BOK-VQA: Bilingual Outside Knowledge-based Visual Question Answering via Graph Representation Pretraining

MinJun Kim[1*]**, SeungWoo Song**[1*]**, YouHan Lee**[2]**, Haneol Jang**[1]**, KyungTae Lim**[3†]

[1] Hanbat National University
[2] Kakao Brain
[3] Seoul National University of Science and Technology
{20181796, 20181620}@edu.hanbat.ac.kr, youhan.lee@kakaobrain.com, hejang@hanbat.ac.kr, ktlim@seoultech.ac.kr

Abstract

The current research direction in generative models, such as the recently developed GPT4, aims to find relevant knowledge information for multimodal and multilingual inputs to provide answers. Under these research circumstances, the demand for multilingual evaluation of visual question answering (VQA) tasks, a representative task of multimodal systems, has increased. Accordingly, we propose a bilingual outside-knowledge VQA (BOK-VQA) dataset in this study that can be extended to multilingualism. The proposed data include 17K images, 17K question-answer pairs for both Korean and English and 280K instances of knowledge information related to question-answer content. We also present a framework that can effectively inject knowledge information into a VQA system by pretraining the knowledge information of BOK-VQA data in the form of graph embeddings. Finally, through in-depth analysis, we demonstrated the actual effect of the knowledge information contained in the constructed training data on VQA.

Introduction

Visual question answering (VQA) (**?**) refers to a system that provides answers to natural language questions related to input images. In recent years, an extended form of VQA, called knowledge based VQA (KB-VQA), has developed that utilizes external knowledge sources, such as Wikipedia or knowledge graphs (KGs), to enable higher-level reasoning to answer complex questions. This knowledge-based approach in VQA leverages various types of information, including biographical (K-VQA) (**?**), common sense (FVQA) (**?**), (KB-VQA) (**?**), Wikipedia abstracts (OK-VQA) (**?**), and scene graphs (GQA) (**?**).

However, despite these advancements, the existing versions of KB-VQA encounter various challenges. First, owing to the cost of construction, the size of training data is either relatively small (FVQA and KB-VQA) or generated automatically based on metadata (GQA). Second, most datasets are constructed predominantly in English. Given these language biases, a recent approach for addressing this issue introduced

---

xGQA (**?**), a multilingual VQA evaluation dataset for a few-shot evaluation, however, xGQA is limited by the lack of sufficient training data.

Further, KB-VQA encounters challenges in the construction of training data for less-resourced languages owing to the need for accompanying information (e.g., Wikipedia). The consideration of this issue gives rise to the major research question of this study, i.e., can we leverage the rich external knowledge of highly-resourced languages to construct KB-VQA training data for less-resourced languages? We attempted to address this question in this study. Considering that it is not necessary to rely on a specific language if utilizing the knowledge representation of highly-resourced languages in the form of entity embeddings, we proposed a bilingual dataset and method for representing knowledge through KG Embeddings (KGEs).

The proposed dataset possessed the following characteristics: (1) We constructed a relatively large-scale training dataset using 17K Korean-English question-answer pairs and 280K instances of information. (2) It is compiled with the structure of bilingual outside-knowledge VQA (BOK-VQA), which utilizes knowledge base (KB) information from English, a representative high-resource language, in conjunction with other languages.

KGs, which typically consist of object and relation information, can be used to represent complex relationships. Thus, KGE methods (**??????**) have been proposed as techniques for compressing and representing data while preserving the structure and information of KGs. This leads us to the final question of this study, i.e., how can knowledge embeddings be efficiently fused into VQA? For that, we propose a graph-embedded learning-based VQA (GEL-VQA) model. This model acquires knowledge information from KGs through triple prediction and subsequently leverages this information to address VQA problems. The proposed multitasking model is designed to minimize the loss of information across different tasks by predicting the information required to solve VQA questions through triple prediction and concurrently conducting VQA training.

