UserRNN の追加による話者ごとの発話情報を考慮したマルチターン対話生成

大西孝宗
岡山理科大学大学院総合情報研究科情報科学専攻
i20im02ot@ous.jp

椎名広光
岡山理科大学総合情報学部情報科学科
shiina@mis.ous.ac.jp

1 はじめに

ニューラルネットワークを用いた対話生成において、Encoder-Decoder(Seq2Seq) モデル [1, 2, 3] の応用が Vinyals ら [4] により提案されている．対話は一般的に話者の交代が複数回行われるマルチターンとなっているが、Encoder-Decoder モデルはひとつの入力に対しひとつの出力が対応するため、マルチターンの対話を扱う場合には複数の発話をひとつにまとめることで入力としている．このため、発話の始めのあたりの情報が失われやすく、しばしば会話の流れに沿わない応答を生成してしまうことが報告されている．

これに対し、Serban らによる Hierarchical Recurrent Encoder-Decoder (HRED) モデル [5] は Encoder-Decoder モデルを複数個重ねて階層構造をつくることでマルチターンの対話に対応したモデルとなっている．更に Serban らは HRED モデルを改良し、潜在変数を追加することで多様な応答を生成するモデルとして、Latent Variable Hierarchical Recurrent Encoder-Decoder (VHRED) モデル [6] を提案している．

一方で、これらのモデルは発話したユーザを考慮していないため、同一の対話のなかで一貫性がない応答を生成することが課題としてあげられる．Li ら [7] や Bak ら [8] は、応答の生成時にユーザの埋め込みベクトルを用いることでユーザごとの発話の一貫性を持たせる手法を提案している．しかし、ユーザの埋め込みベクトルを用意出来ない場合には既存のモデルと同様に一貫性のない応答を生成してしまう．

本研究ではユーザの埋め込みベクトルを用いるのではなく、UserRNN を追加することで話者ごとの発話情報を保持し、一貫性のある応答生成を行っている．

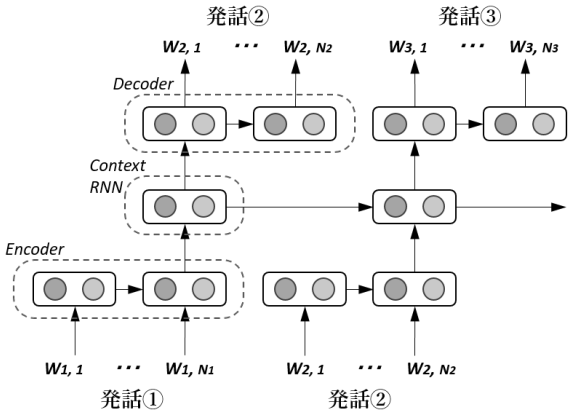

図 1 HRED モデル

2 関連研究

2.1 HRED モデル

HRED モデルはマルチターンの対話のために Encoder-Decoder モデルを拡張した対話生成モデルであり、発話 1 から発話 n-1 までを入力として発話 n を予測し生成する．

HRED モデルの構造は、Encoder-Decoder モデルを複数個重ねた階層構造となっており、階層構造中の各 Encoder-Decoder モデル間は ContextRNN によって結ばれている．ContextRNN は対話の流れを保持している．これによって過去の発話を考慮した応答の生成を可能としている．HRED モデルの概略図を図 1 に図示する．

