Transfer Learning for Closed Domain Question Answering in COVID-19

Nur Rachmawati, Evi Yulianti
Faculty of Computer Science, Universitas Indonesia, Depok, Indonesia

***Abstract*—COVID-19 has been a popular issue around 2019 until today. Recently, there has been a lot of research being conducted to utilize a big amount of data discussing about COVID-19. In this work, we conduct a closed domain question answering (CDQA) task in COVID-19 using transfer learning technique. The transfer learning technique is adopted because a large benchmark for question answering about COVID-19 is still unavailable. Therefore, rich knowledge learned from a large benchmark of open domain QA are utilized using transfer learning to improve the performance of our CDQA system. We use retriever-reader framework for our CDQA system, and propose to use Sequential Dependence Model (SDM) as our retriever component to enhance the effectiveness of the system. Our result shows that the use of SDM retriever can improve the F-1 score of the state-of-the-art baseline CDQA system using BM25 and TF-IDF+cosine similarity retriever by 3,26% and 32,62%, respectively. The optimal parameter settings for our CDQA system are found to be as follows: using 20 top-ranked documents as the retriever's output, five sentences as the passage length, and BERT-Large-Uncased model as the reader. In this optimal parameter setting, SDM retriever can improve the F-1 score of the state-of-the-art baseline CDQA system using BM25 by 5,06 % and TF-IDF+cosine similarity retriever by 24,94 %. Our last experiment then confirms the merit of using transfer learning, since our best-performing model (double fine-tune SQuAD and COVID-QA) is shown to gain eight times higher accuracy than the baseline method without using transfer learning. Further fine-tuning the transfer learning model using closed domain dataset (COVID-QA) can increase the accuracy of the transfer learning model that only fine-tuning with SQuAD by 27, 26%.**

***Keywords*—COVID-19; closed domain question answering; sequential dependence model; transfer learning; BERT**

I. Introduction

Corona virus disease 2019, known as COVID-19, has been a popular topic in recent period, from 2019 until today. COVID-19 can cause severe respiratory disease in humans [1]. This disease initially came from Wuhan city in China, and was detected for the first time on December 29, 2019 [2]–[4].

In this COVID-19 pandemic, obtaining information about COVID-19 is really necessary. A specific question answering (QA) system on COVID-19, referred to as closed domain question answering (CDQA) system, is beneficial to provide direct answers to many questions that may arise related with COVID-19 issue. This system will be useful for health practitioners as well as public in general to have deeper knowledge on COVID-19. This motivates us to implement a question answering system on specific COVID-19 topic in this study.

To build an effective CDQA system in COVID-19 using state-of-the-art deep learning approach, a large benchmark consisting of pairs of questions and answers together with a huge collection of documents about COVID-19 issue, is needed. A previous research has built CORD-19 dataset which includes a collection of one million articles about COVID-19 [5]. However, a large number of pairs of questions and answers about COVID-19 from that data are still unavailable to train a deep learning-based QA model. A previous work has paid some attention to this matter and put some effort to build 2019 question-answer pairs in COVID-QA dataset [6]. This number, however, is still relatively low to learn a deep learning-based QA model from scratch.

To tackle the above problem, we adopt transfer learning technique to build more effective CDQA system in COVID-19. Transfer learning is commonly used to make use of knowledge extracted from high-resource data to solve the task on the low-resource data. It becomes our intuition to use transfer learning for our case of low resource COVID-19 QA benchmark. Transfer learning approach enables us to exploit rich knowledge learned from an available large benchmark of open domain QA system, i.e., SQuAD dataset [7] to extract answers about COVID-9 topic from low-resource closed domain COVID-19 dataset, i.e., CORD-19 [5].

Some research has used transfer learning method to build CDQA system [8] [9]. Alzubi et al. [8] and Yang et al. [9] both implemented CDQA system using dual architecture (retriever-reader system) and transfer learning principle. Here, retriever component will find subset of relevant documents, and reader component trained on high-resource data will find the span of text from those documents as the answer to the given question. In this work, we aim to improve the accuracy of CDQA system of Alzubi et al. [8] and Yang et al. [9] by enhancing the retriever component of CDQA system. By improving the accuracy of retriever, it will result in the increasing the accuracy of the reader in extracting answers. While Alzubi et al. [8] and Yang et al. [9] used TF-IDF vectorizer with cosine similarity, and BM25, respectively, as their retriever, we propose to use Sequential Dependence Model [10] as the retriever component in the CDQA system. Here, the retriever model of Alzubi et al. [8] and Yang et al. [9] will be changed to the CDQA systems. Our intuition in using SDM is because it optimizes the scoring functions of documents by modelling question term dependencies, which therefore can capture different variations of question terms in the documents. While many previous work has demonstrated the effectiveness of this

model for an ad-hoc retrieval [10]–[12], to the best of our knowledge, none of previous studies have investigated the use of SDM for QA system, more specifically dual architecture CDQA system. This becomes a research gap that is fulfilled by this work.

In this paper, there is some following research question (RQ) that has been investigated:

RQ1: Does the use of SDM as retriever in the retriever-reader architecture improve the performance of state-of-the-art transfer-learning-based CDQA systems in COVID-19 using TF-IDF+cosine and BM25 retriever?

RQ2: What are the effects of tuning some parameters (top n retrieved documents, passage length, and reader variations) on the accuracy of CDQA system in COVID-19? What are the optimal parameter settings for our system?

RQ3: To what extent the use of transfer learning technique using SDM as retriever in reader-retriever architecture can improve the CDQA system in COVID-19 that does not use transfer learning technique?

Our contribution in this work are as follows: (1) we propose to use Sequential Dependence Model as retriever in retriever-reader architecture of CDQA system in COVID-19; (2) we conduct empirical evaluation on the effectiveness of our proposed method compared to the state-of-the-art CDQA method used by Alzubi et al. [6] and Yang et al. [9]; we also perform evaluation to show the merit of using transfer learning technique by comparing our method with the method that does not use transfer learning.

