Kashihara, A. et al. (Eds.) (2024). Proceedings of the 32nd International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

A Proposal for a Quantitative Evaluation Model for Error Image Generation in L2 Vocabulary Learning

Kazuki SUGITA[a*], Wen GU[b], Koichi OTA[b], Prarinya SIRITANAWAN[a] & Shinobu HASEGAWA[b^]
[a]*Graduate School of Advanced Science and Technology, Japan Advanced Institute of Science and Technology, Japan*
[b]*Center for Innovative Distance Education and Research, Japan Advanced Institute of Science and Technology, Japan*
*s2320028@jaist.ac.jp; ^hasegawa@jaist.ac.jp

Abstract: Vocabulary learning that incorporates visual information has become widely recognized as an alternative to context-based methods. However, few studies focus on learners' incorrect answers. On the other hand, fossilization caused by repeated errors has been a concern. Our proposed system, L-VEIGe, effectively prevents repeated errors by visualizing learners' incorrect answers through image generation, which encourages introspection. However, there exists a 'Feature Disappearance' problem, where the generated images for incorrect answers lack sufficient information for comprehension. This study proposes a method for quantitatively evaluating these error images from a cognitive perspective.

Keywords: Vocabulary Learning, Cognitive Fidelity, Quantitative Evaluation of Error Images, Learning from Error

1. Introduction

The importance of English in global communication is well-known. Vocabulary is essential for L2 learners and crucial in daily interactions, faces challenges in traditional methods relying on direct translations due to cultural differences in usage (Imai, 2016). While context-based vocabulary learning methods are recognized as effective, they can be challenging for beginners, leading to increased attention on visual information for supporting learning (Clark & Paivio, 1991). Additionally, there is the issue of fossilization, where repeated errors by learners become ingrained as part of their language use (Vigil & Oller, 2006). Despite this, few studies have focused on addressing errors in English vocabulary learning. The L-VEIGe, which applies error image generation to vocabulary learning support, has demonstrated effectiveness in promoting learners' reflection on their mistakes and preventing repeated errors (Sugita et al., 2023). However, L-VEIGe faces a limitation known as the Feature Disappearance problem, where insufficient information is provided for learners to fully understand their errors. This study aims to address the Feature Disappearance problem by developing a quantitative evaluation method for error-generated images.

2. Related Work

2.1 Vocabulary Learning Using Images

The Dual Coding Theory posits that memory retention and retrieval are more effective when both visual and verbal information are encoded together (Vigil & Oller, 2006). According to this theory, visual and verbal information create separate memory codes, which can trigger one another, leading to more effective memory retention and recall. However, Poor-quality images or images that are not directly related to the textual content can hinder learning by functioning

as visual noise (Li et al., 2022). This distracts learners and disrupts cognitive processing. The difficulty of ensuring high-quality images is considered a limitation, and there are few studies that have attempted to automatically and quantitatively evaluate these images.

2.2 Learning from Errors

Noticing facilitates the conversion of input into correct forms and promoting error awareness can lead to error correction (Schmidt, 1990). Error-Based Simulation (EBS) emphasizes the importance of cognitive conflict, which occurs when learners identify and attempt to correct their own errors (Hirashima, 2003). Kunichika et al. developed a system that uses animation to visualize grammatical errors in English (Kunichika et al., 2008). However, this approach is limited by the need for manual definitions, restricting the range of learning targets. Our proposed L-VEIGe system, which focuses on vocabulary learning, overcomes these limitations by automatically visualizing errors using image generation models. Figure 1 (left) shows L-VEIGe in action, where learners choose what they believe is the correct word based on an image (a) and a related question with a blank (b). If they choose incorrectly an image (c) and the sentence with the error are displayed. However, as shown in Figure 1 (right), a "Feature Disappearance" problem arises, where images for incorrect answers lack key information. This study aims to resolve this issue by automatically evaluating the generated images.

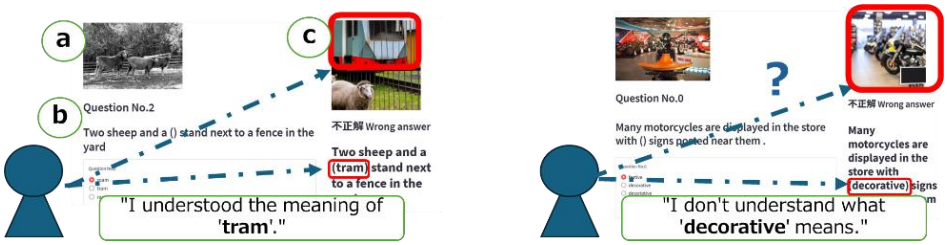

Figure.1 The overview of L-VEIGe and the Feature Disappearance problem

3. Methodology

3.1 Cognitive Fidelity for Error Images

In learning from errors, Cognitive Fidelity is crucial; it is important not only to ensure physical accuracy but also to present information in a way that is easily noticeable to learners (Hirashima, 2003). To effectively prompt learners to reflect on their errors without increasing cognitive load, we define evaluation criteria aligned with Cognitive Fidelity: impressionability and imaginability. As shown in Figure 2, imaginability is further divided into context similarity and error concept similarity. Impressionability is defined by the error concept surface ratio, which compares the error concept with other areas, and error concept color contrast.

