令和元年度 情報数理工学プログラム卒業論文概要

平成 28 年度 入学	学籍番号 1610630
指導教員 植野 真臣	氏名 村田 哲啓
題目 問題非依存の特徴量を学習する深層学習自動採点手法	

概要

自動採点手法として, 近年, 人手での特徴量設計を必要としない深層学習ベースの自動採点手法が多数提案されている.

深層学習ベースの自動採点手法として最も代表的な手法は, LSTM を利用したモデルである. このモデルは, 2016 年に提案されて以降, 近年の深層学習自動採点手法の基礎モデルとして広く利用されており, 様々な拡張手法も提案されてきた.

このような深層学習自動採点モデルを利用するためには, 事前に収集された採点済み答案のデータセットを用いてモデルを学習しておく必要がある. しかし, 実際の試験においては, 出題する問題ごとに大量の採点済み答案データを用意することは難しい場合が多い.

この問題を解決する最も単純なアプローチとしては, 出題する問題に関する少量の採点済み答案データと過去に出題・採点された問題に対する大量の採点済み答案データを同時に用いて自動採点モデルを学習する方法が考えられる. しかし, この方法では, 対象問題に固有の特徴をモデルに反映することが難しいため, 対象の問題が他の問題と異なる特徴を持つ場合に予測精度が低下すると予測できる.

この問題を解決するために, 本研究では, 問題非依存の特徴量と問題依存の特徴量を同時に学習できる新たな深層学習自動採点モデルを提案する. そして評価実験より, 提案手法が, 従来手法とは異なり, 精度向上に有効である問題に依存しない特徴量を学習できることと, 複数の問題の答案データからも問題に依存する特徴量を学習できることを示した.

電気通信大学情報理工学域
I 類 (情報系) 情報数理工学プログラム卒業論文

問題非依存の特徴量を学習する深層学習自動採点手法

2020 年 2 月 28 日

情報数理工学プログラム

学籍番号 1610630

村田　哲啓

指導教員 植野 真臣

目次

図目次

表目次

1　はじめに

近年, 記述・論述式試験を用いた能力評価のニーズが世界的に高まっている. 日本の大学入試でも, 受験者の思考力・判断力・表現力などの高次の能力の評価を目指して, 全問マーク式の大学入試センター試験に代わり, 記述式問題を導入した新たな試験設計が検討されていた [1]. しかし, 大学入試のように受験者数が大規模となる試験では, 採点に要する時間的・経済的なコストが高く，運用が困難となる また, 大規模試験では一般に多数の評価者で分担して採点作業が行われるが, そのような場合, 得られる評価点が採点者ごとの主観や特性に依存してしまい [2], 評価の公平性・信頼性を担保することが難しいという問題も指摘されている [3]. このような問題を解決する手法の一つとして, 自動採点技術の実用化が期待されており, これまでに数多くの研究が行われてきた.

自動採点手法として，これまでは，事前に人手で設計した特徴量を使うアプローチ（特徴量ベース・アプローチと呼ぶ）が主流であった [4,5,6,7,8,9,10,11,12,13]. 特徴量としては, 例えば, 総単語数や文ごとの平均語数, スペルミスや文法エラーの数, 接続表現の種類や数など, 様々なものが利用されてきた. このアプローチは, 実装が比較的容易であり, 一度実装すれば様々な試験に簡単に適用できるという利点がある. 他方で, このアプローチは性能が特徴量設計に強く依存することが知られており, 高精度を実現するためには十分な特徴量チューニングが必要となる. この問題を解決するアプローチの一つとして, 近年, 人手での特徴量設計を必要としない深層学習ベースの自動採点手法が多数提案されている.

深層学習ベースの自動採点手法として最も代表的な手法は，時系列データを処理する深層学習モデルである RNN（Recurrent Neural Network）の一種である LSTM（Long short-term memory）[14] を利用したモデルである [15,16]. このモデルは，2016 年に提案されて以降，近年の深層学習自動採点手法の基礎モデルとして広く利用されており，様々な拡張手法も提案されてきた. 例えば，文章の一貫性の獲得を目指した手法 [17,18,19] や答案文に加えて問題文の情報も扱える手法 [20], モデル学習法を強化学習に変更した手法 [21], 人手で作成した特徴量も組み込む手法 [22,23] などが代表的な拡張手法であり，いずれも高精度な自動採点を実現している.

このような深層学習自動採点モデルを利用するためには, 事前に収集された

採点済み答案のデータセットを用いてモデルを学習しておく必要がある．深層学習モデルの学習には一般に大量のデータが必要となる．しかし，実際の試験においては, 出題する問題ごとに大量の採点済み答案データを用意することは難しい場合が多い.

この問題を解決する最も単純なアプローチとしては, 出題する問題に関する少量の採点済み答案データと過去に出題・採点された問題に対する大量の採点済み答案データを同時に用いて自動採点モデルを学習する方法が考えられる. しかし, この方法では, 対象問題に固有の特徴をモデルに反映することが難しいため，対象の問題が他の問題と異なる特徴を持つ場合に予測精度が低下すると予測できる.

この問題を解決するために，本研究では, 問題非依存の特徴量と問題依存の特徴量を同時に学習できる新たな深層学習自動採点モデルを提案する．具体的には，全ての問題の答案データを処理するネットワークと問題ごとに処理するネットワークの二種類のネットワーク構造を内部的に有する深層学習モデルとして定式化する．このようにモデル化することで，全答案データを処理するネットワークでは問題に依存しない共通的な特徴量を学習し，問題ごとのネットワークでは個別の問題に固有の特徴量を学習できると期待できる．これにより，対象とする問題の特徴が他の問題と異なる場合の予測精度の向上が期待できる.

