Language	Head	Relation	Tail	Overall
Bilingual	63.82	82.76	70.73	47.17
English	66.24	81.93	68.51	47.59
Korean	65.37	85.39	75.08	50.43

Table 3: Triple prediction performance of the GEL-VQA model

(Language-independent KG Utilization) Table 3 summarizes the performance of the proposed triple prediction method. Despite the use of English KGs, the performance of predicting triples based on images and questions in Korean was similar to or slightly better than that in English. These results imply that KGs are independent of language. However, it should be noted that the English question data were constructed based on Korean questions, which were constructed first; thus, Korean sentences may have been more natural in terms of quality and structure.

Analysis

In this section, we analyze the quality of the proposed dataset and methods of knowledge utilization used in the model.

Robustness Test of Proposed Dataset

Eval Method	BASELINE	GEL-VQA
Raw Question	21.51±1.81	45.08±0.94
Punctuation	21.09±0.81	44.02±0.93
Antonyms	20.94±0.50	37.23±1.62
Synonym(Verb)	21.07±0.77	43.36±0.90
Synonym(Noun)	20.97±0.59	42.79±0.91
Hypernym(Noun)	20.76±0.23	36.24±0.76
Hyponym(Noun)	20.83±0.33	38.23±1.01

Table 4: Model robustness test results through question transformation.

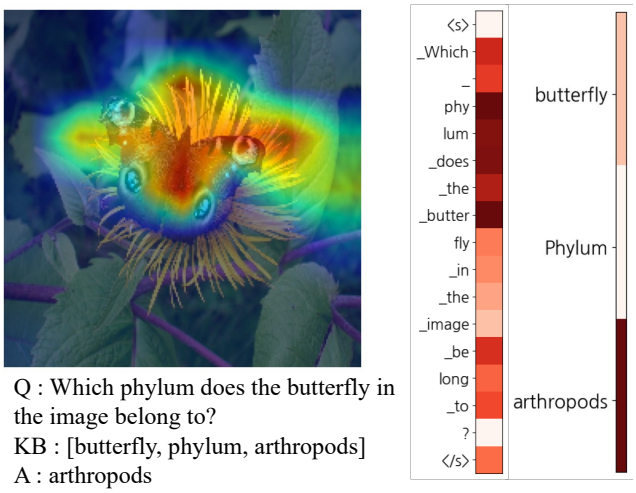
High-quality data should be unbiased and robust. How, then, can we analyze the quality of the BOK-VQA dataset proposed in this study? CARETS (**?**) and VALSE (**?**) are methods used to determine the consistency and robustness of a model by transforming questions into VQA. For example, a model using consistent data should consistently answer the questions Who is Barack_Obamas **wife**? and Who is Barack_Obamas **spouse**?, while data with errors or conflicts would not. To verify the consistency and robustness of our model, we conducted robustness experiments through six types of question modifications based on WordNet (**??**), as shown in column Eval Method in Table 4.

Question transformation involves changing a word in a question using the aforementioned six methods. For example, in the case of noun synonyms, the system sequentially searches for nouns in a question and replaces them with synonyms from WordNet, if available. Table 4 lists the results of experiments on data robustness. When nouns are replaced with hyponyms or hypernyms, the performance significantly drops in comparison to that of the original question, which is used as the baseline. Specifically, in the case of hypernyms, replacing wife with person increases the ambiguity when searching for triples in the transformed sentence, resulting in performance deterioration.

In contrast, when nouns are replaced with synonyms, such as changing wife to spouse, the performance remains relatively similar to that of the original question. This indicates that preserving meaning through synonym replacement does not adversely affect the systems ability to retrieve appropriate information from KGs. Transformations, including punctuation and verb changes, demonstrated similar performance as the original question test.

In conclusion, performance deterioration due to noun replacement showed that nouns in questions have a strong correlation with objects in images or heads in KGs. When the meaning of a noun changes significantly, the KG system may struggle to find the desired information in the KG owing to changes in a semantic association. Finally, when comparing the quality with the VQA v2.0 dataset (**?**), we observed a performance drop of 12.03% in the ontological transformation data proposed by CARETS, whereas our dataset showed a 7.85% decrease; this could indicate that our dataset is more robust.

Triple Attention Score Statistics



Q : Which phylum does the butterfly in the image belong to?
KB : [butterfly, phylum, arthropods]
A : arthropods

Figure 5: Visualizing attention score of H-Given case.

Case	H attention	R attention	T attention
H-Given	0.3239	0.2587	0.4174
T-Given	0.3280	0.2790	0.3930
HT-Given	0.3329	0.2819	0.3852

Table 5: Average attention scores based on the triple component.

