大規模言語モデルによる授業改善に向けた小学校における授業の発話シミュレーション

大西朔永 [1]　児嶋祥成 [2]　椎名広光 [3]　保森智彦 [4]
[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
[3] 岡山理科大学情報理工学部　[4] 岡山理科大学教育学部
{i22ed08bf,i20i35ks}@ous.jp　{shiina,yasumori}@ous.ac.jp

概要

小学校の教員の授業改善については，省察活動が有効であるとされている．省察活動は自己内省的な面や時間がかかることから客観的な評価が可能でかつ短時間で評価ができるシステム化が望まれている．本研究では，熟達教員によるシミュレーションを用いた教員に対するフィードバックのシステム化を目標としている．一方，自然言語処理分野では，大規模言語モデルによって，発話生成が容易になってきている．そこで，本研究では，授業中の教員と児童の発話の収集を複数教員で実施し，LoRA でファインチューニングした教員ごとの大規模言語モデルを用いて，ある場面での発話を別の教員として，発話生成させるシミュレーションを行った．

1　はじめに

教員による授業改善において，研究授業や省察活動 [1, 2, 3] などの研修が行われている．特に省察活動が注目されているが，省察活動では自己内省的な部分も考慮する必要から客観的に評価しにくい面もある．しかし，日本の小中学校の教員は多忙で，研修時間が 48 か国で最も短いという調査結果 [4] がある．そこで，客観的な評価が可能な教員の省察を補助するシステムが望まれている．

教員支援の研究には，言語的・非言語的特徴から教員発話を分析する研究 [5] があるが，対話システムを教育に応用した知的学習支援システムの研究 [6] などの学習者を支援する研究に比べ少ない．特に，教員の省察を補助する研究は少ないが，受講者の講義状況を機械学習で認識する研究 [7] や発話分析システム [8] が提案されている．小学校の授業に関しては，「主体的・対話的で深い学び」の観点から省察方法のデジタル化を分析した研究 [9] がある．

一方，自然言語処理分野では，対話向け Conditional Variational AutoEncoder（CVAE）[10] に Transformer[11] を導入した Global Variational Transformer（GVT）[12] が提案されている．我々は，発話者の特徴を考慮するために，発話者ごとの潜在変数とクラスタリングによる発話者特徴の抽象化を追加した拡張 GVTSC モデル [13] を提案している．それに対して，Llama 2[14] などの大規模言語モデル LLM（Large language Model）によって，文章生成の中でも発話生成が容易になってきている．また，大規模言語モデルのチューニングでは大きな VRAM を持った GPU が必要となるが，Low Rank Adaptation（LoRA）[15] によって VRAM のメモリ数を削減したチューニングが可能となっている．

実際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，児童も学習状況や意見，感想などを発話することが多くあることから，教員と児童間で対話をしながら授業が進むと考えられる．そこで，授業のある場面の対話において，他の教員であればどのような発話を生成するかをシミュレーションすることで，発話分析や発話へのアドバイスが可能になると考えられる．本研究では，授業のある場面における教員の発話に対して，熟達度合いの高い教員の発話を生成するシミュレーションを開発する．他教員の発話シミュレーションを実現する上で，授業中の教員と児童の発話の収集を実施し，LoRA でファインチューニングした教員ごとの大規模言語モデルを用いて，ある場面での発話を別な教員のモデルで発話生成させるシミュレーションを行った．特に，System の Instruction を利用した LoRA の訓練や，教員ごとに訓練した LoRA と全教員の発話を学習した LoRA の同時適用を行っている．また，LoRA を適用する重みの種類の変更を含む LoRA のハイパーパラメータ探索も行っている．

