採点基準を利用した記述式答案の自動採点

王 天奇 [1,2]　井之上 直也 [1,2]　水本 智也 [2]　大内 啓樹 [2,1]　乾 健太郎 [1,2]
[1] 東北大学　[2] 理化学研究所 AIP センター
{outenki,naoya-i,inui}@ecei.tohoku.ac.jp
{tomoya.mizumoto,hiroki.ouchi}@riken.jp

1 はじめに

Short Answer Grading (SAG) は、試験などの記述式問題に対する学生の解答の正しさを自動的に評価するタスクである [8]。教育の分野では、特に教師の数が限られる場合に、SAG は非常に有益な技術である [9]。SAG の先行研究では、機械学習を用いた様々な手法が提案されてきた。典型的なアプローチは、学生の解答と教師が採点した解答のスコアのペアを用いて、回帰モデルを学習するものである。近年、他の自然言語処理のタスクと同様に、ニューラルネットワークを用いた、特徴量学習に基づく手法の有効性が示されている [12]。

人間の採点者は、設問に対する採点基準に基づいて学生からの解答に点数を付けるため、人間の採点者にとって採点基準は重要な情報である。採点基準の例を図 1 に示す。この設問は、学生にタンパク質合成の過程 (steps involved in protein synthesis) を答えるよう要求するものである。採点基準 (Rubric) では、key elements と呼ばれる、解答における重要な要素が定義され（例えば *mRNA exits nucleus via nuclear pore.*）、最終的な解答の点数は、解答の中に存在する key elements の数によって定まるよう規定している。しかしながら、これまでの SAG の先行研究では、採点基準の情報の有効性が検証されてこなかった。

本研究の貢献は、以下の通りである:

- 採点基準の情報をニューラル SAG モデルに組み込む手法を提案した最初の研究である。
- 既存のニューラル SAG モデルを、採点基準の情報を参照するコンポーネントで拡張する、一般的なフレームワークを提案する。

2 先行研究

SAG の研究コミュニティの主な関心は、解答の特徴表現と解答間の類似性測度を探求することにある。これまで、Latent Semantic Analysis (LSA) [8]、編集距離ベースの類似度、WordNet を用いた知識ベースの類似度、および単語分散表現ベースの類似度など、さまざまな手法が検討されてきた。近年は、ニューラルネットワークに基づく特徴表現学習が SAG に有効であることが報告されている [12]。

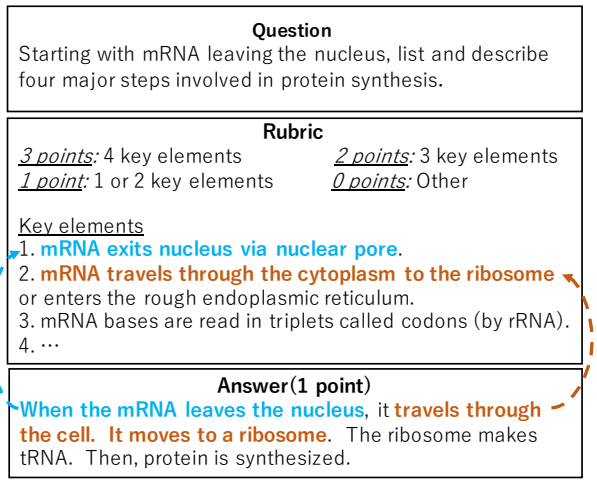

図1: ASAP-SAS データセットにおける設問と採点基準の一例。

これに対して、本研究は SAG における採点基準の情報の有効性を探求するものである。Sakaguchi ら [13] は、手作業で作成された解答の特徴量テンプレート（例えば、単語 n-gram などの疎な離散特徴量）と採点基準（解答と採点基準項目の間の、BLEU [10] に基づく類似度）が SAG に有効であることを示した。しかしながら、(i) 近年有効とされている、ニューラルネットワークに基づく特徴量学習パラダイムのもとで、採点基準の利用が有効か、(ii) 採点基準の情報を効率的に利用するために、どのようなニューラルネットワークのアーキテクチャを採用すべきか、は自明でない。

低リソース設定における SAG の研究がいくつかある。Heilman ら [4] は、手作業で作成された疎な離散特徴量を用いた非ニューラル SAG モデルに対する訓練データのサイズの重要性を調査している。Horbach ら [5] は、アクティブ・ラーニングにより、SAG の訓練事例を効率的に獲得できることを示している。これらの研究では採点基準の有効性は検証されておらず、本研究とこれらの研究を組み合わせることで、さらに高精度な自動採点を行うことが期待できる。

