Neural Automated Essay Scoring Incorporating Handcrafted Features

宇都 雅輝[†]

1 はじめに

本稿では、COLING'2020 に採択された論文 “Neural Automated Essay Scoring Incorporating Handcrafted Features” (Uto et al. 2020) について解説する。本採択論文では、小論文自動採点タスクにおける得点予測の精度向上を目指し、深層学習に基づく自動採点モデルに伝統的な特徴量ベースの手法（人手で設計した特徴量を用いる手法）を統合するフレームワークを提案した。後述するように、提案手法は単純なアイデアではあるが精度改善に大きく寄与し、従来の様々な深層学習自動採点モデルに容易に組み込むことができる。次章では、採択論文の紹介に先立ち、著者が本研究の着想に至った経緯や COLING'2020 に投稿するまでの経緯について紹介する。

2 着想・論文投稿の経緯

大学院からの著者の主な研究内容は、ベイズ統計や機械学習手法に基づく教育評価のための数理手法の開発であり、自然言語処理分野の研究歴は比較的短い。実際、言語処理学会年次大会での研究発表は 2019 年が初めてであり、言語処理系の国際会議への投稿は本採択論文が初めてである。そのような著者が、COLING への投稿に至るまでの経緯を簡単にご紹介したい。

著者が長年行ってきた研究の一つとして、記述・論述式試験やスピーキング試験、実技試験などのパフォーマンス形式の試験（以降、パフォーマンス評価と呼ぶ）の信頼性に関わる研究が挙げられる。パフォーマンス評価は、論理的思考力や創造力、表現力などの実践的な能力を測定するニーズの高まりに伴い、近年様々な学習・評価場面において活用されている。他方で、パフォーマンス評価の問題として、人間評価者の主観採点を伴うことにより、得られるスコアが評価者の甘さ/厳しさなどの特性差に依存してしまい、評価の信頼性が低下する点が指摘されてきた。著者は、この問題を解決するために、評価者の特性差の影響を考慮して対象の真のスコアを推定できる数理モデルを開発してきた (e.g., Uto and Ueno 2020)。これらのモデルは、テスト理論の一つとして教育評価・心理測定の分野で広く知られる項目反応理論 (Item Response

[†] 電気通信大学

Theory: IRT) に基づくモデルとして定式化している。著者は、そのような基礎モデルを開発するとともに、様々な応用研究や実試験への適用などを通して、これらのモデルが評価者バイアスの影響を取り除いた高精度なスコアリングに有効であることを示してきた。

著者が当初開発したモデルでは、評価者の特性を表すパラメータと対象の真スコアを表す潜在変数を、評価者が与えた得点のデータセットのみから推定する。しかし、研究を進める中で、観測される得点データが少ない場合やスパースな場合、これらのパラメータや潜在変数の推定が難しくなることが明らかになった。そこで著者は、この問題を解決するために、対象を小論文試験に限定し、自然言語処理技術を活用したモデル拡張を考えた。具体的には、潜在ディリクレ配分法 (LDA) を用いて各小論文のトピック分布を推定し、そのトピック分布を項目反応モデルで推定される真スコアに反映させることで、評価者が与える得点データと小論文の文章情報を同時に加味して真スコアを推定できるモデルを開発した (Uto 2019)。振り返ると、この研究が自然言語処理分野に参入するきっかけになったと考える。

得点データと文章情報からスコアを推定する上記のモデルは、運用方法によっては一種の自動採点モデルとしても利用できる。他方でその当時、国内では大学入試改革が話題になっており、記述式試験の導入に関して様々な議論がなされていた。記述式試験を含むパフォーマンス評価の信頼性について研究していた著者は、そのような議論の中で、測定の信頼性だけでなく、大量の答案を短期間で採点することの重要性を改めて強く認識し、自動採点の今後のニーズ拡大を予想した。そこで自動採点の研究に焦点化して調査を進めたところ、当時は深層学習を用いた自動採点モデルが提案されて間もない時期であり、様々な拡張モデルが提案され始めていた。著者も深層学習自動採点モデルの研究を始めたが、やや後発であったこともあり、深層学習のアーキテクチャのみで新規性・有効性の高い技術を開発することはすぐには難しいと感じた。そこで、視点を変えて新しいフレームワークを開発することを考えた。その中心的な発想は、小論文のテキスト情報のみを深層学習モデルに入力するのではなく、その他の情報も加味できるようにすれば性能が向上するのではないかというものである。本採択論文も、専門家が設計した特徴量を深層学習モデルに組み込むというものであり、この発想に基づいている。他にも、上述したような項目反応理論で得られる情報を深層学習自動採点モデルに組み込んだ研究なども行っている (e.g. Uto and Okano 2020)。

