多次元項目反応理論と深層学習に基づく複数観点同時自動採点手法の精度改善

柴田拓海 [1]　宇都雅輝 [1]
[1] 電気通信大学大学院
{shibata,uto}@ai.lab.uec.ac.jp

概要

深層学習を用いた小論文自動採点手法の一つとして，全体得点と複数の評価観点別得点を同時に予測する手法が近年注目され，高精度を達成しつつある．しかし，従来手法は得点予測の根拠について解釈性が低いという問題があった．Shibata & Uto はこの問題を解決するために多次元項目反応理論を組み込んで解釈性を高めた複数観点同時自動採点手法を提案している．しかし，この手法は従来手法と比較して僅かながら予測精度の低下がみられる．そこで本研究では，精度低下の原因を分析し，予測精度を改善した手法を提案する．

1　はじめに

近年，小論文試験の採点をコンピュータを用いて自動化する小論文自動採点（Automated Essay Scoring；AES）手法が多数提案されている．自動採点を実現する手法は特徴量ベースと深層学習ベースの手法に大別される．これまでは，特徴量ベースの手法が一般的であったが（e.g., [1–3]），近年では深層学習を用いた自動採点モデルが多数提案されている（e.g., [4–16]）．深層学習自動採点モデルは文章の単語系列を入力として，データから自動で複雑な特徴量を学習でき，高精度を達成している．

従来の自動採点モデルの多くは，全体得点のみを予測する採点場面を想定している（e.g., [5–12]）．しかし，学習評価場面などで小論文試験を運用する場合，より詳細なフィードバックを受検者に行うために，論理構成力や文章表現力などの評価観点別の得点付けを行いたい場面も少なくない [13]．そこで，複数の評価観点に対応する得点を同時に予測できるモデルが近年提案されている（e.g., [13–16]）．

現時点では，Ridley ら [16] の複数観点自動採点モデルが最高精度を達成しているが，このモデルには解釈性の観点から次のような問題がある．(1) 評価観点ごとに複雑な多層ニューラルネットワークを持つため予測根拠を解釈することが難しい．(2) 一般に評価観点は，背後に測定したい能力尺度を想定し，それを測定できるように設計されるが [17]，このモデルでは複数評価観点の背後に想定される能力尺度を解釈することができない．

これらの問題を解決するために，Shibata & Uto [18, 19] は，項目反応理論を利用した手法を提案している．具体的には，評価観点の特性を考慮した多次元項目反応モデル [20] を出力層とし，それ以外を Ridley らのモデルを元にした評価観点共通のニューラルネットワークとしたモデルを開発している．この手法の利点は以下の通りである．(1) 評価観点固有の出力層は，識別力と困難度と呼ばれる項目反応理論で一般的な 2 種類のパラメータのみで説明されるため，それらのパラメータ値に基づいて観点ごとの特性を定量的に解釈できる．(2) 多次元項目反応モデル層の能力次元数を最適化してパラメータを分析することで，複数評価観点の背後に想定される能力尺度を解釈できる．

このモデルは上記の点において解釈性を向上させることに成功したが，僅かに Ridley らの手法と比較してモデルの予測精度が低下することがわかっている．この原因を調査するために評価観点別の精度評価実験を行ったところ，特に全体得点の予測精度が低いことがわかった．そこで本研究では Shibata & Uto の手法において，全体得点の予測精度を改善する手法を提案する．

2　項目反応理論

ここでは，本研究で扱う自動採点手法で利用している項目反応理論について説明する．項目反応理論（Item Response Theory；IRT）[21] は，近年のコンピュータ・テスティングの発展に伴い，様々な分野

