項目採点技術に基づいた和文英訳答案の自動採点

菊地正弥 [1,2*] 尾中大介 [1,2*] 舟山弘晃 [1,2*] 松林優一郎 [1,2] 乾健太郎 [1,2]
[1] 東北大学 [2] 理化学研究所
{taisuke.onaka.q7, seiya.kikuchi.t3}@dc.tohoku.ac.jp
hiroaki@ecei.tohoku.ac.jp {y.m, inui}@tohoku.ac.jp

1 はじめに

国内の英語教育では、短い和文に対する自由記述形式の英訳問題（短文和文英訳問題と呼ぶ）が盛んに使われている。この形式の問題は出題者が問う語学的知識を限定しやすく、特定の文法・語彙知識の習熟訓練・反復学習方法として有効である。一方でこうした記述式の答案に対する採点と助言の返却は教育者側の負担が大きく、アウトプットと反復学習機会が重要な語学教育においては、その限定的な試行回数が課題となる。

そこで、本稿では短文和文英訳問題に対する自動採点技術の応用を試みる。日本人学習者を対象にした英作文の自動採点を扱う研究はこれまでほとんど行われていないが、先行研究として、正答例を列挙する表現法を提案し効率的にマッチングを行うことで自動採点を実現しているものや [1]、学習者のエッセイライティングを対象にして文法誤りを検出し解説文を生成する手法 [2] がある。また、近接する分野として、和文英訳問題の採点は文法的正しさの評価という観点で Grammatical Error Correction (GEC) と関連する。短文からなる答案に対して採点基準に基づき評価を行うという観点では Short Answer Scoring (SAS) と見なすこともできる。

本研究では、高校の学習課程での運用実態を参考に、短文和文英訳問題の採点を、作題者が確認したい複数の部分要素を採点項目として反映した項目別の SAS の問題として定式化し、NN ベースのモデルにより自動採点を実現する。この我々の目的に必要となる学術利用可能な日本人学習者の和文英訳文採点データは存在しないため、我々はまず、和文英訳問題に対する自動採点モデルの学習と評価を可能とするデータ形式を設計し、答案の収集・アノテーションを行いデータセットを構築した（図 1）。このデータ中では、明確な複数の採点基準に基づき基準ごとに答案の採点が行われる。また、各採点項目の根拠箇所が付与されており採点の解釈性が高い。

図 1 データセットに含まれる答案及び採点基準の例。

実験では、我々が作成したデータセットに対し [3] が提案する項目採点モデルを適用し、その採点精度と今後の課題を検証する。加えて、疑似データ生成による性能向上についても検証する。一般に SAS におけるデータ作成は高コストなため、データ数が少ないときの性能の維持が課題である [3, 4]。そこで、GEC において成功している疑似的な文法誤り生成のアプローチ [5] を用いて、和文英訳問題採点での効果を検証する。実験の結果、大半の採点基準に対して F 値で 80%以上の精度で採点可能であることが明らかになった。また、GEC で用いられている手法 [5] を用いて生成した疑似データを学習に用いることにより、学習データが少ない評価ラベルに対する採点精度の向上を確認した。

2 和文英訳問題データセット

本研究を行うにあたり作成したデータセットについて説明する。本データセットは四天王寺高等学校・中学校および増進堂・受験研究社の協力のもと作成したものであり、和文英訳問題の答案として初

* equal contribution

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

表 1 答案データの統計値

	答案数	項目数	○	△	×
問 1	159	9	923	114	235
問 2	172	8	652	98	454
問 3	77	8	357	40	142
問 4	69	9	356	76	120
問 5	102	9	387	161	268
問 6	79	12	701	14	154
問 7	90	10	534	72	204

