

L-VEIGe : 誤答画像生成を活用した語彙学習支援システム

— 画像の示唆性に対する指標の検討 —

杉田 一樹^{*1}, 太田 光一^{*1}, 谷 文^{*1}, 長谷川 忍^{*1}

^{*1} 北陸先端科学技術大学院大学

L-VEIGe: Vocabulary Learning Support System

Using Error Image Generation

-A study of criteria for image suggestibility -

Kazuki SUGITA^{*1}, Koichi OTA^{*1}, Wen GU^{*1}, Shinobu HASEGAWA^{*1}

^{*1} Japan Advanced Institute of Science and Technology

The L-VEIGe (Learning-Vocabulary Error Image Generation) system for supporting vocabulary learning has been shown to promote learners' introspection and prevent repeated mistakes by visualizing errors through image generation. Learners' impressions and memorability are considered important criteria for the generated images. This paper examines a quantitative method for learning efficiency and addresses the problem of feature disappearance in the generated images that is necessary to understand the vocabulary in the errors.

キーワード: 語彙学習支援システム, 第二言語習得, 誤答画像の示唆性, 生成画像評価指標

1. はじめに

英語語彙学習は英語を第二言語とする L2 学習者に於いて重要とされ, 語彙はコンテキストから学習する方法が有効であるとされている⁽¹⁾. 語彙学習では, 画像を用いた研究が多く存在し, 画像から対応する文を自動で生成するイメージキャプションを活用した語彙学習支援システムが存在する⁽²⁾. ここでは, 画像をコンテキストと見做すことが可能だとし, 膨大な読書量に対する画像の代替の可能性を示している. しかしながら, これらは学習時の誤答については言及されていない. 一方, 英語学習に於いて, 誤った認識のまま繰り返し用いる事によることに依る化石化「fossilization」の問題がある⁽³⁾. 化石化では, その誤りが一度定着すると, その修正は困難であるとされている. つまり, 学習者の語彙に対する誤りを効果的に認識させることが重要であると捉えられる. そこで筆者らは, 英語語彙補充問題を基本とし, 学習者の誤答

を画像生成により可視化することにより学習者の誤答に対する内省を促す, 語彙学習支援システム L-VEIGe(Learning-Vocabulary Error Image Generation)を提案している.

本稿では, 提案手法と評価実験の概要について述べ, 生成される画像の示唆性に於ける想起性に関する課題である特徴消失について議論し, その解決策として想起性の定量的評価指標を提案する⁽⁴⁾.

2. 提案手法

ここでは, L-VEIGe の構成について述べる. 既存のデータセットである MS-COCO データセット⁽⁵⁾を活用する事により, 以下のタスクを定義し, ①~③を実現した.

- ① 質問生成タスク: 学習用教材として網羅的である問題の自動生成を実現,
- ② 回答支援タスク: 語彙学習に於いて近似したスベルが間違いを誘発するとされることから, 近

