applied sciences

MDPI

Article

An Efficient Document Retrieval for Korean Open-Domain Question Answering Based on ColBERT

Byungha Kang †, Yeonghwa Kim † and Youhyun Shin *

Department of Computer Science and Engineering, Incheon National University, Incheon 22012, Republic of Korea; docdoca@inu.ac.kr (B.K.); 112movie@naver.com (Y.K.)
* Correspondence: yhshin@inu.ac.kr
† These authors contributed equally to this work.

Abstract: Open-domain question answering requires the task of retrieving documents with high relevance to the query from a large-scale corpus. Deep learning-based dense retrieval methods have become the primary approach for finding related documents. Although deep learning-based methods have improved search accuracy compared to traditional techniques, they simultaneously impose a considerable increase in computational burden. Consequently, research on efficient models and methods that optimize the trade-off between search accuracy and time to alleviate computational demands is required. In this paper, we propose a Korean document retrieval method utilizing ColBERT's late interaction paradigm to efficiently calculate the relevance between questions and documents. For open-domain Korean question answering document retrieval, we construct a Korean dataset using various corpora from AI-Hub. We conduct experiments comparing the search accuracy and inference time among the traditional IR (information retrieval) model BM25, the dense retrieval approach utilizing BERT-based models for Korean, and our proposed method. The experimental results demonstrate that our approach achieves a higher accuracy than BM25 and requires less search time than the dense retrieval method employing KoBERT. Moreover, the most outstanding performance is observed when using KoSBERT, a pre-trained Korean language model that learned to position semantically similar sentences closely in vector space.

Keywords: natural language processing; deep neural networks; question answering; document retrieval

Citation: Kang, B.; Kim, Y.; Shin, Y. An Efficient Document Retrieval for Korean Open-Domain Question Answering Based on ColBERT. *Appl. Sci.* **2023**, *13*, 13177. https://doi.org/10.3390/app132413177

Academic Editor: Dimitris Mourtzis

Received: 31 October 2023
Revised: 20 November 2023
Accepted: 6 December 2023
Published: 12 December 2023

1. Introduction

Open-domain question answering (ODQA) aims to answer queries based on large knowledge corpora, like the web, without explicit evidence. The typical framework for ODQA is structured as a two-stage process, consisting of a retriever and a reader [1]. The retriever selects a candidate set relevant to a given question from a large corpus, and the reader predicts answers to the question from the retrieved set. The performance of the retriever is typically crucial for the overall QA performance, as it determines the quality of the candidate set. Therefore, there has been extensive research conducted to enhance the retriever's performance [1–3].

Traditional retrievers mainly use sparse representations through TF-IDF and BM25 [2,3], often encountering term-mismatch problems due to the difference in terms used in the query and those found in the documents. Furthermore, there are limitations in their performance, as it is difficult to reflect the semantic relationship between documents and queries. Recently, the emergence of dense retrieval utilizing deep learning has shown significant improvements in search performance. Pre-trained language model-based search, particularly BERT [4], has been very effective. However, BERT-based search requires supplying each query–document pair through a large neural network to calculate relevance scores, increasing the computational cost and inference time by tens of thousands of milliseconds (ms)

compared to previous approaches [5]. Although BERT-based search has brought significant performance improvements, it has also led to substantial computational costs.

To address this issue, varied research on retriever models and techniques that pursue a balance between accuracy, computational cost, and latency have been conducted [6]. In particular, the model for ODQA deals with exceptionally vast datasets, leading to the challenge of retrievers taking considerable time to find question-relevant information or documents. The Learning Index for Learning Passage Retrieval (LIDER) [7] achieves a balance between search speed and accuracy during training by dynamically adapting the corpus index, as opposed to locality-sensitive hashing (LSH) [8] and inverted file (IVF) [9]. While LSH and IVF involve research efforts utilizing efficient approximate nearest neighbor (ANN) methods for enhanced search speed, they often suffer from significant degradation in search accuracy [10]. ColBERT is an innovative ranking model designed to enhance retrieval by efficiently adapting deep language models (specifically BERT) for document ranking. It employs a novel late-interaction architecture that independently encodes queries and documents using BERT, followed by a cost-effective interaction step to model their fine-grained similarity. This approach not only leverages the power of deep language models but also speeds up query processing by enabling pre-computation of document representations. ColBERT's efficiency makes it competitive with existing BERT-based models, outperforming non-BERT baselines while being significantly faster and more resource-efficient.

While ColBERT presents itself as a highly efficient retrieval model, its application in the Korean language context remains limited. Research on natural language processing in Korean has not gained as much widespread attention as it has in English, leading to Korean being often referred to as a language with limited resources [11]. Consequently, this results in a shortage of Korean ODQA data, and the scale of the available data is indeed limited. In this paper, we construct a large-scale dataset for training Korean ODQA, surpassing the existing machine reading comprehension (MRC) datasets. We conduct an analysis of our data, including different question types. Additionally, we apply the efficient ColBERT model to our dataset, conducting comparative experiments with conventional retrieval methods to validate the effectiveness of an efficient retrieval model on a large-scale Korean dataset. The ColBERT approach demonstrates a performance surpassing term-based retrieval BM25 and exhibits significantly faster search times compared to KoBERT. Furthermore, utilizing various pre-trained Korean language models for ODQA tasks, we identify KoSBERT as the most effective and efficient pre-trained Korean language model. Lastly, through performance comparisons based on tokenizers in the ablation study, we underscore the necessity for efficient models tailored to the complexities of the Korean language.

This paper's main contributions are outlined as follows:

- We adapt and propose the application of ColBERT to Korean document retrieval, demonstrating its efficiency and effectiveness in open-domain Korean question answering.
- We construct and analyze a large-scale Korean open-domain question answering dataset, filling a gap in Korean natural language processing resources.
- Through extensive evaluations using various pre-trained language models, we demonstrate that KoSBERT significantly outperforms other models in document retrieval tasks, emphasizing the effectiveness of KoSBERT specifically for Korean language retrieval tasks.