のデータであるだけでなく日本人英語学習者の SAS 形式の英文データとしても初めてのものである。

本データセットには問題が 7 問含まれており、それぞれの問題に複数の明確な採点項目と採点基準が存在する。問題文、採点基準は専門家により作成されたものである。答案は各採点項目ごとに採点され、項目評価値（○, △, × のいずれか）および評価根拠箇所（答案中の単語位置）が付与されている。答案例と採点基準の一部を図 1 に示す。問 1, 2 は四天王寺高等学校において行われた試験問題と生徒の答案である。また、増進堂・受験研究社の協力により新たに 5 問を作題し、問題 1, 2 と合わせてクラウドソーシングにより答案の収集を行った。収集においては、回答者のレベルを高校生と同等水準に揃えるため、英語検定試験のスコア (付録参照) で募集をかけ、能力チェック用設問の正答率によって回答者の選定を行なった。さらに、収集した答案の品質維持のため、著しく内容が逸脱した回答をルール (付録参照) および目視により除外した。その後、収集した答案に対して採点基準に基づき項目評価値と根拠箇所のアノテーションを行なった。アノテーションは、作題者である専門家がクラウドソーシングで収集した答案を閲覧し精緻化した採点基準書にしたがって、著者ら監督のもと別の作業者が実施した。

7 問に対するデータの統計を表 1 に示す。高校の答案を含む問 1,2 で 160-170 程度、その他は 70-100 程度の答案を含む。各問題には 10 前後の採点項目が存在する。本データセットは学術利用目的に限り公開する予定である[1)]。

3 和文英訳答案の自動採点

3.1 問題設定

本研究では和文英訳問題の採点を複数の採点項目を持った SAS 問題として定式化する。ある和文英訳問題に対して、その採点項目の集合を C とする。入力は答案テキスト $(w_1, w_2, ..., w_n)$ および採点項目 $c \in C$ であり、出力は c に対する 3 段階の評価値 $s_c \in \{○, △, ×\}$ とする。採点は各採点項目ごとに出力し、答案に対する総合的な評価は行わない。

3.2 項目採点モデル

本研究では、SAS のために提案された [6] の採点モデルを採点項目ごとに個別採点するよう拡張したモデル [3] によって和文英訳問題の採点を行う。ただし、我々のモデルは得点に対する回帰問題ではなく、3.1 に記述した 3 値分類問題を学習する。

本研究で用いるモデルを説明する。まず、入力された答案テキストは単語ベクトルの列 $\mathbf{x} = (x_1, x_2, ..., x_n)$ に変換される。その後、このベクトル列は Bi-LSTM に入力され、順方向と逆方向の中間状態ベクトルの和を取ることにより、D 次元の中間表現の列 $\mathbf{h} = (h_1, h_2, ..., h_n)$ にエンコードされる。これを用いて、各採点項目ごとに個別の文ベクトルを計算する。ただし、中間表現の列 $\mathbf{h}$ を作る Bi-LSTM は全採点項目で共有されることに注意されたい。$c \in C$ に対する文ベクトル $\tilde{h_c}$ は中間表現の重み付き和として次のように計算される。

$$\tilde{h_c} = \sum_{i=1}^{n} \alpha_i^c h_i \tag{1}$$

α_i^c は採点基準 c に対する i 番目の単語の重みであり、式 (2) に示す注意機構によって算出される。

$$\begin{aligned} t_i^c &= h_i M_c V_c \\ \alpha_i^c &= \frac{\exp(tanh(t_i^c))}{\sum_{k=1}^{n} \exp(tanh(t_k^c))} \end{aligned} \tag{2}$$

$M_c \in \mathbb{R}^{D \times D}$、$V_c \in \mathbb{R}^D$ は学習するパラメータである。最後に項目 c に対する評価値 s_c を次式により得る。

$$\begin{aligned} p(s_c|\mathbf{x}) &= \mathtt{softmax}(W\tilde{h_c} + b) \\ s_c &= \underset{s_c \in \{○, △, ×\}}{\arg\max} \{p(s_c|\mathbf{x})\} \end{aligned} \tag{3}$$

$W \in \mathbb{R}^{3 \times D}$、$b \in \mathbb{R}^3$ は学習するパラメータである。

3.3 学習

項目採点モデルは各採点項目の評価結果について負の対数尤度 (NLL) を最小化するように学習する。

$$L_{score} = \sum_{c \in C} \mathrm{NLL}(p(s_c|\mathbf{x}), \hat{s_c}) \tag{4}$$

