NLM at BioASQ Synergy 2021: Deep Learning-based Methods for Biomedical Semantic Question Answering about COVID-19

Mourad Sarrouti, Deepak Gupta, Asma Ben Abacha and Dina Demner-Fushman

U.S. National Library of Medicine, National Institutes of Health, Bethesda, MD, USA

Abstract

The COVID-19 outbreak has heightened the need for systems that enable information seekers to search vast corpora of scientific articles to find answers to their natural language questions. This paper describes the participation of the U.S. National Library of Medicine (NLM) team in BioASQ Task Synergy on biomedical semantic question answering for COVID-19. In this work, we exploited the pre-trained Transformer models such as T5 and BART for document re-ranking, passage retrieval, and answer generation. Official results show that among the participating systems, our models achieve strong performance in document retrieval, passage retrieval, and the "ideal answer" generation task.

Keywords

Question Answering, Document Retrieval, Passage Retrieval, Answer Extraction, Natural Language Processing, Deep Learning, COVID-19, BioASQ

1. Introduction

The global response to COVID-19 has yielded thousands of new scientific articles about COVID-19 and other related topics [1, 2]. The COVID-19 outbreak has emphasized the need for sophisticated systems that enable querying large volumes of scientific articles to find answers to questions expressed in natural language. Therefore, to provide information seekers with relevant and precise information about COVID-19, more sophisticated and specialized tools are needed [3, 4]. Question Answering (QA), aiming at answering natural language questions from textual documents, is a potential approach that could help information seekers to identify the precise information readily [5, 6, 7, 8].

This paper presents the participation of the U.S. National Library of Medicine (NLM) team in BioASQ[1] Task Synergy on Biomedical Semantic QA for COVID-19. For given COVID-19 related questions, this task aims at **(1)** retrieving the relevant documents, **(2)** retrieving the most relevant passages, and **(3)** extracting/generating the exact and ideal answers from a corpus of scientific articles. To address these problems, we exploited natural language processing techniques and pre-trained language models for document retrieval, passage retrieval, and

CLEF 2021 – Conference and Labs of the Evaluation Forum, September 21–24, 2021, Bucharest, Romania

mourad.sarrouti@nih.gov (M. Sarrouti); deepak.gupta@nih.gov (D. Gupta); asma.benabacha@nih.gov (A. Ben Abacha); ddemner@mail.nih.gov (D. Demner-Fushman)

0000-0002-3739-4192 (M. Sarrouti)

[1] http://www.bioasq.org/

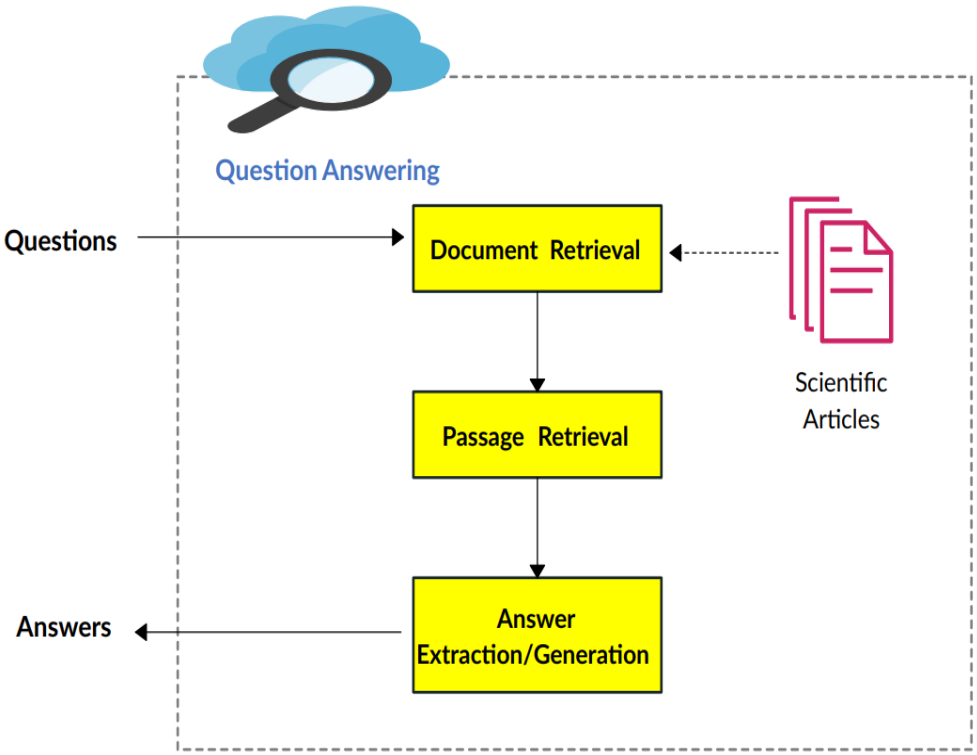

Figure 1: The pipeline of our QA system

answer extraction/generation. Figure 1 shows the pipeline of our proposed QA system. We first index the COVID-19 Open Research Dataset (CORD-19) and retrieve the top-n relevant documents for each question using BM25 as a retrieval model. We then re-rank the retrieved documents using the Text-to-Text Transfer Transformer (T5) relevance-based re-ranking model, and select the top-k documents. Once the k top-ranked documents are retrieved, we then retrieve the relevant passages using T5 as a re-ranker model. We finally extract and generate the "ideal answers" (i.e., a paragraph-sized summary of relevant information,) using T5 and BART models.

