発話者分離学習を用いた対話モデルによる小学校の授業発話の分析

大西朔永 [1]　椎名広光 [2]　保森智彦 [3]
[1] 岡山理科大学大学院総合情報研究科　[2] 岡山理科大学情報理工学部
[3] 岡山理科大学教育学部
i22ed08bf@ous.jp　{shiina,yasumori}@ous.ac.jp

概要

小学校段階の授業では，教員の説明や促進，質問等の発話と児童の返答が多くあり，一種の対話がなされている．これらの対話から教員の発話の種類や児童の学びの状況を分析することで，学びに関する分類を教員へフィードバックすることが可能と考えられる．自然言語処理分野では，文脈を捉えることが比較的可能となっている BERT の他に，対話に適したモデルとして対話応答生成で用いられる VHRED や GVT が提案されている．GVT に対して，本研究では発話者ごとの特徴に加えて，対話処理の内部に反映させる前に対話を事前分類する手法（拡張 GVTSC）を提案し，授業の発話の分析を行う．

1　はじめに

日本の小学校の授業では，教員は児童の状況を見ながら授業を進めており，児童は学習状況や意見，感想を発話することが多くあり，対話をしながら授業は進むと考えられる．教員と児童や児童間では，一種の対話が成り立っており，授業の理解を促したり示したりする発話や対話を自動的に分析することができれば，教員に対して多くのフィードバックが可能となる．省察のデジタル化を分析した研究 [1] やシステムによる発話の分析 [2] がある．日本の教員は省察等の時間が 48 か国で最も短く [3]，システムによる分析手法の開発は急務である．

一方，自然言語処理分野では，Transformer[4] を用いた文脈を考慮した言語処理が可能な BERT[5] が提案されている．対話処理に適したモデルでは，翻訳において用いられる RNN[6] をベースにした Encoder-Decoder モデル [7] を対話応答生成へ応用した研究 [8] がある．長い入力系列に対応するために階層構造を取り入れた HRED[9] が提案され，CVAE[10] の手法を取り入れた VHRED[11] では，潜在変数を用いることで応答生成に多様性を与えている．また，対話向け CVAE に Transformer を導入した GVT[12] が提案されている．我々は発話者ごとの特徴を考慮するために，発話を発話者ごとに分離して入力し，発話者ごとの潜在変数を用いるように拡張を行った拡張 GVT モデル [13] を提案している．

本研究では，事前にクラスタリングを用いて発話者の特徴を抽象化する拡張 GVTSC モデルを提案すると共に，小学校の授業の対話に対して，拡張 GVTSC モデルによる発話の分析を行っている．具体的には，小学校の算数の授業を録画し，教員と児童の発話に対して文字起こしを行った対話形式のテキスト情報を分析している．分析には，あらかじめ授業中の発話に対して学びに関連する複数のラベルを人手で付与した後に，その発話と近い距離にある発話を提案する拡張 GVTSC を用いて抽出している．

2　発話者のクラスタリングを追加した拡張 GVTSC モデル

対話応答生成には，様々な対話に対しての応答と成り得る無難な応答を生成するために，応答の多様性が低くなるという課題が存在する [14, 15]．GVT モデルは，Decoder の入力にサンプリングした潜在変数を利用する手法であるが，発話者の特徴を潜在変数で表現し，サンプリングを行うことで，応答の多様性をもたらしていると考えられる．しかし，先行研究では潜在変数が生成された応答の一貫性を低下させる傾向があることが示されている [16]．そこで，発話者ごとの特徴を考慮するために，クラスタリングを用いて発話者の特徴を抽象化し，Encoder においてその発話者の特徴を加味することで，一貫性と多様性を向上させる．本研究では，GVT モデル及び，話者の分離を行っている拡張 GVT モデル