ここで、$\hat{s_c}$ は採点基準 c に対する教師信号のラベル（評価値）である。また、2 節に示したように本

1） https://aip-nlu.gitlab.io/projects/sas-j

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

データセットには答案に対して採点項目ごとに採点根拠箇所 $\hat{\alpha^c} = (\hat{\alpha_1^c}, \hat{\alpha_2^c}, ..., \hat{\alpha_n^c})$ が付与されている。$\hat{\alpha_i^c} \in [0, 1]$ は答案中の i 番目の語が根拠となるかを表す数であり、採点根拠となるトークンが答案中に k 個ある場合、その位置には $1/k$ が、それ以外には 0 が付与されている。[3] に従い、本研究においても、以下の損失関数によりアテンションの教師有り学習を行う。

$$L_{att} = \sum_{c \in C} \sum_{i=1}^{n} (\alpha_i^c - \hat{\alpha_i^c})^2 \quad (5)$$

よって、全体の損失 L は以下の式で表される。

$$L = L_{score} + L_{att} \quad (6)$$

3.4 疑似データ生成

表 1 に示したように、本データセットに含まれる答案数は各問につき高々 170 件程度であり、NN モデルを用いる他のタスクのデータセットと比べて著しく少ないが、問題ごとにモデルの訓練が必要であるという SAS の特徴をふまえると大規模なデータ収集により答案数を増やすことは現実的ではない。

そこで、我々は GEC における疑似データ生成手法の一つである [5] をベースにした手法によって間違い事例の増強を検討する。和文英訳においては正答、部分正答の答案パターンは限られている一方で、間違い方は多様である。しかし表 1 のとおり、× 評価は ○ 評価に比べて答案の分布として収集が難しい傾向にある。これは、通常、和文英訳問題が問題に適した学力水準の被験者を対象として出題されることと関係する。したがって、× 評価の答案を効率的に収集する何らかの手法が必要である。

本研究では、まず数個の模範解答例について人手で項目別採点を行い、その各採点項目の採点根拠となっているトークンに対して、実際の間違い事例の類型化に基づいて設計した以下の品詞カテゴリの置き換え操作によって疑似データの生成を行う。

- 冠詞、前置詞、関係詞、所有限定詞
- 動詞の人称・時制
- 同一語幹の形容詞と副詞
- スペルが似た語

上記の置き換えが発生する箇所を採点根拠としている採点項目について、トークンを置き換えた後の評価値を × とすることで、疑似的に × の事例を増強する。「スペルが似た語」ではレーベンシュタイン距離が 4 以下でジャロウィンクラー距離が 0.8 を超える単語に置き換えた。なお、高校生の語彙レベルに合わせるために、置き換え語の単語は 1993 年から 2017 年まで 4 年ごとのセンター試験および追試験の電子化データ[2] より抽出した 4463 語に絞った。生成された疑似データの例は図 2 に示す。

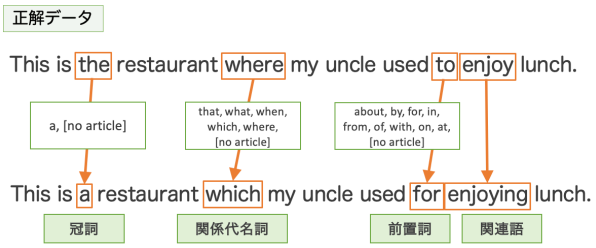

図 2 疑似データ生成例。この例では冠詞、関係詞、前置詞および関連語を置き換えることにより負例を疑似的に生成している。

4 実験

4.1 設定

本研究では ○, △, × の 3 値で評価を行っているため、自動採点の精度評価には F 値を用いた。モデルの学習においては、Adam [7] を用いて最適化を行った。また、学習率として 0.001 を用いた。また、実験は各問についてデータセットを訓練:開発:評価 =3:1:1 と分割する 5 分割交差検定により行なった。評価は採点項目ごとに行い、それぞれ 50 エポック回した上で開発セットに対して最も性能が高かった時のパラメータを用いた。また、疑似データの混入にあたっては、混入量をもとの訓練データの量の $\{10\%, 20\%, 30\%, 40\%, 50\%\}$ と変化させた時に開発セットにおいて最も性能が高かった割合を用いたモデルにより評価を行った。

