E-022

ニューラルネットワークを利用した日本語小論文の自動採点の検討
Automatic scoring of Japanese essays using neural networks

清野光雄 Mitsuo Kiyono
竹内孔一 Koichi Takeuchi

1. はじめに

2020 年に現在施行されているマーク方式の「大学入試センター試験」が廃止され，記述式問題を含む「大学入学希望者学力評価テスト」への変更が検討されている．記述式問題のひとつである小論文問題を採点するにあたり，現在では小論文問題は人手による採点が行われている．しかしこれには非常に多くの労力がかかり，また採点者の身体的，精神的疲労などにより公正な採点を行うのが難しい．よって小論文問題の採点者を支援するシステムが必要となる．

本研究では，機械学習により小論文を自動で採点するシステムを提案する．機械学習を用いて短答式の試験で成功している Neural Attention モデル [1] を小論文を自動で採点するシステムに適用し，どの程度有効かを明らかにする．

2.Neural Attention モデルを利用した自動採点システムの構築

本稿で提案する小論文自動採点システムでは，人手で採点した点数と機械学習の手法を用いて，未知の小論文を採点する．こうした採点済みのデータを用いる手法は，既に英語圏で利用されている e-rater や IntelliMetric[2] でも用いられており，広く採用されている．

自動採点システムの入力は，形態素解析された日本語小論文である．この形態素を学習に利用できる形にするため，200 から 300 次元のベクトルである分散表現ベクトルに変換する．形態素解析には MeCab*を利用する．形態素を分散表現に変換する際には nwjc2vec[3] を利用する．nwjc2vec とは国立国語研究所†が収集している約 1 億ページの日本語 Web コーパスに対して fasttext[4] を利用することで得た形態素のベクトル表現である．

入力形態素列の特徴を取り込むニューラルネットワークのモデルとして，ゲート付き RNN の 1 つである LSTM を用いる [5]．LSTM の特徴として，勾配消失が起きにくく，長期的な依存関係を記憶セルに保存できることが挙げられる．また各 LSTM の出力には，Attention 機構で重みづけを行う．Attention とは，Encoder の出力の中から重要な情報に注意を向けさせる仕組みである．各 LSTM の中間層から得られた隠れ状態の中で，重要な情報であるかの重みを付け，LSTM の最終出力に加える．

自動採点システムは 1 から 5 まで整数で点数を出力する．教師信号として人手で採点された 1 から 5 点のスコアを 5 次元の One-Hot ベクトルとして与え，最終出力層との積で誤差量を求める．

*https://taku910.github.io/mecab/
†https://www.ninjal.ac.jp/

表 1: 各小論文のテストデータに対する Accuracy による精度評価

小論文の課題		Attention	LSTM	IDF
グローバリゼーションの光と影	1	0.439	0.439	0.314
	2	0.500	0.667	0.214
	3	0.545	0.394	0.263
自然科学の構成と科学教育	1	0.458	0.407	0.339
	2	0.552	0.569	0.398
	3	0.661	0.661	0.236
東アジア経済の現状	1	0.655	0.621	0.297
	2	0.534	0.397	0.344
	3	0.603	0.293	0.330
批判的思考とエセ科学	1	0.690	0.741	0.566
	2	0.397	0.397	0.360
	3	0.466	0.345	0.259

3. 評価実験の設定

今回実験に使用する小論文データは，2016 年から 2017 年にかけて岡山大学にて行った 4 種類の模擬試験のものである．この小論文を課題提案者が定める 4 つの評価基準について採点した．今回小論文のスコアとして用いるのはその中でも「理解力」として付けた点数である．

学習したニューラルネットワークに対して, 学習に利用していない小論文を入力し，理解力の点数を推定し、人手で採点したスコアと比較する．ニューラルネットワークの出力は 1 から 5 のクラス分類のため，出力は 1 点から 5 点の 5 種類であり，人手によるスコアも 1 点から 5 点の整数で付与されていることから，評価尺度として, スコアが一致するかどうかの Accuracy が利用できる．全小論文の数を N，システムが推定したスコアと人手で付けたスコアが一致した小論文の数を A とすると，Accuracy(Acc) の計算式は下記の（1）式である．

$$Acc = \frac{(\text{推定点数と人手の点数の一致数})}{(\text{全小論文の数})} \quad (1)$$

4. 実験結果

学習に Neural Attention モデルを利用したものと，LSTM の最終出力のみを用いて学習したものの 2 種類の自動採点システムを用意し，テストを行った．また，同様の小論文データに対して学習データを利用せずに Wikipedia の IDF 値を利用した大野らの手法 [6] による Accuracy を参考として掲載する‡．これらの結果を表 1 に示す．

