大学入学共通テスト試行調査における短答式記述答案の完全自動採点

岡 知樹 [1]　Hung Tuan Nguyen[2]　Cuong Tuan Nguyen[2]　中川 正樹 [2]　石岡 恒憲 [3]
[1] 東京大学　[2] 東京農工大学　[3] 大学入試センター
oka-haruki497@g.ecc.u-tokyo.ac.jp　ntuanhung@gmail.com　ntcuong2103@gmail.com
nakagawa@cc.tuat.ac.jp　tunenori@rd.dnc.ac.jp

概要

自然言語処理の教育分野への応用タスクとして、短答式記述問題の自動採点に関する研究が行われている。実際の教育現場では、短答式記述問題の解答用紙はほとんどが手書きであり自動採点の実応用での障壁となっている。本研究では手書き文字認識と自然言語処理を用いて、短答式問題の手書き回答を完全に自動採点するシステムを開発した。本研究で提案した完全自動採点システムでも、人間の採点に匹敵する高い精度で採点できることを確認した。

1　はじめに

現在の教育現場では，言語学で培った能力を適切に評価するために記述式問題が多く導入されている．採点の効率化や安定化のために，近年では人工知能による自動採点の研究が進んでいる．英語を対象とした短答式記述答案の自動採点は，深層学習を用いた手法が提案されて以来，その性能が向上してきた [1, 2, 3, 4, 5] ．特に，近年では Transformer ベースの言語モデルを用いた短答式記述答案の自動採点 [6, 7, 8, 9, 10, 11, 12, 13] が考案されている．こうした背景から，実際の模擬試験のデータを使用した最新の研究も存在する [14, 9].

しかし，これらの研究には 2 つの問題が存在する．1 つ目は，短答式記述答案の自動採点には余分な手作業が必要なことである．記述答案の多くは手書きであるため，手書きデータを電子媒体に変換するには手間がかかる．また，従来の方法では，正確性を確保するために，採点の目安となるアノテーションを付与している．教育現場での実用化を考えると，これらの手間を省くための改良が必要である．本研究では，アノテーション付与や手書きの答案をテキストデータに変換といった，データ処理を確実に省略できる全自動採点システムを開発した．2 つ目は，実際の教育現場で扱うデータが過小であり，大規模な検証ができなかったことである．プライバシーの観点からデータ数が限られていた．本研究では大学入学共通テスト試行調査のデータを用いて実験を行い，大規模な教育現場のデータでも高い採点精度を保証できることを明らかにした．

2　共通テスト試行調査のデータ

2.1　概要

2017 年と 2018 年に実施された大学入学共通テスト試行調査の国語の短答式記述問題を使用する．試験問題は本番と同じ方法で作成し，試験問題の質は厳密に検証されている．今回は日本の約 38 %の高校がこの試行調査に参加し，約 6 万人が受験した．

2.2　国語の短答式記述問題

試行調査の国語は，5 つの大問で構成されている．そのうちの 1 つが短答式記述問題であり，3 つの小問で構成されている．2017 年では，それぞれ 50 字，25 字，120 字の，2018 年では，30 字，40 字，120 字の字数制限が設けられた．図 1 に 2018 年に実施された短答式記述問題の例（問 1，30 字）を示す．

問題文
“. . . ことばの全く通じない国に行って，相手に
なにかを頼んだり尋ねたりする状況を考えてみよう．
この時には，**指差しが魔法のような力を発揮する**はずだ. . . ”

設問
“指差しが魔法のような力を発揮する” とはどういうことか?
三十字以内で書け．

答案例
ことばを用いなくても意思が伝達できること．

Score: 3/3

図 1　2018 年に実施された短答式記述問題の例

3 方法

3.1 タスク設定

手書き文字認識モデルを用いてテキストデータに変換した短答式記述問題の答案を入力し，対応する予測得点を出力する．次に，採点基準に基づいた人手による採点結果と比較することで，本研究の採点モデルがスコアを正しく予測できるかを検証する．