項目反応理論を活用した自動採点技術の研究は、教育評価の技術を利用しているため、著者が毎年投稿している国際会議 AIED (人工知能技術の教育応用に関するトップ国際会議の一つ) で発表した。他方で、本稿で紹介する論文は、言語処理技術のみの研究であったため言語処理分野の国際会議が適当と考え、AIED と同ランク (CORE ランク=A) であり、投稿スケジュールの都合が良かった COLING に投稿した。冒頭にも記述したように言語処理分野の国際会議は投稿自体が初めてであったため不安もあったが、無事採択となり、査読や発表を通して様々な意見をもらうことができた。本採択論文の内容を次章で紹介する。

3 本研究の紹介

既存の小論文自動採点技術は大きく 2 つのアプローチに分類できる (Ke and Ng 2019). 一つは専門家が設計した特徴量 (単語数やスペルミス数など) を利用する特徴量ベースのアプローチであり, 古くから様々な研究がなされてきた. もう一つは深層学習に基づくアプローチであり, 小論文の単語系列を入力することで, 事前の特徴量設計を行うことなく得点予測を行うことができる. 深層学習ベースの自動採点モデルは, 2016 年に初期モデル (Taghipour and Ng 2016) が提案されて以降, 活発に研究がなされている.

一般に特徴量ベースと深層学習ベースの手法は独立に研究されてきたが, これらの二つの手法はそれぞれに異なる利点を有しており, 本来は対立する手法ではない (Ke and Ng 2019). そこで, 近年, これらのアプローチを統合したハイブリッド手法が Dasgupta ら (2018) によって提案された. このハイブリッド手法は, 深層学習モデルで内部的に得られる特徴表現と人手で設計した特徴量を同時に利用して予測得点を計算するというものであり, いずれか片方のみを用いるよりも性能を改善できることが示された. Dasgupta らのモデルでは, 文単位で特徴量を設計・抽出し, その文単位の特徴量の系列をリカレントニューラルネットワークに入力して固定次元の分散表現ベクトルを得る. そして, そのベクトルを, 単語系列を入力とする従来の深層学習自動採点モデルから得られる分散表現ベクトルと統合して得点予測を行う. しかし, このモデル設計には大きく次の問題がある. 1) 特徴量ベース手法の研究で有効性が示されてきた文章レベルの特徴量を活用できない. 2) 単語系列と文単位の特徴量系列の二種類の時系列入力を処理する深層学習モデルを内部に持つため, 従来の深層学習自動採点モデルと比べて, 学習パラメータやチューニングパラメータが大幅に増加する.

3.1 提案手法

上記の問題を解決するために, 本研究では, 文章レベルの特徴量を利用できる新たなハイブリッド手法を提案した. 提案手法は, 従来の深層学習自動採点モデルで得られる分散表現ベクトルに, 文章レベルで設計した特徴量を直接結合する手法として定式化した. 具体的には, 従来モデルにおいて出力層 (全結合層) の直前で得られる固定次元の分散表現ベクトルに, 人手で設計した文章レベルの特徴量を連結 (Concatenate) し, そのベクトルを出力層に入力して予測得点を求めるという構成になっている. 提案手法の概念図を図 1 に示す.

既存の深層学習自動採点モデルの多くは, 最終層の直前で分散表現ベクトルを生成する構成となっているため, 提案手法を容易に適用できる. また, 提案手法では特徴量ベース自動採点の研究で有効性が示されてきた様々な文章レベル特徴量を利用できる. さらに, 既存モデルの出力層に特徴量を加えるだけでよいため, 増加するパラメータは採用する特徴量と同数の出力層の重みパラメータのみであり, 新たなチューニングパラメータも必要としない.