Through in-depth experiments, we revealed that the performance of VQA varied widely from 21.16 to 55.48 depending on the application of KGE and triple prediction. This indicates that the proposed dataset represents a problem that can be solved by utilizing KB information, and KG-based

methods have a substantial impact. In addition, to analyze the impact of KGE methods on the GEL-VQA model, we evaluate the performance of the proposed model based on different KGE methods. The contributions of this study are given below:

- We propose a multilingual extensible KB-VQA training dataset named BOK-VQA.
- We propose a language-independent GEL-VQA training framework for KG information.
- Through in-depth analysis, we prove the actual effect of language-independent KGs on BOK-VQA.

Related Works

VQA was first introduced in 2015 by **?**, and the dataset comprised questions that could be answered with only a shallow understanding of images.

Moving beyond simple questions, VQA that used more complex external knowledge was scarcely explored until **?** proposed the FVQA dataset. FVQA employed DBpedia, ConceptNet, and Webchild to incorporate common sense and object-related knowledge into question-answering. However, despite the use of KGs, the questions mainly focused on object recognition in images, and object-related question-answering was similar to traditional knowledge-based question-answering. To address these issues, outside knowledge base VQA (OK-VQA) (**?**), which is less reliant on KBs, was proposed. OK-VQA includes information based on web texts and other sources, offering greater flexibility, but at the cost of lower knowledge utilization. Subsequently, scene graph-based VQA (GQA) (**?**) was introduced in 2019, which utilized scene graph information from the Visual Genome dataset (**?**) to automatically generate a large volume of VQA datasets. Efforts to expand VQA into multilingual settings began in 2021. Because most VQA-related datasets are based on monolingual (English) sources, there is a lack of research on VQA for low-resource languages. Therefore, xGQA was proposed to evaluate multilingual VQA. However, since it was mainly constructed for evaluation in various languages through few-shot learning rather than as training data, using it for training purposes is challenging.

Additionally, VQA datasets have been proposed for domain-specific purposes, such as MovieQA (**?**), TVQA (**?**), DramaQA (**?**), and ActivityNet-QA (**?**). MovieQA is a video-based VQA dataset comprising 15,000 question-answer pairs obtained from captures, plots, lyrics, and scripts from 400 movies. TVQA specializes in six TV show domains and contains 152,000 question-answer pairs from 925 episodes. DramaQA is a question-answering dataset for the Korean drama "Another Miss Oh" characterized by its hierarchical structure based on question difficulty. Finally, ActivityNet-QA is an action-oriented VQA dataset with 58,000 question-answer pairs extracted from 5,800 videos.

Proposed BOK-VQA Dataset

In this study, we propose a BOK-VQA dataset comprising 17,836 samples and 282,533 knowledge triples. Each sample contained of an image, question, answer, and k external

Q(en) : Which phylum does an animal belonging to Scyphozoa in the image belong to?

Q(ko) : 이미지에 나오는 해파리강에 속하는 동물은 어느 문에 속하나요?

KB : [Jellyfish, phylum, Cnidaria]

A(en) : Cnidaria

A(ko) : 자포동물

Figure 1: Example of BOK-VQA sample

knowledge IDs that are necessary to solve a question. Figure 1 presents an example of the BOK-VQA dataset. It is challenging to answer the question, "Which phylum does an animal belonging to Scyphozoa in the image belong to?" without using external knowledge. The answer accuracy can be enhanced using external knowledge given in the form of a triple "['Jellyfish', 'phylum', 'Cnidaria']" with the image-question.

Considering multilingual extensibility, how should knowledge information in KB-VQA be efficiently structured? Data with a unified ontological system for objects expressed in different languages can be advantageous in this case. External knowledge that meets these criteria can be linked to data wherein concepts in different languages are connected through URIs. Accordingly, repositories such as ConceptNet (**?**) and DBpedia (**??**), which were proposed as common sense KBs in FVQA, can be potentially used.

