文法項目の多様性と誤り情報を利用したエッセイ自動採点

土肥 康輔　須藤 克仁　中村 哲
奈良先端科学技術大学院大学
{doi.kosuke.de8, sudoh, s-nakamura}@is.naist.jp

概要

本研究では，エッセイ自動採点において文法特徴を考慮することの効果について検証した．エッセイの分散表現に加えて，学習者が正しく使えている文法項目の情報や，文法誤りの情報をモデルの入力に用いることで，モデル性能が向上した．また，文法特徴の利用に加えて，エッセイの総合スコアと文法スコアでマルチタスク学習を行うと，モデル性能がより大きく向上した．しかし，学習者が正しく使えている文法項目の情報と文法誤りの情報を組み合わせて用いても，それぞれを単独で用いたときと同等のモデル性能となり，両者を同時に用いることの相乗効果は見られなかった．

1　はじめに

エッセイ自動採点 (AES) は，学習者が書いたエッセイにスコアやレベル等の評価値を付与するタスクである．外国語教育においては，近年のライティングやスピーキングといった産出能力の育成も重視する流れを受けて，エッセイ課題が用いられる場面も増えている．しかし，エッセイの採点には，大きな人的・時間的コストがかかる．採点自体に時間がかかることに加えて，複数の採点者間での評価の信頼性を保つためには，採点者の訓練が必要である．また，採点者が 1 名であっても，疲労などから同一のエッセイに異なる評価をする場合があることが報告されている．このような課題を背景に，自動採点研究が行われている．

エッセイ課題の採点方法は，エッセイに対して 1 つのスコアを付与する総合的評価と，文法や内容等の複数の観点に対してそれぞれスコアを付与する分析的評価に大別される．ただし，総合的評価においても，分析的評価で用いられる観点での出来栄えが基準となっていることが一般的である．

それらの観点の中で，土肥ら [1] は，CEFR の基準特性研究 [2] を参考に，総合スコアを推定する AES モデルに文法特徴を明示的に利用する手法を提案した．具体的には，BERT [3] により得られるエッセイの分散表現に，文法誤りの情報，または学習者が正しく使えている文法項目の情報を結合したものを全結合層への入力とし，エッセイの総合スコアを推定した．[1] では文法誤りの数をエッセイの分散表現を入力として推定していたが，その推定精度に課題があった．また，正しく使えている文法項目の情報は，モデル性能の向上につながらなかった．

そこで本研究では，文法誤り訂正モデル [4] の訂正結果に基づく文法誤り数を特徴量として用いる．[1] で正しく使えている文法項目の情報の効果がなかった原因としては，(1) エッセイの分散表現（768 次元）に対して特徴量の次元数（最大で 25 次元）が小さかったこと，(2) 実験に使用したデータセットは複数のエッセイ課題に基づく答案が含まれており，使用される文法項目が学習者の習熟度よりも，エッセイ課題の影響を強く受けていたことが可能性として挙げられる．そこで本研究では，256 次元の新たな特徴量を用い，Automated Student Assessment Prize (ASAP)[1] を実験に用いる．特徴量とデータセットの詳細については，それぞれ 3 節，4.1 節で述べる．

2　関連研究

初期の AES モデルでは，人手で作成された特徴量が用いられていた ([5, 6] を参照)．例えば e-rater [7] は，文法誤りや語彙の複雑さの指標を含む 12 種類の特徴量を用いている．[8] は様々な言語特徴の中で，文法の複雑さや誤りに関する指標に高い重みが割り振られていたことを報告している．[9] は文法誤り検出タスクとのマルチタスク学習を行うことで，AES の性能が向上することを示した．

近年では，深層学習に基づく手法が主流となっている．RNN や Bi-LSTM を用いたモデル [10, 11] や，BERT 等の事前学習済み言語モデルに基づくモデル [12, 13, 14, 15] が提案されている．また，深層学習