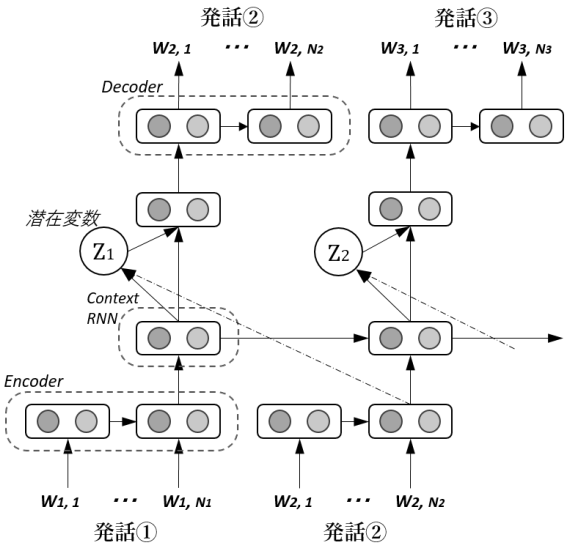

図 2　VHRED モデル

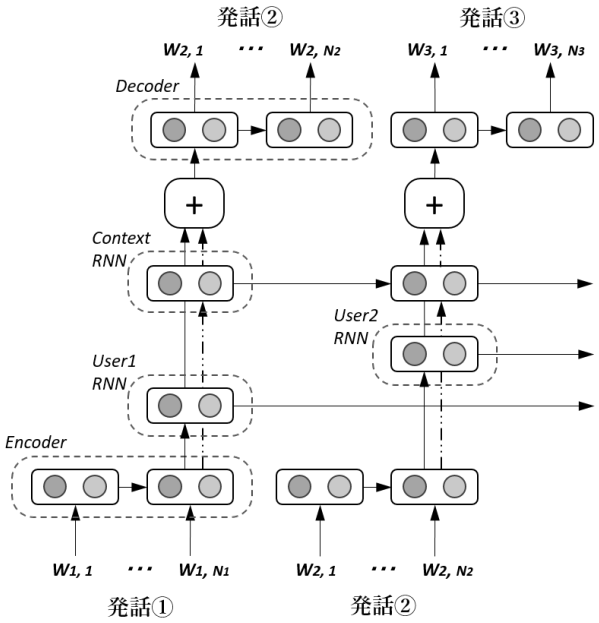

図 3　提案モデル 1：HRED (Context RNN + User RNN)
(HRED モデルの構成を ContextRNN と UserRNN の出力の和をデコーダの入力とするように変更)

2.2　VHRED モデル

VHRED モデルは HRED モデルを改良したモデルであり、HRED モデルでは“I don't know”のような無難な応答を生成してしまうことが多かったが、VHRED モデルでは潜在変数を加えることで ContextRNN に対して確率的なノイズを与え、多様な応答を生成することを可能にしている．特に、HRED モデルに比べて長文の応答を生成しやすいことが報告されている．VHRED モデルの概略図を図 2 に図示する．

3　発話者ごとの対話情報を考慮するための提案手法

本研究では既存のモデルである HRED モデルおよび VHRED モデルに対して、話者の発話情報を保持する UserRNN を追加することで、話者ごとの発話の一貫性を保った応答の生成を試みている．

3.1　UserRNN を追加した提案モデル

HRED モデルおよび VHRED モデルに対して UserRNN を追加した提案モデルについて説明する．構成の異なる 3 種類のモデルを提案する．

(1) ContextRNN と UserRNN の出力の和をデコーダの入力とするモデル　エンコーダからの出力を UserRNN および ContextRNN それぞれに入力し、その出力の和をデコーダへの入力としている．UserRNN は ContextRNN と同一の構造としている．(提案モデル 1：HRED (Context RNN + User RNN) 図 3 および提案モデル 4：VHRED (Context RNN + User RNN))

(2) UserRNN のみの構成に変更したモデル　ContextRNN を除き、UserRNN からの出力のみをデコーダへの入力としている．(提案モデル 2：HRED (User RNN only) 図 4 および提案モデル 5：VHRED (User RNN only))

(3) UserRNN の出力を ContextRNN に入力するモデル　提案モデル 3 の概略図を図 5 に図示する．エンコーダからの出力を UserRNN に入力し、さらにその出力を ContextRNN への入力とする．デコーダへの入力は ContextRNN からの出力のみとしている．(提案モデル 3：HRED (User RNN → Context RNN) 図 5 および提案モデル 6：VHRED (User RNN → Context RNN))