The rest of the paper is organized as follows: Section 2 presents the most relevant prior work and describes the datasets used in BioASQ Task Synergy. Section 3 presents our systems for document retrieval, passage retrieval, and "ideal answer" extraction/generation. Official results for all models are presented in Section 4. Finally, the paper is concluded in Section 5.

2. Related Work

- **Document Retrieval:** Neural-based models have shown promising results in a variety of IR tasks. Xiong et al. [9] developed a kernel pooling technique by customizing word embeddings that learn to encode the relevance preferences. This approach was further enhanced by Dai et al. [10] who proposed a convolutional model to consider n-gram

representations of the word. Traditional models, such as BM25 and query likelihood are known to be successful retrieval models [11]. These models are based on the exact matching of query and document words, which might limit the available information for the ranking model, which, in turn, may lead to a vocabulary mismatch issue. Models for statistical translation have tried to overcome this limitation. They model the relevance of query documents with a pre-computed translation matrix describing the similarities between word pairs. Zamani et al. [12] accentuated the effectiveness of neural ranking models and developed a neural model to retrieve documents from a very large dataset. Recently, the pre-trained Transformer models (such as BERT) have also demonstrated their efficacy in ranking tasks. Nogueira and Cho [13] showed that the BERT model was highly effective in the passage re-ranking task on the MS-MARCO and TREC CAR [14] datasets. MacAvaney et al. [15], Yang et al. [16] utilized the BERT model to predict the answer spans for a given question. Other studies have also explored BERT-based representations for document ranking.

- **Extractive Summarization:** The recent progress in the development of neural models and pre-trained Transformer models has led to significant growth in extractive document summarization [17]. The majority of the existing summarization models are built upon sequence-to-sequence frameworks [18, 19, 20], recurrent neural networks [20, 21], and Transformers [22, 23]. Cheng and Lapata [18] and Nallapati et al. [24] developed approaches that aim to decide whether a given sentence will qualify for the summary or not. Nallapati et al. [20] proposed SummaRuNNer that adds more lexical features to the sequence-to-sequence model. First, SummaRuNNer predicts the extraction probability score for each sentence, and then it performs sentence selection to select the top sentences for the summary. Chen and Bansal [25] followed a similar line of study and exploited the pointer generator network to sequentially select sentences from the document to generate a summary. Other decoding techniques, such as ranking [26], have also been utilized for content selection. Recently, several studies have explored pre-trained language models in summarization for contextual word representations [27, 23].
- **Abstractive Summarization:** The availability of large-scale training data has boosted the development of abstractive summarization techniques in the open domain. Rush et al. [28] proposed a sequence-to-sequence model with attention for abstractive sentence summarization. Later, Li et al. [29] utilized the sequence-to-sequence models in multi-sentence document summarization. Nallapati et al. [30] utilized the copy-mechanism to generate or copy words either from the source document or vocabulary. See et al. [31] introduced the coverage mechanism in the pointer generator network to generate non-hallucinated summaries. Few other works [32, 33] have proposed different techniques to generate factually-correct summaries. Studies conducted by Falke et al. [34], Kryściński et al. [35], Wang et al. [36] have utilized the natural language inference and question answering tasks to obtain factually-correct summaries. Other methods [37, 38, 39, 40, 41] based on reinforcement learning (RL) were developed to improve the quality of the generated summaries. Pasunuru and Bansal [38] proposed RL-based optimization on the modified version of the ROUGE score that considers readability. Zhang and Bansal [39] addressed the semantic drift issue in question generation, proposing question-paraphrase and question-answering probability rewards. Yadav et al. [42] introduced question-focus

and question-type based semantic rewards that enforce the model to generate semantically valid and factually correct question summaries.
Recently, abstractive summarization was used for the summarization of various medical and clinical texts, such as radiology reports [41, 43, 44], consumer health questions and medical answers [45, 46, 47, 48, 49], and biomedical documents [50].

3. Approach

3.1. Background

BM25. BM25 algorithm [51] is a bag-of-words retrieval function that ranks a set of documents based on the query terms appearing in each document. The BM25 score between a query term $Q = \{w_1, w_2, \ldots, w_n\}$ and document D is computed as:

$$\text{Score}(D, Q) = \sum_{i=1}^{n} \text{IDF}(w_i) \cdot \frac{f(w_i, D) \cdot (k_1 + 1)}{f(w_i, D) + k_1 \cdot (1 - b + b \cdot \frac{|D|}{\text{avgdl}})} \tag{1}$$

where $f(w_i, D)$ is w_i's term frequency in the document D, $|D|$ is the length of the document (in words), and $avgdl$ is the average document length in the document set. k_1 and b are the hyperparamerts.

$$\text{IDF}(w_i) = log \frac{N - n(w_i) + 0.5}{n(w_i) + 0.5} \tag{2}$$

where N is the total number of candidate documents, $n(w_i)$ is the number of document containing w_i.

Text-to-Text Transfer Transformer (T5). This is a pre-trained model developed by Raffel et al. [52] who explored the transfer learning techniques for NLP by introducing a unified framework that converts all text-based language problems into a text-to-text format. This approach is inspired by previous unifying frameworks for NLP tasks, including casting all text problems as question answering [53] or language modeling [54]. The T5 model is an Encoder-Decoder Transformer with some architectural changes discussed in detail in Raffel et al. [52].

Bidirectional and Auto-Regressive Transformers (BART). BART [55] is a denoising autoencoder built with a sequence-to-sequence model. Due to its bidirectional encoder and left-to-right decoder, it can be considered as generalizing BERT [56] and GPT [54], respectively. BART pretraining has two stages: (1) a noising function is used to corrupt the input text, and (2) a sequence-to-sequence model is learned to reconstruct the original input text.