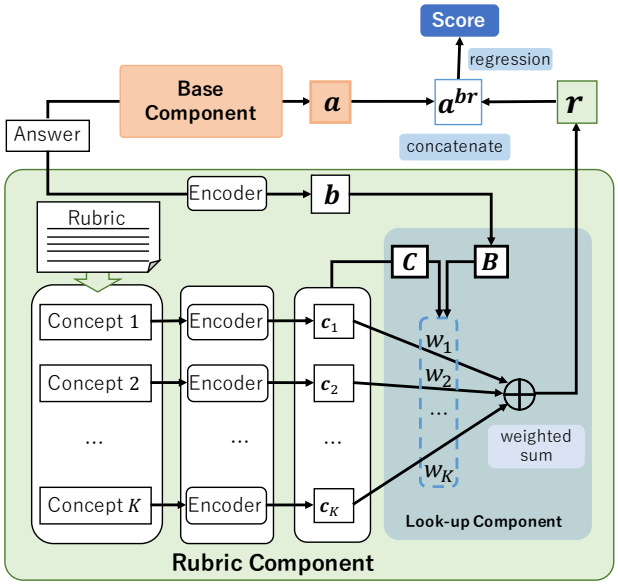

図2: 採点基準の情報を用いる提案モデルのアーキテクチャ。

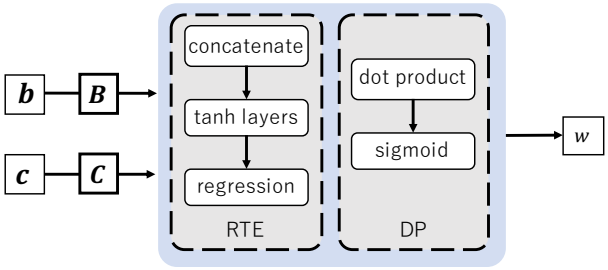

図3: RTE また内積より Key concept に対するアテンションの計算。$\boldsymbol{b}$ と $\boldsymbol{c}$ は解答と key concept のベクトルであり、$B \in \mathbb{R}^{2h \times M}$ と $C \in \mathbb{R}^{2h \times M}$ はモデルによって学習される変換行列であり、w はアテンションである。

3 提案モデル

3.1 キーアイデア

図 1 は、解答に必要な情報を *key concept* として規定している。訓練事例における各学生の解答は、複数の key concept の組み合わせとみなすことができる。したがって、SAG モデルの訓練時には、モデルはそれぞれの解答に含まれる key concept を認識し、key concept の点数への寄与度合いを学習することが必要である。しかし、訓練事例が十分にない状況（低リソース設定）では、このような学習は困難と考えられるため、採点基準の使用による性能の向上が期待できる。

本研究のキーアイデアを図 2 に示す。まず、学生の解答を特徴ベクトル $\boldsymbol{a}$ に変換し、$\boldsymbol{a}$ から解答の点数を推定する**ベースコンポーネント**（Base Component）を仮定する。次に、**採点基準コンポーネント**（Rubric Component）を構築する。与えられた採点基準は、自然言語で一連の key concept を規定していると仮定する。採点基準コンポーネントは、採点基準に規定されている key concept を格納し、学生の解答に関連する key concept を探す。このために、メモリネットワーク [15] を用いて採点基準コンポーネントを実現する。まずは、文エンコーダ（Encoder）を用いて各 key concept を特徴ベクトルに変換する。次に、アテンションに基づく**ルックアップコンポーネント**（Look-up Component）を使用し、所与の解答に関連する key concept （例えば、解答に含まれる key concept）のみをピックアップし、選択された key concept の情報を集約し、特徴ベクトルを生成する。

最後に、学生の解答の特徴ベクトルと集約された key concept の特徴ベクトルの両方を利用し、点数を予測する。提案モデルの利点は、訓練において、モデルが学生の解答に含まれる key concept を暗黙的に推定するため、key concept の明示的なアノテーションが不要である点である。また、提案するフレームワークはエンコーダに依存しない点も利点として挙げられる。すなわち、固定長の特徴ベクトルを生成する任意のエンコーダであれば何でも、採点基準の文エンコーダとして利用することができ、文エンコーダの分野の研究成果を容易に取り入れることができる。

3.2 ベースコンポーネント

ベースコンポーネントとしては、公開されたモデルの中で state-of-the-art の SAG システムであるニューラル SAG モデル [12] を利用する。