1） https://www.kaggle.com/c/asap-aes

に基づくアプローチと人手で作成された特徴量を組み合わせたモデルも提案されている [16, 17]．[18] は，BERT から得られるエッセイの分散表現に文法誤りの情報を組み合わせたモデルを提案した．しかし，文法誤り情報の追加は，話し言葉自動採点でモデル性能を向上させた一方で，AES では効果がなかった．[1] は文法誤り情報，および正しく使えている文法項目の情報を用いたが，後者はモデル性能の向上につながらなかった．

大規模言語モデルを活用した研究も行われており，[19] は GPT-3 で言語特徴量を用いることで，採点精度が向上することを示した．[20] は，GPT-4 に少数の採点例を与えることで，85 の言語特徴に基づき数十万のデータで学習したモデルと同等の性能を得られることを報告している．

本研究は，深層学習と人手で作成した特徴量を組み合わせるアプローチにおいて，[1, 18] で用いられていた文法特徴量を改良し，AES で文法特徴を明示的に考慮することの効果を検証するものである．

3 文法特徴量

3.1 CEFR と基準特性

CEFR [21] は言語能力を評価する国際指標で，習熟度を A1 (初級) 〜 C2 (上級) の 6 段階に区分する．どのような文法や語彙が CEFR のレベルごとに使えるようになっていくかは言語ごとに研究されており，英語では English Profile Programme によりそのような項目（基準特性）が明らかになっている [2]．基準特性は CEFR の各レベルで特徴的な言語項目群であり，あるレベル以上の学習者が正しく使うことができる positive linguistic features (PF) と，あるレベルの学習者が間違えやすい negative linguistic features (NF) がある．人間の採点者は，これらの項目を探しながら学習者のパフォーマンスを評価していると言われており [2]，AES モデルにおいて文法特徴を明示的に利用することで，モデル性能を向上させることが期待できる．以下の節で，実験に用いる PF と NF について具体的に説明する．

3.2 Positive Linguistic Features

PF の抽出には，CEFR-J Grammar Profile 文法項目頻度分析プログラム [22, 23] を用いる．同プログラムは，501 種類の文法項目のテキスト中での頻度を，正規表現に基づいて算出できる．[1] では，エッセイ中で使われない文法項目が多いことによりベクトルがスパースになることを防ぐため，最大で 25 次元に集計した特徴量を用いたが，モデル性能の向上にはつながらなかった．原因として，エッセイの分散表現に対して特徴量の次元数が小さかった可能性があるため，本研究では「CEFR-J Grammar Profile 教員版」に基づき集計した 256 次元のベクトルを特徴量として用いる (表 1 上段)．各次元は文法項目に対応しており，エッセイ中で使われていれば 1，使われていなければ 0 となる．

表 1 使用する文法特徴量
誤り数は 100 語あたりの相対頻度．

文法特徴量	説明
type256	使用の有無，256 項目
err24	誤り数，誤りタグで集計（24 種類）
err54	誤り数，誤りタグと誤りタイプの可能な組み合わせ（54 種類）

3.3 Negative Linguistic Features

NF には，100 語あたりの文法誤り数を用いる．具体的には，ERRANT [24] によって付与される誤りタグに基づき，24 次元 (err24)，および 54 次元 (err54) の特徴量を作成する (表 1 下段)．エッセイ中の文法誤りの訂正には GECToR-large [4] を用いる．

4 実験

実験では，[1, 18] と同様に，BERT に基づく AES モデルにおいて，文法特徴を用いることの効果について検証する．

4.1 データセット

実験には ASAP と ASAP++ [25] を用いる．ASAP には 8 つのエッセイ課題に関する答案が含まれており，課題 1〜6 には総合スコア，課題 7〜8 には観点別スコアが付与されている．課題 7〜8 では観点別スコアの合計が総合スコアとなる．データの概要を表 2 に示す．ASAP++は ASAP の課題 1〜6 の答案に対して観点別スコアを付与したデータセットである．

本研究では各エッセイ課題の総合スコアを予測するモデルを構築した．観点別スコアからは，文法に関連するスコア[2)]のみをマルチタスク学習の補助タスク（4.2 節参照）において用いた．

モデルの性能評価は，先行研究に従い 5 分割交差検証を用いて行った．[10] の分割に基づき，60%，

2) 課題 1〜2，7〜8 は Conventions，課題 3〜6 は Language.