3.2 評価実験

本論文では、提案手法の有効性を評価するために、小論文自動採点のベンチマークデータセットとして知られる Automated Student Assessment Prize を用いて得点予測精度の評価を行った。このデータセットは8つのプロンプト（小論文課題）に対する採点済み小論文データで構成される。得点予測精度の評価は二次重み付きカッパ係数を評価指標として5分割交差検証法で行った。自動採点モデルには、LSTM ベースのモデルと BERT ベースのモデル、従来のハイブリッドモデルを利用し、文章レベル特徴量を組み込んだ提案手法と特徴量を用いない従来手法とで比較を行った。詳細な実験条件やモデルの設定については論文を参照してほしい。

実験結果の一部を表 1 に示す。表では「+文章レベル特徴量」の行が提案手法の結果を表す。表から、すべてのベースモデルにおいて提案手法を用いて特徴量を加えたときに平均の精度が向上していることが確認でき、提案手法が自動採点の精度向上に有効であることがわかる。なお、紙面の都合上ここでは結果の一部のみを示した。詳細な結果は原論文を参照されたい。

図 1 提案手法の概念図

	プロンプト（小論文課題）								平均	p-value
	1	2	3	4	5	6	7	8		
LSTM (MoT)	0.72	0.52	0.62	0.78	0.80	0.78	0.75	0.56	0.69	0.02
+文章レベル特徴量	0.82	0.65	0.62	0.79	0.79	0.81	0.79	0.69	0.75	
双方向 LSTM	0.48	0.42	0.50	0.78	0.74	0.72	0.63	0.22	0.56	0.01
+文章レベル特徴量	0.78	0.60	0.58	0.78	0.76	0.77	0.76	0.66	0.71	
BERT	0.829	0.391	0.76	0.89	0.88	0.58	0.82	0.54	0.71	0.02
+文章レベル特徴量	0.85	0.65	0.80	0.89	0.89	0.82	0.86	0.65	0.80	
従来のハイブリッド手法	0.73	0.64	0.63	0.79	0.80	0.79	0.77	0.69	0.73	0.07
+文章レベル特徴量	0.82	0.67	0.60	0.80	0.79	0.81	0.81	0.71	0.75	

表 1 実験結果

4 おわりに

本稿では, COLING'2020 に採択された論文 “Neural Automated Essay Scoring Incorporating Handcrafted Features” について, 投稿に至るまでの経緯も含めて紹介した.

参考文献

- Dasgupta, T., Naskar, A., Dey, L., and Saha, R. (2018). “Augmenting Textual Qualitative Features in Deep Convolution Recurrent Neural Network for Automatic Essay Scoring.” In *Proceedings of the Workshop on Natural Language Processing Techniques for Educational Applications, Association for Computational Linguistics*, pp. 93–102.
- Ke, Z. and Ng, V. (2019). “Automated Essay Scoring: A Survey of the State of the Art.” In *Proceedings of the International Joint Conference on Artificial Intelligence*, pp. 6300–6308.
- Taghipour, K. and Ng, H. T. (2016). “A Neural Approach to Automated Essay Scoring.” In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 1882–1891.
- Uto, M. (2019). “Rater-effect IRT Model Integrating Supervised LDA for Accurate Measurement of Essay Writing Ability.” In *Proceedings of the International Conference on Artificial Intelligence in Education*, pp. 494–506.
- Uto, M. and Okano, M. (2020). “Robust Neural Automated Essay Scoring Using Item Response Theory.” In *Proceedings of the International Conference on Artificial Intelligence in Education*, pp. 549–561.
- Uto, M. and Ueno, M. (2020). “A Generalized Many-facet Rasch Model and Its Bayesian Estimation using Hamiltonian Monte Carlo.” *Behaviormetrika, Springer*, **47** (2), pp. 469–496.
- Uto, M., Xie, Y., and Ueno, M. (2020). “Neural Automated Essay Scoring Incorporating Handcrafted Features.” In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 6077–6088. International Committee on Computational Linguistics.

略歴

宇都 雅輝: 2013 年電気通信大学大学院情報システム学研究科博士後期課程修了. 博士 (工学). 現在, 電気通信大学准教授. AIED Best paper award, 日本行動計量学会奨励賞, 人工知能学会研究会優秀賞, 日本テスト学会大会発表賞, 教育システム情報学会全国大会奨励賞など受賞.