4.2 結果

3.1 節で説明したモデルを用いた際の採点精度、および、3.4 節で説明した手法により疑似データを生成し、学習に用いた時の採点精度の変化を測定した。結果を表 2 に示す。実験結果は各問ごとに全採点基準のマイクロ平均取った時の F 値を示した。表 2 より、問題 1, 3, 4, 6 に対しては疑似データ追加を行う前の段階で F 値 0.9 ポイント程度と非常に高い精度で採点できることがわかった。問題 2, 5, 7 については他の問よりも採点精度は低く、特に問題 2 の

2) https://21robot.org/dataset.html

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

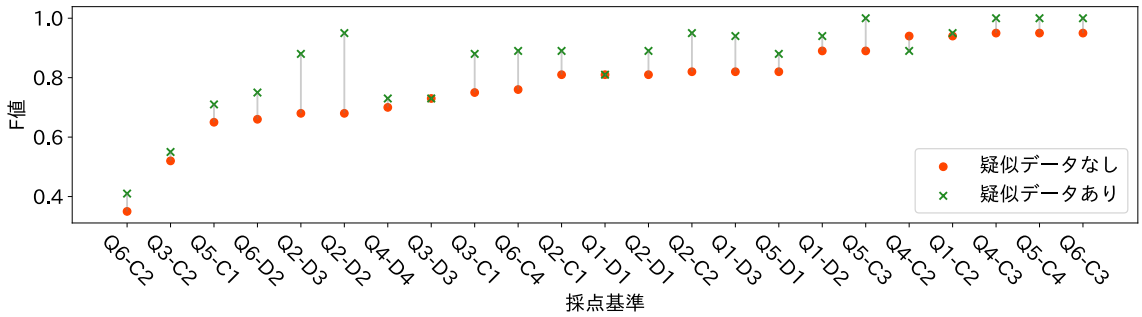

図 3　× 評価に対する項目ごとの F 値

表 2　問 1〜問 7 に対する項目別自動採点の F 値

	問題 1	問題 2	問題 3	
NN Model	0.915	0.736	0.892	
+ 疑似データ	**0.931**	**0.903**	**0.902**	
	問題 4	問題 5	問題 6	問題 7
NN Model	0.914	0.841	0.906	0.834
+ 疑似データ	0.906	**0.881**	**0.921**	**0.920**

採点精度については 0.73 ポイント強にとどまっている。次に、疑似データを混入した時の精度に注目する (表 2 中の+疑似データ)。疑似データを用いることで問題 4 を除いたすべての問に対して採点精度の向上が確認できた。特に、問題 2, 5, 7 においてはそれぞれ F 値で 0.17, 0.040, 0.096 ポイントの大きな精度向上が見られた。このことから、問題 2, 5, 7 のように元の採点精度が低い問題に対しては疑似データの生成が有効に働くと考えられる。

我々の疑似データ生成は × 評価に対する採点精度向上を狙ったものである。この効果を確認するため、図 3 に疑似データ生成の対象となった採点項目の × 評価に対する F 値を示した。ただし、正確な評価を行うために、× 評価の答案が 5 件以下の採点基準については評価を行っていないため、図 3 には掲載されていない。全体の傾向として、疑似負例の導入より × 評価に対する採点精度が向上することが分かる。特にオリジナルのデータのみを学習した場合に精度が低い採点項目（Q2-D3, Q2-D2）などにおいては、疑似データを学習することでモデルの性能は大きく向上している。ここから、容易に答案数を増やせない状況においても、疑似データ生成により低コストでモデルの精度向上が可能だということが確認できた。一方で、疑似データを学習したことにより性能が下がるケースも見受けられる（Q4-C2）が、その影響はわずかであった。

