The 35th Annual Conference of the Japanese Society for Artificial Intelligence, 2021

写真描画問題における自動採点手法の検討

Automatic Scoring for Picture Description Problems

田中健斗 *1 Kento Tanaka　西村太一 *1 Taichi Nishimura　白井圭佑 *1 Keisuke Shirai　亀甲博貴 *2 Hirotaka Kameko　森信介 *2 Shinsuke Mori

*1京都大学大学院情報学研究科
Department of Intelligence Science and Technology, Kyoto University

*2京都大学学術情報メディアセンター
Academic Center for Computing and Media Studies, Kyoto University

In language learning, training output skill such as speaking and writing is vital in order to retain the learned knowledge. However, scoring descriptive questions by humans would be costly, and this is why automatic scoring systems attract attention. In this research, we try to realize an automatic scoring system for picture description. Concretely, (i) we first analyze the trends of errors that English learners would make, (ii) then create a pseudo dataset by artificially mimicking the errors, and (iii) finally consider a model that judges whether a given pair of a picture and a sentence is valid or not. In experiments, we trained the model with the created pseudo data and evaluate it with the answers provided by actual learners. From experimental results, we found that our model outperforms a random agent.

1. はじめに

語学学習において,「話す」,「書く」といった言語の産出能力の訓練は学習した知識を定着させる上で重視されている．しかしながら，近年の教育現場での教師不足問題を背景に，記述式問題の採点には多大なコストを要する．写真描画問題は，産出能力を訓練する記述式問題の一つで，TOEIC をはじめとした英語の試験で採用されており，与えられた語句を用いて写真の様子を適切に表現する能力が求められる．問題の例を図 1 に示す．自由記述式のエッセイなどと異なり，写真を選出するだけで作問可能なので，教材作成コストが抑えられる．学習者は写真描画問題をはじめとした記述式問題を解き，誤りに対する適切なフィードバックを受けることで効果的な学習が期待できる．

本研究では，TOEIC のライティング試験 *1 を参考に表 1 に示す 3 つの評価基準を設け，総合的に写真描画問題における学習者の解答を採点することを考えた．評価基準の (2) については，指定の語句が適切に使われているかどうか判断可能である一方で，(1) と (3) については学習者の解答に沿った採点が求められる．本研究では (1) の基準に着目し，学習者の解答が写真に関連する内容であるか判断する手法を提案し，写真描画問題における自動採点手法の検討を行う．

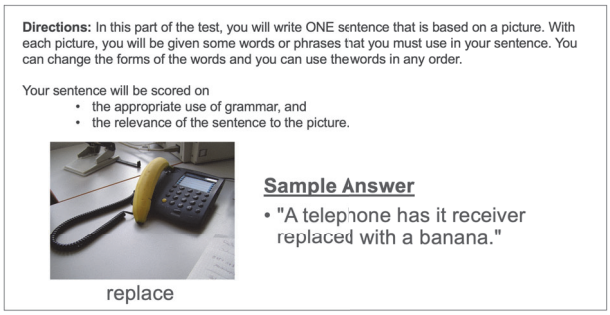

図 1: 写真描画問題.

表 1: 写真描画問題の評価基準.

評価基準	
(1)	写真と関連する内容が記述されている
(2)	与えられた語句を適切に使っている
(3)	文法的誤りがない

2. 関連研究

語学学習支援に関わる研究は盛んに行われており，学習者の書いた文の文法的な誤りを検知し訂正する文法誤り訂正（Grammatical Error Correction; GEC） [Omelianchuk 20] や，自由記述文の添削 [Alikaniotis 16] などがある．GEC では文法的に誤りのある文を入力として受け取り，誤り箇所を訂正した文を出力する．特に機械翻訳の手法を用いて誤りのある文から訂正文に翻訳する sequence-to-sequence の手法で高い精度が得られることが確認されている [Chollampatt 18, Zhao 19]．一方で，機械翻訳では大量の対訳コーパスが必要とされるが，GEC における学習者コーパスは規模が小さく，学習者の誤り傾向から疑似的に誤りのある文を生成することで，学習データを拡張する研究が盛んに行われている．