The rest of this paper is organized as follows: Section 2 discusses the related work, providing a background and context for this study. Section 3 delves into the adaptation of ColBERT for Korean open-domain question answering, including details on the query encoder, document encoder, and late interaction mechanism. Section 4 covers the Korean open-domain question answering dataset construction, detailing the creation of both the training and evaluation datasets. Section 5 presents the experimental setup and results, including the datasets used, baseline comparisons, evaluation metrics, computational

details, results, and an ablation study. Section 6 discusses the implications of the study, highlighting how the ColBERT model's effectiveness in Korean language processing can be extended to various languages. Finally, Section 7 concludes the paper with a summary of the findings and suggestions for future research directions.

2. Related Work

The architecture of open domain question answering (ODQA) systems primarily combines information retrieval (IR) and reader modules. IR plays a crucial role in searching for evidence passages within a vast knowledge corpus to answer questions. Traditional non-neural models utilized in IR include TF-IDF [2] and BM25 [3]. These methods rely on lexical information, leading to significantly lower performance when there is a mismatch between the query and passage terminology. However, with advancements in deep learning and the introduction of powerful pre-trained language models such as BERT [4], methods surpassing lexical-level information have been proposed. These methods leverage semantic correlations between questions and passages for more effective document retrieval. The interaction-based retriever illustrated in Figure 1a inputs the query and context in the format [CLS] Query [SEP] Context [SEP] into BERT, enabling the computation of scores between queries and documents through derived representations [12,13].

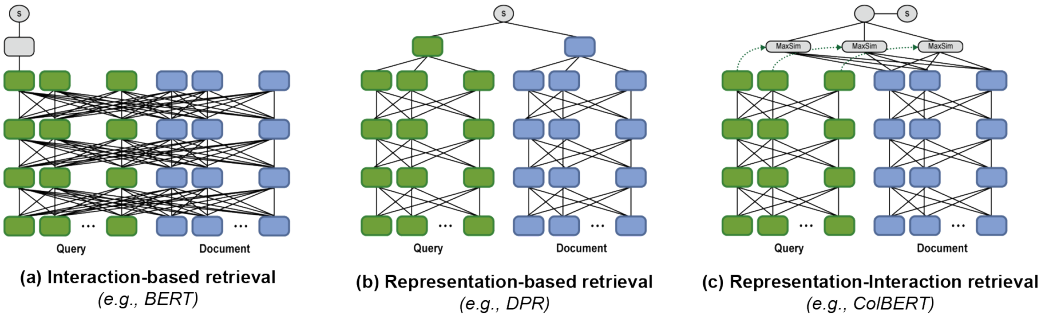


Figure 1. Three types of dense retrievers [14]. Given a question (query) and a document as inputs, calculate the degree of relevance (s).

Nevertheless, Transformer-based pre-trained language models (PLMs) tend to be computationally intensive and slow when dealing with longer input lengths. In response, dense passage retrieval (DPR) [3] independently encodes queries and passages for relevance score calculation, as shown in Figure 1b. Despite its efficiency, DPR does not consider the interaction between query and passage, which can limit its performance.

To balance efficiency and accuracy, ColBERT [5] has been proposed. ColBERT, a representation–interaction retriever model, is depicted in Figure 1c. After encoding questions and documents separately using BERT-based encoders, it calculates token embedding scores for each question across all documents. Subsequently, all scores are summed to form the final relevance score between the question and document. ColBERT has demonstrated an effective balance between search performance and processing speed.

Recent research efforts aim to develop efficient ODQA systems [6]. Techniques like brute search [15], hierarchical navigable small world graphs (HNSW) [16], approximate nearest neighbor (ANN) [17], locality-sensitive hashing (LSH) [8], and inverted file (IVF) [9] have been applied in various studies. Additionally, approaches such as removing the reader from the retrieval–reader structure [18] or directly generating answers from given questions using generative-only methods [19–21] have been proposed to enhance ODQA efficiency.

This study aims to apply ColBERT, which shows a balanced performance in terms of search speed and accuracy, to the Korean language. In the ColBERT structure, BERT-based models are used to derive embeddings for queries and passages. BERT has significantly outperformed traditional methods in various tasks, including review classification [22]. Several versions of BERT exist in Korea [23], and open-source PLMs have been trained in different environments and datasets, necessitating efforts to find models suitable for

specific tasks. Attempts to apply Korean BERT in document classification [24], comment analysis [25], and medical fields have been proposed [26].

3. ColBERT for Korean Open-Domain Question Answering

As shown in Figure 2, the ColBERT model used in this paper can be divided into three main components: the query encoder, document encoder, and late interaction. The given question and document generate their respective vector representations through the KoBERT-based encoders. Since the question and document encoders share a single KoBERT model, special tokens [Q] and [D] are added to each input to differentiate between questions and documents. The resulting question and document representations are passed through the late interaction, which calculates the final relevance score using the MaxSim operation, as in [5]. This involves calculating the maximum cosine similarity between the document and question embeddings and summing them up to produce the final score.

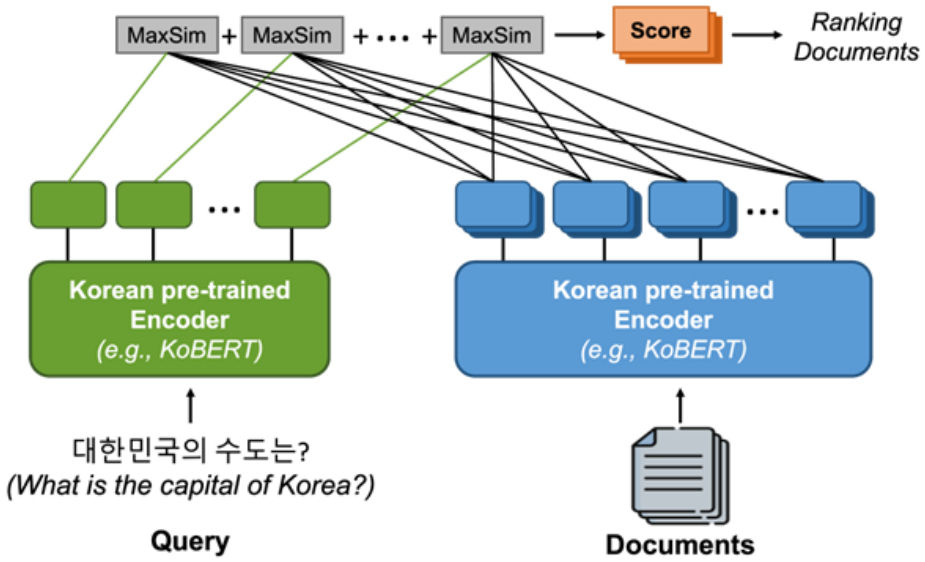


Figure 2. Document retrieval model architecture using KoBERT-based ColBERT.