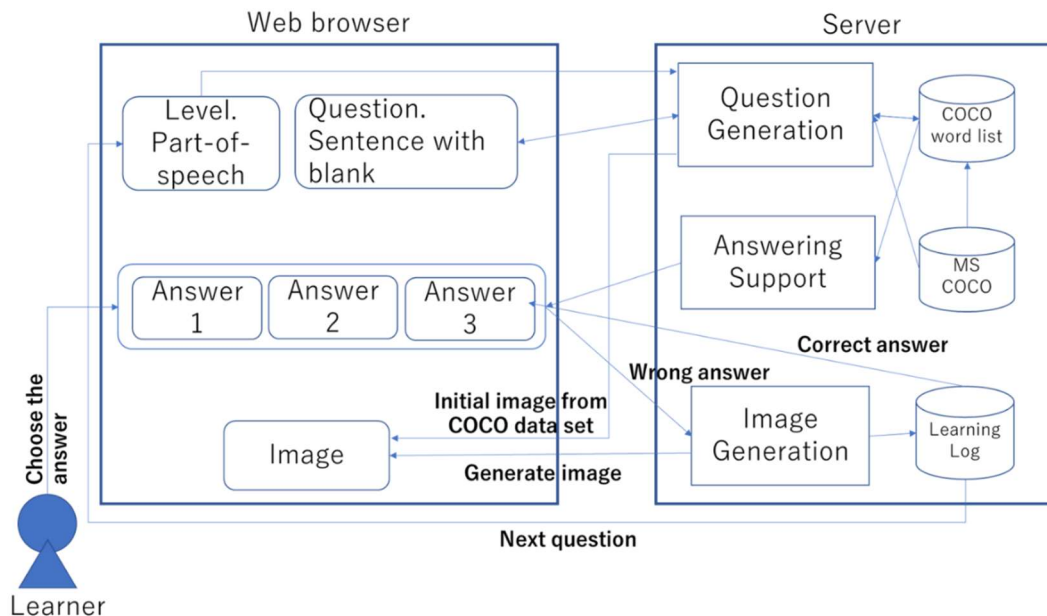


図 1 L-VEIGe ユースケース図

似したスペルを有する語彙を学習対象語彙の回答の候補として自動的に抽出し、誤答を促す環境を構築。

- ③ 画像生成タスク：文に応じた画像を生成可能とする DALL-E⁽⁶⁾を活用し、学習者の誤答に応じた画像生成を実現。

これら①～③は、図 1 に対応する。図 2 は学習者の誤答に応じ、画像が変化した様子を表す。ここでは、学習者は、初めに左画像とその画像に対応する文「Two sheep and a () stand next to a fence in the yard」に対し、正答が ram となるが、学習者が tram を選択した為、誤答を含む文に対応した画像が提示される。

3. 評価実験

繰り返しの誤答の回避、誤答した語彙に対する理解を評価する為に、英語を第二言語とする TOEIC 500～900 点を有する、8 カ国 24 名の理系大学院生を対象とし実験を行った⁽⁷⁾。実験では、システムエラーに依るデータ欠損の為、被験者 22 名から得られたデータを評価対象とした。被験者は、誤答時の画像生成を有する Condition1 と画像生成を有さない Condition2 の両方に依る学習を連続で行い、2 つの条件での連続した学習を 1 タームとし、1 ターム目と同じ問題、条件を繰り返す合計 2 タームに依る実験を行った。被験者は同一の問題が与えられ、条件の順序、問題の順序に

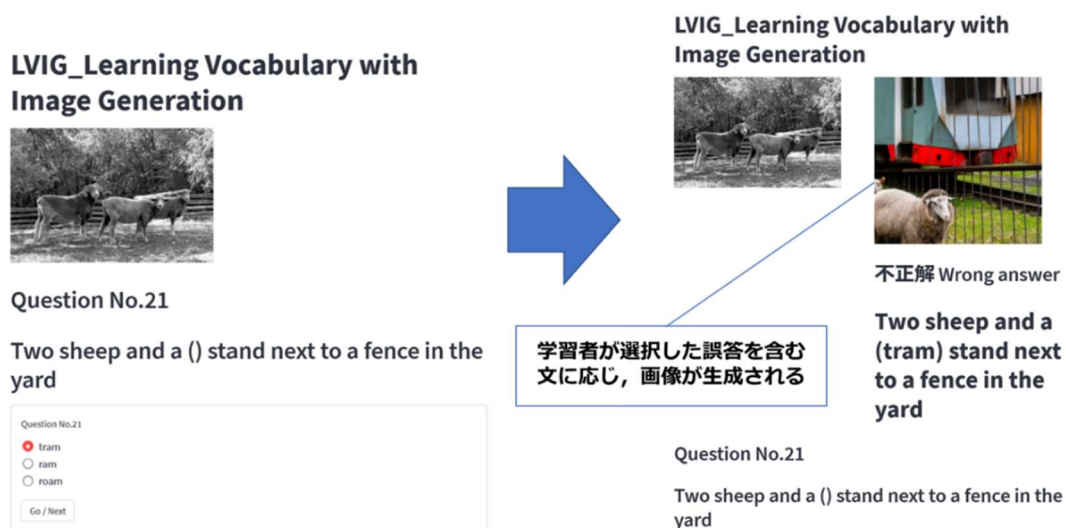


図 2 L-VEIGe システム操作例

依る効果を解消した被験者内実験に依るカウンターバランスを取った。又、学習対象者が繰り返しの誤答を避けられたか判定する為、意図的に 20% すべての選択肢が誤りである問題を混在させた。又、生成画像を評価する上で、画像に対する、印象、想起性の二つの評価軸を定義し、画像の示唆性と定義した。被験者は初期画面に表示される画像、誤答時に生成される画像に対して、