表 2　ASAP データセットの概要

課題	エッセイ数	スコア範囲
1	1,783	2-12
2	1,800	1-6, 1-4[3]
3	1,726	0-3
4	1,772	0-3
5	1,805	0-4
6	1,800	0-4
7	1,569	0-30
8	723	0-60

20%，20%をそれぞれ学習，開発，テストデータとした．評価は課題ごとに独立して行い，評価値には 2 次の重み付きカッパ係数 (QWK) を用いた．

4.2　採点モデル

提案モデルは，エッセイの分散表現と文法特徴量を入力とし，エッセイの総合スコアを予測する．ベースラインとしては，エッセイの分散表現のみを入力とするモデルを準備した．エッセイの分散表現には BERT の CLS トークンを用いた．文法特徴量は，図 1 に示す 4 種類の設定でモデルの学習に利用した．cat はエッセイの分散表現と文法特徴のベクトルを結合したものを順伝播型ニューラルネットワーク (FFNN) への入力とした．net は文法特徴を FFNN に通した後でエッセイの分散表現と結合した．multi は net に加えて，総合スコアと文法スコアでマルチタスク学習を行った．multi の FFNN にはタスク固有層を設けず，共通層のみから成る構造とした[4]．dual は文法特徴のネットワークから文法スコアを推定する構造とした．

エッセイのスコアは −1 から 1 の範囲に正規化した．BERT は HuggingFace 社が公開している bert-base-uncased を用い，学習データの最大入力長は 512 トークンとした．損失関数は平均二乗誤差を用い，FFNN と BERT 層の両方のパラメータを更新した．文法特徴のネットワークの隠れ層の数は 3，隠れ層のノード数は文法特徴の次元数の 2 分の 1 とした．FFNN の隠れ層の数は，cat では {1, 2, 3, 4, 5, 7, 10}，net では {1, 2, 3}，multi と dual では {2, 3} を探索し，課題 1 の開発セットで QWK が最も高くなったものを選択した．FFNN の隠れ層のノード数は 512 とした．multi と dual での主タスクの損失

3)　課題 2 は Domain 1（内容や構成など）と Domain 2（文法など）の 2 種類のスコアが付与されている．ドメインごとにモデルを学習してそれぞれ評価し，その平均値が課題 2 の評価値となる．

4)　タスク固有層を持つ構造も試したが，共通層のみのときよりも QWK が低くなった．

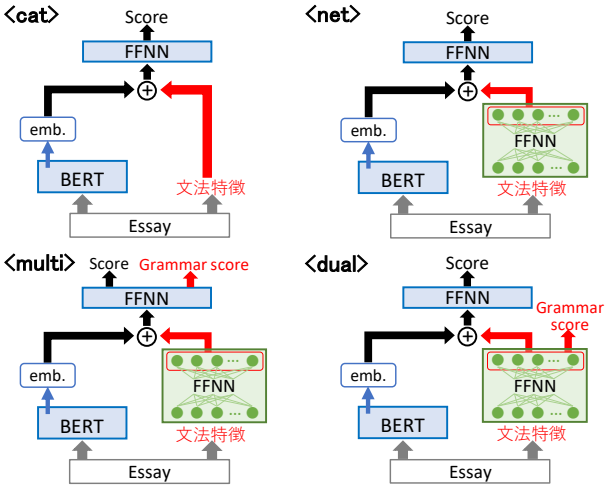

図 1　採点モデルでの文法特徴量の利用方法

関数の重みは {0.8, 0.6} を試した．文法特徴のネットワークと FFNN における活性化関数には relu を用いた．最適化アルゴリズムは Adam，学習率 1e-5，dropout = 0.2 とし，バッチサイズ {4, 8, 16, 32} で 10 エポック学習を行った．4.3 節では，各エッセイ課題において開発セットでの QWK が最も高くなったバッチサイズでの結果を報告する．

