

Article

Using Multiple Monolingual Models for Efficiently Embedding Korean and English Conversational Sentences

Youngki Park ¹ and Youhyun Shin ^{2,*}¹ Department of Computer Education, Chuncheon National University of Education, Chuncheon 24328, Republic of Korea; ypark@cnue.ac.kr² Department of Computer Science and Engineering, Incheon National University, Incheon 22012, Republic of Korea

* Correspondence: yhshin@inu.ac.kr

Abstract: This paper presents a novel approach for finding the most semantically similar conversational sentences in Korean and English. Our method involves training separate embedding models for each language and using a hybrid algorithm that selects the appropriate model based on the language of the query. For the Korean model, we fine-tuned the KLUE-RoBERTa-small model using publicly available semantic textual similarity datasets and used Principal Component Analysis (PCA) to reduce the resulting embedding vectors. We also selected a highly-performing English embedding model from available SBERT models. We compared our approach to existing multilingual models using both human-generated and large language model-generated conversational datasets. Our experimental results demonstrate that our hybrid approach outperforms state-of-the-art multilingual models in terms of accuracy, elapsed time for sentence embedding, and elapsed time for finding the nearest neighbor, regardless of whether a GPU is used. These findings highlight the potential benefits of training separate embedding models for different languages, particularly for tasks involving finding the most semantically similar conversational sentences. We expect that our approach will be used for diverse natural language processing-related fields, including machine learning education.

Keywords: natural language processing; multiple monolingual models; sentence transformers



Citation: Park, Y.; Shin, Y. Using Multiple Monolingual Models for Efficiently Embedding Korean and English Conversational Sentences. *Appl. Sci.* **2023**, *13*, 5771. <https://doi.org/10.3390/app13095771>

Academic Editor: Francisco García-Sánchez

Received: 15 April 2023

Revised: 29 April 2023

Accepted: 3 May 2023

Published: 7 May 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Sentence-BERT (SBERT) [1] has emerged as one of the most efficient BERT-based [2] sentence embedding approaches. SBERT not only outperforms other approaches such as GloVe [3], InferSent [4], and Universal Sentence Encoder [5] in semantic textual similarity tasks but also solves the problem of slow inference time of cross-encoder (BERT)-based sentence embedding approaches. SBERT typically involves using pre-trained models such as BERT-base, BERT-large, RoBERTa-base, or RoBERTa-large [6], which are fine-tuned on datasets specific to downstream tasks such as semantic textual similarity.

While most SBERT models focus on the English language, there is a need to embed both English and Korean sentences. To the best of our knowledge, one of the best approaches for embedding both Korean and English sentences using SBERT is through the use of the “Knowledge Distillation” [7] method. This method involves a teacher model and a student model, where the student model is trained to imitate the teacher model using translated sentences. Surprisingly, the authors of this approach stated that it outperforms even state-of-the-art Korean monolingual models for Korean semantic textual similarity tasks.

In this paper, we present a novel and efficient approach for embedding both Korean and English. Instead of using a single multilingual model that maps both language sentences into the same embedding space, our main idea is to prepare each small and efficient monolingual model (Korean and English) by selecting appropriate pre-trained models and training datasets. During inference, we select the appropriate monolingual model based on the language of a query sentence. To create an efficient Korean model, we use

the KLUE-roberta-small [8] pre-trained model and fine-tune it on public semantic textual similarity datasets. We then use PCA to reduce the dimensionality of the resulting vectors for efficient inference. For the efficient English model, we select the model that is best suited for our purposes. Figure 1 shows the overview of our approach.

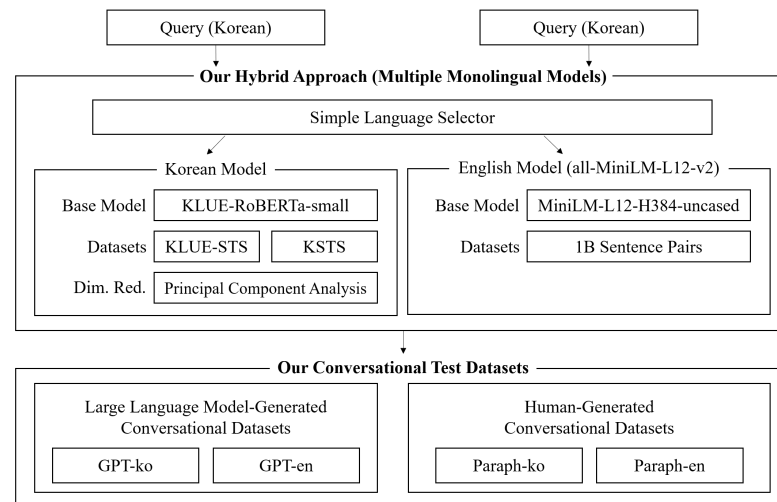


Figure 1. Overview of our approach based on multiple monolingual models.

