論理構造グラフを用いた自動採点モデル

加藤　嘉浩
ベネッセ教育総合研究所
y-kato@mail.benesse.co.jp

概要

本研究では，エッセイの論理構造のグラフ表現を考慮した自動採点モデルを提案する．エッセイの論理構造は採点結果に影響する要因の一つと考えられ，これまでに論理構造の要素（主張や前提となる節や句，単語の連なり．以降，論理要素）の統計量や要素間の関係を考慮したモデルが提案されている．しかし，従来手法の論理要素とその関係を提案モデルに入力してもあまり精度が向上しなかった．本研究では，論理要素を含む文全体を論理要素として用い，文間の関係を推定し線形計画法で論理構造グラフを抽出する．そして，グラフ構造を考慮可能な Graph Attention Network により特徴量を抽出し自動採点モデルに結合する．ASAP データセットでの実験から提案モデルはベースラインおよび従来手法よりも高精度な結果となった．また，論理構造推定手法を検証するため，ランダム構造やグラフのエッジ削除前の構造を入力した場合とも比較することで提案手法の有効性を示した．

1　はじめに

エッセイのように学習者に自身の考えを表現させる解答構築型テストが古くから多くのテストや教材で利用されている．しかし，採点および指導にかかるコストが高く，採点品質（精度や再現性）の向上と効率化（採点者の訓練費用と人件費の低減，採点期間短縮）が求められる．そのため，自然言語処理や機械学習手法による自動採点を使いこれを実現することが考えられてきた．計算機に採点を代行させる自動採点研究の歴史は古く，これまでに多くの手法が提案されている [1, 2, 3].

自動採点は，記述式解答データをテキスト化したもの（以降，解答データ）と採点者がつけた得点のペアを入力データとして用いる．自動採点の基本的な方法は解答データの特徴量（例えば，単語数，語彙数，文章数などの統計量）を抽出し，分類や回帰モデルを用いて得点を予測する [4]．深層学習手法の進歩が目覚ましい近年では，Transformer[5] や BERT[6] などの言語モデルを用いて入力文章を固定長ベクトル（文レベルの埋め込み表現，文脈ベクトル）にし，分類器や回帰モデルに入力し得点を予測する手法 [7, 8, 9] が提案されている．また，従来手法のように解答データから推定した統計量（単語数や語彙数など）を追加するモデル [10, 11, 12]，Convolutional Neural Network（CNN）や Long Short Term Memory (LSTM）など別の Neural Network(NN) からの出力を結合したモデル [13] が提案されている．

本研究では，言語モデルを用いた自動採点モデルを基礎とし，そこに補助情報（統計量や別の NN からの特徴量）を追加することで精度向上を目指す．自動採点モデルに追加する補助情報は，特徴量ベースの自動採点モデル [4] や補助情報を追加したモデル [10, 11, 12, 13, 14] が参考になる．特徴量ベースの自動採点モデルでは，単語数，語彙数，誤字や文法誤りの割合，話題，得点が高い解答データとの類似度などが挙げられる．

エッセイのような長文は複数の文章を含み，文章数が多くなるほど構成や論理構造などが複雑になり得点に大きく影響すると考えられる．そこで，本研究では，論理構造の情報を自動採点モデルに活用することで，エッセイの得点の予測精度を向上させたい．Nguyen と Litman[10] は，論証マイニングを用いてエッセイの論理要素を推定し，自動採点に活用する手法を提案している．ただし，このモデルは主張や前提の数などの統計量をモデルに追加しており，論理構造は利用していない．Yamaura ら [14] は，言語モデルの Attention に論理構造情報を作用させた言語モデルと文脈ベクトルを結合することで精度向上を達成している．

本研究では，論証マイニングを用いて解答データから論理構造のグラフ情報を推定し，そのグラフ構造を特徴として用いることができる Graph Attention

Network（GAT）を用いた自動採点モデルを提案する．本研究での実験では，従来の論理構造推定手法で得られる論理構造グラフを用いても高精度な結果とはならなかった．従来手法では節や句，単語の連なりを論理要素とするが，意味をなさない単語の連なりが推定されることが多かった．そこで，論理要素を含む文を論理要素として，文間の関係を推定し線形計画法により論理構造グラフを決定し提案モデルに入力する．