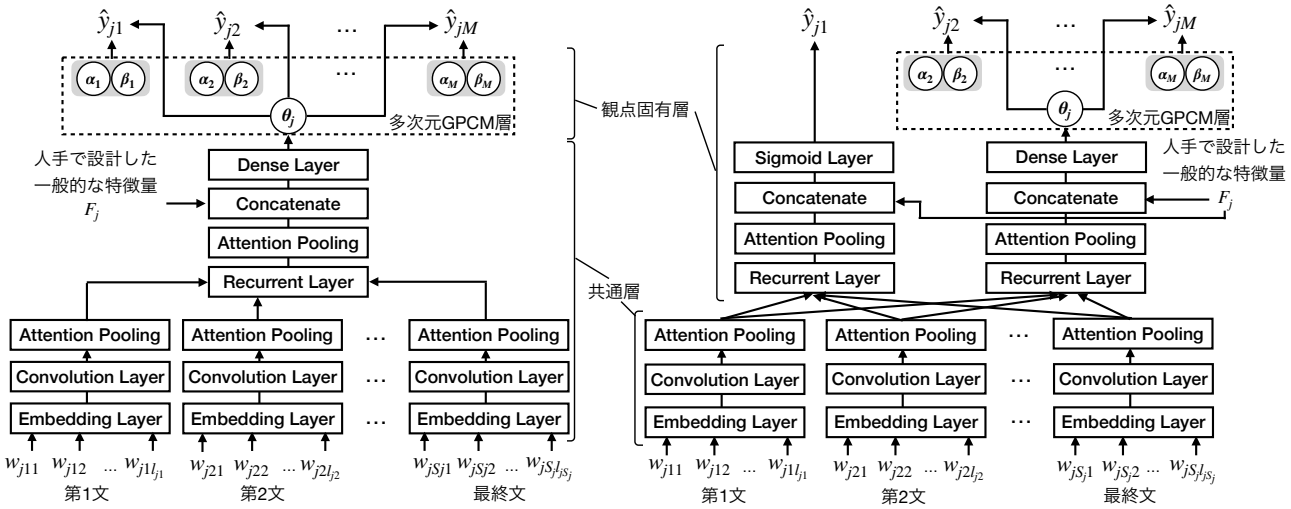

図 1　Shibata & Uto の複数観点同時自動採点モデル（左）と提案モデル（右）の概念図

で実用化が進められている数理モデルを用いたテスト理論の一つである．IRT モデルは正誤データなどの 2 値型のデータを前提とするものが多いが，多段階の得点データに対応した IRT モデルも多数提案されている．また一般的な IRT モデルでは，測定対象の能力に 1 次元性を仮定しているが，測定される能力に多次元性を仮定できるモデルも提案されている．

本研究では，代表的な多次元多値型 IRT モデルである多次元一般化部分採点モデル（Generalized Partial Credit Model；GPCM）[22] を使用する．ここでは，先行研究 [17,20] のように各評価観点を項目とみなして多次元 GPCM を適用する．具体的には，受検者 j が評価観点 m において，得点 $k \in \{1,2,\ldots,K_m\}$ を得る確率を次式で与えるモデルを適用する．

$$P_{jmk} = \frac{\exp(k\boldsymbol{\alpha}_m^T\boldsymbol{\theta}_j + \sum_{u=1}^{k}\beta_{mu})}{\sum_{v=1}^{K_m}\exp(v\boldsymbol{\alpha}_m^T\boldsymbol{\theta}_j + \sum_{u=1}^{v}\beta_{mu})} \tag{1}$$

ここで，$\boldsymbol{\theta}_j = (\theta_{j1},\theta_{j2},\ldots,\theta_{jd})$ は受検者 j の d 次元の能力を表すパラメータベクトルであり，ベクトルの各要素は各次元の能力値を表す．$\boldsymbol{\alpha}_m = (\alpha_{m1},\alpha_{m2},\ldots,\alpha_{md})$ は $\boldsymbol{\theta}_j$ に対応した評価観点 m の d 次元識別力，β_{mu} は評価観点 m においてカテゴリ $u-1$ から u に遷移する困難度を表すパラメータである．K_m は，評価観点 m における得点段階数を表す．なお，モデルの識別性のために，$\beta_{m1} = 0 : \forall m$ を所与とする．