Note that our research focuses on developing a model for identifying the most semantically similar “conversational” sentences, such as questions or commands. To verify the effectiveness of our approach, we construct new test sets composed of conversational sentences. Based on these test sets, we demonstrate that our multiple monolingual models can more accurately and quickly identify the most similar sentences than existing multilingual models. We anticipate that our approach will fulfill the needs of various artificial intelligence applications, including chatbot development, natural language processing education [9–11], and others.

The main contributions of our paper are as follows:

- First, we introduce two types of new conversational test sets to evaluate sentence similarity methods. These sets capture a wider range of conversational contexts (Section 3).
- Second, we propose a sentence embedding model specifically designed for Korean conversational sentences. To develop this model, we select a pre-trained model and fine-tune it on a Korean or Korean-translated corpus (Section 4).
- Third, we compare existing public SBERT models to identify the most efficient English sentence embedding model for this task (Section 5).
- Finally, we present a hybrid approach for embedding both Korean and English sentences, which outperforms existing multilingual approaches in terms of accuracy, elapsed time, and model size (Section 6).

2. Related Work

Efficient semantic sentence embedding is crucial for many natural language processing tasks. While several approaches have been proposed, Sentence BERT (SBERT)-based methods [1] have gained popularity due to their superior performance in terms of accuracy compared with existing methods such as GloVe [3], InferSent [4], and Universal Sentence Encoder [5]. Additionally, SBERT-based approaches are much faster than BERT-based cross-encoder approaches, making them a promising solution for large-scale NLP applications.

SBERT is a BERT [2]-based approach that uses two siamese BERTs and pooling layers to generate sentence embeddings. The network is fine-tuned to ensure that similar sentences have close embedding vectors. Intuitively, the Semantic Textual Similarity (STS) dataset is one of the appropriate datasets for fine-tuning SBERT, as it consists of two sentence

pairs with their semantic similarity scores. One of the most widely used STS datasets for fine-tuning and testing is the SemEval STS dataset [12–17]. Note that there are many different versions of BERT, such as BERT-large, RoBERTa-base [6], etc. The performance of SBERT greatly depends on not only the datasets used but also the pre-trained model used.

While most sentence embedding techniques are designed for English, there are also widely used approaches for Korean, such as KLUE [8]. KLUE offers pre-trained Korean models and datasets, including models based on BERT and RoBERTa, which are trained on extensive Korean corpora. The Korean corpora, which have been tokenized using a morpheme analyzer and a BPE-based tokenizer, amount to approximately 62.65 GB. This size is larger than the data used to train the original BERT (16GB) but smaller than the data used to train RoBERTa (160GB). Additionally, KLUE-STs provides training and development sets for STS tasks. KorSTS [18] is also a popular Korean STS dataset, which provides training, development, and test datasets for Korean STS tasks by translating the STSb [12] English dataset.

There are also approaches to create a single multilingual model that can map sentences from multiple languages into the same embedding space, such as Knowledge Distillation [7]. The Knowledge Distillation approach first prepares an English model (teacher model) and then trains the multilingual model in such a way that the translated sentences are mapped to the same embedding space as the original sentences. Surprisingly, it has been found that this multilingual approach outperforms even the Korean monolingual model for the STS task.

3. Dataset Preparation

3.1. Training Data

Fine-tuning pre-trained models using STS (Semantic Sentence Similarity) datasets is a popular method for training Sentence BERT models. These datasets typically contain pairs of sentences, along with their corresponding semantic similarity scores ranging from 0 to 5. Table 1 provides a summary of the datasets that were utilized to train our proposed models.

Table 1. Summary of the data used for training our proposed models.

Language	Dataset Name	# of Sentences
Korean	KLUE	24,374
	KorSTS	17,250
	KSTS (STS 2012~2017, Translated)	18,874
	KSICKR (Translated)	15,612
English	STSb (STS Benchmark)	16,080
	STS (STS 2012~2017)	35,804
	SICKR	19,854

To the best of our knowledge, among these datasets, the STSb (STS Benchmark) [12], STS 2012–2017 [12–17], and SICK-R datasets are the most popular STS datasets. Thus, we consider them for training English sentence embedding models. For training Korean sentence embedding models, we consider three types of datasets: (1) KLUE-STs [8], which is one of the most popular Korean STS datasets, consisting of Policy News (articles produced by ministries, etc.), ParaKQC (conversational sentences about smart home devices), and Airbnb Reviews. (2) KorSTS [18] is a translated version of STSb, but the translated sentences were refined. (3) KSTS and KSICKR are the translated versions of STS (STS 2012–2017) and SICK-R, respectively, using gpt-3.5-turbo. When translating using gpt-3.5-turbo, we exclude some of the sentences that are not correctly translated for a number of reasons. Note we use the notation “SICKR” instead of “SICK-R” and “KLUE” instead of “KLUE-STs” to express the dataset names concisely.

