A Smart System to Generate and Validate Question Answer Pairs for COVID-19 Literature

Rohan Bhambhoria[*†], **Luna Feng**[*‡], **Dawn Sepehr**[*‡], **John Chen**[§¶], **Conner Cowling**[‡], **Sedef Akinli Kocak**[¶], **Elham Dolatabadi**[§¶]
Queen's University[†], Thomson Reuters[‡], University of Toronto[§], Vector Institute[¶]
r.bhambhoria@queensu.ca, johnc@cs.toronto.edu,
{luna.feng,dawn.sepehr,conner.cowling}@thomsonreuters.com,
{sedef.kocak, elham.dolatabadi}@vectorinstitute.ai

Abstract

Automatically generating question answer (QA) pairs from the rapidly growing coronavirus-related literature is of great value to the medical community. Creating high quality QA pairs would allow researchers to build models to address scientific queries for answers which are not readily available in support of the ongoing fight against the pandemic. QA pair generation is, however, a very tedious and time consuming task requiring domain expertise for annotation and evaluation. In this paper we present our contribution in addressing some of the challenges of building a QA system without gold data. We first present a method to create QA pairs from a large semi-structured dataset through the use of transformer and rule-based models. Next, we propose a means of engaging subject matter experts (SMEs) for annotating the QA pairs through the usage of a web application. Finally, we demonstrate some experiments showcasing the effectiveness of leveraging active learning in designing a high performing model with a substantially lower annotation effort from the domain experts.

1 Introduction

Building a QA system is a complex process requiring advanced text mining approaches (Jothi et al., 2015) and domain expertise for model evaluation. Accordingly, automatically generating question-answer pairs using recent advances in natural language processing (NLP) models has gained much attention from researchers and has achieved impressive results on various publicly available datasets. (Yang et al., 2018; Rajpurkar et al., 2016). In this work, we explore the COVID-19 Open Research Dataset (CORD-19) (Wang et al., 2020) first introduced in a Kaggle Competition[1]. The competition has been launched as a call to action for machine learning researchers to assist the medical community in developing answers to high-priority scientific questions related to COVID-19. A major challenge in dealing with a large semi-structured dataset (i.e., scholarly articles) is the lack of gold data which we aim to address in this work.

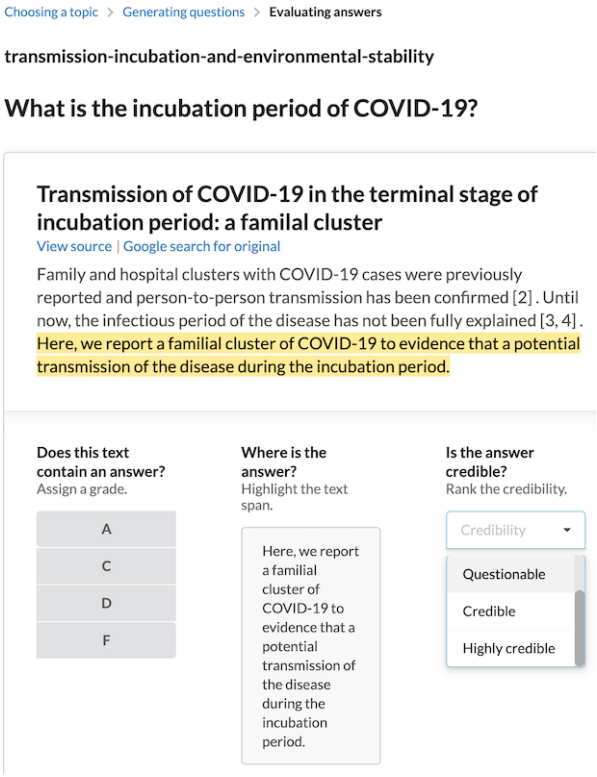


Figure 1: User Engagement App: SMEs are provided with information including the question, the title of the article, and some context from the article to grade the answer, highlight the exact answer, and rate the credibility of the source.

