画像描写問題における学習者作文の誤り訂正

田中健斗 [1]　西村太一 [1]　南條浩輝 [2]　白井圭佑 [1]　亀甲博貴 [2]
[1] 京都大学大学院情報学研究科　[2] 京都大学学術情報メディアセンター
[1]tanaka.kento.s07@kyoto-u.jp
[1]{nishimura.taichi43x,shirai.keisuke.64x}@st.kyoto-u.ac.jp
[2]nanjo@media.kyoto-u.ac.jp　[2]kameko@i.kyoto-u.ac.jp

概要

本研究では，画像描写問題における学習者作文の誤りを訂正する新しいタスクを提案する．本タスクでは，文法的な誤りだけでなく，画像描写問題特有の意味論・語用論に関する誤りの訂正が求められる．画像と学習者作文を入力に取り，誤りを訂正した文を生成するベースラインを構築した．実験結果では，文法誤り訂正 (Grammatical Error Correction; GEC) で訂正が困難な意味論・語用論の誤りを訂正できることを確認し，画像を参照する提案手法の有用性を確認した．

1　はじめに

外国語教育において，コミュニケーション能力を重視した 4 技能の育成が重視されている．4 技能のうちの「話す」「書く」という産出能力の評価は難しいタスクである．これは，学習者の多様な発言・作文に対応する必要があるためである．本研究では，学習者に画像を見せて関連する文を産出させる「画像描写問題」[1)]に着目し，学習者作文の自動評価の研究を行う．

学習者は画像描写問題を通して，即時的に考えを述べる能力を身につけることができる．また，画像により作文の自由度が制限されるので，添削基準が確立され，添削コストは削減される．しかし，画像描写問題は明確な利点が多くある一方で，自動評価に取り組んだ研究例は少ない．

本稿では，画像描写問題における学習者作文の誤りを訂正するシステム (図 1) を作成したので，それについて述べる．

1）　外国語の試験などでは「写真描写問題 (picture description)」と表されることが多いが，情報処理分野では「写真 (picture)」を「画像 (image)」と表現するため，このように呼称する．

図 1　画像描写問題の誤りを訂正するシステム．

2　関連研究

近年，学習者作文の誤り訂正のための Shared task である CoNLL-2014 [1] や BEA-2019 [2] が整備され，文法誤り訂正 (Grammatical Error Correction; GEC) が盛んに研究されている．Zhao ら [3] は，誤りを含む文を訂正文に翻訳する機械翻訳ベースの GEC 手法において，訂正文は入力文の多くがコピーされる傾向に着目し，Transformer にコピー機構を備えた手法を提案している．一方で Omelianchuk ら [4] は，コーパス構築にコストを要する機械翻訳ベースの GEC 手法の精度向上が難しい点に着目し，系列ラベリングの手法を用いた GECToR を提案した．

GEC は文法上の誤りの訂正を目指すものであり，意味論または語用論に関する誤りは原理的に訂正できない．例えば，“there is a bench on glass” には文法上の誤りはないため，GEC では訂正できない．“glass” を “grass” と訂正するには，世界に対する背景知識やコンテクストが必要である．コンテクストには，この文の前にある文章だけではなく，書かれた状況（画像）も含まれる．画像描写問題における誤り訂正は，コンテクストとして画像の情報を利用することで，このような誤りの訂正を目指すものである．

表 1　データセットの統計情報.

	文数	平均単語数
学習者作文	651	7.12
専門家による添削	644	7.54

3　データセットの構築

画像描写問題における学習者作文訂正のために，学習者作文とその訂正文のペアを収集した．具体的には，画像描写問題における英作文を収集し，英文添削の専門家による添削を実施した．

3.1　学習者作文の収集

本研究では，主に日本の高校生に画像描写問題に取り組んでもらい，学習者作文を収集した[2]．画像描写問題では，画像と画像内の枠部分 (赤枠で囲む) を提示し，その部分について 1 文の英語で描写することを求めた．

問題に使った画像は，RefCOCOg [5] から選出した．RefCOCOg は，MSCOCO [6] の 26,711 枚の画像の中の 54,822 個の物体に対する 85,474 個の参照表現[3]からなるデータセットである．問題に使用する画像，及び特定の物体は画像のカテゴリ (e.g., 人，食べ物，車) に偏りが生まれないように 120 枚を人手で選出した．この 120 枚の画像に対し，651 文の学習者作文を得た．