4　対話の生成実験と評価

各モデルについてマルチターンの対話応答生成を行う．データセットには Ubuntu Dialogue Corpus [9] および Cornell Movie-Dialogs Corpus [10] を用いた．Ubuntu Dialogue Corpus はインターネットリレーチャットの Ubuntu チャンネルから 1 対 1 の対話を抽出したデータセットとなっており、約 100 万件の対話データが含まれている．Cornell Movie Corpus は映画の対話を抽出したデータセットであり、約 22 万件の対話データとなっている．

表 1　各モデルの Embedding-based Metrics による評価

Model	Cornell Average	Cornell Greedy	Cornell Extrema	Ubuntu Average	Ubuntu Greedy	Ubuntu Extrema
1-turn						
HRED モデル	0.559	0.412	0.358	0.557	0.404	0.332
提案モデル 1：HRED (Context RNN + User RNN)	0.562	**0.415**	0.364	0.544	0.392	0.323
提案モデル 2：HRED (User RNN only)	0.560	0.412	0.361	0.538	0.384	0.321
提案モデル 3：HRED (User RNN → Context RNN)	0.561	0.402	**0.381**	0.549	0.397	0.327
VHRED モデル	0.553	0.404	0.370	0.553	0.397	0.306
提案モデル 4：VHRED (Context RNN + User RNN)	0.554	0.407	0.359	0.569	0.408	0.305
提案モデル 5：VHRED (User RNN only)	0.552	0.406	0.355	**0.571**	**0.414**	**0.339**
提案モデル 6：VHRED (User RNN → Context RNN)	**0.563**	0.407	0.375	0.550	0.394	0.306
5-turn						
HRED モデル	**0.594**	**0.436**	0.371	0.606	0.425	0.346
提案モデル 1：HRED (Context RNN + User RNN)	0.590	0.434	0.364	0.572	0.389	0.336
提案モデル 2：HRED (User RNN only)	0.579	0.410	**0.381**	0.570	0.395	0.323
提案モデル 3：HRED (User RNN → Context RNN)	0.579	0.421	0.372	**0.614**	**0.435**	**0.352**
VHRED モデル	0.581	0.417	0.371	0.599	0.424	0.322
提案モデル 4：VHRED (Context RNN + User RNN)	0.582	0.426	0.359	0.574	0.393	0.301
提案モデル 5：VHRED (User RNN only)	0.571	0.424	0.348	0.497	0.337	0.297
提案モデル 6：VHRED (User RNN → Context RNN)	0.590	0.426	0.386	0.593	0.417	0.318

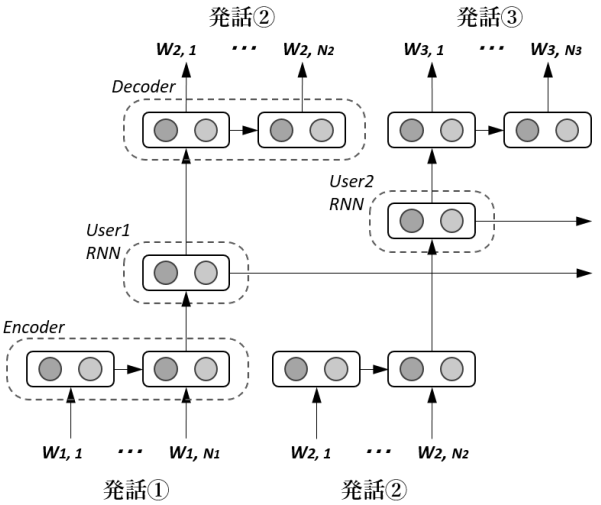

図 4　提案モデル 2：HRED (User RNN only)
(HRED モデルの構成を UserRNN のみへ変更)

4.1　評価手法

生成した応答文と実際の応答との関連性を評価するため、Liu らが提案した Embedding-based Metrics [11] を用いて自動評価を行う．Embedding-based Metrics は事前学習済みの単語ベクトルを用いて文の類似性を評価するものであり、Embedding Average, Greedy Matching, Vector Extrema の 3 つの算出方法が提案されている．事前学習済みの単語ベクトルには、Google News Corpus で学習させた Word2Vec の単語ベクトルを用いた．

