採点項目に基づく国語記述式答案の自動採点

水本智也† 磯部順子† 関根聡† 乾健太郎‡†
† 理化学研究所 AIP センター ‡ 東北大学
{tomoya.mizumoto, yoriko.isobe satoshi.sekine}@riken.jp
inui@ecei.tohoku.ac.jp

1 はじめに

記述式の問題は論理的思考力・判断力・表現力を養うために効果的とされている．このような力を養うために多くの記述式問題を解き，添削してもらいフィードバックを得る必要がある．しかし，添削が可能な教師の数は限られており，学習者が記述式問題学習のためのフィードバックを得る機会は多くない．そのため学習者が記述式問題を学習する際，なんらかの補助が必要となる．

記述式問題の学習を補助する方法の一つとして，自動採点がある．自動採点では，文／文章が入力として与えられた時に，それらの良さを点数として出力する．この自動採点システムを使うことで，学習者は自分が書いた解答がどの程度の到達度かを知ることができる．近年では，機械学習／深層学習を使った手法が提案されており，自動採点の性能も向上している [9, 8, 12]．また，[4] のように採点する側を助けるためのシステムも開発されている．一方で，記述式問題の学習補助という点から見ると，現状の自動採点のように点数を出力するだけでは不十分である．学習補助のためには，「なぜその点数なのか」や「どう改善すれば点数を上げることができるか」といった根拠・理由などのコメントが必要である．

そこで，国語読解問題の記述式答案の自動採点を対象に，その第一ステップとなる研究を行う．国語の読解記述式問題の多くは，項目が細かく分かれておりそれに対応する内容が書かれているかで点数を付ける形式である（図 1，詳しい説明は次節）．採点項目ごとに点数を出力できるだけで，全体の点数しかわからない場合よりも情報量が増えるため，学生自身で点数を見てどの部分が不十分であるかを知ることができる．言い換えると，複数の項目のうち点数の低いところが不十分であるということを学習者自身で理解することができる．

採点項目ごとのスコアを出力する問題を，深層学習を使った自動採点のモデル [10] を拡張することで実現する．通常のモデルでは全体の点数を 1 つ出力するだけのところを，提案モデルは同じネットワークをシェアして項目ごとに点数を予測するモデルに拡張する．提案モデルでは項目ごとにアテンション機構を用意し，注目するべきところを学習させることで項目ごとの点数予測を実現する．項目ごとに点数を予測するモデル構築のために，項目ごとの点数の付いたデータを作成するが，大量の解答に項目ごとの点数を付けることは労力を要する．そこで，少量の項目に点数が付いたデータと大量にある全体の点数が付いたデータの両方を使うモデルの構築も行う．実験の結果，全体の点数が付いたデータを使うことで，項目ごとの性能が向上することを確かめた．また，本稿で提案する深層学習ベースのモデルではアテンション機構を使っているため，どの箇所に注目してシステムが点数を付けたかを知ることができる．アテンション機構の重みを分析することで，記述式問題の学習支援の助けになるかを調査する．

表 1 代ゼミデータの詳細

	論説 1	論説 2	論説 3	随筆 1	随筆 2	小説
解答数	5,181	5,200	4,849	5,182	4,733	5,273
文字数	70	70	70	50	60	60
配点	16	15	15	12	14	12
平均点	6.68	5.33	4.49	4.23	5.27	5.39
(SD)	3.68	2.84	2.55	1.84	3.09	2.06
項目数	4	3	3	4	3	4

2 国語長文読解の記述式問題データ

データの説明をする前に本研究で扱う問題について整理する．本研究で扱うタスクは，Short Answer Scoring (SAS) であり，長文記述を採点する Essay Scroring [3, 1] とは異なる．SAS の中には，コンピュータの知識を問うようなもの [7] で採点項目が 1 つのものもあるが，本研究では採点基準が細かく分類されている，つまり採点項目が複数あるものを対象とする．また，全体の点数がその合計で計算される問題を扱うこととする．

本研究では，代々木ゼミナールの国語長文読解の記述式問題 6 題のデータ (以下，代ゼミデータ) を用いる．このデータは学生の書いた答案，採点者によって付けられた点数からなる．表 1 に本データの詳細を示す．6 題は全て現代文の問題であり、論説問題が 3 題，随筆問題が 2 題，小説問題が 1 題である．解答数は採点の必要のない白紙のものは除いている．解答数はおよそ 5,000，平