このモデルは、三つのレイヤからなる。解答 $a = (x_1, x_2, ..., x_n)$（x_i は a の i 番目の単語）が与えられたとき、Embedding レイヤは、各単語のベクトル $\boldsymbol{x}_i \in \mathbb{R}^D$ を出力する。次に、Bi-directional Long Short-Term Memory (BiLSTM) [14] レイヤは、単語の文脈を加味した単語のベクトル表現 $\boldsymbol{a}_i = [\overrightarrow{\boldsymbol{h}_i}; \overleftarrow{\boldsymbol{h}_i}]$ を生成する。ここで、$\overrightarrow{\boldsymbol{h}_i} \in \mathbb{R}^h$ と $\overleftarrow{\boldsymbol{h}_i} \in \mathbb{R}^h$ は、それぞれ BiLSTM によって生成された順方向と逆方向の隠れ状態である。

最後に、これらのベクトルに mean-over-time 関数 mot を適用し、解答の特徴量を得る: $\boldsymbol{a} = \mathrm{mot}(\boldsymbol{a}_1, \boldsymbol{a}_2, ..., \boldsymbol{a}_n) \in \mathbb{R}^{2h}$。

3.3 採点基準コンポーネント

図 1 のように、所与の解答に含まれる key concept を特定したい。このために、key concept の特徴ベクトル $\boldsymbol{c}_1, \boldsymbol{c}_2, ..., \boldsymbol{c}_K \in \mathbb{R}^{2h}$ へのアテンションを、解答の特徴ベクトルと照らし合わせながら計算する。図 3 に示すように、本研究ではアテンションを計算するために以下の二つの手法を用いる。

- Recognizing Textual Entailment（RTE）タスクと見なし、[1] で提案された手法を利用する (図 3 の RTE)。B と C によって fine tuning した解答と key concept の特徴量（$\boldsymbol{b}$ と $\boldsymbol{c}$）を concatenate し、複数の tanh レイヤーと回帰レイヤーにより、アテンションを計算す

表1: 訓練データのサイズを変化させた場合の性能（QWK）。

データサイズ	13%	25%	50%	100%
ベース	.564	.627	.697	.732
＋採点基準（RA）	.557	.623	.687	.720
＋採点基準（RW）	**.583**	**.639**	.695	.732
＋採点基準（DA）	.577	**.639**	.699	.733
＋採点基準（DW）	.572	**.639**	**.705**	**.736**

表2: 13% の訓練データを利用する場合の各設問に対する提案モデルの QWK（上）と RMSE（下）。

設問	1^K	2^K	3^A	4^A	5^K	6^K	7^A	8^A	9^A	10^K
ベース	.675 .743	.426 .892	.603 .522	.524 .539	.600 .509	**.650** .544	**.524** .727	.351 .828	.644 .576	.639 .475
+RA	.659 .733*	.403 .888	.569 .531	.472 .565	.672 .438	.709 .448	.496 .718	.322 .831	.623 .576	.642 .471
+RW	**.687** .703*	.424 .888	**.641** .509*	.481 .528	**.689** .427	**.688** .454	.515 .697*	.382 .789*	.668 .557	.654 .462
+DA	.679 .725*	.432 .869*	.630 .491*	**.540** .513*	.575 .492*	.620 .494*	.513 .698*	**.397** .797	**.703** .522*	**.681** .464
+DW	.680 .722	**.471** .872*	.633 .512*	.485 .538	.595 .467	.623 .480*	.520 .694	.378 .812	.672 .547*	.667 .467

る。回帰レイヤーの活性化関数に sigmoid 関数を用いる。RTE の各レヤーは SNLI [1] データセットで事前学習し、訓練フェーズで fine tuning する。

- 学習するパラメーターを減らすため、図 3 の DP のように、内積によってアテンションを計算する: $w_i = \text{sigmoid}(\boldsymbol{b}B \cdot (\boldsymbol{c}_i C)^\top)$。正規化関数としては、解答と関連のある key concept がない可能性があるため、softmax 関数ではなく、sigmoid 関数を使用する。

最後に、採点基準の特徴ベクトル $\boldsymbol{r}$ を計算する。図 1 のように、key concept ベクトルの重み付き平均を計算する: $\boldsymbol{r} = \frac{1}{K}\sum_{i=1}^{K} w_i \boldsymbol{c}_i, \boldsymbol{r} \in \mathbb{R}^{2h}$。学習するパラメータを減らすため、計算されたアテンションを $\boldsymbol{r}$ として使うのも手法の一つである：$\boldsymbol{r} = [w_1, w_2, ..., w_K], \boldsymbol{r} \in \mathbb{R}^K$。本研究では両方を使い、性能を比較する。