本研究は，解答データの論理構造に着目する点でNguyen と Litman[10] や Yamaura ら [14] のモデルと類似している．特に，Yamaura ら [14] のモデルと同様に言語モデルの出力に論理構造の情報を追加する点が類似している．本研究は論理構造の推定方法と論理構造グラフを GAT の入力とする点が従来手法と異なる．著者の調査では，これまでに論理構造のグラフを用いた手法は自動採点では確認できなかった．

ベンチマークデータで提案モデルの精度検証を行い，ベースモデルおよび従来モデルよりも高精度な結果となることを確認した．ベンチマークデータには論理構造の正解データが付与されておらず提案手法が論理構造が正しく推定できているか確認できないため，ランダムな論理構造を用いた場合も検証し，提案手法の方が高精度な結果となったことから有効性を示した．

本研究の貢献は以下の通りである．

1. 言語モデルに論理構造グラフを用いた自動採点モデルを提案した．
2. 論理構造グラフを用いた自動採点モデルがベースモデルおよび従来モデルよりも高精度な結果を得た．

2 関連研究

2.1 Argument Mining

論理構造は，論証マイニング（Argument Mining）と呼ばれる手法で推定する．論証マイニングは，文章や節などを主張（Claim）や前提（premise）といった論理要素を推定し，支持（support），否定（attack），無関係（neutral）など要素間の関係性を推定することで論理構造を推定する．論理要素をグラフの点（ノード），ノード間を結ぶ矢印（エッジ）の重みを構造推定時の推定値とし，エッセイの論理構造をグラフ化できる．一般的に，論理構造は以下の手順で推定される [15, 16]．（１）論理要素を推定しノードと見做す要素抽出（Argument Component Identification）を行い，（２）論理要素を結論（Major Claim），主張（Claim），前提（Premise）に分類する要素分類，（３）ある論理要素が他の論理要素を支持する場合，ノード間に有向エッジを引く（否定の場合は逆向きのエッジ）関係分類，（４）要素分類と関係分類の結果から論理構造を推定する．これまでに（１）から（４）を独立なタスクと見做すもの [15, 17, 16]，すべてを 1 つのモデルで行うものが提案されている [18, 19, 20]．

本研究では，従来手法と独自の方法それぞれで論理構造を推定し，その結果を自動採点モデルに入力し比較検証した．従来手法として Morio ら [19] の方法を採用した．Morio らは，より長い入力文章を扱える Longformer を用いて要素分類（要素抽出含む）と関係分類を同時に行うモデルを提案し，複数のデータセットでファインチューンすることで従来モデル [18] よりも高精度な結果を報告している．

従来手法は節や句だけでなく数語の意味をなさないものも推定される．しかし，それらを構成要素とする論理構造が，エッセイの得点に影響を及ぼすとは考えにくい．そのため，本研究では，節や句ではなく文を要素として採用し構造推定した（以降，sentArg）．具体的には，（１）を BERT を使った系列ラベリングし，論理要素を含む文を抽出する．（２）の要素分類は行わず，（３）の関係分類を（１）で推定したすべての文のペアに網羅的に行った．また，推定時に算出される予測確率をエッジの重みとして（４）の線形計画法に用いて構造を決定した．（３）の関係分類は，Jo ら [21] の LogBERT を用いた．Jo らは文間の関係性を４つの論理関係（事実の一貫性，感情の一貫性，因果関係，規範的関係）に分類し，確率的ソフトロジック（Probabilistic Soft Logic）を用いて定式化した．この結果を BERT を用いた分類モデルの学習に流用した LogBERT を提案し，従来手法よりも高精度な結果を示している．関係分類は，文１と文２の関係を support，attack，neutral のいずれかに分類する．support の場合は文１から文２に，attack の場合は文 2 から文 1 へエッジが引かれる．neutral はエッジは引かない．（４）の線形計画法は，循環がなく，自分へのエッジは引かない，複数のノードにエッジを引かない制約を課し構造推定する．

