論述構造解析技術を用いたニューラル小論文自動採点手法

山浦美里 [1]　福田樹 [1]　宇都雅輝 [1]
[1] 電気通信大学大学院
{yamaura,fukuda,uto}@ai.lab.uec.ac.jp

概要

近年，深層学習を用いた小論文自動採点モデルが高精度を達成しつつあるが，従来の深層学習自動採点モデルは文章の論理構造を明示的には考慮できない．本研究では，論述構造解析技術を用いて推定される文章の論理構造を考慮できる新たな深層学習自動採点モデルを提案する．

1　はじめに

論理的思考力や表現力が新しい時代に求められる資質として注目される中，そのような能力の評価法の一つとして小論文試験が広く活用されている [1, 2, 3]．しかし，小論文試験には，採点の公平性担保の困難さや人手採点に伴うコストの増大などの懸念がある [4]．これらの問題を解決する手段の一つとして自動採点技術が近年注目されている [5, 6]．

既存の小論文自動採点手法は大きく 2 つに分類できる [6, 7]．一つは人手で設計した特徴量を用いる手法であり，文章から抽出した特徴量（総単語数や接続詞数，文法エラー率など）を線形回帰モデルや決定木などの機械学習モデルに入力して得点を予測する [5, 8]．二つ目は深層学習を用いる手法であり，入力が小論文の単語系列，出力が得点となる深層学習モデルを構築して自動採点を実現する [6, 9]．近年では，BERT（Bidirectional Encoder Representations from Transformers）[10] などを用いた深層学習自動採点モデルが高精度を達成している [11, 12]．

従来の深層学習自動採点モデルでは，単語の意味や単語間の関係性は考慮できるが，文章の論理構造は直接的には考慮できない [13, 14]．論理構造は小論文の質に関わる本質的な要素の一つであるため，論理構造をモデルに明示的に与えることができれば，更なる自動採点の精度向上が期待できる．このような背景から，Nguyen & Litman[15] は，論理構造を考慮できる小論文自動採点手法を提案している．この手法では，自然言語処理分野において近年高精度化が進む論述構造解析技術（Argument Mining）[16, 17, 18, 19, 20, 21] を用いて文章の論理構造を推定し，得られた論理構造から特徴量（論理構造の構成要素数やそれらの要素間のエッジの数，各要素中の単語数など）を抽出して特徴量ベースの自動採点手法を構築している．しかし，実験の結果，論理構造に関する特徴量の追加による自動採点精度の改善は限定的であったことを報告している．この要因としては，論理構造の情報を表層的な特徴量に縮約してしまったため，論理構造の情報を十分に活用できなかったことが考えられる．

そこで本研究では，論述構造解析で求めた論理構造を人手で設計した特徴量に変換せずに処理できる深層学習手法を開発し，その手法を組み込んだ新たな深層学習自動採点手法を提案する．具体的には，BERT モデルの Self-Attention 機構を拡張することで，論理構造を考慮できる深層学習モデルを構築し，その処理結果を従来の深層学習自動採点モデルに統合する方法を開発する．さらに，ベンチマークデータを用いた実験を通して提案手法の有効性を示す．

2　提案手法

2.1　論述構造解析による論理構造推定

提案手法では，Nguyen & Litman [15] と同様に，論理構造の推定に論述構造解析技術を用いる．論述構造解析では，図 1 のように，まず文章中から論理構造のノードに対応する文や文節（論理要素と呼ぶ）

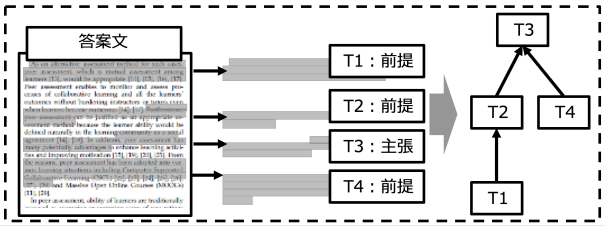

図 1　論述構造解析の概念図

を抽出し，それらの論理要素間の論理関係を木構造制約を満たすように推定することで，文章の論理構造を推定する [18]．本研究では，近年高精度を達成している深層学習ベースの論述構造解析手法の一つである Eger et al.[19] の手法を用いる．