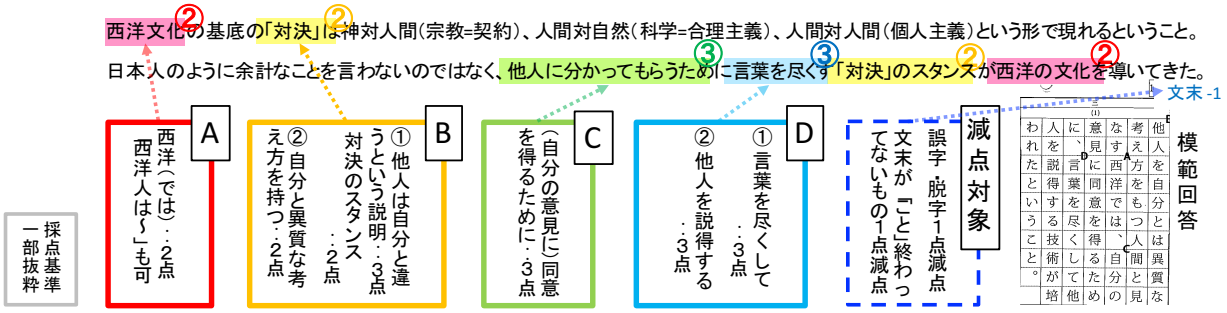

図 1　採点項目ごとのアノテーション．この例では 4 つの項目に分かれている．丸付き数字が項目点，マーカーの箇所が項目点対象箇所である．どの文も「西洋では」と同義の表現が書かれているため項目点 A が与えられる．

均点は 4.23～6.68 点，標準偏差は大きいもので 3.68 であり，問題によっては点数のばらつきが大きい．採点項目数はどの問題も 3 つもしくは 4 つからなる．

採点項目は，国語記述式の問題で採点を行うときにベースとなるものである．国語記述式の問題の採点の多くは，採点項目ごとに満たすべき内容が書かれているかで点数を付け，その合計を設問の点数とする（以下，全体点）．[*1]図 1 の例の 1 文目は採点項目の A と B で点数が付き，赤のマーカで塗られている箇所が項目 A の点数を与える部分になる．本稿では，この点数を項目点，また項目点の付く箇所を項目点対象箇所と呼ぶ．

代ゼミデータには，全体点は付与されているが，項目点は付与されていない．そこで，採点基準を参考に図 1 のように項目点と項目点対象箇所をアノテーションした．アノテーション作業は著者の 2 人でそれぞれ 2 問，4 問に対して行った．論説 1 に対しては 200 解答，それ以外には 100 解答アノテーションした．新しく付与した項目点の合計と元の点数を比較したところ，一部異なる点数になった．この原因については，4.3 節で議論する．項目点と項目点対象箇所を付けたデータを項目点付データ，全体点のみ付いたデータを全体点データと呼ぶ．

3　採点項目を考慮した自動採点モデル

本研究では，文献 [9] で提案された深層学習のモデルを拡張することで採点項目ごとに点数を出力するモデルを構築する．図 2 に提案モデルを示す．オレンジ色の破線で囲んである部分が，項目点を予測する部分である．提案モデルでは，一つのネットワークから項目の数だけ点数を出力する．解答中の単語を入力として，それらを分散表現 (ベクトル) に変換する．このベクトルを LSTM に渡す．各項目で重要箇所に注目させるために項目ごとにアテンション機構を用意し，LSTM の出力に対して適用する．最後に sigmoid 関数を使って項目ごとにスコアを出力する．このモデルを学習する際，項目ごとに正解点を教師として与えることで学習する．

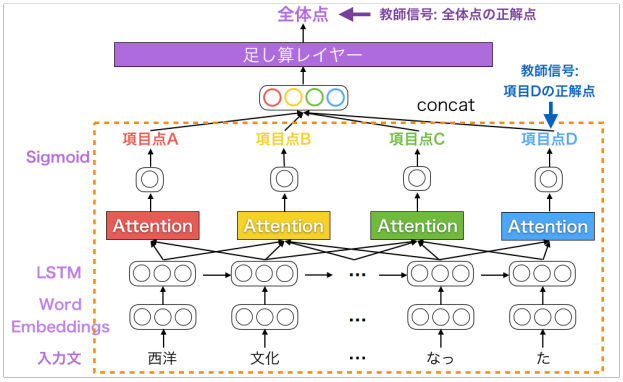

図 2　提案モデルのイメージ図．採点項目ごとに点数を予測し，その足し算によって全体のスコアを予測．

項目点付データは全体点データの 50 分の 1 ほどしかない．そのため，項目点を出力するモデルを拡張して，全体点データを使えるようにする．本研究では，上述したモデルで出力した項目点を足し算する層を用意し，全体点を予測できるモデルを構築した（図 2 の破線部より上）．足し算層の入力は要素数が項目数のベクトルで，出力は全体点が 1 つである．学習の際は，全体点の正解を教師信号として与える．このモデルの拡張により，全体点データを使って，項目点を予測するモデルの性能向上も可能になる．最初の項目点を予測するモデルを項目予測モデル，項目点の足し算によって全体点を予測するモデルを項目 + 全体予測モデルと呼ぶ．