[*]Equal contributions, listed alphabetically

[1]https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge

Proceedings of the First Workshop on Scholarly Document Processing, pages 20–30

https://doi.org/10.18653/v1/P17

Existing methods developed by different groups in the Kaggle competition have mainly used clustering approaches (Kanungo et al., 2002) coupled with statistical methods (Blei et al., 2003) in order to group articles together and discover keywords from the resulting representation of the scholarly articles, respectively. Other researchers made use of transformer-based QA models and BERTserini (Yang et al., 2019) to retrieve relevant answers to keywords extracted from a question, after which resulting solutions were ranked by unsupervised embedding methods (Cer et al., 2018). Finally, top-ranking results were combined and summarized (Chipman et al., 2010). Other QA models made use of BERT (Devlin and Toutanova, 2019), adapted variations of BERT models (Huang et al., 2019; Beltagy et al., 2019), and BERT-like models (Lan et al., 2020) to produce semantically meaningful sentence embeddings from abstracts to answer important questions raised by the healthcare community. Developing QA systems for COVID-19 was not limited to the Kaggle competition; in (Oniani and Wang, 2020), a hybrid approach based on GPT-2 (Radford et al., 2019) was proposed to generate responses for different COVID-19 related questions. In general, developing QA generation has achieved promising progress recently. However, answering a question in specific domains such as health domain is still challenging, due to the requirement of expert knowledge and lack of high-quality training data. For example, Walonoski et al. (Walonoski et al., 2018) focused on generation of dataset from the state transition of patient records. Recently, Shen at al. (Shen et al., 2020) introduced structure information of QA pairs generation in medical domain. They proposed an unsupervised detector to automatically explore external materials for the validity of generated QA pairs. Despite all these attempts and solutions by various researchers, lack of annotated data for the CORD-19 dataset presents a challenge to automatically verify the correctness of the created QA pairs and also prevents us from leveraging supervised techniques.

To help address the shortcomings of previous approaches, we aim to create gold data related to COVID-19 which in turn can serve the purpose of training and evaluating supervised models. We employ transformer and rule-based methods to automatically generate a set of QA pairs, which we call Silver QA, and then leverage various active learning selection strategies to present samples to SMEs for annotation. To the best of our knowledge, this is the first work which explores the potential use of generative models to create QA pairs for quality verification by SMEs. Our proposed approach can serve as a practical foundation for the creation of a QA system for any complex semi-structured dataset requiring the employment of domain knowledge experts to maintain the standard of the generated QA pairs.

To provide a better user experience during the annotation process, we build a web application, which we call the User Engagement App (UEA), shown in Figure 1. The UEA presents a batch of QA pairs once the SMEs select their expertise of a specific domain and topics of interest (e.g., vaccines and therapeutics, virus genetics, origin, and evolution). It also allows the SMEs to grade the QA pairs, select exact answers, and rank the credibility of the source. While developing the web application, we engaged two medical students to obtain their feedback. We summarize their feedback into two major issues when annotating the QA pairs, details of which are outlined in Section 2. From these sets of feedback, we may firstly conclude that the SMEs require several hours to review a small batch of QA pairs due to the scientific complexity of the questions, and secondly, a high degree of domain-specific knowledge (e.g., virology, molecular genetics) is required to answer these scientific questions. We address the first feedback from the SMEs by introducing an active learning strategy which provides a method on how to select a limited number of samples. This in turn reduces the annotation efforts required to develop a practical QA system. In the future work section, we also provide some directions for the second feedback based on the results obtained from the QA pairs that we have generated.

Figure 2 illustrates the overview of how we integrate these different strategies to create the QA pairs and obtain gold data provided by the SMEs via the web application. We explain these steps in more details in Sections 2 and 3 and also provide the experimental results in Section 4.

2 Datasets

In this section, we introduce the publicly available biomedical datasets that we use for our experiments. We also describe various methods conducted in this work to generate QAs specifically but not limited to COVID-19.

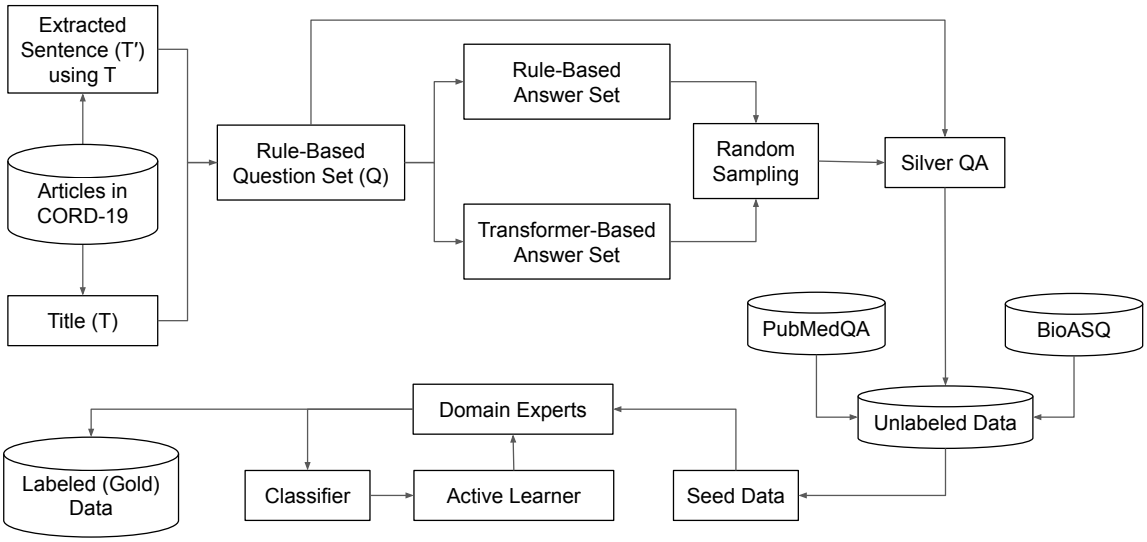


Figure 2: Generation of the Silver QA Data and the Process of Obtaining Gold Data using Active Learning