3.1. Query Encoder

Given a question q, it is tokenized into q1, q2, . . . , ql using the SentencePiece-based KoBERT tokenizer [27], and a special token [Q] is added after the [CLS] token to distinguish it from the document sequence. If the length of the question is less than the predetermined maximum length Nq, it is padded with BERT's special [MASK] token to reach Nq; otherwise, it is truncated. The token sequence padded with masked tokens passes through KoBERT to compute the contextualized representation of each token. This padding and masking strategy for questions can serve as a query augmentation, which can improve the model's performance [5]. The representations obtained through KoBERT pass through a linear layer without an activation function, adjusting the dimensionality of the embeddings. This dimensionality reduction may impose constraints on the question encoding but significantly benefits the runtime. The output embeddings are then normalized using L2 norm. The cosine similarity between two embeddings, in the range of $[-1, 1]$, is calculated using the dot product.

3.2. Document Encoder

Similar to the query encoder, given a document d, it is tokenized into d1, d2, . . . , dk using the KoBERT tokenizer, and a special token [D] is added after the [CLS] token to distinguish it from the question sequence. However, unlike the query encoder, the document encoder does not use the [MASK] token for padding based on document length.

The document embeddings produced by KoBERT pass through a linear layer. The number of document embeddings is then reduced through punctuation filtering.

3.3. Late Interaction

To compute the relevance score between a query q and a document d, ColBERT employs a late interaction between their bags of embeddings. The query and document are separately input into KoBERT to obtain the token-level embeddings for each query and document token. Next, for each query token, the MaxSim operation is performed by calculating the similarity between the query token and all document tokens and selecting the maximum similarity value. The similarity between tokens is calculated through cosine similarity or squared L2 distance. Finally, the relevance score between query q and document d is obtained by summing the result values from the MaxSim operation conducted for all tokens in the query.

$$\mathbf{Score}_{q,d} := \sum_{i \in [|E_q|]} \max_{j \in [|E_d|]} \mathbf{E}_{q_i} \cdot \mathbf{E}_{d_j}^T \tag{1}$$

Afterward, the KoBERT encoders are fine-tuned, and the added special token parameters are trained. Given a triple <q, d+, d−>, consisting of a question, a relevant document, and an irrelevant document, the relevance scores for each document with respect to the question are calculated. ColBERT is optimized using Adam while calculating the pairwise softmax cross-entropy loss for each pair.

Top-k Re-ranking Re-ranking involves reordering a pre-indexed set of k documents (e.g., k = 1000) for a given query q. First, ColBERT computes the embedding matrix Eq for the given query q and calculates embeddings for the k documents to form a three-dimensional tensor D, which is then moved to the GPU memory. Next, the dot product between Eq and D is computed across multiple mini-batches. The resulting three-dimensional tensor represents a collection of cross-match matrices between query q and each document. To calculate scores for each document, we reduce the computed matrix through max-pooling across document tokens (MaxSim operation) and then compute the relevance score by summing over query tokens. The documents are re-ranked based on the calculated scores for the k documents.

4. Korean Open-Domain Question Answering Dataset Construction

4.1. Training Dataset

To construct a dataset for Korean open-domain question answering, we used four distinct domains datasets provided by AI-Hub [28]: *general knowledge* [29], *machine reading comprehension* [30], *book material machine reading comprehension* [31], and *large-scale web-based Korean corpus* [32]. *General knowledge* is the used MRC dataset containing knowledge from the Korean Wikipedia, with a query count of 67,775, a query length in characters ranging from 1 to 92, and a document count of 43,499. The document length in characters ranges from 46 to 1782. *Machine reading comprehension* is a news-based MRC dataset and contains news article data in nine fields (politics, economy, society, etc.). It consists of a query count of 440,322, a query length of 0–254, a document count of 89,164, and a document length of 44–26,747. *Book material machine reading comprehension* is an MRC dataset utilizing book material on various topics. It consists of a query count of 949,332 with a query length of 6–231, a document count of 181,198, and a document length of 343–678. *Large-scale web-based Korean corpus* for document augmentation is based on website-based large-scale text data. It contains more than 1 billion words of text data and includes news from 17 different fields (IT/tech, culture/fashion/beauty, international, etc.). The overall construction process of the dataset is depicted in Figure 3, and the code to generate the data is available at https://github.com/movie112/ColBERTforKorean (accessed on 20 November 2023).

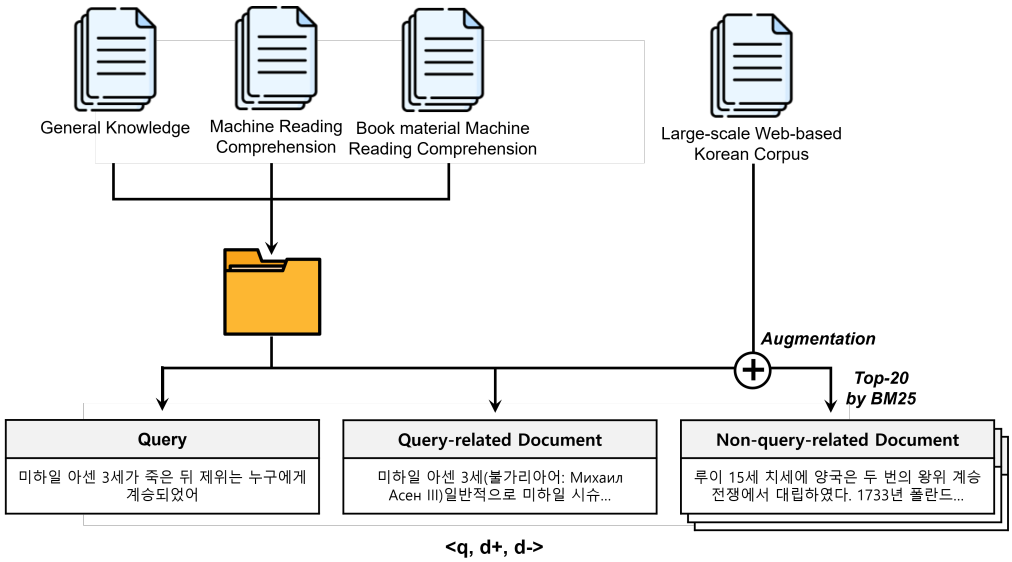


Figure 3. Process of constructing a dataset for open-domain question answering.

