大規模言語モデルを用いた講義振り返りテキストからの学生の成績推定

青野広太郎　　笹野遼平
名古屋大学大学院情報学研究科
aono.kotaro.i1@s.mail.nagoya-u.ac.jp, sasano@i.nagoya-u.ac.jp

概要

本研究では、講義を受けた学生が質問に回答する形式で講義内容について記述したテキストである「振り返りテキスト」を用いた学生の成績推定に取り組む。具体的には、2 人の学生の初回講義の振り返りテキストから、最終的にどちらがより高い成績を収めるかを予測する成績上下予測を、Meta-Llama-3.1-8B、Llama3.1-Swallow-8B-v0.2、Llama-3-ELYZA-JP-8B、Ruri-large の 4 つの言語モデルを利用して行う。九州大学の 3 つの講義における振り返りテキスト及び成績データを使用した実験の結果、複数のモデルの出力を多数決によって集約することで、およそ 60%の精度で成績の上下を判別することができた。

1　はじめに

大学教育において、学習意欲の低下や成績不振による学生のドロップアウトを防止するために、早期に学習状況を把握し、適切なサポートを提供することは重要である。そのため、累積 GPA[1, 2] や出席 [3]、課題の成績 [4] など、試験以外の様々な要素を用いて学生の成績を予測する取り組みが行われている。しかし、出席データは十分な量が蓄積されるまでは有効な予測材料にはなりにくく、また課題の採点には多大な労力が必要であり、受講者および採点者の負担が大きい。そこで、本研究では成績予測の根拠として「振り返りテキスト」に注目する。振り返りテキストとは、学生が講義について表 1 に示す 5 つの質問に答える形式で記述したテキストである。毎回の講義で学生自身が理解した内容や理解できなかった点が記述されており、これらは成績を予測する上で重要な手がかりとなるとともに、その内容を基にしたフィードバックの提供が可能であると考えられる。しかし、振り返りテキストは成績に直接結びつく記述ではないため、人間が内容を一目で見て成績を予測することは困難である。また、振り返りテキストのような自由記述文が成績予測にとって有用な情報を含むかどうかは、現時点では十分な分析はされていない。

表 1　振り返りテキストにおける 5 つの質問

番号	質問内容
1	今日の内容を自分なりの言葉で説明してみてください
2	今日の内容で、分かったこと・できたことを書いてください
3	今日の内容で、分からなかったこと・できなかったことを書いてください
4	質問があれば書いてください
5	今日の授業の感想や反省を書いてください

近年、GPT-4 [5] や Llama-3 [6] などの大規模言語モデル (LLM) はテキスト生成タスクにおいて高い性能を示し、これらのモデルの言語理解能力が注目を集めている。また、成績推定タスクにおいても言語モデルを利用する試みが行われており、例えば Qu ら [7] は BERT [8] を特徴抽出器として利用し、講義に対するコメントを成績推定のための特徴量の 1 つとして活用している。Llama-3 などの LLM は BERT よりも高次元の埋め込みを利用しているため、より多くの情報を抽出できると考えられる。また、LoRA（Low-Rank Adaptation）[9] などの効率的なファインチューニング手法により、計算コストを抑えつつタスクに特化したモデルを構築することが可能となっている。

そこで、本研究では、振り返りテキストに含まれる情報を LLM を利用して抽出し、成績推定を試みる。具体的には、2 人の学生の初回講義の振り返りテキストを LLM に入力することによって、2 人のうちどちらの最終的な成績が上になるか予測する、成績上下推定に取り組み、振り返りテキストが成績予測にとって有用な情報を含んでいるか検証する。

表 2　講義ごとの成績分布

成績	講義 1	講義 2	講義 3
A	36	41	151
B	38	177	16
C	12	102	12
D	2	20	7
F	3	31	39
合計	91	371	225

2　振り返りテキストからの成績推定

2.1　データセット

本研究では、九州大学の 3 つの講義における振り返りテキストと成績データを使用する。本稿では、各講義をそれぞれ講義 1、講義 2、講義 3 と呼ぶ。講義 1 はサイバーセキュリティに関する講義であり、パスワード設定から情報倫理や暗号まで、サイバーセキュリティに関する内容を扱っている。講義 2 は情報科学に関する講義であり、情報通信技術からデータサイエンスや人工知能まで、情報に関わる幅広い技術の基礎理論を学ぶ講義である。講義 3 は信号処理に関する講義であり、デジタル信号の変換、伝達に用いられる技術について演習を交えながら進めていく講義である。