本論文の構成は以下の通りである．まず, 第二章では関連研究をいくつか紹介する．次に, 第三章では本研究の基礎モデルとして利用する深層学習自動採点モデル [15] について概説する．第四章では提案手法の詳細について説明し, 第五章では評価実験について述べる．最後に, 第六章でまとめと今後の課題について整理する.

2　関連研究

本研究の関連研究として, 本章では 2.1 で特徴量ベース手法, 2.2 で深層学習自動採点手法について紹介する.

2.1　特徴量ベース手法

本節では, 特徴量ベースの自動採点手法の概要について説明する. 特徴量ベースの手法は, 事前に人の手で設計した特徴量を使い, 機械学習モデルにより得点を予測する手法である. 初の自動採点機は, この手法により実装された PEG(Project Essay Grade)[9] である. PEG の最初のバージョンでは, 表層的な特徴量 (総単語数や前置詞の数など) に基づく重回帰モデルにより, 得点予測を行っていた. しかし, 予測精度が悪いことに加え, 受験者に仕組みが暴露するとその仕組みを悪用することで容易に高得点を得ることが可能であるという問題点が指摘されていた. その後, 改良が加えられ, 現在では商用化されている.

他にも, 多数の特徴量ベースの採点機が開発されている. 例えば, Landauer ら [10] により開発された IEA(Intelligent Essay Assessor) は, Latent Semantic Indexing によるエッセイの意味的な内容の一致を重視して, 得点を予測する手法である. また, Rudner ら [11] によって開発された BETSY(Bayesian Essay Test Score sYstem) は, 2 種類のベイジアンモデルによる確率計算を利用して, 得点を予測する手法である. Elliot ら [12] によって開発された IntelliMetric は, ルール発見アルゴリムにより, 採点済み答案データから人間の採点ルールを学習し, 得点を予測する手法である. 以降では, 特徴量ベース手法の代表例である e-rater(Electronic Essay Rater)[13] を紹介する.

e-rater は, ETS(Educational Testing Service) の Burstein ら [13] が開発したシステムで, TOEFL や GMAT(Graduate Management Admission Test) などで既に実用化されている. 厳選された 12 の特徴量に基づく重回帰モデルにより, 採点予測を行う. なお, 特徴量の重みパラメータは固定されている. e-rater の 12 の特徴量を表 1 に示す.

表 1　e-rater の特徴量

1	総単語数に対する文法誤りの割合
2	総単語数に対する語の使用法についての誤りの割合
3	総単語数に対する手順の誤りの割合
4	総単語数に対するスタイルに関する誤りの割合
5	談話ユニット数
6	各ユニットにおける平均単語数
7	エッセイを 6 点法で採点する際の語彙のコサイン類似度が最大となる点数
8	最高点を得たエッセイとのコサイン類似度
9	全単語数に対する異なる語彙の割合
10	語彙の困難度
11	平均単語長
12	総単語数

2.2　深層学習自動採点手法

前節の特徴量ベース手法の利点は, 実装が比較的容易であることと, 一度実装すれば様々な試験に適用できるという汎用性である. しかし, 高精度を実現するためには, 対象とするデータセットに適した特徴量のチューニングや再設計が必要となる. この問題を解決する手法の一つとして, 近年, 人手での特徴量設計を必要としない深層学習ベースの自動採点手法が多数提案されている. 本節では, この深層学習ベースの自動採点手法を用いたいくつかの自動採点モデルについて説明する.

深層学習自動採点機の代表例は, LSTM を用いた手法 [15] である. この手法については次章で詳細に述べるが, この手法では, 答案の単語系列を入力として受け取り, 多層のニューラルネットワークを通して得点の予測値を出力する. この手法が提案されて以降, 近年の深層学習自動採点手法の基礎モデルとして広く利用されており, 様々な拡張手法が提案されてきた. 以降では, 拡張手法の例の概要を紹介する.

2.2.1 単語埋め込み手法の拡張

Alikaniotis ら [16] は, 答案の得点予測に有効である単語が存在することを指摘し, その課題を解決するために, 新たな単語埋め込みを学習する手法を提案している. 単語埋め込みは Collobert ら [24] の手法を利用することで学習できるが, Alikaniotis らは, 単語が出現する答案の得点に対応した出力を用いてこの手法を拡張し, 問題依存の単語埋め込みを学習させている. これにより, 答案の採点に有益な単語を判別できるため, 得点の予測精度を向上できる.

2.2.2 文書の階層構造の活用

Dong[25] らは, 単語の連結により文が構成され, 文の連結により文書が構成されることに着目し, 二種類の畳み込み層によって階層構造をモデル化した. このモデルでは, 畳み込み層のうち, 一つは単語レベルの特徴量を, もう一つは文レベルの特徴量を学習するように設計されている.

2.2.3 Attention の導入

答案の得点予測の際, 予測に有益な文字・単語・文章に注意して採点することは, 自動採点機の性能の向上につながる. Dong ら [26] は, 有益な文字・単語・文章を識別する Attention 構造 [27] をネットワークに組み込むことを提案している.

2.2.4 文章の一貫性に着目した拡張

一貫性は答案の質を表現するのに重要な特徴量であるため, これを明示的にモデル学習させることで, 自動採点機の性能を向上できると考えられる. そこで, Tay ら [17] は一貫性が類似性と正の相関を持つことを利用し, 異なる LSTM の出力から類似性を演算する層を追加し, 演算結果 (一貫性を表す特徴量) と LSTM から最終的に得られた特徴量の二つを用いて得点を予測することで, 自動採点機の性能を向上させている.