The rest of this paper is organized as follows. Section II describes about the related works. Section III explains our research methodology such as the dual architecture of our CDQA system, dataset, and baseline methods. Section IV and V presents our experiment details and results. Section VI discusses about the limitation and the challenges of this work. Finally, Section VII concludes this study and answers all research questions outlined above.

II. Related Work

A. Closed Domain Question Answering (CDQA)

Question Answering is a system that is used to answer the question given by users. The input to the system is a question and a list of documents. The system tries to find the answer by finding the start/end positions where the answer is located within the text. Question answering can be divided into two parts: closed domain and open domain. Closed domain question answering (CDQA) is a question answering system that has ability in answering questions regarding the specific domain by exploiting mainly domain-specific knowledge [8]. An example of this closed domain question answering is COBERT implemented by Alzubi et al. [8]. They used a specific domain of COVID-19 when building such CDQA system. Meanwhile, open domain question answering is a question answering system that has an ability to answer questions about any domains and rely on general ontology and world knowledge [10]. In open domain question answering, they often use Wikipedia as the unique knowledge source when looking for answers [13]. An example of this open domain question answering is DrQa [13]. In this paper, our work focus is on building CDQA system in COVID-19 domain.

B. Machine Reading Task for Question Answering

Machine reading task for question answering has been accelerated in recent years. This technique is about how machines can read and learn from articles that have been given to the QA system. Some dataset has been built for this machine reading text such as SQuAD [7], WikiQA [14], CoQA [15], QACNN [16], etc.

Some previous researches have paid attention about this topic. DrQA has been built for open domain question answering using a retriever-reader architecture. The retriever is a module using bigram hashing and TF-IDF matching that will retrieve relevant articles from Wikipedia. Then, the reader is RNN model that detects answer spans from the relevant documents. In general, DrQA pipeline combines bigram hashing with TF-IDF matching retriever and bidirectional RNN paragraph reader [13].

Some recent methods have utilized BERT for question answering, as a result of the effectiveness of BERT that has been shown in various text processing tasks in previous work [17], [18]. Alzubi et al. [8] proposed COBERT, COVID-19 Question Answering System Using BERT. The system uses a retriever-reader architecture and CORD-19 dataset [5] as the input. They used TF-IDF vectorizer with cosine similarity, implemented using scikit-learn [19] tool, to get the top *N* most relevant documents. These documents are then split into passages to be inputted into the reader. The reader used BERT model that was fine-tuned with SQuAD dataset [7] to identify the answer spans based on the final score. These answers were then ranked based on this score.

Yang et al. [9] proposed end-to-end open domain question answering with BERTserini, a combining BERT with Anserini toolkit. With the Wikipedia article and SQuAD dataset [7] as the input, BERTserini tried to use dual architecture (retriever-reader). In retriever, they used Anserini IR toolkit with BM25 [20]. The retriever will retrieve a set of documents and produce a set of text segments (passages) to be inputted into the reader. They used BERT reader that has been fine-tuned with SQuAD dataset [7] to identify the answer span. The system was shown to outperform the DrQA system [13]. Nogueira et al. [21] also adopted BM25 and BERT model to perform passage / answer retrieval.

Semnani et al. [22] proposed Mindstone, a domain-specific question answering system using Wikipedia and Snowflake text corpora. In the Mindstone pipeline, there are three components: retriever, ranker, and reader. In retriever, they used Anserini based on Lucene version 8.0, and Okapi BM25. Then, Neural RM3 (Relevance based language models) is used as ranker to expand the query and re-rank the documents based on the new score given by the ranker. For reader, they used BERT-base model to rank the retrieved answers. The system outperformed the BERTserini system in terms of EM (Exact Match) dan F1-scores.

In this paper, we use retriever-reader architecture in machine reading task for question answering, similar to Alzubi et al. [8] and Yang et al. [9]. However, we use different

retriever using Sequential Dependence Model (SDM) to improve the performance of CDQA system.

C. Transfer Learning

Transfer learning is a method using machine learning / deep learning that is commonly used to improve the performance on low-resource tasks. Transfer learning describes the learning schemes when information in source task is used to achieve some improvement in target task performance [23].

Some previous researches on question answering system used this method. Akdemir et al. [24] proposed transfer learning for Biomedical Question Answering. They used this method because available datasets in biomedical question answering are limited. Therefore, with transfer learning schema, they want to transfer information learned in high-resource tasks which is a similar domain with the source task, into a low-resource target task. They hope by applying transfer learning can improve the performance on their question answering system. Other research that used transfer learning method is Syed et al. [25]. They used transfer learning by using SQuAD dataset [7] to fine-tune the question answering system. Alzubi et al. [8] also used transfer learning method to build closed domain question answering system in COVID-19. They used transfer learning because available dataset in COVID-19 question answering is still limited. In this paper, transfer learning method will be used too. It is the same approach with Alzubi et al. [8] and Yang et al. by utilizing SQuAD [7] dataset to fine-tune the reader component of our CDQA system. However, we propose to replace the retriever component of Alzubi et al. and Yang et al. with SDM (Sequential Dependence Model) that has been shown to achieve satisfactory results in previous work on document ranking [10]–[12].

III. Methodology

A. System Architecture

In this work, we use dual architecture (retriever-reader) of question answering system similar to Alzubi et al. [8] and Yang et al. [9]. The difference is that the architecture does not use TF-IDF vectorizer or BM25 as retriever. We implement SDM (Sequential Dependence Model) as retriever using PyTerrier tool [26]. Fig. 1 illustrates the architecture of our system. Initially, the collection of COVID-19 articles, CORD-19, are preprocessed, and then indexed using PyTerrier. The retriever is implemented using SDM and the reader is implemented using BERT model that was fine-tuned using SQuAD dataset. The flow of the process in our system is as follows: Given a question input, the SDM retriever will retrieve a top-N relevant documents that further will be split into passages. These passages will be inputted into the reader that will extract the answers from the passages and rank them to generate the ranked list of answers.