3.2. Document Retrieval

For a given question, the document retrieval task at BioASQ Synergy aims at retrieving a list of 10 most relevant scientific articles $(d_1, d_2, ..., d_{10})$ from the COVID-19 Open Research Dataset

(CORD-19). To address this challenge, we first retrieved the relevant scientific articles from the CORD-19 collection using the BM25 model and the Terrier[2] search engine. We then re-ranked the top-1000 documents with the Text-to-Text Transfer Transformer (T5) [52] relevance-based re-ranking model and selected the top-10 relevant articles. T5 with traditional Transformer architecture and BERT's masked language modeling [56], was shown to be effective on newswire retrieval and MS MARCO [57]. In contrast to BERT that is pre-trained on a Masked LM (MLM) and Next Sentence Prediction (NSP) objective and then, fine-tuned on specific tasks, the T5 model casts all natural language processing tasks (e.g. natural language inference, question answering) into a text-to-text format. We adopted the T5 approach to document re-ranking by using the following input sequence:

$$Question : q \; Document : \; d \; Relevant : \tag{3}$$

The T5 model was fine-tuned on (1) MS MARCO passage ranking dataset [58] and (2) TREC-COVID[3] dataset by maximizing the log probability of generating the output token "true" when the document is relevant, and the token "false" when the document is not relevant to the query [57]. Once fine-tuned, we first apply a softmax only on the logits of the "true" and "false" generated tokens, and then re-rank the documents using the probabilities of the "true" token. More details about this approach appear in [57].

3.3. Passage Retrieval

The passage retrieval task at BioASQ Synergy consists of retrieving a set of at most 10 relevant text passages/snippets $(p_1, p_2, ..., p_{10})$ from the abstracts or titles of the documents returned by the document retrieval method. To address this problem, we used the T5 relevance-based re-ranking model [52] that we also used for document re-ranking. To do so, we first split the abstracts of the documents retrieved for a given question into sentences/chunks (i.e. passages) using NLTK[4], and then ranked these passages based on the relevance score that determined how relevant a candidate passage was to the question. The passages were ranked by a pointwise re-ranker that used T5. We adapted the T5 approach presented in the previous section (cf. Section 3.2) to passage re-ranking by using the following input sequence:

$$Question : q \; Sentence : \; S \; Relevant : \tag{4}$$

We first applied a softmax only on the logits of the "true" and "false"tokens generated by T5 that was fine-tuned on MS MARCO and TREC-COVID datasets. We then re-ranked the passages/snippets using the probabilities of the "true" tokens.

3.4. Ideal Answer Generation

The "ideal answer" is defined as a single paragraph-sized text summarizing the most relevant information from the passages. To generate the ideal answer for a given question in BioASQ

[2] http://terrier.org/
[3] https://ir.nist.gov/covidSubmit/data.html
[4] https://www.nltk.org/

Synergy, we explored extractive and abstractive summarization approaches based on pretrained language models.

1. **Extractive approach.** We formed the ideal answer to a question by rejoining the selected top-3 passages returned for the passage retrieval task by the T5 relevance-based re-ranking model.
2. **Abstractive approach.** We utilized the COVID-19 Open Research Dataset (CORD-19) [1] to fine-tune the BART model. We trained the answer summarization model by considering various sections of the biomedical article as the Source and the article's abstract as the Target.

3.5. Additional Datasets

Document and passage retrieval. For the document and passage retrieval tasks, we used the following datasets to fine-tune the T5 model:

- **MS MARCO Passage** [58] is a large dataset for passage ranking. It contains 8.8M passages retrieved by the Bing search engine for around 1M natural language questions.
- **TREC-COVID** [59] is a large test collection created to evaluate ad-hoc retrieval of documents relevant to COVID-19[5].

Ideal answer generation.

- **CORD-19** [1] is a collection of scientific papers on COVID-19 and related coronavirus research. These scientific papers are processed to remove the duplicate entries and collect the relevant metadata. The rich collection of these structured data is used to develop the text-mining and information retrieval systems.

3.6. Evaluation metrics

The performance of the document retrieval and passage retrieval systems was evaluated using the typical evaluation measures used in information retrieval: mean precision, mean recall, mean F-measure, mean average precision (MAP) and geometric mean average precision (GMAP). The ideal answers were automatically evaluated using ROUGE-2 and ROUGE-SU4. Detailed descriptions of these evaluation metrics appear in [60]. The BioASQ challenge also provided manual scores in terms of readability, recall, precision, and repetition for the ideal answers.

4. Experimental Results and Discussion

Document retrieval. We submitted the following runs for the document retrieval task:

1. **NLM-1** : In this run, we fine-tuned T5 on the MS MARCO passage ranking dataset.

[5] https://ir.nist.gov/covidSubmit/data.html

Table 1
Official results of BioASQ Task Synergy: NLM runs for the document retrieval task. Our Best run and the best participants' run are selected based on the MAP metric.

