様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　６ 年　 ６ 月 １８ 日現在

機関番号：８２６１６
研究種目：基盤研究(B)（一般）
研究期間：2020～2022
課題番号：２０Ｈ０４３００

研究課題名（和文）教師採点データに拠らない深層学習に基づく記述式自動採点システムの開発

研究課題名（英文）An Development of automated short-answer scoring system based on deep learning without using supervised scoring data

研究代表者

石岡　恒憲（Ishioka, Tsunenori）

独立行政法人大学入試センター・研究開発部・教授

研究者番号：８０３１１１６６

交付決定額（研究期間全体）：（直接経費）　　12,400,000円

研究成果の概要（和文）：近年、自然言語での記述文を順番のある時系列データと見なし、これを入力データとして処理するリカレントニューラルネットワークと呼ばれる深層学習手法、特にバートなどのトランスフォーマーの研究が進み、その性能の良さが証明されてきた。そこで平成29年と30年に実施した共通テスト試行調査12万件による記述回答データを文字認識から一気通貫でバートによる自動採点までを行うことを試みた。我々の共同研究グループは、通常の採点システムが用いる人手による補助輪をしない実運用で平均96%，最低でも93%の一致率を確保した。また各問６万件という膨大なデータにより、深層学習に必要な標本サイズについても新たな知見を得た。

研究成果の学術的意義や社会的意義
いままでの研究では学習データに用いるサンプルはせいぜい２千件程度であり、どの程度のサンプルがあれば十分な予測ができるかの目安は与えられていなかった。さらに九大グループでは意味的埋め込みと呼ばれる異なったアプローチによる方法を試みた。これら結果については本科研で3件の学会表彰（日本計算機統計学会第35回大会，学生研究発表賞；Duolingo Award for IMPS 2021；SMASH22 Winter Symposium,準優秀賞）を受け、その成果については日本教育新聞や日経新聞教育面に大きく掲載された。その後、教育工学のトップ国際会議AIED 2022でも論文採択された。

研究成果の概要（英文）：In recent years, research into deep learning methods called recurrent neural networks, especially transformers such as BART, has progressed, and their excellent performance has been proven. Here, we consider written sentences in natural language as time-series data with an order, and process this as input data. We attempted to process written response data from 120K common test trial surveys conducted in 2017 and 2017, from character recognition to automatic scoring using Bart, all at once. Our collaborative research group achieved an average agreement rate of 96% and a minimum of 93% in real-world operations without the manual training wheels used in conventional scoring systems. Additionally, by using a huge amount of data containing 60K questions for each question, we gained new knowledge about the sample size required for deep learning.

研究分野：情報数理

キーワード：自然言語処理　自動採点　機械学習　深層学習　トランスフォーマー　手書き文字認識

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

記述式テストはいままでの多肢選択テストに比べ、より正統あるいは真正(authentic)で信頼できると広く考えられており、採点のための技術的な課題が克服されてさえいれば、これを使う潜在的な需要は（センター試験に限らず）計りしれない。著者らはエッセイタイプの問題に対して日本語で初めての小論文自動採点システムJess (Ishioka, 2006)を開発した。成果は朝日新聞夕刊の一面トップ（2005年2月）に大きく掲載されたほか、Yahoo!インターネットガイド（2006年6月）、韓国KBSテレビ（2007年2月）等、多くのマスコミで紹介された。

その後、模範解答との意味的合致や含意の判定が必要となる短答式記述採点システムの開発に科研費などを得て着手した。2011 年より国立情報学研究所が主催する「ロボットは東大に入れるか」（東ロボ）プロジェクトが動き出し、オールジャパンでロボットがセンター試験や東大の２次試験に合格すべき、問題を理解し解答を作成する技術を競った。研究代表者（石岡）は東ロボプロジェクトの共同研究者の一人として参加した。世界史や日本史などの問題は試験問題の正解が教科書に書かれているので、試験問題が教科書の文章との同義あるいは含意の判定が正しくできれば正解を導くことができることになる。東ロボは 2016 年まで実施されたが、その結果わかったことは、完全な含意関係認識技術すなわち正しい意味理解は現時点では困難であり、いくつかの手法を組み合わせて半ばアドホックに解くことで、あたかも人間が解答したかのようにみせかけることがせいぜいであるという事実であった。それでもセンター試験で 100 点満点 76 点をマークすることができたので、大学入試試験レベルの短答式記述試験の（ある程度の）自動採点と人間による採点を支援するシステムを試作・実装した。国立情報学研究所が主催する競争型の国際シンポジウム NTCIR(エンティサイル)－13 の質問応答タスク(QALab-3)に参加し、本システムの自動採点の性能は国内外の 11 機関の中で同率トップと評価された（2017 年 12 月）。ただ近年の人工知能による自然言語処理技術の進歩は目覚ましく、最新の技術を踏まえてシステムの改良を行うことは急務であった。