1) Retriever system: Sequential dependence model (SDM) was proposed by Metzler et al. [10] to optimize the document scoring function by including proximity in the query based on the occurrences of single terms, ordered phrases, and unordered phrases. The SDM method has an assumption that all pairs of sequential terms extracted from the query are dependent [27]. It assumes that the occurrences of adjacent query terms are related.

$$score_{SDM}(Q,D) = \lambda_T \sum_{qCQ}^{|Q|-1} f_T(q,D) + \lambda_O \sum_{i=1}^{|Q|-1} f_O(q_i, q_i + 1, D) + \lambda_U \sum_{i=1}^{|Q|-1} f_U(q_i, q_i + 1, D) \quad (1)$$

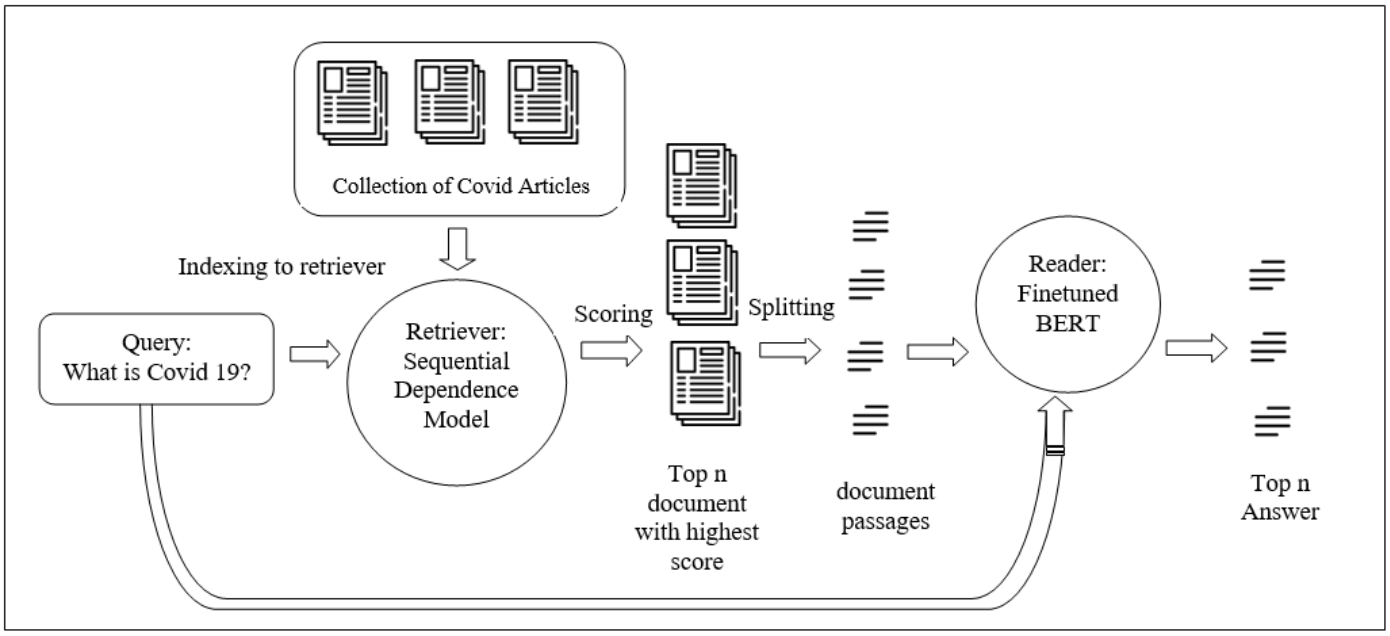

Fig. 1. Architecture System.

The SDM scoring function is described in Eq 1. Given a query Q, the function will calculate the SDM score for each document D in the collection. There are three types of features in SDM: single term features which are standard unigram language model features (fT), exact ordered phrase features which are words appearing in sequence (fO), and unordered window features which require words to be close together, but not necessarily in an exact sequence order (fU) [28]. fT, fO, fU are term frequency for the respective features, and λT , λO, λU are weight of the respective features.

System will rank documents in the collection based on the result of SDM scoring function, and top-n documents will be retrieved. We further split each of these documents into passages and choose eight sentences as the passage length, following the setting used by Alzubi et al. [8].

2) Reader system: The reader system receives a set of passages from top-n documents retrieved by the retriever system. There are three model used as the reader. They are BERT-base-uncased, BERT-large uncased, and BERT-large-word-whole-masking.

- BERT-base-uncased: BERT-base-uncased is variation of BERT, a transformers model pretrained on a large corpus of English language [29], using base and uncased version. Uncased means that it does not make any difference between lowercase and uppercase.
- BERT-large-uncased: BERT-large-uncased is another variant from BERT using large and uncased version. This model was pretrained using higher number of parameters and attention heads than base version of BERT model as a result of using a higher number of encoder layers. BERT-base has 12 encoder layers and 768 hidden layers with 12 attention heads and 110 million parameters. On the other hand, BERT-large has 24 encoder layers and 1024 hidden layers with 16 attention heads with 340 million parameters [11].

The architecture of the single encoder from BERT can be seen from Fig. 2. We can see that the architecture of single encoder consist of input embedding, positional encoding, and N block encoder. In input encoding, word will be converted into vector with some steps: tokenization, numericalization and word embedding generation. After converting into vector, the next part is positional encoding. In this part, it will add positional information. Then, the data will go through N encoder blocks in iterative process. This process will capturing more complex relationships between words in the input sequence.

