発話者ごとの潜在変数学習による対話応答生成の一貫性向上

大西孝宗 [1]　椎名広光 [2]
[1] 岡山理科大学大学院 総合情報研究科　[2] 岡山理科大学 総合情報学部
i20im02ot@ous.jp　shiina@mis.ous.ac.jp

概要

対話応答生成において CVAE を応用したモデルは潜在変数の違いによって多様性のある応答生成を可能にしている．一方で，コンテキスト全体からサンプリングされた潜在変数は，発話者の特徴が希薄となり，生成した応答の一貫性低下に影響すると考える．この課題を解決するため，発話者ごとの潜在変数を用いることで，一貫性のある応答を生成することを試みている．また，手法の有効性を検証するため，英語と日本語のコーパスを用いて応答の生成実験を行い，自動評価指標を用いて応答の多様性や参照応答との類似性，コンテキストとの類似性について評価を行った．

1　はじめに

近年，ニューラルネットワークを用いた対話応答生成の研究が盛んに行われおり，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．対話データを用いて，Encoder-Decoder モデルに発話から応答へ変換することを学習させる手法であるが，無難な応答を生成する傾向にあることが課題として報告されている．

この課題に対して，Conditioned Variational Autoencoder(CVAE) をベースとした手法が提案されている [5][6]．また，計算の高速化を目的とし，RNN を Transformer[7] に置き換えたモデルである，Global Variational Transformer (GVT) モデル [8] が登場している．これらの手法では，コンテキスト・応答から生成した事前・事後分布より，サンプリングした潜在変数をデコーダの入力に付加することを行っている．これにより，応答の生成において，潜在変数の違いから多様性をもたらすことに成功している．

一方で，潜在変数は発話者を区別することなく，コンテキスト全体を条件としてサンプリングされる．このため，発話者それぞれの主張が混在してしまうことや，発話者の特徴が希薄となり生成した応答の一貫性低下に影響すると考える．

本研究では，応答の一貫性を改善するため発話者ごとの潜在変数を GVT モデルに対して追加する．これにより，発話者ごとの特徴を捉える効果を狙い，一貫性のある応答を生成することを試みている．

2　関連研究

2.1　対話応答生成のための CVAE

対話生成タスクの向けの CVAE は，Encoder-Decoder モデルを拡張したモデルである．RNN ベースの Encoder を 2 つ持つ構造であり，コンテキストと応答を別々にエンコードする．さらに，コンテキストと応答それぞれの Encoder の隠れ層のベクトルに従う事前・事後分布を多層パーセプトロン (MLP) によって近似した，Prior Net・Recognition Net から潜在変数をサンプリングする．最後に，Decoder はコンテキストを処理した Encoder の隠れ層ベクトルと潜在変数 z を用いて，応答の生成を行う．

また，学習が進むにつれて Decoder が潜在変数 z の情報を考慮しなくなる KL vanishing 問題のため，KL アニーリング [9]，Bag-of-Words(BoW) loss を適用している．

2.2　GVT モデル

Global Variational Transformer (GVT) モデルは対話向け CVAE モデルをベースに，Transformer 化したモデルである．Encoder および Decoder において，RNN の代わりに Transformer を用いている．CVAE モデルでは事前・事後分布の MLP による近似である Prior Net・Recognition Net の計算に Encoder の隠れ層の状態を用いていたが，GVT モデルでは CLS トークンの位置の出力ベクトルを用いる．CLS トークンは入力シーケンスの先頭に追加され，Self-attention を介して出力ベクトルが計算される．したがって，CLS トークンの出力ベクトルは，入力全体の表現

