E-010

Bidirectional LSTM を用いた誤字脱字検出による採点項目別自動採点の改善
Improving automatic scoring according to grading item by detecting misspellings and omissions using Bidirectional LSTM

倉田 基成† 作田 航平† 横尾 拓未† 森 康久仁‡ 須鎗 弘樹‡
Motonari Kurata Kohei Sakuta Takumi Yokoo Yasukuni Mori Hiroki Suyari

1. はじめに

記述式問題は採点におけるコストの高さと採点の難しさから，自動採点システムによる採点支援が求められている．機械学習による記述式問題自動採点の先行研究として，水本らは採点項目ごとに点数を予測するニューラルネットワークモデル(以降，従来モデルと呼ぶ)を提案している[1]．しかし，水本らが提案した手法を元に実際に自動採点を行った結果，誤字脱字等の誤りに対する減点項目(以降 Miss Score と呼ぶ)の採点精度が，他の採点項目の採点精度に比べ低いということが分かった．したがって，誤字脱字等の誤りに対する減点項目の採点精度を改善することができれば，自動採点システム全体の精度向上が見込まれる．

誤字脱字の検出方法として，高橋らは Bidirectional LSTM(BiLSTM) を用いた誤字脱字検出システムを提案している[2]．この誤字脱字検出システムの中で使用されている言語モデルは，文章の文脈から単語の誤りを予測するため，多様な誤りのパターンに対応することができる．したがって，本研究は Miss Score を言語モデル，その他の項目点を従来モデルで採点することにより，自動採点モデルの採点精度を改善する手法を提案する．

2. 記述式問題の自動採点

本節では我々が取り扱う記述式問題と，水本らが提案した自動採点モデルによる採点精度について説明する．本研究にて取り扱う記述式問題は，採点項目が複数あり，採点項目ごとに点数を付与する問題を対象とする．解答とその採点例を図 1 に示す．この例では採点項目が A-D，減点 A，減点 B の 6 つあり，それぞれ定められた内容を満たしているかで点数が付与されている．このうち，減点 A が Miss Score にあたる．

水本らが提案したニューラルネットワークモデルは，採点項目ごとに Attention 機構を用意することで，採点項目ごとに重要な箇所を学習し，採点が行える．加えて，採点の根拠となる箇所の情報や全体点を教師データとして学習，予測を行うことができる．また， Attention 機構による解答の重みづけを分析することにより，モデルが項目ごとに重要であると予測した箇所を明示できるため，学習支援として活用できることも示唆されている．

従来モデルの採点精度を詳しく知るため，水本らの提案手法を参考にモデルを構築し，自動採点を行う．採点する記述式問題として，理化学研究所が公開している理研記述式問題採点データセット[3]から，国語記述式問題 6 題の解答データを用いた．6 題の解答データから，学習データ，

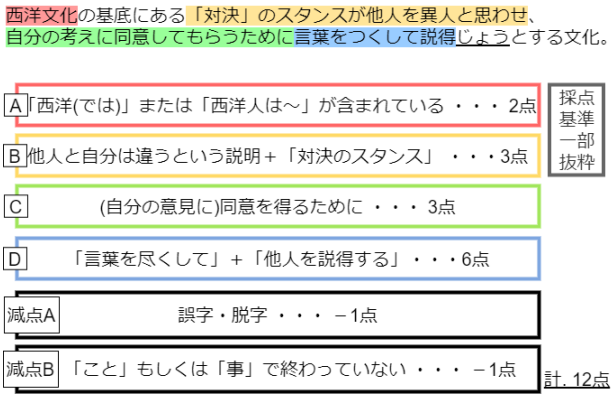

図 1 記述式問題の解答例

表 1 従来モデルの実験結果

	Q1	Q2	Q3	Q4	Q5	Q6
全項目平均	0.768	0.744	0.656	0.498	0.634	0.589
Miss Score	0.047	0.195	0.198	0.292	0.098	0.043

検証データ，評価データをそれぞれ 100 解答， 250 解答，250 解答ずつ抽出し，項目点を教師データとして学習，評価を行った．評価尺度として，Quadratic Weighted Kappa (QWK) を用いた．表 1 は 各問題における従来モデルの採点結果を示す．結果はすべて 20 回試行した平均を示す．