Note that, in the experiments described in the following sections, we used all of the training, development, and test sets for training. For instance, although KorSTS includes

separate datasets for training, development, and testing, we trained our models on all of the KorSTS datasets (including the training, development, and test sets) when reporting our results using the KorSTS dataset.

3.2. Test Data

We propose two types of test datasets to evaluate the effectiveness of Korean and English SBERT models: one made by AI and another made by humans. Our conversational dataset, called “GPT-ko”, was generated by ChatGPT (using GPT-3.5 and GPT-4). The other dataset, called “Paraph-ko”, consists of 10,000 conversational sentences extracted from ParaKQC, which was generated by humans. Although ParaKQC is also used for KLUE-STS, “Paraph-en” and KLUE-STS do not have common sentence pairs. “GPT-en” and “Paraph-en” are the English versions of “GPT-ko” and “Paraph-ko”, respectively, translated by the gpt-3.5-turbo model. Table 2 summarizes the test data used in our experiments.

Table 2. Summary of the test data used in our experiments.

Language	Dataset Name	# of Sentences	Description
Korean	GPT-ko	2000	Conversational sentences (across 34 topics) generated by ChatGPT.
	Paraph-ko	2000	Sentences extracted from an existing paraphrase dataset.
English	GPT-en	2000	English translations of the GPT-ko dataset, generated using the gpt-3.5-turbo.
	Paraph-en	2000	English translations of the Paraph-ko, generated using the gpt-3.5-turbo.

To construct the GPT-ko dataset, we followed the algorithm below:

1. We asked ChatGPT to recommend appropriate topics for generating datasets. Based on its response, we selected 10 topics (Culture, Travel, Science, Sports, Education, Food, Health, Technology, History, Humanities) and 38 subtopics.
2. We manually selected 34 subtopics from the 38, eliminating the very similar ones.
3. For each of the remaining 34 subtopics, we asked ChatGPT to randomly generate conversational sentences in Korean.
4. For each conversational sentence, we asked ChatGPT to generate one paraphrase sentence in Korean that differed in syntax as much as possible.
5. Although the generated sentences were mostly of high quality, for some subtopics, ChatGPT did not generate enough diverse sentences. In these cases, we tried switching from GPT version 3.5 to 4.0. However, generating diverse sentences was still difficult for some subtopics. For those subtopics, we stopped generating and moved on to the next ones.

We have created the GPT-ko dataset based on this algorithm, which includes 2000 conversational sentences generated in Korean across 34 subtopics. Furthermore, we have used GPT-3.5-turbo to translate the GPT-ko dataset into English, resulting in the GPT-en dataset. Table 3 displays the statistics and examples of the generated sentences. Due to limitations in table space, we present only one example sentence per topic in this paper. However, the complete dataset is available on our GitHub page (<https://github.com/tooeeworld/multiple-monolingual-model>, accessed on 2 May 2023).

Unlike the GPT-ko and GPT-en datasets, the Paraph-ko and Paraph-en datasets are generated by humans. To create these datasets, we extracted 2000 sentences from the existing ParaKQC paraphrase dataset, which contains 10,000 sentences organized into 1000 groups, with each group consisting of 10 semantically identical sentences. We randomly selected two sentences from each group to obtain the 2000 sentences in the Paraph-ko dataset. The Paraph-en dataset is the English-translated version of the Paraph-ko dataset, which was translated using gpt-3.5-turbo.

Table 3. Conversational sentence dataset generated for each of the 34 topics generated through ChatGPT. Each sentence has one paraphrase sentence within the same dataset.