3.2　専門家による添削

次に，日本語を十分に理解する英文添削の専門家に，収集した学習者作文の誤りの訂正を依頼した．その際，元の作文に限りなく近い構文での，文法上の誤りの訂正（文法的な添削）および画像の内容にふさわしくない表現誤りの訂正（写真との関連性における添削）を依頼した．誤りが著しく多く添削が困難な文が 7 文あり，それらは訂正されなかった．表 1 に学習者作文と添削結果の統計情報を示す．図 2 に訂正例を示す．

4　画像描写の誤り訂正

本研究では，3 節で述べたデータセットを用いて，学習者による画像描写の誤りを訂正するベースラインモデルを構築し，その評価を行う．具体的には，画像 I と学習者作文 L から，適切に訂正された文 C を出力する問題とし，エンコーダ・デコーダモデルでモデル化する．モデルの概要を図 3 に示す．詳細は次に述べる．

2)　一部のデータは web 上で匿名ユーザに対して実施して収集した．

3)　参照表現とは特定の物体を他の複数の物体から識別する言語表現である．

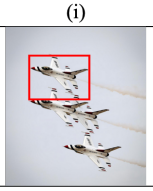
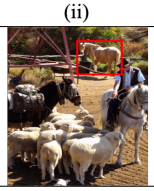
	(i)	(ii)
		
学習者作文	That air is flying of sky.	A month is lonely
専門家による添削	A jet is flying in the sky.	A horse is lonely.

図 2　専門家による添削例．（赤枠で囲まれた部分について描写した作文を学習者に求めた）

4.1　エンコーダ

画像エンコーダ． 画像エンコーダでは画像 I の特徴量を抽出する．学習者は複数の物体を描写することが多く，各物体を基にした特徴量を抽出するために，注意機構 [7, 8] をモデルに取り入れる．画像 I から Faster-RCNN [9] を用いて物体を抽出し，ResNet-101 [10] を用いて，物体領域の特徴量 $\boldsymbol{O} = (\boldsymbol{o}_1, \ldots, \boldsymbol{o}_k, \ldots, \boldsymbol{o}_K) \in \mathbb{R}^{d_v \times K}$ を獲得する．また，物体の位置関係を考慮するために，4 次元の位置座標 (バウンディングボックス) を d_e 次元の特徴量 $\boldsymbol{P} = (\boldsymbol{p}_1, \ldots, \boldsymbol{p}_k, \ldots, \boldsymbol{p}_K)$ に拡張した．抽出した物体 $\boldsymbol{V} = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k, \ldots, \boldsymbol{v}_K) \in \mathbb{R}^{d_e \times K}$ は，$\boldsymbol{O}$ と $\boldsymbol{P}$ の合計を取って，次のように表現できる．

$$\boldsymbol{V} = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_K) \tag{1}$$

$$= (\boldsymbol{W}_o \boldsymbol{o}_1 + \boldsymbol{p}_1, \ldots, \boldsymbol{W}_o \boldsymbol{o}_K + \boldsymbol{p}_K), \tag{2}$$

$\boldsymbol{W}_o \in \mathbb{R}^{d_e \times d_v}$ は全結合層の重みを表す．ここで述べた画像の特徴量抽出方法を本論文では，**Bottom-up** と呼ぶ．また，ResNet-152 を用いて，画像全体から特徴量 $\boldsymbol{v}' \in \mathbb{R}^{d_e}$ を抽出する手法も **Global** として試した．

学習者作文エンコーダ． 事前学習済みの BERT [11] を用いて，学習者作文 L から特徴量 $\hat{\boldsymbol{e}} \in \mathbb{R}^{d_e}$ を抽出する．学習者作文 L は WordPiece Tokenizer を用いてサブワードに分割し，BERT により特徴量 $(\boldsymbol{e}_1, \ldots, \boldsymbol{e}_n, \ldots, \boldsymbol{e}_N) \in \mathbb{R}^{d_e \times N}$ を獲得する．各特徴量を平均プーリングすることで文レベルの特徴量 $\hat{\boldsymbol{e}}$ とする．

$$\hat{\boldsymbol{e}} = \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{e}_i. \tag{3}$$