4.3　実験結果

4.3.1　モデル構造の探索結果

文法特徴量に type256 を用い，各モデル構造で最適なハイパーパラメータの探索を行った．表 3 は FFNN の隠れ層の数を変化させたときの，課題 1 の開発セットでの QWK の結果である．cat で隠れ層の数を 1 としたときの QWK (.792) は，ベースライン (.813) の値を下回っていた．QWK は隠れ層の数を 2 としたときに最も高くなったが，隠れ層の数を増やしていくと徐々に低下していった．net では FFNN の隠れ層の数を 2 としたときに QWK が最も高くなった．multi と dual では，ともに主タスクの重みを 0.8，FFNN の隠れ層の数を 3 としたときに QWK が最も高くなった．以降の実験では，これらのハイパーパラメータを用いて学習を行った．

4.3.2　各エッセイ課題での評価結果

PF の効果　各モデル構造で文法特徴量として type256 を用いたときの，テストセットにおけるエッセイ課題別の QWK スコアを表 4 に示す．type256 を用いることで，全エッセイ課題の平均 QWK スコアが，全てのモデル構造においてベース

表 3　FFNN の隠れ層の数による比較 (課題 1，QWK dev)

	# of hidden layers						
Model	1	2	3	4	5	7	10
cat	.792	**.825**	.814	.813	.801	.766	.722
net	.812	**.824**	.817	–	–	–	–
multi (0.8)	–	.819	**.827**	–	–	–	–
multi (0.6)	–	.804	.812	–	–	–	–
dual (0.8)	–	.816	**.824**	–	–	–	–
dual (0.6)	–	.820	.819	–	–	–	–

表 4　モデル構造による比較 (type256，課題別，QWK test)

	課題番号								
Model	1	2	3	4	5	6	7	8	avg.
baseline	.807	.659	.671	.805	.799	.803	.819	.749	.764
+ type256									
cat	.818	.675	.673	.819	.806	.808	.832	.734	.771
net	**.822**	**.684**	.685	.811	.804	.813	.834	.746	.775
multi	.812	.682	.694	.817	.808	**.814**	**.837**	.749	.777
dual	.820	.675	**.700**	**.820**	**.809**	.806	.831	**.763**	**.778**

ラインよりも向上した（表 4 の avg.）．エッセイ課題ごとのスコアを見ても，課題 8 の `cat`，`net`，`multi` を除く全てで QWK スコアが向上していることから，本研究で用いた PF はモデル性能の向上に効果があると考えられる．

また，[1] で用いられていた 9 次元の特徴量を ASAP データセットにおいて用いたとき，モデル性能が向上する場合があった（例：文法項目の延べ使用数を CEFR-J レベルごとに集計した特徴量，モデル構造=`dual`，課題 1=.827，課題 8=.755）．[1] で PF を用いてもモデル性能が向上しなかった原因として，本研究では (1) 特徴量の次元数が小さかったこと，(2) 実験に使用したデータセットに複数のエッセイ課題に基づく答案が含まれていたことの 2 つを仮説として設定していたが，上記の結果は (2) を支持するものである．すなわち，エッセイ課題が同一であれば，エッセイのスコア帯によって用いられている文法項目が異なっていることが示唆される．

`type256` を用いたとき，モデル構造の中では `dual` が最も高い QWK スコアを達成した（表 4）．そのため，以降の実験では `dual` を用い，NF 単独，および PF と NF を組み合わせたときの効果について検証を行った．