従来モデルによる自動採点の結果，全採点項目の項目点の平均 QWK と比較して，Miss Score の QWK が低いことが分かった．原因として，従来モデルでは採点項目の内容に関わらず，各項目点の学習，予測に同一のネットワークを使用している．しかし，Miss Score に該当する誤字脱字などの文章の誤りは多様であり，誤り自体を Attention 機構によって学習することは難しい．したがって，Miss Score の採点に対応するネットワークを，誤字脱字等の誤りを検出する手法に変更することで，採点精度が向上すると考えられる．

3. 提案手法

3.1 提案モデルの概要

提案モデルは Attention 機構を用いて解答と項目点を学習，予測する項目モデルと，誤りのない解答文の文脈を学習し，誤りである単語を予測することで，Miss Score の採点を行う言語モデルの 2 つのモデルに大きく分けられる．図 2 に提案モデルの概念図を示す．項目モデル，言語モデル共通の処理として，解答文中の各単語は word2vec，BiLSTM を通すことにより，単語の意味や語順を考慮した分散表現(ベクトル)に変換される．その後，項目モデルの Attention 機構や，言語モデルの全結合層へと入力される．

† 千葉大学大学院融合理工学府 Graduate School of Science and Engineering, Chiba University
‡ 千葉大学大学院工学研究院 Graduate School of Engineering, Chiba University

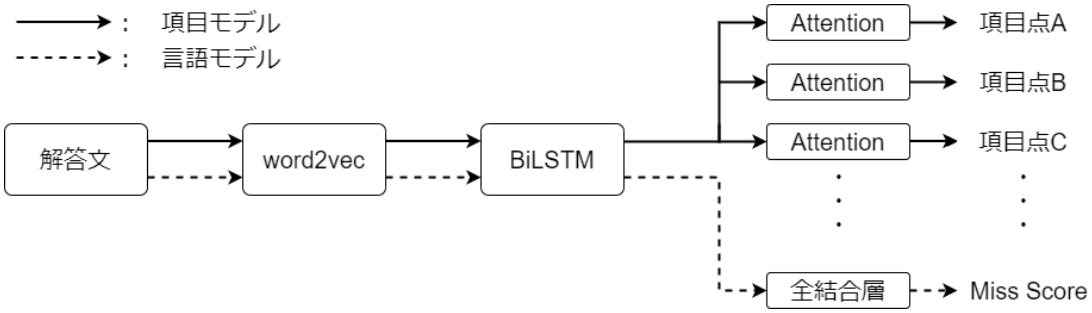

図 2 提案モデルの概念図

3.2 項目モデル

図 3 に項目モデルを示す．項目モデルと言語モデルでは BiLSTM の出力に対する扱いが異なる．したがって，項目モデルと言語モデルの共通部分である，解答文と word2vec, BiLSTM も図 3 に示している．BiLSTM が出力した単語ごとのベクトルに対して，採点項目ごとに Attention 機構が重要な箇所に重みづけを行う．出力された隠れ層を全結合層，Sigmoid 関数に通すことで点数を出力する．Sigmoid 関数を用いていることから，出力される点数は 0 から 1 の範囲になる．そのため，学習時は項目ごとに実際の点数を配点で割ることで， 0 から 1 の範囲に正規化を行い，教師データとしている．損失関数には Mean Squared Error(MSE)を使用し，各採点項目の誤差の合計を 1 文の誤差として，最小化するように学習を行う．評価時にはモデルの出力に項目ごとの配点を掛け，四捨五入をした値を用いて評価する．

3.3 言語モデル

図 4 に言語モデルを示す．図 3 と同様に，項目モデルと言語モデルの共通部分である，解答文と word2vec, BiLSTM も図 4 に示している．このモデルでは直接点数を出力するのではなく，解答文中の各単語が正常であるかどうかを確率として出力，それをもとに誤りである単語や Miss Score を予測する．