- BERT-large-uncased-whole-word-masking: BERT-large-uncased-whole-word-masking is another BERT version which was pretrained using new technique, called as 'whole word masking' [29]. In this technique, all tokens that are associated with a word will be masked only at once. Meanwhile, for the overall masking rate will remain the same.

The reader component in our CDQA system is each of the BERT model above that is fine-tuned with SQuAD version 1.1 [7]. The CDQA tool is used to fine-tuned the reader. To fine-tune the reader, every document passage will be paired with the query and be transformed into BERT format. Then, the reader will be calculated start logit score and end logit score from every word based model. The logit score itself is the logarithm of the odds from p/(1-p) where p is the probability. Then, the start index is determined based on the start logit value and end index based on end logit value. Once found, predictions are taken based on the start index and end index.

Predictions are taken with several criteria: predictions cannot exceed the maximum answer length, the start index cannot exceed the end index, and some other criteria. Then reader will determine the probability value of each prediction by calculating the softmax function. After that, the results are issued in the form of a tupple by giving three best answer and the order of the answers based on final score. Final score is calculated from retriever score and reader average of start and end logits.

B. Dataset

Table I summarizes all dataset used in this work. In general, there are three datasets utilized in our experiment: SquAD (Stanford Question Answering Dataset) version 1.1, CORD-19 (COVID-19 Open Resource Dataset), and COVID-QA (Question Answering Dataset for COVID-19).

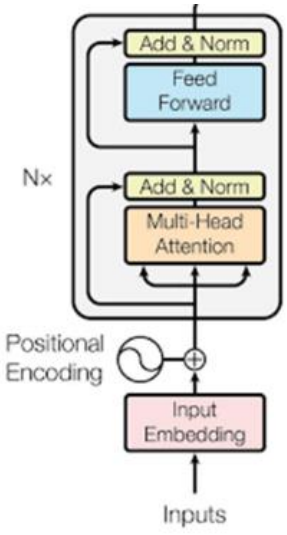

Fig. 2. Architecture from Single Encoder.

TABLE I. STATISTIC DATASET

Dataset	Total Documents	Total Question-Answer Pairs	Topic
SQuAD v.1.1	442	87.599	Open domain
CORD-19	368.618	-	COVID-19 (closed domain)
COVID-QA	147 scientific articles in CORD-19.	2.019	COVID-19 (closed domain)

In order to do transfer learning in our CDQA system, we use SQuAD dataset version 1.1. (SQuAD) as an auxialiary resource. SQuAD will be fine-tune to our reader system in BERT model. SQuAD itself is a reading comprehension dataset. This dataset consists of questions that have been posed by crowd workers from Wikipedia articles. The answer of every question in this dataset can be as a segment of text / span from the corresponding reading passage. This dataset contains over 100.000 question-answer pairs on over 500 articles [7]. This dataset split into training set, development set, and test set. But in this work, we just use training set to fine-tune the reader. So, the squad dataset that we use is 87.599 question-answer pairs on 442 articles. This dataset can be downloaded from https://raw.githubusercontent.com/rajpurkar/SQuAD-explorer/ master/dataset/train-v1.1.json.

CORD-19 dataset use as input in this work. CORD-19 will be indexed by our retriever in order to produce document ranking for every given question to our CDQA system. This dataset is a collection of articles/academic paper in COVID-19, SARS-CoV-2, and related coronaviruses. This dataset was collected by the Semantic Scholar team at the Allen Institute for AI. This dataset consists of over 1,000,000 scholarly articles [5]. But in this research we use 368.618 document that contain the fulltext. This dataset can be downloaded from https://huggingface.co/datasets/cord19.

For evaluation part, COVID-QA dataset [6] will be used because it contains question-answer pairs from the CORD-19 dataset [5] that can provide ground truth answers for the questions. All of COVID-QA dataset will be used to evaluate our CDQA sytem with the baseline. We also use this dataset to evaluate the effectiveness of transfer learning by split this dataset into 70% for fine-tuning and 30% for evaluation. COVID-QA dataset itself consist of 2.019 question-answer pairs. This dataset annotated by volunteer biomedical experts. They selected 147 scientific articles that mostly correlated with COVID-19 from dataset CORD-19 when building this dataset [6]. This dataset can be download in https://github.com/deepset-ai/COVID-QA.

C. Baselines

Three baselines will be used to test the effectiveness of our CDQA system:

- COBERT (TF-IDF+Cosine+BERT) [8]: COBERT [8] is a closed domain question answering that uses TF-IDF Vectorizer with cosine similarity as retriever in order to get top N documents that are related to query.
- BM25+BERT [9]: This baseline uses BM25[9] [20] as retriever to retrieve top N documents that are related to query.
- DLM+BERT: This baseline use Dirichlet Language Model [30], [31] as Retriever to retrieve top n documents that are related to query.

All the above baselines using retriever-reader architecture of CDQA system. They all use BERT model that are finetuned using SQuAD dataset. They only differ in terms of the method used in the retriever component. After the top N documents are split into passages (with the size of passages is eight sentence), the reader then predict / extract answers from the passages and rank them as the results.

IV. Experiment

A. Pre-processing

In the indexing process using PyTerrier, we do some pre-processing in CORD-19 dataset, such as: removing some punctuation, removing stopwords, and lowercasing the text. Then the next process is replacing multiple space into a single space.

B. Hardware

In this work, Google Colab Pro + is used as the machine. This machine use 1 GPU with Tesla T4 type running on CUDA 11.2. Because of technical issue, when doing evaluation part, we move to DGX-1, machine from Tokopedia AI Centre, Universitas Indonesia. It uses 1 GPU with Tesla V100.

C. Finetuning Process

The next process is fine-tuning the reader component based on pretrained language model BERT. Here, BERT-base-uncased, BERT Large Uncased, and BERT-large-uncased-word-whole-masking models are fine-tuned on open domain QA dataset, i.e., SQuAD dataset version 1.1. The hyperparameter that is used to finetune the reader can be seen in Table II.