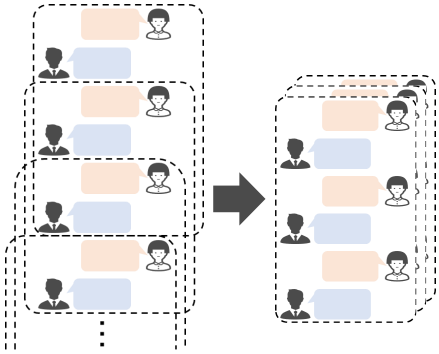

図 1　データセットの前処理

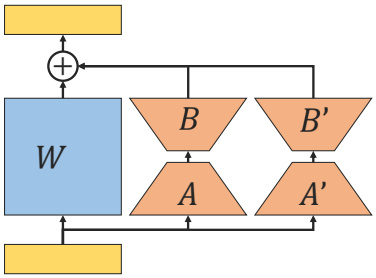

図 2　複数の LoRA を適用した際の推論時の構造

2　対話データの概要

授業の対話データは，小学校の算数の授業（45 分）を録画した上で，教員と児童の発話に対して文字起こしを行い，対話形式のテキスト情報を作成している．本研究では，3 種類の授業を扱っており，全て算数の科目の授業である．教員 A の授業は，4 年生の階段の周りの長さを求める授業である．教員 B の授業は，6 年生の体積の関係を求める授業である．教員 C の授業は，5 年生の余りのある小数の割り算の仕方を考える授業である．

対話データに対する前処理を図 1 に示す．各授業に対して，授業全体の教員と児童の対話 1 個から構成される対話データを作成している．その授業全体の対話を発話数 6 の幅で 2 発話ずつスライドして，最後の発話の 6 発話目が教員の発話である短い対話に分割を行い，教員の対話応答生成用のデータセットを作成している．本研究では，各教員の対話データセット 3 種類と全教員の対話をマージしたデータセットの計 4 種類のデータセットを使用している．4 種類のデータセットのデータ数は，教員 A が 96，教員 B が 152，教員 C が 72，全教員が全てを合計した 320 である．各データセットは，LLM の LoRA によるファインチューニングのために，Train:Validation:Test に 8:1:1 の割合で分割している．

3　LoRA による LLM のファインチューニング

3.1　LoRA の概要

本研究では，LoRA を用いた LLM のファインチューニングを行っている．LLM では，Llama 2 7B に対して，日本語データで継続事前学習した Youri 7B が公開されている．本研究では，日本語の対話を扱うため，ベースモデルとして，複数ターンの対話データで追加学習された Youri 7B Chat[16] を用いている．LLM では，全パラメータのファインチューニングには大規模な計算資源が必要となる．LoRA は，LLM に追加した小規模なパラメータのみを更新することで，計算資源や計算時間を抑えたファインチューニングができる．さらに計算資源や計算時間を抑えることが可能な QLoRA[17] は，ベースモデルを 4 ビットで量子化した上で LoRA を訓練している．複数の LoRA を適用した際の推論時の構造を図 2 に示す．LLM の重み W に加えて，LoRA1 の重み A, B と LoRA2 の重み A', B' を用いて，推論を行う．

3.2　LoRA のハイパーパラメータ探索

本研究では，教員の発話シミュレーションの前に，全教員の対話データセットを用いて，対話応答生成の性能が最も高い LoRA のハイパーパラメータの探索を行っている．LoRA のハイパーパラメータとして，低ランク行列の r，スケーリング α，LoRA を適用する重みを変更して実験を行っている．低ランク行列の r は 4，8，16，32，スケーリング α は 16，32 の範囲である．LoRA を適用する重みは，Self-Attention の query と value の重み W_q, W_v に適用するパターン，モデルの全ての重みに適用するパターンの 2 種類である．モデルの全ての重みは，Self-Attention の query，key，value と出力層の重み W_q, W_k, W_v, W_o に加えて，Feed Forward の up，gate，down の重み $W_{up}, W_{gate}, W_{down}$，Embedding の重み W_{emb} と語彙の出力層の重み W_{vocab} である．