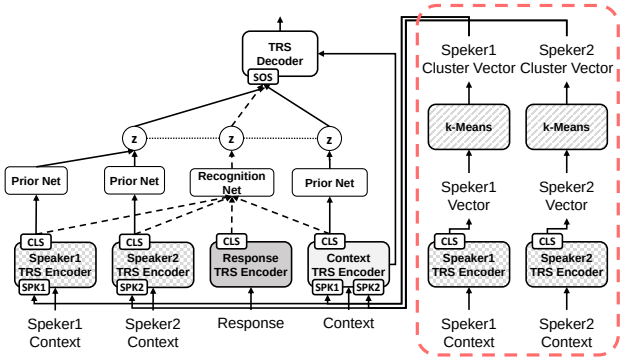

図 1　拡張 GVTSC モデルの構造

に対して，事前に発話者を分類するクラスタリングを追加した 2 種類のモデル（GVTSC，拡張 GVTSC）を提案し，評価を行っている．

拡張 GVTSC モデルの概要を図 1 に示す．拡張 GVT モデルに対して，発話者の特徴ベクトルをクラスタリングを用いて作成する部分（図 1 の点線部分）を追加し，コンテキストのエンコードにおいて利用している．

次に，拡張 GVTSC モデルの処理について述べる．最初に，発話者の特徴ベクトルをクラスタリングを用いて作成する．コンテキストは対話を行う 2 者の発話をまとめたものであり，各発話者ごとに分割することが可能である．そこで，対話のコンテキストを発話者ごとに分割し，発話者ごとに処理を行っている．発話者ごとの処理は同様であるため，発話者 1 の場合を述べる．Speaker1 TRS Encoder で発話者 1 のコンテキストをエンコードする．TRS Encoder では，入力系列の先頭に CLS トークンを付加しており，Transformer によって出力ベクトルが計算される．発話者 1 のコンテキストのベクトル（Speaker1 Vector）として，CLS トークンのベクトルを取得する．Speaker1 Vector に対して，クラスタリングを行う．本研究では，クラスタリングに k-Means を使用している．クラスタ数 k については，ハイパーパラメータ同様実験において決定する必要がある．クラスタリングの結果，Speaker1 Vector が属するクラスタを予測し，そのクラスタの中心ベクトル（Speaker1 Cluster Vector）を取得する．発話者 2 に対しても，発話者 1 と同様の処理を行い，Speaker2 Cluster Vector を取得する．ここで，クラスタリングに使用している TRS Encoder は，応答生成で訓練している TRS Encoder を共有している．ただし，クラスタリングの処理では，誤差逆伝搬による訓練は行われない．

対話のコンテキスト全体は，Context TRS Encoder に入力し，出力ベクトルを得る．コンテキストのエンコードでは，入力系列に発話者ごとのトークン（SPK1, SPK2）を追加し，SPK1 に Speaker1 Cluster Vector，SPK2 に Speaker2 Cluster Vector を入力する．また，各発話者ごとにコンテキストを分割し，それぞれを各 Speaker TRS Encoder に入力し，出力ベクトルを得る．その際には，Speaker1 TRS Encoder には入力系列に発話者 1 のトークン（SPK1）を追加し，Speaker1 Cluster Vector を入力し，Speaker2 TRS Encoder には入力系列に発話者 2 のトークン（SPK2）を追加し，Speaker2 Cluster Vector を入力する．発話者ごとのコンテキストのエンコードにおいて，各発話者の特徴ベクトルを利用することで，より発話者の特徴を考慮したエンコードを図っている．

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 z をサンプリングする．Prior Net は，Speaker TRS Encoder または，Context TRS Encoder の CLS トークンの出力ベクトルを基に，MLP によってコンテキストのベクトルの平均と分散を推定する．その平均と分散に従う正規分布より潜在変数 z をサンプリングする．Recognition Net では，Speaker TRS Encoder と Context TRS Encoder に加えて Response TRS Encoder の CLS トークンの出力ベクトルも用いて，MLP により対話全体のベクトルの平均と分散を推定する．Prior Net と同様に推定した平均と分散に従う正規分布から潜在変数 z のサンプリングを行う．TRS Encoder の CLS トークンの出力ベクトルは入力全体の表現したベクトルとみなすことができるため，CLS トークンの出力ベクトルから事前・事後分布を生成し，潜在変数 z をサンプリングしている．

TRS Decoder では，入力系列の先頭の SOS トークンに通常の潜在変数に加えて，応答の発話者の潜在変数を入力することで，潜在変数を応答の生成に利用している．この際，TRS Decoder は，学習時には Recognition Net からサンプリングした潜在変数を利用し，生成時には Prior Net からサンプリングした潜在変数を利用する．