各講義における成績の分布を表 2 に示す。各講義では全 15 週分の振り返りテキストが収集されているが、本研究では早期検出を目的とし、第 1 回講義のテキストのみを利用する。表 4 に講義 1 において成績の良かった学生 X、良くなかった学生 Y の振り返りテキストの例をそれぞれ示す。学生 X は質問 1、「今日の内容を自分なりの言葉で説明してみてください」に対して講義内容に沿った説明をしているが、学生 Y は講義内容の説明というよりは講義に対する感想を述べている。また、質問 3、「今日の内容で、分からなかったこと・できなかったことを書いてください」に対しても、学生 X は回答で講義内容に触れているが、学生 Y は講義内容には触れず自身の反省を述べている。

本研究における実験では、講義ごとに、学生を単位とする 5 分割交差検証を実施する。具体的には、講義ごとに学生を学習用、開発用、テスト用に 3:1:1 の割合で分割し、各分割内で、各学生とその学生とは異なる成績であった学生をランダムな 10 人選び、それらの学生のデータを組み合わせ、各組み合わせの成績の上下を予測する。この際には、第一回の振り返りテキストに何も記入していない生徒は除外した。また、学生の分割後にある学生と異なる成績を持つ学生が 10 人存在しない場合は、存在する全員と組み合わせた。表 3 に講義ごとの学習用、開発用、テスト用データ数の平均値と分割された学生の数を示す。

表 3　講義ごとのデータ数の平均値 (括弧内は分割された学生の人数を表す)

分割	講義 1	講義 2	講義 3
訓練	449 (54)	2087 (218)	1193 (133)
開発	94 (18)	634 (73)	309 (44)
テスト	94 (18)	634 (73)	309 (44)

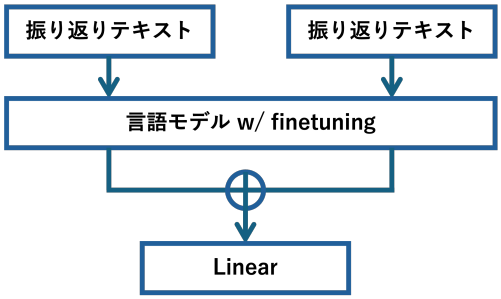

図 1　分類モデルの概観

2.2　モデル

成績上下予測モデルの概要を図 1 に示す。比較する二人の学生の初回講義の振り返りテキストを、それぞれファインチューニングした言語モデルに入力し、それぞれの埋め込みを組み合わせた後、全結合層に入力して判別を行う。本研究では、言語モデル部分に以下の 4 つのモデルを利用した。

- Meta-Llama-3.1-8B [6]
- Llama-3.1-Swallow-8B-v0.2 [10, 11]
- Llama-3-ELYZA-JP-8B[1)]
- Ruri-large [12]

ELYZA-llama3-8B と Swallow-13B は Llama をベースに日本語に特化した継続事前学習を行ったモデルである。Llama などの単方向言語モデルの文末トークンは文全体の情報を考慮できる[2)] ことから、本研究では振り返りテキストの末尾トークンの最終層における表現をその埋め込みと解釈する。Ruri-large は BERT [8] を基盤にに日本語での大規模事前対照学習を施した日本語汎用埋め込みモデルであり、分類タスクにおいて高い性能を示す。Ruri-large の埋め込みは 1024 次元、それ以外のモデルの埋め込みは 4096 次元である。

1）　https://huggingface.co/elyza/Llama-3-ELYZA-JP-8B
2）　https://github.com/hppRC/llm-lora-classification