学習時には誤りのない解答文の語順を学習する．解答文中の t 番目の単語を予測する場合，順方向 LSTM の t-1 番目の単語における出力と，逆方向 LSTM の t+1 番目の単語における出力を結合し，全結合層，Sigmoid 関数に与える．これにより，評価時に t 番目の単語が学習データにない誤りのパターンであっても，予測を行うことができる．最終的な出力は，使用データ 1 問に含まれる単語の語彙数分の次元を出力する．言語モデルの学習データには誤字脱字などの誤りを含まない解答文のみを使用する．損失関数には Cross Entropy を使用し，各単語の誤差を足し合わせた値を 1 文の誤差として，最小化するように学習を行う．評価時は，最終的な出力から解答を構成している単語に対する確率を取得し，この取得した結果を利用して Miss Score の採点を行う．

3.4 学習方法

項目モデルと言語モデルは学習に用いるデータ， 出力の次元数，損失関数が異なるため，項目モデルと言語モデルを 1epoch ごとに交互で予測を行うことで提案モデルの学習を行う．

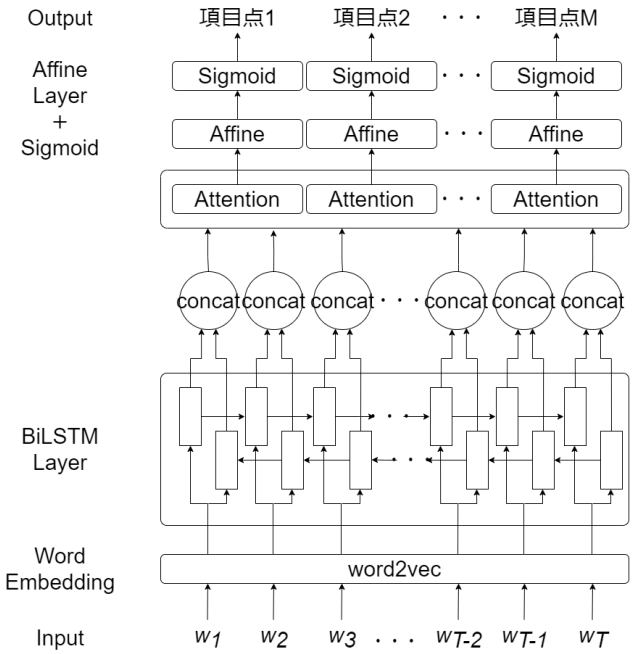

図 3 項目モデルのモデル図

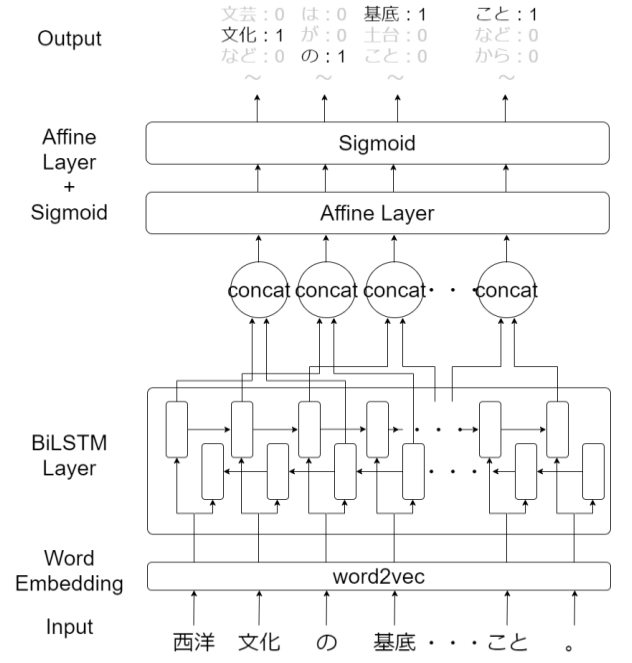

図 4 言語モデルのモデル図

4. 実験

4.1 実験データ

実験には理研記述問題データセット[3]に含まれる，代々木ゼミナールの国語記述式問題 6 題のデータを使用する．このデータは高校生を対象とした模擬試験における学生の解答と，解答に対して採点者が付与した全体点，採点項目ごとの項目点が含まれている．また，各解答はあらかじめ MeCab により単語ごとに分かち書きされている．表 2 にデータの詳細を示す．解答数は各題 2100 解答ずつある．各問題には加点項目と減点項目がそれぞれ決められており，合わせて 4 つから 6 つの採点項目がある．各採点項目は採点基準によって定められており，加点項目の多くは解答文中に満たすべき内容がどの程度含まれているかで点数がつけられる．Miss Score は減点項目の 1 つであり，すべての問題の採点項目に含まれている．解答文中に誤字脱字や文法的ミスなどが含まれているかで点数がつけられており，誤字脱字や文法的ミスなどの誤りが含まれていない場合は 0 点，含まれている場合は -1 点となる．

