項目反応理論に基づく能力推定値を活用した短答記述式問題自動採点手法

内田 優斗　　宇都 雅輝
電気通信大学
{uchida, uto}@ai.lab.uec.ac.jp

1 はじめに

近年，論理的思考力や判断力，表現力などの高次の能力を測定する手法のひとつとして短答記述式問題の利用が注目されている [3]．短答記述式問題は，TOEFL や GMAT（Graduate Management Adimossion Test）などの世界の様々な大規模試験で導入されており，日本でも大学入試センター試験に代わる大学入学共通テストで導入が予定されていた[1]．他方で，これらの試験のように受験者数が数十万人規模となる大規模試験に短答記述式問題を導入する場合，採点の信頼性や時間的・金銭的コストなどの点で課題が残る．これらの課題を解決する手法の一つとして自動採点が注目されている [15]．

短答記述式問題の自動採点技術は古くから研究されてきた [1]．自動採点のアプローチとしては，専門家が設計した特徴量を用いる手法が伝統的に利用されてきた．代表的なモデルとしては，世界最大のテスト機関である Educational Testing Service（ETS）が開発した手法 [2] が知られている．他にも様々な手法 [6, 5] が提案されており，様々な試験や研究で活用されてきた．

他方で，近年では，深層学習を用いたアプローチが多数提案され，高い精度を達成している．このアプローチでは，採点済み解答文のデータセットから得点予測に有効な特徴量を自動で獲得するため，特徴量の設計やチューニングを行うことなく，問題ごとの高精度な自動採点を実現できる．深層学習を利用した代表的な自動採点モデルとしては，畳み込みニューラルネットワーク（CNN）を用いた手法 [14] や Long Short Term Memory（LSTM）を用いた手法 [12, 4, 3] が知られている．さらに，直近では，自己注意機構（Self attention）で構成される Transformer モデルの一種である Bidirectional Encoder Representations from Transformers（BERT）を用いた手法も提案されている [7, 8]．これらの自動採点モデルは，多段階得点の予測においては 70%以上の正答率を達成し [4, 3]，正誤の 2 値分類においては 90%を超える正答率を達成する [14, 12, 7, 8] ことも多い．

このように深層学習を用いた自動採点モデルは高精度な得点予測を実現しているが，特に大学入試や資格試験などのハイステークス試験においては，わずかな採点ミスでも多数の受験者に深刻な影響を与える可能性があるため，自動採点にはさらなる精度向上が求められる．そこで，本研究では，深層学習自動採点モデルの精度を改善することを目標とする．

精度改善のために，本研究では，短答記述式問題が一般に客観式問題を含むテストの一部として出題されることに着目する．テストは特定の能力を測定するツールであるため，同一テスト上の短答記述式問題と客観式問題は共通の能力を測定していると仮定できる．すなわち，同一テスト内の客観式問題から推定される各受験者の能力値は，短答記述式問題の得点を予測するための有益な補助情報になると考えられる．そこで本研究では，項目反応理論（IRT : Item Response Theory）を用いて客観式問題から求められる受験者の能力推定値を組み込んだ新たな深層学習自動採点モデルを提案する．提案アプローチは，いずれの深層学習自動採点モデルでも適用できるが，本研究では，最も標準的に利用されている LSTM に基づくモデルへの組み込みを行う．また本研究では，実データ実験により，提案手法が精度向上に有効であることを評価する．

2 深層学習を用いた自動採点手法

本研究では，最も代表的な深層学習自動採点モデルである LSTM に基づくモデルを基礎モデルとして用いる．LSTM に基づくモデルには様々なバリエーションが存在するが，ここでは最も標準的な Riordan ら [4] のモデルについて説明する [9]．