3　Shibata & Uto の自動採点モデル

ここでは上述した多次元 GPCM を組み込んだ Shibata & Uto [18,19] の複数観点同時自動採点モデルについて説明する．モデルの概念図を図 1（左）に示した．このモデルは，受検者 $j \in \mathcal{J} = \{1,2,\ldots,J\}$ の小論文を入力とし，評価観点 $m \in \mathcal{M} = \{1,2,\ldots,M\}$ に対応する得点 $\hat{y}_{jm}$ を出力する．ここで J は受検者数，M は評価観点数を表し，特に $m = 1$ は全体得点を表すものとする．また，受検者 j の小論文は単語系列として，$\{w_{jsl}|s \in \{1,2,\ldots,S_j\}, l \in \{1,2,\ldots,l_{js}\}\}$ と表せる．w_{jsl} は受検者 j の小論文における s 番目の文の l 番目の単語であり，S_j はその小論文の文数，l_{js} は s 番目の文の単語数である．

このモデルでは，まず受検者の小論文が文ごとに Embedding 層，Convolution 層，Attention Pooling 層 [9] に入力され，文単位の分散表現の系列に変換される．次に，この出力系列に対して，Recurrent 層 [23]，Attention Pooling 層が適用される．さらに単語数や可読性，文章の複雑さなどを表す人手で設計した特徴量のベクトル F_j [16] を結合することで文章単位の分散表現 $\boldsymbol{c}_j$ が得られる．このモデルでは，このベクトル $\boldsymbol{c}_j$ を全結合層に入力し，多次元 IRT における能力値ベクトル $\boldsymbol{\theta}_j$ に対応する値を $\boldsymbol{\theta}_j = \boldsymbol{W}_t\boldsymbol{c}_j + \boldsymbol{b}_t$ で求める．ここで，$\boldsymbol{W}_t$ は重み行列，$\boldsymbol{b}_t$ はバイアスベクトルを表す．

最後に，得られた $\boldsymbol{\theta}_j$ を用いて多次元 GPCM 層で式（1）を計算することで，各評価観点 m に対する得点の出力確率が得られる．得点予測の際には，期待得点 $\sum_{k=1}^{K_m} kP_{jmk}$ を予測得点とする．

モデルの学習は以下の多クラス交差エントロピー（Categorical Cross-Entropy；CCE）誤差を損失関数として誤差逆伝播法で行われる．

$$\mathcal{L}_{CCE} = -\frac{1}{JM}\sum_{j=1}^{J}\sum_{m=1}^{M}\sum_{k=1}^{K_m} y_{jmk}\log(P_{jmk}) \tag{2}$$

表 1　課題別の平均 QWK スコア

モデル	課題番号								
	1	2	3	4	5	6	7	8	Avg.
Ridley	.685	**.655**	**.660**	.720	.706	**.750**	.694	.568	**.680**
Shibata–Uto	.669	.623	.627	.717	**.715**	.731	.697	**.604**	.673
提案モデル	**.691**	.640	.652	**.721**	.709	.745	**.702**	.577	**.680**

表 2　課題 1，2 の評価観点別 QWK スコア

課題	モデル	評価観点					
		全体	Cont	Org	WC	SF	Conv
1	Ridley	**.823**	**.700**	.653	.669	.643	.624
	Shibata–Uto	.801	.677	.634	.654	.635	.613
	提案モデル	.813	.689	**.665**	**.671**	**.663**	**.644**
2	Ridley	.686	**.632**	**.653**	**.660**	**.645**	**.651**
	Shibata–Uto	.639	.619	.616	.633	.622	.608
	提案モデル	**.691**	.630	.626	.642	.631	.620

表 3　課題 3，4，5，6 の評価観点別 QWK スコア

課題	モデル	評価観点				
		全体	Cont	PA	Lang	Nar
3	Ridley	**.668**	**.674**	**.671**	**.616**	**.673**
	Shibata–Uto	.603	.647	.649	.594	.643
	提案モデル	.658	.662	.665	.605	.671
4	Ridley	.769	**.751**	.735	.625	.722
	Shibata–Uto	.742	.741	.737	**.629**	**.734**
	提案モデル	**.794**	.740	**.739**	.617	.716
5	Ridley	**.797**	.713	.696	.646	.680
	Shibata–Uto	.796	**.730**	**.710**	.653	**.683**
	提案モデル	.791	.716	.707	**.654**	.677
6	Ridley	**.810**	**.822**	**.787**	**.642**	**.690**
	Shibata–Uto	.777	.802	.771	.640	.666
	提案モデル	.803	.812	.781	**.642**	.687