また, Farag ら [19] の手法も Tay ら [17] と同様に一貫性に着目した研究である. Farag らは, LSTM を用いた自動採点機 [15] と一貫性を採点する自動採点機の二つを統合し, 同時に学習することで頑健な自動採点機の開発をしている. 一貫性の学習には, 答案の文の順序をシャッフルし, それらを一貫性のない答案として利用することを提案している.

2.2.5　人手で設計した特徴量の組み込み

特徴量ベースには, 深層学習ベースだけでは表現の難しい特徴を表現できるという利点がある. その利点を活用するために, Dasgupta ら [22] は, 特徴量ベースの特徴量と深層学習ベースの特徴量の二つを同時に利用することで, 自動採点機の性能を向上させている. 具体的には, 答案の単語系列を入力として受け取るネットワークと, 文単位で抽出した特徴量ベースの特徴量を入力とするネットワークの二種類のネットワークにより構成されている.

3　LSTM 自動採点モデル

本研究では, 2.2 で紹介したモデルの中で基礎モデルとして最も広く利用されている LSTM 自動採点モデル [15] を利用する．このモデルは, 答案の単語系列を入力として受け取り, 多層のニューラルネットワークを通して得点の予測値を出力する．モデルの概念図を図 1 に示し，各層の詳細を以降で説明する．

一層目の Word Embeddings 層 (埋め込み層) では, 個々の入力単語を埋め込み表現と呼ばれる潜在的な意味空間にマッピングする．具体的には, 答案集合で利用される語彙数を V とし, $\boldsymbol{w}_t$ を入力単語系列の t 番目の単語 $(t = 1, \cdots, N)$ を表す V 次元のワンホットベクトル，$\boldsymbol{E}$ を $V \times D$ 次元の埋め込み行列とすると，各単語 $\boldsymbol{w}_t$ は $\boldsymbol{E}$ との内積で埋め込み表現 $\boldsymbol{x}_t$ に変換される．具体的には,

$$\boldsymbol{x}_t = \boldsymbol{E}\boldsymbol{w}_t \tag{1}$$

で変換できる．

二層目の LSTM 層では, Word Embeddings の出力 $\boldsymbol{x}_t$ を入力として受け取り, 得点予測に効果的な特徴量を抽出する．LSTM は時系列データを処理する深層学習モデルである RNN の一種であり, 主に記憶セルを表す $\boldsymbol{c}_t$ と三つのゲート (入力ゲートを表す $\boldsymbol{i}_t$, 忘却ゲートを表す $\boldsymbol{f}_t$, 出力ゲートを表す $\boldsymbol{o}_t$) に

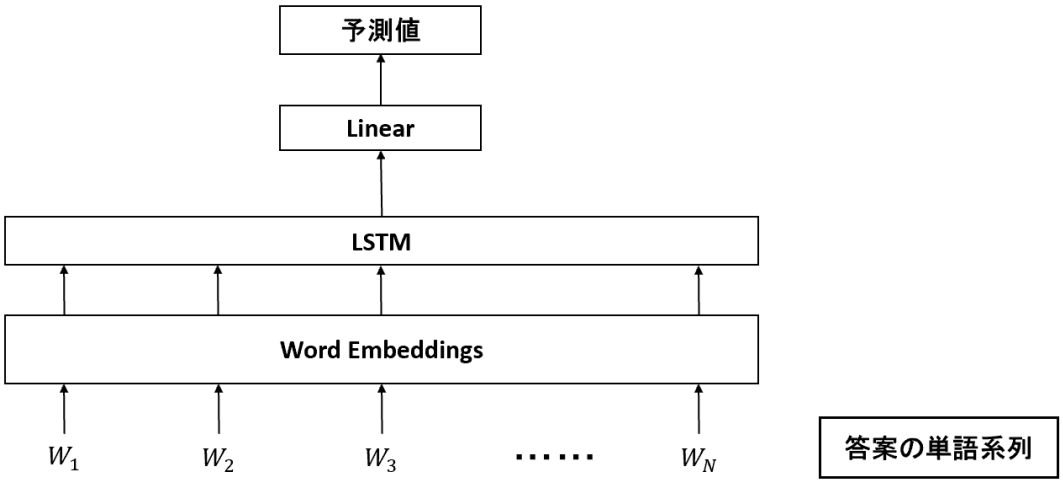

図 1　LSTM モデル

よって構成される．そして $\boldsymbol{c}_t$ が保持する情報を三つのゲートによって調整することで, 情報の長期的な依存関係を学習する．具体的には, 次式の演算により, $\boldsymbol{h}_t$ を出力する．

$$
\begin{aligned}
\boldsymbol{i}_t &= \sigma\ (\boldsymbol{W}_i \boldsymbol{x}_t + \boldsymbol{R}_i \boldsymbol{h}_{t-1} + \boldsymbol{b}_i) \\
\boldsymbol{f}_t &= \sigma\ (\boldsymbol{W}_f \boldsymbol{x}_t + \boldsymbol{R}_f \boldsymbol{h}_{t-1} + \boldsymbol{b}_f) \\
\boldsymbol{o}_t &= \sigma\ (\boldsymbol{W}_o \boldsymbol{x}_t + \boldsymbol{R}_o \boldsymbol{h}_{t-1} + \boldsymbol{b}_o) \\
\boldsymbol{c}_t &= \boldsymbol{i}_t \circ tanh(\boldsymbol{W}_c \boldsymbol{x}_t + \boldsymbol{R}_c \boldsymbol{h}_{t-1} + \boldsymbol{b}_c) + \boldsymbol{f}_t \circ \boldsymbol{c}_{t-1} \\
\boldsymbol{h}_t &= \boldsymbol{o}_t \circ tanh(\boldsymbol{c}_t)
\end{aligned}
\tag{2}
$$