$\boldsymbol{c}_i$ と $\boldsymbol{b}$ を得るための文エンコーダとしては、さまざまな選択肢がある [3, 7, etc.]。本研究では、訓練済みの Universal Sentence Encoder [2] を用いる。[*1]

4 評価実験

4.1 データセット

本研究では、SAG の分野で広く利用されている ASAP-SAS[*2]を用いる。このデータセットには 10 題の設問が含まれており、それぞれの設問に詳細な採点基準が添付されている。各設問に対する解答数の平均は、2,225 である。解答の点数の値域は設問によって異なるが、三段階（0 点から 2 点まで）または四段階（0 点から 3 点まで）である。このうち、10% を開発データ、10% をテストデータ、80% を訓練データとして用いる。

設問 1、2、5、6、10 の採点基準には、採点における重要な要素である key elements が含まれており（例えば、図 1）、これを key concept として用いる。それ以外の設問については、key elements が与えられておらず、受験者に文章（問題文）を読ませて、解答とその根拠を問題文から抽出する設問である。このため、問題文のそれぞれの文を key concept として利用する。

4.2 実験設定

提案モデルは、各設問ごとに訓練する。まず、ベースコンポーネントを訓練し、次にモデル全体を訓練する。この際、ベースコンポーネントのパラメタは固定する。

単語ベクトルについては、Wikipedia と Gigaword5 上で学習された 300 次元 ($D = 300$) の GloVe 埋め込み [11][*3]により初期化し、訓練時にアップデートする。ハイパーパラメタについては、$h = 256$、Dropout Probability $= 0.5$ を用いる。採点基準コンポーネントについては、$M = 512$ を用いる。B と C を単位行列より初期化する。

RTE を利用場合に、M を 100 に設置し、B と C をランダムで初期化する。tanh レヤーを一つ用いる。SNLI のデーターは「neutral」、「contradiction」と「entailment」三つのラベルがある。本研究では「entailment」を 1 に設置して、その以外のラベルを 0 に設置する。

損失関数としては、Mean Squared Error (MSE) を用い、最適化には Adam [6]（学習率 0.001、バッチサイズ 32）を用いる。訓練は 50 epochs 実施し、開発データの上で最良のモデルを選択する。[*4] 訓練データの点数は、訓練時には $(0, 1)$ の範囲に正規化する。

モデルの性能は、SAG のコミュニティで標準的な評価指標である Quadratic Weighted Kappa (QWK) を用いて評価する。初期値のランダム性を考慮するために、0 から 5 までの初期ランダムシードを用いて実験を 6 回繰り返し、平均性能を最後の結果として評価する。低リソース設定における性能を評価するため、訓練データを様々なサイズに変化させてモデルを訓練・評価する。

4.3 結果と分析

異なるサイズのデータセットを用いて、全部の問の平均性能をを表 1 に示す。「ベース」はベースコンポーネントだけを使う場合の性能に対応し、「＋採点基準」は

[*1] https://tfhub.dev/google/universal-sentence-encoder/2

[*2] https://www.kaggle.com/c/asap-sas/data

[*3] https://nlp.stanford.edu/projects/glove/

[*4] Keras (https://keras.io/) Tensorflow Backend を用いてモデルを実装した。

ベースコンポーネントの上に、採点基準コンポーネントを用いて拡張した場合の性能を示す。アテンションの計算手法として RTE（R）、内積（D）二つの手法があり、採点基準コンポーネントの特徴ベクトルとしてアテンションベクトル（A）、key concept の重み付き平均（W）の二つを比較した。

全ての訓練事例を用いて訓練したベースコンポーネントの ASAP-SAS に対する性能は、Riordan ら [12] が報告していた最高性能の 0.723 と同等であり、Riordan らのモデルを再現できたと言える。この場合、ベースコンポーネントは、採点基準の情報がない場合でも、多くの訓練事例の解答から直接的に key concept を学習できていると言える。採点基準の情報を用いた提案モデル（「+ 採点基準」）は、ベースコンポーネントと同等の性能を示している。この結果より、十分な訓練事例が利用できる場合でも、提案モデルはベースモデルの性能を傷つけないことがわかる。

低リソース設定（≤ 50%）では、データサイズの減少に伴って、採点基準の利用による性能向上（「+ 採点基準 RW」と「+ 採点基準 DA」）の効果が高くなることが確認できる。この結果より、ベースコンポーネントは、限られた訓練事例から直接的に key concept を学習するのが難しいことがわかる。また、このような状況では、採点基準の導入が SAG モデルの性能改善に有効であることがわかる。この結果は、我々の研究の非ニューラル版に対応する Sakaguchi ら [13] の結論とも一致する。