‡テストデータとして対象とする文章は異なる

Attention を用いた手法，LSTM を用いた手法共に IDF を用いた手法よりも全体を通して高い Accuracy を得た．Attention と LSTM を比較すると，一部では Attention が LSTM の Accuracy より低くなっているが，全体で見れば Attention を用いた手法の Accuracy が高い箇所の方が多い．また，設問によっては Attention，LSTM 共に Accuracy に大きな差があることが分かる．例えば，設問「批判的思考とエセ科学」では問 2 について両モデルとも Accuracy が 0.4 を下回る低い値となっている．一方で，問 1 では 0.7 に近い値が得られた．これは問 1 が 100 文字程度と短いのに対し，問 2 は 400 文字程度と長いことが要因の 1 つと考えられる．一方で，各講義内容の自由度が比較的高い問 3 の Accuracy は Attention モデルでは最低でも 0.466 とそれほど低下していないことが分かる．これより長文の採点において Attention モデルは識別精度を下げないと考えられる．

5. 考察

5.1.IDF を用いた手法との比較

IDF を利用した手法は，小論文に対して講義内容の文章との単語のマッチングにより採点を行う．よって設問毎の違いや，人手の採点結果などが考慮されていない．一方で，本提案手法は事前の人手による点数を学習させるため，より高い精度の採点が実行できている．

しかし，全ての採点結果の 8 割を学習に利用しているにもかかわらず，ニューラルネットワークによる推定では 90%を越える点数の一致はできていない．理由の一つとして 1 から 5 点のクラス分類は人手でも難しい部分があり，ニューラルネットワークでも完璧にこなすのは難しかったことが考えられる．逆に，学習データを利用しない IDF による手法でも Accuracy はそれほど低くない．例えば，自然科学の構成と科学教育の設問 1 では LSTM の Accuracy が 0.407 のとき，大野らの手法でも 0.339 と近い値を示している．これは，設問 1 は事前の講義内容で話された自然科学に関する 100 文字の課題であり，解答すべき内容が限定的となる．このような問題には，学習データを利用せず，参照データとの IDF による単語の一致度が有効であることが分かる．

一方で，同講義の設問 3 は 500 字以上 800 字以内と自由度の高い記述問題である．このような場合，書く内容の自由度も高く，単純な IDF によるマッチはあまり有効には働いていない．それに対して，Neural Attention，LSTM 両モデルとも安定して約 66%の精度を出ており，教師データによる学習が有効であるといえる．よって自由度の高い小論文課題では，学習を用いる手法が特に有効であることが分かる．

5.2.LSTM と Neural Attiontion モデルを用いた手法との比較

LSTM のみを用いた手法よりも Neural Attention モデルを用いた手法の Accuracy が高いものは，設問 12 個のうち 6 個に対してのみであり，逆に低い場合は 3 個のみである．よって Neural Attention モデルを用いた手法の方が精度が高いことが分かる．しかし，LSTM のみを利用した手法との差は大きくない．また同点の設問が 3 個であることから，設問によっては Attention 機構が有効でないことが考えられる．

この原因の 1 つとして考えられるのは学習データの少なさと解答文章の長さである．短答式で成功している先行研究 [1] では 100 字以下の課題を 5000 事例学習させている．しかし，本研究では最長 800 文字の課題にもかかわらず，約 250 件前後のデータしか学習しないため，事例が不十分であった可能性が高い．よって精度の差が大きくなかったのではないかと考えられる．

%

6. まとめ

本研究では，事前に人手で採点された小論文学習データから，ニューラルネットワークを利用した自動採点システムを構築した．Neural Attention モデルを導入し，人手で採点された小論文データを学習した．Neural Attention モデルの入力として，各形態素を 200 次元の分散表現ベクトルに変換した nwjc2vec を利用した．再帰型ニューラルネットワークとして LSTM を利用し，1 から 5 点のクラス分けを学習させた．その結果，先行研究の人手による採点データを利用しない手法と比べて高い精度を示した．また，LSTM と Neural Attention モデルとの比較を行ったところ，Neural Attention モデルの方が全体的に高い精度を示した．今後，より少ない学習データで識別精度の高いモデルの構築を検討する予定である．

謝辞として、今回の研究の遂行にあたり岡山大学学務部にご協力いただきました．深く感謝いたします．

参考文献

[1] 水本智也, 磯部順子, 関根聡, 乾健太郎. 採点項目に基づく国語記述式答案の自動採点. 言語処理学会第 24 回年次大会発表論文集, pp. 552–555, 2018.

[2] 石岡恒憲. コンピュータ上で実施する記述式試験―エッセイタイプ，短答式，マルチメディア利用について―. 電子情報通信学会誌, Vol. 99, No. 10, pp. 1005–1011, 2016.

[3] 浅原正幸, 岡照晃. nwjc2vec:『国語研日本語ウェブコーパス』に基づく単語の分散表現データ. 言語処理学会第 23 回年次大会, pp. 94–97, 2017.

[4] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of Tricks for Efficient Text Classification. *arXiv preprint arXiv:1607.01759*, 2016.

[5] 岡谷貴之. 深層学習. 講談社, 2015.

[6] 大野雅幸, 竹内孔一, 泉仁宏太, 小畑友也, 田口雅弘, 稲田佳彦, 飯塚誠也, 阿保達彦, 上田均ほか. 参照データと idf を利用した事前採点不要な小論文評価手法. 研究報告自然言語処理 (NL), Vol. 2018, No. 17, pp. 1–6, 2018.