表 2 実験データの詳細

	Q1	Q2	Q3	Q4	Q5	Q6
形式	論説	随筆	論説	小説	論説	随筆
文字数	70	50	70	60	70	60
配点	16	12	15	12	15	14
平均点	6.79	4.03	5.46	5.27	4.63	5.51
標準偏差	3.47	1.84	2.67	2.06	2.62	3.14
項目数	6	6	5	5	4	4

4.2 実験手法

実際の現場において，採点項目ごとに採点を行うことはコストが高く，大量の項目点付きデータを用意することは難しい．そのため，本研究では用意できる学習データが少ないことを想定し，各問題から学習データをそれぞれ 25 解答，50 解答，100 解答，200 解答として実験を行う．検証データと評価データは各問題それぞれ 250 解答，250 解答ずつ使用する．ベースラインとして水本らが提案した従来モデルを使用し，項目点を教師データとして与えた場合の結果と比較する．word2vec は Wikipedia のダンプデータにより事前学習を行ったものを使用する．また，エポック数は 100 とした．今回用いる実験データでは，Miss Score が 0 点または-1 点の 2 値で点数が付けられている．したがって，言語モデルによる Miss Score の採点では，解答を構成している全単語の確率最低値と閾値を比較し，閾値以上であれば 0 点，閾値未満であれば-1 点と採点する．この閾値は，検証データでの Miss Score の QWK が最も高くなる場合の閾値を用いる．結果は全て 20 回試行した平均の値であり，検証データで全採点項目の平均 QWK が最大となった epoch のモデルで評価した時の値である．

5. 結果と考察

提案モデルによる Miss Score の採点結果を表 3 に示す．従来モデルと提案モデルを比較すると，全学習データ数について Q1，Q2，Q5，Q6 の 4 題の記述式問題に対して QWK の値が向上した．これは学習データに含まれていない，もしくは含まれる回数が少ないために，従来モデルによる学習，検出ができなかった誤字脱字のパターンを，言語モデルが検出できたためと考えられる．

表 3 Miss Score の採点結果

	Q1	Q2	Q3	Q4	Q5	Q6
学習データ 25						
従来モデル	0.017	0.051	0.111	0.183	0.011	0.017
提案モデル	0.081	0.174	0.007	0.001	0.061	0.051
学習データ 50						
従来モデル	0.026	0.140	0.183	0.274	0.041	0.012
提案モデル	0.121	0.236	0.014	-0.007	0.120	0.076
学習データ 100						
従来モデル	0.047	0.195	0.198	0.292	0.098	0.043
提案モデル	0.168	0.296	0.044	0.003	0.157	0.106
学習データ 200						
従来モデル	0.089	0.339	0.262	0.386	0.185	0.102
提案モデル	0.189	0.350	0.112	0.031	0.189	0.173

解答例 1
試合に出れず，心だけがまだ終わっていない心情.

解答例 2
自分は試合に出れなかったので，高校野球は終わった気がせず，ケリのつけ方がわからなくて，おさまりがつかない複雑な心情.

図 5 従来モデルの Miss Score の採点例

一方で，Q3，Q4 では提案モデルの QWK よりも従来モデルの QWK の方が高い値になった．Q4 において Miss Score が-1 点の解答に対して，従来モデルは正しく採点をしたが，提案モデルは誤って採点をした例を図 5 に示す．傍線部は，従来モデルの Attention 機構が Miss Score の採点において高い重みづけをした表現を示す．“出れず”や“出れなかった”のような，ら抜き言葉と考えられる表現が含まれる解答に対して，従来モデルでは正しく採点したが，提案モデルでは誤って採点をした場合が見られた．この内，従来モデルが共通して注目している，“出れ”について，“出れ”が含まれる解答数の内訳を表 4 に示す．