TABLE II. Hyperparameter to Fine-tune Reader

Hyperparameter	Rincian
Training_batch_size	4
Learning_rate	3e-5
Num_train_epoch	1
Optimizer	adamw
Max_seq_length	384
Max_query_length	64
Gradient_acumulation_step	3
Doc_stride	128

The resulting models are then used to extract and rank answers in closed domain dataset about COVID-19, i.e., CORD-19. By conducting this process, the transfer learning approach is applied by utilizing knowledge learned from large open domain QA dataset to solve the closed domain QA task.

D. Evaluation Metric

In evaluation part, there are two metrics that will be used. They are exact match and F1-score. Exact Match (EM) measures if a resulting answer is exactly similar to the ground truth answer. The score will be 1 if the predicted answer similar with ground truth. Otherwise, the score will be 0.

F1 score is more general metric, that combine together the precision and recall scores. Precision is the ratio that is calculated by the number of the shared words to the total number of words in the prediction. While recall is the ratio that is calculated by the number of shared words to the total number of words in the ground truth. To calculate F1-score, we use the Eq. below (2).

$$F_1 = \frac{2}{recall^{-1} + precision^{-1}} = 2\,x\,\frac{precision - recall}{precision + recall} \quad (2)$$

V. RESULT

A. The Effectiveness of SDM as Retriever in Transfer-Learning-based CDQA System

In the evaluation part, we compare our system with three transfer-learning-based CDQA systems that use retriever-reader architecture as baselines. They are TF-IDF+Cosine+BERT that has been proposed by Alzubi et al. [8], BM25+BERT that has been proposed by Yang et al. [9], and DLM+BERT. Table III shows the comparison of the EM and F1-scores of our method against all baseline methods. The models presented in the table using top 10 document as the retriever output and eight sentences as the passage length.

Our result in Table III show that the use of SDM retriever can improve the F-1 score of the state-of-the-art baseline CDQA system using DLM, BM25 and TF-IDF+Cosine similarity retriever by 3,26% , 3,26% and 32,62%, respectively. Besides, there is improving in EM score when using SDM retriever. SDM retriever can improve EM by 63,06 % for TF-IDF+Cosine similiarity, 2,96% for DLM and BM25. From the result, it turns out that our CDQA system using SDM retriever outperforms all baseline methods. It related with how SDM work. SDM uses various query representations by rewriting each input query in close proximity so that results will be more accurate.

We also explore the answer from two question from all retriever randomly based on Table III in Table IV.

Table IV shows us that one of three answers from SDM retriever and BM25 retriever close to the ground truth both of two questions. Because of that, their F-1 score and EM become one. While, DLM gets the third position, because DLM gets wrong in all answer in first question but true in next question because one of the answer is close to ground truth. While TF-IDF+Cosine similiarity gets the bad result in both of two question because EM Score and F1-score is 0. This cases show us that our SDM retriever is better than TF-IDF+Cosine similiarity practically.

B. The Effect of Tuning the Top N Retrieved Documents, Passage Length, and Reader Variations on the Effectiveness of CDQA System

Table III shows us that SDM becomes the best model among others. Because of that, we try to explore some variation analysis in SDM retriever.

1) Top n document variation: We vary SDM retriever with reader bert based uncased with retriever setting: top five document, top 10 and top 20 document.

Table V shows us that if SDM-BERT-base-uncased sets in top 10 document, it will improve the Top5-SDM-BERT-base-uncased by 10,73% in EM and 3,53% in F-1 score. Then, changing top n document into 20 document can improve Top 5-SDM-BERT-base-uncased by 21,46% in EM and 9,55% in F-1 score. In conclude, setting parameter into top 20 document will make CDQA system become the best result among others.

TABLE III. EVALUATION ON TOP 10 DOCUMENT IN COVID-QA.

Model	EM (%)	F1 Score (%)
Top10-TF-IDF+Cosine-BERT-based-uncased	5.35	17.66
Top10-DLM-BERT-base-uncased	8.42	22.68
Top10-BM25-BERT-base-uncased	8.42	22.68
Top10-SDM-BERT-base-uncased (ours)	**8.67**	**23.42**

TABLE IV. EVALUATION IN TWO QUESTIONS IN TOP 10 DOCUMENT IN COVID-QA

No.	Question			
1.	Question	What is the most common species of Human Coronavirus among adults?		
	Answer (Ground Truth)	HCoV-OC43		
	Method	**3 Answer**	**EM**	**F1-score**
	Top10-TF-IDF+Cosine-BERT-base-uncased	• bats, • obesity • Measles	0	0
	Top10-DLM-BERT-base-uncased	• Bats • bats • Rotaviruses	0	0
	Top10-BM25-BERT-base-uncased	• Rhinovirus • HCoV-OC43 • HCoV-229E	1	1
	Top10-SDM-BERT-base-uncased (ours)	• Rhinovirus • HCoV-OC43 • HCoV-OC43	1	1
2.	Question	What is a natural reservoir of coronavirus?		
	Answer (Ground Truth)	Bats		
	Method	**3 Answer**	**EM**	**F1 score**
	Top10-TF-IDF+Cosine-Bert Based Uncased	• Host Cell Receptor GRP78 • largemouth bass Micropterus salmoides • These reservoirs are located in east Texas	0	0
	Top10-DLM-Bert Based Uncased	• bats • sooty mangabeys • Wild rodents	1	1
	Top10-BM25-Bert Based Uncased	• bats • sooty mangabeys • Chinese horseshoe bats are considered to be the natural reservoirs of SARS-CoV. MERS-CoV	1	1
	Top10-SDM-Bert Based Uncased (ours)	• bats • SARS-CoV • one or more of them may serve as the natural reservoir of SARS-CoV and/or its progenitor virus	1	1

TABLE V. EVALUATION ON TOP N DOCUMENT VARIATION IN COVID-QA

Model	EM (%)	F1 Score (%)
Top5-SDM-BERT-base-uncased	7,83	22,62
Top10-SDM-BERT-base-uncased	8,67	23,42
Top20-SDM-BERT-base-uncased	**9,51**	**24,78**

2) Passage length variation: In Table III, IV and V, we set passage length to eight sentence in every passage based on Alzubi et al. [8] that used eight sentence too. Then, we think about making a deep analysis with some variation of passage length.