However, because gathering sufficient KG data is not feasible for low-resource languages, we propose a novel method to use only English KGs. It is noteworthy that because entity and relation information are independent of language, well-trained KGEs, which contain knowledge, are language-agnostic. For example, the surface form of English triple <Barack_Obama, Spouse, Michelle_Obama> and Korean triple <_, Spouse, _> is presented differently in two distinct languages, but the meaning of entity and relation information <head, relation, tail> is entirely identical. Therefore, by using English KGs as embeddings for Korean KB-VQA, the need for Korean KGs can be eliminated. Specifically, in this study, we pretrained the English KGs as embeddings using ConvKB (**?**) and used it for multiple languages. Then the only remaining question is how can a Korean VQA question be linked to a related English triple? We address this problem by utilizing a module that predicts related English triple based on an image and a Korean question. For instance, to answer the Korean question " ? (Who is Barack Obama's wife?)", we predicted the English triple <Barack_Obama, Spouse, Michelle_Obama> using the given image and Korean question. We then extracted its embedding, which was pretrained using English KB data from the predicted triple.

Considering the aforementioned aspects, we assembled 282,533 triple knowledge entries comprising 1,579 objects and 42 relations from English ConceptNet and DBpedia. The selection criteria for the objects and relations were principally based on the 500 objects and 10 relations used in the FVQA dataset. In addition, considering the usage frequency, we incorporated 1,079 objects derived from ImageNet (**?**) and

Criterion	FVQA	OK-VQA	KB-VQA	KVQA	BOK-VQA(Ours)
Number of images	2,190	14,031	700	24,602	17,836
Number of question-answer pairs	5,826	14,055	2,402	183,007	17,836
Average question length(words)	9.5	8.1	6.8	10.1	10.8
Number of questions per image	2.66	1.00	3.43	7.44	1.00
Number of knowledge sources	3	1	1	1	2
Knowledge source	DBpedia +ConceptNet +Webchild	Wikipedia	DBpedia	Wikidata	DBpedia +ConceptNet
Multilinguality	✗	✗	✗	✗	✓

Table 1: Key statistics for KB-VQA dataset.

supplemented 32 additional relations. Table 1 presents a basic statistical comparison between the proposed BOK-VQA and other KB-VQA datasets. BOK-VQA utilizes a relatively large number of images and composes longer questions with a bilingual context. The object list and relation statistical information utilized in BOK-VQA, along with detailed data samples, can be found in Appendix A.

A total of 87 skilled annotators created the question-answer data for BOK-VQA. After the creation of the Korean question data, English questions and answers were generated through neural translation. To ensure the quality of the bilingual QA, two professional translators deeply reviewed the translated English data. Based on this review process, high-quality bilingual data were created. However, there may still be a limitation in the standardization of sentence structures because we revised the translated sentences despite the review process.

In summary, BOK-VQA inherited the data schema of FVQA, with significantly enhanced reliability and scalability. First, in terms of data reliability, the BOK-VQA dataset was constructed by imposing several constraints during the question generation process, ensuring that questions required knowledge information for solving. For instance, questions like 'Who invented the instrument in the image?' were selected only if three different reviewers deemed appropriate knowledge information necessary to address the question. Second, in terms of scalability, the question-answer data were constructed as bilingual training data; moreover, because only English is required for KGs, they can be extended to multiple languages. Further, with the availability of object lists, knowledge information related to objects can be easily accessed in every language.

Model Description

In this section, we introduce the proposed GEL-VQA model that combines VQA and KGs by converting them into a representational layer. This model focuses on answering visual questions and acquires external information by predicting KG triples. The proposed model broadly contains three parts. First, the **VQA module** is designed to train on VQA tasks using image and question features. This module serves as the BASELINE model. Second, the **triple prediction module**, a component intended for knowledge acquisition from the KB, is given a pair of images and questions as inputs. This module acquires the information associated with an image by predicting external knowledge contained in triples. Third, the **KGE module** that offers a method for incorporating KGs into the

VQA module; it uses KGE techniques to generate embedding containing relation information of KGs. The embedding values obtained through this module were fed into the proposed model, which enhanced the model's performance.
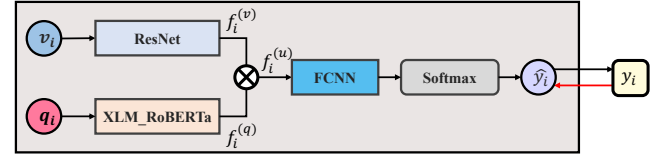
VQA Module

Figure 2: Architecture of VQA module.

