画像キャプションを用いた日本語学習支援の検討

川端 公貴 † 南條 浩輝 ◇ 亀甲 博貴 ◇ 森 信介 ◇
† 京都大学 総合人間学部 ◇ 京都大学 学術情報メディアセンター
kawabata.kouki.76c@st.kyoto-u.ac.jp nanjo@media.kyoto-u.ac.jp
{kameko, forest}@i.kyoto-u.ac.jp

1 はじめに

深層学習の発展によって，自然言語と視覚といったマルチモーダルな情報を融合する研究が近年広がりを見せている．その中でも画像を入力として説明文を出力する画像キャプション生成については Show and Tell [1] 等を皮切りに研究が進んでおり，この分野の中心的な存在である．画像キャプション生成技術を応用した研究には，レシピ生成 [2] や自動運転 [3] などがあり，また画面付きスマートスピーカー Echo Show なども自分が手に持っている物体を認識して音声で伝える機能を視覚障害者向けに提供している [4]．しかしながら，語学教育の分野に適用させた応用研究の例はいまだ少ない．

我々は本研究を画像キャプションを用いた語学学習者支援システム構築に向けての基礎研究と位置づけ，画像を用いた語学教材のあり方やその効果，また学習者に対してふさわしいフィードバック支援の方法について研究している．本論文では，その第一ステップとして，画像に付与された日本語キャプション文の一部を空白にして穴埋め問題を作成し，日本語母語話者と日本語学習者の双方に対して欠落している言葉を埋めさせるタスクを検討した．そこで得られた回答結果をもとに，画像の有無による正答率の比較，学習者の誤り傾向の分析を行い，語学教育への応用可能性を検討した．

2 関連研究

日本語学習支援システムに関する研究には，ライティング学習支援についていえば，学習者の書いた作文を自動で文法誤り検出/訂正するもの，作文する際にキーワードを推薦するもの，学習者の誤用例文を検索するものなどがある．一方，一部が欠落した文章と画像/動画をもとに空欄を埋めるのに適切な語を生成するタスク設定の研究 [5, 6] も存在する．これらは 1-best の単語を生成することを目的としているのに対して，本研究の目的は学習者のあらゆる回答に対して自動評価および適切なフィードバックを与えることである．画像情報は言語非依存であるため，教材作成においても学習者の母語によって一つ一つ作り分ける必要がないのが画像を用いた教材の大きな利点と言える．

3 実験

本研究の目的は，画像情報をメディアとして用いた語学学習者の産出支援である．これには (1) 画像を提示して文章を産出させるための教材の自動作成，(2) 産出させた文章の自動評価（採点）の 2 段階が含まれる．

はじめに予備実験として，日本語学習者に画像を提示して自由に作文させるタスクを試みた．しかし，収集した作文には「おしゃれな部屋の写真です」のような当たり障りのない短文が目立ち，画像を与えられても何を描写すればいいのかよく分からないという意見が多く見られた．そこで，今回はキャプション文の一部を抜いた穴埋め問題のタスクとした．こうすることで学習者がどの部分を描写すればいいか明確になり，また産出能力のみならず視覚情報を加味した外国語理解能力をも訓練する手助けになると期待できる．以下に今回の手法の詳細を示す．

3.1 問題セットの作成

本研究では日本語学習を対象とした．そのため，問題セットの作成に日本語画像キャプションデータセットを利用した．既存のデータセットには主に YJ Captions 26k Dataset [7] と STAIR Captions [8] の二種類が存在し，いずれも画像データに MS-COCO [9] を用いている．画像数，キャプション数に関して，前者が 26,500 画像-131,740 キャプションに対し，後者が 123,287 画像-616,435 キャプションと前者を大きく上回っている．本論文では，日本語学習者に対する文体を考慮して「です・ます体」を採用する YJ Captions 26k Dataset を問題セット作成に使用することにした．このなかから 60 組の画像-キャプション対を抽出し，各キャプション文の一箇所を空白にして穴埋め問題を作成した．穴埋め対象とした語は名詞，動詞，格助詞の 3 種類とし，それぞれ 20 ずつの合計 60 問を作成した．動詞に関しては，表 1 に示す 3 パターンのいずれかにマッチす

表 1: 問題作成（動詞）のテンプレート

1:	[動詞] [語]{0,2} て い ま
2:	[動詞] [語]{0,2} て い る
3:	[動詞] [語]{0,2} た [名詞]