Topics	# of Sentences	Examples
Traditional Korean Culture	72	What traditional holidays are celebrated in Korea?
K-POP	76	Please tell us how K-Pop idols interact with their fans.
Korean Drama	60	Please recommend a Korean drama with a prominent romance storyline.
Domestic Travel Destinations	36	Where are the travel destinations in Korea with abundant things to see?
Overseas Travel Destinations	38	What are the things to consider when planning an overseas trip?
Travel Tips	44	It's better to pack your travel bag lightly.
Science and Technology	74	How is biotechnology being utilized in the medical field?
Environmental Issues	40	Electric cars are an effective alternative for reducing air pollution.
Space Exploration	80	What are the criteria for selecting astronauts?
Soccer	74	Please explain the concept and rules of penalty kicks in soccer games.
Baseball	86	What position do you want to play after becoming a baseball player?
Basketball	54	What is your favorite NBA team?
K-12 Education	48	What do you think is the most important thing you learned in school?
Student Issues	172	What are some ways to increase self-confidence?
Education System	56	What are the issues with South Korea's university entrance system?
Korean Cuisine	28	I want to try Korean food. Do you have any recommended dishes?
Western Cuisine	28	What kitchen tools are necessary to cook Western-style cuisine?
Chinese Cuisine	28	What are the ingredients commonly used in Chinese cuisine?
Japanese Cuisine	28	What are the characteristics of the rice used in Japanese cuisine?
Bakery	22	I will find out what kind of cake is popular at the bakery.
Exercise	96	How many minutes of exercise is appropriate per day?
Diet Therapy	86	Please tell me know about suitable types and intake amounts of fats for diet.
Health Management	24	Please tell me about various ways to manage stress.
Disease Prevention	114	What are the ways to prevent and manage high blood pressure?
Artificial Intelligence	64	What is natural language processing technology in artificial intelligence?
Blockchain	72	Please explain the concept and role of tokens in blockchain.
Smartphone	60	What kind of camera technology is available on smartphones?
Korean History	50	What was the social class system like during the Joseon Dynasty?
World History	70	Please explain the collapse of the Roman Empire and its causes.
Historical Events and Figures	36	What type of movement did Mahatma Gandhi serve as a leader?
Literature	40	The author can convey various messages to readers through their own works.
Philosophy	60	Does scientific progress make humans happier?
Religion	56	What role does religion play in human emotions?
Psychology	28	Why do we form a sense of self?

Note that in all of the generated test sets (GPT-ko, GPT-en, Paraph-ko, and Paraph-en), there are 2000 conversational sentences. Each sentence has exactly one semantically identical sentence (paraphrased sentence) within the dataset. When evaluating the models, each model found the most semantically similar sentence for each query sentence within the dataset, and we counted the number of times it accurately finds it. Because we set every sentence to be a query sentence, there are 2000 query sentences. Given a query sentence, there are 1999 candidates, and among them, only one sentence is the answer. Obviously, it is a very difficult task to find the answer, and we will show how well existing and our approach models perform.

4. An Efficient Korean Embedding Model

In this section, we aim to create an efficient Korean sentence embedding models by utilizing KLUE-RoBERTa-base, a pre-trained model based on a large Korean corpus. The pre-trained models are fine-tuned using the Korean datasets introduced in Section 3.1. We select different combinations of Korean datasets to train and evaluate our models, and the results are reported in Table 4. Additionally, we present the results of publicly available Korean SBERT models from Hugging Face and GitHub. The table presents accuracy as a percentage, where a value of 50.00 indicates that the corresponding model correctly

identifies the most semantically similar sentence for 1000 query sentences out of a total of 2000 query sentences.

Table 4. Comparison of existing Korean sentence embedding models (the first three models) and our embedding models trained with different datasets.

Korean SBERT Models	Accuracy (%)				
	GPT-ko	Paraph-ko	GPT-en	Paraph-en	Average
snunlp/KR-SBERT-V40K-klueNLI-augSTS	65.65	79.40	30.50	45.90	55.36
jhgan/ko-sroberta-multitask	80.95	86.40	53.40	68.00	72.19
Huffon/sentence-klue-roberta-base	85.50	89.00	55.25	70.95	75.18
(KLUE-RoBERTa-base) KLUE	87.00	89.05	54.10	69.90	75.01
(KLUE-RoBERTa-base) KorSTS	80.55	84.85	57.25	69.55	73.05
(KLUE-RoBERTa-base) KSTS	84.05	85.95	60.00	69.55	74.89
(KLUE-RoBERTa-base) KSICKR	63.60	76.05	46.90	66.10	63.16
(KLUE-RoBERTa-base) KLUE+KorSTS	88.10	88.80	55.95	71.05	75.98
(KLUE-RoBERTa-base) KLUE+KSTS	88.30	89.85	61.30	72.05	77.88
(KLUE-RoBERTa-base) KLUE+KSICKR	87.60	89.65	58.20	72.35	76.95
(KLUE-RoBERTa-base) KLUE+KSTS+KorSTS	88.35	89.40	62.40	73.50	78.41
(KLUE-RoBERTa-base) KLUE+KSTS+KorSTS+KSICKR	87.40	89.15	60.75	69.40	76.68