The dataset is structured to reflect the format of the MS MARCO Ranking [33] dataset used in ColBERT. First, we limit the data to documents with lengths of 1000 or less, based on the maximum document length in [33], using three datasets: general knowledge, machine reading comprehension, and book material machine reading comprehension. The dataset is constructed in the form of triples <q, d+, d−> for training, following the format of [33]. The query and relevant document pairs <q, d+> are derived from the general knowledge, machine reading comprehension, and book material machine reading comprehension datasets. To obtain low-relevance documents d−, we extracted the top 20 documents using the BM25 algorithm from the entire document collection, which is augmented with AI-Hub's large-scale web-based Korean corpus dataset.

To assess performance by query length, we partitioned datasets based on maximum query lengths. Queries start from a length of 5, set to reduce noise, and datasets were partitioned with max query lengths of 20, 50, and the overall longest query.The statistical information of the final dataset used for training is shown in Table 1.

Table 1. ColBERT training triple <q, d+, d−> dataset statistics information.

Divisions	Min, Max Query Length	Avg Query Length	Avg Document Length	Queries	d+ Documents	d− Documents
20	5~20	16.7	582.2	122,328	69,001	413,711
50	5~50	28.8	544.9	126,574	97,254	460,440
200	5~183	31.3	544.7	126,725	98,714	461,452

4.2. Evaluation Dataset

In our study, we employ the BM25 algorithm as a preliminary step in constructing the test set for evaluating our proposed methodology. BM25 identifies and ranks documents based on their relevance to a query, which is achieved by balancing the term frequency with the term's inverse document frequency, adjusted for document length. The score of a document D for a query Q using BM25 is computed as follows:

$$\text{Score}(D,Q) = \sum_{i=1}^{n} IDF(q_i) \times \frac{f(q_i,D) \times (k_1+1)}{f(q_i,D) + k_1 \times (1 - b + b \times \frac{|D|}{\text{avgdl}})} \quad (2)$$

This formula incorporates the term frequency $f(q_i, D)$ and inverse document frequency $IDF(q_i)$, and adjustments for the document's length and the average document length in the corpus, with the hyperparameters k_1 and b, respectively.

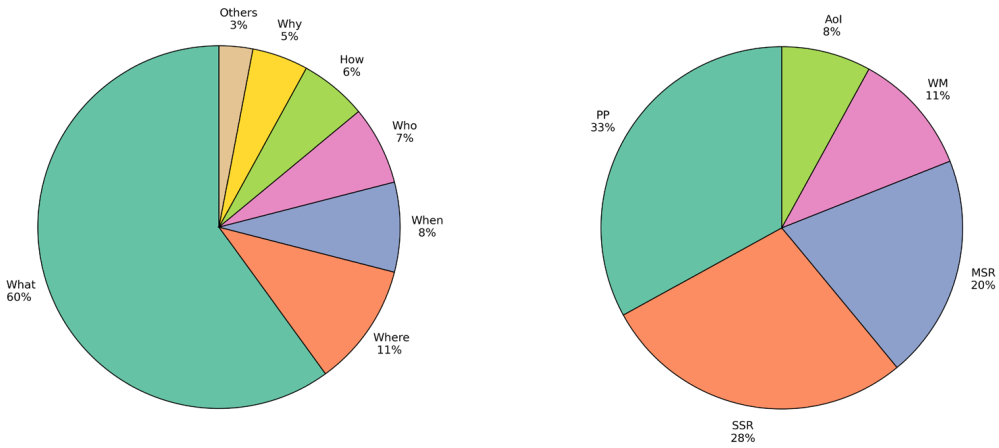
Using BM25, we generated a pre-indexed set of documents for each of the 2999 queries in our dataset. This involves selecting the top 1000 documents based on their BM25 relevance scores. These pre-selected documents constituted the test set for our subsequent evaluation with ColBERT.

Among the 1000 documents, the gold-label ranking for the relevant document d+ in each query is set to 1. The remaining 999 documents are considered low-relevance documents d−. This configuration aims to assess the model's ability to rank the most relevant document as the top result. The statistical information of the dataset used for the re-ranking performance evaluation experiment is shown in Table 2.

Table 2. ColBERT top-k re-ranking evaluation dataset statistics information.

Divisions	Min, Max Query Length	Avg Query Length	Avg Document Length	Queries	k
20	5~20	16.7	601.6	2999	1000
50	5~50	28.9	608.9	2999	1000
200	5~124	31.3	609.7	2999	1000

We analyzed our constructed dataset from the perspectives of question types and reasoning types. The distribution of these question and reasoning types is visualized in Figure 4. We manually annotated the question and reasoning types for a randomly sampled 100 questions. We adopted the classification scheme of CMRC [34], categorizing them into seven types: Who, What, When, Where, Why, How, and Others. The 'What' type constitutes approximately 60%, which is similar to the 54% for 'What' questions in SQuAD [35]. For reasoning types, following the criteria set by Hill et al. [36] and Nguyen et al. [37], the questions were annotated into five different categories: word matching (WM), paraphrasing (PP), single-sentence reasoning (SSR), multi-sentence reasoning (MSR), and ambiguous/insufficient (AoI). The proportion of reasoning types in our dataset stands at 48%, which is higher than the 21% in SQuAD and 34% in NewsQA [38]. This highlights that our dataset is more challenging compared to SQuAD and NewsQA.



(a) Question Type **(b) Reasoning Type**

Figure 4. Proportions of question and reasoning types from a randomly sampled subset of 100 examples in the evaluation dataset.