ここで σ や $tanh$ はそれぞれ活性化関数のシグモイド関数とハイパボリックタンジェントを表し, $\boldsymbol{b}_*$ はバイアスを表すベクトルを, $\boldsymbol{W}_*$ や $\boldsymbol{R}_*$ は重み行列を表す．また $\circ$ はアダマール積を表す．

LSTM の出力は個々の単語に対して得られるが, 次の層への入力としては, 最後の単語 $\boldsymbol{w}_N$ に対応する出力 $\boldsymbol{h}_N$ を利用する．本研究ではこの方法を利用するが，他のアプローチとして，出力系列の平均値を使う手法や注意機構（Attention）[26] を利用する方法なども提案されている．

最終層である Linear 層では, 入力として受け取ったベクトルをスカラーで表される得点に変換する．具体的には, LSTM の出力 $\boldsymbol{h}_N$ を線形変換し，その値をシグモイド関数 σ を用いて次式のように変換することで，(0,1) の範囲の予測得点値 $\hat{y}$ を出力する．

$$
\hat{y} = \sigma\ (\boldsymbol{W} \boldsymbol{h}_N + \boldsymbol{b}) \tag{3}
$$

ここで, $\boldsymbol{W}$ は重みを, $\boldsymbol{b}$ はバイアスを意味するパラメータである．

Linear 層の出力はシグモイド関数によって (0,1) の値となるが，実際のデータの得点尺度はこれと異なる場合がある．その場合には，$\hat{y}$ を一次変換し実データの得点尺度に合わせる．例えば, 実際の得点尺度が 0 から C の $C+1$ 段階得点の時, $C\hat{y}$ と変換を行う．

3.1 学習手法

上記のモデルを自動採点に利用するためには, 事前にモデルのパラメータを採点済み答案データから学習する必要がある．モデルの学習は, 深層学習で一般に利用される誤差逆伝播法を用いて行う．誤差逆伝播法は勾配法を利用した

最適化手法であり, 入力データに対する予測値と真値との誤差を任意の損失関数で求め, その誤差が減少するようにパラメータを更新していく．具体的には，$\boldsymbol{W}_{ij}$ を更新対象のパラメータとすると，更新式は次式である．

$$\Delta \boldsymbol{W}_{ij} = -\eta \frac{\partial E}{\partial \boldsymbol{W_{ij}}} \tag{4}$$

ここで E は任意の損失関数で求められる誤差を, η は変更されるパラメータの大きさを表す学習率を意味する．

本章で紹介したモデルの損失関数としては平均二乗誤差 (mean squared error:MSE) が利用されている．具体的には, 答案 k に対する予測値を $\hat{y}_k$, t_k を真値とすると, MSE は次式で求められる．

$$E(\hat{y}_k, t_k) = \frac{1}{N} \sum_{k=1}^{N} (\hat{y_k} - t_k)^2 \tag{5}$$

ただし，問題ごとに得点尺度が異なる可能性があるため, 学習時の尺度を統一するために尺度の正規化を行う．具体的には, t_k を正規化後の得点, y_k を正規化前の得点, y_{min} と y_{max} を正規化前の得点尺度の最小値と最大値とすると, 次式で [0, 1] に正規化される．

$$t_k = \frac{y_k - y_{min}}{y_{max} - y_{min}} \tag{6}$$

3.2 問題点

本章で紹介したような深層学習自動採点モデルは, 採点対象問題ごとに収集された採点済み回答文のデータセットを用いて学習される．具体的には, Q 種類の問題 $q \in \{1, \cdots .Q\}$ が存在し，それぞれの問題に対応する学習データセットを $D_q \in \mathcal{D}= \{D_1, \cdots, D_Q\}$ とすると，問題 q の採点に利用するモデルは D_q を用いて学習する．しかし, 1 章でも述べたように，実際の試験においては, 出題する問題ごとに大量の採点済み答案データを事前に用意することは難しい場合が多いといえる．

この問題を解決する最も単純なアプローチとしては, 出題する問題に関する少量の採点済み答案データと過去に出題・採点された問題に対する大量の採点済み答案データを同時に用いて自動採点モデルを学習する方法が考えられる. 具体的には, 問題 q の採点に利用するモデルを，データ D_q を含む全てのデータ $\mathcal{D}$ を用いて学習するという方法である．しかし, この方法では, 対象問題に固有の特徴をモデルに反映することが難しいため，対象の問題が他の問題と異なる特徴を持つ場合に予測精度が低下すると予測できる.

4　提案手法

前節で挙げた問題点を解決するために, 本研究では, 問題非依存の特徴量と問題依存の特徴量を同時に学習できる新たな深層学習自動採点モデルを提案する.

図 2 にモデルの概略図を示す. 本研究では, 答案の特徴量を, 問題に依存しない共通の特徴量 S_C と, 問題に依存する特徴量 S_P の二つに分けられると仮定し, 提案モデルを全ての問題の答案データを処理するネットワークと問題ごとに処理するネットワークで構成される深層学習モデルとして定式化する. 具体的な各層の詳細を以降で説明する.