As presented in Figure 2, the VQA module serves as the foundational model and uses images and questions as input data. Given a dataset $D = \{v_i, q_i, y_i\}_{i=1}^{N}$, where v_i represents the i-th image, q_i is the i-th question, and y_i is the corresponding answer to v_i and q_i, if we denote the VQA module as P_{vqa}, our problem can be defined as the following probabilistic model utilizing the method of maximum likelihood:

$$\theta^* = \max_\theta \prod_{i=1}^{N} P_{vqa}(y_i|v_i, q_i; \theta) \qquad (1)$$

Here, θ represents the set of model parameters, and θ^* denotes the optimal parameters of the probabilistic model. $P_{vqa}(y_i|v_i, q_i; \theta)$ represents the probability of accurately predicting the target value (y_i), given the input data (v_i, q_i) and the model parameters (θ). Accordingly, the cross-entropy loss function can be defined as:

$$\mathcal{L}(\theta) = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{C} y_{i,j} \log(P_{vqa}(j|v_i, q_i; \theta)) \qquad (2)$$

Where $y_{i,j}$ is the one-hot encoded representation of the actual class of the i-th data sample and C is the total number of classes. In this study, to implement the probability model (P_{vqa}) as a multilingual model, we generate a d-dimensional feature vector $f_i^{(q)} \in \mathbb{R}^d$, for the input question q_i through the multilingual-based XLM-RoBERTa (**?**). For input image v_i, we generate a d-dimensional feature vector, $f_i^{(v)} \in \mathbb{R}^d$, via ResNet50 (**?**). Finally, to combine these two modality features, we create an unified d-dimensional feature vector

$f_i^{(u)} \in \mathbb{R}^d$ through element-wise multiplication (denoted as $\otimes$) as follows:

$$f_i^{(q)} = \text{XLM-RoBERTa}(q_i) \tag{3}$$

$$f_i^{(v)} = \text{ResNet}(v_i) \tag{4}$$

$$f_i^{(u)} = f_i^{(v)} \otimes f_i^{(q)} \tag{5}$$

In addition, we employ a multi-layer perceptron (MLP) to transform the unified feature vector $f_i^{(u)}$ into a logit value $f_i^{(u')}$. To introduce non-linearity into the process, the ReLU function is utilized.

$$f_i^{(u')} = W^{(\alpha)}\text{ReLU}(W^{(\beta)}f_i^{(u)} + b^{(\beta)}) + b^{(\alpha)} \tag{6}$$

If we denote c as a specific class, where $c \in \{0, 1, 2, ..., C\}$, the final predicted value $\hat{y}_i$ can be obtained by selecting the class with the highest probability after applying the Softmax transformation. This can be expressed as follows:

$$\hat{y}_i = \arg\max_c (\text{Softmax}(f_{i,c}^{(u')})) \tag{7}$$

Triple Prediction Module

Figure 3: Architecture of Triple Prediction Module.

Figure 3 demonstrates the triple prediction module. It predicts the head, relation, and tail from the given image and question. Considering the features of each triple, the proposed model implements shared parameters for the image and question feature extractors. Concurrently, it maintains distinct parameters for each classifier responsible for predicting the head, relation, and tail. In other words, the classifier processes each head, relation, and tail prediction using a different linear layer. As mentioned in the section of Proposed BOK-VQA Dataset, the triple prediction module demonstrates language-independent characteristics. This is because the proposed method yields one-hot encoded target values regardless of their source language.