- A. 初期画像の想起性：初めに表示される画像は文を想起させるものか、
- B. 生成画像の想起性：生成された画像は文を想起させるものか、
- C. 生成画像の印象：生成された画像は印象に残るか、

の 3 点について 5 段階評価を行った。又 3 日後に被験者が誤答した語彙に対し理解度を問うアンケートを行った。

3.1 繰り返しの誤答の回避に対する評価

ここでは、学習者が繰り返しの誤答を避けられたか否かの評価を行い、ターム 1、ターム 2 に於ける、Condition1, 2 による誤答の繰り返し数を基に評価を行う。ここでは、評価対象者 22 人の結果に対して、ノンパラメトリックな手法である Wilcoxon の符号順位検定を行う。ここでは、帰無仮説 H_0 を Condition1 と Condition2 の繰り返しの誤答の回数に差がないとする。検定により導出された 2 群間のデータの差の絶対値を W と定める。サインランク表を用いた検定により、片側有意水準 $\alpha=0.05$ 、棄却限界値 53、 $W=45$ 、 $W<53$ より、帰無仮説を棄却する。図 3 に示す度数分布表より、提案手法である画像生成を伴うモデルが繰り返しの誤答を防ぐうえで有効である事を示した。

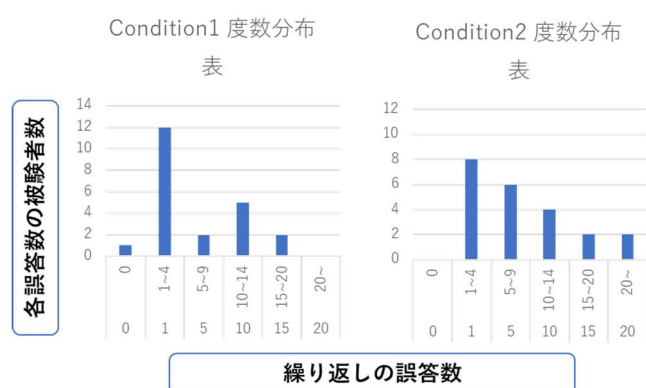


図 3 印象と繰り返しの誤答数との相関関係

3.2 生成画像の特性と繰り返しの回避の関係

学習者に依る定性評価による、A. 初期画像の想起性、B. 生成画像の想起性、C. 生成画像の印象のうち、誤答に関する評価である、B,C について、繰り返しの誤答数との相関をノンパラメトリックな指標である、Spearman の順位相関係数に依る検定を行った。想起性と繰り返しの誤答数についての検定では、帰無仮説 H_0 を想起性と繰り返しの誤答数に相関関係はないとする。図 4 左では、y 軸を画像の想起性、x 軸を誤答の繰り返し回数を表している。図 4 右ではスピアマンの順位相関係数である ρ 、P 値を示す。ここでは、 $\rho=5.57$ 、 $P=0.99>0.05$ より帰無仮説を採択する。つまり、想起性と繰り返しの誤答数については有意な相関を得られなかった。印象と繰り返しの誤答数についての検定では、帰無仮説 H_0 を印象と繰り返しの誤答数に相関関係はないとする。図 5 では、 $\rho=-0.24$ 、 $P=0.03>0.05$ より帰無仮説を棄却する。ここでは、印象と繰り返しの誤答数に於いて、弱い負の相関が得られた。これらの両者の比較では、生成画像の想起性より、印象の度合いがより繰り返しの誤答に影響を与えていることが伺える。