ただし，J は訓練データの小論文数，M は評価観点数，y_{jmk} は受検者 j の評価観点 m における真得点が k のときに 1 をとり，それ以外のときに 0 をとるダミー変数である．

4　性能評価

ここでは，Shibata & Uto のモデルの性能を詳細に分析するために，評価観点別の予測精度の評価を行う．本実験では，実データとして，Automated Student Assessment Prize (ASAP) と，ASAP++ [24] を用いる．ASAP は AES 研究の分野で広く使用されるデータセットである．ASAP と ASAP++には 8 つの小論文課題に関する答案が含まれており，それぞれの答案に対して全体得点と評価観点別の得点が与えられている．小論文数の課題ごとの平均は約 1622，平均単語数は 275 である．なお課題 1 と課題 2 は，Content（Cont），Organization（Org），Word Choice（WC），Sentence Fluency（SF），Conventions（Conv），課題 3 から課題 6 は Cont，Prompt Adherence（PA），

表 4　課題 7 の評価観点別 QWK スコア

モデル	評価観点				
	全体	Cont	Org	Conv	Style
Ridley	.765	**.766**	.658	.625	**.656**
Shibata–Uto	.774	.755	.668	**.632**	.653
提案モデル	**.785**	.755	**.688**	.629	.652

表 5　課題 8 の評価観点別 QWK スコア

モデル	評価観点						
	全体	Cont	Org	WC	SF	Conv	Voice
Ridley	.656	.559	.589	.556	.656	.559	.589
Shibata–Uto	**.660**	**.605**	**.624**	**.599**	**.660**	**.605**	**.624**
提案モデル	.644	.564	.578	.539	.644	.564	.578

Language（Lang），Narrativity（Nar），課題 7 は Cont，Org，Conv，Style，課題 8 は課題 1 と 2 に共通する評価観点に加えて Voice といった評価観点でそれぞれ得点付けされている．

本論文では，Shibata & Uto のモデルの次元数をそれぞれ 1，2，3 と変化させて得点予測精度を評価する実験を行った．またベースラインとして Ridley ら [16] のモデルの精度評価も行った．なお各モデルにおける Embedding 層では，共通して 50 次元の GloVe [25] による事前学習済みの単語埋め込みを利用した．モデルの性能評価は，5 分割交差検証を用いて行った．5 分割交差検証は課題ごとに独立して実施し，エポック数は全てのモデルで 50 とした．評価関数には，2 次の重み付きカッパ係数（Quadratic Weighted Kappa；QWK）を用いた．

実験結果を表 1 に示す．表 1 では観点ごとに QWK スコアを計算し，その平均スコアを課題ごとに示している．各条件で最も精度が高い手法の結果を太字で示してある．なお，表には Shibata & Uto のモデルの中で最も精度が高かった 3 次元のモデルの数値のみ示している．表 1 の平均 QWK スコアから，Shibata & Uto のモデルは Ridley らのモデルと比較して予測精度が低下していることが読み取れる．この原因をより詳しく分析するために，各評価観点ごとの QWK スコアを表 2-5 に示す．これらの表から特に課題 1-4 と 6 における全体得点の予測精度が低い傾向が読み取れる．このことは全体得点とその他の観点別の得点を単一の IRT モデルで表現することが困難な場合があることを示唆いている．そこ

で，本研究では Shibata & Uto のモデルに対して全体得点を個別に予測する構造を持たせたモデルを提案する．

5　提案モデル

本研究では，従来の Shibata & Uto モデル（以下，従来モデルと呼ぶ）に全体得点を予測する固有のニューラルネットワークを追加したモデルを提案する．提案モデルの概念図を図 1（右）に示す．また，従来モデルは全体得点 $\hat{y}_{j1}$ を観点別得点の一つとして扱ったが，提案モデルでは全体得点と観点別得点を区別する点に注意されたい．