訓練事例が少ない場合、学習するパラメーターが少ない DA の性能は DW よりよくなる。一方で RA と RW を比較した場合、少ない訓練事例でもパラメータ数の多い RW が RA を上回る結果となった。これは SNLI によりパラメータを事前学習しているためだと考える。

低リソース設定に関するさらなる知見を得るため、13%の訓練事例を利用した場合の、各設問に対する性能を表 2 に示す。Key elements（K）また問題文の文（A））を key concept として用いた場合を示す。* はベースモデルと統計的有意差（Wilcoxon's signed rank test, $p < 0.05$）があることを示す。興味深いことに、設問 1、3、5、6 と 10 の「+ RW」の性能はベースコンポーネントより性能の向上ができた。その内、設問 1、5、6、と 10 では key elements は明示的に提供されている。この結果から、RTE モデルは key element が提供される設問にはより有効であると言える。

5 おわりに

採点基準は SAG にとって重要な役割を果たすが、SAG のコミュニティでは、その有効性がほとんど検証されてこなかった。本研究では、採点基準の情報をニューラル SAG モデルに取り込む手法を提案した。ベースコンポーネントとして最先端のニューラル SAG モデルを実装し、採点基準 (key concept) を取り込むための採点基準コンポーネントによってモデルを拡張した。実験により、ニューラルネットワーク SAG モデルに対する採点基準の利用が有効であることを明らかにした。一方で、SAG タスクにおける訓練データのサイズは限られているため、学習するパラメータの少ないモデル、または事前学習できるモデルが重要だとわかった。また key element 以外の採点基準の有効な利用手法も今後の重要な課題である。自動採点の次のステップとして、key concept のアテンションに基づくフィードバックの生成手法も検討する予定である。

参考文献

[1] Samuel R Bowman et al. "A large annotated corpus for learning natural language inference". In: *arXiv preprint arXiv:1508.05326* (2015).

[2] Daniel Cer et al. "Universal sentence encoder". In: *arXiv preprint arXiv:1803.11175* (2018).

[3] Alexis Conneau et al. "Supervised learning of universal sentence representations from natural language inference data". In: *arXiv preprint arXiv:1705.02364* (2017).

[4] Michael Heilman and Nitin Madnani. "The impact of training data on automated short answer scoring performance". In: *Proceedings of the Tenth Workshop on Innovative Use of NLP for BEA*. 2015, pp. 81–85.

[5] Andrea Horbach and Alexis Palmer. "Investigating Active Learning for Short-Answer Scoring". In: *Proceedings of the 11th Workshop on Innovative Use of NLP for BEA*. 2016, pp. 301–311.

[6] Diederik P Kingma and Jimmy Ba. "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980* (2014).

[7] Ryan Kiros et al. "Skip-thought vectors". In: *Advances in neural information processing systems*. 2015, pp. 3294–3302.

[8] Michael Mohler, Razvan Bunescu, and Rada Mihalcea. "Learning to grade short answer questions using semantic similarity measures and dependency graph alignments". In: *Proceedings of the ACL*. ACL. 2011, pp. 752–762.

[9] Michael Mohler and Rada Mihalcea. "Text-to-text semantic similarity for automatic short answer grading". In: *Proceedings of the EACL*. ACL. 2009, pp. 567–575.

[10] Kishore Papineni et al. "BLEU: a method for automatic evaluation of machine translation". In: *Proceedings of the 40th annual meeting on ACL*. ACL. 2002, pp. 311–318.

[11] Jeffrey Pennington, Richard Socher, and Christopher Manning. "Glove: Global vectors for word representation". In: *EMNLP*. 2014, pp. 1532–1543.

[12] Brian Riordan et al. "Investigating neural architectures for short answer scoring". In: *Proceedings of the 12th Workshop on Innovative Use of NLP for BEA*. 2017, pp. 159–168.

[13] Keisuke Sakaguchi, Michael Heilman, and Nitin Madnani. "Effective Feature Integration for Automated Short Answer Scoring". In: *Proceedings of NAACL*. 2015, pp. 1049–1054.

[14] Mike Schuster and Kuldip K Paliwal. "Bidirectional recurrent neural networks". In: *IEEE Transactions on Signal Processing* 45.11 (1997), pp. 2673–2681.

[15] Sainbayar Sukhbaatar, Jason Weston, Rob Fergus, et al. "End-to-end memory networks". In: *Advances in neural information processing systems*. 2015, pp. 2440–2448.