In the proposed GEL-VQA+ATTN model, self-attention (**?**) was applied to the head, relation and tail as mentioned in the section of Experiment Settings. Consequently, when solving

the VQA problem, we could examine the information (between $e_{h_i}, e_{r_i}, e_{t_i}$) that the model focused on more. Figure 5 shows which information between <head, relation, tail> the model focused on for a single sample. Including this example, in most cases, the attention scores were the highest for the tail. This is because the 'tail' often represents the objective or outcome of the "relation"; hence, the model primarily sought outcome information through external knowledge. Interestingly, when the question included information from <h, r, t>, the model tended to focus slightly more on triple components not included in the question. For instance, Figure 5 shows a question that includes head information; in this case, the model tended to focus relatively more on information other than the head. We analyzed the results presented in Table 5 based on three criteria (H-, T-, and HT-Given) to verify the extent of knowledge utilization in the model. Table 5 lists the average attention scores for the entire dataset under three cases.

- **H-Given**: This refers to cases in which the head information was included in a question. The attention score was the highest for the tail, which suggests that because the model had already obtained information about the head from the question, it focused more on obtaining tail information from external knowledge.
- **T-Given**: This refers to cases in which the tail information was included in the question. The attention scores for both head and tail were relatively evenly distributed. This suggests that because the model had already obtained information about the tail from the question, it tended to focus on the head and relation in the external knowledge.
- **HT-Given**: This refers to cases in which both head and tail information were included in the question. In this case, the model tended to focus more on the part corresponding to relation in the embeddings given by external knowledge. Therefore, compared to H-Given and T-Given, the attention score for relation appeared to be higher.

In conclusion, although the differences in attention scores were marginal, the triple information on which the model focuses varied depending on the type of question.

Analysis of KGE Model Impact on GEL-VQA Performance

The foremost contribution of this study is the construction of a language-independent KB-VQA using KGEs. Various KGE methods have been proposed for this purpose, which raises the question: what impact does the choice of different KGE training methods have on VQA, and which KGE is the most efficient? Previous studies utilized FB15K (**?**), a large relational graph dataset, to evaluate the performance of different KGE training methods. FB15K employs link and entity prediction methods for the intrinsic evaluation of KGE models using Hit@10 as an evaluation metric that determines whether the correct answer is included in the top 10 predicted results. The FB15K column in Table 6 presents the performance evaluation of various KGE training methods using FB15K, specifically the Hit@10 score. In this study, the performance of ConvKB, used as a baseline KB, was significantly different (more than two-fold) in comparison to that of recently proposed algorithms under the FB15K dataset. Conversely, GEL-VQA (Table 6), which represents the intrinsic evaluation of corresponding KGEs in VQA, exhibited a relatively lower difference in performance between ConvKB and other models. The Pearson correlation coefficient between the scores of FB15K(H@10) and proposed GEL-VQA-IDEAL model had a high value of 0.85. This implies a correlation between the performance of the two benchmarks, signifying that the choice of the KGE model significantly affects the performance of both GEL-VQA and GEL-VQA-IDEAL. These results reveal the potential applicability of KGs and indicate the need for further research to effectively harness KGEs.

KGE	GEL-VQA	GEL-VQA(IDEAL)	FB15K
TransE	51.22±1.02	79.56±0.77	0.847
TorusE	50.51±1.32	79.15±0.28	0.839
HolE	48.64±0.75	78.91±1.08	0.867
DistMul	47.28±0.80	72.40±1.60	0.863
ConvKB	45.08±0.94	66.01±1.83	0.408

Table 6: Impact of KGE methods on GEL-VQA.

Conclusion

In this study, we proposed a method to effectively construct external knowledge VQA training data for less-resourced languages using the abundant external knowledge of high-resource languages. We constructed large-scale training data using 17K Korean-English question-answer pairs and 280K instances of information. Furthermore, we demonstrated a language-independent approach to KG utilization using a GEL-VQA model that performed VQA and KGE training in a multitasking manner. Our experiments demonstrated the acceptable performance of a bilingual VQA model using the proposed BOK-VQA and GEL-VQA. Through the results of this study, we anticipate that the benefits of multilingual training and our evaluation approach can serve as performance metrics for future multilingual VQA research. Nonetheless, there are several unresolved limitations in this study. While the knowledge utilized to construct a question involves K $\geq$ 1 pieces of information, our study predicts and utilizes only one piece of knowledge. Thus, it remains inadequate for solving problems that require the complex utilization of multiple pieces of knowledge.

Acknowledgements

This research was supported by the National Research Foundation of Korea (2021R1F1A1063474) for KyungTae Lim. This research used datasets from The Open AI Dataset Project (AI-Hub) (No. 2022–41, 2023–93). Sint aut temporibus pariatur quisquam, recusandae accusamus ducimus nisi totam eaque dolores, modi quasi error officia cumque perspiciatis, quidem aspernatur aliquam officiis, mollitia beatae dignissimos tempora id repellendus nam omnis vel libero vitae esse?Quis ad maiores fugiat consequuntur voluptatum dolor eum reprehenderit, illo quod culpa labore, aliquam placeat asperiores aliquid enim ad qui alias quam dicta nam, doloribus deleniti dolorum assumenda incidunt asperiores voluptas delectus, illum inventore autem porro officia.