図 2 にタスクの流れを示す．答案データの文字を修正することなく，また，答案にアノテーションを加えることなく，スコアを出力し精度を評価する．

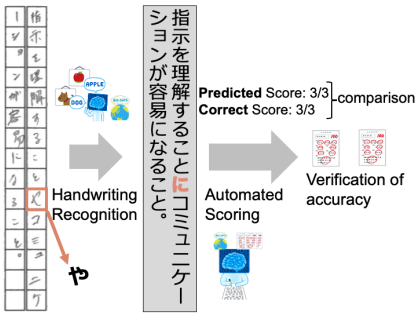

図 2 タスクの流れ

3.2 手書き文字認識モデル

ETL（Extracting, Transforming, and Loading）データベースは，オフラインの日本語手書き文字を収録したデータベースである．このデータベースは，異なる条件で収集された 9 つのデータセットから構成されている [15]．収集されたサンプルは，大学入試の答案用紙のように区切られたボックスに書かれているため，ETL データベースはオフラインの日本語手書き文字認識モデルを構築するのに適している．

本研究では複数の畳み込みニューラルネットワーク（CNN）のアンサンブル学習した文字認識モデルを使用する．文字認識モデルは，VGG（Visual Geometric Group)[16]，MobileNet[17]，ResNet（Residual Network)[18]，ResNext[19] であり，それぞれ 16 層，24 層，34 層，50 層で構成されている．

これらの CNN の学習には，回転，剪断，拡大縮小，ぼかし，コントラスト，ノイズ付加などの変換を適用し，サンプル数が 100 万程度であったため，過学習の問題を回避している．ETL データベースを使ってこれらの CNN を学習した後，手書き答案の中から手動でラベル付けされた 100 個のサンプルを使って，CNN をファインチューニングする．

訓練された文字認識モデルは，確率の C 次元ベクトルとして予測出力を与える．ここで C は各文字サンプルのカテゴリー数である．これらの予測出力は，1.0 の均等な重みで平均化され，アンサンブル予測出力となる．このようにして，アンサンブル予測出力の中で最も確率の高いカテゴリがトップの予測となる．図 3 は，16 層，24 層，50 層の CNN を判定する手順を示したものである．

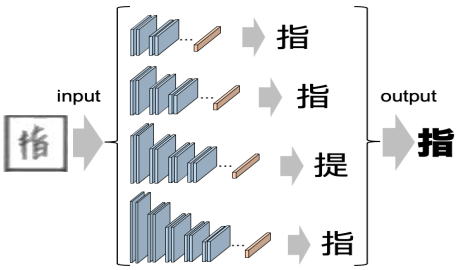

図 3 アンサンブル学習した CNN の文字認識モデル

また，文字の中には曖昧なものもあるため，N-gram 言語モデルを用いて，言語的な文脈を用いて誤認識した文字を修正する．各文字について，認識スコアと各文字の言語スコアを合わせたスコアを計算する．認識スコアはアンサンブル学習するそれぞれの文字認識モデルによって生成された，文字の確率積である．

次に，言語スコアは日本語 Wikipedia で事前に学習された 5-gram の日本語言語モデルに基づいて認識された文字の確率積である．

3 つ目の複合スコアは，認識スコアと言語スコアを線形結合したもので，$\alpha \in [0, 1]$ の重みがついている．複合スコアに応じて，ビーム幅 10 の文字列に沿ったビームサーチアルゴリズムを採用し，複合スコアの高い上位 10 個の候補が抽出されるようにする．本研究では最も高いスコアの候補の文字のみを自動採点に使用する．

3.3 自動採点モデル

日本語で同様の短答式記述答案の自動採点を行う舟山ら [9] や水本ら [14] の手法は，Bi-LSTM に attention を追加したものである．これらの手法では，各採点基準やルーブリックに基づいて予測スコアを出力する．しかし，本研究の手法では，各採点基準のスコアを蓄積するのではなく，全体のスコアを予測する．本研究では，日本語 wikipedia で事前学習した BERT[20][1] をファインチューニングする