エッセイタイプについてはこれまでも、特にアメリカにおいて多くのシステムが開発され実用に供されてきた。アメリカのビジネススクール入学のための共通テストであるGMATにおける作文テストでは、1998年よりe-raterが、2006年よりはIntelliMetricが採点をおこなっている。他にも商用のシステムとして、PEG（Project Essay Grade）やIEA（Intelligent Essay Assessor）、CRASEなどがあり利用に供されている。

しかし今ここで開発しようとしている短答式テストの自動採点については、その重要性は認められているものの技術的にさまざまな課題が未解決のままである。Vigilante(1999) は世界最大のテスト機関である ETS とニューヨーク大学とで、この分野における共同研究を行い、最初の報告をした。Leacock & Chodorow (2003)は、ETS が開発した c-rater の最新の仕様について報告している。Pulman & Sukkarieh (2005)は、情報抽出技術に隠れマルコフモデルなどの自然言語処理を用いて、システムが用意する正解文と同じ意味の文を幾つか自動生成する試みについて述べている。しかしその性能は、エッセイの自動評価採点システム(e-rater)に比べ、人間による専門家との一致率は 10%以上も小さい。このため海外においても短答式テストの自動採点システムは未だ研究段階であり、ハイステークスな試験においては未だ利用されていなかった。現在もそうである。

２．研究の目的

短答式記述テストをコンピュータで自動採点し、その自動採点に至った根拠を示すことで人間（採点者）が自動採点を修正することのできる採点支援システムを実装する。採点は設問ごとに作題者が用意した「採点基準」に従いシステムがある程度（90%以上程度）の精度をもった採点計算（自動採点）を基本とし、その結果を人間が確認・修正できるものとする。システムは「（予め用意された）模範解答(部分点解答を含む)」と「（被験者の実際の）記述解答」との（ある程度の）意味的同一性や含意性を判定するほか、プロンプトと呼ばれる素材文と解答文との意味的近似性なども考慮にいれる。

採点者が自動採点の結果に同意するならば、採点者は標準（ディフォルト）で表示される採点結果を承認するだけで採点をすすめることができる。記述テストの採点支援について文科省には現時点で答案のクラスタリングしか案がなく、ここで実装されるシステムはまさに人工知能の活用の名に値する新規性・独創性のある課題達成といえる。

このシステムの最大の特徴は、意味的同一性や含意性の判定に採点済みの教師データを使わないことにある。予め別に用意された新聞や教科書、Wikipediaなど別のコーパスなどから自動構築した言語モデルによって判定を行う。教師学習を行う限り、学習モデルに含まれない（＝学習データとして与えられていない）データについてシステムは適切な予想を返すことはない。仮に何万件もの採点データが得られたとしても、センター試験のような全国レベルでの試験では予期せぬ解答は無視できない確率で生じる。人間による採点済みデータで全ての解答パター

ンを被覆することはそもそも無理なことであり、

採点するデータサイズ ≪ 採点されたデータサイズ

が成り立つ、すなわち「採点するデータサイズ」が「採点されたデータサイズ」に比べ圧倒的に小さい場合に限って機械学習は有効である。もちろん我々のシステムは、教師データがある場合にはそれを言語モデルに反映することもできるものとする。これにより、その場合は採点精度を向上させることができる。