In these experiments, we used the three popular Korean embedding models as comparators:

- The public model snunlp/KR-SBERT-V40K-klueNLI-augSTS (<https://github.com/snunlp/KR-SBERT>, accessed on 2 May 2023) is based on the KR-BERT-V40K pre-trained model. It was fine-tuned using the KLUE-NLI and KorSTS datasets. The KorSTS dataset was augmented by Augmented SBERT [19];
- The public model jhgan/ko-sroberta-multitask (<https://github.com/jhgan00/ko-sentence-transformers>, accessed on 2 May 2023) is based on the KLUE-RoBERTa-base pre-trained model. It was fine-tuned using the KorNLI and KorSTS datasets. This model is considered the best among the models introduced on the GitHub page;
- The public model Huffon/sentence-klue-roberta-base (<https://huggingface.co/Huffon/sentence-klue-roberta-base>, accessed on 2 May 2023) is based on the KLUE-RoBERTa-base pre-trained model. It was fine-tuned using the KLUE-STs dataset.

For a fair comparison, we did not use any development sets or other datasets related to the test sets. We trained each of our approaches for only one epoch.

The experimental results indicate that the highest average accuracy is achieved when using KLUE-RoBERTa-base as a pre-trained model and fine-tuning with three datasets: KLUE, KSTS, and KorSTS. Additionally, the best results for Korean test sets are obtained when only the KLUE and KSTS datasets are used. Based on these findings, we have identified three important factors. Firstly, using an appropriate pre-training model is crucial. The accuracy difference between the KR-SBERT-V40K based model and those based on KLUE-RoBERTa-base is significant. Secondly, dataset quality is more important than the dataset's quantity. Unfortunately, some popular datasets, especially KSICKR, do not lead to improved performance and can even degrade performance. Finally, although high-accuracy results are achieved on Korean test sets, the results on the English test sets are relatively lower. In the following sections, we describe novel ways to address these limitations.

5. An Efficient English Embedding Model

In this section, we conduct experiments to find the best model for conversational sentence embeddings in English. Our approach follows the same experimental setup as described in Section 4, but with English datasets instead of Korean datasets. Additionally, to compare our approach with existing methods, we use all of the selected English models

available on sbert.net (https://sbert.net/docs/pretrained_models.html, accessed on 2 May 2023). The experimental results are presented in Table 5.

Table 5. Comparison of existing English sentence embedding models (the first nine models) and our embedding models trained with different datasets.

English SBERT Models	Accuracy (%)				
	GPT-ko	Paraph-ko	GPT-en	Paraph-en	Average
paraphrase-albert-small-v2	1.30	0.25	78.60	72.55	38.18
all-MiniLM-L6-v2	12.60	17.20	81.25	77.95	47.25
paraphrase-MiniLM-L3-v2	14.35	21.85	77.65	76.70	47.64
multi-qa-mpnet-base-dot-v1	16.10	23.30	80.00	72.95	48.09
all-mpnet-base-v2	16.50	21.25	80.45	74.40	48.15
all-MiniLM-L12-v2	15.05	25.05	80.55	79.30	49.99
multi-qa-MiniLM-L6-cos-v1	18.60	25.40	80.70	76.10	50.20
multi-qa-distilbert-cos-v1	17.45	26.80	80.80	76.35	50.35
all-distilroberta-v1	18.60	30.50	76.60	75.85	50.39
(KLUE-RoBERTa-base) STSb	74.45	83.65	65.45	69.95	73.38
(KLUE-RoBERTa-base) STS	82.60	85.30	66.65	69.65	76.05
(KLUE-RoBERTa-base) SICKR	52.15	74.25	47.60	58.90	58.23
(KLUE-RoBERTa-base) STSb+STS	83.30	85.90	69.15	71.40	77.44
(KLUE-RoBERTa-base) STSb+STS+SICKR	81.50	85.15	66.05	70.50	75.80

In these results, one of our approaches (fine-tuning on STSb and STS English datasets) achieves the best average accuracy. An interesting thing is that the Korean pre-trained model “KLUE-RoBERTa-base” performed well for English sentence embedding. Our hypothesis is that, during its pre-training phase, the Korean datasets used to train the model may have contained some English expressions. This could have enabled the pre-trained model to also learn how to process English expressions effectively. However, although it is helpful for our approaches to train on English training datasets, state-of-the-art SBERT models still significantly outperform our approaches in the English test datasets. Surprisingly, the best performing English models are very small models: all-MiniLM-L6-v2 performs the best in the GPT-en dataset, and all-MiniLM-L12-v2 performs the best in the Paraph-en dataset and on average in the English test datasets.