1） https://github.com/huggingface/transformers

ことで，マルチラベル分類モデルを明示的に利用する．その手順は以下の通りである（図 4）．

1. $x = \{x_1, x_2, \ldots, x_n\}$ を，手書き文字認識によってテキストデータに変換された答案文として事前学習した BERT に入力し，その答案文に対する予測スコア $s \in C = \{0, \ldots, N\}$ をラベルの出力として与える．
2. BERT 全 12 層のうち，隠れ層の最後の 4 層の [CLS] トークンのベクトルを抽出する．これらを組み合わせることで，最終層の [CLS] トークンのベクトルのみを使用する場合と比較して，採点精度が向上した．モデルの最適化には Adam を使用した．バッチサイズは 16，エポック数は 5 とした．
3. 連結した [CLS] トークンのベクトルを分類器に入力し，予測スコア s を出力する．

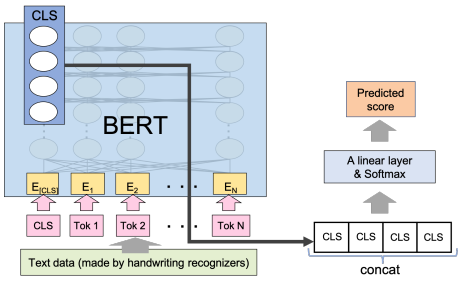

図 4　自動採点モデル

4　実験

4.1　解答データ

以上の条件や分類方法に基づいて 2017 年と 2018 年にそれぞれ 3 問ずつを含む 6 問の自動採点を行った．答案数は，2017 年，2018 年ともに約 6 万件であり，表 1 に，各設問の採点の統計を示す．設問 ID，答案数，スコア幅，得点の平均（±標準偏差），許容文字数の順に示した．BERT に使用したデータを，3 : 1 : 1（＝ 60 % : 20 % : 20 %）に分けて，トレーニングセット，開発セット，評価セットとした．採点精度の評価は，Quadratic Weighted Kappa (QWK) [21] を用いて行った．

4.2　実験結果

今回の実験では答案の文字数が比較的多く，内容も平易ではない．このような場合，推定精度を保証するためにはどの程度のサンプルサイズが必要なのかを知る必要がある．そこでサンプルサイズの大きさを 50,000，10,000，5,000，1,000，500 と変え，QWK がどのように変化するかを観察した．約 60,000 個のフルサイズデータを含めた結果を表 2 に示す．太字は最良の値を示す．

表 1　各設問の統計量

設問 ID	答案数	スコア幅	mean	文字数
2017 #Q1	62,222	0 – 6	4.46 ± 1.67	– 50
2017 #Q2	61,777	0 – 2	1.51 ± 0.86	– 25
2017 #Q3	59,791	0 – 5	0.43 ± 1.10	80 – 120
2018 #Q1	67,332	0 – 3	2.51 ± 0.88	– 30
2018 #Q2	66,246	0 – 3	1.87 ± 1.14	– 40
2018 #Q3	58,159	0 – 3	0.76 ± 1.07	80 – 120

表 2　各設問の QWK

設問 ID	全数	50,000	10,000	5,000	1,000	500
2017 #Q1	0.978	**0.979**	0.967	0.946	0.883	0.679
2017 #Q2	**0.963**	0.949	0.934	0.922	0.818	0.884
2017 #Q3	**0.866**	0.836	0.705	0.680	0.473	0.276
2018 #Q1	**0.976**	0.968	0.974	0.914	0.863	0.820
2018 #Q2	**0.954**	0.945	0.923	0.903	0.796	0.724
2018 #Q3	**0.944**	0.929	0.916	0.894	0.783	0.753

1. 設問の種類に関わらず，6 つの設問すべてで精度が高く保たれている．精度が一番低い 2017 年の Q3 でも，QWK は 0.86 となった．
2. 基本的には，サンプルサイズが大きいほど精度は高くなる．つまり，精度が収束しないという予想外の結果になった．60,000 というサンプルサイズは，一般的なテストでは十分な大きさだと考えられるが，予測の精度を高めるためにはより多くの答案数が必要であることを示している．
3. 問題が簡単なほど得点率が高く，推定精度も高い．2017 年，2018 年ともに Q1 が最も簡単で Q3 が最も難しい．Q1 の推定精度は Q3 の推定精度よりも高い．この傾向は得点のカテゴリ数には依存しない．