一層目の Word Embeddings 層では, $\boldsymbol{w}_{tq}$ を問題 q の入力単語系列の t 番目の単語を表す $\boldsymbol{V}$ 次元のワンホットベクトルとすると, 前章の (1) 式と同様に, 個々の入力単語が埋め込み表現 $\boldsymbol{x}_{tq}$ に変換される.

$$\boldsymbol{x}_{tq} = \boldsymbol{E}\boldsymbol{w}_{tq} \tag{7}$$

二層目では, 問題が Q 種類の場合, Word Embeddings の出力先に $Q+1$ 種類の LSTM を用意する. 一つは, $\boldsymbol{S}_C$ を出力する $LSTM_C$ であり, 全問題に対

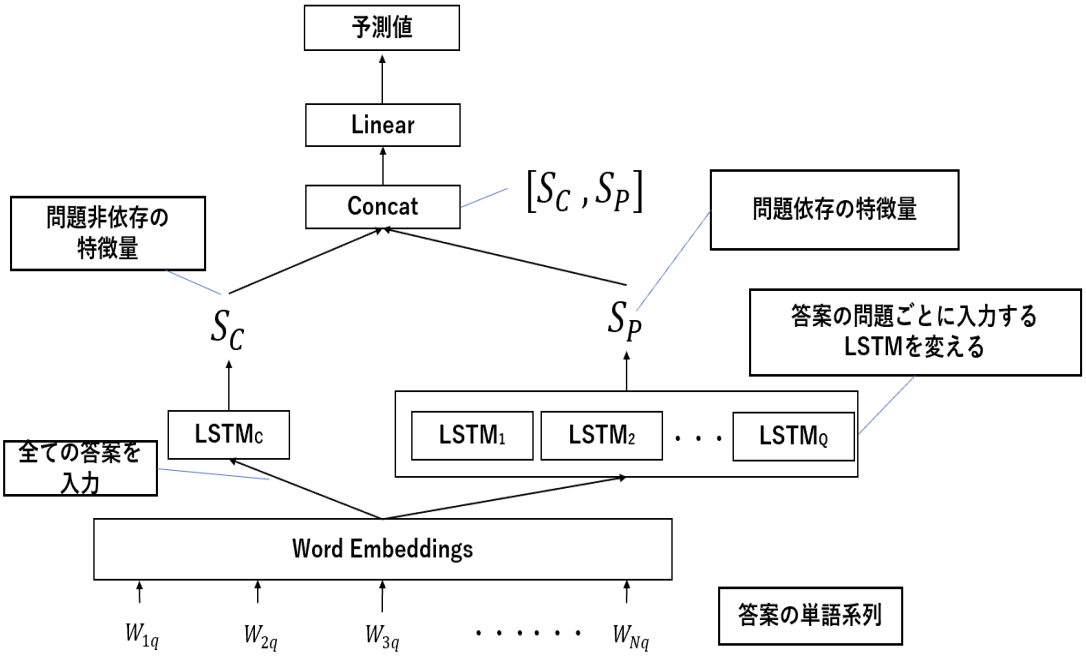

図 2　提案モデル

する答案データを入力する．それに対し, 残りの Q 種類の LSTM は, $\boldsymbol{S}_P$ を出力する $LSTM_q$ であり, それぞれ対応づけられている問題の答案データのみを入力する．ここで, Word Embeddings の出力 $\boldsymbol{x}_{tq}$ を入力として受け取った LSTM の演算を $\boldsymbol{h}_{tq} = LSTM(\boldsymbol{x}_{tq})$ と定義する．$\boldsymbol{h}_{tq}$ は出力を表す.

前章と同様に, LSTM の出力のうち, 次の層への入力としては, 最後の単語 $\boldsymbol{w}_{Nq}$ に対応する出力 $\boldsymbol{h}_{Nq}$ を利用する．ここで, $LSTM_C$ と $LSTM_q$ は, 次式で問題非依存の特徴量 $\boldsymbol{S}_C$ と問題依存の特徴量 $\boldsymbol{S}_P$ に変換される.

$$\boldsymbol{S}_C = LSTM_C(\boldsymbol{x}_{Nq}) \tag{8}$$
$$\boldsymbol{S}_P = LSTM_q(\boldsymbol{x}_{Nq}) \tag{9}$$

このようにモデルを構成することにより, 全答案データを処理するネットワークでは問題非依存の特徴量 $\boldsymbol{S}_C$ を学習し, 問題ごとのネットワークでは問題依存の特徴量 $\boldsymbol{S}_P$ を学習することが期待できる.

最後は, 二つの特徴量を Concat 層で統合 (Concatenation) し, Linear 層でスカラーの値に対応させ, 採点予測を行う. 具体的には, $\boldsymbol{G}_1$ 次元の $\boldsymbol{S}_C$ と $\boldsymbol{G}_2$ 次元の $\boldsymbol{S}_P$ を統合し, $\boldsymbol{G}_1 + \boldsymbol{G}_2$ 次元の特徴量 $h'_N = [\boldsymbol{S}_C, \boldsymbol{S}_P]$ に変換する．その後, h'_N を入力とし, 前章の (3) 式と同様に以下の演算を行い, (0,1) の範囲の予測得点値 $\hat{y}$ を出力する.

$$\hat{y} = \sigma\left(\boldsymbol{W}'\boldsymbol{h}'_N + \boldsymbol{b}'\right) \tag{10}$$

ここで, $\boldsymbol{W}'$ と $\boldsymbol{b}'$ は, それぞれ重みとバイアスを意味するパラメータである. つまり, 提案モデルでは, $\boldsymbol{W}'$ によって問題非依存の特徴量と問題依存の特徴量の得点予測への重みも最適化されることになる.

4.1 学習手法

提案モデルの学習は, 前章の手法と同様に, 得点尺度を [0,1] に調整したデータ $\mathcal{D}$ を用いて誤差逆伝播法で行う．損失関数には, MSE を利用する.

5　評価実験

本章では, ベンチマークデータセットを用いて, 提案手法の有効性を評価する.