5. Experimental Setup and Results

5.1. Datasets

In this paper, we evaluated the top-k re-ranking performance of Korean ColBERT using the evaluation dataset constructed in Section 4.2. To assess the performance based on different query lengths, training was carried out based on those lengths, and the evaluation was then conducted using the corresponding datasets defined by the maximum query lengths. Detailed statistics of the dataset used for the evaluation can be found in Table 2.

5.2. Baselines

In this study, with the aim of applying ColBERT to Korean, we used the traditional IR model BM25 and the dense retrieval model KoBERT as baseline models for comparison. BM25 is a traditional statistics-based information retrieval (IR) model that calculates relevance scores by computing the word weights between the query and documents. KoBERT is a dense retrieval-based model that calculates relevance scores through the [CLS] token embedding obtained by inputting query–document pairs into KoBERT. The proposed ColBERT with KoBERT (ColBERT w/KoBERT) is a model that calculates the relevance scores between the query and documents through the late interaction.

5.3. Evaluation Metrics

$$\text{MRR} = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{\text{Rank}_i} \tag{3}$$

In this paper, we used MRR (mean reciprocal rank) and recall to measure the search performance of the open-domain question answering system and latency to measure the speed. MRR is the average reciprocal rank of the most relevant document for a query Q. The reciprocal rank is the inverse of the rank of the most relevant document in the re-ranked document list for query Q. MRR@10 and MRR@100 represent the average reciprocal rank targeting only the top 10 and 100 documents, respectively. If there is no relevant document for the query in the target documents, the reciprocal rank for that question is calculated as 0. MRR@10 and MRR@100 allow us to measure how accurately the model places the most relevant documents at the top for a query. Recall@k is the number of relevant documents in the top k results divided by the total number of relevant documents. Recall@50 and Recall@200 represent the number of relevant documents included in the top 50 and 200 results, respectively, divided by the total number of relevant documents among 1000 documents. Recall@50 and Recall@200 allow us to measure how many relevant documents the model can find in the top search results. Latency is an indicator to measure the search speed of the open-domain question answering system, representing the average delay time that occurs during the re-ranking process of the query-related documents. We can evaluate the overall performance and efficiency of the search system by measuring the search accuracy through MRR and recall and the search speed through latency.

5.4. Computation Details

As mentioned before, when calculating the similarity between the bag of contextualized embeddings for both query and document, the cosine similarity or squared L2 distance can be used for computation. In this experiment, the cosine similarity was used as the similarity measure. However, the computation was implemented as a dot product. This is due to embedding normalization.

In the top-k re-ranking batch computation, the batch size was set to 128. Given that k = 1000, for each query, 1000 documents were divided into a total of 8 mini-batches for computation. By partitioning the calculations across multiple mini-batches, efficient GPU memory utilization was achieved while maintaining a rapid computational speed.

Across all experiments, we used a single A100 GPU that has 40 GB of memory on a server with two Intel Xeon Gold 6226R CPUs, each with 16 physical cores, and 495 GB of RAM.

5.5. Results

We evaluated the efficiency of ColBERT by training it using the triples dataset <q, d+, d−>, as described in Section 4.1, and then testing its re-ranking performance using the test dataset from Section 4.2. The dataset for re-ranking was constructed by extracting 1000 documents for each query using the BM25 algorithm. While BM25, a term matching-based document retrieval method, may not be as accurate as neural approaches, it effectively filters documents with at least some relevance to the query. Re-ranking these documents demands not just retrieving relevant documents but also precisely identifying key documents that directly answer the query. Therefore, through our re-ranking experiment, we aimed to demonstrate the precision and efficiency of our proposed method by measuring MRR and latency.

Table 3 shows the performance of re-ranking documents based on their relevance score to each question in a set of k (k = 1000) ranked documents, with the performance varied depending on the maximum length of the query. The xperimental results show that KoBERT-based ColBERT (ColBERT w/KoBERT) achieves about 1.8 times, 1.1 times, and 1.1 times higher performance in MRR@10 compared to the traditional IR model BM25 when the question length is 20, 50, and 200, respectively. Additionally, the search speed is improved by more than 3.5 times. Although ColBERT w/KoBERT has a slightly lower MRR@10 performance by about 0.1 times than the dense retrieval KoBERT for question lengths of 50 and 200, it shows a search speed that is more than ten times faster. For question lengths of 20 or less, it demonstrates about 2.1 times better performance in MRR@10 and more than ten times faster search speed. These experimental results confirm that, while baselines may exhibit slightly better performance due to their ability to compute more complex interactions between the query and the document, ColBERT w/KoBERT is an efficient model that still exhibits excellent search performance while maintaining a fast search speed.

The experimental results indicate that the performance gap in MRR between dense retrieval and the Korean ColBERT is larger compared to the original ColBERT paper [10]. This could be attributed to the elevated level of reasoning required by the dataset we have constructed. As can be seen in Figure 4, queries involving single-sentence reasoning (SSR) and multi-sentence reasoning (MSR) inference types make up nearly 48% of our dataset, almost half. Unlike other inference types, these categories do not have direct evidence for the answers within the document, requiring a deeper contextual understanding to find the correct answers. This might explain why dense retrieval exhibits superior performance in our experiments. Nevertheless, ColBERT still demonstrates markedly higher performance than BM25 and offers a search speed that is ten times faster than dense retrieval.

Table 4 compares the re-ranking performance when using various pre-trained Korean language models other than KoBERT to generate question (question length 200) and document embeddings in ColBERT. We selected PLMs that performed well on the STS (semantic textual similarity) task. The experimental results show that the Korean-adapted Sentence-BERT (SBERT) [39], KoSBERT [40], has the best performance. ColBERT with KoSBERT (ColBERT w/KoSBERT) shows the best performance among ColBERT models utilizing pre-trained Korean language models. Although the MRR@10 performance of ColBERT w/KoSBERT is slightly lower by about 0.1 times compared to the dense retrieval KoSBERT, the search speed is improved by 23 times. The experiments confirm that using sentence-level embeddings like SBERT, instead of token-level embeddings like BERT, for generating question and document embeddings in ColBERT results in more accurate and faster search performance.

Table 3. Comparison of re-ranking performance by query length.