短答式試験の完全なる自動評価採点支援システムには以下のメリットがある：

(1) 経済性：もし人間による評価採点の大半をコンピュータに置き換えることができれば、試験の評価に対するコストを大幅に削減することができる。アメリカのビジネススクールでの入学共通試験であるGMATにおける作文試験での適用の仕方と同様に、最終判定点の決定が人間側に委ねられてさえいれば、コンピュータ採点は一般には受け入れ易いと考えられる。

(2) 即時性：短答式の自動採点はシステムがよくできてさえいれば、即座のフィードバックを返すことができる。本システムでも採点基準への適合の程度をシステムが自動的に判定するから、初心者の採点者はこれをみることで採点行為そのもののやり方を学習訓練することができる。これは個人教育的な側面を有しているともいえる。

(3) 説明責任：近年、採点の論拠を示すことの重要性はますます増大してきている。本システムはこれらの社会的要請へ応えるものといえる。

３．研究の方法

本科研では、「システム性能の実用に耐えうるだけの性能を確保すること」と、「採点アルゴリズムの汎化（試験問題のタイプが変わってもシステムの採点アルゴリズムを変更しなくても済むようにすること）」を目標としたい。そのために近年、ディープラーニング（深層学習）の研究で注目されているリカレントニューラルネット（RNN）及びその改良であるロングショートタームメモリ（LSTM）をこの短答式自動採点に取り込むことを目指した。本応用の当初考えた達成課題は以下の通りである。

(1) 文章生成が目的ではなく入力文（解答文）が模範解答と同一、あるいはその言い換えになっているかの判定こそが重要である。これは従来法、すなわち入力文を分散表現によりベクトル化し、これと模範解答のベクトル表現の関連度による評価でもある程度は達成できると思われるが、出力である生成文のベクトル表現との関連度も用いることで更なる性能向上が期待される。関連度評価にはコサイン近似度を用いる。

(2) 採点基準には部分点（解答文のある一部が合致していれば得点を与える）があり、その判定が必要である。このためにアテンション機構を用意し、RNN やLSTMの出力に対して適用する。

(3) 解答文が論理的に意味のある文になっているかの判定が必要である。入力文を分散表現によりベクトル化し、その分散表現の繋がりの滑らかさをコーパス（新聞等の文書集体）を用いて評価する。

(4) 理由を答えるべき設問に対して解答がそのような回答形式になっているか、あるいは名詞や名詞句を問う設問に対してそのように答えているかの判定がときには必要である。このために、理由を答える文の形式をコーパスから学習し、そのための辞書を作成する。

研究課題に対する大よその分担は以下の通りである：

石岡（研究代表者）：総括、自然言語処理、システム設計、機械学習

峯（研究分担者）：自然言語処理、システム構築、評価手法の妥当性評価

宮澤（研究分担者）：システム実装、CGI プログラミング

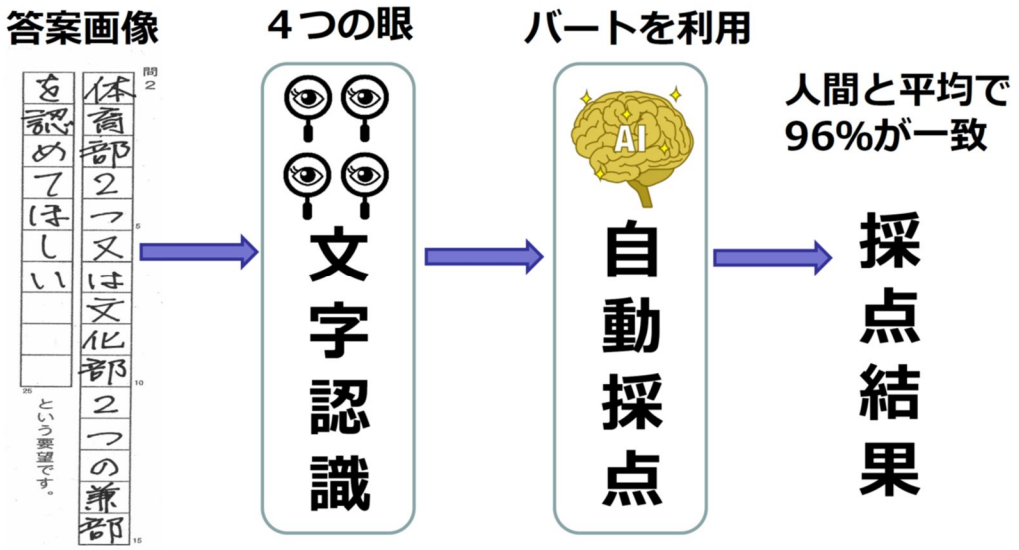
本システムは Web アプリケーションとして提供することを想定し、全体設計や Web システムの構築については石岡が担当する。評価手法の妥当性評価を峯が行う。ユーザインターフェイスや画面設計周辺については宮澤が担当する。またユーザや外部からの助言を得て、システム設計における PDCA サイクルを回す。システム開発のための研究員を雇用する。