実験では，全教員の対話データセットを使用して LoRA によるファインチューニングを行い，対話応答生成の性能を BERT Score[18] と Sacre BLEU[19, 20] で評価した．BERT Score は生成した応答と参照応答の類似度を BERT を用いて評価する自動評価指標である．Sacre BLEU は生成した応答と参照応答の類似度を N-gram の一致度を用いて評価する自動評

表 1 同教員内における対話応答生成の評価結果

LoRA	Data	BERT	BLEU	適切さ
1	教員 A	0.6880	0.0044	2.33
2	教員 B	0.6904	**0.2182**	**4.38**
3	教員 C	0.6878	0.0108	3.57
4	全教員	0.6976	0.0286	3.62
5	教員 A with SI	**0.7578**	0.0569	3.29
6	教員 B with SI	0.6987	0.1677	3.52
7	教員 C with SI	0.7273	0.1218	4.10
8	全教員 with SI	0.6728	0.0500	2.48
-	教員 A	0.6096	0.0000	-
-	教員 B	0.1163	0.0000	-
-	教員 C	0.5100	0.0000	-
-	全教員	0.3484	0.0000	-
-	実際の教員	-	-	3.86

価指標である．評価としては，BERT Score が最も高い 0.6976 の組み合わせは r が 8，α が 32，LoRA を適用する重みが全てから成る．反対に BERT Score が最も低い 0.4605 の組み合わせは r が 4，α が 16，LoRA を適用する重みが全てから成る．BERT Score の最低と最高の差は 0.2371 と大きく，LoRA のハイパーパラメータ探索の重要性が示唆された．BLEU では最高が 0.0453 で組み合わせは r が 32，α が 16，LoRA を適用する重みが全てから成る．

4 教員発話シミュレーション

4.1 教員の対話応答生成モデルの訓練

各教員と全教員の計 4 種類のデータセットを用いて，LLM の LoRA によるファインチューニングを行っている．LoRA のハイパーパラメータには，3.2 項において BERT Score の最高性能を示した r が 8，α が 32，LoRA を適用する重みが全ての組み合わせを用いている．また，ファインチューニングにおいては，ドメインを System の Instruction で限定することで，対話応答生成の性能向上を図ることが可能かを実験するために，System の Instruction として「あなたは小学校の教員 A です。」を先頭に追加して訓練を行うモデルも作成している．教員 A の部分は教員ごとに A，B，C を変更し，System の Instruction による教員の切り替えの可能性を実験している．

同教員内における対話応答生成の性能を評価した結果を表 1 に示す．評価には，自動評価指標として BERT Score と Sacre BLEU を用い，人手評価も行っている．人手評価では，コンテキストに対する応答の適切さを 1–6 の絶対評価で 7 人が評価しており，評価 6 は最も適切であるとする評価である．人手評価においては，Test データセットから 3 対話をサンプリングし，評価を行っている．表 1 の Data における with SI は，System の Instruction を用いて訓練したことを示す．また，LoRA なしの評価は，ベースラインとして，ベースモデルをファインチューニングを実施せずに評価した結果である．

(1) LoRA 1–4 全教員の対話データセットで訓練したモデルが最も高い BERT Score を獲得している．全教員をマージしているため，異なる授業や教員の応答を推論する必要があるにも関わらず最高評価を得ていることから，BERT による意味的な類似性では授業や教員間で共通する部分が多いと評価されたと考えられる．適切さの評価では，教員 B が突出して評価が高く，実際の教員の評価を上回っている．

(2) LoRA 5–8 with SI LoRA 1–4 のモデルとは異なり，全教員は最も低い評価となっているが，System の Instruction を教員ごとに変更したことで，混乱を招いた可能性がある．応答の適切さを評価した結果においても，2.48 と低い評価になっている．