KGE Module

KGE is an approach for transforming KGs from the triple to embedding form. Hence, similar to word embeddings, KGE must be pretrained before training. The KGE pretraining process can be divided into two steps, i.e., sampling and training. First, in the sampling step, assuming that the triple information included in the training data is a valid triple set $\mathcal{K}$, an invalid triple set $\mathcal{K}'$ is generated by randomly replacing the head or tail. The final valid/invalid triple set can be defined as $\mathbb{K} = \{\mathcal{K} \cup \mathcal{K}'\}$. Second, in the training step, the created invalid triples are trained using a loss function ($\mathcal{L}_c$) that

effectively distinguishes between valid and invalid triples as follows :

$$\mathcal{L}_c = \sum_{(h,r,t) \in \mathbb{K}} \log(1 + e^{g(h,r,t)}) + \frac{\lambda}{2}||w||_2^2 \tag{8}$$

$$g(h,r,t) = \begin{cases} f(h,r,t), & \text{for } (h,r,t) \in \mathcal{K} \\ -f(h,r,t), & \text{for } (h,r,t) \in \mathcal{K}' \end{cases} \tag{9}$$

$$f(h,r,t) = W \cdot (\text{ReLU}([e_h; e_r; e_t] * \Omega)) + b \tag{10}$$

In this study, we used the ConvKB approach, which captures the patterns of individual entities and relations present in a KG via convolution operations. The aforementioned function f signifies the scoring function employed by ConvKB. Within this context, $*$ represents the convolution operation, while e_h, e_r, e_t denote head, relation, and tail embeddings, respectively. They are randomly initialized in the beginning, similar to word embeddings. Furthermore, Ω represents the set of weights used in the convolutional filter. We trained ConvKB over our entire KG (280k) for 50,000 iterations.

Aligning KG for VQA

When KGEs are successfully trained using the KGs over our dataset, pretrained knowledge embeddings corresponding to triples h_i, r_i, and t_i which are necessary for the answer from the image v_i and question q_i can be represented as e_{h_i}, e_{r_i}, and e_{t_i}, respectively, as follows:

$$emb_{kb_i} = \{e_{h_i}, e_{r_i}, e_{t_i}\} = ConvKB(h_i, r_i, t_i) \tag{11}$$

$$f_i^{(z)} = f_i^{(u)} \otimes emb_{kb_i} \tag{12}$$

Where $f_i^{(u)}$ denotes the fusion feature vector of modalities, which is an element-wise product of image feature vector $f_i^{(v)}$ and question feature vector $f_i^{(q)}$.

GEL-VQA

Figure 4: GEL-VQA architecture: (a) VQA module, (b) triple prediction module, and (c) pretrained KGE.

In the context of VQA that uses external knowledge, it is unrealistic to assume that one possesses external knowledge pertaining to the given images and questions. Consequently, in this study, we proposed the GEL-VQA model that employs a multitask learning approach to perform triple prediction

Language	BASELINE	GEL-VQA(IDEAL)	GEL-VQA	GEL-VQA + TF	GEL-VQA + TF + ATTN
Bilingual	21.51±1.81	66.01±1.83	45.08±0.94	48.07±1.33	48.11±1.50
English	21.90±2.34	66.68±1.15	47.83±0.56	51.64±0.88	50.74±0.90
Korean	21.16±1.50	72.25±1.29	50.30±2.24	53.40±2.73	55.48±1.89

Table 2: Performance of five models across three languages: Bilingual, English, and Korean.

and uses the predicted triples as external knowledge. To predict and use triples in real time, the model utilizes the predicted triples $\{\hat{h}_i^{(\tau-1)}, \hat{r}_i^{(\tau-1)}, \hat{t}_i^{(\tau-1)}\}$, computed using updated weights from the previous time step $\tau - 1$. Figure 4 illustrates the structure of GEL-VQA. At time step $\tau \geq 1$, the given data v_i, q_i and KGE $\mathcal{E}_i^{(\tau-1)} = [e_{\hat{h}_i^{(\tau-1)}}; e_{\hat{r}_i^{(\tau-1)}}; e_{\hat{t}_i^{(\tau-1)}}]$ predicted through triple prediction at time-step $\tau - 1$ can be used as external knowledge inputs. Where $\mathcal{E}_i^{(0)}$ contains random embeddings of head, relation, and tail, and ';' denotes the concatenation of its embeddings.