2.1 Existing Biomedical Datasets

We use three existing biomedical datasets:

- PubMedQA (Jin et al., 2019): a biomedical QA dataset collected from the abstracts of PubMed articles. PubMedQA has 1k expert-annotated, 61.2k unlabeled and 211.3k artificially generated QA instances.
- BioASQ (Tsatsaronis et al., 2015): a biomedical QA dataset consisting of 2,747 questions in the "Training 7b" dataset and 500 questions in the "7b golden enriched" test dataset. The questions are provided with their relevant articles, snippets, concepts and Resource Description Framework (RDF) triples, "exact" and "ideal" answers.
- CORD-19 (Wang et al., 2020): a resource of over 158K scholarly articles (released on April 16th, 2020), including over 75K with full text, about COVID-19, SARS-CoV-2, and related coronaviruses collected by the White House and a coalition of leading research groups and it was released along the Kaggle Competition.

2.2 Generating the Silver QA Dataset

We limit our question generation procedure to transformer and rule-based methods, refraining from the usage of other neural question generation methods such as (Du et al., 2017; Krishna and Iyyer, 2019) in order to create simple baselines for the comparison and evaluation of subsequent sections. We recognize the importance of exploring neural methods to further assess the quality of the generated QA pairs and leave this exploration to the future work. The details of the steps taken to create the Silver QA are explained in this section and are also illustrated at the top part of Figure 2. In order to engage the SMEs effectively, we use two approaches. In our first approach, we create "QA-pre" by considering only the titles of the CORD-19 dataset starting with "Do/Does" and "Is/Are", i.e., titles with these prepended keywords, in order to be consistent with the schema of PubMedQA in which questions can be answered by "yes/no/maybe". However, this would result in a very small dataset. We therefore also include questions prefixed with "Wh". Including all the three question types our dataset contained only 553 titles with these prefixes which can serve as questions. To further generate a larger dataset, we consider all titles with verbs and incorporate the usage of a POS-tagger to formulate questions from titles. The resulting titles are grammatically incorrect in several instances and a potential direction would be to solely prepend "Do/Does" and "Is/Are" to titles containing verbs to formulate questions. This would require the SMEs to correct questions if they do not match the potential answer based on the choices available to them. Due to the limited size of "QA-pre", we introduce our second approach using a siamese BERT structure (SBERT) (Reimers and Gurevych, 2019).

To create a structured dataset which is practical for the SMEs to provide annotations, we follow a

procedure of creating valid QA pairs using:

1. Titles of CORD-19 scholarly articles which we consider, for n articles in the CORD-19 dataset as the set $\{t_1, t_2, ...t_n\} \in T$ where T contains all titles obtained from the scholarly articles, $\{d_1, d_2, ...d_n\} \in D$ where d_i represents a single article from which a corresponding title, t_i is obtained.

2. Sentences with high cosine similarity to the titles of CORD-19 scholarly articles taken from the abstracts and conclusions from CORD-19 represented by $\{t_1', t_2', ..., t_n'\} \in T'$, obtained using SBERT as the encoder.

Sentences from T and T' are used as inputs for generating a set of questions $\{q_1, q_2, ...q_n\} \in Q$ and answers $\{a_1, a_1', a_2, a_2', ..., a_n', a_n\} \in A$. The set of questions, Q, is generated from these sentences solely using rule-based methods which make use of a syntactic parser to refactor them into questions by prepending "Wh" interrogative words to T and T' (Heilman and Smith, 2009). Similarly, the answers, $a_i \in A$, are generated based on rule-based methods, also making use of a syntactic parser for matching or refactoring sentences to a specific structure, maintaining an answer-like format. The subset of answers, $a_i' \in A$ are generated from the recently released Text-to-Text Transfer Transformer model, T5 (Raffel et al., 2019). This becomes possible as T and T' contains sufficient context for any given scholarly article in D. At the time of writing this paper, T5 is the highest ranking encoder-decoder structured model on a wide variety of NLP tasks, including the GLUE benchmark (Wang et al., 2018), and the extractive, context-based question answering task (Rajpurkar et al., 2016). Moreover, T5 also shows good performance on *closed-book question answering*, a question answering task that involves generating answers to questions when no context is supplied (Roberts et al., 2020).

The generation of $a_i' \in A$ involves finetuning of a pre-trained "large" configuration of the T5 model (770M parameters) on a mixture of three datasets: the TriviaQA dataset (Joshi et al., 2017), the Natural Questions dataset (ignoring the available context) (Kwiatkowski et al., 2019), and finally a domain specific COVID-19 dataset[2] with human-annotated answers. The model is finetuned for 25,000 steps and greedy decoding is performed.