Method	MRR@10	MRR@100	Recall@50	Recall@200	Latency (ms)
Max Query Length			**20**		
BM25	23.18	23.91	51.81	64.48	5586
KoBERT	20.05	21.09	55.01	70.29	16,328
ColBERT w/KoBERT (Ours)	**41.15**	**41.79**	**69.22**	**74.99**	**1581**
Max Query Length			**50**		
BM25	51.34	51.9	79.52	85.66	5869
KoBERT	**65.05**	**65.46**	**86.89**	**89.76**	16,465
ColBERT w/ KoBERT (Ours)	58.95	59.46	84.39	89.32	**1538**
Max Query Length			**200**		
BM25	55.25	55.77	81.19	86.96	5883
KoBERT	**68.09**	**68.38**	**88.49**	**90.76**	16,466
ColBERT w/KoBERT (Ours)	62.55	62.98	86.29	89.99	**1578**

Table 4. Comparison of re-ranking performance using various pre-trained Korean language models.

Method	MRR@10	MRR@100	Recall@50	Recall@200	Latency (ms)
			Baseline		
BM25	55.25	55.77	81.19	86.96	**5883**
KoBERT	68.09	68.38	88.49	90.76	16,466
KoSBERT	**73.71**	**73.92**	**90.63**	**91.39**	20,411
			Ours		
ColBERT w/KoBERT	62.55	62.98	86.29	89.99	1578
ColBERT w/KLUE-RoBERTa	66.04	66.40	86.66	90.26	886
ColBERT w/KoBigBird	67.29	68.07	87.16	90.50	1240
ColBERT w/KLUE-BERT	67.91	68.24	86.99	**90.79**	1322
ColBERT w/KoSBERT	**68.15**	**68.48**	**87.26**	90.63	**873**

These experimental results demonstrate the effectiveness of SBERT in ODQA tasks. SBERT is trained to discern the semantic similarity between two sentences by inputting each sentence into BERT separately and using the obtained representations. Therefore, SBERT's representations may be more suitable for similarity-based tasks compared to those of standard BERT. This is evidenced by KoSBERT exhibiting superior performance over KLUE-BERT. These findings can be instrumental in guiding the selection of pre-trained language models for future ODQA tasks.

Table 5 presents additional comparative experimental results examining the re-ranking performance of KoSBERT for each task. In [39], BERT's sentence embeddings are improved by mapping semantically similar sentences close together in the vector space through the natural language inference (NLI) task, which discriminates between implication, contradiction, and neutrality by calculating embeddings for two sentences separately, and the semantic textual similarity (STS) task, which predicts similarity scores between 0 and 5 using the cosine similarity of two sentence embeddings. KoSBERT is a pre-trained Korean language model trained using the SBERT structure proposed in [39] on the KLUE-BERT model. KoSBERT-STS is a model fine-tuned with the KorSTS dataset [41], KoSBERT-NLI is a model fine-tuned with the KorNLI dataset [42], and KoSBERT-Multi is a model fine-tuned with KorNLI and further fine-tuned with KorSTS. The experiments show that the best performance is achieved when using KoSBERT-NLI among the various KoSBERT models. This implies that the model trained on inferring semantic relationships between sentences has a significant impact on exhibiting good search performance.

Table 5. Comparison of re-ranking performance by KoSBERT for different tasks.

Method	MRR@10	MRR@100	Recall@50	Recall@200	Latency (ms)
ColBERT w/KoSBERT-STS	67.51	67.85	**87.29**	**90.66**	1618
ColBERT w/KoSBERT-NLI	**68.15**	**68.48**	87.26	90.63	**873**
ColBERT w/KoSBERT-Multi	67.95	68.28	87.16	90.63	1368

5.6. Ablation Study

In our experiment, we tokenized the queries and the entire document set using the Kobert tokenizer, then extracted 1000 documents per query via the BM25 algorithm to evaluate the re-ranking performance of ColBERT. However, unlike English, Korean is an agglutinative language, where grammatical functions are determined by appending various suffixes or particles to words or phrases. Due to this characteristic, it can be more effective to process words by first separating them into morphemes, the smallest units bearing meaning, rather than using the words as they are. Hence, we utilized the Korean morpheme analyzer Mecab [43] to tokenize the queries and the entire document set at the morpheme level, then re-applied the BM25 to extract 1000 documents per query, conducting an additional experiment.

Table 6 compares the performance of re-ranking when using documents extracted by the BM25 after tokenizing with the Kobert tokenizer and when using documents extracted by the BM25 after tokenizing at the morpheme level. The experimental results demonstrate that re-ranking using documents extracted by the BM25 after tokenizing at the morpheme level yielded a higher performance across all metrics. This highlights the significance of morphological analysis when dealing with agglutinative languages like Korean.

Table 6. Comparison of re-ranking performance for different tokenization methods on top 1000 documents extracted by BM25.

Method	MRR@10	MRR@100	Recall@50	Recall@200	Latency (ms)
Re-ranking 1000 docs extracted by BM25 (KoBERT tokenization)					
ColBERT w/KoBERT	62.55	62.98	86.29	89.99	1578

Table 6. *Cont.*

Method	MRR@10	MRR@100	Recall@50	Recall@200	Latency (ms)
ColBERT w/KLUE-RoBERTa	66.04	66.40	86.66	90.26	886
ColBERT w/KoBigBird	67.29	68.07	87.16	90.50	1240
ColBERT w/KLUE-BERT	67.91	68.24	86.99	**90.79**	1322
ColBERT w/KoSBERT	**68.15**	**68.48**	**87.26**	90.63	**873**
Re-ranking 1000 docs extracted by BM25 (morpheme-based tokenization)					
ColBERT w/KoBERT	63.46	63.94	88.06	92.13	2188
ColBERT w/KLUE-RoBERTa	66.80	67.18	88.72	92.43	1286
ColBERT w/KoBigBird	68.72	69.03	89.16	92.46	1858
ColBERT w/KLUE-BERT	68.82	69.17	89.23	**92.66**	1286
ColBERT w/KoSBERT	**69.19**	**69.56**	**89.26**	92.59	**1263**

6. Implications

The results of this study have practical implications applicable to Korean open-domain question answering systems. First, our research confirms the effectiveness of the ColBERT architecture in Korean, an agglutinative language. This architecture independently processes embeddings for both queries and passages and employs a late interaction mechanism for efficient semantic correlation in document retrieval. Furthermore, the study reveals that SBERT, when employed within the ColBERT framework, outperforms traditional BERT models in retrieval tasks. This improvement is realized by using various open-source Korean BERT-based pre-trained language models. These findings are crucial in aiding the selection of PLMs for developing representations of queries and passages in ODQA research, especially with Transformer-based pre-trained language models. The experimental outcomes suggest that the adoption of alternative PLMs to traditional BERT models can lead to improved retrieval performance.