４．研究成果

2017 年と 2018 年に実施された共通テストのための試行調査のうち、国語の記述回答（年ごとに各３問、２年間で計６問）の手書き文字解答データ（各年６万件、２年で１２万件）を、農工大・中川教授の研究グループの貢献によってデジタル化した。この文字認識データを、2018 年にグーグルが開発したバートと呼ばれる、現在の最新の言語モデルを用いて採点させた。全体の８割の解答を人間が採点した上で、その採点結果を学習させてから残りの解答について自動採点させた。その結果、３段階から７段階の評価で、人間の採点結果と平均で 96%が一致した。最も一致率の低い場合でも 93%だった。白紙の答案は含まない。

図は AI 自動採点の模試図である。手書き認識では、四種類の深層学習を行う AI が事前に百万件を超える手書き文字データを学習し、それぞれの観点で文字の候補を絞ったのち、その確信度に応じて最終的に文字を決定する。人間でいえば、いわば四人の眼で見て総合的に合議の上、判断する。「三人寄れば文殊の知恵」のことわざの通り、複数の眼を通すことでその精度を上げる。学習用の手書き文字データは縦書きを想定しておらず、今回の試行調査データのような縦書き

文字の認識はより難しい。その処理速度は、一般的な研究室にある深層学習用マシンで、25 文字の解答に対して 0.059 秒、80 から 120 文字の解答に対しては 0.29 秒である。



図：AI 自動採点の仕組み

我々の研究の技術的な革新は二つある。一つは手書き文字認識から自動採点までを一気通貫で行い、そこにコンピュータが採点メカニズムを理解するための人手による「補助輪」を一切用いないことである。採点時間に制限のある大規模試験では重要なことである。革新の二つ目は、12 万件というこの分野においては極めて大量の採点データの利用である。試験の採点データは一般には非公開とされ、利用できるデータ数は限られている。国内外の過去の研究においても２千件程度である。このような大規模な自動採点は、おそらく世界でも初めてである。我々は十分に大きなデータを扱うことにより、採点に必要な標本サイズに対しても幾つかの知見を得ることができた。

我々の試作システムとその性能については、国内で多くの評判をよび、日経新聞に２千字の寄稿が、また日本教育新聞には１面トップで掲載された。またその成果については多くの学会賞を受けた（日本計算機統計学会第 35 回大会，学生研究発表賞；Duolingo Award for IMPS 2021；SMASH22 Winter Symposium,準優秀賞；言語処理学会第 28 回年次大会，若手奨励賞）。また、教育工学のトップカンファレンスである AIED 等の国際会議で成果を発表した。

その後、研究分担者である農工大・中川教授の研究グループでは共通テスト試行調査における手書き数式認識という更に難しい問題にチャレンジしている。同じく分担者である千葉大・須鎗教授の研究グループでは少ない解答サンプルでも十分な性能を確保できるよう、効果的な学習データのサンプリングについての研究に着手している。

５．主な発表論文等

〔雑誌論文〕　計3件（うち査読付論文　3件／うち国際共著　3件／うちオープンアクセス　0件）

１．著者名 Nguyen, H.T., Nguyen, C. T., Oka, H., Ishioka, T. & Nakagawa, M.	４．巻 LNCS 13639
２．論文標題 Handwriting recognition and automatic scoring for descriptive answers in japanese language tests	５．発行年 2022年
３．雑誌名 International Conference on Frontiers in Handwriting Recognition, ICFHR 2022	６．最初と最後の頁 274-284
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 該当する