[2]https://github.com/xhlulu/covid-qa

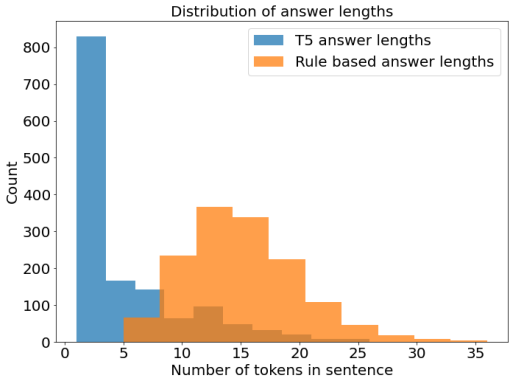


Figure 3: T5 answers are generally composed of fewer tokens than those generated via the rule-based approach.

As titles may not be written in the form of natural sentences, we add an additional layer of validation by asking four individuals without domain-specific expertise to provide manual annotations. The purpose of this validation step is to filter out grammatically incorrect questions generated by the rule-based methods. We measure the inter-annotator agreement of these annotations using the Cohen's kappa coefficient (McHugh, 2012) on 100 unique questions from T' and summarize these statistics in Table 1. In general higher values of this coefficient confirms a higher level of agreement between the annotators (Landis and Koch, 1977). We observe that the inter-annotator agreement scores for these annotations are low and we suspect this is due to the lack of detailed instructions for the annotation process. We would like to address this issue by setting up comprehensive instructions in the future work. As T' contains a natural sentence structure in contrast to T, QA pairs formulated from these sentences are directly taken into consideration without the extra validation step for filtering.

Annotator ID	**1**	**2**	**3**	**4**
1	1	0.21	0.35	0.16
2	0.21	1	0.53	0.52
3	0.35	0.53	1	0.4
4	0.16	0.52	0.4	1

Table 1: Inter-annotator agreement scores calculated based on Cohen's kappa coefficient

The resulting dataset, which we denote as Silver QA, contains all questions from Q and randomly sampled answers from A. The reason for random sampling of answers which are generated

by transformer and rule-based methods is to avoid bias which may be introduced by either method on producing answers. A subset of the samples in the Silver QA dataset are shown in Table 2. We find that the two approaches generally do not produce similar answers: neither method produces the exact same answer for any given input question. The rule-based method, which extracts chunks of text, produces longer answers, while the T5 model, which generates answers token by token via greedy decoding, produces more terse responses as seen in Figure 3. When looking at fuzzy matches (as computed by the fuzzywuzzy[3] Python package), the answers have a fuzzy match score (Levenshtein distance similarity ratio) (Levenshtein, 1966) of 0.196, indicating on average that a large number of edits are required to transform one answer to another. Finally, they also share a low average cosine similarity of 0.07 on a simple bag-of-words encoding.

3 An Active Learning Strategy for Data Selection

Active learning (AL) strategies are shown to be effective in reducing the number of samples a machine learning model requires to achieve comparable performance to the case where a large amount of data is annotated (Aggarwal et al., 2014). Here, the main idea is to ask the SMEs to annotate strategically picked samples in small batches to minimize their efforts while encouraging the creation of a successful QA system.

At a high level, we start by randomly choosing a small subset of the samples from the unlabeled pool, details are in the next section, as the seed data to be annotated by the SMEs. Using the annotated seed data, we train a binary classifier to differentiate between different samples. The next step is to choose which unlabeled data points should be sent to the SMEs in the next iteration. After obtaining the annotations, the labeled samples are added to the pool of labeled data and further used to retrain the classifier. This iterative process, as illustrated at the bottom part of Figure 2, is repeated until either the annotation budget runs out or all the samples in the unlabeled pool have been annotated. In the following subsections, we explain how we formulate this problem as a binary classification task and introduce the different sampling strategies we have implemented.

[3]https://github.com/seatgeek/fuzzywuzzy

3.1 Problem Formulation

To show the effectiveness of the AL strategies, we consider the following human-annotated QA pairs from the datasets introduced in Section 2: 1) the questions and long answers from the PubMedQA expert-annotated dataset considering only the yes/no answers; and 2) the questions and ideal answers from the BioASQ dataset. We label these QA pairs as "valid" since the answers are the expected results for the questions. We also consider the QA pairs created in the Silver QA as valid. However, to differentiate between the valid QA pairs already annotated by the SMEs and the ones for which we would like to get the SMEs feedback using the UEA, we assign different weights to these samples in the AL strategy. We explain the details of these strategies in Section 3.2. It is noteworthy that we consider the PubMedQA and BioASQ datasets in our experiments since our models can benefit from these larger publicly available structured datasets in the biomedical domain which is similar to the domain of the COVID-19.