Test set	System	Mean precision	Recall	F-Measure	MAP	GMAP
Batch 1	NLM-1	0.4773	0.3251	0.3383	0.2946	0.0459
	NLM-4	0.4438	0.3310	0.3078	0.2735	0.0635
	Our Best Run	0.4773	0.3251	0.3383	0.2946	0.0459
	Best Participants	0.4963	0.3795	0.3457	0.3375	0.0829
	Average Participants	0.3653	0.27615	0.2516	0.2420	0.0321
Batch 2	NLM-1	0.3500	0.3360	0.2762	0.3179	0.0714
	NLM-4	0.3088	0.2854	0.2387	0.2845	0.0556
	Our Best Run	0.3500	0.3360	0.2762	0.3179	0.0714
	Best Participants	0.4039	0.4108	0.3205	0.4069	0.1586
	Average Participants	0.2940	0.2874	0.2294	0.2829	0.0520
Batch 3	NLM-1	0.2977	0.3177	0.2378	0.2489	0.0418
	NLM-4	0.2523	0.2687	0.2015	0.2008	0.0186
	Our Best Run	0.2977	0.3177	0.2378	0.2489	0.0418
	Best Participants	0.3451	0.3226	0.2628	0.3257	0.0484
	Average Participants	0.2192	0.2100	0.1640	0.1861	0.0183
Batch 4	NLM-1	0.2604	0.2752	0.2124	0.2294	0.0302
	NLM-4	0.2473	0.2465	0.1983	0.1956	0.0318
	Our Best Run	0.2604	0.2752	0.2124	0.2294	0.0302
	Best Participants	0.3027	0.3169	0.2375	0.2983	0.0573
	Average Participants	0.2322	0.2187	0.1758	0.1990	0.0227

2. **NLM-4** : For this run, we first fine-tuned T5 on the MS MARCO passage ranking dataset and then TREC-COVID.

We have shown the detailed performance evaluation based on different metrics in Table 1. We achieved the best results with our NLM-1 run in all batches. The in-domain dataset (TREC-COVID) did not help to improve the performance of T5 in NLM-4 run. This is mainly due to the limited number of queries in TREC-COVID.

Passage retrieval. We submitted the following runs for the passage retrieval task:

1. **NLM-1** : In this run, we fine-tuned T5 on the MS MARCO passage ranking dataset. We considered the NLTK sentence length as a passage length.
2. **NLM-2** : For this run, we fine-tuned T5 on the MS MARCO passage ranking dataset. We considered a chunk of two sentences as a passage length.
3. **NLM-3** : This run for batch #2, #3 and #4 is similar to the NLM-2 run for the batch #1. For batch #1, NLM-3 is similar to the NLM-4 run.
4. **NLM-4** : In this run, we first fine-tuned T5 on the MS MARCO passage ranking dataset and then TREC-COVID. We considered the NLTK sentence length as a passage length.
5. **NLM-5** : We first fine-tuned T5 on the MS MARCO passage ranking dataset and then TREC-COVID. We considered a chunk of two sentences as a passage length.

Table 2
Official results of BioASQ Task Synergy: NLM runs for the passage retrieval task. Our Best run and the best participants' run are selected based on the MAP metric.

Test set	System	Mean precision	Recall	F-Measure	MAP	GMAP
Batch 1	NLM-1	0.3927	0.1798	0.2153	0.2676	0.0206
	NLM-2	0.4157	**0.2584**	**0.2712**	0.2107	0.0197
	NLM-3	0.3557	0.1714	0.1903	0.2652	0.0176
	NLM-4	0.3608	0.2355	0.2315	0.2068	0.0190
	Our Best Run	0.3927	0.1798	0.2153	0.2676	0.0206
	Best Participants	0.4248	0.2008	0.2194	0.3127	0.0307
	Average Participants	0.3177	0.1660	0.1762	0.2279	0.0142
Batch 2	NLM-1	0.2685	0.1688	0.1634	0.2422	0.0193
	NLM-3	0.2523	**0.2265**	**0.1885**	0.2043	0.0177
	NLM-4	0.2172	0.1230	0.1246	0.1991	0.0106
	NLM-5	0.2154	0.1442	0.1409	0.1574	0.0065
	Our Best Run	0.2685	0.1688	0.1634	0.2422	0.0193
	Best Participants	0.2981	0.1992	0.1858	0.3201	0.0349
	Average Participants	0.2059	0.1393	0.1283	0.2032	0.0151
Batch 3	NLM-1	0.2459	0.1808	0.1645	0.2378	0.0147
	NLM-3	0.2426	**0.2408**	0.1940	0.1722	0.0145
	NLM-4	0.1962	0.1428	0.1280	0.1859	0.0071
	NLM-5	0.1840	0.1685	0.1339	0.1306	0.0041
	Our Best Run	0.2459	0.1808	0.1645	0.2378	0.0147
	Best Participants	0.2986	0.2297	0.2026	0.3186	0.0351
	Average Participants	0.1978	0.1550	0.1331	0.1926	0.0138
Batch 4	NLM-1	0.2225	0.2045	0.1703	0.2219	0.0136
	NLM-3	0.2228	**0.2455**	**0.1909**	0.1582	0.0087
	NLM-4	0.1804	0.1333	0.1268	0.1689	0.0063
	NLM-5	0.1869	0.1700	0.1461	0.1363	0.0061
	Our Best Run (MAP)	0.2225	0.2045	0.1703	0.2219	0.0136
	Best Participants	0.2453	0.2229	0.1826	0.2842	0.0210
	Average Participants	0.1685	0.1450	0.1229	0.1604	0.0082

The results obtained by our submissions and the best participants' results are shown in Table 2. In terms of MAP and GMAP, our NLM-1 run achieved the best performance among our submissions on all testing batches. NLM-3 achieves the best recall and F1 scores on all batches. We note that the NLM-2 run in batch #1 is similar to the NLM-3 in batch #2, #3, and #4. The results showed that the passage length has an impact on the performance of our passage retrieval models. As in the document retrieval task, we found that the in-domain dataset (TREC-COVID) did not improve the performance for the passage retrieval task.