NF の効果　表 5 は，`dual` の設定で異なる文法特徴を用いたときの，テストセットにおけるエッセイ課題別の QWK スコアを示している．NF を単独で用いたとき，全エッセイ課題の平均 QWK スコアはベースラインよりも向上した（`err24` = .774，`err54` = .777）．また，エッセイ課題ごとのスコアを見ても，全ての課題で QWK スコアがベースラインよりも向上している．より詳細な誤り情報を持つ特徴量（err54）のほうが高いスコアを達成していることは，[1] での結果の傾向と一致している．

表 5　文法特徴量による比較 (dual，課題別，QWK test)

	課題番号								
Features	1	2	3	4	5	6	7	8	avg.
baseline	.807	.659	.671	.805	.799	.803	.819	.749	.764
type256	.820	.675	.700	.820	.809	.806	.831	.763	**.778**
err24	.814	.666	.690	.817	.809	.805	.833	.761	.774
err54	.820	.682	.677	.820	.803	.823	.836	.758	.777
type256 + err24	.821	.674	.686	.827	.799	.809	.833	.751	.775
type256 + err54	.821	.679	.690	.816	.816	.810	.835	.754	**.778**
Yang+ 2020 [13]	.817	**.719**	.698	.845	**.841**	**.847**	.839	.744	**.794**
Cao+ 2020 [14]	.824	.699	**.726**	**.859**	.822	.828	**.840**	.726	.791
Wang+ 2022 [15]	**.834**	.716	.714	.812	.813	.836	.839	**.766**	.791

PF + NF の効果　PF と NF を組み合わせたときも，全エッセイ課題の平均 QWK スコアはベースラインよりも向上した（`type256 + err24` = .775，`type256 + err54` = .778）．しかし，そのスコアは PF または NF を単独で用いたときと同程度となっており，PF と NF を両方用いることによる相乗効果は見られなかった．加えて，先行研究のモデル [13, 14, 15] に対しては，全エッセイ課題の平均 QWK スコアで .01〜.015 程度のビハインドがある．本研究では，PF と NF のベクトルを結合したものを文法特徴のネットワークの入力として全結合層に渡していたが，それぞれを別のネットワークに渡すなど，より効果的な PF と NF の組み合わせ方を検討することで QWK スコアを向上させることが今後の課題である．また，異なる文法特徴量で学習したモデルをアンサンブルすることも考えられる．

5　おわりに

本研究では，AES で文法特徴を考慮することの効果について検証し，PF と NF は共にモデル性能の向上に寄与することを示した．また，同一のエッセイ課題であれば，エッセイのスコア帯によって異なる文法項目が使われている可能性が示唆された．しかしながら，PF と NF を組み合わせたときの相乗効果は見られなかったため，より効果的な組み合わせ方を検討することが今後の課題である．また，文法誤りや使われている文法項目とスコアとの関係性や，提案モデルで文法特徴量に割り振られている重みを分析することで，学習者へより具体的なフィードバックを行う方法についても検討したい．

謝辞

本研究の一部は JSPS 科研費 JP21H05054，JST 科学技術イノベーション創出に向けた大学フェローシップ創設事業 JPMJFS2137 の助成を受けたものである．

参考文献

[1] 土肥康輔, 須藤克仁, 中村哲. エッセイ自動採点における文法特徴と学習者レベルの関係. 言語処理学会第 29 回年次大会発表論文集, pp. 211–216, 2023.

[2] John A Hawkins and Luna Filipović. **Criterial Features in L2 English: Specifying the Reference Levels of the Common European Framework**. Cambridge University Press, 2012.

[3] 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, 2019.

[4] Maksym Tarnavskyi, Artem Chernodub, and Kostiantyn Omelianchuk. Ensembling and knowledge distilling of large sequence taggers for grammatical error correction. In **Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 3842–3852, 2022.

[5] Zixuan Ke and Vincent Ng. Automated essay scoring: A survey of the state of the art. In **Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence**, 2019.

[6] Masaki Uto. A review of deep-neural automated essay scoring models. **Behaviormetrika**, Vol. 48, No. 2, pp. 459–484, 2021.