拡張 GVTSC モデルは，c をコンテキスト，c_{s1} を発話者 1 のコンテキスト，c_{s2} を発話者 2 のコンテキスト，x を応答，z を潜在変数，v_{s1} を発話者 1 のクラスタベクトル，v_{s2} を発話者 2 のクラスタベクトルとして，下記の ELBO を最大化することでモデル

の最適化を行う．

$$
\begin{aligned}
&\mathcal{L}_{ELBO}(x,c)\\
&=\log p(x|c)\\
&\geqq \mathbb{E}_q(z|x,c,v_{s1},v_{s2})[\log p(x|z,c,v_{s1},v_{s2})]\\
&\quad - KL(q(z|x,c,v_{s1},v_{s2})\|p(z|c,v_{s1},v_{s2}))\\
&\quad - KL(s(z|x,c_{s1},c,v_{s1},v_{s2})\|r(z|c_{s1},v_{s1}))\\
&\quad - KL(s'(z|x,c_{s2},c,v_{s1},v_{s2})\|r'(z|c_{s2},v_{s2}))
\end{aligned}
\tag{1}
$$

ここで，KL は分布間の KL divergence であり，事前分布 p, r, r' は，下記の式で定義される．

$$p(z|c,v_{s1},v_{s2}) \sim \mathcal{N}(\mu_p,\sigma_p^2) \tag{2}$$
$$r(z|c_{s1},v_{s1}) \sim \mathcal{N}(\mu_r,\sigma_r^2) \tag{3}$$
$$r'(z|c_{s2},v_{s2}) \sim \mathcal{N}(\mu_{r'},\sigma_{r'}^2) \tag{4}$$

ここで，

$$\left[\mu_p,\log(\sigma_p^2)\right] = \mathrm{MLP}_p(c,v_{s1},v_{s2}) \tag{5}$$
$$\left[\mu_r,\log(\sigma_r^2)\right] = \mathrm{MLP}_r(c_{s1},v_{s1}) \tag{6}$$
$$\left[\mu_{r'},\log(\sigma_{r'}^2)\right] = \mathrm{MLP}_{r'}(c_{s2},v_{s2}) \tag{7}$$

事後分布 q, s, s' は，下記の式で定義される．

$$q(z|x,c,v_{s1},v_{s2}) \sim \mathcal{N}(\mu_q,\sigma_q^2) \tag{8}$$
$$s(z|x,c_{s1},c,v_{s1},v_{s2}) \sim \mathcal{N}(\mu_s,\sigma_s^2) \tag{9}$$
$$s'(z|x,c_{s2},c,v_{s1},v_{s2}) \sim \mathcal{N}(\mu_{s'},\sigma_{s'}^2) \tag{10}$$

ここで，

$$\left[\mu_q,\log(\sigma_q^2)\right] = \mathrm{MLP}_q(x,c,v_{s1},v_{s2}) \tag{11}$$
$$\left[\mu_s,\log(\sigma_s^2)\right] = \mathrm{MLP}_s(x,c_{s1},c,v_{s1},v_{s2}) \tag{12}$$
$$\left[\mu_{s'},\log(\sigma_{s'}^2)\right] = \mathrm{MLP}_{s'}(x,c_{s2},c,v_{s1},v_{s2}) \tag{13}$$

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 [17]，BoW（Bag-of-Words）loss[10, 18] の手法を取り入れている．KL アニーリングは式 1 の KL divergence の値について，学習が進むに連れて 0 から 1 に線形に増加する重みをつける手法である．BoW loss は応答に含まれる単語の集合を潜在変数から推定するサブタスクを追加する手法であり，潜在変数と応答中の単語の関連性を強くすることを目的としている．

3　対話モデルの評価実験

データセットにはおーぷん 2 ちゃんねる対話コーパス [19]，小学校の授業対話データを用いた．前処理として SentencePiece を用いてサブワードへの分割を行っている．コンテキストの長さについては，3-turn までの対話応答を評価している．自動評価手法として，Dist-n[20] を用いて，生成した応答の多様性，BERT Score[21] を用いて参照応答との類似性について評価する．Dist-N は，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指標である．BERT Score は，事前学習した BERT の埋め込みを使用して，モデルが生成した応答と参照応答の類似性を評価する手法である．