4　実験

提案した採点項目ごとに点数を予測するモデルの性能を確かめるために実験を行う．

4.1　実験設定

2 節で説明した代ゼミデータを実験に使用する．各問題，全体点データから開発と評価に各 500 解答を使用し，残りを学習に使った．ベースラインとして [10] の

[*1] 誤字や文末表現で減点される場合もある．

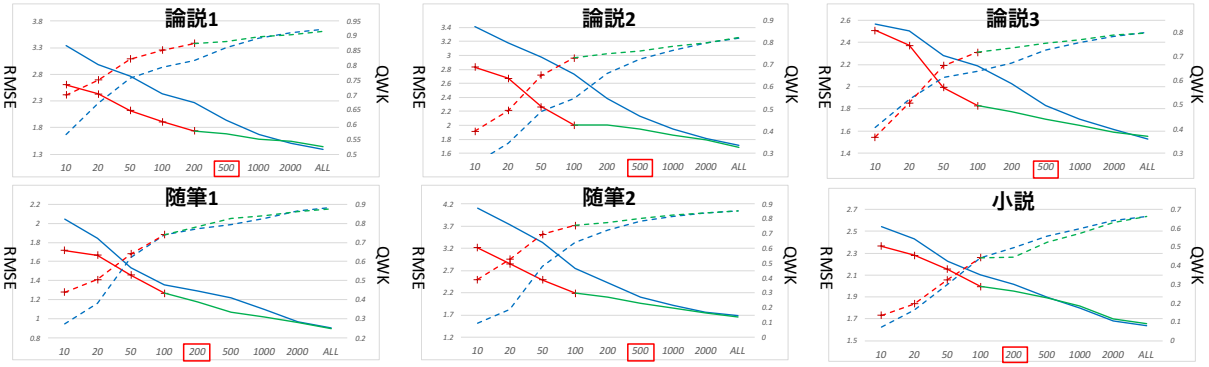

図 3　データサイズを変化させた場合の性能．実線が RMSE，破線が QWK に対応する．また，青がベースライン，赤が項目予測モデル，緑が項目 + 全体予測モデルの結果に対応．

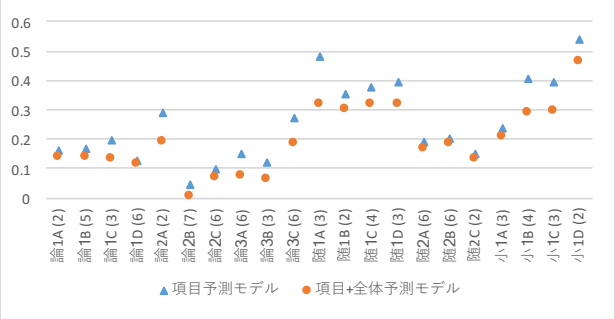

図 4　採点項目ごとの RMSE の比較．各項目に対して配点で正規化を行った．括弧内は配点を示す．

uni-directional LSTM のモデルと比較する．[*2]ニューラルネットモデルの設定は [10] を元にした．ただし，Word Embeddings は 100 次元とし，Wikipedia のダンプデータで word2vec によって学習したものを使用した．単語分割には MeCab を使用した．また，項目予測モデルのエポック数は 250 とした．項目 + 全体予測モデルは，項目予測モデルの pre-train した重みを使用し，項目点予測の性能を下げないため，項目点の予測と全体点の予測を交互に行うことでモデルの学習を行った．評価尺度として，Root Mean Squared Error (RMSE)、Quadratic Weighted Kappa (QWK) を用いた．報告する値は，全て 10 回実験を回した平均の値であり，開発セットで QWK がもっとも高かった時のエポック数のモデルで評価した時の値である．

4.2　実験

提案モデルの全体点予測性能の調査のために，ベースラインと提案モデルで，データサイズを変えた場合の性能を比較する．図 3 に結果を示す．どの問題に対しても，項目予測モデル（赤線）はベースライン（青線）より高い性能である．また，データを増やした際の性能向上の傾きもベースラインと同等である．その後，全体点データを使った場合も性能向上は続き，さらにデータを増やすことで性能向上も見込むことができる．