2.2 論理構造を処理する深層学習モデル

提案手法では，論理構造を処理する深層学習モデルの基礎モデルとして BERT を用いる．BERT は，Self-Attention 機構をコアとする Transformer と呼ばれる構造を 12 層重ねた深層学習モデルである．本研究では，BERT への入力として，採点対象文の先頭に [CLS] という特殊タグを挿入した単語系列 $\{w_0, w_1, \ldots, w_T\}$（ただし，$w_0$ は [CLS] タグ，w_t $(t > 1)$ は対象文の t 番目の単語，T は対象文の単語数を表す）を考え，[CLS] タグに対応する出力ベクトルを入力文に対する分散表現ベクトルとみなす．このとき，BERT の l 層目の Self-Attention は次式で求められる．

$$\mathbf{H}_l = \mathbf{A}_l \cdot \mathbf{V}_l \tag{1}$$

ここで，$\mathbf{H}_l$ は l 層目の Self-Attention 層の出力であり，$(T+1) \times D$ の行列（ただし，D は BERT の潜在変数ベクトルの次元数を表す）である．なお，$\mathbf{H}_l$ の行数が対象文中の単語数 T ではなく $(T+1)$ となっているのは，上記の通り，BERT への入力の先頭に特殊タグ [CLS] を付与しているためである．また，$\mathbf{V}_l$ は一つ前の層の出力 $\mathbf{H}_{l-1}$ に基づいて計算される $(T+1) \times D$ の行列である．さらに，$\mathbf{A}_l$ は $(T+1) \times (T+1)$ のアテンション重みを表す行列であり，通常の BERT では，次式で求められる．

$$\mathbf{A}_l = softmax(\frac{\mathbf{Q}_l \cdot \mathbf{K}_l^\top}{\sqrt{d}}) \tag{2}$$

ここで，$\mathbf{Q}_l$ と $\mathbf{K}_l$ は一つ前の層の出力 $\mathbf{H}_{l-1}$ に基づいて計算される $(T+1) \times D$ の行列であり，d は調整係数である．

アテンション重み行列 $\mathbf{A}_l$ は，文章中のある単語の分散表現ベクトルを計算するために，その他の単語の情報をどれだけ参照するかをコントロールする機能を持つ．例えば，$\mathbf{A}_l$ の t 行 t' 列目の要素 $a_{ltt'}$ は単語 w_t の分散表現ベクトルを計算する際に，$w_{t'}$ の情報をどれだけ重み付けして加算するかを表現する．本研究では，このアテンション重みを調整することで，論理構造を明示的に反映することを目指す．具体的には，論理関係のある論理要素間では単語間の情報参照を許可し，論理関係がない論理要素間では単語間での情報参照を行わせないようにするために，Visible Matrix と呼ぶ $(T+1) \times (T+1)$ の行列 $\mathbf{M} = \{m_{tt'} \mid t, t' \in \{0, \ldots, T\}\}$ を導入し，アテンション重み行列 $\mathbf{A}_l$ を次のように計算する．

$$\mathbf{A}_l = softmax(\frac{\mathbf{Q}_l \cdot \mathbf{K}_l^\top}{\sqrt{d}} + \mathbf{M}) \tag{3}$$

この式では，Visible Matrix $\mathbf{M}$ の t 行 t' 列目の要素 $m_{tt'}$ が $-\infty$ になると，Softmax 関数により，アテンション行列の t 行 t' 列目の要素 $a_{ltt'}$ が 0 となるため，式（1）の Self-Attention の計算において t 番目の単語の分散表現ベクトルの計算時に t' 番目の単語の情報を無視させることができる．そこで，本研究では，論理関係がない単語間について $m_{tt'} = -\infty$，論理関係がある単語間について $m_{tt'} = 0$ になるように，次のように Visible matrix を構築する．

今，採点対象文に論述構造解析を適用した結果，P 個の論理要素が得られたとする．ここで，p 番目の論理要素を $\mathbf{C}_p$ （ただし，$p = \{1, \ldots, P\}$）とし，$\mathbf{C}_p$ はその論理要素に含まれる連続する単語の集合として $\{w_i, w_{i+1}, \ldots, w_{I_p}\}$ で表すとする．ただし，i は対象文中における論理要素 $\mathbf{C}_p$ の開始位置，I_p は $\mathbf{C}_p$ に含まれる単語数とする．また，論理要素間の関係は $P \times P$ の対称行列 $\mathbf{R} = \{r_{uv} \mid u, v \in \{1, \ldots, P\}\}$ で表し，u 番目の論理要素 $\mathbf{C}_u$ と v 番目の論理要素 $\mathbf{C}_v$ に論理関係があるとき $r_{uv} = r_{vu} = 1$，そうでないときに $r_{uv} = r_{vu} = 0$ とする．以上の定義のもとで，本研究では，以下の 2 条件を満たすときに $m_{tt'} = 0$，そうでないときに $m_{tt'} = -\infty$ となるように，Visible matrix $\mathbf{M}$ を作成する．