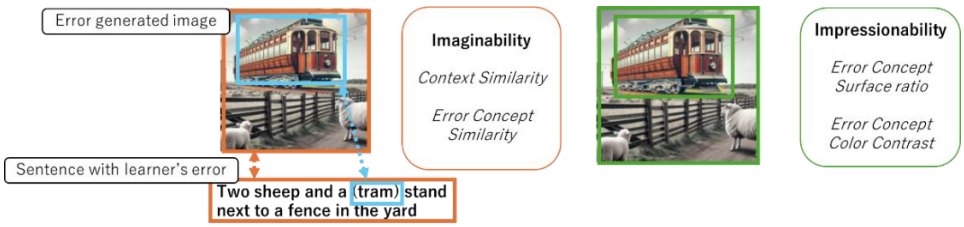

Figure.2 Correspondence between images and Imaginability and Impressionability

3.2 Evaluation Model

3.2.1 Imaginability Evaluation Model

The Imaginability Evaluation Model extends from (Sugita et al., 2023). This model evaluates two key metrics for understanding errors: Context Similarity and Error Concept Similarity. Study on evaluate the alignment between images and text, according to Sebastian's classification, falls into two categories: Embedding-based and Content-based methods (Hartwig et al., 2024). Embedding-based vectorizes both images and captions to evaluate their similarity. On the other hand, Content-based methods, such as those utilizing VQA (Visual Question Answering), focus on image evaluation. For text-focused evaluation, methods assess the similarity to reference captions.
In this study, we adapt evaluation methods for imaginability, addressing model biases. We developed an ensemble combining two models: (1) an Image Captioning Model from our previous study and (2) a VQA Model using GPT-4o (OpenAI, 2024). The resulting four evaluation scores train the Imaginability Evaluation Model, as shown in Figure 3.

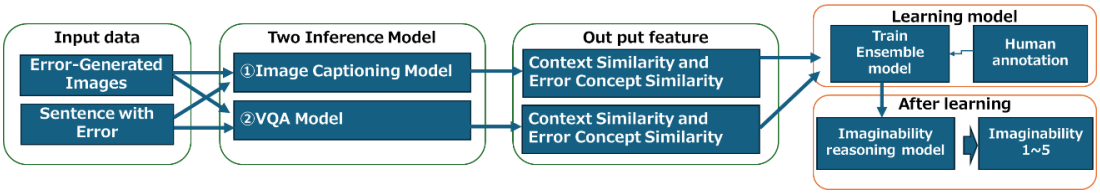

Figure.3 Imaginability Evaluation Model

3.2.2 Impressionability Evaluation Model

The Impressionability Inference Model evaluates two metrics: Error Concept Surface Ratio and Error Concept Color Contrast. To extract relevant areas, we use CLIPSeg (Lüddecke & Ecker, 2022) for object extraction based on both the image and prompt. CLIPSeg extracts error concepts, and the Surface Ratio is calculated by analyzing the extracted areas at the pixel level. For Color Contrast, the extracted regions are converted to HSV color space, and the contrast is measured by the average saturation difference between the error concept and surrounding areas. These features train the Impressionability Evaluation Model, as shown in Figure 4.