表 1　各モデルについての自動評価結果

Model	Diversity Dist-1	Diversity Dist-2	Diversity Dist-3	Simirality BERT
おーぷん 2 ちゃんねるコーパス				
GVT	0.007	0.361	0.911	0.643
拡張 GVT	**0.017**	0.413	0.886	**0.654**
GVTSC	0.008	0.405	0.931	0.644
拡張 GVTSC	**0.017**	**0.515**	**0.949**	0.652
実際の応答	0.017	0.552	0.926	-
小学校の授業対話データ				
GVT	0.484	0.720	0.739	0.654
拡張 GVT	0.530	0.810	0.821	0.655
GVTSC	0.563	0.877	0.910	0.662
拡張 GVTSC	**0.640**	**0.950**	**0.975**	**0.672**
実際の応答	0.647	0.947	0.963	-

各モデルが生成した応答の自動評価結果を表 1 に示す．GVTSC モデルは，GVT モデルに対して，拡張 GVTSC モデルと同様に発話者のクラスタリングを追加したモデルである．

まず，おーぷん 2 ちゃんねる対話コーパスによる評価について述べる．なお，クラスタリング（k-Means）のクラスタ数は，予備実験における結果から，GVTSC モデルは 8，拡張 GVTSC モデルは 3 としている．表 1 の GVTSC モデルは，GVT に対して，多様性の評価では，全て向上している．また，拡張 GVT モデルと比較すると，GVTSC モデルは Dist-3 で約 0.045 高い評価を得ている．拡張 GVTSC モデルでは，拡張 GVT モデルや GVTSC モデルと比較して，類似性の評価はほとんど差が見られない．しかし，多様性の評価では，特に Dist-2 で約 0.1 以上評価が向上している．

次に，小学校の授業対話データによる評価について述べる．なお，GVTSC モデルと拡張 GVTSC モデルのクラスタリング（k-Means）では，予備実験にお

表 2　拡張 GVTSC による 1 発話における類似発話

ラベルあり発話	付与ラベル	ラベルなし発話	Cos 類似度	JW 距離
2たす1は？○○君。これはさあ。表をどんな風に見た考え方なんかね？	数学的な見方，関数，深い学び	これが 1 段。これが 1 段だったらさ，2 段はどうなる？	**0.964**	0.468
ええ！まじかよ。ささっと書いて終わればいいんよ。あと 300 秒しかないんよ？	児童の主体的な学びとは逆行している	みんなさあ，20 段の時さあ，はい，注目！先生と目線が合うようにしてください。	**0.744**	0.377

表 3　小学校の授業対話データに対する対話応答の生成例

コンテキスト
発話 1: 12。
発話 2: 4 段の時は?
発話 3: 16。
発話 4: 5 段の時は?
発話 5: 24。24,24,24,28。えー。
応答
GVT: そうそうそう。はい, ここまでいいかな?同じです。
拡張 GVT: この式, 合ってそう?違ってそう?
GVTSC: 合ってそうです。さあ, ここまでいいかな?え?式一個じゃないん?
拡張 GVTSC: この式, 合ってそう?違ってそう?
参照: これで合ってる?

ける結果からクラスタ数を 8 としている．GVTSC は，多様性の評価で GVT と拡張 GVT より高い評価を得ており，類似性の評価でも約 0.008 向上している．拡張 GVTSC は，多様性の評価では全てのモデルより高い評価を得ており，実際の応答の多様性に近い評価となっている．類似性の評価においても，最も高い評価となっており，GVT と比較すると約 0.018 向上している．Encoder におけるエンコード過程で発話者の特徴を考慮することで，潜在変数のサンプリングや Decoder での Attention に利用する Encoder の各トークンの出力ベクトルに影響を与えていると考えられる．

小学校の授業対話データの評価において生成された応答の例を表 3 に示す．全てのモデルがコンテキストに関連した応答を生成することができており，多様性がある応答となっている．

4　拡張 GVTSC による発話の分析

本研究では拡張 GVTSC モデルを用いて対話のベクトル化を行っている．対話データを用いて訓練した拡張 GVTSC モデルは，対話応答を生成するために必要なコンテキストの発話をベクトル化する能力を得ている．対話のベクトルは，拡張 GVTSC モデルにコンテキストとして対話を入力し，拡張 GVTSC モデルの Context TRS Encoder と Speaker TRS Encoder が出力した CLS トークンのベクトルの和を計算することで作成している．