[1]採点の信頼性担保の困難さや採点期間の短さなどが問題視され，延期を余儀なくされた．

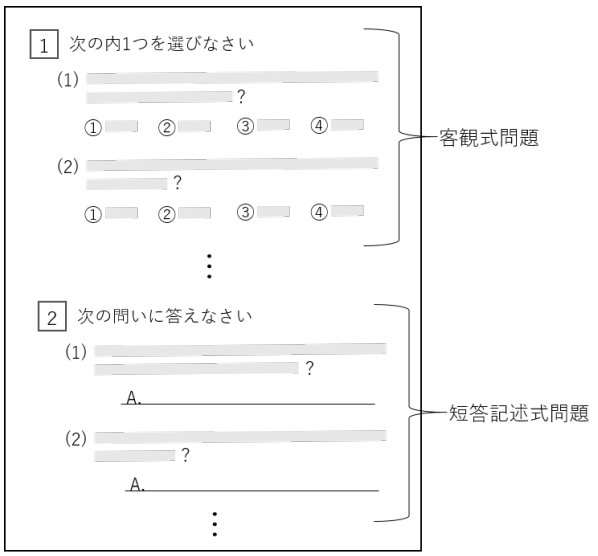

図 1: 客観式問題，短答記述式問題を含むテストの例

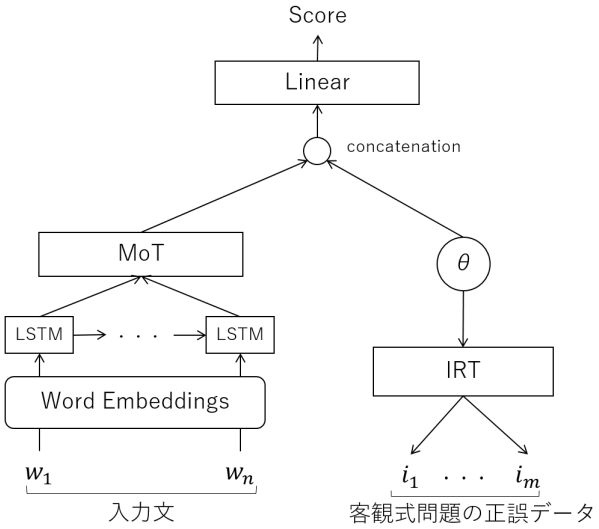

図 2: 提案モデル

このモデルは，解答文の単語系列を入力とし，LSTM を含む 4 つの層で構成される多層ニューラルネットワークを通して得点を出力する．具体的には，1 層目の Lookup table layer で入力単語を埋め込み表現（word embeddings）に変換し，2 層目の Recurrent layer で埋め込みベクトルの系列を LSTM に入力する．次に，3 層目の Mean over Time layer（MoT）で，LSTM の出力系列 $\{\mathbf{h_1}, \mathbf{h_2}, ..., \mathbf{h_N}\}$ の平均 $\mathbf{M} = \frac{1}{N}\sum_{n=1}^{N} \mathbf{h_n}$ を求め，4 層目の Linear layer with Sigmoid activation で，MoT の出力を $s = \sigma(\mathbf{WM} + \mathbf{b})$ で得点 s に変換する．ここで $\mathbf{W}, \mathbf{b}$ はそれぞれ重み行列とバイアスを表すパラメータである．この手順で得られた s は 0 から 1 の範囲の値となるため，多段階得点を扱う場合には，s を実際の採点尺度に線形変換する必要がある．モデル学習は，平均二乗誤差を損失関数とする誤差逆伝播法を用いて行う．

本研究では，この深層学習自動採点手法の精度改善を目標とする．本研究の主なアイデアは，図 1 のように，短答記述式問題が一般に客観式問題を含むテストの一部として出題されることに着目し，客観式問題への正誤データから求められる各受験者の能力推定値を短答記述式問題の得点予測に利用することにある．本研究では，この能力推定に IRT を利用する．

3 項目反応理論

IRT はコンピュータ・テスティングの普及とともに，近年様々な分野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 [10]．IRT は特定の受験者集団やテスト項目に依存しない形で受験者のパラメータや項目の性質を推定することが可能である [13, 11]．IRT の利点としては以下のような点が挙げられる [10]. 1）能力推定の低い異質項目の影響を小さくして能力推定を行うことができる．2）異なる項目への受験者の反応を同一尺度上で評価できる．3）欠測データから容易にパラメータを推定できる．

本研究では，代表的な IRT モデルである，2 母数ロジスティックモデル（2PLM）を用いる．2PLM は受験者 j が項目 i に正答する確率を次式で表す．

$$P_{ij} = \frac{exp\left(a_i(\theta_j - b_i)\right)}{1 + exp\left(a_i(\theta_j - b_i)\right)} \tag{1}$$

ここで，a_i, b_i は項目 i の識別力，困難度を表し，θ_j は受験者 j の能力を表す．困難度 b_i は正答確率が 0.5 となる能力値 θ，識別力 a_i は能力値 $\theta = b_i$ 付近の能力をどの程度の精度で識別できるかを表す．