表 4　成績の良あった学生 X の振り返りテキスト (左) と成績の良くなかった学生 Y の振り返りテキスト (右) の例

質問番号	学生 X の回答	学生 Y の回答
1	サイバーインシデントとその対策としてのパスワードやセキュリティバイデザインについて。	オンライン授業での初回ということもあり不安な点はあったが、今後の進め方についても理解でき授業への意欲が高まった。
2	様々なサイバーインシデントとそれに対する企業や政府の取り組みを学んだ。そしてパスワードを変える重要性とセキュリティバイデザインのメリットとデメリットを知った。	ipa という組織がセキュリティに関する情報を発信していることや、パスワードを変えることの重要性を学んだ。
3	政府の取り組みの具体例がよく分からなかった。	事前の学習で線を引くなどの予習ができていなかったので次回からは授業前にしっかり資料を確認したい。
4	[無回答]	特に無いです。
5	事前に資料をもっと詳しく見ておくとより理解出来ただろうと感じたので、次回からは詳しく見ていきたい。	オンライン授業ではあるが、集中力など切れないように今後も学習していきたいと思う。

3　実験

3.1　実験設定

プロンプト　プロンプトとして振り返りテキストのうち回答部分のみを入力した。また、本タスクでは振り返りテキストの文字数情報が大きな手掛かりとなることから、末尾に「文字数は n 文字です。」という形式で、振り返りテキストの文字数情報を付け加えた設定でも実験を行った。

全結合層への入力　全結合層には、2 つの振り返りテキスト埋め込み (v_1, v_2) と差分 $(|v_1 - v_2|)$ を連結して入力した。

ファインチューニング　Ruri-large 以外の言語モデルの訓練には LoRA[9] を用い、Ruri-large についてはフルファインチューニングを実施した。各モデルの訓練可能な全層を、講義ごとに 10 エポックずつファインチューニングし、開発データにおける正解率が最も高かったエポックでのモデルをテストに用いた。バッチサイズは Ruri-large 以外のモデルでは 16、Ruri-large では 32 とし、学習率は Ruri-large 以外のモデルでは 5×10^{-4}、Ruri-large では 5×10^{-5} とした。また、最適化手法として Adam [13] を採用した。

評価方法　モデルの性能は正解率で評価した。データ分割による性能の揺れを抑えるため、分割方法を変えて 5 分割交差検証を 3 回行い、得られた正解率のマイクロ平均を評価に用いた。

ベースライン　ベースラインとして、文字数の多い振り返りテキストを書いた学生の方が成績が高いとする文字数ベースラインを採用した。たとえば、前述の学生 X の振り返りテキストは 192 文字、学生 Y は 203 文字であるため、学生 Y の方が成績が良いと判定する。この場合は、実際には学生 X の方が成績が良いことから、文字数ベースラインの判定は誤りとなる。

統合モデル　深層学習に基づくモデルでは、複数のモデルを統合することで、より高い性能が得られる場合があることが知られている。そこで、文字数ベースラインと 4 つの言語モデルに基づくモデルの合わせて 5 つのモデルの出力の多数決により最終的な判定を行う、多数決モデルを用いた評価も実施した。

有意差検定　多数決モデルを除くモデルの判別結果が文字数ベースラインより統計的に有意に優れているかを検証するため、試行回数 10 万回の並び替え検定を行った。有意水準は 5% とし、多重比較の補正にはボンフェローニ法 [14] を用いた。具体的には、文字数情報を使う場合と使わない場合、それぞれ 4 つの言語モデルを用いた実験を行うことから、有意水準を 8 で割った 0.625%に補正して判定した。

3.2　実験結果

表 5 にプロンプトに振り返りテキストの回答部分のみを利用した場合と、文字数情報を付加した場合、それぞれの場合の正解率を示す。下線を付した正解率は、文字数ベースラインとの差を検定した結果、p 値が補正後の有意水準である 0.625%未満となり、文字数ベースラインより有意に高い正解率であると判定されたことを示す。

Ruri-large はどの設定でも文字数ベースラインを超え、単なる文字数情報では捉えきれない成績判定

表 5　成績上下推定の正解率 (± の後の数字は標準偏差を表す)