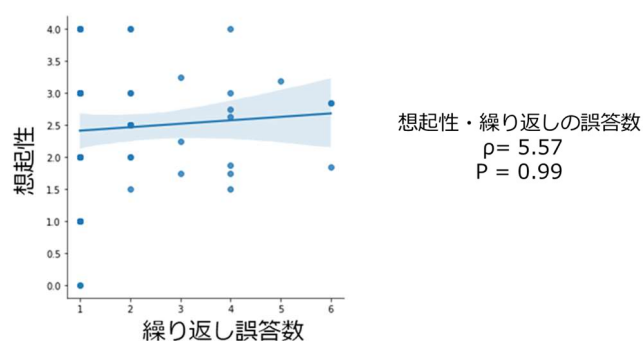


図 4 想起性と繰り返しの誤答数との相関関係

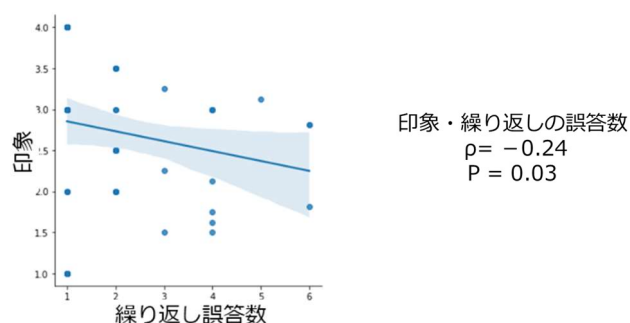


図 5 印象と繰り返しの誤答数との相関関係



図 6 特徴消失した生成画像例

3.3 誤答に対する理解に対する評価

遅延テストでは、誤答に対する理解度として、被験者ごとに誤った語を提示し、それらの語彙に対して、“既に知っている”、“想像できる”、“想像できない”のいずれかを回答する方式で行った。被験者に於ける各コンディションの理解度の評価として、“想像できる”と回答した語彙数から“想像できない”と回答した語彙数の差を誤った語の理解度の評価値とした。先述した繰り返しの誤りの回避に対する評価同様、両者に於ける誤答の理解度に対する評価として、ノンパラメトリックな手法である Wilcoxon の符号順位検定を行う。ここでは、帰無仮説 H_0 を Condition1 と Condition2 の繰り返しの誤答の回数に差がないとする。サインランク表を用いた検定により、片側有意水準 $\alpha=0.05$, 棄却限界値 41, $W=49$, $W>41$ より、帰無仮説を採択する。つまり、誤答した語彙の理解度について、有意な差は見られなかった。繰り返しの誤りの回避に対する評価の結果、今回の結果より誤答への効果的な印象を与え、繰り返しの誤答を防げたが、誤った語を理解するという点に於いては、不十分である可能性を有していることが伺える。

3.4 生成画像と誤答の理解の関係

3.4.1 特徴消失問題

誤った語に対する被験者の回答で、“想像できなかった”という回答が 2 回以上選ばれた語彙について生成画像より分析を行う。“yawn”, “ecstatic”, “fence” を例示する。図 6 では、それら誤った語を入力とした際に生成される画像、及び問題を示している。青色で記された語彙は、その時の正答を示す。図 6 下部の値については、それらの語の誤りの“繰り返し回数”、“想

起性”、“印象”の値を示す。ここでは、これらの生成画像に於いて、画像とセンテンスの対応を表す、“想起性”に於いて 2 以下の低い数値を示した。ここでの“想像できなかった”とされる画像に共通している点として、図 6 より空所として補充した語彙が画像に反映されない特徴消失が発生していることが伺える。

3.4.2 特徴消失問題に依る誤解の誘発

特徴消失によって、明示的に誤答の意味が画像に示されない場合、学習者は、誤答であることを正しく認識できないまま語彙を捉えることが危惧される。図 7 では、学習者が左図の画像に対する回答として、「decorative」を選択することによって、生成された画像を示している。一方、右図では、生成された画像は四角枠で示す通り黒い板が生成され、これによって、学習者は「decorative」の意味を黒い板と関連して記憶する可能性を有する。ここでは、「decorative」の本来の意味は「装飾的な」等の意味を有することから異なる認識を与える原因となりうる。



図 7 誤解を誘発する生成画像例

a:学習者の誤答に応じた生成画像



b:学習者の誤答を含む問題文

Two sheep and a (tram) stand next to a fence in the yard



c:生成されたキャプション

1. A couple of groups of brown sheep standing on the side of a track.
2. A sheep looking at a water trough while walking with a person behind it. and the
3. A flock of white and green sheep walk by a barn and two black sheep looking out
4. A group of brown and black sheep looking toward a group of white sheep outside a barn
5. The two sheep are standing next to others on the side of a field.