注意機構． エンコーダで獲得した物体の特徴量 V と学習者作文の特徴量 $\hat{\boldsymbol{e}}$ を基に，注意機構は各物体

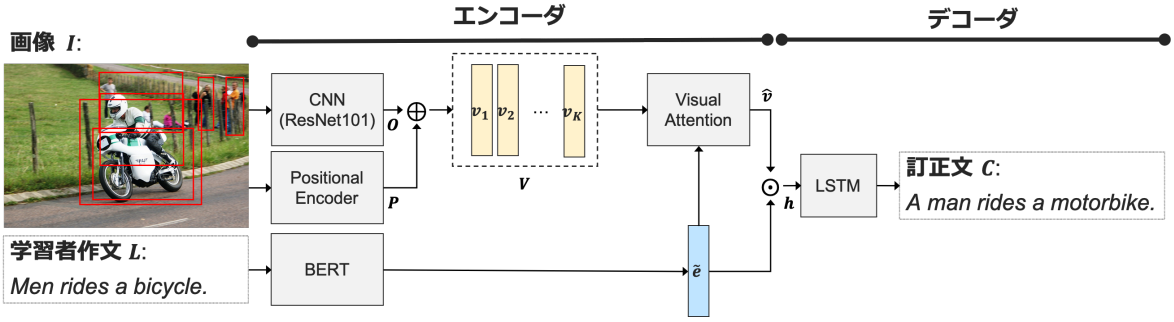


図 3　提案する誤り訂正モデルの概要図．（画像中の赤枠は画像エンコーダが特徴量を取り出す際に参照した領域）

の学習者作文に関する重み $\hat{\boldsymbol{v}}$ を算出する．[12] らの手法に倣い，画像の特徴量 $\hat{\boldsymbol{v}}$ は，

$$\hat{\boldsymbol{v}} = \sum_{k=1}^{K} \alpha_k \boldsymbol{v}_k. \tag{4}$$

のように算出する．また，重み α_k は，

$$\alpha_k = \frac{\exp(\tau_k)}{\sum_{j=1}^{K} \exp(\tau_j)}, \tag{5}$$

$$\tau_k = \boldsymbol{W}_{ve}(\mathrm{ReLU}(\boldsymbol{W}_v \boldsymbol{v}_k) \odot \mathrm{ReLU}(\boldsymbol{W}_e \hat{\boldsymbol{e}})), \tag{6}$$

のように算出する．ただし，$\odot$ はアダマール積を指し，$\boldsymbol{W}_v \in \mathbb{R}^{d_e \times d_e}$，$\boldsymbol{W}_e \in \mathbb{R}^{d_e \times d_e}$，$\boldsymbol{W}_{ve} \in \mathbb{R}^{1 \times d_e}$ は全結合層の重みを表す[4]．また，画像エンコーダに “Global” を用いた場合，注意機構は適用されず，画像の特徴量は $\hat{\boldsymbol{v}}$ の代わりに $\boldsymbol{v}'$ となる．

4.2　デコーダ

学習者作文と画像の特徴量 $(\hat{\boldsymbol{e}}, \hat{\boldsymbol{v}})$ を基に，LSTM を用いて訂正文を生成する．LSTM の隠れ状態 $\boldsymbol{h}$ の初期値は次のように算出する．

$$\boldsymbol{h} = \mathrm{ReLU}(\boldsymbol{W}_f \hat{\boldsymbol{v}}) \odot \mathrm{ReLU}(\boldsymbol{W}_g \hat{\boldsymbol{e}}), \tag{7}$$

$\boldsymbol{W}_f, \boldsymbol{W}_g \in \mathbb{R}^{d_e \times d_e}$ は全結合層の重みを指す．

4.3　損失関数

画像 I と学習者作文 L を入力に取り，訂正文 C を出力するように，次の損失関数 $\mathcal{L}$ を最小化する．

$$\mathcal{L} = -\sum_{\mathcal{D}_{train}} \log p(C|\boldsymbol{h}; \boldsymbol{\theta}), \tag{8}$$