Ideal answer extraction/generation. We submitted the following runs for the ideal answer extraction/generation task:

1. **NLM-1** : In this run, we form the summary by rejoining the top-2 ranked passages returned by the NLM-1 run of the passage retrieval task.

Table 3
Automatic scores of NLM runs at the "ideal answer" generation in BioASQ Task Synergy. Our Best run and the best participants' run are selected based on the R-SU4 (F1) metric.

Test set	System	R-2 (Rec)	R-2 (F1)	R-SU4 (Rec)	R-SU4 (F1)
Batch 2	NLM-1	0.0934	0.0669	0.1047	0.0720
	NLM-2	0.0554	0.0423	0.0681	0.0495
	NLM-3	**0.0956**	0.0690	**0.1080**	0.0743
	NLM-4	0.0289	0.0197	0.0389	0.0266
	NLM-5	0.0437	0.0304	0.0548	0.0376
	Our Best Run	0.0956	0.0690	0.1080	0.0743
	Best Participants	0.0758	0.0726	0.0779	0.0749
	Average Participants	0.0506	0.0421	0.0572	0.0467
Batch 3	NLM-1	0.1039	0.0709	0.1150	0.0778
	NLM-2	0.0809	0.0551	0.0926	0.0631
	NLM-3	0.0881	0.0622	0.0996	0.0685
	NLM-4	0.0365	0.0252	0.0488	0.0341
	NLM-5	0.0593	0.0437	0.0707	0.0518
	Our Best Run	0.1039	0.0709	0.1150	0.0778
	Best Participants	0.1120	0.1139	0.1150	0.1170
	Average Participants	0.0808	0.0678	0.0891	0.0737
Batch 4	NLM-1	0.1119	0.0854	0.1220	0.0916
	NLM-3	0.0948	0.0711	0.1077	0.0787
	NLM-2	0.0733	0.0581	0.0840	0.0659
	NLM-4	0.0380	0.0265	0.0513	0.0364
	NLM-5	0.0604	0.0459	0.0737	0.0562
	Our Best Run	0.1119	0.0854	0.1220	0.0916
	Best Participants	0.1169	0.1215	0.1208	0.1254
	Average Participants	0.0849	0.0723	0.0938	0.0790

2. **NLM-2** : For this run, we use BART to generate a summary from the set of passages returned by the NLM-1 run of the passage retrieval task. The BART model is fine-tuned by considering the introduction, conclusion, and results sections of the scientific articles in the CORD-19 dataset as the Source and the abstract as the Target.
3. **NLM-3** : We form the summary by rejoining the top-2 ranked passages returned by the NLM-2 run of the passage retrieval task.
4. **NLM-4** : The BART model is used to generate the summary from the set of passages returned by the NLM-4 run of the passage retrieval task. It is fine-tuned by considering the introduction and discussion sections of the scientific articles in the CORD-19 dataset as the Source and the abstract as the Target.
5. **NLM-5** : The summary is generated by BART which is fine-tuned by considering all sections of the CORD-19 scientific articles (except the abstracts) as the Source and the abstract as the Target. It is generated from the passages that were retrieved by the NLM-5 run in the passage retrieval task.

Table 4
Manual scores of NLM runs at the "ideal answer" generation in BioASQ Task Synergy.

Test set	System	Readability	Recall	Precision	Repetition
Batch 2	NLM-1	3.51	3.62	3.36	3.34
	NLM-2	2.91	3.00	2.94	3.66
	NLM-3	3.51	**3.68**	3.55	3.58
	NLM-4	2.45	1.40	2.00	3.26
	NLM-5	3.09	2.96	3.15	3.47
	Our Best Run	3.51	3.68	3.55	3.58
	Best Participants	3.92	3.38	3.75	3.64
	Average Participants	2.83	2.41	2.61	2.92
Batch 3	NLM-1	3.54	3.50	3.09	3.71
	NLM-2	3.11	3.08	2.83	3.70
	NLM-3	3.45	3.51	3.06	3.64
	NLM-4	2.70	1.43	1.93	3.23
	NLM-5	3.28	2.91	2.69	3.70
	Our Best Run	3.54	3.50	3.09	3.71
	Best Participants	4.39	3.94	4.00	4.41
	Average Participants	3.46	3.10	3.06	3.70
Batch 4	NLM-1	3.27	3.27	3.02	3.43
	NLM-3	3.12	3.16	2.80	3.16
	NLM-2	2.86	2.66	2.64	3.33
	NLM-4	2.57	1.36	1.81	3.01
	NLM-5	2.93	2.70	2.70	3.27
	Our Best Run	3.27	3.27	3.02	3.43
	Best Participants	3.76	3.42	3.42	3.71
	Average Participants	3.16	2.81	2.79	3.35

Table 3 and Table 4 present the automatic and manual scores of the ideal answer generation task. For Batch #2 of the "ideal generation" task, we obtained the best results across all the evaluation metrics with our NLM-3 run. Similarly, for Batch #3 and Batch #4 our NLM-1 run outperformed the remaining runs across all the evaluation metrics. We observe that the extractive summary generation approach (rejoining the top-k ranked passages returned in the passage retrieval task) performed better than the abstractive summary generation approach across all the test batches. The NLM-2 run, has shown better performance across all the metrics amongst all the abstractive runs: NLM-2, 4 and 5. Table 5 presents examples of extractive and abstractive summaries.