モデル	講義 1	講義 2	講義 3	平均	p 値
文字数	0.549±0.002	**0.611**±0.011	0.517±0.006	0.559±0.048	
文字数情報なし					
Meta-Llama-3.1-8B	0.573±0.050	0.556±0.040	0.518±0.024	0.549±0.028	1.000
Llama3.1-Swallow-8B-v0.2	0.635±0.003	0.545±0.014	0.502±0.042	0.561±0.068	0.418
Llama-3-ELYZA-JP-8B	0.601±0.029	0.538±0.002	0.481±0.013	0.540±0.060	1.000
Ruri-large	0.640±0.020	0.583±0.014	0.506±0.033	0.576±0.067	0.004
多数決	0.648±0.009	0.589±0.015	0.501±0.028	0.577±0.061	
文字数情報あり					
Meta-Llama-3.1-8B	0.583±0.044	0.531±0.030	**0.557**±0.02	0.557±0.026	1.000
Llama3.1-Swallow-8B-v0.2	0.640±0.008	0.540±0.015	0.548±0.015	0.576±0.056	0.008
Llama-3-ELYZA-JP-8B	0.621±0.006	0.515±0.028	0.546±0.026	0.561±0.055	0.416
Ruri-large	**0.648**±0.026	0.581±0.022	0.507±0.038	**0.579**±0.071	0.002
多数決	0.662±0.007	0.568±0.016	0.554±0.020	0.594±0.048	

に有用な情報を抽出できていると考えられる。LLM ベースのモデルも一部は設定次第で文字数ベースラインを上回り、講義によっては Ruri-large を上回る精度を示す場合もあったが、平均では Ruri-large が最も高い精度であった。統合モデルは文字数情報ありの設定で精度が大きく向上し、平均で 0.6 近い精度となった。

講義別の精度に着目すると、講義 2 は文字数ベースラインが強いにもかかわらず、文字数情報の追加による性能向上は見られなかった。一方、講義 3 は文字数ベースラインは弱いものの、文字数情報を取り入れることで性能が向上した。この差には付録 A.1 の表 7 に示す文字数と成績の関係が影響していると考えられる。講義 1 および 2 では文字数と成績に弱い正の相関があるが、講義 2 では成績 B の平均文字数が A を上回るなど例外的なパターンが散見されるため、文字数単独の予測はある程度可能であるが、モデルの判別には十分寄与しなかったと推察される。一方、講義 3 では、D や F の平均文字数が相対的に多く、文字数が増えるほど成績が下がる傾向がわずかに見られるため、モデルがその特徴を学習し、文字数情報の追加で性能が向上したと考えられる。

3.3　各質問項目の重要性

振り返りテキスト中の質問 1〜5 のうちどの質問が最も精度に影響を及ぼすかを調査するために、3.2 節で最も判別精度が高かった設定である Ruri-large に文字数情報を付与した設定に対し、質問を 1 つずつ除外して実験するアブレーション分析を行った。結果を表 6 に示す。

全体的に質問 1 または 2 を削除した場合に正解率が低下したことから、これらの質問が成績予測に重要である可能性が高いと考えられる。講義別の傾向としては、講義 1 では質問 1「今日の内容を自分なりの言葉で説明してみてください」を削除することによる影響、講義 2、3 では、質問 2「今日の内容で、分かったこと・できたことを書いてください」を削除することによる影響がもっとも大きかった。

表 6　アブレーション分析の結果

質問	講義 1	講義 2	講義 3	平均
1	0.522±0.036	0.585±0.008	0.507±0.040	0.538±0.041
2	0.607±0.044	0.558±0.012	0.494±0.018	0.553±0.057
3	0.640±0.039	0.604±0.017	0.521±0.032	0.588±0.061
4	0.643±0.043	0.572±0.023	0.510±0.017	0.575±0.067
5	0.663±0.027	0.564±0.028	0.504±0.026	0.577±0.080
なし	0.648±0.026	0.581±0.022	0.507±0.038	0.579±0.071

4　おわりに

本研究では、大規模言語モデルを用いて、大学講義の初回振り返りテキストを用いて、2 名の学生の最終的な成績の上下を推定した。実験の結果、単一のモデルでは Ruri-large モデルに振り返りテキストおよびその文字数を入力する設定が最も高い予測精度を示した。また、文字数を入力し、かつ複数のモデルを組み合わせた統合モデルを利用することによっておよそ 60%の精度を達成した。本研究の成果は、振り返りテキストという成績に直結しない情報源からでも、成績予測において埋め込みモデルの有効性を示すものである。今後の研究としては、まず、より最近実施された講義を含むデータセットを作成することで、さらに正確かつ汎用性の高い分析を行えると考えられる。次に、より大規模かつ高性能な埋め込みモデルを用いて学習を行うことで、さらなる性能向上が期待される。