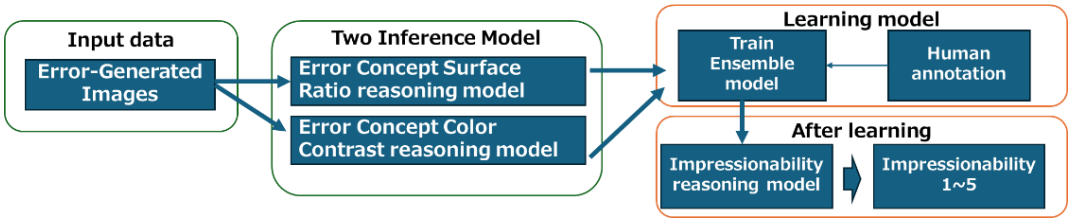

Figure.4 Impressionability Evaluation Model

4. Data Collection

The Imaginability and Impressionability Evaluation Models are trained using labels provided by human annotators. In line with Sheng et al., using multiple annotators improves labeling quality through majority voting, even when individual labels are noisy or from non-experts (Sheng et al., 2008). Thus, in this study, three annotators label the same dataset to ensure more accurate and reliable results. For Imaginability Evaluation, annotators provide a 5-point rating based on how well the generated image evokes the correct sentence. For Impressionability Evaluation, the rating reflects how effectively the image prompts learners to recognize the error. Figure 5 illustrates the annotation environment. In this process, learners first view the presented image (1), and then a fill-in-the-blank question containing the correct answer (2), (3), they label four generated images according to the prompts at the bottom of the interface (4), (5). The annotation environment is currently under construction.

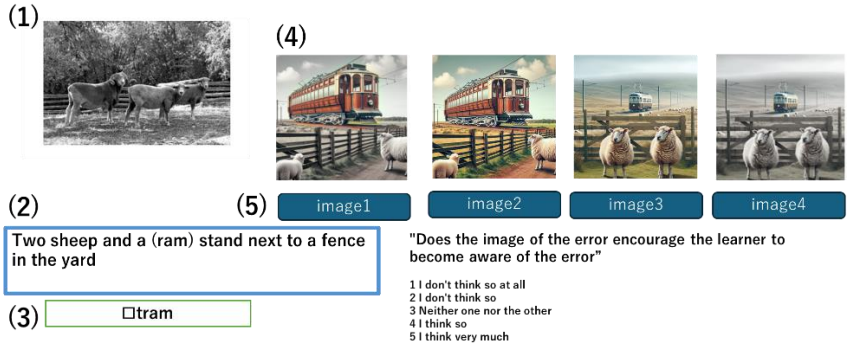

Figure.5 Annotation environment for impressionability

5. Conclusion

This study addressed the Feature Disappearance problem in the error image generation component of L-VEIGe, where cognitive fidelity is defined based on English vocabulary errors, and a quantitative evaluation method has been established. The main challenge now is designing prompts for a model that does not support long inputs, requiring us to select the key part of the sentence. In future work, we will finalize the Imaginability and Impressionability models through human annotation and evaluate how these models' outputs relate to vocabulary learning effectiveness. Our goal is to understand how L-VEIGe's cognitive fidelity influences learners' recognition of error images and their learning outcomes.

Acknowledgements

This work was supported by JST SPRING, Grant Number JPMJSP2102.

References

Clark, J., & Paivio, A. (1991). Dual Coding Theory and Education. Educational Psychology Review, 3(2), 149–210.

Hartwig, S., et al. (2024). A Survey on Quality Metrics for Text-to-Image Models. arXiv, arXiv:2403.11821. Retrieved from https://arxiv.org/abs/2403.11821

Hirashima, T. (2003). Learning Support System to Activate Metacognition. In Proceedings of the 17th Annual Conference of the Japanese Society for Artificial Intelligence (pp. 1-4).

Imai, T. (2016). The Effects of Explicit Instruction of 'Image English Grammar for Communication' on Tertiary English Classes. ARELE: Annual Review of English Language Education in Japan, 27, 137–152.

Kunichika, H., et al. (2008). English Composition Learning Support through Error Visualization. The IEICE Transactions on Information and Systems, 91(2), 210–219.

Li, W., et al. (2022). Dual Coding or Cognitive Load? Exploring the Effect of Multimodal Input on English as a Foreign Language Learners' Vocabulary Learning. Frontiers in Psychology, 13, Article 834706. https://doi.org/10.3389/fpsyg.2022.834706

Lüddecke, T., & Ecker, A. (2022). Image Segmentation Using Text and Image Prompts. arXiv preprint arXiv:2112.10003v2. Retrieved from https://arxiv.org/abs/2112.10003

OpenAI. (2024). GPT-4 Technical Report. arXiv preprint arXiv:2303.08774v6. Retrieved from https://arxiv.org/abs/2303.08774

Schmidt, R. (1990). The role of consciousness in second language learning. Applied Linguistics, 11(2), 129-158.

Sheng, V. S., et al. (2008). Get another label? Improving data quality and data mining using multiple, noisy labelers. In Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 614-622).

Sugita, K., et al. (2023). L-VEIGe: Vocabulary Learning Support System Using Error Image Generation - A Study of Criteria for Image Suggestibility. JSiSE Research Report, 38(2), 118-124.

Vigil, N., & Oller, J. (2006). Rule Fossilization: A Tentative Model. Language Learning, 26, 281–295. https://doi.org/10.1111/j.1467-1770.1976.tb00258.x