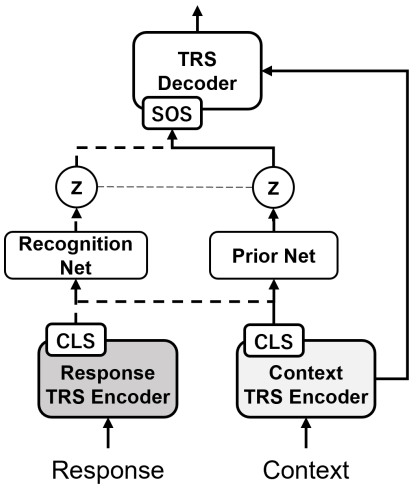

図 1　Global Variational Transfo (GVT) モデル

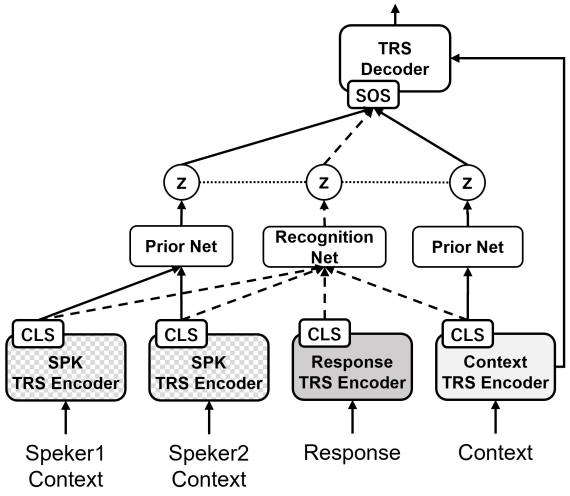

図 2　各発話者ごとの入力を処理する Encoder を GVT モデルに追加した提案モデル

したベクトルとみなすことができる．Prior Net・Recognition Net からサンプリングされた潜在変数は，Decoder の入力の先頭に追加し，応答の生成に利用される．また，CVAE と同様に潜在変数 z を Decoder に考慮させるため，KL アニーリング及び BoW loss を用いて学習を安定させている．GVT モデルの概略図を図 1 に図示する．

3　各発話者ごとの潜在変数の生成を行う提案手法

コンテキスト全体から生成した潜在変数では，発話者ごとの特徴が希薄になり一貫性にかける応答を生成する一因であると考えられる．そこで本研究では，発話者ごとの潜在変数を GVT モデルに対して追加し，モデルが生成する発話の一貫性を向上させる効果を狙う．

提案モデルでは，GVT モデルと同様にコンテキストと応答の入力を処理し，Prior Net・Recognition Net から潜在変数 z をサンプリングする．加えて，各発話者ごとの入力からも Prior Net を生成し潜在変数 z をサンプリングする．

コンテキストは対話を行う 2 者の発話を一つにまとめたものであり，それぞれの発話者ごとに分割することができる．各発話者ごとにコンテキストを分割し，入力シーケンスの先頭に CLS トークンを付加した上で，それぞれ Encoder へ入力する．さらに，CLS トークンの位置の出力ベクトルから事前分布を生成し，各発話者ごとの潜在変数をサンプリングする．通常の潜在変数に加えて，応答の発話者の潜在変数を Decoder の入力の先頭に追加し，応答の生成において利用する．また，GVT モデルと同様に KL アニーリングおよび BoW loss を学習安定のために用いる．提案モデルの概略図を図 2 に図示する．

4　応答の生成実験と評価

4.1　実験設定

ベースラインである GVT モデルと提案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 [10] およびおーぷん 2 ちゃんねる対話コーパス [11] を用いた．前処理として SentencePiece[1)] を用いてサブワードへの分割を行っている．

4.2　評価手法

応答の多様性の自動評価　モデルによって生成された応答の多様性を測定するため，dist-n[12] を用いる．n-gram の総数に対して n-gram の種類数が占める割合を算出し，この比率が高いほど，多様性が高いことを示す．

応答の類似性の自動評価　モデルによって生成された応答と参照応答の類似性について，両者の埋め込みベクトル間のコサイン類似度を計算することで評価する．2 種類の計算手法を用いる．

Embedding-based-Metrics[13] 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手法であり，Embedding Average, Greedy Matching, Vector Extrema の 3 種類の算出法が提案されている．本研究では Embedding Average を評価指標として用いた．事前学習済みの単語ベクトルとして，Ubuntu

1）　https://github.com/google/sentencepiece

表 1　各モデルの自動評価結果

Ubuntu Dialogue コーパス					
	Diversity			Simirality	
Model	**dist-1**	**dist-2**	**dist-3**	**EMB_{w2v}**	**BERT Score**
GVT モデル	0.010	0.087	**0.267**	**0.671**	**0.835**
提案モデル: GVT + SpeakerEncoder	**0.014**	**0.096**	0.256	0.626	0.825
おーぷん２ちゃんねるコーパス					
	Diversity			Simirality	
Model	**dist-1**	**dist-2**	**dist-3**	**EMB_{w2v}**	**BERT Score**
GVT モデル	0.011	0.228	0.705	0.604	0.652
提案モデル: GVT + SpeakerEncoder	**0.015**	**0.364**	**0.853**	**0.662**	**0.657**

表 2　参照応答の発話者の発話履歴と生成した応答についての BERT Score

Ubuntu Dialogue コーパス	
Model	**BERT Score**
GVT モデル	0.832
提案モデル: GVT + SpeakerEncoder	0.818
おーぷん２ちゃんねるコーパス	
Model	**BERT Score**
GVT モデル	0.652
提案モデル: GVT + SpeakerEncoder	0.657