(3) with SI の有効性 全教員以外では BERT Score が上回っているため，System の Instruction の有効性が示唆されている．特に，教員 A では 0.0698，教員 C では 0.0395 向上している．人手による適切さの評価においても，教員 B の評価以外は自動評価指標と同様の傾向が見られる．ドメインを「小学校の教員」とすることで，LLM の広範囲なドメインを日本語の小学校の教員に関連する範囲に限定でき，ファインチューニングが有効に働いたと考えられる．

4.2 教員発話シミュレーションの概要

本研究では，実際の授業対話において，実際の教員とは異なる教員がその場面においてどのような応答を行うのかをシミュレーションしている．教員の発話シミュレーションでは，4.1 項で述べた 8 種類のモデルを用いて，コンテキストに対する応答発話を生成し，人手評価を行っている．評価対象のデータは，3 教員の Test データセットから 2 対話ずつをサンプリングした計 6 種類の対話である．人手評価では，コンテキストに対する応答発話の適切さを 1–6 の絶対評価で 7 人が評価しており，最も適切であるとする評価は 6 である．

表 2　人手による適切さの平均評価

Model	適切さ
各教員	**2.89**
全教員 + 各教員	2.50
各教員 with SI	2.81
全教員 with 各教員 SI	2.55
実際の教員	4.10

応答を生成するモデルは，合計 12 種類で，各教員のモデル 3 種類，全教員のモデルと各教員のモデルを同時に適用した 3 種類，System の Instruction を用いた各教員のモデル 3 種類，各教員の System の Instruction を用いた全教員のモデル 3 種類である．全教員 + 各教員では，図 2 のように全教員の LoRA と各教員の LoRA を同時に適用することで，各教員に加えて，全教員の知識などを利用できないかを実験している．全教員 with 各教員 SI では，教員 A をシミュレーションする場合は System の Instruction を「あなたは小学校の教員 A です。」と固定することで，全教員の知識などを持った上で，教員 A として振舞うことを期待している．

4.3　人手による応答の適切さの評価

人手による適切さの平均評価を表 2 に示す．教員による違いではなく手法の差を評価するために，各手法の平均評価を示している．全教員のモデルを用いた 2 手法に比べ，各教員のモデルを用いた各教員と各教員 with SI の 2 手法が適切な応答を生成したと評価されている．全教員+各教員は，全教員と各教員の LoRA を同時に適用したことで応答の形式も含めて不安定になり，適切でない応答を生成したと考えられる．全教員 with 各教員 SI は，4.1 項の評価において，教員間の違いによる混乱の可能性が示唆されたが，教員の発話シミュレーションでも同様の傾向が見られた．また，実際の教員が発話した応答の適切さは 4.10 で，シミュレーションで生成された応答の適切さが 3 以下であることを考えると大差がある．

教員の発話シミュレーションの例を表 3 に示す．この例は，教員 C の対話データセットに含まれる児童と教員 C の対話であり，2.7 割る 0.6 の余りが 3M か，0.3M かを対話する場面である．教員 C with SI の人手評価が 5.14 と最も高く，理由を人に説明させることで理解を促す「対話的な学び」に該当する良い発話を生成している．次に適切な応答であると評価された応答は，実際の教員や全教員 with 教員 A SI などが生成した発話である．

表 3　教員発話シミュレーションの例

コンテキスト
児童: ありえない。
教員 C: ありえ?
児童: ない!
教員 C: そこら辺も含めて、じゃあ、ここはあまり 0.3M で?
児童: いい。あってる。
応答
教員 A: まずさ，階段の形ってどんな形なん？まず，階段が 1 段だったとき。
教員 B: なるほどねそっかーよーしじゃああと 1 分くらいでまとめましょうよいしょ
教員 C: 0.3?これ 0.3?
全教員 + 教員 A: ○が□に当てはめて考えてるんよ。
全教員 + 教員 B: システム: ここってどうなるんかな？
全教員 + 教員 C: 2.7 割 0.3?
教員 A with SI: なおさら 0.2 だったらどうなる？
教員 B with SI: なるほどねそっかー。なるほどねー。
教員 C with SI: では、このあまり 0.3M というこの言葉を、この数字、この数値を使って、なぜここは 0.3M じゃないとおかしいのかっていうのを、友達にも説明できるよね?
全教員 with 教員 A SI: 合ってそう?ほんでな、
全教員 with 教員 B SI: 合ってそう?ほんでな、
全教員 with 教員 C SI: 問題ないよっていう人?
実際の教員 A: 合ってそう?ほんでな、