$\mathcal{D}_{train}$ は訓練データを指し，$\boldsymbol{\theta}$ はモデルのパラメータを指す．

5　実験

画像描写作文の誤りを訂正する実験を実施する．

4）　式中ではバイアス項を省略して表記する．

5.1　実験設定

擬似データによる事前学習．　誤り訂正の精度を高めるために，大規模で容易に入手可能な画像キャプションに擬似的な誤りを生成することで，訓練データを拡張する．本研究では，GEC の擬似的誤り生成手法 [13] を用いて，MS-COCO のキャプションから誤り文を生成した．ここで構築した擬似データ[5] を用いてモデルの事前学習を実施する．3 節で収集したデータセットは，事前学習されたモデルの fine-tuning に用いる．データは画像ごとに訓練データ/開発データ/テストデータに 6:2:2 の割合で分割した．

ハイパーパラメータ．　最適化手法には Adam [14] を使用し，ミニバッチサイズは事前学習時に 64，fine-tuning 時に 8 とした．次元 d_v と d_e は，ResNet と BERT の出力する特徴量の次元に合わせて，それぞれ 2,048 と 768 としている．また，学習率は 5.0×10^{-5} とした．

評価尺度．　GEC の標準的な評価尺度である ERRANT [15] と GLEU [16] を用いた評価を実施する．これらの評価は，システムが画像を参照した訂正の評価については十分でないため，今後改善する必要があると考えている．

モデル．　提案手法と比較するモデルとして，既存の GEC モデルの GECToR を用いる．また，提案手法では，次のモデルを用いて実験を実施した．

- **L-C**: 学習者作文 L のみを用いて，訂正文 C を生成する[6]．
- **LI-C**: 学習者作文 L と画像 I を用いて，訂正文 C を生成する．4.1 で述べた通り，画像の特徴量抽出は，Global と Bottom-up をそれぞれ適用する．

5）　訓練用に 画像 40,186 枚, キャプション 201,059 文，開発用に 画像: 5,000 枚, キャプション 25,014 文を用いた．

6）　L-C モデルの fine-tuning 時のみ学習率を 1.0×10^{-3} とした．

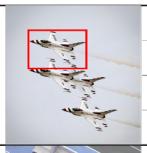
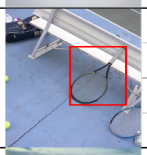
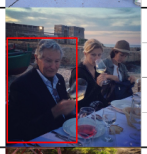
(a)		学習者作文	That air is flying of sky.
		専門家による添削	A jet is flying in the sky.
		GECToR	That air is flying of sky.
		提案手法 LI-C (Bottom-up)	A fighter is flying in the sky.
(b)		学習者作文	A tennis racket is black clour on the bench.
		専門家による添削	The tennis racket is black and under the bench.
		GECToR	A tennis racket is black clour on the bench.
		提案手法 LI-C (Bottom-up)	A black tennis racket is by the bench.
(c)		学習者作文	A old man is having wain grass.
		専門家による添削	An old man is holding a wine glass.
		GECToR	An old man is having wain grass.
		提案手法 LI-C (Bottom-up)	A man is having a conversation.
(d)		学習者作文	A month is lonely
		専門家による添削	A horse is lonely.
		GECToR	A month is lonely
		提案手法 LI-C (Bottom-up)	This is a very big.

図 4　提案手法による訂正文と既存の GEC モデル GECToR が出力する訂正文の比較．学習者作文との差分を赤字で示す．（画像中の赤枠を中のものを参照した作文を学習者に求めた）

5.2　定量評価．

表 2 に，ERRANT で算出した適合率，再現率，$F_{0.5}$ 値と GLEU の値を示す．

GECToR との比較．　GECToR と比較すると，LI-C モデルは ERRANT の再現率，GLEU において性能を上回った一方で，ERRANT の適合率，$F_{0.5}$ 値では下回る結果となった．これは，GECToR が訂正箇所を最小限に抑える一方で，LI-C モデルは学習者作文を改変し，画像を基にした文を生成する傾向にあることが原因にあると考えられる．学習者作文の構文・語彙に近い訂正を実現するために，コピー機構 [3] を取り入れることが解決策の一つにある．

また，GECToR と L-C モデルは，どちらも画像を活用しないが，精度に差が生まれた．これは，GECToR が大規模なコーパスを用いて学習されていることから明らかである．

画像の特徴量．　LI-C モデルは全ての評価値において，L-C モデルの性能を上回った．これは，誤り訂正における画像の情報の有用性を示す．また，画像の特徴量抽出では，“Bottom-up” が 全ての評価値において “Global” の性能を上回った．画像内の物体を考慮した特徴量の抽出が，学習者作文を訂正する上で有用であったことがわかる．