一方で，写真を用いた語学学習支援の事例は多くない．古屋ら [古屋 20] は画像キャプション生成手法から得られる文の生成確率を用いた低コストな自動採点方法を検討し，綴り誤りや軽微な文法誤りを含む文など，画像キャプション生成手法をそのまま適用できない場合に改善が必要であることを示した．川端ら [川端 20] は画像キャプション生成手法から得られた文の一部を空白にして穴埋め問題を作成した．解答データを収集して誤り傾向の分析を行い，画像キャプション生成手法を用いた語学学習支援システム構築の為の基礎研究を実施している．

連絡先: 田中健斗，京都大学大学院情報学研究科，京都府京都市左京区吉田本町，tanaka.kento.27n[at]st.kyoto-u.ac.jp

*1 https://www.iibc-global.org/toeic/test/sw/about/format.html

表 2: 写真描画問題解答者の CEFR と母国語.

グループ A		グループ B	
CEFR	母国語	CEFR	母国語
C1	Chinese	C2	English
C1	German	B2	Japanese
B2	Japanese	B2	Japanese
B2	Japanese	B1	Japanese
B2	Japanese	B1	Japanese
B1	Japanese	B1	Japanese
B1	Japanese	A2	Japanese
A2	Japanese	A2	Japanese
A1	Japanese		

3. 写真描画試験の実施

3.1 試験作成と実施

本研究では，写真描画問題における学習者の誤り傾向の分析を行う．(i) 写真の選出，(ii) 解答者に与える語句の抽出，(iii) 写真描画問題の試験の実施，の 3 つのプロセスで写真描画問題の試験を実施する．

問題セット作成には MS-COCO [Lin 14] の写真とキャプションを利用し，動物，食べ物，スポーツなど，写真のジャンルに偏りが生まれないよう 20 枚を選出し，10 枚ずつの 2 グループ（グループ A, B）に分割した．また，本研究では，解答者に提示する語句を動詞に限定し，MS-COCO のキャプションから抽出した．

外国語の学習・教授・評価のためのヨーロッパ言語共通参照枠（Common European Framework of Reference for Languages; CEFR） *2 が示す外国語熟達度の基準をもとに，英語の初学者から母国語話者を含む合計 17 名を対象に写真描画問題の試験を実施した．CEFR は言語の枠や国境を越えて，学習者や評価者が外国語の熟達度を同一の基準で判断することができ，A1，A2，B1，B2，C1，C2 の 6 段階に分けられる．また，A1，A2 は基礎段階の言語使用者，B1，B2 は自立した言語使用者，C1，C2 は熟練した言語使用者と定義されている．解答者をグループ A に 9 名，グループ B に 8 名を割り当て，それぞれのグループの問題セットを web 上で解いてもらい解答データを収集した．グループ A とグループ B それぞれの解答者の CEFR レベルと母国語を表 2 に示す．

3.2 誤り傾向分析

誤り傾向を分析するために，収集した解答データの人手評価を実施した．評価基準は表 1 に従い，写真と関連する内容が記述されているか，与えられた語句が適切に使えているか，文法的に誤りがないか，それぞれの項目ごとに正しければ 1 点を誤りがあれば 0 点と評価した．解答者の誤り傾向を図 2 に示す．写真描画問題における誤りの多くは文法的な誤りにあり，動詞の活用や時制，冠詞，スペル，前置詞の誤りが確認できた．一方で，解答文と写真の関連性に関する誤りは全体の 15 %で，内容語に関わる誤りが多く確認された．内容語の誤りの例を図 3 に示す．なお，文法的に正しい表現であるが写真の内容にそぐわない名詞や動詞が含まれている場合は内容語の誤りとしている．図 3 の (1) では電話の受話器を “receiver” と表現するべきだが，“picker” としてしまっている．(2) ではゾウを “danbo” とし，“2 years old” という写真から判断できない情報を追加している．(3) では川沿いの鳥は立っているため

*2 https://www.cambridgeenglish.org/exams-and-tests/cefr/

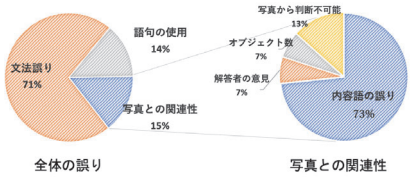

図 2: 写真描画問題における誤り傾向分析.