Furthermore, to create the set of QA pairs labeled as "invalid", for each question in the valid QA pairs, we randomly select a text snippet from the articles in the CORD-19 dataset and use it as the answer assuming that there is a small chance that the text snippet actually answers the question. Following this procedure, we build a dataset with 23,208 valid and invalid QA pairs with a 50% split between the two classes. We keep nearly 5% of the samples which results in 1,000 samples in each of the validation and test datasets with an equal split between the two classes and use the rest for training.

At this point, we can utilize a binary classifier to distinguish between the valid and invalid QA pairs to pick which samples should be sent to the SMEs. This trained binary classifier can be further used in our QA system to retrieve answers that are more likely to be labeled as correct by the SMEs, thus, reducing the cost of the annotation process even further. We choose XGBoost (Chen and Guestrin, 2016) as the classifier due to its efficiency in speed and performance in the AL experiments. Furthermore, we use sentence embeddings produced by transformer-based models (Vaswani et al., 2017) such as BERT (Devlin and Toutanova, 2019), and BioBERT (Lee et al., 2019) as the features for each QA pair. Specifically, we concatenate each question and answer separated by a blank space and

Question	Answer (Rule-Based)	Answer (Transformer-Based)
What has played a significant role in controlling measles in China?	The live-attenuated measles virus vaccine based on the Hu191 strain has played a significant role in controlling measles in China	the chinese government has taken proactive steps to reduce the spread of the disease
What are respiratory and enteric bovine coronavirus strains distinctive in?	It is unclear whether respiratory and enteric bovine coronavirus strains are distinctive in biological, antigenic and genetic characteristics	they are not conspecific
What is Pneumonia an inflammatory disease of?	Pneumonia is an inflammatory disease of the lung, responsible for high morbidity and mortality worldwide	lungs
what causes lower respiratory tract infections?	Background: Human metapneumovirus causes lower respiratory tract infections, particularly in young children and the elderly	bacteria, viruses, and protozoa
What mediates viral entry into host cells?	The filovirus surface glycoprotein mediates viral entry into host cells	a complex interaction between the virus and host cell membranes

Table 2: Qualitative Assessment of QA Pairs. The first column contains Questions from the set Q, the second and third columns contain generated rule-based and transformer-based answers from the set A respectively.

then obtain its embedding. More details of these experiments are reported in Section 4.

3.2 Design of the Sampling Strategies

We propose leveraging different AL strategies to sample unlabeled QA pairs from the pool of valid and invalid QA pairs to be annotated by the SMEs. A baseline strategy in comparison with any AL approach is choosing the samples randomly according to a uniform probability distribution and we also use this baseline to compare the performance of our proposed methods.

The first AL strategy that we implement, denoted by AL-Uncertainty, is based on the uncertainty of the classifier. In this case, the probability of the labels predicted by the classifier is used as a measure of uncertainty and the samples for which the binary classifier is the least certain about their labels are selected for annotation. Despite its simplicity, this technique has been successfully used in many different applications and has been one of the ubiquitous AL strategies to select the most informative samples for a model to be annotated by the SMEs (Fu et al., 2013; Aggarwal et al., 2014; Konyushkova et al., 2017). One can formulate this strategy as follows

$$x^* = \arg\min_{x_i \in \mathcal{U}} P(y_i = y \mid x_i) \qquad (1)$$

where $P(y_i = y \mid x_i)$ is the probability of the predicted class y for sample x_i, $\mathcal{U}$ is the pool of unlabeled samples, and x^* is the sample picked for annotation. In our simulations, in order to differentiate between the already human-annotated QA pairs and the samples in the Silver QA dataset, we consider different weights for different data sources. Indeed, we rank the samples after considering their class predicted probability based on the weight of their data source. Thus, we can write the following

$$R(x_i) = w_i \, r_{\mathrm{P}}(x_i) \qquad (2)$$

where $R(x_i)$ is the final rank of sample x_i, w_i is the weight assigned to the source of x_i, and $r_{\mathrm{P}}(x_i)$ is the rank of x_i using the probability P over all samples in $\mathcal{U}$. The AL-Uncertainty strategy picks a number of samples equal to the batch size which have the lowest final rank R, thus, samples with a lower source weight have a higher chance of being selected for annotation.

As our second AL strategy, we propose promoting sample diversity to the uncertainty approach in order to improve the performance of the AL-Uncertainty as explained in (Fu et al., 2013). Inspired by (Shuyang et al., 2018) and denoted by AL-Clustering, this strategy is based on clustering the samples. This method clusters the samples, represented by the features obtained from the embeddings of the QA pair, and then picks one sample

within each cluster for which the classifier is the least confident about its predicted label. We can formulate this strategy as follows

$$x^* = \arg\min_{x_i \in \mathcal{U}_{c_i}} P(y_i = y \mid x_i) \quad (3)$$

where $\mathcal{U}_{c_i}$ is the cluster that x_i belongs to in the pool of unlabeled samples, and x^* is the sample picked for annotation in cluster c_i. We would like to emphasize that unlike the method in (Shuyang et al., 2018), we perform the clustering in each iteration to rearrange the samples in different clusters as we get more annotated samples. We set the number of clusters in each iteration equal to the batch size and also employ the same weighting scheme described for the AL-Uncertainty strategy which results in