d: b - cの近似度

1. 0.72919154
2. 0.6544069
3. 0.7812311
4. 0.72477007
5. 0.891336

図 8 コンテキストの類似度の推定

4. 誤答に対する想起性の指標

評価実験より、誤答時の効果的な印象付けが期待でき繰り返しの誤答を回避する上で有効であることを示す一方で、印象と想起性の二つの軸により定義した画像の示唆性について、特に想起性について特徴消失の問題を有する。誤答への概念理解という観点では、生成される画像が誤答を理解するに足る情報を有している必要があり、従来の手法では、生成される画像の特徴の消失に対する定量的な基準がない為に不十分である。ここでは想起性に対する定量的な評価基準として、コンテキストの類似度、誤り概念の類似度、の二つの評価尺度を定義し、それぞれの方法について提案する。

4.1 関連研究

画像から生成されたキャプションに対する定量的な評価指標として、CLIP スコア⁽⁹⁾が存在する。CLIP スコアでは、画像とキャプションをそれぞれベクトル化し、その一致度を指標とすることにより、画像とキャプションの評価を可能とする。従来の手法では、正解となるキャプションと生成されたキャプションの差異に依る評価が一般的であったが、CLIP スコアでは、正解となるキャプションを必要とせず評価が可能となる。そのため、生成画像と文の全体の一致という観点では評価が可能であるが、ベクトル化された表現は可読性が低いために、特定の語彙の特徴の評価が困難である。

4.2 想起性に対する定量的評価手法

本研究では、文の特徴、誤答となる学習対象語の特徴に注目し、それらの特徴を画像が含意しているか評

価するにあたり、以下の 2 つの評価尺度を定義する。

1. コンテキストの類似度：語彙を学ぶ上でコンテキストが重要と捉えられるため、生成画像に含意する情報に十分なコンテキストを有しているか評価する。ここでは、画像に含まれる情報を自然言語化し、対象となる文と抽出した文の一致度が閾値以上であるとき、生成画像は十分な特徴を有すると定義する。
2. 誤り概念の類似度：学習者が誤答した際に選択した語彙を学習対象語と捉え、その語彙を理解するに足る情報が生成画像に有しているか評価する。誤答となる学習対象語と 1 のプロセスで生成された文中に含まれる単語の一致度を測定し、特定の語彙が閾値以上であるとき、生成画像十分な情報を有すると定義する。

上記 2 点を満たす生成画像を想起性に於ける定量的評価値として定める。

4.2.1 コンテキストの類似度

画像と文の全体の一致度を測定する上で、初めに、画像から自然言語化する方法として、イメージキャプションニングを活用し、画像の特徴を記述した文を得る。ここでは、エンコーダーとデコーダーを事前学習した CLIP⁽⁹⁾をベースに構築された、antarctic-captions⁽¹⁰⁾をモデルとして適用した。図 8 は学習者の誤答に応じた生成文と、生成画像から抽出したキャプションの一致度を表す。a は、対象となる画像を表し、b は、生成された画像の基となる学習者の誤答を含む文を示す。c では、イメージキャプションニングによって生成された画像から抽出されたキャプションを表す。次に d では、抽出した文と、元となる文の一致度を示す。ここでは、画像に含まれる情報を自然言語化する手法とし