5. Conclusion

In this paper, we described our participation in Task Synergy at BioASQ 2021 that aims to answer questions about COVID-19 using scientific articles. We explored the T5-relevance-based

Table 5
Examples of extractive and abstractive summaries.

Question	Extractive summary	Abstractive summary
Describe the role of neuropilin-1 (NRP1) in COVID-19	Neuropilin-1 (NRP-1) is a multifunctional transmembrane receptor for ligands that affect developmental axonal growth and angiogenesis. In addition to a role in cancer, NRP-1 is a reported entry point for several viruses, including severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), the causal agent of coronavirus disease 2019 (COVID-19). In-silico studies were carried out to understand the role of its bioactive constituents in COVID-19 treatment and prevention. Firstly, the disease network was prepared by using ACE2 (Angiotensin-II receptor), as it is the entry site for virus.	Neuropilin-1 (NRP-1) is a multifunctional transmembrane receptor for ligands that affect developmental axonal growth and angiogenesis. In addition to a role in cancer, neuropilins, heparan sulfate and sialic acids and the putative alternative receptors, such as CD147 and GRP78, are reported entry points for several viruses, including Severe Acute Respiratory Syndrome-related Coronavirus-2 (SARS-CoV-2), the causal agent of coronavirus disease 2019 (COVID-19
What Covid-19 viral protein or proteins do the vaccines target?	Our study proposes a detailed and comprehensive immunoinformatic approach that can be applied to the currently available coronavirus protein data in the online server for vaccine candidate development. We have identified the receptor binding domain (RBD) of structural spike protein (S1) as a potential target for immunity against COVID- 19 infection. To develop vaccine, we target S- protein, expressed on the virus surface plays important role in COVID-19 infection. We identified 12 B-cell, 9 T-helper and 20 Cytotoxic T-cell epitope based on criteria of selection.	The ongoing COVID-19 pandemic caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has resulted in more than 7,000,000 infections and 400,000 deaths worldwide to date. A key target of these efforts is the spike (S) protein, a large trimeric class I fusion protein that mediates the host cell entry by binding to the angiotensin-converting enzyme 2 (ACE2). In this study, immunoinformatics approach was employed to design a novel multi-epitope vaccine using receptor-binding domain (RBD) of S

re-ranking model for document and passage retrieval. We also exploited T5 and BART for extracting and generating "ideal answers". The official results show that our models achieve strong performance compared to the participants' systems. We found that augmenting the training data with relevance judgments obtained from related TREC-COVID tasks did not improve the performance of our systems in the passage retrieval task. We also found that extractive summarization performed better than abstractive summarization for the generation of ideal answers. In the future, we would like to explore suitable datasets and techniques for abstractive summarization to improve the performance of the ideal answer generation task.

Acknowledgments

This work was supported by the intramural research program at the U.S. National Library of Medicine, National Institutes of Health.

References

[1] L. L. Wang, K. Lo, Y. Chandrasekhar, R. Reas, J. Yang, D. Eide, K. Funk, R. Kinney, Z. Liu, W. Merrill, et al., Cord-19: The covid-19 open research dataset, ArXiv (2020).

[2] Y. Mrabet, M. Sarrouti, A. B. Abacha, S. Gayen, T. Goodwin, A. Rae, W. Rogers, D. Demner-Fushman, Nlm at trec 2020 health misinformation and deep learning tracks (2020).

[3] C. Wise, M. Romero Calvo, P. Bhatia, V. Ioannidis, G. Karypus, G. Price, X. Song, R. Brand, N. Kulkani, COVID-19 knowledge graph: Accelerating information retrieval and discovery for scientific literature, in: Proceedings of Knowledgeable NLP: the First Workshop on Integrating Structured Knowledge and Neural Networks for NLP, Association for Computational Linguistics, Suzhou, China, 2020, pp. 1–10. URL: https://www.aclweb.org/anthology/2020.knlp-1.1.

[4] Z.-H. Lu, J. X. Wang, X. Li, Revealing opinions for COVID-19 questions using a context retriever, opinion aggregator, and question-answering model: Model development study, Journal of Medical Internet Research 23 (2021) e22860. URL: https://doi.org/10.2196%2F22860. doi:10.2196/22860.

[5] M. Sarrouti, A. Lachkar, A new and efficient method based on syntactic dependency relations features for ad hoc clinical question classification, International Journal of Bioinformatics Research and Applications 13 (2017) 161–177.

[6] M. Sarrouti, S. O. E. Alaoui, A passage retrieval method based on probabilistic information retrieval model and UMLS concepts in biomedical question answering, Journal of Biomedical Informatics 68 (2017) 96–103. URL: https://doi.org/10.1016%2Fj.jbi.2017.03.001. doi:10.1016/j.jbi.2017.03.001.

[7] M. Sarrouti, S. O. E. Alaoui, SemBioNLQA: A semantic biomedical question answering system for retrieving exact and ideal answers to natural language questions, Artificial Intelligence in Medicine 102 (2020) 101767. URL: https://doi.org/10.1016%2Fj.artmed.2019.101767. doi:10.1016/j.artmed.2019.101767.

[8] E.-d. El-allaly, M. Sarrouti, N. En-Nahnahi, S. O. El Alaoui, Mttlade: A multi-task transfer learning-based method for adverse drug events extraction, Information Processing & Management 58 (2021) 102473.

[9] C. Xiong, Z. Dai, J. Callan, Z. Liu, R. Power, End-to-end neural ad-hoc ranking with kernel pooling, in: Proceedings of the 40th International ACM SIGIR conference on research and development in information retrieval, 2017, pp. 55–64.