5.1　データ

本実験の実データには, Kaggle が公開している Automated Student Assessment Prize(ASAP)[*1] と呼ばれる英語のエッセイデータセットを用いる．ASAP は自動採点の性能評価に広く用いられているベンチマークデータセットである．データは，8 種類のエッセイ問題に対するエッセイ文とそれらに対する得点で構成されている．表 2 にデータの基礎統計量を示し，表 3 に各問題の内容を示す．表 2 からデータ数や平均単語数は問題ごとに異なっており，得点の尺度も 4 段階から 60 段階まで多様であることがわかる．また，問題のタイプは，大きく課題文型とテーマ型の二種類に分けられることがわかる．

課題文型では, 論述の材料となる詳細な資料が与えられ, 受験者はその内容に基づいてエッセイを書くことが求められる．ASAP では, 問題 3,4,5,6 が課題文型となっている．一方，テーマ型では, 論述の材料となる詳細な資料は与えらず，与えられたテーマに関して自身の意見を論述する．ASAP では, 問題 1,2,7,8 がテーマ型である.

表 2　ASAP データセット

問題番号	データ数	得点尺度	問題タイプ	平均単語数
1	1783	2-12	テーマ型	350
2	1800	1-6	テーマ型	350
3	1726	0-3	課題文型	150
4	1772	0-3	課題文型	150
5	1805	0-4	課題文型	150
6	1800	0-4	課題文型	150
7	1569	0-30	テーマ型	250
8	723	0-60	テーマ型	650

[*1] https://www.kaggle.com/c/asap-aes

表 3　ASAP の問題内容

問題番号	内容
1	地元の新聞の読者に, コンピュータが人々に与える影響についての考えを納得させることを求めている.
2	自身の経験を踏まえ, 特定のメディアを図書館から禁止するべきかどうかについての意見を求めている.
3	自転車で旅行している人に関する長文を踏まえ, その環境が語り手にどのような影響を与えるかを説明することを求めている.
4	テキストの最終段落を指し, 著者がこの段落で物語を結論づけた理由を説明するよう求めている.
5	与えられた回想録から著者の気持ちを説明することを求められている.
6	テキストに基づき, ある建設業者が直面した障害について説明するように求められている.
7	”忍耐”について論述することを求めている.
8	笑いがあった実話を書くことを求めている.

5.2　得点予測精度の評価

本節では, 提案モデルの得点予測精度を評価する. 具体的には, 3 章で紹介した手法を用いて, 対象問題のみのデータで学習した自動採点機を既存モデル (対象問題), 8 つの全ての問題のデータで学習した自動採点機を既存モデル (全問題) とし, この二つのモデルと提案モデル間の得点予測精度を比較する.

得点予測精度の評価には, 自動採点の評価指標で最も一般的な二次重みカッパ係数 (quadratic weighted kappa coefficient:QWK) を用いた. 二次重みカッパ係数は, 実数値と予測値の一致度を測る評価指標である. -1 に近いと一致度は低く, 1 に近いと一致度は高い. 本研究では, Python のライブラリの一つである scikit-learn[*2]により二次重みカッパ係数を求めた.

なお実験は二分割交差検証で行い, ニューラルネットワークのフレームワー

[*2] https://scikit-learn.org/stable/index.html

クの一つである Chainer[*3]により LSTM や Linear を実装した．また本研究の深層学習モデルのハイパーパラメータは, 先行研究 [15] に合わせ, LSTM の出力のパラメータ数は 300 に, Linear 出力のパラメータ数は 1 にしてシグモイド関数と組み合わせた．また各 LSTM の出力にはドロップアウト率 0.5 を実装し, エポック数は 50, ミニバッチサイズは 32 とした．Word Embeddigs も先行研究と同様に, Zou ら [28] が公開している次元数 V が 50 の事前学習された埋め込み層 (Word Embeddings) を用いた．検証データサイズは各問題 100 とし, パラメータの最適化アルゴリズムには Adam[29] を利用した．

表 4　二次重みカッパ係数による比較

問題番号	提案モデル	既存モデル (全問題)	既存モデル (対象問題)
1	**0.714**	0.671	0.576
2	**0.554**	0.491	0.358
3	**0.746**	0.741	0.718
4	**0.722**	0.720	0.720
5	0.722	0.735	**0.736**
6	0.655	0.697	**0.704**
7	0.644	0.554	**0.645**
8	**0.308**	0.165	0.141
平均	**0.633**	0.598	0.575

実験結果を表 4 に示した．表 4 では, 問題ごとに精度が一番よいものを太字で示した．

各問題ごとにモデルの性能を比較すると, 既存モデル (全問題) が既存モデル (対象問題) よりも高い精度を示す問題については, 提案モデルも既存モデル (対象問題) の精度を上回っている．具体的には, 問題 1,2,3,8 でその傾向が確認できる. 既存モデル (全問題) と提案モデルは大量の過去データから問題非依存の特徴量を学習できるという利点を有しており，このことが性能向上に寄与したと解釈できる．

また, 提案モデルと既存モデル (全問題) を比較したとき, 提案モデルが高い性能を示した問題 1,2,7,8 の共通点として, 問題タイプがテーマ型であることが

[*3] https://chainer.org/

挙げられる．一方, 提案モデルの性能が同程度あるいは低い場合の問題 3,4,5,6 の問題タイプは, 課題文型である．二つの問題タイプの違いは, 論述の材料となる詳細な資料の有無である. 資料が与えられる課題文型は, エッセイの書き方が資料の文章に影響されやすいのに対し, テーマ型は, エッセイ執筆の自由度が高いため, 課題文型より多様性があり, 問題非依存の特徴量だけでは表現しきれないと推測できる．このため, テーマ型の問題では, 問題依存の特徴を柔軟に表現できる提案手法の精度が大きく向上したと考えられる．