Our experiments showed that, although the English models achieved exceptional performance on the English datasets, they performed poorly on the Korean datasets. This suggests that using two separate monolingual models, one for English and one for Korean, in collaboration may lead to better results. In the next section, we will explore this possibility further.

6. A Hybrid Approach

In the previous sections, we conducted an analysis to identify the best-performing embedding models for each language in our conversational datasets. Based on our findings, we aim to answer the following research questions:

- Can we achieve better accuracy in our test datasets by using our Korean and English sentence embedding models together, compared with using state-of-the-art multilingual models?
- If so, can we achieve both higher accuracy and faster processing times in our test datasets by using even smaller versions of the two monolingual models in combination, compared with using state-of-the-art multilingual models?

In order to answer these questions, we have configured our hybrid approach as follows:

- We utilized the KLUE-RoBERTa-small pre-trained model for developing a Korean monolingual model. It is worth noting that this model is smaller than the KLUE-

RoBERTa-base utilized in Section 4. Hence, it may achieve lower accuracy but perform faster. To fine-tune our model, we employed the KLUE+KSTS datasets, which demonstrated the best performance on Korean test datasets, as introduced in Section 4.

- To speed up the inference time, we applied PCA (Principal Component Analysis) to reduce the dimensionality of the resulting vectors. We followed a similar approach used in sbert.net: we randomly extracted 20,000 sentences from the KLUE-NLI dataset (which is not relevant to our test datasets) to generate sample embeddings for PCA. The original dimensionality of the Korean model was 768, but our resulting vectors, applied after PCA, had a dimensionality of 384. We added the “PCA layer” to our model, and the PCA process ran very quickly.
- For the English monolingual model, we used the all-MiniLM-L12-v2 model, which was the best performing model on average in the English test sets. Because this model was already a very small model, we did not need to make the model smaller. Additionally, the output of the model is also small (384-dimensional vectors), so we did not apply PCA to reduce the dimensionality.
- For combining these two monolingual models, we used a very simple hybrid approach: if a query sentence has at least one Korean letter, then the Korean model makes the corresponding 384-dimensional vector. Otherwise, the English model makes the corresponding 384-dimensional vector.

Table 6 displays the experimental results of existing multilingual models and our approaches. The comparators include two types of models. Firstly, the first five models in this table are selected multilingual models from sbert.net (https://sbert.net/docs/p/retrained_models.html, accessed on 2 May 2023), which are based on English sentence embedding models and were generated using Knowledge Distillation. Secondly, the next three models in this table are our own models that use both Korean and English training datasets to fine-tune the KLUE-RoBERTa-base model. Finally, the last model is generated by our hybrid approach, which combines two small monolingual models. The experimental results show that our hybrid approach demonstrates consistent performance across the four test datasets and outperforms the other approaches on average.

Table 6. Comparison of existing multilingual embedding models (the first five models) and our embedding models trained with different datasets. Note our hybrid approach uses two small monolingual models.

Multilingual SBERT Models	Accuracy (%)				
	GPT-ko	Paraph-ko	GPT-en	Paraph-en	Average
clip-ViT-B-32-multilingual-v1	57.20	56.25	65.60	67.05	61.53
distiluse-base-multilingual-cased-v2	51.35	65.10	60.55	74.60	62.90
distiluse-base-multilingual-cased-v1	56.00	73.70	61.00	75.30	66.50
paraphrase-multilingual-MiniLM-L12-v2	74.40	62.10	82.60	76.90	74.00
paraphrase-multilingual-mpnet-base-v2	78.55	76.85	83.00	77.70	79.03
(KLUE-RoBERTa-base) KLUE+KSTS+STSb	87.65	88.30	67.05	71.50	78.63
(KLUE-RoBERTa-base) KLUE+KSTS+STS	87.35	88.50	67.60	72.15	78.90
(KLUE-RoBERTa-base) KLUE+KSTS+STSb+STS	87.15	88.30	66.35	70.80	78.15
Our Hybrid Approach	85.60	87.35	80.55	79.30	83.20

Based on our previous experiments, only two existing models had an average accuracy of 75% or higher: “Huffon/sentence-klue-roberta-base” and “paraphrase-multilingual-mpnet-base-v2”. Therefore, we compare our approach with these models in detail. We compare not only accuracy but also model size, encoding time (CPU/GPU), and time for finding the most similar sentences (CPU/GPU). These elapsed times are crucial in many AI-related applications, particularly in AI educational programming environments. Table 7 shows that our hybrid approach outperforms the existing state-of-the-art approaches in

terms of total model size, encoding time, time for finding the most similar sentences, and average accuracy. Note that no efficient techniques, such as numba (<https://numba.pydata.org/>, accessed on 2 May 2023), were used in this process. As a result, the absolute values presented may not be meaningful. The elapsed times are only provided for a relative comparison.