注 1: 形態素解析として KyTea [10] を使用
注 2: [語] は KyTea で分割された任意の単語
注 3: X{0,2} は X の 0 回以上 2 回以下のくり返し

るものを穴埋め対象とした．格助詞に関してはデータセットの中で使用頻度の多い順に 5 種類（が，の，に，を，で）を対象とした．作成された問題の一例を次に示す．

1. 名詞: 草原の川をシマウマの____がわたっています
2. 動詞: 皿にクラッカーとサラダが____ています
3. 助詞: 飛行機の機内から隣の飛行機____見えます

本論文では，画像を与えたときの作文を評価する．比較として画像を与えないときの作文も評価する．このため，用意した 60 問を 30 問ずつ（名詞，動詞，格助詞，各 10 問）の 2 セット（セット A，B）に分割し，被験者も 2 グループ（グループ 1，2）に分割する．グループ 1 には画像ありの問題セット A と画像なしの問題セット B，グループ 2 には画像なしの問題セット A と画像ありの問題セット B をそれぞれ与えて穴埋めさせ，評価を行う．問題の回答の際には，辞書や翻訳サービス等を利用したり他人に頼ることをしないよう指示し，画像ありの場合は「画像に対する適切な説明文となるように空欄を埋めてください」というインストラクションのみを．画像なしの場合は「画像はありません，文章のみを見て空欄を埋めてください」というインストラクションのみを与え，埋める品詞については言及しない．ただし格助詞の問題については「が，で，に，の，を， のどれかで答えてください」というインストラクションを付け加えた．

3.2 回答データの収集

グループ 1， 2 それぞれに日本語母語話者 3 名と日本語学習者 8 名の合計 22 名を割りあて，穴埋めさせた．問題は WEB 上で回答させた．回答画面（画像なし）の例を図 1 に示す．なお，画像ありの場合は，このレイアウトのままで画像がスペースに挿入され，インストラクションの文言が画像ありバージョンに変わる．回答画面のインストラクションは，日本語または

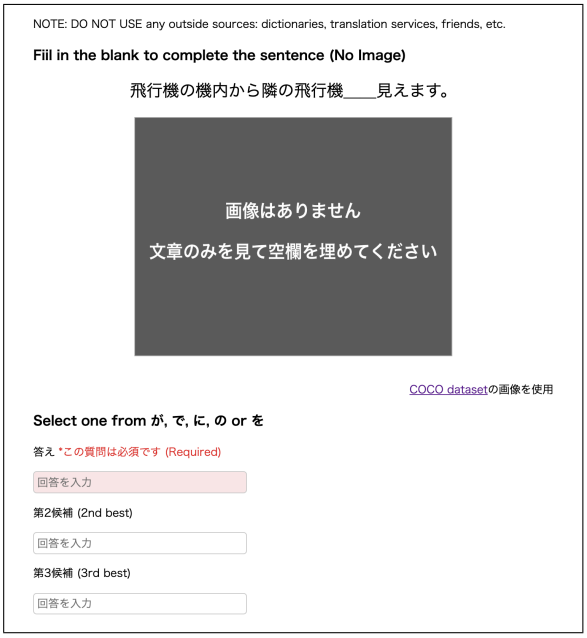

図 1: 実際の回答画面（画像なし）

英語で行った（被験者が切替可能）．被験者は日本語入力を使用して，漢字，かな，カナなどで回答した．1 つの問題につき答えを 3 つまで回答できる（1 つは必ず回答）こととした．グループ 1 の日本語学習者の母語の内訳は，中国語（諸方言を含む）5 名，英語，ロシア語．イタリア語が各 1 名，グループ 2 の日本語学習者の母語の内訳は，中国語（諸方言を含む）6 名，英語．タイ語が各 1 名であった．日本語能力試験のレベルは中国語語母語話者の 1 名（未受験）を除き全員 N3 以上だった．

3.3 回答結果の評価

回答結果ははじめに日本語母語話者により人手で評価を行う．評価基準は機械翻訳で用いられている翻訳文の人手評価の方法 [11] を参考にし，流暢さ (fluency) と適切性 (adequacy) の 2 側面でレベルを判定することにした．採点者は表 2 に示す通りの「採点基準」となる語句と点数を与えられ，自身の直感で点数化する．