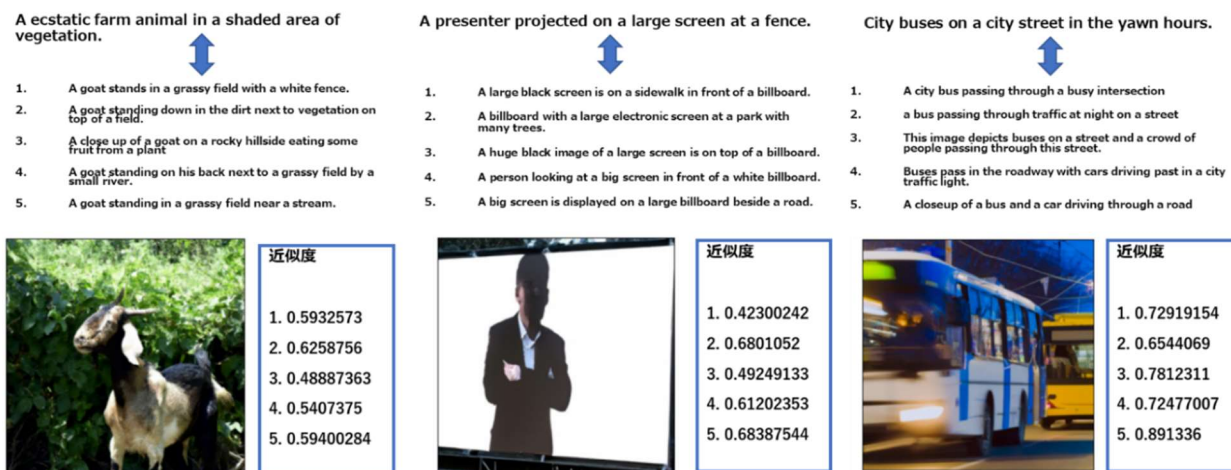


図 9 特徴消失した生成画像に於けるコンテキストの類似度

て、文脈を加味した単語の分散表現を得ることが可能である。BERT⁽¹¹⁾の拡張として、文の分散表現を得ることが可能である。Sentence BERT⁽¹²⁾を用いることで、文の一致度についてコサイン類似度を基に推定する。この時、任意の閾値を超えたとき、生成された画像は、入力に対する文の特徴を有していると評価が可能である。図 9 では、3.4 で言及した特徴消失が発生した画像に対する、生成画像と抽出した文の近似度を示す。ここでは、図 8 で示した生成画像と比較し、特徴消失を有する画像は一部のキャプションに於いて、高い数値を示す。つまり、学習対象語の特徴を有していないが、元の文のコンテキストは類似していると捉えられる。

4.2.2 誤り概念の類似度

学習対象語の一致度を測定する上で、先述した BERT のツールの一つである、BERT Tokenizer を利用し、前後の文脈を加味した対象単語の分散表現を得る方法を検討する。図 10 では、学習者の誤答を含む「Two sheep and a (tram) stand next to a fence in the yard」の文中の、誤答である学習対象語「tram」について、BERT Tokenizer により前後の文脈を加味

した単語の分散表現を取得し、その後、1 のプロセスで得た文に含まれる単語全てに対して、一致度を推定する。この時、最も一致度が高い数値が任意の閾値を超えた場合に、画像が学習対象語の特徴を含意しているか否かの定量的な評価が可能であると捉えられる。

4.2.3 評価方法の考察

現在はコンテキストの類似度に於いて、生成画像より 5 つのキャプションを抽出し、それぞれの文に対する類似度を基の画像と比較し評価した。一方、図 8 の例では、5 番目の文が最も類似度が高いとされるが、一方、学習対象語となる「tram」の特徴を有するという観点では、1 番目の生成文書に「track」という単語を有する点からも、誤りの類似度という観点では、1 番目の文書のほうが類似度が高いと言える。つまり抽出した各文書に対する類似度の推定では、特徴の見落としが発生する可能性を有する為、抽出したキャプションすべての特徴と、元となる文の特徴の比較という方法も考えられる。又、この観点からも 5 つのキャプションで十分であるかは今後検討の必要があり、どの程度のキャプションの生成が必要であるか今後の検討が必要である。