[10] Z. Dai, Z. Yang, Y. Yang, J. Carbonell, Q. V. Le, R. Salakhutdinov, Transformer-xl: Attentive language models beyond a fixed-length context, arXiv preprint arXiv:1901.02860 (2019).

[11] R. Aly, T. Demeester, S. Robertson, Probabilistic models in IR and their relationships, Inf. Retr. 17 (2014) 177–201. URL: https://doi.org/10.1007/s10791-013-9226-3. doi:10.1007/s10791-013-9226-3.

[12] H. Zamani, M. Dehghani, W. B. Croft, E. Learned-Miller, J. Kamps, From neural re-ranking to neural ranking: Learning a sparse representation for inverted indexing, in: Proceedings of the 27th ACM international conference on information and knowledge management, 2018, pp. 497–506.
[13] R. Nogueira, K. Cho, Passage re-ranking with bert, arXiv preprint arXiv:1901.04085 (2019).
[14] L. Dietz, M. Verma, F. Radlinski, N. Craswell, Trec complex answer retrieval overview., in: TREC, 2017.
[15] S. MacAvaney, A. Yates, A. Cohan, N. Goharian, Cedr: Contextualized embeddings for document ranking, in: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2019, pp. 1101–1104.
[16] W. Yang, H. Zhang, J. Lin, Simple applications of bert for ad hoc document retrieval, arXiv preprint arXiv:1903.10972 (2019).
[17] S. Yadav, M. Sarrouti, D. Gupta, Nlm at mediqa 2021: Transfer learning-based approaches for consumer question and multi-answer summarization, in: Proceedings of the 20th Workshop on Biomedical Language Processing, 2021, pp. 291–301.
[18] J. Cheng, M. Lapata, Neural summarization by extracting sentences and words, in: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2016, pp. 484–494.
[19] A. Jadhav, V. Rajan, Extractive summarization with swap-net: Sentences and words from alternating pointer networks, in: Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: Long papers), 2018, pp. 142–151.
[20] R. Nallapati, F. Zhai, B. Zhou, Summarunner: A recurrent neural network based sequence model for extractive summarization of documents, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 31, 2017.
[21] Q. Zhou, N. Yang, F. Wei, S. Huang, M. Zhou, T. Zhao, Neural document summarization by jointly learning to score and select sentences, arXiv preprint arXiv:1807.02305 (2018).
[22] M. Zhong, D. Wang, P. Liu, X. Qiu, X. Huang, A closer look at data bias in neural extractive summarization models, arXiv preprint arXiv:1909.13705 (2019).
[23] Y. Liu, M. Lapata, Text summarization with pretrained encoders, arXiv preprint arXiv:1908.08345 (2019).
[24] R. Nallapati, B. Zhou, M. Ma, Classify or select: Neural architectures for extractive document summarization, arXiv preprint arXiv:1611.04244 (2016).
[25] Y.-C. Chen, M. Bansal, Fast abstractive summarization with reinforce-selected sentence rewriting, arXiv preprint arXiv:1805.11080 (2018).
[26] S. Narayan, S. B. Cohen, M. Lapata, Ranking sentences for extractive summarization with reinforcement learning, arXiv preprint arXiv:1802.08636 (2018).
[27] M. Zhong, P. Liu, D. Wang, X. Qiu, X. Huang, Searching for effective neural extractive summarization: What works and what's next, arXiv preprint arXiv:1907.03491 (2019).
[28] A. M. Rush, S. Chopra, J. Weston, A neural attention model for abstractive sentence summarization, arXiv preprint arXiv:1509.00685 (2015).
[29] W. Li, X. Xiao, Y. Lyu, Y. Wang, Improving neural abstractive document summarization with explicit information selection modeling, in: Proceedings of the 2018 conference on empirical methods in natural language processing, 2018, pp. 1787–1796.
[30] R. Nallapati, B. Zhou, C. Gulcehre, B. Xiang, et al., Abstractive text summarization using

sequence-to-sequence rnns and beyond, arXiv preprint arXiv:1602.06023 (2016).
[31] A. See, P. J. Liu, C. D. Manning, Get to the point: Summarization with pointer-generator networks, arXiv preprint arXiv:1704.04368 (2017).
[32] L. Lebanoff, J. Muchovej, F. Dernoncourt, D. S. Kim, L. Wang, W. Chang, F. Liu, Understanding points of correspondence between sentences for abstractive summarization, arXiv preprint arXiv:2006.05621 (2020).
[33] L. Huang, L. Wu, L. Wang, Knowledge graph-augmented abstractive summarization with semantic-driven cloze reward, arXiv preprint arXiv:2005.01159 (2020).
[34] T. Falke, L. F. Ribeiro, P. A. Utama, I. Dagan, I. Gurevych, Ranking generated summaries by correctness: An interesting but challenging application for natural language inference, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 2214–2220.
[35] W. Kryściński, B. McCann, C. Xiong, R. Socher, Evaluating the factual consistency of abstractive text summarization, arXiv preprint arXiv:1910.12840 (2019).
[36] A. Wang, K. Cho, M. Lewis, Asking and answering questions to evaluate the factual consistency of summaries, arXiv preprint arXiv:2004.04228 (2020).
[37] R. Paulus, C. Xiong, R. Socher, A deep reinforced model for abstractive summarization, arXiv preprint arXiv:1705.04304 (2017).
[38] R. Pasunuru, M. Bansal, Multi-reward reinforced summarization with saliency and entailment, arXiv preprint arXiv:1804.06451 (2018).
[39] S. Zhang, M. Bansal, Addressing semantic drift in question generation for semi-supervised 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), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 2495–2509. URL: https://www.aclweb.org/anthology/D19-1253. doi:10.18653/v1/D19-1253.
[40] D. Gupta, H. Chauhan, R. T. Akella, A. Ekbal, P. Bhattacharyya, Reinforced multi-task approach for multi-hop question generation, in: Proceedings of the 28th International Conference on Computational Linguistics, 2020, pp. 2760–2775.
[41] Y. Zhang, D. Merck, E. Tsai, C. D. Manning, C. Langlotz, Optimizing the factual correctness of a summary: A study of summarizing radiology reports, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 5108–5120.
[42] S. Yadav, D. Gupta, A. B. Abacha, D. Demner-Fushman, Reinforcement learning for abstractive question summarization with question-aware semantic rewards, arXiv preprint arXiv:2107.00176 (2021). arXiv:2107.00176.
[43] Y. Zhang, D. Y. Ding, T. Qian, C. D. Manning, C. P. Langlotz, Learning to summarize radiology findings, in: Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis, 2018, pp. 204–213.
[44] Y. Li, X. Liang, Z. Hu, E. P. Xing, Hybrid retrieval-generation reinforced agent for medical image report generation, in: Advances in neural information processing systems, 2018, pp. 1530–1540.
[45] A. Ben Abacha, D. Demner-Fushman, On the summarization of consumer health questions, in: A. Korhonen, D. R. Traum, L. Màrquez (Eds.), Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August