提案モデルは，入力層から一層目の Attention Pooling 層までは従来モデルと同じ構造であり，このネットワークを用いて文単位の分散表現の系列を生成した後，全体得点のみを予測するネットワークと観点別得点を予測するネットワークに分岐する．ここで，観点別得点を出力するネットワークは従来モデルと同じ構造となっている．他方，全体得点固有のネットワークは Recurrent 層から Concatenate 層までは観点別得点固有のネットワークと同じ構造であるが，出力層の設計が異なる．具体的には，Concatenate 層までで全体得点の予測に使用する文章単位の分散表現 $\tilde{\boldsymbol{c}}_j$ を算出し，この $\tilde{\boldsymbol{c}}_j$ に対してシグモイド関数を活性化関数に持つ全結合層を適用させ，受検者 j の全体得点 $\hat{y}_{j1}$ を $\hat{y}_{j1} = \sigma(\boldsymbol{W}_o\tilde{\boldsymbol{c}}_j + b_o)$ と予測する．ここで，σ はシグモイド関数，$\boldsymbol{W}_o$ は重み，b_o はバイアスを表す．なお，この出力層の設計は，全体得点を予測する一般的な自動採点モデル（e.g., [6, 8, 9]）と同様の設計である．この出力層は得点予測にシグモイド関数を使用しているため，$\hat{y}_{j1}$ は 0 から 1 の間の値をとる．実際の得点尺度がこれと異なる場合には，$\hat{y}_{j1}$ を一次変換して実際の得点尺度に合わせる．

提案モデルは全体得点に関する誤差を平均二乗誤差（Mean Squared Error；MSE）で算出し，他の観点別得点に関する誤差は従来モデルと同様に CCE 誤差で算出する．つまり，提案モデルは損失関数を次式のように MSE 誤差と CCE 誤差の和として定義し，誤差逆伝播法で学習を行う．

$$\mathscr{L}_{total} = \frac{1}{J}\sum_{j=1}^{J}(\hat{y}_{j1} - y_{j1})^2 + \mathscr{L}'_{CCE} \tag{3}$$

ここで y_{j1} は受検者 j の真の全体得点を，$\mathscr{L}'_{CCE}$ は式（2）において全体得点を含めない場合の CCE 誤差を表す．また，モデルの各種ハイパーパラメータは先行研究 [18, 19] に合わせ，最適化アルゴリズムには RMSProp [26] を用いる．

なお提案モデルの解釈の方法については従来モデルとほとんど同じであるため，詳しい解釈方法は文献 [18, 19] を参照されたい．

6　得点予測精度の評価実験

ここでは提案モデルの精度評価実験を行う．実験の手順は 4 章と同様である．表 1 に最も精度が高かった 3 次元の提案モデルの結果を示した．表 1 の平均 QWK スコアより，提案モデルは，従来モデルの精度を上回っていることが読み取れる．さらに提案モデルは Ridley らのモデルと比較しても同等の精度を達成していることがわかる．このことから，提案モデルは予測精度を改善できたといえる．

次に，各課題に関する評価観点別の QWK スコアを表 2-5 に示す．これらの表より，特に課題 1-4 と 6 において全体得点の予測精度が大きく向上していることがわかる．この結果から提案モデルの全体得点固有のネットワークが有効に機能していることが確認できる．

7　まとめ

本研究では，従来の複数観点同時自動採点モデルの解釈性を保持しつつ，予測精度を高めたモデルを提案した．提案モデルを用いた実験から，提案モデルは全体得点の予測精度向上に寄与することがわかった．

なお実験結果から全体得点と観点別得点を単一の IRT モデルで表現していることが従来の Shibata & Uto モデルの精度低下の原因であると考えられる．しかし，本研究で利用している多次元 GPCM は一つの観点に対して完全に独立な能力尺度を与えることが可能であるため，理論的には多次元 IRT における能力次元数を十分に多く設定すれば，尺度構成に関わらず柔軟に適合できるはずである．しかし実際には従来モデルの予測精度は高々 3 次元で十分であることがわかっており，次元数が増加しても従来モデルの予測精度は改善しない．そのため従来モデルの全体得点における予測精度の低さは，多次元 GPCM 層よりも下の層における表現力の低さに起因すると考えられる．この仮説を検証する方法として，従来モデルを各能力次元に固有のニューラルネットワークを持たせることが考えられる．この手法については今後の課題としたい．

謝辞

本研究は JSPS 科研費 19H05663, 20K20817, 21H00898 の助成を受けたものです.