Table 7. Comparison of sentence encoding and nearest neighbor search times for models with an average accuracy of 75% or higher, using different hardware configurations. Sentence encoding time per sentence (ENC) and nearest neighbor search time per sentence (NNS) are measured in seconds.

Multilingual SBERT Models	Test Sets	Elapsed Time (Seconds)				Accuracy (%)
		ENC-GPU	NNS-GPU	ENC-CPU	NNS-CPU	
Huffon/sentence-klue-roberta-base (Total Model Size: 443 MB)	GPT-ko	0.0102	0.0727	0.0324	0.0890	85.45
	Paraph-ko	0.0097	0.0725	0.0309	0.0884	89.00
	GPT-en	0.0104	0.0728	0.0349	0.0882	55.25
	Paraph-en	0.0100	0.0732	0.0365	0.0887	70.95
	Average	0.0101	0.0728	0.0337	0.0886	75.16
paraphrase-multilingual-mpnet-base-v2 (Total Model Size: 1.11 GB)	GPT-ko	0.0098	0.0731	0.0320	0.0896	78.55
	Paraph-ko	0.0100	0.0732	0.0327	0.0893	76.85
	GPT-en	0.0100	0.0734	0.0312	0.0895	83.00
	Paraph-en	0.0099	0.0731	0.0268	0.0828	77.70
	Average	0.0099	0.0732	0.0307	0.0878	79.03
Our Hybrid Approach (Korean+English) (Total Model Size: 393 MB)	GPT-ko	0.0055	0.0380	0.0167	0.0475	85.60
	Paraph-ko	0.0057	0.0384	0.0162	0.0466	87.35
	GPT-en	0.0086	0.0380	0.0115	0.0429	80.55
	Paraph-en	0.0084	0.0379	0.0123	0.0431	79.30
	Average	0.0071	0.0381	0.0142	0.0450	83.20

It is interesting that the Knowledge Distillation paper showed that the multilingual approach outperformed the Korean monolingual approach for Korean test sets, which is opposite to our study's results. Instead, our results are somewhat in line with the study using two monolingual models for intent classification and slot filling [20], which showed they were better than multilingual models for the same task. It is also interesting that we can make models work faster while maintaining higher accuracy by combining small models.

7. Conclusions

In this paper, we present an effective hybrid approach for embedding Korean and English sentences in order to find the most similar sentence to a given query. Our main idea is to use two small but effective monolingual models, one for each language, and to choose the appropriate model based on the query sentence. For the Korean sentence embedding model, we use the small pre-trained model (KLUE-roberta-small) for fast encoding and fine-tune it on the KLUE-STS and KSTS datasets. To speed up the process of finding the most similar sentence using cosine similarity, we apply a principal component analysis to reduce the dimensionality by half. The English sentence embedding model is carefully chosen based on experimental results and model size. To verify the effectiveness of our approach, we constructed two types of datasets: one created by artificial intelligence across 34 topics and another created by humans. Our experimental results demonstrate that our hybrid approach outperforms existing multilingual and monolingual approaches in terms of average accuracy, encoding time, and time for finding semantically similar sentences on both CPU and GPU. We expect that our approach will have broad applications in natural language processing, including artificial intelligence education.