また表 4 から, 精度の平均は提案モデルが最も高いことが確認できる．ここで, 提案モデルと既存モデル (対象問題), 既存モデル (全問題) との間の二次重みカッパ係数の平均に有意な差があるかどうかの確認のため, 対応のある t 検定を行った．表 5 は t 検定の結果である．表 5 より, 提案モデルは他モデルと 10% で有意に精度が高いことが確認できる．

表 5　t 検定の結果

提案モデル vs 既存モデル (対象問題)	提案モデル vs 既存モデル (全問題)
p= 0.059	p= 0.072

5.3　問題依存の特徴量の有効性評価

本節では, 提案モデルが，既存モデル (全問題) とは異なり, 問題依存の特徴量を学習できているかを評価する．このために, 図 2 における $\boldsymbol{S}_P$ のみでの採点予測の精度を評価した． 具体的には, 問題非依存の特徴量を使わずに採点予測するために, $\boldsymbol{S}_C = [0, 0, \cdots, 0]$ で固定した上で $\boldsymbol{S}_P$ と統合し, それぞれの問題に対する採点精度を評価した.

実験結果を表 6 に示す．表 6 の n 行 m 列目は, 問題 n の答案を, m 番目の問題に関する答案データから学習した $LSTM_m$ で採点予測した場合の精度を表す．表 6 より, 精度は学習と予測を同じ問題のデータで行った時の精度が高く, それ以外では低い傾向であることを確認できる．具体的には, $LSTM_1$ から $LSTM_7$ は, それぞれ対応する問題の場合の精度が最も高い．また $LSTM_8$ は, 問題 8 の時が二番目に高い精度である．このことから, 提案モデルは, 問題依存の特徴量をそれぞれ学習出来ていると考えられる．

以上の実験から, 提案手法により, 問題非依存の特徴量と依存する特徴量の二

表 6　問題依存の特徴量の有効性評価

$LSTM_q$ / 問題番号	1	2	3	4	5	6	7	8
1	0.335	0.050	0.139	0.082	-0.019	0.053	-0.023	-0.015
2	0.068	0.088	0.063	0.012	0.005	0.063	-0.034	0.023
3	0.149	0.043	0.569	0.113	-0.035	0.032	-0.002	-0.009
4	0.140	0.022	0.216	0.505	-0.003	0.047	-0.005	-0.010
5	0.196	0.047	0.432	0.154	0.379	0.027	0.015	0.000
6	0.094	0.053	0.042	0.013	-0.031	0.252	0.009	-0.007
7	0.071	0.030	0.257	0.110	-0.019	0.015	0.288	0.010
8	0.007	-0.005	0.055	0.032	0.008	0.009	0.147	0.013

つを学習し, 問題に依存せずに学習することが確認できた.

6　まとめと今後の課題

自動採点は世界的に数多くの研究が行われている．これまでは, 事前に人手で設計した特徴量を使うアプローチが主流であった．しかし, この手法では, 高精度を実現するためには十分な特徴量チューニングが必要となる．この問題を解決するアプローチとして, 近年, 人手での特徴量設計を必要としない深層学習ベースの自動採点手法が注目されている．

深層学習自動採点モデルを利用するためには, 事前に収集された採点済み答案のデータセットを用いてモデルを学習しておく必要がある．深層学習モデルの学習には一般に大量のデータが必要となる．しかし, 実際の試験においては, 出題する問題ごとに大量の採点済み答案データを用意することは難しい場合が多い．

この問題を解決する最も単純なアプローチとしては, 出題する問題に関する少量の採点済み答案データと過去に出題・採点された問題に対する大量の採点済み答案データを同時に用いて自動採点モデルを学習する方法が考えられる．しかし, この方法では, 対象問題に固有の特徴をモデルに反映することが難しいため, 対象の問題が他の問題と異なる特徴を持つ場合に予測精度が低下すると予測できる．

この問題を解決するために, 本研究では答案の特徴量は, 個別の問題に依存しない共通の特徴量と問題に依存する固有の特徴量の二つから構成されていると仮定し, 複数の問題の答案データから二つの特徴量を同時に学習できる新たな深層学習自動採点モデルを提案した．具体的には, 全ての問題の答案データを処理するネットワークと問題ごとに処理するネットワークの二種類のネットワーク構造を内部的に有する深層学習モデルとして定式化した．そして評価実験より, 提案手法が, 従来手法とは異なり, 問題に依存しない特徴量を学習でき, 精度向上に有効である傾向が確認できた．また, 提案手法が, 従来手法とは異なり, 複数の問題の答案データを用いた学習においても, 問題に依存する特徴量を学習できることが確認できた．

提案モデルには問題非依存の特徴量があるため, 訓練データがない, もしくは少量しかない問題に対する自動採点の成果も期待できる．本研究では，この点を直接には評価できなかったが，今後はこの実験も行なって提案の有効性を評価したい．また, 本実験では, 問題に依存しない特徴量と依存する特徴量の組み

合わせの比重は, 問題に関わらず同じ比重で行った. この比重を問題ごとに最適化出来れば, 更なる精度の向上が見込める. 今後はこの実験も行って, 提案の性能を向上させたい.

謝辞

本論文の作成にあたり, 丁寧かつ熱心なご指導を頂いた植野真臣教授, 宇都雅輝助教に感謝の意を表します.

参考文献

[1] 中央教育審議会. ”新しい時代にふさわしい高大接続の実現に向けた高等学校教育, 大学教育, 大学入学者選抜の一体的改革について (答申)”. 文部科学省.2014

[2] 宇都雅輝, 植野真臣. ”パフォーマンス評価のための項目反応モデルの比較と展望”. 2016. 日本テスト学会誌, Vol.12, No.1, pp.55-75.