本研究では，この IRT モデルを用いて客観式問題への正誤データから推定される能力値を補助情報として利用できる新たな LSTM 自動採点モデルを提案する．

4 提案手法

提案モデルの概念図を図 2 に示す．提案モデルは，客観式問題への正誤データから IRT を用いて推定された能力値 θ と解答文の単語系列を入力として得点を予測する．具体的には，2 で紹介したモデルと同様に，各解答文の単語系列から Lookup table layer，Recurrent layer，MoT layer を通してベクトル表現 $\mathbf{M}$ を求め，それを各受験者の能力推定値 θ と結合（Concatenation）

表 1: データの基礎統計量

	問題 1			問題 2			問題 3		
	合計	観点 1	観点 2	合計	観点 1	観点 2	合計	観点 1	観点 2
得点	1.783	0.832	0.951	2.223	1.290	0.933	2.838	1.432	1.405
	（1.179）	（0.652）	（0.999）	（1.205）	（0.906）	（0.860）	（1.375）	（0.899）	（0.836）
文字数	27.1 （7.51）			33.0 （10.6）			49.6 （13.6）		

したのち，Linear layer with Sigmoid activation で得点を算出する．

2 で紹介したモデル同様，多段階得点のデータを扱う場合には s を実際の採点尺度に線形変換する．また，モデル学習は，平均二乗誤差を損失関数として誤差逆伝搬法で行う．IRT モデルの学習には，Expectation Maximization（EM）アルゴリズムやマルコフ連鎖モンテカルロ (MCMC) アルゴリズムなどの任意の方法を利用する．

5 実験

本章では，実データを用いて提案手法の有効性を評価する．

5.1 実データ

本研究では，ベネッセ教育総合研究所で開発している読解テストの解答データを実験データとして扱う．本データは短答記述式問題 3 問分に対する 511 人分の解答文情報と，それらに対して評価者が与えた得点，および同受験者の客観式問題 44 問への正誤データで構成される．短答記述式問題の得点は，2 つの評価観点について 3 段階評価で与えられ，その 2 つを足したものを合計点として扱う．採点は 2 名の評価者で行われ，採点が一致しなかった場合には，もう一名の評価者が確認して最終得点を決定した．短答式問題の得点と解答文の文字数に関する記述統計量を表 1 に示す．記述統計量は平均（標準偏差）で記述する．以降，それぞれの合計，観点を項目と呼称する．

5.2 実験手順

5 分割交差検証法により既存の LSTM を用いたモデルと提案手法の性能を比較する．また，提案手法において能力値の代わりに，客観式問題への正誤データベクトルを利用した場合についても比較を行う．

IRT の学習には MCMC アルゴリズムを利用した．MCMC のサンプリング数は 10,000 とし，バーンイン 5,000 以降のサンプルを 100 刻みで抽出し，その期待値を点推定値とした．深層学習モデル学習は，epoch 数を 100，バッチ数を 10，Word Embedding，LSTM の出力ベクトルの次元数をともに 100 とした．これらチューニングパラメータはグリッドサーチにより決定した．また，本実験は観点別に行った．

評価指標には自動採点の性能評価において一般に使われている Quadratic Weighted Kappa（QWK）を使用した．モデル学習はシードを変えて 30 回行い，その平均値を求めた．また，提案手法と従来手法の差異を確認するために，対応のある t 検定を行った．

5.3 実験結果

本実験の結果を表 2 に示す．表中の太字はそれぞれの項目の最大値を示す．また,「提案手法（素点）」は，提案手法において能力値の代わりに正誤データを利用したモデルを意味する．表 2 より，提案手法はすべての場合で従来手法より QWK が改善しており，最大では問題 1 の観点 1 で 4.4%と大幅に改善している．また，t 検定の結果，提案手法が従来手法より有意水準 5%で高い精度を示したことが確認できた．

提案手法において IRT の利用の有無を比較すると，性能に大きな差異はなく，検定でも有意差は認められなかった．このことは，正誤データの背後に存在する「能力値」という潜在情報が，記述式問題の得点予測においては本質的な情報となっていることを示唆している．3 章で述べたように，IRT には項目特性を考慮した高精度な能力推定が可能であり，異なる問題に解答した受験者の能力を同一尺度上で比較できるなどの利点がある．さらに，提案モデルに正誤データを直接利用すると，解答に欠測がある場合にモデル学習が困難になるが，IRT の能力値は欠測データからも推定できるため工夫なくモデルを学習できる．