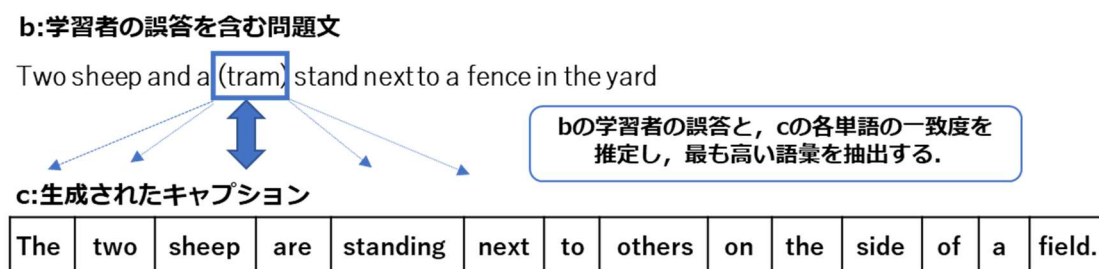


図 10 誤り概念の類似度の推定

5. まとめ

提案手法では、誤答時の画像変化が繰り返しの誤答を防ぐうえで有効である事を示した。又、画像に含まれる、印象、想起性、の二つの評価軸が生成画像の示唆性の指標として必要であると定義した。実験結果と考察より、想起性より印象の値が繰り返しの誤答を避ける上で有効であったが、想起性が欠如する特徴消失問題は、誤答を理解するに足る情報が不足する問題を孕む。そこで、誤答時の画像変化に依る語彙学習支援に於ける有効な画像を評価する上で、印象と想起性に於ける定量的評価指標の必要性に触れ、特に想起性に於ける定量的評価指標について検討した。ここでは、それぞれ、1. コンテキストの類似度、2. 誤り概念の類似度と定義し、方法論について言及した。一方、生成画像に含まれる想起性の評価指標として、1, 2 の数値に於ける閾値を定義する必要がある。今後は、1, 2 の数値と学習効果の関係性から、それぞれの数値の重みを定義する事で、生成画像の想起性に対する評価指標を策定する必要がある。又、効果的な学習という観点から、印象に対する評価指標の策定も必要となることから、今後検討していきたい。誤答時の生成画像に於ける定量的な評価指標の策定は、今後画像生成を活用した語彙学習支援システムによる基準となり、多義語、類義語といった様々な特徴を有する語彙の可視化の指標となることを期待している。

謝辞

本研究は、JST 次世代研究者挑戦的研究プログラム JPMJSP2102 の支援を受けたものです。

参考文献

- (1) Krashen, S. D. “Principles and practice in second language acquisition”, 1st ed. Pergamon, Oxford; New York, (1982)
- (2) Hasnine, M. N, Flanagan, B., Akcapinar, G., et al. “Vocabulary Learning Support System Based on Automatic Image Captioning Technology,” in Distributed, Ambient and Pervasive Interactions, vol.11587, N. Streitz and S. Konomi, eds. Springer International Publishing, Cham, pp.346–358, (2019)
- (3) Vigil N. and Oller, J. “Rule fossilization: A tentative

model,” Language Learning, vol.26, pp.281–295, Oct. (2006)

- (4) Sugita, K., Gu, W., and Hasegawa, S., “Error-Driven Learning: Development of Vocabulary Learning Support System that Utilizes Suggestibility of Error by Image Generation,” IIAI Letters on Informatics and Interdisciplinary Research, vol.3, p.1, (2023)
- (5) Lin, T.-Y., Maire, M. Belongie, S. et al. “Microsoft COCO: Common Objects in Context,” arXiv, (20-Feb-2015)
- (6) Ramesh, A., Dhariwal, P. Nichol, A., et al. “Hierarchical Text-Conditional Image Generation with CLIP Latents,” arXiv, (12-Apr-2022)
- (7) Sugita, K., Gu, W., Ota, K., and Hasegawa, S., “L-VEIGe : Development of L2 Vocabulary Learning Support System with Suggestibility of Error by Image Generation’ IEICE Trans. Inf. & Syst (under review)
- (8) Hessel, J., Holtzman, A., Forbes, M. et al. “CLIPScore: A Reference-free Evaluation Metric for Image Captioning,” Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Online and Punta Cana, Dominican Republic, pp.7514–7528, (2021)
- (9) Radford, A., Kim, J. W., Hallacy, C. et al. “Learning Transferable Visual Models From Natural Language Supervision,” arXiv, (26-Feb-2021)
- (10) Xu, K., Ba, J.,Kiros, R., et al. “Show, Attend and Tell: Neural Image Caption Generation with Visual Attention,” arXiv, (19-Apr-2016)
- (11) Devlin, J., Chang, M.-W., Lee, K., et al, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” arXiv, (24-May-2019)
- (12) Reimers, N. and Gurevych, I., “Sentence-BERT: Sentence Embeddings using Siamese BERT Networks,” (arXiv, 27-Aug-2019)