2.2 Graph Attention Network

本研究では，論理構造グラフを自動採点モデルに結合するため，グラフを特徴ベクトルに変換する必要がある．グラフ表現された構造データを深層学習で扱う手法として Graph Neural Network (GNN)[22] がある．GNN はノード間のエッジに基づき近傍ノードの集約やエッジに沿った畳み込み [23] によりグラフ情報から特徴量を抽出する．本研究では，ノードは論理要素となるテキストデータであり，ノードの集約や畳み込みを行う場合は BERT からの特徴ベクトルを用いる．ノードの情報をどのように取捨選択し集約するかが GNN の性能に影響すると考えられ，言語モデルに用いられる attention を GNN に導入した Graph Attention Networks(GAT)[24] が提案されている．GAT は，任意のノードとその近傍からの情報を集約する際，ノードの重要度を attention を用いて決定し，重要なノードの情報をより多く取り入れる手法である．さたに，Brody ら [25] は，ノードの特徴とその近傍の特徴を考慮しノード間の相互作用を効率的に学習する GATv2 を提案し，GAT よりも高精度な結果を示している．本研究では，Brody らの GATv2 を採用する．

3 提案モデル

BERT を用いた自動採点モデルと提案モデルについて説明する．j 番目の解答データを $x_j = \{w_{j0}, \ldots, w_{jn_j}\}$ と表す．解答データは改行を削除し 1 つの文にし，BERT の入力形式に合わせ文頭に [CLS]，文末に [SEP] を追加する．w_{ji} は x_j の i 番目のトークン，n_j はエッセイ j の総トークン数を示す．BERT を用いた自動採点は，BERT の最終層の出力の入力の [CLS] に対応する特徴ベクトル $\mathbf{h}_{[CLS]}$ を回帰もしくは分類モデルに入力し得点を予測する．多くの先行研究で回帰モデルが採用されているため本研究もそれに倣う．x_j の得点を s_j，線形変換を $Lin(\cdot)$，シグモイド関数を $\sigma(\cdot)$ として，BERT を用いた自動採点モデル（ベースライン）は式（2）のように表せる．式（1）の s_j と推定値 $\hat{s}_j$ との平均二乗誤差（Mean Squared Error）を最小化することでモデルを最適化する．

$$\text{argmin} \frac{1}{N} \sum_{i=1}^{N} (s_i - \hat{s}_i)^2 \tag{1}$$

$$s_j = \sigma\left(\text{Lin}\left(\mathbf{h}_{[CLS],j}\right)\right) \tag{2}$$

解答データの論理構造グラフとノードの特徴量を GATv2 に入力して得た特徴ベクトルを h_{GATv2} とする．本研究のグラフのノードはテキストデータのため，BERT の最終層の [CLS] ベクトルをノードの特徴量とした．提案モデル（BERT+GAT）は式（3）のように表現できる．

$$s_j = \sigma\left(\text{Lin}\left(\mathbf{h}_{[CLS],j} + \mathbf{h}_{GATv2,j}\right)\right) \tag{3}$$

関連研究で挙げた補助情報を追加する自動採点モデルの多くは，ベースとなる特徴量（本研究では BERT からの文脈ベクトル）に補助情報となる特徴ベクトルを式（3）のように直接結合している．異なるアーキテクチャからの特徴ベクトルを直接結合し 1 つのベクトルとすることが最適である裏付けはない．本研究では，BERT および GATv2 からの特徴ベクトルをそれぞれ線形変換後に結合し線形変換を行うモデル（BERT Linear3+GAT）も検証した（式（4））

$$s_j = \sigma\left(\text{Lin}\left(\text{Lin}\left(\mathbf{h}_{[CLS],j}\right) + \text{Lin}\left(\mathbf{h}_{GATv2,j}\right)\right)\right) \tag{4}$$

4 実験

実験に使うデータセットは Automated Student Assessment Prize (ASAP)，BERT は HggingFace の Transformers に公開されている bert-base-uncased[1)] を使用した．データ数，ジャンル，エッセイの平均文章長，得点値域を表 1 に示す．ASAP データセットは，指定されたテーマについて自分の意見を正当化する議論型，与えられた文書に関する問に回答する出典依存型，特定のテーマについて物語を語ってもらう物語型の 3 つのエッセイ タイプで構成されている．

表 1 Kaggle ASAP データセットの統計量

問題 ID	データ数	ジャンル	平均長	得点値域
1	1783	議論型	350	2-12
2	1800	議論型	350	1-6
3	1726	出典依存型	150	0-3
4	1772	出典依存型	150	0-3
5	1805	出典依存型	150	0-4
6	1800	出典依存型	150	0-4
7	1569	物語型	250	0-30
8	723	物語型	650	0-60

実験は，先行研究 [7] を参考に 5-fold 交差検証を用いた．学習は，最適化アルゴリズム Adam，バッチサイズは計算機環境の制約から 4 とし，earlystopping を採用した．dropout 率は BERT から線形変換への入力と線形変換から線形変換への入力に対し 0.5 と