5　追加実験

本研究では，採点精度への影響を 2 つの観点から検討した．1 つ目の観点は，手書き文字認識モデルによる影響である．認識モデルの変化が全体の採点精度にどのように影響するかを調べた．2 つ目の視点は，言語処理モデルによる影響である．BERT の 12 層から抽出する情報の位置を変更し，全体の採点精度にどのような影響を与えるかを検証した．

5.1 手書き文字認識モデルの効果検証

採点精度に与える影響を調べるために，4 つの文字認識モデルをアンサンブル学習したモデルと他の手法とを比較した．手法は以下の通りである．

1. No LM：N-gram 言語モデルによる誤認識の補正を行わない文字認識モデル
2. VGG のみ：アンサンブル学習を行わない単一の文字認識モデル
3. DenseNet のみ：アンサンブル学習を行わない単一の文字認識モデル
4. Esm5: 5 つの文字認識モデルをアンサンブル学習した文字認識モデル

表 3 異なる文字認識モデルを用いた際の QWK の比較

設問 ID	手書き文字認識モデル				
	Original	No LM	VGG	DenseNet	Esm5
2017 #Q1	0.978	0.975	0.977	0.974	**0.980**
2017 #Q2	**0.963**	0.957	0.957	0.952	0.959
2017 #Q3	**0.866**	0.847	0.844	0.820	0.830
2018 #Q1	**0.976**	0.973	0.972	0.970	0.970
2018 #Q2	**0.954**	0.950	0.952	0.953	0.953
2018 #Q3	**0.944**	0.937	0.933	0.935	0.941

表 3 は，それぞれの出力結果を用いて自動採点した結果の QWK を比較したものである．その結果，複数の文字認識モデルを用いてアンサンブル学習を行ったものは，単一の文字認識モデルを用いたものよりも総合的な精度が高いことが明らかとなった．また，言語モデルによる修正を加えたモデルの方が，加えないモデルよりも精度が高いという結果が出た．加えて，アンサンブル学習のモデルの数を増やしても精度に大きな変化はなかった．これらの結果から，全体的な精度は言語モデルの変更と文字認識モデルの品質の両方に影響されることが判明した．同時に，文字認識モデルの品質向上による総合精度の向上には限界があることも判明した．

5.2 BERT から取得した情報の効果検証

BERT から取得した言語情報が採点精度に及ぼす影響を調査した． BERT は全 12 層で構成されており，各層が異なる情報を保有することが知られている [22]．具体的には入力部に近い層は形態素情報，中間部の層は構文情報， 出力部に近い層は意味情報に焦点を当てた情報を保有している．本研究では BERT を入力部に近い層，中間部の層，出力部に近い層の 3 つに分け，それぞれの層での採点精度の違いを調べた．入力部から 1〜4 層，5〜8 層，9〜12 層のベクトルを抽出する．各精度の結果を表 4 に示した．各問題とも 9〜12 層の情報を抽出したときに，最も高い採点精度が得られることがわかった．

表 4 情報抽出した層の部分による QWK の比較

設問 ID	BERT の中間層		
	1-4	5-8	9-12
2017 #Q1	0.977	0.977	**0.978**
2017 #Q2	0.952	0.955	**0.963**
2017 #Q3	0.830	0.832	**0.866**
2018 #Q1	0.969	0.972	**0.976**
2018 #Q2	0.951	0.950	**0.954**
2018 #Q3	0.936	0.939	**0.944**

この結果から，システムが自動採点作業を行う際に意味的な情報を重視していることが推察される．特に，2017#Q3 の QWK は，全設問の中で 3.0 以上の差があり，推定精度の低下が際立っている．

6 おわりに

本研究では手書き文字認識モデルを用いた短答式記述問題の完全自動採点手法を検討し，大規模な全国テストでその性能を評価した．「完全」とは，人手による採点データへの注釈付けや手書き文字の変換が不要であることを意味している．大学入学共通テストの 2 回の試行調査で実施された前例のない大量のデータを使用し，事前学習済みの BERT を用いて採点を行った．その結果，以下のことが分かった．

