価の際のトークン化には MeCab [12] を用い，GPT-2 のファインチューニングの際には，rinna/japanese-gpt2-small の事前学習で使用されるトークナイザーによってトークン化を行った．

4.3　評価尺度

生成応答に対し，自動評価および人手評価を行う．自動評価指標としては，参照応答との類似度を測定する BLEU [13]，応答の多様性を測定する dist-n [14]，応答の平均長を用いた．

人手評価では，過去の対話セッションを選択する提案手法とベースライン間での性能比較を行った．3 人のアノテータ（著者を含まない大学院生）が，直前までの発話履歴と選択された過去の対話セッション履歴を踏まえた上で，ベースラインと提案手法のどちらの応答が妥当かを選択した．その上で，提案手法による応答が妥当と選択された問題数が，妥当でないと選択された問題数に対してどの程度大きいかを示すゲインを測定し，95%信頼区間を求めた．

5　実験結果

5.1　自動評価

表 1 に自動評価による結果を示す．過去の対話セッションの情報を付与した 2 つの提案手法はベースラインに比べ，高い BLEU-2 を達成した．提案手法はより参照応答に似た応答を生成すると考えられる．生成応答の単語の多様性を測定する dist-2 における評価では，いずれの提案手法でもベースラインを若干下回った．また，提案手法が生成する応答はトークン数が若干多かった．これは，過去の対話の情報にのみ登場する，音符や三点リーダなどの語尾やユーザの名前といった追加情報を取り入れ，応答を生成する傾向が観測されたことと，関連があると考えられる．

5.2　人手評価

表 2 に人手評価による結果を示す．表に掲載されている全てのゲインが正であることから．提案手法がベースラインに比べ，妥当な回答だと選択されていることがわかる．また，提案手法において対話セッションを選択した際に，現在の対話との類似度が低いサンプルについての性能に比べ，類似度が高いサンプルの方が性能が高いことがわかる．この結果から，現在の対話に関連した過去の対話の情報が，応答生成に有用な情報を与えていると考えられる．なお，3 人のアノテータの評価について，フライスのカッパ係数を用いて相関を測定したところ，0.3866 とやや低い値となった．これは，挨拶のやりとりなど，過去の対話セッションを用いても応答に差がつきにくく，妥当性の判断が難しい対話が含まれていたためと考えられる．

6　おわりに

本研究では，膨大な過去の対話セッションから有用な情報を離散的に選択し，応答生成を行う雑談対話システムを提案した．進行中の対話セッションの履歴に加え，過去の対話セッションから関連性の高いものを選択，もしくは重要語を抽出し，GPT-2 のファインチューニングの入力とした．実験では，Twitter 上の対話ログを用いて訓練を行い，生成された応答に対し自動評価および人手評価によって，過去の対話履歴が発話生成に及ぼす効果を検証した．

2）　https://huggingface.co/rinna/japanese-gpt2-small

謝辞

この研究は国立情報学研究所（NII）CRIS と LINE 株式会社とが推進する NII CRIS 共同研究の助成を受けています．また，人手評価にご協力くださった同研究室のメンバーに深く感謝申し上げます．

参考文献

[1] Lifeng Shang, Zhengdong Lu, and Hang Li. Neural responding machine for short-text conversation. In **Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)**, pp. 1577–1586, Beijing, China, July 2015. Association for Computational Linguistics.

[2] Alessandro Sordoni, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao, and Bill Dolan. A neural network approach to context-sensitive generation of conversational responses. In **Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 196–205, Denver, Colorado, May–June 2015. Association for Computational Linguistics.

[3] Iulian V. Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. Building end-to-end dialogue systems using generative hierarchical neural network models. In **Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence**, AAAI'16, p. 3776–3783. AAAI Press, 2016.

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

[5] Braden Hancock, Antoine Bordes, Pierre-Emmanuel Mazare, and Jason Weston. Learning from dialogue after deployment: Feed yourself, chatbot! In **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics**, pp. 3667–3684, Florence, Italy, July 2019. Association for Computational Linguistics.

[6] Gaurav Pandey, Danish Contractor, Vineet Kumar, and Sachindra Joshi. Exemplar encoder-decoder for neural conversation generation. In **Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 1329–1338, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[7] Sahisnu Mazumder, Bing Liu, Shuai Wang, and Nianzu Ma. Lifelong and interactive learning of factual knowledge in dialogues. In **Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue**, pp. 21–31, Stockholm, Sweden, September 2019. Association for Computational Linguistics.

[8] Chien-Sheng Wu, Andrea Madotto, Zhaojiang Lin, Peng Xu, and Pascale Fung. Getting to know you: User attribute extraction from dialogues. In **Proceedings of the 12th Language Resources and Evaluation Conference**, pp. 581–589, Marseille, France, May 2020. European Language Resources Association.

[9] Abdurrisyad Fikri, Hiroya Takamura, and Manabu Okumura. Stylistically user-specific response generation. 自然言語処理, Vol. 28, No. 4, pp. 1116–1140, 2021.

[10] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing. In **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations**, pp. 38–45, Online, October 2020. Association for Computational Linguistics.

[11] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In **8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020**. OpenReview.net, 2020.

[12] T. KUDO. Mecab : Yet another part-of-speech and morphological analyzer. **http://mecab.sourceforge.net/**, 2005.

[13] 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.

[14] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. A diversity-promoting objective function for neural conversation models. In **Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 110–119, San Diego, California, June 2016. Association for Computational Linguistics.

A　正誤表

表 2 の計算手順に誤りがあったため、再度計算をし直した結果をここに記載する。なお、口頭発表時には修正後の結果を報告している。

（誤）

	ゲイン (%)	95%信頼区間
全サンプル	12.83	[1.579, 24.08]
類似度が低いサンプル	11.66	[-16.04, 39.37]
類似度が高いサンプル	14.31	[7.674, 20.94]

表 3　人手評価の結果（修正前）

（正）

	ゲイン (%)	95%信頼区間
全サンプル	0.67	[-6.91, 8.24]
類似度が低いサンプル	-1.33	[-11.78, 9.11]
類似度が高いサンプル	2.67	[-8.42, 13.75]

表 4　人手評価の結果（修正後）