$$R^{c_i}(x_i) = w_i \, r_P^{c_i}(x_i) \quad (4)$$

where $R^{c_i}(x_i)$ is the final rank of sample x_i in its cluster c_i, w_i is the weight assigned to the source of x_i, and $r_P^{c_i}(x_i)$ is the rank of x_i using the probability P over all samples in $\mathcal{U}$ which belong to cluster c_i. The AL-Clustering strategy picks one sample in each cluster which has the lowest final rank R, thus, similar to the AL-Uncertainty strategy, samples with a lower source weight have a higher chance of being selected for annotation.

4 Experiments and Results

We evaluate the performance of the different AL strategies by reporting both the accuracy and F1 score on the test dataset. To have a fair assessment of the performance of each method, we run the experiments 5 times using different random seeds and average the results as illustrated in Figure 4. We discuss the details of the implementation and setup for all of these experiments in Section 4.1 and discuss the details of the results in Section 4.2

4.1 Experimental Setup

We randomly choose 20 QA pairs from the unlabeled pool of the training set as the initial seed data. The batch size of the samples to be selected per iteration is 5 and the total number of iterations is set to 50 which amounts to 1.2% of the entire training dataset. As aforementioned, we use the XGBoost classifier to predict whether a QA pair is valid or not. The sentence embedding of each QA pair is employed as its input features. We experiment with two settings of embeddings that are produced by the pre-trained "bert-base-cased"[4] and "biobert-base v1.1"[5] transformer-based models. For both of these models, the output of the network for the [CLS] token is used to represent the input sentence.

For the weights of data sources utilized in the AL-Uncertainty and AL-Clustering strategies, we follow this scheme: X weights to the samples from PubMedQA and BioASQ, $3X$ weights to the samples from Silver QA, and $2X$ weights to the samples from CORD-19. The intuition behind this setting is that, at this point, the samples in Silver QA and CORD-19 have not been validated by the SMEs yet. However, in the real world scenario when we use the UEA, reversing the setting of the weights (i.e., lower values given to QA pairs from Silver QA and CORD-19) would result in a higher probability for selection of the QA pairs for which the model is unable to make a clear judgment about their labels. Also, due to the fact that the QA pairs from PubMedQA and BioASQ have already been annotated, we can filter them out in the UEA if they are selected by the active learner.

4.2 Results and Discussion

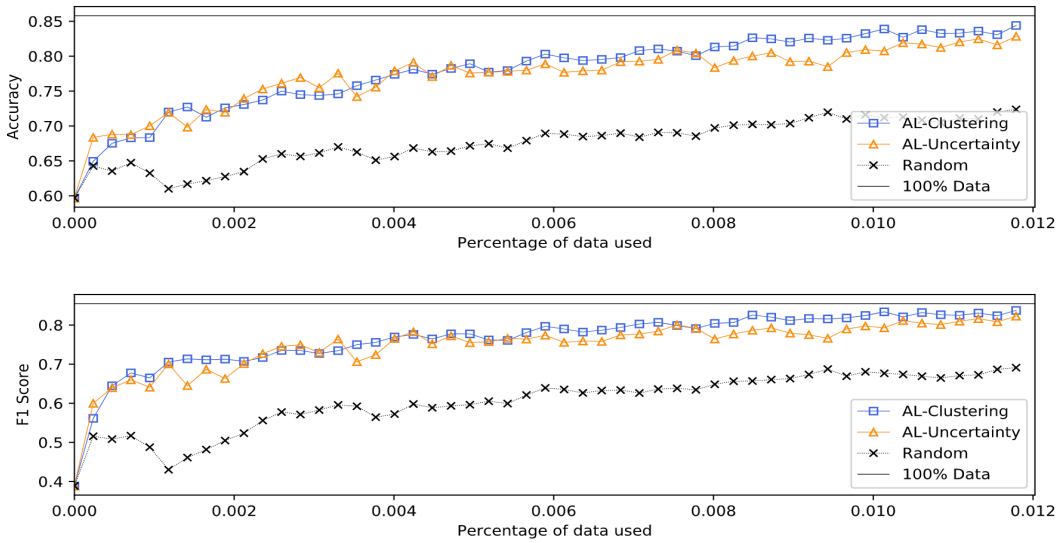
We empirically compare the performance of the three selection strategies described in Section 3.2. We also compare the achievable performance of the two XGBoost-based models with "DistilBERT-base-cased"[4], a BERT-based classifier, when trained on the entire dataset, as reported in Table 3. We observe that the DistilBERT model (Sanh et al., 2019) outperforms the XGBoost-based models substantially due to its more advanced architecture in which the embedding layers of the network are also updated during the training whereas the sentence embeddings used in the XGBoost-based models are static. However, incorporating the DistilBERT model in our experiments is both time and resource intensive since DistilBERT runs 36 times slower than XGBoost per AL iteration on a K80 GPU. Therefore, in this work, we only experiment with