$$\mathcal{L}_{\text{VQA}}^{(\tau)}(\theta) = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{C}$$
$$y_{i,j} \log(P_{vqa}(j|v_i, q_i, \mathcal{E}_i^{(\tau-1)}; \theta)) \quad (13)$$

$\mathcal{L}_{\text{VQA}}^{(\tau)}$ in equation (13) represents the loss function of the VQA module at time-step τ, where image v_i, question q_i, and embedding $\mathcal{E}_i^{(\tau-1)}$ of the triple predicted in the previous time-step $\tau - 1$ are used as inputs. Similarly, $\mathcal{L}_{\text{T}}(\theta)$ in equation (14) denotes the loss function of the triple prediction module. Finally, the total loss function of our GEL-VQA, $\mathcal{L}^{(\tau)}(\theta)$, can be represented as:

$$\mathcal{L}^{(\tau)}(\theta) = \mathcal{L}_{\text{VQA}}^{(\tau)}(\theta) + \mathcal{L}_{\text{T}}(\theta) \quad (14)$$

Experiments

Our research objective involves exploring the feasibility of flexible multilingual VQA using KGs of high-resource languages. This can be accomplished through experiments using the proposed BOK-VQA dataset and GEL-VQA model to address the following questions: (1) Does the proposed BOK-VQA dataset require real-world external knowledge? (2) Does the proposed GEL-VQA model utilize KGs independent of language? (3) Is there a performance degradation in relatively less-resourced languages based on the proposed GEL-VQA?

Experimental Settings

We performed experiments using five models for three different language combinations. The dataset was split into 60% for training and 20% each for validation and testing, with five-fold validation. The proposed dataset contains one question-answer pair per sample; therefore, all evaluation metrics used in the experiment utilized accuracy. First, for the Korean evaluation, we used only 17K Korean questions and an English KB for training. Second, for English, we used the English KB and 17K English questions. Conversely, in the bilingual combination, both 17K Korean and 17K English

question sets were used in conjunction with the English KB for training. As introduced in KGE module, to utilize triple embedding in VQA experiments, all experiments were pre-trained on English KB data using the ConvKB algorithm for 50,000 iterations. Detailed settings of the hyperparameters applied in the experiments are presented in Appendix C. The proposed VQA models are described below.

- BASELINE: A basic VQA model that utilizes image-question pairs without employing external knowledge.
- GEL-VQA(IDEAL): Given an image-question corresponding to a triple, this model uses ConvKB to embed triples. The model operates under the assumption that it is aware of all triples in the training/testing data and accordingly infers VQA answers. Owing to this assumption, it may not be suitable for real-world applications. However, by examining the difference between GEL-VQA(IDEAL) and BASELINE this model demonstrated that the BOK-VQA dataset used in this study requires external knowledge.
- GEL-VQA: This model is an evolved version of GEL-VQA(IDEAL) through triple prediction and utilization of predicted triples.
- GEL-VQA + TF: This model applies the concept of teacher forcing to the GEL-VQA model using ground-truth triples for training data during VQA model training. Note that the test phase used predicted triples.
- GEL-VQA + TF + ATTN: This model strengthens selective knowledge acquisition by applying multi-head attention to the <head, relation, tail> of KGE (see Appendix D).

Experiment Results

(Overall) Table 2 summarizes the experimental results of each model on the datasets used for training in terms of language. Overall, BASELINE, which does not utilize external knowledge, exhibits the lowest performance, whereas GEL-VQA(IDEAL), the ideal case, exhibits the highest performance. The performance gap between the BASELINE and GEL-VQA(IDEAL) models suggests that the BOK-VQA dataset is more effective when external knowledge is used. Additionally, the performance of the proposed model, GEL-VQA, is 25.97 points higher than that of BASELINE, indicating that it could effectively utilize the external knowledge provided by KGs.

(Performance in Less-resourced Languages) Table 2 indicates that despite the use of English KB, the performance in Korean, a relatively less-resourced language was on average 3.7% higher than that in English. This indicates that the training performance of less-resourced languages does not deteriorate, and the quality of question construction is more important.

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

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.

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.

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 ≥ 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.