7. Conclusions

In this paper, we propose an efficient Korean document retrieval method for open-domain question answering using the ColBERT structure. We construct a Korean dataset for open-domain question answering and evaluate the performance of the proposed method on this dataset. The experimental results show that our proposed KoBERT-based ColBERT achieves about 13% higher performance in MRR@10 compared to the traditional term-based retrieval model BM25 and reduces the search speed by approximately 3.7 times. KoBERT-based ColBERT shows about 8% lower performance in MRR@10 compared to dense retrieval KoBERT, but the search speed is shortened by more than ten times. Furthermore, when utilizing the KoSBERT-based ColBERT method, the sentence-embedding ability for both questions and documents is improved with the use of KoSBERT, achieving the best search performance and the shortest search time.

In this study, the ColBERT method is not only applicable to open-domain question answering tasks but also holds potential for situations requiring both search accuracy and speed simultaneously, such as search engines and recommendation systems. Additionally,

given that search performance is improved when utilizing sentence embeddings trained to capture semantic similarity between sentences, future research will investigate approaches to enhance embeddings for effectively capturing semantic similarity between query and document, with the goal of further improving search performance.

Author Contributions: Conceptualization, B.K., Y.K. and Y.S.; methodology, B.K. and Y.K.; software, B.K. and Y.K.; validation, B.K. and Y.K.; formal analysis, B.K. and Y.K.; investigation, B.K. and Y.K.; resources, Y.S.; data curation, B.K. and Y.K.; writing—original draft preparation, B.K. and Y.K.; writing—review and editing, Y.S.; visualization, B.K.; supervision, Y.S.; project administration, 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 a National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (No. 2021R1G1A1012766).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data and related information used in this study can be found at the https://github.com/movie112/ColBERTforKorean.

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

Abbreviations

The following abbreviations are used in this manuscript:

BERT	Bidirectional Encoder Representations from Transformers
ColBERT	Contextualized Late Interaction over BERT
IR	Information Retrieval
TF-IDF	Term Frequency–Inverse Document Frequency
MRR	Mean Reciprocal Rank
SBERT	Sentence-BERT
NLI	Natural Language Inference
STS	Semantic Textual Similarity
KLUE	Korean Language Understanding Evaluation
RoBERTa	A Robustly Optimized BERT Pre-training Approach

References

1. Chen, D.; Fisch, A.; Weston, J.; Bordes, A. Reading Wikipedia to Answer Open-Domain Questions. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL), Vancouver, BC, Canada, 30 July–4 August 2017.
2. Mao, Y.; He, P.; Liu, X.; Shen, Y.; Gao, J.; Han, J.; Chen, W. Generation-augmented retrieval for open-domain question answering. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, (Volume 1: Long Papers), Virtual Event, 1–6 August 2021.
3. Karpukhin, V.; Oğuz, B.; Min, S.; Lewis, P.; Wu, L.; Edunov, S.; Chen, D.; Yih, W.-T. Dense passage retrieval for open-domain question answering. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Online Event, 16–20 November 2020; pp. 6769–6781.
4. Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; Liu, P.J. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv* **2019**, arXiv:1910.10683.
5. Khattab, O.; Zaharia, M. ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT. In Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR'20, Virtual Event, 25–30 July 2020; Association for Computing Machinery: New York, NY, USA, 2020; pp. 39–48.
6. Zhang, Q.; Chen, S.; Xu, D.; Cao, Q.; Chen, X.; Cohn, T.; Fang, M. A Survey for Efficient Open Domain Question Answering. 2022. Available online: https://livrepository.liverpool.ac.uk/3170624/ (accessed on 20 November 2023).
7. Wang, Y.; Ma, H.; Wang, D. Z. LIDER: An Efficient High-dimensional Learned Index for Large-scale Dense Passage Retrieval. *arXiv* **2022**, arXiv:2205.00970.
8. Neyshabur, B.; Srebro, N. On symmetric and asymmetric lshs for inner product search. In Proceedings of the 32nd International Conference on International Conference on Machine Learning—Volume 37, ICML'15, Lille, France, 6–11 July 2015; pp. 1926–1934.
9. Sivic, J.; Zisserman, A. Video Google: A Text Retrieval Approach to Object Matching in Videos. In Proceedings of the Ninth IEEE International Conference on Computer Vision, Washington, DC, USA, 13–16 October 2003; IEEE: New York, NY, USA, 2003; Volume 2, pp. 1470–1477.