(1) Embedding Average　モデルが生成した発話文中の単語ベクトルの平均と、実際の応答文中の単語ベクトルの平均をそれぞれの文のベクトルとして両者のコサイン類似度を算出しスコアとする．

(2) Greedy Matching　生成文と実際の応答文に含まれる単語ベクトルを比較した際に、最もコサイン類似度が高くなる単語の組についてそれぞれコサイン類似度を算出し、その平均をスコアとする．

Greedy Matching は次の式で表される．

$$G(r,\hat{r}) = \frac{\sum_{w\in r} \max_{\hat{w}\in\hat{r}} cos\_sim(e_w, e_{\hat{w}})}{|r|} \quad (1)$$

$$GM(r,\hat{r}) = \frac{G(r,\hat{r}) + G(\hat{r},r)}{2} \quad (2)$$

なお、r を実際の応答、$\hat{r}$ を生成した応答とする．

(3) Vector Extrema　モデルが生成した発話文および実際の応答文について、それぞれの文に含まれる単語ベクトルの各次元ごとの最大値を用いて文のベクトルをつくり、生成文と実際の応答文のコサイン類似度を算出しスコアとする．

表 2　Ubuntu Dialogue Corpus を用いた各モデルの対話生成例 ("→"は話者の交代を表す)

Context	Response
hello, guys! i want to know, if i have a debian vps, how can i install ubuntu to replace debian? → hmm.. wipe and reinstall is the safest way to do it. but it's not the only way. but it is the only way to be absolutely sure. → i've only got ssh access to it :(→ why're you downgrading any how ?	**HRED モデル:**i 'm trying to upgrade from 8.04 to 8.10 **提案モデル 1:**i want to upgrade to breezy **提案モデル 2:**i m trying to get my <unk> to work with the latest version of the kernel and the <unk> <unk> <unk> **提案モデル 3:**because i want to install a new version of ubuntu , and i want to install a newer version of ubuntu , and i want to install ubuntu on
hello, guys! i want to know, if i have a debian vps, how can i install ubuntu to replace debian? → hmm.. wipe and reinstall is the safest way to do it. but it's not the only way. but it is the only way to be absolutely sure. → i've only got ssh access to it :(→ why're you downgrading any how ?	**VHRED モデル** because the ubuntu server is only updated, which are you currently in the boot order ? **提案モデル 4:**what do you mean **提案モデル 5:**because i don't want to upgrade to edgy , because i can't get it to work **提案モデル 6:**because i want to upgrade to the latest version of apt get , and it 's just that they have broken dependencies , so if it fails

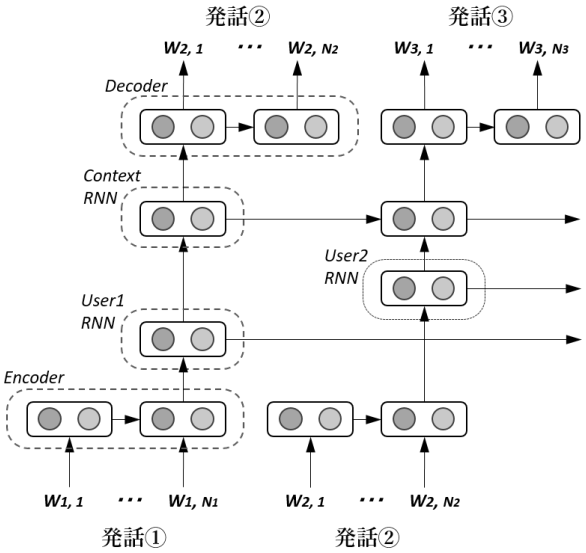

図 5　提案モデル 3：HRED (User RNN → Context RNN) (HRED モデルの構成を UserRNN の出力を ContextRNN に入力するよう変更)