4 分析

回答データを集計すると全 60 問で計 605 個の異なる回答が得られた．その際，動詞問題における明らかな入力の間違い（語尾の “て” を余分に書いている等）は修正し，空白や「わかりません」などの回答は “NONE” に置き換えている．

表 2: 採点基準

流暢さ (fluency)：画像と関係なく，穴埋めされたキャプション文自体が，日本語としてどの程度「自然な」表現であるか
5.　まったく問題ない
4.　良い
3.　非母語的
2.　不自然
1.　理解不能

適切性 (adequacy)：穴埋めされた言葉によって，画像の情報がどの程度「正しく」説明されているか
5.　完全に正しい
4.　ほとんど正しい
3.　多くは正しい
2.　少し正しい
1.　正しくない

表 3: 採点者間の一致率（重み付きカッパ係数）

	流暢さ	適切性
名詞	0.689 ± 0.114	0.742 ± 0.111
動詞	0.587 ± 0.116	0.665 ± 0.088
助詞	0.727 ± 0.147	0.602 ± 0.310
全て	0.684 ± 0.111	0.670 ± 0.144

4.1　予備実験

3.2 節で述べた被験者による回答の収集を行う前に，評価基準の妥当性を確かめるための予備実験を実施した．3.1 節で述べた方法で，異なる画像に対して名詞，動詞，格助詞の問題を 8 組ずつ選択して全 24 問の問題セットを作成した．回答データについては学習者の実際の答えや Show, Attend and Tell [12] のモデルにおいて生成確率の比較的高かったものから筆者が妥当な答えと思われるものを抽出して合計 200 個の擬似回答データを収集した．

これに対して筆者を含めた 5 名の日本語母語話者が 3.3 節の方法に基づいて採点を行った．採点者間の一致度を調べるために，採点者のペア全 10 組について二次の重み付けカッパ係数 (Quadratic Weighted Kappa) を求め，その平均と標準偏差を求めた．結果を表 3 に示す．流暢さ，適切性のそれぞれについて，一致率は 0.684，0.670 であり，標準偏差は 0.1 程度であった．このことより，人間評価は少なくとも moderate な一致率があるといえ，本論文ではこの評価方法を採用する．

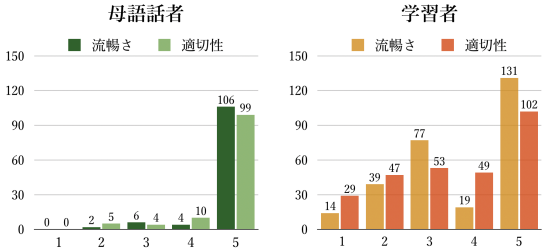

図 2: 点数分布

4.2　本実験の採点結果

予備実験で moderate 以上の一致率が得られたため，3.3 節の方法で本実験の回答データに対して筆者（1 名）が採点を行った．画像ありの場合の全ての種類の回答に関して，点数の分布を日本語母語話者 (計 118 個) と日本語学習者 (計 280 個) についてそれぞれ調べた．結果を図 2 に示す．学習者の回答は点数のばらつきが大きいことが見てとれる．

加えて母語話者と学習者間の画像の有無による点数の比較を行ったところ，表 4 のような結果となった．平均点の算出に用いたものは第 1 候補の回答のみで，回答 “NONE” は 1 点として計算した．流暢さにおいて，母語話者と比べ，学習者の画像ありに対する画像なしの点数の下がり幅が大きいという結果が得られた．特に名詞穴埋め問題においてその傾向が顕著に現れている．これは「____をかけた野球選手がいます」といった問題に対して，「かける」に共起する名詞を画像なしに想起するのは学習者にとって困難であることが原因として挙げられる（“人生”，“電話” などが見られた）．また助詞問題の平均点が，画像なしに比べ画像ありが全体的に上昇しているのは興味深い結果である．

4.3　誤り傾向

本研究のタスク設定として，学習者は穴埋めする語句を画像情報に即して適切に回答することが重要となってくる．全回答者の画像ありでの回答は計 324 種類存在し，うち「適切性」が 3 以下であるものが 133 あった．これについて表 5 のように 4 種類に誤り傾向の分類を行った．タイプ A は画像の着目すべき対象がずれていて画像情報を正しく表現できていないもの，タイプ B は助詞の誤り，動詞/助動詞の活用に関する誤りを指す．タイプ C は画像の着目すべき対象は把握できているが日本語としての単語選択が不適当なもの，タイプ D は語彙としては正しいが表記の上で誤っているものを指す．