5　おわりに

本研究では国内の英語教育で盛んに利用されている自由記述形式の和文英訳問題の採点に焦点を当て、この自動化技術の構築に取り組んだ。高校現場での試験とその採点の運用実態を参考に、タスクの定式化を行い、この形式にもとづいて和文英訳答案の採点データセットを構築した。また、既存の NN ベースの項目点別採点モデルを適用し、約半数の問題では F 値 0.9 ポイント程度の高精度で採点が可能な一方で、一部の問題では採点精度がそれ未満にとどまることがわかった。この主要な理由は、間違い方の種類が多岐に渡る項目において訓練データ中のカバレッジが不足することにある。こうした問題においても疑似的に生成した負例を用いることで × ラベルに対する採点精度を向上させることができ、結果として F 値を 0.9 ポイント付近に引き上げられることを確認した。今後は数十問規模でデータセットを拡充し、現場レベルでの実証実験につなげていくことを予定している。また、学習効果の向上のためには、学習者は提出した答案に対して評価だけではなく学習上のアドバイスを得られる事が重要であると考えられる。したがって、答案の評価を推定するだけではなく、答案に合ったアドバイスを出力することを今後の課題として検討している。

謝辞

データセットの作成にあたって四天王寺高等学校・中学校および株式会社 増進堂・受験研究社にご協力を頂きました。また、Yahoo!クラウドソーシングにおいてご協力頂いたクラウドワーカーの皆様へ深く感謝を申し上げます。本研究は JSPS 科研費 JP19H04162、JP19K12112 の助成を受けたものです。

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

参考文献

[1] Norihisa NISHIMURA, Kentaro MEISEKI, and Michiaki YASUMURA. Development and evaluation of system for automatic correction of english composition. *Transactions of Information Processing Society of Japan*, Vol. 40, No. 12, pp. 4388–4395, dec 1999.

[2] Kazuaki HANAWA, Ryo NAGATA, and Kentaro INUI. Generation control methods for reliable feedback comment generation. *Proceedings of the Annual Conference of JSAI*, Vol. JSAI2020, No. 0, pp. 2D1GS903–2D1GS903, 2020.

[3] Tomoya Mizumoto, Hiroki Ouchi, Yoriko Isobe, Paul Reisert, Ryo Nagata, Satoshi Sekine, and Kentaro Inui. Analytic score prediction and justification identification in automated short answer scoring. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pp. 316–325, Florence, Italy, August 2019. Association for Computational Linguistics.

[4] Tianqi Wang, Naoya Inoue, Hiroki Ouchi, Tomoya Mizumoto, and Kentaro Inui. Inject rubrics into short answer grading system. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, pp. 175–182, Hong Kong, China, November 2019. Association for Computational Linguistics.

[5] Yujin Takahashi, Satoru Katsumata, and Mamoru Komachi. Grammatical error correction using pseudo learner corpus considering learner's error tendency. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pp. 27–32, Online, July 2020. Association for Computational Linguistics.

[6] Brian Riordan, Andrea Horbach, Aoife Cahill, Torsten Zesch, and Chong Min Lee. Investigating neural architectures for short answer scoring. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pp. 159–168, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[7] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

A　付録

A.1　回答者の選抜基準

回答者のレベルを高校生と同等水準に揃えるため、英語検定試験のスコアによって募集者の選抜を行った。表 3 にその際に用いた基準を示す。これらの基準の策定にあたっては、国内の英語検定資格と CEFR の対応[3)]や専門家の意見などを参考にした。また、回答者の年齢を 17〜39 歳に絞った。

表 3　回答者の選抜基準

検定名	対象範囲（点）
センター試験（英語）	140-
TOEIC L&R	550-750
TOEFL iBT	55-70

A.2　無効答案のフィルタリング

収集した答案の中には、題意に沿って回答していない答案や機械翻訳機などを用いて機械的に生成された答案などが散見された。そこで、以下の４項目からなるフィルタリング規則によって、無効答案のフィルタリングを行った。

1. 3 単語未満の回答を除去
2. Google 翻訳と同文の回答を除去
3. 設問にて回答条件が設けられている場合、条件を満たさない回答を除去
4. 正答例との類似度が著しく低い回答を除去

ここで、回答条件とは文頭の数単語が設問中に予め与えられている場合などの、設問で指定されている答案が絶対に満たすべき条件を指す。

3）　https://www.mext.go.jp/b$_{m}enu/houdou$/30/03/$_{icsFiles/afieldfile}$/2019/01/15/1402610$_{1}.pdf$

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).