１．著者名 Bo Wang, Billy Dawton, Tsunenori Ishioka and Tsunenori Mine	４．巻 20th
２．論文標題 Optimizing Answer Representation using Metric Learning for Efficient Short Answer Scoring	５．発行年 2023年
３．雑誌名 Pacific Rim International Conference on Artificial Intelligence (PRICAI)	６．最初と最後の頁 236-248
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 該当する

〔学会発表〕　計9件（うち招待講演　1件／うち国際学会　0件）

１．発表者名 岡知樹, N.T.Hung, N.T.Cuong, 中川正樹, 石岡恒憲
２．発表標題 大学入学共通テスト試行調査における記述式問題の自動採点
３．学会等名 日本計算機統計学会第35回大会, 学生研究発表賞
４．発表年 2021年

１．発表者名 Oka,H., Hung,N.T., Cuong,N.T., Nakagawa,M., Ishioka,T.
２．発表標題 Short answer scoring of the trial test for Japanese Common University Entrance Examination,
３．学会等名 IMPS, Duolingo Award
４．発表年 2021年

１．発表者名
Wang,B., Ishioka,T., Mine,T.

２．発表標題
Automated Short Answer Grading with Rubric-based Semantic Embedding Optimization

３．学会等名
SMASH22 Winter Symposium

４．発表年
2022年

１．発表者名
石岡恒憲，岡知樹，N.T.Hung，N.T.Cuong，中川正樹

２．発表標題
共通テストの試行調査国語記述解答データを用いた自動採点のアルゴリズムとその評価

３．学会等名
日本テスト学会第19回大会発表論文抄録集，124-125.

４．発表年
2021年

１．発表者名
岡知樹，N.T.Hung，N.T.Cuong，中川正樹，石岡恒憲

２．発表標題
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

３．学会等名
言語処理学会第28回年次大会，E3-5，若手奨励賞

４．発表年
2022年

１．発表者名
Hung Tuan Nguyen, Cuong Tuan Nguyen (TUAT), Haruki Oka (UTokyo), Tsunenori Ishioka (The National Center for University Entrance Examinations) , Masaki Nakagawa (TUAT)

２．発表標題
Fully automatic scoring of handwritten descriptive answers in Japanese language tests

３．学会等名
電子情報通信学会 研究会PRMU2021-32，45-50.

４．発表年
2022年

１．発表者名
Ishioka, T.

２．発表標題
AI-based+ Automated Short-answer Scoring System

３．学会等名
Digital World 2020（招待講演）

４．発表年
2020年

１．発表者名
加藤博之，石岡恒憲，峯恒憲

２．発表標題
短答式試験における自動採点のための概念辞書を用いたデータ拡張手法の提案

３．学会等名
信学技報，vol. 120，no. 344，AI2020-15，pp. 7-12

４．発表年
2021年

１．発表者名
Ung, H. Q., Nguyen, C. T., Oka, H., Ishioka, T. & Nakagawa, M.

２．発表標題
Visual constraints for generating multi-domain offline handwritten mathematical expressions

３．学会等名
IEICE technical report, PRMU2021-69, pp. 54-59.

４．発表年
2022年

〔図書〕　計1件

１．著者名 石岡恒憲（石井雄隆・近藤悠介(編)）	４．発行年 2020年
２．出版社 ひつじ書房	５．総ページ数 157
３．書名 自動採点研究のこれから.「英語教育研究における自動採点　現状と課題」	

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考
研究分担者	中川　正樹 (Nakagawa Masaki) (10126295)	東京農工大学・学内共同利用施設等・特任教授 (12605)	
研究分担者	峯　恒憲 (Mine Tsunenori) (30243851)	九州大学・システム情報科学研究院・准教授 (17102)	
研究分担者	須鎗　弘樹 (Suyari Hideki) (70246685)	千葉大学・大学院工学研究院・教授 (12501)	
研究分担者	宮澤　芳光 (Miyazawa Yoshimitsu) (70726166)	独立行政法人大学入試センター・研究開発部・准教授 (82616)	

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関