1. w_t を要素に持つ論理要素 $\mathbf{C}_p$ と $w_{t'}$ を要素に持つ論理要素 $\mathbf{C}_{p'}$ が存在する．
2. $\mathbf{C}_p$ と $\mathbf{C}_{p'}$ に論理関係が存在する．すなわち $r_{pp'} = 1$ である．

なお，個々の論理要素について，論理要素内の単語同士では情報の参照ができるように，$r_{pp} = 1; \forall p$ とする．また，本研究では，[CLS] に対応する出力を論理構造を考慮した全体の分散表現ベクトルとすることを想定しているため，Visible Matrix $\mathbf{M}$ の 0 行目の要素は 0，すなわち $m_{0t} = 0; \forall t$ とし，全体の情報が [CLS] に集約されるようにする．反対に，[CLS] タグに集約される情報を各単語が参照できてはいけないため，$\mathbf{M}$ の 0 列目の要素は 0 行目を除いて全て $-\infty$ とする．本研究では，以上のように作

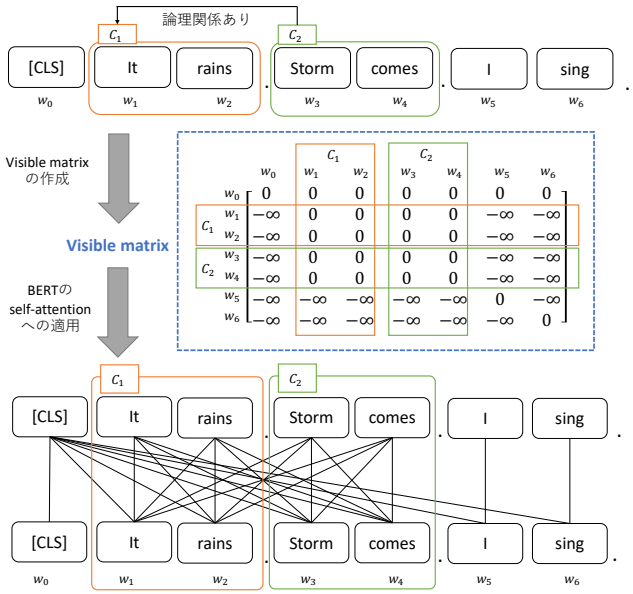

図 2　Visible Matrix 適用の概念図

成した Visible Matrix を用いて，論理構造の情報を反映できるように Self-Attention を拡張した BERT を「BERT-LS」と呼ぶ．

図 2 に Visible Matrix の作成と BERT-LS における Self-Attention の概念図を示す．図 2 の下部で図示しているように，BERT-LS の Self-Attention では，各単語の分散表現ベクトルを計算する際に，その単語と論理的に関係がある論理要素の情報のみが考慮され，論理関係がない論理要素間では単語間の情報参照がおこらないようになっている．そのため，全単語間の関係を考慮する通常の BERT と比べて，BERT-LS では論理構造を強調した情報処理が実現できると考えられる．また，図からもわかるように，BERT-LS では，[CLS] タグに対応する分散表現ベクトルに全体の情報が縮約されている．よって，本研究では，BERT-LS の最終層の [CLS] タグに対応する出力ベクトル $\boldsymbol{x}'_0$ を論理構造を考慮した分散表現ベクトルとして採用する．

2.3　論理構造を考慮した深層学習自動採点手法

ここでは，BERT-LS を用いて処理した論理構造情報を，従来の深層学習自動採点モデルに加味させる手法を提案する．従来の深層学習自動採点モデルとしては，様々なモデルが利用できるが，ここでは，近年ベースラインとして広く利用される BERT を用いた深層学習自動採点モデルを基礎モデルとして利用する．