2, 2019, Volume 1: Long Papers, Association for Computational Linguistics, 2019, pp. 2228–2234. URL: https://doi.org/10.18653/v1/p19-1215.

[46] S. Yadav, M. Sarrouti, D. Gupta, NLM at MEDIQA 2021: Transfer learning-based approaches for consumer question and multi-answer summarization, in: Proceedings of the 20th Workshop on Biomedical Language Processing, Association for Computational Linguistics, Online, 2021, pp. 291–301. URL: https://www.aclweb.org/anthology/2021.bionlp-1.34.

[47] Y. He, M. Chen, S. Huang, damo_nlp at MEDIQA 2021: Knowledge-based preprocessing and coverage-oriented reranking for medical question summarization, in: Proceedings of the 20th Workshop on Biomedical Language Processing, Association for Computational Linguistics, Online, 2021, pp. 112–118. URL: https://www.aclweb.org/anthology/2021.bionlp-1.12.

[48] M. Sänger, L. Weber, U. Leser, WBI at MEDIQA 2021: Summarizing consumer health questions with generative transformers, in: Proceedings of the 20th Workshop on Biomedical Language Processing, Association for Computational Linguistics, Online, 2021, pp. 86–95. URL: https://www.aclweb.org/anthology/2021.bionlp-1.9.

[49] S. Yadav, D. Gupta, A. Ben Abacha, D. Demner-Fushman, Question-aware transformer models for consumer health question summarization, arXiv preprint arXiv:2106.00219 (2021).

[50] A. Cohan, F. Dernoncourt, D. S. Kim, T. Bui, S. Kim, W. Chang, N. Goharian, A discourse-aware attention model for abstractive summarization of long documents, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), 2018, pp. 615–621.

[51] S. E. Robertson, S. Walker, S. Jones, M. M. Hancock-Beaulieu, M. Gatford, et al., Okapi at trec-3, NIST SPECIAL PUBLICATION SP 109 (1995) 109.

[52] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, Journal of Machine Learning Research 21 (2020) 1–67. URL: http://jmlr.org/papers/v21/20-074.html.

[53] B. McCann, N. S. Keskar, C. Xiong, R. Socher, The natural language decathlon: Multitask learning as question answering, arXiv preprint arXiv:1806.08730 (2018).

[54] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, Language models are unsupervised multitask learners, OpenAI blog 1 (2019) 9.

[55] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, L. Zettlemoyer, Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, arXiv preprint arXiv:1910.13461 (2019).

[56] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://www.aclweb.org/anthology/N19-1423. doi:10.18653/v1/N19-1423.

[57] R. Nogueira, Z. Jiang, R. Pradeep, J. Lin, Document ranking with a pretrained sequence-to-sequence model, in: Findings of the Association for Computational Linguistics: EMNLP 2020, Association for Computational Linguistics, Online, 2020, pp. 708–718. URL: https://www.aclweb.org/anthology/2020.findings-emnlp.63. doi:10.18653/v1/

2020.findings-emnlp.63.

[58] P. Bajaj, D. Campos, N. Craswell, L. Deng, J. Gao, X. Liu, R. Majumder, A. McNamara, B. Mitra, T. Nguyen, M. Rosenberg, X. Song, A. Stoica, S. Tiwary, T. Wang, MS MARCO: A Human Generated MAchine Reading COmprehension Dataset, 2018. arXiv:1611.09268.

[59] E. Voorhees, T. Alam, S. Bedrick, D. Demner-Fushman, W. R. Hersh, K. Lo, K. Roberts, I. Soboroff, L. L. Wang, Trec-covid: constructing a pandemic information retrieval test collection, in: ACM SIGIR Forum, volume 54, ACM New York, NY, USA, 2021, pp. 1–12.

[60] G. Balikas, I. Partalas, A. Kosmopoulos, S. Petridis, P. Malakasiotis, I. Pavlopoulos, I. Androutsopoulos, N. Baskiotis, E. Gaussier, T. Artieres, et al., Evaluation framework specifications, Project deliverable D 4 (2013).