“sitting” ではなく “standing” と表現するのが適切である．(4) では男性が来ている服を “write short” と表現している．

また，飛行機が着陸する写真に対し，“Welcome skyland” と描画する，存在しない単語を含む内容語の誤りや，男性がサーフィンをする写真に対して “I like surf” といったように，解答者自身の意見を含めた誤りが確認された．

4. 提案手法

本節では，前節で行った分析をもとに，写真描画問題における解答文と写真との関連性を評価する正誤判定モデルを提案する．提案するモデルを図 4 に示す．

本研究で提案する正誤判定モデルは 2 つのプロセスからなる．(i) で訓練データを疑似的に生成する．(ii) で解答文と写真の特徴量を抽出する．最後に (iii) で (ii) で得られた特徴量を結合し入力に，正答もしくは誤答ラベルを出力する 2 値分類モデルの学習を行う．

4.1 疑似誤り生成

解答データの収集はコストが高く，使用できる学習者コーパスには限りがある．一方で，正誤判定モデルを学習するには大規模なデータが必要となる．従って，正しい文から疑似的に誤りを生成することで，誤り文を獲得する疑似誤り生成を行う．

写真描画問題における，解答文と写真との関連性に関する誤り傾向は，写真内の情報を適切に表現できていない内容語の誤りが最も多いことが確認された．この誤り傾向から，本研究では，正しいキャプションに複数含まれる名詞を適切でない名詞に置換し，不適切な内容語を含む誤り文を生成する．キャプション内の名詞の特定に Natural Language Toolkit (NLTK) [Loper 02] を使用した．さらに，自然言語処理の多くのタスクにおいて高い精度を示している言語モデルの Bidirectional Encoder Representations from Transformers (BERT) [Devlin 18] を用いて，キャプション内の名詞を空所としたときの空所補充問題を解くことで，不適切な名詞に置換する．ただし，元のキャプション内の名詞やその名詞と類似度の高い語を補充しないように，WordNet を用いて，類似度を計算する．類似度の計算には，以下に示す wup-similarity [Pedersen 04] を用いる．

$$\text{wup-similarity}(s_1, s_2) = \frac{2 \times \text{depth}(\text{lcs}(s_1, s_2))}{\text{depth}(s_1) \times \text{depth}(s_2)} \quad (1)$$

s は Synset(概念) を，depth は最上位概念までの距離を示す．また，lcs(Lowest Common Synset) は最小共通概念を示す．

4.2 モデルの学習

図 4 に従って，写真，解答文それぞれの特徴量を抽出する．写真からは ImageNet [Deng 09] で学習済みの ResNet50 [He 16] を用いて特徴量 $\boldsymbol{V}$ を得る．また，解答文からは事前学習済みの SentenceBERT [Reimers 19] を用いて特徴量 $\boldsymbol{A}$ を得る．

そして，写真の特徴量 $\boldsymbol{V}$，解答文の特徴量 $\boldsymbol{A}$ を結合する．結合した特徴量を 2 層の全結合層に入力し，最終層の出力に

	(1)	(2)	(3)	(4)
画像				
語句	replace	kiss	take / sit	hold
解答文	A banana is replaced picker.	The danbo that is only 2 years old kiss handsome man.	A woman is taking a picture of a bird sitting near a river.	The man who wear write short is hold rainbow color balloon.
MSCOCO	A telephone has it receiver replaced with a banana.	A man being kiss by a baby elephant with it's trunk.	A waterway under a bridge with people sitting down and a woman taking a photo.	A man holds strings connected to a large parachute or kite on the beach.