表 3　Ubuntu Dialogue Corpus およびおーぷん２ちゃんねる対話コーパスを用いた各モデルの対話生成例

Ubuntu Dialogue コーパス	
Context	**発話 1:**hi any expert for grub and boot here? **発話 2:**sorry whats the problem? **発話 3:**my system isn't boot from the second hdd
Response	**GVT モデル:**I have to edit the grub partition in ubuntu i will try? **提案モデル:**the drive is not supported with grub **参照応答:**it only boots from the first drive.
おーぷん２ちゃんねる対話コーパス	
Context	**発話 1:**ついに茶が出るようになったか **発話 2:**ゆーめーじん? **発話 3:****コテで有名
Response	**GVT モデル:**(^ q ^) クッソ難しいから **提案モデル:**じゃあ全然有名じゃないです w **参照応答:**なーる。まぁ。いーやー!

コーパスの評価には Google News Corpus で学習させた Word2Vec の単語ベクトルを用いた．おーぷん２ちゃんねるコーパスでは，日本語 Wikipedia で学習させた Word2Vec の単語ベクトルを用いた．

BERT Score[14] は，事前学習した BERT[15] の埋め込みを使用して，モデルによって生成された応答と参照応答の類似性を評価する手法である．

発話の一貫性　発話の一貫性を調査するため，モデルによって生成された応答と参照応答の発話者の発話履歴について，BERT Score を用いて類似性を算出する．

5　実験結果

5.1　自動評価結果

各モデルが生成した応答の自動評価結果を表 1 に示す．生成した応答の多様性においては，Ubuntu コーパス及びおーぷん２ちゃんねるコーパスのどちらにおいても提案モデルのスコアが向上している．一方で参照応答との類似性については，Ubuntu コーパスとおーぷん２ちゃんねるコーパスで反対の結果となった．これは，データセットの性質が影響したと考えられる．おーぷん２ちゃんねるコーパスは多様性に富んだ応答を数多く含むデータセットであるため，多様性の高い応答を生成する提案モデルがスコアを伸ばしている．逆に，Ubuntu コーパスはドメイン固有のデータセットであるため，応答の多様性は高くなく，提案モデルはスコアを落としている．

各モデルが生成した応答と参照応答の発話者の発話履歴の類似性について自動評価結果を表 2 に示す．こちらも提案モデルは，おーぷん２ちゃんねるコーパスではスコアが伸び，Ubuntu コーパスではスコアを下げており，生成した応答と参照応答の類似性スコアと同じ傾向となっている．

5.2　応答生成例

Ubuntu コーパス及びおーぷん２ちゃんねるコーパスから各モデルが生成した応答例を表 3 に示す．Ubuntu コーパスの生成例では，両モデルともやや見当違いの生成をしている．おーぷん２ちゃんねるコーパスの例では，両モデルの生成例とも文法的な

違和感はない．また，提案モデルが生成した応答は参照応答とは異なるが，対話の流れとして比較的自然である．

6 おわりに

本研究では，対話応答生成タスクにおける応答の一貫性を向上させるため，GVT モデルに対して発話者ごとの潜在変数をサンプリングする手法を追加したモデルを提案した．提案モデルでは，発話者ごとの潜在変数の効果によって多様性が向上し，おーぷん２ちゃんねるコーパスでは参照応答との類似度向上につながった．しかしながら，自動評価指標のみでは応答の一貫性について評価しきれていない．

今後の課題として，発話者ごとの潜在変数追加が応答の一貫性にあたえた影響についてより深い検証を行うことと，Decoder において潜在変数をうまく活用するための改善に取り組みたいと考えている．

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] 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.

[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] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, **Advances in Neural Information Processing Systems**, Vol. 30. Curran Associates, Inc., 2017.

[8] Zhaojiang Lin, Genta Indra Winata, Peng Xu, Zihan Liu, and Pascale Fung. Variational transformers for diverse response generation. **arXiv preprint arXiv:2003.12738**, 2020.

[9] Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew Dai, Rafal Jozefowicz, and Samy Bengio. Generating sentences from a continuous space. In **Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning**, pp. 10–21, Berlin, Germany, August 2016. Association for Computational Linguistics.

[10] 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.

[11] 稲葉通将. おーぷん 2 ちゃんねる対話コーパスを用いた用例ベース対話システム. 第 87 回言語・音声理解と対話処理研究会 (第 10 回対話システムシンポジウム), 人工知能学会研究会資料 SIG-SLUD-B902-33, pp. 129–132, 2019.

[12] 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.

[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] Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In **International Conference on Learning Representations**, 2020.

[15] 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.