Model	F1 Score
BERT + XGBoost	0.85
BioBERT + XGBoost	0.87
DistilBERT	0.98

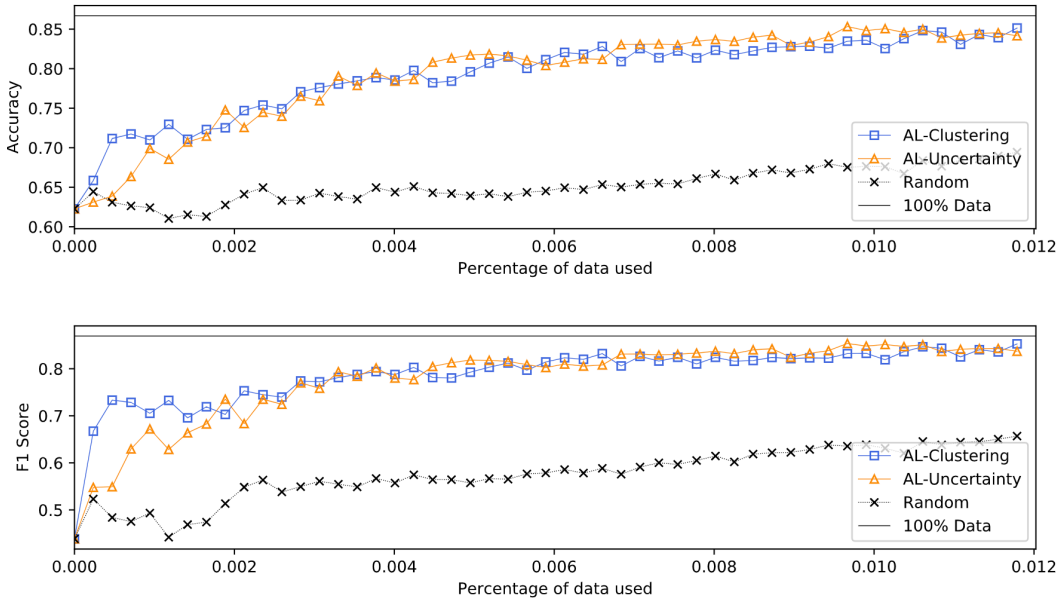
Table 3: F1 Score for three models trained on the entire dataset

[4]https://huggingface.co/

[5]https://github.com/dmis-lab/biobert



(a) XGBoost model with sentence embeddings from pre-trained bert-base-cased model



(b) XGBoost model with sentence embeddings from pre-trained biobert-base v1.1 model

Figure 4: Evaluation of the different AL strategies using accuracy and F1 score measures

the two XGBoost-based models, yet we would like to include the BERT-based classifiers in the experiments for our future work to compare the performance. Also, the XGBoost model with sentence embeddings produced by BioBERT improves the F1 score by 2% compared with the one using BERT since BioBERT is pre-trained on biomedical articles which aligns with the domain of our experiment dataset.

The curves in Figures 4(a)-(b) clearly show that the AL selection strategies outperform the random baseline for both sentence embedding settings. Indeed, we observe that the random strategy using 1.2% of the data achieves a similar performance compared to the AL strategies using less than 0.2% of the data which is a significant improvement since the AL strategies use much less annotated data. We also observe that both of the AL strategies achieve 99% of the achievable performance of the model using only 1.2% of the training dataset. Thus, one can clearly deduce that compared with the random strategy, AL-Uncertainty and AL-Clustering can achieve better performance with much less labeling effort, therefore, justifying our proposed method to obtain the gold data for the CORD-19 dataset using the UEA.

5 Conclusions and Future Work

In this work, we propose a novel strategy consisting of transformer and rule-based methods to generate QA pairs from scientific literature gathered in the CORD-19, while making use of a validation procedure to maintain the quality. We engage SMEs from the medical community and develop a web application to serve the purpose of providing an efficient user interface for annotating the QA pairs generated by our designed system. We also leverage active learning strategies to significantly reduce the required annotation effort from the SMEs.

This work paves the way for several interesting areas which can be explored further in the future. We believe that the engagement app, which will be released to the public soon, would enable the medical community to use it to its full extent as it can incorporate several subjective opinions from different SMEs and researchers. With the foundation laid by this work, we can also investigate better ways to explore the generation process of accurate questions by diving deeper into the task of question generation which is gaining attention in the field of NLP. In order to further benefit from our proposed method to improve the generalizability of the model using only a small annotated dataset, we can provide higher quality QA pairs by removing redundant questions using methods which are proven to work for graphical structures. These methods treat each scholarly article as a node which results in reducing the number of highly interlinked questions. Another plausible direction to explore is the incorporation of tasks such as extreme multi-label classification which would allow us to categorize scholarly articles under areas which may be better suited for annotations and align with the expertise of SMEs. Lastly, as explained earlier, generation of QA pairs in this work was limited to transformer and rule-based methods. We would like to explore the integration of successful neural based methods in our proposed approach.