10. Zhang, Q.; Chen, S.; Xu, D.; Cao, Q.; Chen, X.; Cohn, T.; Fang, M. A Survey for Efficient Open Domain Question Answering. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Toronto, BC, Canada, 9–14 July 2023; Rogers, A., Boyd-Graber, J., Okazaki, N., Eds.; pp. 14447–14465.
11. Vu, D.T.; Yu, G.; Lee, C.; Kim, J. Text Data Augmentation for the Korean Language. *Appl. Sci.* **2022**, *12*, 3425. [CrossRef]
12. Nogueira, R.; Cho, K. Passage Re-ranking with BERT. *arXiv* **2019**, arXiv:1901.04085.
13. Nie, Y.; Wang, S.; Bansal, M. Revealing the Importance of Semantic Retrieval for Machine Reading at Scale. 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 November 2019; Inui, K., Jiang, J., Ng, V., Wan, X., Eds.; pp. 2553–2566.
14. Zhu, F.; Lei, W.; Wang, C.; Zheng, J.; Poria, S.; Chua, T.-S. Retrieving and Reading: A Comprehensive Survey on Open-domain Question Answering. *arXiv* **2021**, arXiv:2101.00774.
15. Zhan, J.; Mao, J.; Liu, Y.; Guo, J.; Zhang, M.; Ma, S. Jointly Optimizing Query Encoder and Product Quantization to Improve Retrieval Performance. In Proceedings of the 30th ACM International Conference on Information & Knowledge Management, CIKM'21, Virtual Event, 1–5 November 2021; Association for Computing Machinery: New York, NY, USA, 2021; pp. 2487–2496.
16. Malkov, Y.A.; Yashunin, D.A. Efficient and Robust Approximate Nearest Neighbor Search Using Hierarchical Navigable Small World Graphs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 824–836. [CrossRef] [PubMed]
17. Johnson, J.; Douze, M.; Jégou, H. Billion-Scale Similarity Search with GPUs. *IEEE Trans. Big Data* **2021**, *7*, 535–547. [CrossRef]
18. Lee, J.; Sung, M.; Kang, J.; Chen, D. Learning Dense Representations of Phrases at Scale. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Virtual Event, 1–6 August 2021; Zong, C., Xia, F., Li, W., Navigli, R., Eds.; pp. 6634–6647.
19. Roberts, A.; Raffel, C.; Shazeer, N. How Much Knowledge Can You Pack Into the Parameters of a Language Model? In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Virtual, 16–20 November 2020; Webber, B., Cohn, T., He, Y., Liu, Y., Eds.; pp. 5418–5426.
20. Lee, J.; Wettig, A.; Chen, D. Phrase Retrieval Learns Passage Retrieval, Too. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Virtual Event, 7–11 November 2021; Moens, M.-F., Huang, X., Specia, L., Yih, S.W.-T., Eds.; pp. 3661–3672.
21. Lewis, P.; Wu, Y.; Liu, L.; Minervini, P.; Küttler, H.; Piktus, A.; Stenetorp, P.; Riedel, S. PAQ: 65 Million Probably-Asked Questions and What You Can Do With Them. *Trans. Assoc. Comput. Linguist.*, **2021**, *9*, 1098–1115. [CrossRef]
22. Bilal, M.; Almazroi, A.A. Effectiveness of Fine-tuned BERT Model in Classification of Helpful and Unhelpful Online Customer Reviews. *Electron. Commer. Res.* **2023**, *23*, 2737–2757. [CrossRef]
23. Park, S.; Moon, J.; Kim, S.; Cho, W.I.; Han, J.Y.; Park, J.; Song, C.; Kim, J.; Song, Y.; Oh, T.; et al. KLUE: Korean Language Understanding Evaluation. In *Neural Information Processing Systems Track on Datasets and Benchmarks*; Volume 1; Vanschoren, J., Yeung, S., Eds.; Curran: Nice, France, 2021. Available online: https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/file/98dce83da57b0395e163467c9dae521b-Paper-round2.pdf (accessed on 20 November 2023).
24. Hwang, S.; Kim, D. BERT-based Classification Model for Korean Documents. *J. Soc. e-Bus. Stud.* **2020**, *25*, 203-214.
25. Lee, J. KcBERT: Korean Comments BERT. In Proceedings of the Annual Conference on Human and Language Technology, Kaunas, Lithuania, 22–23 September 2020; Human and Language Technology: Kaunas, Lithuania, 2020; pp. 437–440.
26. Kim, Y.; Kim, J.H.; Lee, J.M.; Jang, M.J.; Yum, Y.J.; Kim, S.; Song, S. A Pre-trained BERT for Korean Medical Natural Language Processing. *Sci. Rep.* **2022**, *12*, 13847. [CrossRef] [PubMed]
27. Available online: https://github.com/monologg/KoBERT-Transformers/blob/master/kobert_transformers/tokenization_kobert.py (accessed on 14 April 2023).
28. AI-Hub. Available online: https://aihub.or.kr/ (accessed on 14 April 2023).
29. Available online: https://aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=realm&dataSetSn=106 (accessed on 14 April 2023).
30. Available online: https://aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=realm&dataSetSn=89 (accessed on 14 April 2023).
31. Available online: https://aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=realm&dataSetSn=92 (accessed on 14 April 2023).
32. Available online: https://aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=realm&dataSetSn=624 (accessed on 14 April 2023).
33. Nguyen, T.; Rosenberg, M.; Song, X.; Gao, J.; Tiwary, S.; Majumder, R.; Deng, L. MS MARCO: A Human Generated MAchine Reading COmprehension Dataset. *Choice* **2016**, *2640*, 660.
34. Cui, Y.; Liu, T.; Che, W.; Xiao, L.; Chen, Z.; Ma, W.; Wang, S.; Hu, G. A Span-Extraction Dataset for Chinese Machine Reading Comprehension. *arXiv* **2019**, arXiv:1810.07366.
35. Aniol, A.; Pietron, M.; Duda, J. Ensemble approach for natural language question answering problem. In Proceedings of the 2019 Seventh International Symposium on Computing and Networking Workshops (CANDARW), Nagasaki, Japan, 26–29 November 2019.
36. Hill, F.; Bordes, A.; Chopra, S.; Weston, J. The Goldilocks Principle: Reading Children's Books with Explicit Memory Representations. *arXiv* **2015**, arXiv:1511.02301.

37. Van Nguyen, K.; Tran, K.V.; Luu, S.T.; Nguyen, A.G.T.; Nguyen, N.L.T. Enhancing lexical-based approach with external knowledge for Vietnamese multiple-choice machine reading comprehension. *IEEE Access* **2020**, *8*, 201404–201417. [CrossRef]
38. Trischler, A.; Wang, T.; Yuan, X.; Harris, J.; Sordoni, A.; Bachman, P.; Suleman, K. NewsQA: A Machine Comprehension Dataset. *arXiv* **2016**, arXiv:1611.09830.
39. Reimers, N.; Gurevych, I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In Proceedings of the EMNLP/IJCNLP, Hong Kong, China, 3–7 November 2019; Inui, K., Jiang, J., Ng, V., Wan, X., Eds.; pp. 3980–3990.
40. Available online: https://huggingface.co/jhgan/ko-sbert-nli (accessed on 14 April 2023).
41. Available online: https://github.com/kakaobrain/kor-nlu-datasets/tree/master/KorSTS (accessed on 14 April 2023).
42. Available online: https://github.com/kakaobrain/kor-nlu-datasets/tree/master/KorNLI (accessed on 14 April 2023).
43. Available online: https://konlpy.org (accessed on 6 July 2023).