1） https://huggingface.co/bert-base-uncased

表 2　ASAP データセットの QWK（5-fold 交差検証の平均値）

Model	問題 ID								平均
	1	2	3	4	5	6	7	8	
BERT	0.814	0.675	0.649	0.814	0.795	0.811	0.825	0.713	0.762
EASE+Argumentative Features[10]	0.832	0.689	–	–	–	–	–	–	–
BERT+Logical Structures[14]	0.815	0.672	0.693	0.816	0.809	**0.814**	**0.829**	0.717	0.771
BERT+Semantic Features[13]	**0.846**	**0.698**	0.684	0.675	0.795	0.704	0.711	0.668	0.723
GMFRM[26]	0.837	0.679	0.670	0.738	0.797	0.785	0.789	0.710	0.756
Severity-fixed GMFRM[26]	0.831	0.667	0.665	0.738	0.797	0.773	0.788	0.710	0.746
BERT+GAT(sentArg)	0.803	0.682	0.645	0.818	**0.810**	0.810	0.816	**0.729**	0.764
BERT+GAT(mt-am[19])	0.802	0.649	0.643	0.816	0.792	0.804	0.825	0.708	0.755
BERT+GAT (original graph)	0.811	0.673	0.668	0.810	0.790	**0.814**	0.818	0.703	0.761
BERT+GAT (random graph)	0.795	0.670	0.634	0.814	0.797	0.813	0.812	0.717	0.756
BERT linear3+GAT(sentArg)	0.827	0.667	0.697	0.815	0.807	0.807	0.827	0.726	**0.772**
BERT linear3+GAT(mt-am[19])	0.807	0.681	0.683	0.816	0.806	0.809	0.821	0.703	0.766
BERT linear3+GAT (original graph)	0.807	0.686	**0.698**	**0.823**	0.796	0.805	0.826	0.696	0.767
BERT linear3+GAT (random graph)	0.820	0.656	0.685	0.810	0.804	0.808	0.827	0.711	0.765

した．開発データの loss を最小化し，テストデータの評価指標 QWK を算出した．

実験は，提案手法（BERT+GAT と BERT linear3+GAT）を論理構造推定手法（sentArg と mt-am）を変え検証した．sentArg は，線形計画法によりエッジの削除やループにならないように制限を課しているが，それが正しいか不明なため線形計画法に入力前のグラフ（original graph）を用いた場合も検証した．

本研究で使用している ASAP データセットは，論理構造の要素や要素間の関係のアノテーションがないため，推定結果の精度を検証できない．sentArg は，従来手法と異なり論理構造の要素を文としているため先行研究データを用いて精度検証することも難しい．そこで，論理構造の要素および構造をランダムにしたグラフ（random graph）を入力することで提案手法の有効性を検証した．

5　実験結果

実験結果を表 2 に示す．表中の数値は交差検証で算出した問題 ID ごとの QWK の平均値，平均はモデルごとの平均値を示す．BERT linear3+GAT(sentArg) の平均が最も高い結果となった．mt-am よりも oroginal graph，さらに sentArg の方が高い結果となった．また，ランダムな構造（random graph）を追加した手法よりも oroginal graph，sentArg の方が高精度だった．original graph よりも sentArg が高精度になことから，線形計画法で構造決定した方が良いと考えられる．BERT+GAT よりも BERT linear3+GAT の方が高精度なため，言語モデルの特徴量に別の特徴量を追加する場合は直接結合せず間に線形変換を挟んだほうが良いと考えられる．

BERT+GAT は，ベースラインよりも問題 8 つの平均は高い結果となったが，問題 1，3，6，7 はベースラインを下回った．BERT linear3+GAT においてもベースラインよりも平均は高いが，問題 6 が下回った．

本研究と類似したモデルである Yamaura ら [14] の結果と比較すると，問題 1，3，8 と平均値は BERT linear3+GAT, 問題 2，4，5，8 は BERT+GAT が上回る結果となった．

6　おわりに

本研究では，論理構造グラフを考慮した自動採点モデルを提案しその有効性を示した．しかし，論理構造を正しく推定できたか，それが精度にどう影響するかなど十分な分析はできていないため今後の研究課題としたい．

参考文献

[1] Zixuan Ke and Vincent Ng. Automated essay scoring: A survey of the state of the art. In **IJCAI**, Vol. 19, pp. 6300–6308, 2019.