Table VI shows us that our CDQA system will get the bad result if the passage length sets under five sentence: one sentence (7,92 for EM and 21,10 for F1-score), three sentence (8,77 for EM and 23,40 for F1-score). It also get the bad result if the passage length sets upper five sentence: eight sentence (8,67 for EM and 23,42 for F1-score), 10 sentence (7,63 for EM and 22,16 for F1-score). So, we can see from Table V that the best passage length is five sentence with 8,77 EM and 23,45 F1-score.

3) Reader variation: We also vary the reader with some models. They are BERT base uncased, BERT large uncased, and BERT large uncased word whole masking in Table VII.

Table VII shows us that the reader with BERT large uncased gets the best result among others. It improves BERT based uncased by 9,68% in EM and 2% in F-1 score. It also improves BERT large uncased whole word masking by 5,55% in EM and 1,01% in F-1 score.

From the above variation, we conclude that the optimal parameters in our CDQA system is retriever with Sequential Dependence Model that uses the top 20 document when retrieved, having passage length five sentence, and having BERT Large Uncased for the reader. Because of that, we try to set all the retriever methods with the optimal parameter in Table VIII.

Table VIII shows us that by optimal parameter in all of the retrievers (top 20 documents as the retriever results, five sentences as the passage length, and BERT-large as the reader model) our CDQA system outperform the baselines by 37,28% in EM and 24,94% in F-1 score for TF-IDF+Cosine similiarity retriever, 5,59% in EM and 4,67% in F-1 score for DLM retriever, and 5,59% in EM and 5,06% in F-1 score for BM25 retriever.

Our CDQA system with SDM retriever with the optimal parameter achieves 10,20 for EM and 24,90 for F1-score. This value is better than SDM retriever in another parameter setting before in Table III, V, VI, and VII.

C. Transfer Learning Versus Non-transfer Learning

We also think that what is the effect of using transfer learning, with the reader retriever (SDM) method for building closed domain question answering, compared to non-transfer learning method. To answer this, we do some work. Different with experiment in Table II, III, V, VI, VII, and VIII that use all COVID-QA for evaluation part. In this experiment, COVID-QA will be split into 70% for fine-tuning and 30% for evaluation. In this experiment, the configuration in the model with the optimal parameter in our closed domain question answering from Table VIII (Top20-SDM-BERT Large Uncased-5 Sentence) will be used. We compare how if the CDQA system use a transfer learning method where the question answering system is fine-tuned with another dataset, such as SQuAD. Then, how if the CDQA system is fine-tuned again with COVID-QA after fine-tuning with SQuAD, and how if the CDQA system is fine-tuned without transfer learning method, that is fine-tuned to COVID-QA only. Table IX will show some experiments about comparing the CDQA system between transfer learning and non-transfer learning.

TABLE VI. EVALUATION ON PASSAGE LENGTH VARIATION IN COVID-QA

Model	EM (%)	F1 Score (%)
Top10-SDM-BERT-base-uncased-1 sentence	7,92	21,10
Top10-SDM-BERT-base-uncased-3 sentence	8,77	23,40
Top10-SDM-BERT-base-uncased-5 sentence	**8,77**	**23,45**
Top10-SDM-BERT-base-uncased-8 sentence	8,67	23,42
Top10-SDM-BERT-base-uncased-10 sentence	7,63	22,16

TABLE VII. EVALUATION ON READER VARIATION IN COVID-QA

Model	EM (%)	F1 Score (%)
Top10-SDM-BERT-base-uncased	8.67	23.42
Top10-SDM-BERT Large Uncased Word Whole Masking	9,01	23,65
Top10-SDM-BERT Large Uncased	**9,51**	**23,89**

TABLE VIII. EVALUATION ON THE OPTIMAL PARAMETER IN COVID-QA

Model	EM (%)	F1 Score (%)
Top20-TF-IDF+Cosine-BERT Large Uncased-5Sentence	7,43	19,93
Top20-DLM-BERT Large Uncased-5Sentence	9,66	23,79
Top20-BM25-BERT Large Uncased-5Sentence	9,66	23,70
Top20-SDM-BERT Large Uncased-5Sentence	**10,20**	**24,90**

TABLE IX. EVALUATION ON TRANSFER LEARNING AND NON-TRANSFER LEARNING METHOD IN 30% COVID-QA

Model	EM (%)	F1 Score (%)
Top 20-SDM-BERT Large Uncased-5 Sentence-Fine-tuned to CovidQA (Non-transfer learning)	0,0	3,20
Top 20-SDM-BERT Large Uncased-5 Sentence-Fine-tuned to SQuAD (transfer learning)	6,44	23.29
Top 20-SDM-BERT Large Uncased-5 Sentence-Double Fine-tuned(SQuAD+COVID-QA) (transfer learning fine-tuned to COVID-QA)	**10,07**	**29,64**

Table IX shows us that models that use transfer learning are better than models that use non-transfer learning method. Models that use transfer learning try to mitigate question answering system that has low resource question-answering dataset by transferring information from auxiliary dataset to improve their performance. It is especially useful for low-resource tasks such as our closed domain question answering in COVID-19. Models that use non-transfer learning, where the reader is fine-tuned to COVID-QA, achieve bad results. It gets 0.0 in EM and 3,20 in F1-score. It looks like this model becomes underfitting because the question-answer pair dataset

in COVID-QA train is too small. All of us know that SQuAD has 87599 question-answer pairs, while COVID-QA has only 2019 question-answer pair. Then for fine-tuning, 1413 question-answer pair has been used. The amount of the dataset makes the model unable to learn. If COVID-QA dataset as big as SQuAD may be the model will be better.