Author Contributions: Conceptualization, Y.P. and Y.S.; methodology, Y.P. and Y.S.; software, Y.P. and Y.S.; validation, Y.P. and Y.S.; investigation, Y.P. and Y.S.; data curation, Y.P. and Y.S.; writing—original draft preparation, Y.P. and Y.S.; writing—review and editing, Y.P. and Y.S.; funding acquisition, Y.S. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by an Incheon National University Research Grant in 2022.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Access to some of the data used in our paper is available at the following repository: <https://github.com/tooeeworld/multiple-monolingual-model>, accessed on 2 May 2023.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Reimers, N.; Gurevych, I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, 3–7 November 2019; pp. 3982–3992.
2. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the NAACL-HLT, Minneapolis, MN, USA, 2–7 June 2019; pp. 4171–4186.
3. Pennington, J.; Socher, R.; Manning, C.D. Glove: Global vectors for word representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 26–28 October 2014; pp. 1532–1543.
4. Conneau, A.; Kiela, D.; Schwenk, H.; Barrault, L.; Bordes, A. Supervised Learning of Universal Sentence Representations from Natural Language Inference Data. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark, 9–11 September 2017; pp. 670–680.
5. Cer, D.; Yang, Y.; Kong, S.Y.; Hua, N.; Limtiaco, N.; John, R.S.; Constant, N.; Guajardo-Cespedes, M.; Yuan, S.; Tar, C.; et al. Universal sentence encoder. *arXiv* **2018**, arXiv:1803.11175.
6. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. Roberta: A robustly optimized bert pretraining approach. *arXiv* **2019**, arXiv:1907.11692.
7. Reimers, N.; Gurevych, I. Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Online, 16–20 November 2020; pp. 4512–4525.
8. Park, S.; Moon, J.; Kim, S.; Cho, W.I.; Han, J.; Park, J.; Song, C.; Kim, J.; Song, Y.; Oh, T.; et al. Klue: Korean language understanding evaluation. *arXiv* **2021**, arXiv:2105.09680.
9. Park, Y.; Shin, Y. Tooee: A novel scratch extension for K-12 big data and artificial intelligence education using text-based visual blocks. *IEEE Access* **2021**, *9*, 149630–149646. [\[CrossRef\]](#)
10. Park, Y.; Shin, Y. A Block-Based Interactive Programming Environment for Large-Scale Machine Learning Education. *Appl. Sci.* **2022**, *12*, 13008. [\[CrossRef\]](#)
11. Park, Y.; Shin, Y. Text Processing Education Using a Block-Based Programming Language. *IEEE Access* **2022**, *10*, 128484–128497. [\[CrossRef\]](#)
12. Cer, D.; Diab, M.; Agirre, E.E.; Lopez-Gazpio, I.; Specia, L. SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Cross-lingual Focused Evaluation. In Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), Vancouver, BC, Canada, 3–4 August 2017; pp. 1–14.
13. Agirre, E.; Cer, D.; Diab, M.; Gonzalez-Agirre, A. Semeval-2012 task 6: A pilot on semantic textual similarity. In Proceedings of the First Joint Conference on Lexical and Computational Semantics, Montreal, QC, Canada, 7–8 June 2012; pp. 385–393.
14. Agirre, E.; Cer, D.; Diab, M.; Gonzalez-Agirre, A.; Guo, W. SEM 2013 shared task: Semantic textual similarity. In Proceedings of the Second Joint Conference on Lexical and Computational Semantics, Atlanta GA, USA, 13–14 June 2013; pp. 32–43.
15. Agirre, E.; Banea, C.; Cardie, C.; Cer, D.M.; Diab, M.T.; Gonzalez-Agirre, A.; Guo, W.; Mihalcea, R.; Rigau, G.; Wiebe, J. SemEval-2014 Task 10: Multilingual Semantic Textual Similarity. In Proceedings of the 8th International Workshop on Semantic Evaluation, Dublin, Ireland, 23–24 August 2014; pp. 81–91.
16. Agirre, E.; Banea, C.; Cardie, C.; Cer, D.; Diab, M.; Gonzalez-Agirre, A.; Guo, W.; Lopez-Gazpio, I.; Maritxalar, M.; Mihalcea, R.; et al. Semeval-2015 task 2: Semantic textual similarity, english, spanish and pilot on interpretability. In Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), Denver, CO, USA, 4–5 June 2015; pp. 252–263.
17. Agirre, E.; Banea, C.; Cer, D.; Diab, M.; Gonzalez Agirre, A.; Mihalcea, R.; Rigau, G.; Wiebe, J. Semeval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In Proceedings of the 10th International Workshop on Semantic Evaluation, San Diego, CA, USA, 16–17 June 2016; pp. 497–511.

18. Ham, J.; Choe, Y.J.; Park, K.; Choi, I.; Soh, H. KorNLI and KorSTS: New Benchmark Datasets for Korean Natural Language Understanding. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16–20 November 2020; pp. 422–430.
19. Thakur, N.; Reimers, N.; Daxenberger, J.; Gurevych, I. Augmented SBERT: Data Augmentation Method for Improving Bi-Encoders for Pairwise Sentence Scoring Tasks. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Online, 6–11 June 2021; pp. 296–310.
20. Lothritz, C.; Allix, K.; Lebichot, B.; Veiber, L.; Bissyandé, T.F.; Klein, J. Comparing multilingual and multiple monolingual models for intent classification and slot filling. In Proceedings of the 26th International Conference on Applications of Natural Language to Information Systems, Saarbrücken, Germany, 23–25 June 2021; pp. 367–375.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.