発話の分析では，授業の対話データについて，1 発話ごとのデータを作成し，ラベルなし発話について，ラベル付き発話との距離を求めている．距離については，ベクトルで表現できていることから Cos 類似度とジャロ・ウィンクラー距離（JW 距離）を用いた．表 2 に 1 発話における類似発話の例を示す．表 2 の 1 行目では，「数学的な見方」や「深い学び」に関連する発話を抽出できている．表 2 の 2 行目では，「児童の主体的な学びに逆行している」というラベルであるが，授業の終盤で授業を終わらせるための発話を抽出している．表面的な文字列の類似度である JW 距離と Cos 類似度の差は大きく，ベクトルを用いた手法の有効性を示している．定量的な評価として，Cos 類似度と JW 距離の相関係数を計算すると，表 2 の 1 発話の場合は無相関（-0.004）となっている．

5　おわりに

本研究では，事前にクラスタリングを用いて発話者の特徴を抽象化する拡張 GVTSC モデルを提案した．対話応答生成では，多様性の評価で実際の応答に近い評価を獲得した．小学校の授業対話データに対しては，類似性の評価においても最も高い評価となった．そして，小学校の授業の対話に対して，拡張 GVTSC モデルを利用した機械学習の手法で発話の分析を試みた．対話モデルを用いた対話のベクトル化により，表層的な類似性とは異なる傾向の分析が得られた．

今後は，教員と児童の対話に対して，発話者の属性が限られたドメインに適した対話モデルの開発を行い，少量の教師ラベルによるラベル推定を行いたいと考えている．

参考文献

[1] 保森智彦. 「主体的・対話的で深い学び」を実現するための教師の発話の検討, pp. 45–52. B, 人文・社会科学, No. 57. 岡山理科大学紀要, 2021.

[2] yuchen Wang, 大井翔, 松村耕平, 野間春生. 新任教員の授業力向上のための授業振り返りシステムに関する研究. 情報処理学会インタラクション, pp. 753–757, 2021.

[3] 国立教育政策研究所. 教員環境の国際比較 OECD 国際教員指導環境調査 (TALIS)2018 調査報告書. ぎょうせい, 2018.

[4] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In **Advances in neural information processing systems**, pp. 5998–6008, 2017.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. **arXiv preprint arXiv:1810.04805**, 2018.

[6] Klaus Greff, Rupesh K Srivastava, Jan Koutník, Bas R Steunebrink, and Jürgen Schmidhuber. Lstm: A search space odyssey. **IEEE Transactions on Neural Networks and Learning Systems**, Vol. 28, No. 10, pp. 2222–2232, 2016.

[7] 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.

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

[9] 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.

[10] 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.

[11] Iulian Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron Courville, and Yoshua Bengio. A hierarchical latent variable encoder-decoder model for generating dialogues. **Proceedings of the AAAI Conference on Artificial Intelligence**, Vol. 31, No. 1, Feb. 2017.

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

[13] Takamune Onishi, Sakuei Onishi, and Hiromitsu Shiina. Improved response generation consistency in multiturn dialog. In **2022 12th International Congress on Advanced Applied Informatics (IIAI-AAI)**, pp. 416–419, 2022.

[14] 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.

[15] Richárd Csáky, Patrik Purgai, and Gábor Recski. Improving neural conversational models with entropy-based data filtering. In **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics**, pp. 5650–5669, Florence, Italy, July 2019. Association for Computational Linguistics.

[16] Bin Sun, Shaoxiong Feng, Yiwei Li, Jiamou Liu, and Kan Li. Generating relevant and coherent dialogue responses using self-separated conditional variational AutoEncoders. 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. 5624–5637, Online, August 2021. Association for Computational Linguistics.

[17] 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.

[18] Xianda Zhou and William Yang Wang. MojiTalk: Generating emotional responses at scale. In **Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 1128–1137, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[19] Michimasa Inaba. A example based dialogue system using the open 2channel dialogue corpus. **Journal of Japanese Society for Artificial Intelligence**, Vol. 87, pp. 129—-132, 2019.

[20] 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.

[21] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. **In International Conference on Learning Representation**, 2019.