学習データ数に注目し RMSE の値を見ると，項目点を 100 解答付けた場合と全体点データ 500 解答の場合で同等の値である（随筆 1 と小説は 200 解答と同等）．本研究の直接の目的ではないが，実際に自動採点の現場で運用するような場合は，全体点のみ付けるか少し労力をかけて項目点まで付けるかのバランスが重要になる．

次に項目予測モデルと項目 + 全体予測モデルを比較して，全体点データを学習に使った場合でも，項目点予測の性能は向上するのかを確かめる．実験 1 で使った評価データには項目点が付いていないため，項目点付データを 5 分割交差検証によって評価する．図 4 は項目ごとに配点数で正規化を行った RMSE の結果である．項目によって改善している幅に差はあるが，全ての問題・項目に対して，項目 + 全体予測モデルによって項目点が改善している．また，図 4 を見ることで，システムにとって自動採点が難しい項目も明らかになる．小説の項目 D は正規化後の RMSE が約 0.5 であり，自動採点が難しい項目であることがわかる．

4.3　分析と考察

1 節で述べたようにアテンションの重みを見ることで，項目予測モデルがどの箇所に注目して点数を付けたかを知ることができる．アテンションの重みの分析は機械翻訳 [2] や構文解析 [11] でも行われている．本稿ではアテンションの重みと人が点数を付ける時に参考にした箇所を比較することで，項目予測モデルが点数の付く箇所を同定できているかを確かめるとともに，今後の応用可能性を探る．

図 5 にアテンションが人のアノテーションと一致している例を示す．項目予測モデルは順方向の LSTM を使っているため，重みの付く箇所が 1 つずれていると解釈すると，論説 1 の例は点数予測もほぼ正解と一致していた．結果を見ると，人が A を付けた箇所「西洋人」と項目予測モデルのアテンションで注目している箇所は同じであることがわかる．他の B，C，D に対しても同様に

[*2] bidirectional LSTM やアテンションを使ったものより，開発セットでの性能が良かったためこのモデルを使用した．

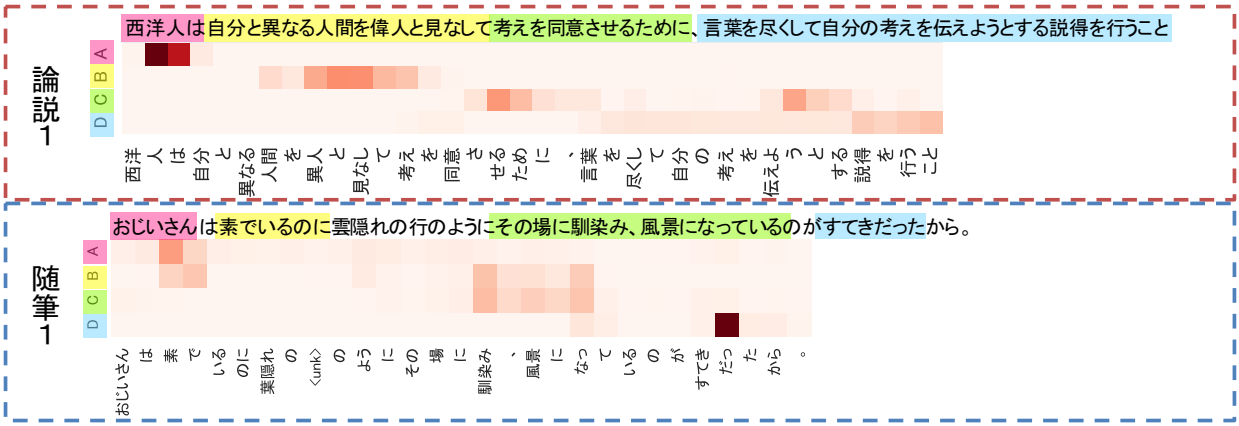

図 5　アテンションの可視化結果．上の例が人手による得点箇所へのアノテーション．下の A，B，C，D はそれぞれ人手の赤，黄色，緑，青に対応．

上手くいっていることがわかる．B の配点が 5 点の時，システムの予測が 3 点であれば，学習者はここの表現を改善する必要があるということを知ることができる．

随筆 1 に対するアテンションも項目 B 以外は対応している．一方，点数予測の結果は人手の点数とずれていた．この結果は，アテンションができても点数予測は正確にできるかわからないことを示唆する．項目 B は人手の結果と異なり，「風景になっている」にも注目している．このような例には，教師ありアテンション [6, 5] を使って，注目する箇所を直接教える改善方法がある．