[3] 河原 宜央. ”国語科の評価問題における記述式問題の採点過程に関する研究 採点基準と採点答案の分析を通して”. 広島県立教育センター. 2017

[4]Torsten Zesch, Michael Wojatzki, Dirk Scholten-Akoun.2015.”Task-Independent Features for Automated Essay Grading”Association for Computational Linguistics.Pages:224 ー 232

[5]Peter Phandi, Kian Ming A. Chai, Hwee Tou Ng. Association for Computational Linguistics. 2015.”Flexible Domain Adaptation for Automated Essay Scoring Using Correlated Linear Regression”Pages:431 ー 439

[6]Ronan Cummins, Meng Zhang, Ted Briscoe. 2016.”Constrained Multi-Task Learning for Automated Essay Scoring”Association for Computational Linguistics.Pages:789 ー 799

[7]Hongbo Chen and Ben He. ”Automated Essay Scoring by Maximizing Human-machine Agreement”. 2013. Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pages 1741 ー 1752, Association for Computational Linguistics

[8]Huy V. Nguyen and Diane J.Litman.”Argument Mining for Improving the Automated Scoring of Persuasive Essays”In Proc. of AAAI, pages 5892 ー 5899, 2018.

[9]Ellis B Page. ”The imminence of Grading Essays by Computer” The Phi Delta Kappan, 47(5):238–243. 1966

[10]Peter W. Fottz, Lynn A. Streeter, Karen E. Lochbaum, and Thomas K Landauer.”Implementation and Applications of the Intelligent Essay Assessor” In Handbook of Automated Essay Evaluation: Current Application and New Directions, pp. 68–88. Routledge, 2013

[11]Lawrence M. Rudner and Tahung Liang. ”Automated Essay

ScoringUsing Bayes' Theorem" The Journal of Technology, Learning, and AssessmentVolume 1, Number 2 · June 2002

[12] Lawrence M. Rudner, Veronica Garcia, and Catherine Welch. "An Evaluation of the IntelliMetric Essay Scoring System" The Journal of Technology, Learning, and AssessmentVolume 4, Number 4 · March 2006

[13]Yigal Attali and Jill Burstein."Automated Essay Scoring With e-rater® V.2", The Journal of Technology, Learning, and AssessmentVolume 4, Number 3 · February 2006

[14]Sepp Hochreiter and Jurgen Schmidhuber. 1997. "LONG SHORT-TERM MEMORY". Neural Computation, 9(8):1735 — 1780.

[15]Kaveh Taghipour and Hwee Tou Ng. 2016. "A Neural Approach to Automated Essay Scoring. Association for Computational Linguistics" Conference on Empirical Methods in Natural Language Processing, pages 1882 — 1891.

[16]Dimitrios Alikaniotis, Helen Yannakoudakis, Marek Rei."Automatic Text Scoring Using Neural Networks"arXiv:1606.04289v2 [cs.CL] 16 Jun 2016

[17]Yi Tay, Minh C. Phan, Luu Anh Tuan, Siu Cheung Hui."SkipFlow: Incorporating Neural Coherence Features for End-to-End Automatic Text Scoring" arXiv:1711.04981v1 [cs.AI] 14 Nov 2017

[18]Farjana Sultana Mim, Naoya Inoue, Paul Reisert, Hiroki Ouchi, Kentaro Inui. 2019."Unsupervised Learning of Discourse-Aware Text Representation"Association for Computational Linguistics.Pages:378 — 385

[19]Youmna Farag, Helen Yannakoudakis, Ted Briscoe. 2018. "Neural Automated Essay Scoring and Coherence Modeling for Adversarially Crafted Input" Association for Computational Linguistics. Pages: 263 — 271

[20]Jiawei Liu, Yang Xu, Lingzhe Zhao. "Automated Essay Scoring based on Two-Stage Learning" arXiv:1901.07744v1 [cs.CL] 23 Jan 2019.

[21]Yucheng Wang, Zhongyu Wei, Yaqian Zhou, Xuanjing Huang," Automatic Essay Scoring Incorporating Rating Schema via Reinforcement Learning" ,Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 791 — 797 Brussels, Belgium, October 31 -

November 4, 2018.

[22]Tirthankar Dasgupta, Abir Naskar, Rupsa Saha and Lipika Dey. "Augmenting Textual Qualitative Features in Deep Convolution Recurrent Neural Network for Automatic Essay Scoring" 2018.Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications, pages 93 — 102.Association for Computational Linguistics.

[23]Cancan Jin, Ben He, Kai Hui and Le Sun. 2018."TDNN: A Two-stage Deep Neural Network for Prompt-independent Automated Essay Scoring" Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Long Papers), pages 1088 — 1097 Melbourne, Australia, July 15 - 20, 2018.

[24]Ronan Collobert and Jason Weston. "A unified architecture for natural language processing: Deep neural networks with multitask learning" In Proc. of ICML, pages 160–167, 2008.

[25]Fei Dong, Yue Zhang. "Automatic Features for Essay Scoring — An Empirical Study" 2016. Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 1072 — 1077,Association for Computational Linguistics.

[26]Fei Dong, Yue Zhang, Jie Yang. "Attention-based Recurrent Convolutional Neural Network for Automatic Essay Scoring" 2017. Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017), pages 153 — 162, Association for Computational Linguistics.

[27]Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. "Sequence to sequence learning with neural networks" In Proc. of NIPS, pages 3104–3112, 2014.

[28]Will Y. Zou, Richard Socher, Daniel Cer and Christopher D.Manning. 2013 ."Bilingual Word Embeddings for Phrase-Based Machine Translation"Association for Computational Linguistics. Pages:1393 — 1398

[29]Diederik P. Kingma, Jimmy Ba. 2014. "Adam: A Method for Stochastic Optimization"arXiv:1412.6980 [cs.LG]