言語誤りについては，従来から用いられている言語モデルや構文解析によってある程度判定できると考えられる．ただし誤って埋めた語が画像の表す内容と一

表 4: 画像の有無による平均点の比較

流暢さ

	母語話者		学習者	
	画像あり	画像なし	画像あり	画像なし
名詞	4.95	**4.88**	4.41	**3.96**
動詞	4.83	4.67	3.96	3.70
助詞	4.97	4.93	4.74	4.65
全て	4.92	**4.83**	4.37	**4.11**

適切性

	母語話者		学習者	
	画像あり	画像なし	画像あり	画像なし
名詞	4.83	3.45	4.23	2.61
動詞	4.82	4.13	3.99	3.15
助詞	4.98	4.95	4.84	4.80
全て	4.88	4.18	4.36	3.52

致していなくても，日本語として適切なものであれば，言語モデルや構文解析だけでは判定が難しいと考えられる．例えば「女性が____を持っています」については，表 5 の例の「笑顔を持つ」は日本語の表現として不自然であり，言語モデル等で検出可能と考えられる．しかし，正解である「グラス」以外，例えば「コップ」「ガラス」などを入れた場合（タイプ C の誤り）は，日本語文としては自然であり，画像がないと誤りと判定できない．画像があれば「女性が____を持っている」「____を被る」という表現が与えられたときの画像に対するアテンションと埋められた語が画像に存在するか，存在すればどこに存在するかなどからも判定できる可能性もある．

タイプ B, C, D の誤りに関しては，作文タスクにおける学習者の典型的な誤用の範疇に含まれるものが多いが，タイプ A の誤りは「____を被る」に対するコロケーションの知識不足などが招くものであり，学習者にとっては共起表現パターンの知識幅を画像の提示によって広げるきっかけとなり得る．

5 おわりに

画像を用いた語学学習者の作文支援研究における初期段階の試みについてまとめた．収集した穴埋め回答データを用い，自動採点モデルの構築について現在研究中である．

謝辞: 本研究は科研費（19K12119）の補助を受けて行われた．

表 5: 誤り傾向の分析

分類	数	誤り例
A	37	球場で**黒い手袋**を被って (ヘルメット) 女性が**笑顔**を持っています (グラス)
B	31	[助詞] 川沿い**に**草原に牛が集まっています [活用] バナナがたくさん**置い**ています
C	58	曲がりくねった**軌道**を一本の電車が走行している (線路) 止まれの**指示**が立っています (標識)
D	7	[カタカナ] シマウマの**ファーミリー** [誤字脱字] 蓋が**空い**ています

参考文献

[1] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In *CVPR*, 2015.

[2] T. Nishimura, A. Hashimoto, and S. Mori. Procedural text generation from a photo sequence. ACL, 2019.

[3] Y. Mori, H. Fukui, T. Hirakawa, J. Nishiyama, T. Yamashita, and H. Fujiyoshi. Attention neural baby talk: Captioning of risk factors while driving. 2019.

[4] M. Coyle. Alexa, what am i holding? `https://blog.aboutamazon.com/devices/alexa-what-am-i-holding`, 2019. (visited on 2019-01-15).

[5] T. Maharaj, N. Ballas, A. Rohrbach, A. Courville, and C. Pal. A dataset and exploration of models for understanding video data through fill-in-the-blank question-answering. In *CVPR*, 2017.

[6] Q. Sun, S. Lee, and D. Batra. Bidirectional beam search: Forward-backward inference in neural sequence models for fill-in-the-blank image captioning. In *CVPR*, 2017.

[7] T. Miyazaki and N. Shimizu. Cross-lingual image caption generation. ACL, 2016.

[8] 吉川友也, 重藤優太郎, 竹内彰一. STAIR Captions: 大規模日本語画像キャプションデータセット. 言語処理学会 第 23 回年次大会, pp. 537–540, 2017.

[9] T. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014.

[10] G. Neubig, Y. Nakata, and S Mori. Pointwise prediction for robust, adaptable japanese morphological analysis. ACL, 2011.

[11] 隅田英一郎, 佐々木裕, 山本誠一. 機械翻訳システム評価法の最前線. 情報処理, pp. 552–557, 2005.

[12] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, 2015.