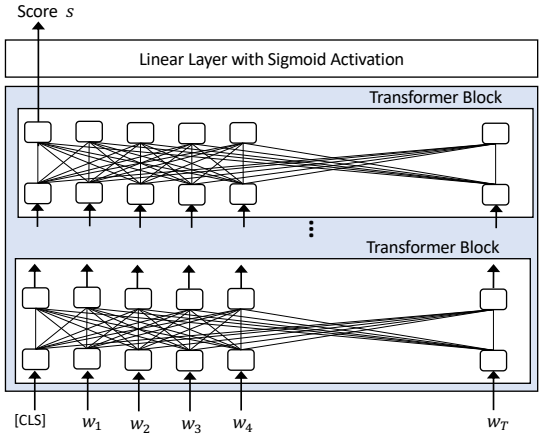

図 3　BERT による自動採点モデルの概念図

BERT を用いた自動採点モデルの概念図を図 3 に示す．モデルへの入力は，BERT-LS と同様に，小論文の先頭に [CLS] タグを挿入した単語系列 $\{w_0, w_1, \ldots, w_T\}$ である．BERT を用いた自動採点モデルでは，[CLS] タグに対応する BERT の出力ベクトル $\boldsymbol{x}_0$ に対して，次式で与えられる Linear Layer with Sigmoid Activation を適用することで，予測得点 s を求める．

$$s = \sigma(\boldsymbol{W}\boldsymbol{x}_0 + b) \tag{4}$$

ここで，σ は Sigmoid 関数を表し，$\boldsymbol{W}$ と b は重みベクトルとバイアスを表すパラメータである．なお，s は 0 から 1 の値を取るため，得点尺度がこれと異なる場合には，s を一次変換し，実際の得点尺度に合わせる．例えば，1〜K の K 段階得点の場合，$Ks + 1$ と変換する．

提案手法では，図 4 に示すように，BERT-LS で得られる分散表現ベクトル $\boldsymbol{x}'_0$（[CLS] に対応する最終

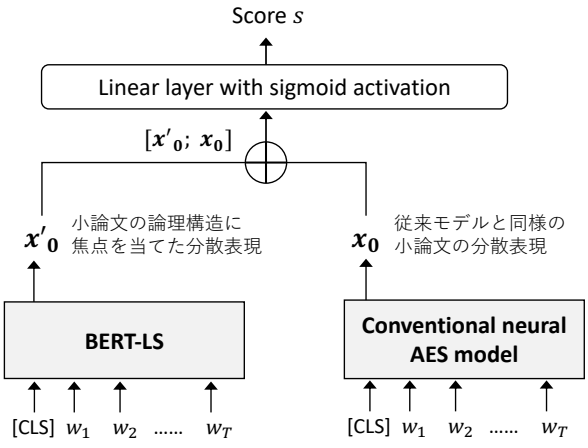

図 4　提案モデルの概念図

表 1　実験結果

ベースモデル	手法	課題 1	課題 2	課題 3	課題 4	課題 5	課題 6	課題 7	課題 8	平均	p 値
BERT	提案手法	**0.815**	**0.672**	**0.693**	**0.816**	**0.809**	**0.814**	0.829	**0.717**	**0.771**	0.009
	従来手法	0.795	0.669	0.671	0.794	0.803	0.806	**0.832**	0.713	0.760	
RoBERTa	提案手法	**0.823**	**0.682**	**0.679**	**0.823**	**0.824**	0.820	**0.837**	**0.736**	**0.775**	0.001
	従来手法	0.812	0.673	0.674	0.814	0.812	**0.821**	0.825	0.717	0.769	
ALBERT	提案手法	**0.817**	**0.679**	0.676	**0.812**	**0.804**	0.807	**0.835**	**0.723**	**0.769**	0.067
	従来手法	0.792	0.665	**0.676**	0.808	0.800	**0.811**	0.834	0.722	0.763	
DistilBERT	提案手法	**0.817**	**0.679**	**0.691**	**0.808**	**0.802**	0.807	0.827	**0.735**	**0.771**	0.048
	従来手法	0.798	0.674	0.653	0.803	0.801	**0.810**	**0.829**	0.726	0.762	
DeBERTa	提案手法	**0.820**	0.670	0.666	0.812	**0.810**	0.806	0.820	**0.719**	0.765	0.373
	従来手法	0.806	**0.671**	**0.680**	**0.817**	0.804	**0.817**	**0.828**	0.710	**0.766**	
LSTM	提案手法	**0.817**	**0.687**	**0.686**	**0.804**	**0.806**	**0.801**	**0.827**	**0.740**	**0.771**	0.007
	従来手法	0.804	0.637	0.656	0.772	0.796	0.800	0.739	0.654	0.732	
BERT-LS のみ		0.821	0.677	0.668	0.807	0.814	0.803	0.821	0.722	0.767	