Because there is no dataset question-answer pair in CORD-19, our work tries to use transfer learning method. Table IX shows two model that use transfer learning. First is a model that reader is fine-tuned to SQuAD only. The second is a model that reader is fine-tuned to SQuAD and fine-tuned again with COVID-QA. We do the double fine-tuned based on Yang et al. [32]. They said that fine-tuning again question answering system will make improvement to the model.

Table IX shows us that our system that use transfer learning (fine-tune to SQuAD) improves the effectiveness of non-transfer learning method with six times higher accuracy than the baseline method without using transfer learning. Further fine-tuning again the reader model that has fine-tuned before with SQuAD into closed domain dataset (COVID-QA) can increase the accuracy of the model by 27,26 % in F-1 score and improve F-1 score of non-transfer learning method with eight times higher. From this result, we can say that transfer learning method will get better result than non-transfer learning when resource of dataset question-answer pair itself is small.

VI. Discussion

Our experimental results show that transfer learning can give a significant improvement to the CDQA system that does not use transfer learning. The transfer learning model gains eight times higher accuracy than the method without using transfer learning. When the model is further fine-tuned with close domain dataset CDQA, the accuracy increased by 27,26%. This accuracy can further increase when a bigger CDQA dataset is available for fine-tuning. Therefore, creating a bigger CDQA dataset will become a research challenge in this case.

This research also has some limitation in terms of the variation of parameters explored in the experiment. For example, for the number of documents retrieved by retriever, we only experimented until 20 documents retrieved for each question. Then, for the length of passages, we only experimented until 10-sentences length for each passage. This restriction is related with the availability of GPU computing resource that is limited, while our model in general requires high computing resource. From this reason, we limit the variation of parameters in our experiment when fine-tuning the reader component. When a higher GPU is available, it may worth experimenting with more parameter values in our system. We can investigate, for example, whether retrieving more than 20 documents may further increase the performance of CDQA system.

VII. Conclusion

This paper explores the use of transfer learning technique to build close-domain question answering (CDQA) system in COVID-19. We propose to use Sequential Dependence Model (SDM) as retriever in the retriever-reader architecture to improve the accuracy of CDQA system. Our experimental results show that the use of SDM as retriever leads to significant performance of the CDQA system. Our model using SDM as the retriever and BERT-base that was fine-tuned on SQuAD benchmark as the reader can outperform F-1 score of the state-of-the-art baseline by 3,26% for BM25, 3,26% for DLM and 32, 62% for TF-IDF+Cosine similiarity.

The hyperparameter tuning on top n documents, passage length, and reader variations are shown to affect the performance of our CDQA model. We found that the best parameter is achieved by using top 20 documents as the retriever results, five sentences as the passage length, and BERT-large as the reader model. Our CDQA system using this parameter setting results in the best-performing model that achieves higher outperforming in F-1 score to the state-of-the-art baseline by 24,94% for TF-IDF+Cosine similiarity retriever, 4,67% for DLM retriever, and 5,06% for BM25 retriever.

Our results also confirm the merit of using transfer learning to build CDQA system in a condition where a large benchmark for CDQA is unavailable. Our CDQA system using transfer learning technique (i.e., using SDM retriever and BERT-large reader that was fine-tuned on SQuAD benchmark) is significantly more effective than the method that does not use transfer learning. Our best-performing model (double fine-tuning SQuAD and COVID-QA) is shown to gain eight times higher accuracy than the baseline method without using transfer learning. Further fine-tuning the transfer learning model using closed domain dataset can increase the accuracy of the transfer learning model that only fine-tuning with SQuAD by 27,26%.

VIII. Future Work

For future works, we consider developing closed domain question answering with Dense Passage Retrieval to get better result than this work. We also want to modify how if all documents have split into passages before sending it to the retriever.

Acknowledgment

This research was funded by the Directorate of Research and Development, Universitas Indonesia, under Hibah PUTI Pascasarjana 2022 (Grant No: NKB-03/UN2.RST/HKP. 05.00/2022).

References

[1] T. Acter, N. Uddin, J. Das, A. Akhter, T. R. Choudhury, and S. Kim, “Evolution of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) as coronavirus disease 2019 (COVID-19) pandemic: A global health emergency,” Science of The Total Environment, vol. 730, p. 138996, Aug. 2020, doi: 10.1016/j.scitotenv.2020.138996.

[2] C. Huang et al., “Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China,” The Lancet, vol. 395, no. 10223, pp. 497–506, Feb. 2020, doi: 10.1016/S0140-6736(20)30183-5.

[3] N. Zhu et al., “A Novel Coronavirus from Patients with Pneumonia in China, 2019,” N Engl J Med, vol. 382, no. 8, pp. 727–733, Feb. 2020, doi: 10.1056/NEJMoa2001017.

[4] P. Zhou et al., “Discovery of a novel coronavirus associated with the recent pneumonia outbreak in humans and its potential bat origin,” Microbiology, preprint, Jan. 2020. doi: 10.1101/2020.01.22.914952.

[5] L. L. Wang et al., “CORD-19: The COVID-19 Open Research Dataset,” 2020, doi: 10.48550/ARXIV.2004.10706.

[6] T. Möller, A. Reina, R. Jayakumar, and M. Pietsch, "COVID-QA: A Question Answering Dataset for COVID-19," in Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020, Online, Jul. 2020. Accessed: Nov. 24, 2022. [Online]. Available: https://aclanthology.org/2020.nlpcovid19-acl.18

[7] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "SQuAD: 100,000+ Questions for Machine Comprehension of Text," in Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin, Texas, 2016, pp. 2383–2392. doi: 10.18653/v1/D16-1264.