5.3　定性評価．

提案手法LI-C (Bottom-up) による訂正文と GECToR が出力する訂正文の比較を図 4 に示す．GECToR が画像の情報を活用できていない一方で，提案手法は画像に関連する学習者の誤りを訂正できていることがわかる ((a)“air” を “fighter” に訂正)．また，提案手法では位置関係の訂正もできた ((b)“under” を “by” に訂正)．一方で，学習者の描写内容を反映できていない誤り訂正の失敗例も確認した ((c)“wine glass” を “conversation” に訂正)．モデルにコピー機構を設けることで，学習者作文の構文・語彙をより考慮した訂正文生成を行うことが解決策の一つに考えられる．

表 2　GEC の評価尺度による評価．最良の結果を太字で示す．

モデル	画像	ERRANT			GLEU
		適合率	再現率	$F_{0.5}$ 値	
GECToR	-	**0.367**	0.116	**0.256**	0.268
L-C	-	0.105	0.126	0.108	0.226
LI-C	Global	0.147	0.168	0.151	0.263
LI-C	Bottom-up	0.165	**0.190**	0.170	**0.284**

6　おわりに

外国語学習者による画像描写の自動評価に取り組んだ．具体的には，学習者作文の誤りを画像情報を用いつつ訂正する方法を研究した．画像と学習者による画像描写，そして，専門家による訂正文から構成されるデータセットを整備し，それを用いて画像描写問題における学習者作文の自動訂正ベースラインを構築した．画像描写の誤り訂正について画像が有用であることを確認した．また，GEC では訂正できない誤りを本手法が訂正できることを確認した．

謝辞

本研究は JSPS 科研費 19K12119 の助成を受けたものである．

参考文献

[1] Hwee Tou Ng, Siew Mei Wu, Ted Briscoe, Christian Hadiwinoto, Raymond Hendy Susanto, and Christopher Bryant. The CoNLL-2014 shared task on grammatical error correction. In **Proc. of CoNLL**, pp. 1–14, 2014.

[2] Christopher Bryant, Mariano Felice, Øistein E. Andersen, and Ted Briscoe. The BEA-2019 shared task on grammatical error correction. In **Proc. of BEA**, pp. 52–75, 2019.

[3] Wei Zhao, Liang Wang, Kewei Shen, Ruoyu Jia, and Jingming Liu. Improving grammatical error correction via pre-training a copy-augmented architecture with unlabeled data. In **Proc. of NAACL**, pp. 156–165, 2019.

[4] Kostiantyn Omelianchuk, Vitaliy Atrasevych, Artem Chernodub, and Oleksandr Skurzhanskyi. GECToR – grammatical error correction: Tag, not rewrite. In **Proc. of BEA**, pp. 163–170, 2020.

[5] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In **Proc. of CVPR**, pp. 11–20, 2016.

[6] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context. In **Proc. of ECCV**, pp. 740–755, 2014.

[7] Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton Van Den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In **Proc. of CVPR**, pp. 3674–3683, 2018.

[8] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In **Proc. of ICML**, pp. 2048–2057, 2015.

[9] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In **Proc. of NeurIPS**, pp. 91–99, 2015.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In **Proc. of CVPR**, pp. 770–778, 2016.

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In **Proc. of NAACL**, pp. 4171–4186, 2019.

[12] Qing Li, Qingyi Tao, Shafiq Joty, Jianfei Cai, and Jiebo Luo. VQA-E: Explaining, elaborating, and enhancing your answers for visual questions. In **Proc. of ECCV**, pp. 552–567, 2018.

[13] Abhijeet Awasthi, Sunita Sarawagi, Rasna Goyal, Sabyasachi Ghosh, and Vihari Piratla. Parallel iterative edit models for local sequence transduction. In **Proc. of EMNLP-IJCNLP**, pp. 4259–4269, 2019.

[14] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In **Proc. of ICLR**, 2015.

[15] Christopher Bryant, Mariano Felice, and Ted Briscoe. Automatic annotation and evaluation of error types for grammatical error correction. In **Proc. of ACL**, pp. 793–805, 2017.

[16] Courtney Napoles, Keisuke Sakaguchi, Matt Post, and Joel Tetreault. Ground truth for grammatical error correction metrics. In **Proc. of ACL-ICJNLP**, pp. 588–593, 2015.