層の出力ベクトル）を，従来の深層学習自動採点モデルで得られる分散表現ベクトル $\boldsymbol{x}_0$ と結合したベクトル $[\boldsymbol{x}_0; \boldsymbol{x}'_0]$ を作成し，それを Linear Layer with Sigmoid Activation に入力して予測得点を算出する．

モデル学習は，次式の平均二乗誤差（mean squared error：MSE）を損失関数として誤差逆伝搬法で行う．

$$\frac{1}{N}\sum_{n=1}^{N}(s_n - s_n^*)^2 \tag{5}$$

ここで，s_n は小論文 n の予測得点，s_n^* は真の得点を表し，N は訓練データ中の小論文の数を表す．なお，出力層に Sigmoid 関数を採用しているため，真の得点は 0 から 1 の値に線形変換する必要がある．

3　実験

提案手法の有効性を確認するために，ベンチマークデータを用いた評価実験を行う．実験には，自動採点研究のベンチマークデータとして広く利用されている Automated Student Assessment Prize（ASAP）を用いた．ASAP は，8 つの異なる小論文課題に対して，英語を母語とする米国の学生が英語で解答した小論文と，それに対する得点で構成されるデータセットである．予測精度の評価実験は，8 つの課題別に 5 分割交差検証法で行い，精度評価指標には 2 次重み付きカッパ係数（QWK; quadratic weighted kappa）を用いた．

ベースとする深層学習自動採点モデル（図 4 の右側のモデル）には，BERT，RoBERTa[22]，ALBERT[23]，DistilBERT[24]，DeBERTa[25]，LSTM（Long short term memory）を中心としたモデル [26] を用い，それぞれのモデルに対して，BERT-LS を統合した提案手法と BERT-LS を使用しない従来手法について予測精度を求めた．さらに，BERT-LS 単体についても同様の実験を行なった．

実験結果を表 1 に示す．太字は提案手法と従来手法で精度の高い方を示している．また，p 値列は，同一のベースモデルにおける提案手法と従来手法の平均精度について，対応のある t 検定を行なった結果を示している．表 1 より，概ね全てのベースモデルで提案手法の平均精度が従来手法より高いこと，また，BERT-LS を単体で使用した場合と比べても，提案手法の精度が高いことが確認できる．

従来手法と比べた提案手法の精度改善は，課題 1，2，8 において大きい傾向が確認できる．ASAP データセットには，自身の意見を論証するタイプの課題（課題 1，2，7，8）と，与えられた長文に対してやや短めの文章で回答する形式の課題（課題 3，4，5，6）が含まれており，精度改善が大きかった課題 1，2，8 は論証タイプの課題である．このことから，提案手法は，論理的な文章構成が必要とされるタイプの小論文課題において，有効性が高い傾向があると考えられる．

4　まとめ

本研究では，論述構造解析を用いて小論文の論理構造を推定し，その情報を深層学習で解析する手法を開発するとともに，その手法を組み込んだ新たな深層学習自動採点モデルを提案した．実データ実験により，提案モデルによって小論文の論理構造を明示的に深層学習モデルに組み込むことが，自動採点の精度改善に有効であることが示された．論理構造を活用した自動採点の先行研究 [15] では，論理構造に基づく特徴量組み込みの有効性は示されなかったが，本研究で提案した方法で組み込みを行えば精度が改善されることが示された．

参考文献

[1] Yousef Abosalem. Assessment techniques and students' higher-order thinking skills. *International Journal of Secondary Education*, Vol. 4, pp. 1–11, 01 2016.

[2] Emily R Lai. Critical thinking: A literature review. *Pearson's Research Reports*, Vol. 6, No. 1, pp. 40–41, 2011.

[3] Ou Lydia Liu, Lois Frankel, and Katrina Crotts Roohr. Assessing critical thinking in higher education: Current state and directions for next-generation assessment. *ETS Research Report Series*, Vol. 2014, No. 1, pp. 1–23, 2014.