謝辞

本研究は、JST CREST（課題番号：JPMJCR22D1）の支援を受けて実施しました。また、実験に使用したデータは、代表機関である九州大学から匿名化されたものを提供いただきました。

参考文献

[1] D.Magdalene Delighta Angeline. Association rule generation for student performance analysis using apriori algorithm. **The SIJ Transactions on Computer Science Engineering its Applications (CSEA)**, Vol. 01, pp. 16–20, 2013.

[2] Mr. M. N. Quadri and Dr. N. V. Kalyankar. Drop out feature of student data for academic performance using decision tree techniques. **Global journal of computer science and technology**, Vol. 10, pp. 1–8, 2010.

[3] Naren.J, Elakia, Gayathri, and Aarthi. Application of data mining in educational database for predicting behavioural patterns of the students. **International Journal of Engineering and Technology**, Vol. vol 5(3), pp. 4469–4472, 2014.

[4] Suhem Parack, Zain Zahid, and Fatima Merchant. Application of data mining in educational databases for predicting academic trends and patterns. In **2012 IEEE International Conference on Technology Enhanced Education (ICTEE)**, pp. 1–4, 2012.

[5] OpenAI. Gpt-4 technical report. **arXiv:2303.08774**, 2024.

[6] Llama Team: AI@Meta. The llama 3 herd of models. **arXiv:2407.21783**, 2024.

[7] Yubin Qu, Fang Li, Long Li, Xianzhen Dou, and Hongmei Wang. Can we predict student performance based on tabular and textual data? **IEEE Access**, Vol. 10, pp. 86008–86019, 2022.

[8] 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 (NAACL-HLT 2019)**, pp. 4171–4186, 2019.

[9] 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 (ICLR)**, 2022.

[10] Kazuki Fujii, Taishi Nakamura, Mengsay Loem, Hiroki Iida, Masanari Ohi, Kakeru Hattori, Hirai Shota, Sakae Mizuki, Rio Yokota, and Naoaki Okazaki. Continual pre-training for cross-lingual llm adaptation: Enhancing japanese language capabilities. In **Proceedings of the First Conference on Language Modeling (COLM)**, 2024.

[11] Naoaki Okazaki, Kakeru Hattori, Hirai Shota, Hiroki Iida, Masanari Ohi, Kazuki Fujii, Taishi Nakamura, Mengsay Loem, Rio Yokota, and Sakae Mizuki. Building a large japanese web corpus for large language models. In **Proceedings of the First Conference on Language Modeling (COLM)**, 2024.

[12] Hayato Tsukagoshi and Ryohei Sasano. Ruri: Japanese general text embeddings. **arXiv:2409.07737**, 2024.

[13] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. **International Conference on Learning Representations (ICLR)**, 12 2014.

[14] Carlo Bonferroni. Teoria statistica delle classi e calcolo delle probabilita. **Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commericiali di Firenze**, Vol. 8, pp. 3–62, 1936.

A Appendix

A.1 文字数と成績の関係

ここでは、本研究で特徴量として利用した文字数と成績の関係について示す。まず、表 7 に、成績 A、B、C、D、F それぞれの平均文字数、さらに成績をそれぞれ数値 4、3、2、1、0 に割り当てたときの文字数と成績の相関係数を示す。また、ここでは第一回講義の振り返りテキストに何も記述していない生徒は除外して計数している。

表 7 各講義における成績ごとの平均文字数と、文字数と成績の相関係数 (括弧内は人数を表す)

成績	講義 1	講義 2	講義 3
A	290 (36)	258 (41)	121 (149)
B	240 (38)	267 (176)	84 (16)
C	211 (12)	188 (100)	98 (12)
D	155 (2)	157 (20)	139 (6)
F	248 (2)	217 (26)	138 (38)
相関係数	0.150	0.187	-0.074