参考文献

[1] Yigal Attali and Jill Burstein. Automated essay scoring with e-rater® v.2. **The Journal of Technology, Learning and Assessment**, Vol. 4, No. 3, pp. 1–30, 2006.

[2] Hongbo Chen and Ben He. Automated essay scoring by maximizing human-machine agreement. In **Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing**, pp. 1741–1752, 2013.

[3] Peter Phandi, Kian Ming A Chai, and Hwee Tou Ng. Flexible domain adaptation for automated essay scoring using correlated linear regression. In **Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing**, pp. 431–439, 2015.

[4] Masaki Uto. A review of deep-neural automated essay scoring models. **Behaviormetrika**, Vol. 48, No. 2, pp. 1–26, 2021.

[5] 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. Association for Computational Linguistics, 2016.

[6] 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.

[7] 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, 2016.

[8] Yi Tay, Minh C Phan, Luu Anh Tuan, and Siu Cheung Hui. Skipflow: Incorporating neural coherence features for end-to-end automatic text scoring. In **Thirty-Second AAAI Conference on Artificial Intelligence**, pp. 5948–5955, 2018.

[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**, pp. 153–162, 2017.

[10] Youmna Farag, Helen Yannakoudakis, and Ted Briscoe. Neural automated essay scoring and coherence modeling for adversarially crafted input. arXiv, 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] 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.

[13] Sandeep Mathias and Pushpak Bhattacharyya. Can neural networks automatically score essay traits? In **Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications**, pp. 85–91, 2020.

[14] Mohamed A. Hussein, Hesham A. Hassan, and Mohammad Nassef. A trait-based deep learning automated essay scoring system with adaptive feedback. **International Journal of Advanced Computer Science and Applications**, Vol. 11, No. 5, pp. 287–293, 2020.

[15] Farjana Sultana Mim, Naoya Inoue, Paul Reisert, Hiroki Ouchi, and Kentaro Inui. Unsupervised learning of discourse-aware text representation for essay scoring. In **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop**, pp. 378–385, 2019.

[16] Robert Ridley, Liang He, Xin-yu Dai, Shujian Huang, and Jiajun Chen. Automated cross-prompt scoring of essay traits. In **Proceedings of the AAAI Conference on Artificial Intelligence**, Vol. 35, pp. 13745–13753, 2021.

[17] Masaki Uto. A multidimensional item response theory model for rubric-based writing assessment. In **Artificial Intelligence in Education**, pp. 420–432. Springer International Publishing, 2021.

[18] Takumi Shibata and Masaki Uto. Analytic automated essay scoring based on deep neural networks integrating multidimensional item response theory. In **Proceedings of the 29th International Conference on Computational Linguistics**, pp. 2917–2926, 2022.

[19] 柴田拓海, 宇都雅輝. 多次元項目反応理論と深層学習を用いた複数観点同時自動採点手法. 電子情報通信学会論文誌 D, Vol. 106, No. 1, pp. 47–56, 2023.

[20] Masaki Uto. A multidimensional generalized many-facet Rasch model for rubric-based performance assessment. **Behaviormetrika**, Vol. 48, No. 2, pp. 425–457, 2021.

[21] F.M. Lord. **Applications of item response theory to practical testing problems**. Erlbaum Associates, 1980.

[22] Lihua Yao and Richard D. Schwarz. A multidimensional partial credit model with associated item and test statistics: An application to mixed-format tests. **Applied Psychological Measurement**, Vol. 30, No. 6, pp. 469–492, 2006.

[23] Jeffrey L Elman. Finding structure in time. **Cognitive science**, Vol. 14, No. 2, pp. 179–211, 1990.

[24] Sandeep Mathias and Pushpak Bhattacharyya. ASAP++: Enriching the ASAP automated essay grading dataset with essay attribute scores. Proceedings of the Eleventh International Conference on Language Resources and Evaluation, 2018.

[25] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In **Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing**, pp. 1532–1543, 2014.

[26] Yann Dauphin, Harm de Vries, and Yoshua Bengio. Equilibrated adaptive learning rates for non-convex optimization. In **Advances in Neural Information Processing Systems**, Vol. 28, pp. 1504–1512. Curran Associates, Inc., 2015.