[8] J. A. Alzubi, R. Jain, A. Singh, P. Parwekar, and M. Gupta, "COBERT: COVID-19 Question Answering System Using BERT," Arab J Sci Eng, Jun. 2021, doi: 10.1007/s13369-021-05810-5.

[9] W. Yang et al., "End-to-End Open-Domain Question Answering with," in Proceedings of the 2019 Conference of the North, Minneapolis, Minnesota, 2019, pp. 72–77. doi: 10.18653/v1/N19-4013.

[10] D. Metzler and W. B. Croft, "A Markov random field model for term dependencies," in Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '05, Salvador, Brazil, 2005, p. 472. doi: 10.1145/1076034.1076115.

[11] E. Yulianti, R.-C. Chen, F. Scholer, W. B. Croft, and M. Sanderson, "Ranking Documents by Answer-Passage Quality," in The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, Ann Arbor MI USA, Jun. 2018, pp. 335–344. doi: 10.1145/3209978.3210028.

[12] E. Yulianti, R.-C. Chen, F. Scholer, and M. Sanderson, "Using Semantic and Context Features for Answer Summary Extraction," in Proceedings of the 21st Australasian Document Computing Symposium, Caulfield VIC Australia, Dec. 2016, pp. 81–84. doi: 10.1145/3015022.3015031.

[13] D. Chen, A. Fisch, J. Weston, and A. Bordes, "Reading Wikipedia to Answer Open-Domain Questions," in Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Vancouver, Canada, Jul. 2017, pp. 1870–1879. doi: 10.18653/v1/P17-1171.

[14] Y. Yang, W. Yih, and C. Meek, "WikiQA: A Challenge Dataset for Open-Domain Question Answering," in Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, 2015, pp. 2013–2018. doi: 10.18653/v1/D15-1237.

[15] S. Reddy, D. Chen, and C. D. Manning, "CoQA: A Conversational Question Answering Challenge," 2018, doi: 10.48550/ARXIV.1808.07042.

[16] K. M. Hermann et al., "Teaching Machines to Read and Comprehend," 2015, doi: 10.48550/ARXIV.1506.03340.

[17] L. F. Simanjuntak, R. Mahendra, and E. Yulianti, "We Know You Are Living in Bali: Location Prediction of Twitter Users Using BERT Language Model," BDCC, vol. 6, no. 3, p. 77, Jul. 2022, doi: 10.3390/bdcc6030077.

[18] E. Yulianti, A. Kurnia, M. Adriani, and Y. S. Duto, "Normalisation of Indonesian-English Code-Mixed Text and its Effect on Emotion Classification," IJACSA, vol. 12, no. 11, 2021, doi: 10.14569/IJACSA.2021.0121177.

[19] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research 12 (2011), pp. 2825–2830.

[20] G. Amati, "BM25," in Encyclopedia of Database Systems, L. Liu and M. T. Özsu, Eds. Boston, MA: Springer US, 2009, pp. 257–260. doi: 10.1007/978-0-387-39940-9_921.

[21] R. Nogueira and K. Cho, "Passage Re-ranking with BERT." arXiv, Apr. 14, 2020. Accessed: Nov. 24, 2022. [Online]. Available: http://arxiv.org/abs/1901.04085

[22] S. J. Semnani, M. Pandey, and M. Pandey, "Domain-Specific Question Answering at Scale for Conversational Systems," 33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada, p. 10.

[23] Akdemir, "Research on Task Discovery for Transfer Learning in Deep Neural Networks," in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, Online, 2020, pp. 33–41. doi: 10.18653/v1/2020.acl-srw.6.

[24] Akdemir and T. Shibuya, "Transfer Learning for Biomedical Question Answering," CLEF 2020, 22-25 September 2020, Thessaloniki, Greece, vol. 2696, p. 15.

[25] Z. H. Syed, A. Trabelsi, E. Helbert, V. Bailleau, and C. Muths, "Question Answering Chatbot for Troubleshooting Queries based on Transfer Learning," Procedia Computer Science, vol. 192, pp. 941–950, 2021, doi: 10.1016/j.procs.2021.08.097.

[26] Macdonald, N. Tonellotto, S. MacAvaney, and I. Ounis, "PyTerrier: Declarative Experimentation in Python from BM25 to Dense Retrieval," in Proceedings of the 30th ACM International Conference on Information & Knowledge Management, Virtual Event Queensland Australia, Oct. 2021, pp. 4526–4533. doi: 10.1145/3459637.3482013.

[27] S. Huston and W. B. Croft, "A Comparison of Retrieval Models using Term Dependencies," in Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, Shanghai China, Nov. 2014, pp. 111–120. doi: 10.1145/2661829.2661894.

[28] Z. J. Zhang et al., "A generic retrieval system for biomedical literatures: USTB at BioASQ2015 Question Answering Task," p. 7.

[29] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," 2018, doi: 10.48550/ARXIV.1810.04805.

[30] Zhai and J. Lafferty, "A study of smoothing methods for language models applied to Ad Hoc information retrieval," in Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, New York, NY, USA, Sep. 2001, pp. 334–342. doi: 10.1145/383952.384019.

[31] J. M. Ponte and W. B. Croft, "A language modeling approach to information retrieval," in Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '98, Melbourne, Australia, 1998, pp. 275–281. doi: 10.1145/290941.291008.

[32] W. Yang, Y. Xie, L. Tan, K. Xiong, M. Li, and J. Lin, "Data Augmentation for BERT Fine-Tuning in Open-Domain Question Answering." arXiv, Apr. 14, 2019. Accessed: Nov. 24, 2022. [Online]. Available: http://arxiv.org/abs/1904.06652