[7] Jill Burstein, Martin Chodorow, and Claudia Leacock. Automated essay evaluation: The criterion online writing service. **AI Magazine**, Vol. 25, No. 3, pp. 27–36, 2004.

[8] Sowmya Vajjala. Automated assessment of Non-Native learner essays: Investigating the role of linguistic features. **International Journal of Artificial Intelligence in Education**, Vol. 28, No. 1, pp. 79–105, 2018.

[9] Ronan Cummins and Marek Rei. Neural multi-task learning in automated assessment. **arXiv**, Vol. arXiv: 1801.06830, , 2018.

[10] Kaveh Taghipour and Hwee Tou Ng. A neural approach to automated essay scoring. In **Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing**, pp. 1882–1891, 2016.

[11] Dimitrios Alikaniotis, Helen Yannakoudakis, and Marek Rei. Automatic text scoring using neural networks. In **Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 715–725, 2016.

[12] Farah Nadeem, Huy Nguyen, Yang Liu, and Mari Ostendorf. Automated essay scoring with discourse-aware neural models. In **Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications**, pp. 484–493, 2019.

[13] Ruosong Yang, Jiannong Cao, Zhiyuan Wen, Youzheng Wu, and Xiaodong He. Enhancing automated essay scoring performance via fine-tuning pre-trained language models with combination of regression and ranking. In **Findings of the Association for Computational Linguistics: EMNLP 2020**, pp. 1560–1569, 2020.

[14] Yue Cao, Hanqi Jin, Xiaojun Wan, and Zhiwei Yu. Domain-Adaptive neural automated essay scoring. In **Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval**, pp. 1011–1020, 2020.

[15] Yongjie Wang, Chuang Wang, Ruobing Li, and Hui Lin. On the use of bert for automated essay scoring: Joint learning of multi-scale essay representation. In **Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 3416–3425, 2022.

[16] Tirthankar Dasgupta, Abir Naskar, Lipika Dey, and Rupsa Saha. Augmenting textual qualitative features in deep convolution recurrent neural network for automatic essay scoring. In **Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications**, pp. 93–102, 2018.

[17] Masaki Uto, Yikuan Xie, and Maomi Ueno. Neural automated essay scoring incorporating handcrafted features. In **Proceedings of the 28th International Conference on Computational Linguistics**, pp. 6077–6088, 2020.

[18] Stefano Bannò and Marco Matassoni. Cross-corpora experiments of automatic proficiency assessment and error detection for spoken English. In **Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)**, pp. 82–91, 2022.

[19] Mizumoto Atsushi and Eguchi Masaki. Exploring the potential of using an ai language model for automated essay scoring. **Research Methods in Applied Linguistics**, Vol. 2, No. 2, p. 100050, 2023.

[20] Kevin P. Yancey, Geoffrey Laflair, Anthony Verardi, and Jill Burstein. Rating short L2 essays on the CEFR scale with GPT-4. In **Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)**, pp. 576–584, 2023.

[21] Council of Europe. **Common European Framework of Reference for Languages: Learning, Teaching, Assessment**. Cambridge University Press, 2001.

[22] 投野 由紀夫 (編) . 平成 24 年度～平成 27 年度科学研究費補助金 (基盤研究 (A)) 研究課題番号 24242017 研究成果報告書 学習者コーパスによる英語 CEFR レベル基準特性の特定と活用に関する総合的研究. 2016.

[23] 石井 康毅. CEFR-J Grammar Profile 文法項目頻度分析プログラム. http://www.cefr-j.org/download.html (2022 年 10 月 20 日ダウンロード), 2020.

[24] Christopher Bryant, Mariano Felice, and Ted Briscoe. Automatic annotation and evaluation of error types for grammatical error correction. In **Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 793–805, 2017.

[25] Sandeep Mathias and Pushpak Bhattacharyya. ASAP++: Enriching the ASAP automated essay grading dataset with essay attribute scores. In **Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)**. European Language Resources Association (ELRA), 2018.