図 3: 内容語の誤り.

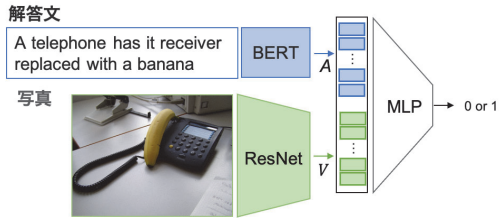

図 4: 提案する正誤判定モデル.

表 3: データセットの統計情報.

	train	valid
写真数	117,755	4,977
写真に対する正例数	5.00	5.00

シグモイド関数を適用することで，解答文が正答であれば 1，誤答であれば 0 を出力し，正誤ラベル $\boldsymbol{L} = (0, 1)$ を得る．また，交差エントロピー誤差関数を用いて学習する．

5. 実験

写真描画問題アンケートで収集した解答に，本手法の正誤判定モデルを適用し提案手法を評価する．

5.1 データセット

モデルの学習に MS-COCO を利用した．表 3 にデータセットの統計情報を示す．提案手法の疑似誤り生成における類似度の閾値は 0.25，0.5，0.75，1.00 と変化させ，閾値よりも小さな類似度を持つ単語に限定し，不適切な名詞に置換した．

モデルの学習は訓練用と検証用で分割された MS-COCO のデータセットを利用した．検証用データにおける損失関数の値が最も低くなった時点のモデルの重みをテストに用いた．

5.2 実験設定

写真は学習済みの ResNet50 から 2,048 次元の特徴量を抽出した．また，解答文は SentenceBERT を用いて 768 次元の特徴量を抽出した．写真と解答文の特徴量を結合した 1,536 次元の特徴量をモデルの入力とする．バッチサイズは 128 とし，最適化手法には Adam を使用した．

表 4: 正誤判定モデルの定量的評価.

判別器	適合率	再現率	F 値
ランダムに出力する判別器	0.112	**0.688**	0.193
提案手法 + 類似度 (< 0.25)	0.375	0.188	0.250
提案手法 + 類似度 (< 0.50)	**0.750**	0.188	0.300
提案手法 + 類似度 (< 0.75)	0.667	0.250	**0.336**

5.3 定量的評価

提案手法を評価するために，写真描画問題アンケートで収集した学習者の実際の解答にモデルを適用した．写真描画問題の試験は 17 人の解答者が 10 の質問に解答しており，170 の解答データのうち写真に関連した内容が表現できていない誤りは 16 個含まれている．人手による採点をもとに正誤判定モデルの誤り検知に対する適合率と再現率，F 値を算出した．また，正誤判定をランダムに出力する判別器の評価も同様に実施する．結果を表 4 に示す．類似度の閾値が 0.75 のとき，F 値は最も高くなった．また，ランダムに出力する判別器と比較して, 提案モデルは高い識別性能を実現することがわかった．一方で，満足な結果が得られなかった原因として，学習者の実際の誤りと疑似誤りの分布の不一致が挙げられる．疑似誤り生成手法において，学習者の誤り傾向を考慮した不適切な名詞への置換などが改善策として考えられる．

5.4 定性的評価

正誤判定モデルを適用した結果，図 3 の (2) の解答文は誤答と判断することができた．また，図 5 に正誤判定モデルを適用した結果の一部を示す．誤答を正答と判断した例の 1 つとして，“ski” と “snow board” の違いを判別できなかった． 本研究では文中の名詞を類似度をもとに不適切な名詞に置換することで内容語の誤りを疑似生成している．“ski” と “snowboard” は類似度の大きさ (0.58) から判別が困難であったと考えられる．また，“train has gone to America” といった写真から判断できない情報を含めている誤りについては，本研究の疑似誤り生成手法では誤りを検知するのが困難であるため，別の手法が必要になると考えられる．一方で，正答を誤答と判断した例もある．一つは写真の状況から推測できる “A vegetable seller” の判断が困難であったと考えられる．また，右の解答文については文法誤りが含まれているので，訂正し，正誤判定モデルを適用した結果，正答と判断することができた．文法誤りが正誤判定モデルの結果に影響する可能性が考えられる．

	誤答を正答と判断した例		正答を誤答と判断した例	
画像				
語句	learn	go	display	look
解答文	A girl learns snow boarding.	The green, black and red train has gone to America.	A vegetable seller displays his apples arranged by color.	Many zebras is look at same spot and is walking.