表 4 “出れ”を含む Q4 の解答の内訳

Miss Score の点数	0 点	-1 点
解答数	20 解答	368 解答

Q4 には“出れず” や“出れなかった”に共通している“出れ”という表現が，Miss Score が-1 点の解答に多く見られた．したがって，従来モデルでは Attention 機構が

“出れ”という表現が重要であると学習し，採点を行ったため，精度が高くなったと考えられる．一方，提案モデルが正しく採点出来なかった原因として，ら抜き言葉が使われている場合でも各単語の使い方や意味は変わらないため，言語モデルがら抜き言葉を検出できず，誤って採点をしたと考えられる．このように，特定の表現が含まれる解答が誤りとなる傾向にある場合，誤りと考えられる表現を学習できる従来モデルの採点精度が高くなりやすい．しかし，その表現が誤りの理由であるかは不明であるため，より精度の高い採点を行うためには，Miss Score に関する採点基準の情報が必要だと考えられる．

言語モデルの導入により Miss Score の採点精度が向上したが，依然として QWK の値は低い．図 6 は Q1 において Miss Score の点数が 0 点の解答を，提案モデルが -1 点と誤って採点した解答を示す．解答文中の傍線部は言語モデルが最も誤りであると予測した単語を示す．

解答例 1
西洋人は自分の考えに他人を同意させる必要があると考え，日本人は他人は自分とおなじ人間と考えるから

解答例 2
西洋人は基本的には他人は自分と異なる人間とみなす．だから自分の考えを他人に同意させる必要があり言葉をつくして自分の考えを伝えようとすること．

図 6 Q1 の解答に対する提案モデルの採点例

“おなじ”や“つくし”といった，漢字で書くことができる単語をひらがなで書いている場合に，誤字として誤って-1 点とされてしまう場合があった．このような記述による表記ゆれに対応するには，平仮名を漢字に変換する前処理が改善の手法として考えられる．

図 7 は Q4 において Miss Score の点数が 0 点の解答を，提案モデルが -1 点と誤って採点した解答を示す．

解答例
神対人間，人間対自然，人間対人間のように様々な民族などが雑居している西洋では，相手を説得するための対決のスタンスが主流であるということ

図 7 Q4 の解答に対する提案モデルの採点例

解答文中の“主流”という単語に対して，言語モデルは最も誤りであると予測していた．この“主流”という単語が用いられているのは，Q4 の全解答を通じて 1 解答のみであるため，誤りと予測されたと考えられる．このような全解答を通じて，現れるのが 1 回のみとなる単語を含む解答に対しては，機械学習による学習，検出が難しいため，今後の課題としたい．

6. まとめ

本研究は誤字脱字等の誤りに対する減点項目の採点精度の向上を目的とした，ニューラルネットワークモデルを提案した．提案手法としては採点項目ごとに重要な箇所に注目して採点を行う従来モデルに，解答文中に含まれる各単語が正しいかの確率を予測して採点を行う言語モデルを導入している．これにより誤字脱字等の誤りに対する減点項目の採点と，その他の採点項目の採点を異なるネットワークで採点をすることができる．提案手法を用いることで国語記述式問題 6 題のうち，4 題に対して Miss Score の採点精度を向上することができた．しかし，QWK の値としては依然として低く，より高い精度で採点を行うことが今後の課題である．その解決方法として，記述で解答することによる多様な表記ずれにも対応するための前処理やモデルの改善が挙げられる．

謝辞

本研究では，国立情報学研究所の IDR データセット提供サービスにより国立研究開発法人理化学研究所から提供を受けた「理研記述問題データセット」を利用した．

参考文献

[1] 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 (BEA 14). pp. 316-325, August 2019.

[2] 高橋 諒, 蓑田 和麻, 舛田 明寛, 石川 信行, “Bidirectional LSTM を用いた誤字脱字検出システム”, 人工知能学会第 33 回全国大会, 2019.

[3] 理化学研究所（2020）：理研記述問題採点データセット．国立情報学研究所情報学研究データレポジトリ．データセット：https://doi.org/10.32130/rdata.3.1