採点項目で注目するべきところがない場合，すなわちいずれかの項目の点数が 0 点の場合について分析する．図 6 に随筆 1 において，項目 A のみ点数が付いた例を示す．この例では，解答文全体で項目 A の点が 3 点を与えられており，アテンションの結果も上手くいっている．一方，点数が与えられない項目 B，C，D のアテンションの結果を見るとどれも後半に重みが付いており，これを見るだけでは学習者の役には立たない．しかし，システムはこれらの項目に対して 0 点を与えているため，このアテンションに意味はないことは分かる．

2 節で述べた「アノテーション結果と元から付いていた点との差異」について議論する．項目点の合計で全体点を計算できる問題として扱ったが，実際の採点はもう少し複雑である．項目間での繋がりが重要であり，例えば項目 A と B の関係を考慮しなければならない場合がある．今後はこのような複雑な条件も扱う必要がある．

5　おわりに

本研究では，学習者の記述式問題の学習支援を目的とした新しい自動採点に取り組んだ．従来の自動採点では，点数を 1 つ出力するだけであったが，項目点を予測することで，学習者に対するフィードバックを可能にした．また，提案モデルのアテンションを分析すること

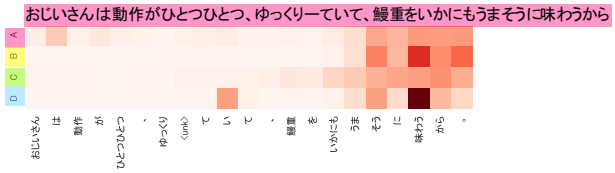

図 6　採点項目を満たす要素が解答文中にない場合．この例では項目 A のみを満たす．

で学習支援に活用できるかを調査した．現状では課題も残るが，活用できる可能性も示唆された．代ゼミデータは，将来的に答案・その採点結果ともに一定の条件の元で、研究用途に配布することを検討している．

謝辞　答案と採点データの入手にご助力いただいた独立行政法人大学入試センターの大久保智哉氏，実際の模試データを提供いただいた学校法人高宮学園代々木ゼミナールに感謝します．

参考文献

[1] Dimitrios Alikaniotis, Helen Yannakoudakis, and Marek Rei. Automatic Text Scoring Using Neural Networks. In *Proc. of ACL*, pp. 715–725, 2016.
[2] Hamidreza Ghader and Christof Monz. What does Attention in Neural Machine Translation Pay Attention to? In *Proc. of IJCNLP*, pp. 30–39, 2017.
[3] Tsunenori Ishioka and Masayuki Kameda. Automated Japanese Essay Scoring System based on ArticlesWritten by Experts. In *Proc. of ACL*, pp. 233–240, 2006.
[4] Tsunenori Ishioka and Masayuki Kameda. Overwritable Automated Japanese Short-answer Scoring and Support System. In *Proc. of WI*, pp. 50–56, 2017.
[5] Hidetaka Kamigaito, Katsuhiko Hayashi, Tsutomu Hirao, Hiroya Takamura, Manabu Okumura, and Masaaki Nagata. Supervised Attention for Sequence-to-Sequence Constituency Parsing. In *Proc. of IJCNLP*, pp. 7–12, 2017.
[6] Lemao Liu, Masao Utiyama, Andrew Finch, and Eiichiro Sumita. Neural Machine Translation with Supervised Attention. In *Proc. of COLING*, pp. 3093–3102, 2016.
[7] Michael Mohler, Razvan Bunescu, and Rada Mihalcea. Learning to Grade Short Answer Questions using Semantic Similarity Measures and Dependency Graph Alignments. In *Proc. of ACL*, pp. 752–762, 2011.
[8] Brian Riordan, Andrea Horbach, Aoife Cahill, Torsten Zesch, and Chong Min Lee. Investigating neural architectures for short answer scoring. In *Proc. of BEA*, pp. 159–168, 2017.
[9] Md Arafat Sultan, Cristobal Salazar, and Tamara Sumner. Fast and Easy Short Answer Grading with High Accuracy. In *Proc. of NAACL*, pp. 1070–1075, 2016.
[10] Kaveh Taghipour and Hwee Tou Ng. A Neural Approach to Automated Essay Scoring. In *Proc. of EMNLP*, 2016.
[11] Oriol Vinyals, Lukasz Kaiser, Terry Koo, Slav Petrov, Ilya Sutskever, and Geoffrey Hinton. Grammar As a Foreign Language. In *Proc. of NIPS15*, pp. 2773–2781, 2015.
[12] Siyuan Zhao, Yaqiong Zhang, Xiaolu Xiong, Anthony Botelho, and Neil T. Heffernan. A Memory-Augmented Neural Model for Automated Grading. In *Proc. of the Fourth ACM Conf. on Learning @ Scale*, pp. 189–192, 2017.