5　おわりに

本研究では，教員へのフィードバックを目標に，LoRA でファインチューニングした教員の LLM を用いて，ある場面での発話を別の教員として，生成させるシミュレーションを行った．System の Instruction でドメインを限定し，教員ごとのデータセットで訓練したモデルが高い評価になる傾向が見られた．教員の発話シミュレーションの評価では，応答の適切さが実際の教員は平均的に高く，モデルが上回る例もあったが，差がある結果となった．

今後の課題は，コンテキストの発話数が 6 と少ないことによる情報不足の改善や訓練データの収集と構築，System の Instruction のチューニング，教員の発話シミュレーションの評価手法である．

謝辞

本研究は JSPS 科研費 JP23K11378 の助成を受けたものです.

参考文献

[1] 秋田喜代美. 変貌する教育学, 教師教育から教師の学習過程研究への転回―ミクロ教育実践研究への変貌―, pp. 45–75. 世織書房, 2009.

[2] 坂本篤史. 授業研究の事後協議会における教師の省察過程の検討 − 授業者と非授業者の省察過程の特徴に着目して − . 教師学研究, Vol. 9, No. 8, pp. 27–37, 2010.

[3] 保森智彦. 算数の授業中と省察の発話プロトコル分析をとおした教師の pck の検討. 日本教科教育学会誌, Vol. 41, No. 1, pp. 59–71, 2018.

[4] 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.

[5] Nicholas Hunkins, Sean Kelly, and Sidney D'Mello. "beautiful work, you're rock stars!": Teacher analytics to uncover discourse that supports or undermines student motivation, identity, and belonging in classrooms. In **LAK22: 12th International Learning Analytics and Knowledge Conference**, LAK22, p. 230–238, New York, NY, USA, 2022. Association for Computing Machinery.

[6] アイエドゥンエマヌエル, 林佑樹, 瀬田和久. 会話エージェントと学習支援. 教育システム情報学会誌, Vol. 36, No. 4, pp. 221–232, 2019.

[7] 小竹原祐希, 角所考, 西口敏司, 飯山将晃, 村上正行. 講義映像に基づく受講者の多様な状況認識のための挙動のクラスタリング. 教育システム情報学会誌, Vol. 37, No. 2, pp. 120–130, 2020.

[8] 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.

[9] 保森智彦. 省察方法のデジタル化に関する一考察 : 「主体的・対話的で深い学び」の観点から, pp. 3–11. 学習開発学研究, No. 14. 広島大学大学院人間社会科学研究科学習開発学領域, 2022.

[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] 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.

[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] Sakuei Onishi, Tomohiko Yasumori, and Hiromitsu Shiina. Classroom utterance analysis using a generative deep neural networks for dialogue model. In **2023 14th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI)**, pp. 560–565, Los Alamitos, CA, USA, jul 2023. IEEE Computer Society.

[14] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023.

[15] Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In **International Conference on Learning Representations**, 2022.

[16] Tianyu Zhao and Kei Sawada. rinna/youri-7b-chat.

[17] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms, 2023.

[18] 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.

[19] Matt Post. A call for clarity in reporting BLEU scores. In **Proceedings of the Third Conference on Machine Translation: Research Papers**, pp. 186–191, Belgium, Brussels, October 2018. Association for Computational Linguistics.

[20] 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.