以上から，提案手法が短答記述式自動採点モデルの精度改善に有効であることが示された．

6 まとめと今後の課題

本研究では，深層学習を用いた短答記述式問題自動採点の精度を改善することを目標とし，IRT により推

表 2: 実験結果

	問題 1			問題 2			問題 3				
	合計	観点 1	観点 2	合計	観点 1	観点 2	合計	観点 1	観点 2	平均	p 値
従来手法	0.640	0.600	0.814	0.845	0.916	0.861	0.719	0.703	0.768	0.763	-
提案手法	**0.666**	**0.625**	**0.830**	**0.847**	0.922	0.862	0.721	0.707	0.775	**0.773**	0.009
提案手法（素点）	0.657	0.616	0.816	0.841	**0.925**	**0.870**	**0.727**	**0.711**	**0.779**	0.771	0.002

定した客観式問題における受験者の能力値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．実データ実験の結果，能力値を組み込むことで予測精度が有意に改善することが確認できた．

今後は，より多様なデータセットを用いて提案手法の有効性を検証していきたい．また本手法は様々な深層学習モデルと組み合わせて利用できるため，今後は他の深層学習自動採点モデルへの適用も行いたい．

謝辞

実験で利用した実データはベネッセ教育総合研究所から提供を受けた．ベネッセ教育総合研究所の堂下雄輝氏と加藤嘉浩氏に感謝の意を表します．

参考文献

[1] Steven Burrows, Iryna Gurevych, and Benno Stein. The eras and trends of automatic short answer grading. *International Journal of Artificial Intelligence in Education*, Vol. 25, No. 1, pp. 60–117, 2015.

[2] Michael Heilman and Nitin Madnani. Ets: Domain adaptation and stacking for short answer scoring. In *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pp. 275–279, 2013.

[3] Tomoya Mizumoto, Hiroki Ouchi, Yoriko Isobe, Paul Reisert, Ryo Nagata, Satoshi Sekine, and Kentaro Inui. Analytic score prediction and justification identification in automated short answer scoring. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pp. 316–325, Florence, Italy, August 2019. Association for Computational Linguistics.

[4] Brian Riordan, Andrea Horbach, Aoife Cahill, Torsten Zesch, and Chong Min Lee. Investigating neural architectures for short answer scoring. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pp. 159–168, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[5] Keisuke Sakaguchi, Michael Heilman, and Nitin Madnani. Effective feature integration for automated short answer scoring. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1049–1054, Denver, Colorado, May–June 2015. Association for Computational Linguistics.

[6] Md Arafat Sultan, Cristobal Salazar, and Tamara Sumner. Fast and easy short answer grading with high accuracy. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1070–1075, San Diego, California, June 2016. Association for Computational Linguistics.

[7] Chul Sung, Tejas Indulal Dhamecha, and Nirmal Mukhi. Improving short answer grading using transformer-based pre-training. In *International Conference on Artificial Intelligence in Education*, pp. 469–481. Springer, 2019.

[8] Chul Sung, Tejas Dhamecha, Swarnadeep Saha, Tengfei Ma, Vinay Reddy, and Rishi Arora. Pre-training bert on domain resources for short answer grading. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 6073–6077, 2019.

[9] 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.

[10] 宇都雅輝, 植野真臣. パフォーマンス評価のための項目反応モデルの比較と展望. 日本テスト学会誌, Vol. 12, No. 1, pp. 55–75, 2016.

[11] 加藤健太郎, 山田剛史, 川端一光. R による項目反応理論. 株式会社 オーム社, 2014.

[12] 高井浩平, 竹谷謙吾, 早川純平, 森康久仁, 須鎗弘樹. Lstm と attention を用いた自動採点及び採点支援の実用化に向けて. 人工知能学会全国大会論文集 一般社団法人 人工知能学会, pp. 2I5J905–2I5J905. 一般社団法人 人工知能学会, 2019.

[13] 笹川智子, 金井嘉宏, 村中泰子, 鈴木伸一, 嶋田洋徳, 坂野雄二. 他者からの否定的評価に対する社会的不安測定尺度 (fne) 短縮版作成の試み: 項目反応理論による検討 (原著). 行動療法研究, Vol. 30, No. 2, pp. 87–98, 2004.

[14] 寺田凛太郎, 久保顕大, 柴田知秀, 黒橋禎夫, 大久保智哉. ニューラルネットワークを用いた記述式問題の自動採点. 第 22 回言語処理学会年次大会発表論文集, pp. 370–373, 2016.

[15] 中島功滋. 短答式記述答案の採点支援ツールの開発と評価. 言語処理学会第 17 回年次大会発表論文集, pp. 611–614, 2011.