Acknowledgments

The authors would like to thank all the organizers of the COVID-19 Open Research Dataset Kaggle Challenge. We would like to thank Vector Institute for making this collaboration possible and providing academic infrastructure and computing support during all phases of this work. We would also like to thank Richard Pito from Thomson Reuters for his invaluable feedback and support throughout this project. Last but not least, special thanks to Dr. Frank Rudzicz and Dr. Xiaodan Zhu for their academic supervision and insights.

References

Charu C. Aggarwal, Xiangnan Kong, Quanquan Gu, Jiawei Han, and Philip S. Yu. 2014. *Active learning: A survey*, pages 571–605. CRC Press.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. 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)*, pages 3606–3611.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3(null):993–1022.

Daniel Cer, Yinfei Yang, Sheng yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Universal sentence encoder.

Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA. ACM.

Hugh A. Chipman, Edward I. George, and Robert E. McCulloch. 2010. Bart: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1):266–298.

Lee Devlin, Chang and Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Xinya Du, Junru Shao, and Claire Cardie. 2017. Learning to ask: Neural question generation for reading comprehension. *CoRR*, abs/1705.00106.

Yifan Fu, Xingquan Zhu, and Bin Li. 2013. A survey on instance selection for active learning. *Knowledge and information systems*, 35(2):249–283.

Michael Heilman and Noah A. Smith. 2009. Question generation via overgenerating transformations and ranking.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2019. Clinicalbert: Modeling clinical notes and predicting hospital readmission.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. 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), pages 2567–2577.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

Neesha Jothi, Nur'Aini Abdul Rashid, and Wahidah Husain. 2015. Data mining in healthcare – a review. *Procedia Computer Science*, 72:306 – 313. The Third Information Systems International Conference 2015.

T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Y. Wu. 2002. An efficient k-means clustering algorithm: analysis and implementation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7):881–892.

Ksenia Konyushkova, Raphael Sznitman, and Pascal Fua. 2017. Learning active learning from data. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4225–4235. Curran Associates, Inc.

Kalpesh Krishna and Mohit Iyyer. 2019. Generating question-answer hierarchies. *CoRR*, abs/1906.02622.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Matthew Kelcey, Jacob Devlin, Kenton Lee, Kristina N. Toutanova, Llion Jones, Ming-Wei Chang, Andrew Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association of Computational Linguistics*.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations. In *International Conference on Learning Representations*.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

VI Levenshtein. 1966. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*, 10:707.

Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282.

David Oniani and Yanshan Wang. 2020. A qualitative evaluation of language models on automatic question-answering for covid-19.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text.

Reimers and Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How much knowledge can you pack into the parameters of a language model? *arXiv preprint arXiv:2002.08910*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter.

Sheng Shen, Yaliang Li, Nan Du, Xian Wu, Yusheng Xie, Shen Ge, Tao Yang, Kai Wang, Xingzheng Liang, and Wei Fan. 2020. On the generation of medical question-answer pairs. In *AAAI*, pages 8822–8829.

Z. Shuyang, T. Heittola, and T. Virtanen. 2018. An active learning method using clustering and committee-based sample selection for sound event classification. In *2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*, pages 116–120.

George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, Yannis Almirantis, John Pavlopoulos, Nicolas Baskiotis, Patrick Gallinari, Thierry Artieres, Axel Ngonga, Norman Heino, Eric Gaussier, Liliana Barrio-Alvers, Michael Schroeder, Ion Androutsopoulos, and Georgios Paliouras. 2015. An overview of the bioasq large-scale biomedical semantic indexing and question answering competition. *BMC Bioinformatics*, 16:138.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Jason Walonoski, Mark Kramer, Joseph Nichols, Andre Quina, Chris Moesel, Dylan Hall, Carlton Duffett, Kudakwashe Dube, Thomas Gallagher, and Scott McLachlan. 2018. Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. *Journal of the American Medical Informatics Association*, 25(3):230–238.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Darrin Eide, Kathryn Funk, Rodney Kinney, Ziyang Liu, William Merrill, Paul Mooney, Dewey Murdick, Devvret Rishi, Jerry Sheehan, Zhihong Shen, Brandon Stilson, Alex D. Wade, Kuansan Wang, Chris Wilhelm, Boya Xie, Douglas Raymond, Daniel S. Weld, Oren Etzioni, and Sebastian Kohlmeier. 2020. Cord-19: The covid-19 open research dataset.

Wei Yang, Yuqing Xie, Aileen Lin, Xingyu Li, Luchen Tan, Kun Xiong, Ming Li, and Jimmy Lin. 2019. End-to-end open-domain question answering with bertserini. *CoRR*, abs/1902.01718.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.