図 5: 正誤判定モデル適用結果.

6. おわりに

本研究では，写真描画問題における自動採点手法の検討を行った．解答文と写真の関連性に着目し，解答文と写真の特徴量から正誤判定を行うモデルを提案した．実験は学習者の誤り傾向分析をもとに生成した疑似誤りデータを用いて学習し，実際の学習者の解答を用いて評価した． 実験結果から，ランダムに出力を行う判別器と比較して，提案モデルは高い識別性能を実現することがわかった．

今後の課題として，写真内のオブジェクトと解答文内の名詞を関連づけて学習を行う正誤判定モデルを検討する．

参考文献

[Alikaniotis 16] Alikaniotis, D., Yannakoudakis, H., and Rei, M.: Automatic Text Scoring Using Neural Networks, in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 715–725, Berlin, Germany (2016), Association for Computational Linguistics

[Chollampatt 18] Chollampatt, S. and Ng, H. T.: A Multilayer Convolutional Encoder-Decoder Neural Network for Grammatical Error Correction, *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32, No. 1 (2018)

[Deng 09] Deng, J., Dong, W., Socher, R., Li, L., Kai Li, , and Li Fei-Fei, : ImageNet: A large-scale hierarchical image database, in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255 (2009)

[Devlin 18] Devlin, J., Chang, M., Lee, K., and Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, *CoRR*, Vol. abs/1810.04805, (2018)

[He 16] He, K., Zhang, X., Ren, S., and Sun, J.: Deep Residual Learning for Image Recognition, in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778 (2016)

[Lin 14] Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L.: Microsoft COCO: Common Objects in Context, in Fleet, D., Pajdla, T., Schiele, B., and Tuytelaars, T. eds., *Computer Vision – ECCV 2014*, pp. 740–755, Cham (2014), Springer International Publishing

[Loper 02] Loper, E. and Bird, S.: NLTK: The Natural Language Toolkit, *CoRR*, Vol. cs.CL/0205028, (2002)

[Omelianchuk 20] Omelianchuk, K., Atrasevych, V., Chernodub, A., and Skurzhanskyi, O.: GECToR – Grammatical Error Correction: Tag, Not Rewrite, in *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pp. 163–170, Seattle, WA, USA → Online (2020), Association for Computational Linguistics

[Pedersen 04] Pedersen, T., Patwardhan, S., and Michelizzi, J.: WordNet::Similarity: Measuring the Relatedness of Concepts, in *Demonstration Papers at HLT-NAACL 2004*, HLT-NAACL–Demonstrations '04, p. 38–41, USA (2004), Association for Computational Linguistics

[Reimers 19] Reimers, N. and Gurevych, I.: Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks, *CoRR*, Vol. abs/1908.10084, (2019)

[Zhao 19] Zhao, W., Wang, L., Shen, K., Jia, R., and Liu, J.: Improving Grammatical Error Correction via Pre-Training a Copy-Augmented Architecture with Unlabeled Data, in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 156–165, Minneapolis, Minnesota (2019), Association for Computational Linguistics

[古屋 20] 古屋 昭拓, 永田 亮, Tomko, J.：写真を用いた英文記述問題の低コストな自動採点方法の検討, 言語処理学会第 26 回年次大会発表論文集, pp. 693–696 (2020)

[川端 20] 川端 公貴, 南條 浩輝, 亀甲 博貴, 森 信介：画像キャプションを用いた日本語学習支援の検討, 言語処理学会第 26 回年次大会発表論文集, pp. 501–504 (2020)