[4] Masaki Uto and Masashi Okano. Learning Automated Essay Scoring Models Using Item-Response-Theory-Based Scores to Decrease Effects of Rater Biases. *IEEE Transactions on Learning Technologies*, Vol. 14, No. 6, pp. 763–776, 2021.

[5] Zixuan Ke and Vincent Ng. Automated Essay Scoring: A Survey of the State of the Art. In *Proc. International Joint Conferences on Artificial Intelligence Organization*, Vol. 19, pp. 6300–6308, 2019.

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

[7] Takumi Shibata and Masaki Uto. Analytic Automated Essay Scoring Based on Deep Neural Networks Integrating Multidimensional Item Response Theory. In *Proc. International Conference on Computational Linguistics*, pp. 2917–2926, Gyeongju, Republic of Korea, October 2022.

[8] Peter Phandi, Kian Ming A Chai, and Hwee Tou Ng. Flexible Domain Adaptation for Automated Essay Scoring Using Correlated Linear Regression. In *Proc. Conference on Empirical Methods in Natural Language Processing*, pp. 431–439, 2015.

[9] Farah Nadeem, Huy Nguyen, Yang Liu, and Mari Ostendorf. Automated Essay Scoring with Discourse-Aware Neural Models. In *Proc. Workshop on Innovative Use of NLP for Building Educational Applications*, pp. 484–493, 2019.

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. 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, June 2019.

[11] Christian Stab, Tristan Miller, Benjamin Schiller, Pranav Rai, and Iryna Gurevych. Cross-topic argument mining from heterogeneous sources. In *Proc. Conference on Empirical Methods in Natural Language Processing*, pp. 3664–3674, 2018.

[12] Masaki Uto, Yikuan Xie, and Maomi Ueno. Neural Automated Essay Scoring Incorporating Handcrafted Features. In *Proc. International Conference on Computational Linguistics*, pp. 6077–6088, 2020.

[13] Harneet Kaur Janda, Atish Pawar, Shan Du, and Vijay Mago. Syntactic, semantic and sentiment analysis: The joint effect on automated essay evaluation. *IEEE Access*, Vol. 7, pp. 108486–108503, 2019.

[14] Patrick Hohenecker and Thomas Lukasiewicz. Ontology reasoning with deep neural networks. *Journal of Artificial Intelligence Research*, Vol. 68, pp. 503–540, 2020.

[15] Huy Nguyen and Diane Litman. Argument Mining for Improving the Automated Scoring of Persuasive Essays. *Proc. Association for the Advancement of Artificial Intelligence Conference on Artificial Intelligence*, Vol. 32, No. 1, 2018.

[16] Raquel Mochales Palau and Marie-Francine Moens. Argumentation Mining: The Detection, Classification and Structure of Arguments in Text. In *Proc. International Conference on Artificial Intelligence and Law*, pp. 98–107, 2009.

[17] Ruty Rinott, Lena Dankin, Carlos Alzate Perez, Mitesh M. Khapra, Ehud Aharoni, and Noam Slonim. Show Me Your Evidence - an Automatic Method for Context Dependent Evidence Detection. In *Proc. Conference on Empirical Methods in Natural Language Processing*, pp. 440–450, 2015.

[18] Christian Stab and Iryna Gurevych. Parsing Argumentation Structures in Persuasive Essays. *Computational Linguistics*, Vol. 43, No. 3, pp. 619–659, 2017.

[19] Steffen Eger, Johannes Daxenberger, and Iryna Gurevych. Neural End-to-End Learning for Computational Argumentation Mining. In *Proc. Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11–22, 2017.

[20] John Lawrence and Chris Reed. Argument Mining: A Survey. *Computational Linguistics*, Vol. 45, No. 4, pp. 765–818, 2020.

[21] Yuxiao Ye and Simone Teufel. End-to-End Argument Mining as Biaffine Dependency Parsing. In *Proc. Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 669–678, 2021.

[22] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[23] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. ALBERT: A lite BERT for self-supervised learning of language representations. *International Conference on Learning Representations*, 2020.

[24] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

[25] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*, 2020.

[26] Kaveh Taghipour and Hwee Tou Ng. A Neural Approach to Automated Essay Scoring. In *Proc. Empirical Methods in Natural Language Processing*, pp. 1882–1891, 2016.