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

[3] Dadi Ramesh and Suresh Kumar Sanampudi. An automated essay scoring systems: a systematic literature review. **Artificial Intelligence Review**, Vol. 55, No. 3, pp. 2495–2527, 2022.

[4] Yigal Attali and Jill Burstein. Automated essay scoring with e-rater® v. 2. **The Journal of Technology, Learning and Assessment**, Vol. 4, No. 3, 2006.

[5] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. **Advances in neural information processing systems**, Vol. 30, , 2017.

[6] 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.

[7] 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.

[8] Fei Dong and Yue Zhang. Automatic features for essay scoring – an empirical study. In **Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing**, pp. 1072–1077, Austin, Texas, November 2016. Association for Computational Linguistics.

[9] Fei Dong, Yue Zhang, and Jie Yang. Attention-based recurrent convolutional neural network for automatic essay scoring. In **Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)**, pp. 153–162, Vancouver, Canada, August 2017. Association for Computational Linguistics.

[10] Huy Nguyen and Diane Litman. Argument mining for improving the automated scoring of persuasive essays. **Proceedings of the AAAI Conference on Artificial Intelligence**, Vol. 32, No. 1, Apr. 2018.

[11] 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.

[12] Sayaka Nakamoto and Kazutaka Shimada. Automated scoring of logical consistency of japanese essays. In **International Conference on Artificial Intelligence in Education**, pp. 652–658. Springer, 2023.

[13] Jianwei Li and Jiahui Wu. Automated essay scoring incorporating multi-level semantic features. In **International Conference on Artificial Intelligence in Education**, pp. 206–211. Springer, 2023.

[14] Misato Yamaura, Itsuki Fukuda, and Masaki Uto. Neural automated essay scoring considering logical structure. In **International Conference on Artificial Intelligence in Education**, pp. 267–278. Springer, 2023.

[15] Christian Stab and Iryna Gurevych. Identifying argumentative discourse structures in persuasive essays. In **Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)**, pp. 46–56, 2014.

[16] Christian Stab and Iryna Gurevych. Parsing argumentation structures in persuasive essays. **Computational Linguistics**, Vol. 43, No. 3, pp. 619–659, 2017.

[17] Marco Lippi and Paolo Torroni. Argumentation mining: State of the art and emerging trends. **ACM Transactions on Internet Technology (TOIT)**, Vol. 16, No. 2, pp. 1–25, 2016.

[18] Makoto Miwa and Mohit Bansal. End-to-end relation extraction using LSTMs on sequences and tree structures. In Katrin Erk and Noah A. Smith, editors, **Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 1105–1116, Berlin, Germany, August 2016. Association for Computational Linguistics.

[19] Gaku Morio, Hiroaki Ozaki, Terufumi Morishita, and Kohsuke Yanai. End-to-end argument mining with cross-corpora multi-task learning. **Transactions of the Association for Computational Linguistics**, Vol. 10, pp. 639–658, 2022.

[20] Jianzhu Bao, Yuhang He, Yang Sun, Bin Liang, Jiachen Du, Bing Qin, Min Yang, and Ruifeng Xu. A generative model for end-to-end argument mining with reconstructed positional encoding and constrained pointer mechanism. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, **Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing**, pp. 10437–10449, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[21] Yohan Jo, Seojin Bang, Chris Reed, and Eduard Hovy. Classifying argumentative relations using logical mechanisms and argumentation schemes. **Transactions of the Association for Computational Linguistics**, Vol. 9, pp. 721–739, 2021.

[22] Jie Zhou, Ganqu Cui, Shengding Hu, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. Graph neural networks: A review of methods and applications. **AI open**, Vol. 1, pp. 57–81, 2020.

[23] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. **arXiv preprint arXiv:1609.02907**, 2016.

[24] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. **arXiv preprint arXiv:1710.10903**, 2017.

[25] Shaked Brody, Uri Alon, and Eran Yahav. How attentive are graph attention networks? **arXiv preprint arXiv:2105.14491**, 2021.

[26] Masaki Uto, Itsuki Aomi, Emiko Tsutsumi, and Maomi Ueno. Integration of prediction scores from various automated essay scoring models using item response theory. **IEEE Transactions on Learning Technologies**, pp. 1–18, 2023.