Vector Extrema は次の式で表される．

$$e_{rd} = \begin{cases} \max_{w \in r} e_{wd} & \text{if } e_{wd} > |\min_{w' \in r} e_{w'd}| \\ \min_{w \in r} e_{wd} & \text{otherwise} \end{cases} \quad (3)$$

4.2　実験結果

4.2.1　Embedding-based Metrics による評価

モデルに入力するコンテキストのターン数が 1 ターンの場合と 5 ターンの場合について評価を行った．各モデルの Embedding-based Metrics によるスコアを表 1 に示す．

1 ターンの場合には既存のモデルに比べ、おおむね提案モデルのスコアが上回っている．特に Ubuntu コーパスでは、VHRED モデルの構成を UserRNN のみに変更を加えた提案モデル 5 のスコアが良い．

一方で、5 ターンの場合ではスコアの改善はあまり見られず、提案モデルのスコアが既存のモデルのスコアを僅かに下回る場合が多い．特に Ubuntu コーパスにおいて VHRED モデルの構成を UserRNN のみに変更を加えた提案モデル 5 のスコアは他のモデルに比べて大きく下回っている．このことは、ContextRNN がないことによって対話全体の流れを保持することが出来ず、結果としてコンテキストのターン数が増えた場合の応答生成が困難になっている可能性を示している．

4.2.2　応答文の生成例について

Ubuntu コーパスを用いた各モデルの応答文の生成例を表 2 に示す．HRED モデルおよび HRED モデルを変更した提案モデル 1,2,3 はコンテキストにまったく沿わない応答や文としておかしい応答が見受けられる．VHRED モデルおよび VHRED モデルを変更した提案モデル 4,5,6 については、比較的にコンテキストにふさわしい応答を生成している．一方で、UserRNN の追加による差は少ない．

5　まとめ

UserRNN の追加によって、生成した応答と実際の応答との関連性の向上に一定の効果は得られた．しかしながら、一貫性のある応答の生成という点は、Embedding-based Metrics では評価が難しく、コンテキストと生成した応答の関連性についての評価や、人手による評価によって、UserRNN の追加による応答生成への影響を詳しく調べる必要がある．

参考文献

[1]Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27 (NIPS 2014)*, pp. 3104–3112, 2014.

[2]Thang Luong, Hieu Pham, and Christopher D. Manning. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 1412–1421, Lisbon, Portugal, September 2015. Association for Computational Linguistics.

[3]Alexander M. Rush, Sumit Chopra, and Jason Weston. A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 379–389, Lisbon, Portugal, September 2015. Association for Computational Linguistics.

[4]Oriol Vinyals and Quoc V. Le. A neural conversational model. In *ICML Deep Learning Workshop*, 2015.

[5]Iulian Vlad Serban, Alessandro Sordoni, Yoshua Bengio, Aaron C. Courville, and Joelle Pineau. Building end-to-end dialogue systems using generative hierarchical neural network models. In Dale Schuurmans and Michael P. Wellman, editors, *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA*, pp. 3776–3784. AAAI Press, 2016.

[6]Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron C. Courville, and Yoshua Bengio. A hierarchical latent variable encoder-decoder model for generating dialogues. In Satinder P. Singh and Shaul Markovitch, editors, *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pp. 3295–3301. AAAI Press, 2017.

[7]Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and Bill Dolan. A persona-based neural conversation model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 994–1003, Berlin, Germany, August 2016. Association for Computational Linguistics.

[8]JinYeong Bak and Alice Oh. Variational hierarchical user-based conversation model. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 1941–1950, Hong Kong, China, November 2019. Association for Computational Linguistics.

[9]Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. The Ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pp. 285–294, Prague, Czech Republic, September 2015. Association for Computational Linguistics.

[10]Cristian Danescu-Niculescu-Mizil and Lillian Lee. Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. In *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics*, pp. 76–87, Portland, Oregon, USA, June 2011. Association for Computational Linguistics.

[11]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.