1. データが十分に大きい場合，本手法はアノテーション作業や手書き答案のテキスト変換を行うことなく高精度で自動採点が可能である．
2. 25〜120 字の答案では，50,000 件のデータサイズでも学習が収束しないことが多い．
3. 手書き文字認識モデルに起因する誤差があっても，現在の技術であればある程度の採点精度は保証される．

本論文では，現在の技術水準において人手を介さない状態での実際の精度を報告した．扱った問題は，答案の文字数や難易度など様々であったが，いずれの場合も高い精度で得点を予測することができた．このことから，本研究の手法はこの範囲のすべての短答式記述問題に対して有効であり，現在の技術で短答式記述答案の自動採点を十分に行えることが示唆された．また，本研究では，手書き文字認識モデルを短答式記述答案の自動採点で利用する手法の有用性を示すことができた．手書き文字を使うことが多い教育現場での応用に向けて，新たな学習方法を設定するきっかけになると考えられる．

謝辞

本研究は JSPS 科研費 JP20H04300, JST A-STEP JPMJTM20ML の助成を受けたものです.

参考文献

[1] 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, Berlin, Germany, August 2016. Association for Computational Linguistics.

[2] 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, Austin, Texas, November 2016. Association for Computational Linguistics.

[3] 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.

[4] 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.

[5] 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.

[6] Chul Sung, Tejas Indulal Dhamecha, and Nirmal Mukhi. Improving short answer grading using transformer-based pre-training. In **Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)**, Vol. 11625 LNAI, pp. 469–481. Springer Verlag, jun 2019.

[7] Leon Camus and Anna Filighera. Investigating transformers for automatic short answer grading. In **Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)**, Vol. 12164 LNAI, pp. 43–48. Springer, 2020.

[8] Jiaqi Lun, Jia Zhu, Yong Tang, and Min Yang. Multiple data augmentation strategies for improving performance on automatic short answer scoring. In **Proceedings of the AAAI Conference on Artificial Intelligence**, Vol. 34, pp. 13389–13396, Apr. 2020.

[9] Hiroaki Funayama, Shota Sasaki, Yuichiroh Matsubayashi, Tomoya Mizumoto, Jun Suzuki, Masato Mita, and Kentaro Inui. Preventing critical scoring errors in short answer scoring with confidence estimation. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop**, pp. 237–243, Online, July 2020. Association for Computational Linguistics.

[10] Brian Riordan, Sarah Bichler, Allison Bradford, Jennifer King Chen, Korah Wiley, Libby Gerard, and Marcia C. Linn. An empirical investigation of neural methods for content scoring of science explanations. In **Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications**, pp. 135–144, Seattle, WA, USA Online, July 2020. Association for Computational Linguistics.

[11] Masaki Uto and Yuto Uchida. Automated short-answer grading using deep neural networks and item response theory. In **Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)**, Vol. 12164 LNAI, pp. 334–339. Springer, jul 2020.

[12] Masaki Uto and Masashi Okano. Robust neural automated essay scoring using item response theory. In **Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)**, Vol. 12163 LNAI, pp. 549–561. Springer, jul 2020.

[13] Zhaohui Li, Yajur Tomar, and Rebecca J. Passonneau. A semantic feature-wise transformation relation network for automatic short answer grading. In **Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing**, pp. 6030–6040, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[14] 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, 2019.

[15] Taiichi Saito, Hakuzo Yamada, and Kazuhiko Yamamoto. On the database ETL9 of handprinted characters in jis chinese characters and its analysis. **Trans IECE Jpn**, Vol. J68-D, No. 4, pp. 757–764, 1985.

[16] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. **arXiv preprint arXiv:1409.1556**, 2014.

[17] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. **arXiv preprint arXiv:1704.04861**, 2017.

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In **Proceedings of the IEEE conference on computer vision and pattern recognition**, pp. 770–778, 2016.

[19] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In **Proceedings of the IEEE conference on computer vision and pattern recognition**, pp. 1492–1500, 2017.

[20] 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.

[21] Jacob Cohen. Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. **Psychological bulletin**, Vol. 7, No. 4, pp. 213–220, 1968.

[22] Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. What does BERT learn about the structure of language? In **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics**, pp. 3651–3657, Florence, Italy, July 2019. Association for Computational Linguistics.