NLM at MEDIQA 2021: Transfer Learning-based Approaches for Consumer Question and Multi-Answer Summarization

Shweta Yadav[*], **Mourad Sarrouti**[*], **and Deepak Gupta**[*]
LHNCBC, U.S. National Library of Medicine, MD, USA
{shweta.shweta, mourad.sarrouti, deepak.gupta}@nih.gov

Abstract

The quest for seeking health information has swamped the web with consumers' health-related questions, which makes the need for efficient and reliable question answering systems more pressing. The consumers' questions, however, are very descriptive and contain several peripheral information (like patient's medical history, demographic information, etc.), that are often not required for answering the question. Furthermore, it contributes to the challenges of understanding natural language questions for automatic answer retrieval. Also, it is crucial to provide the consumers with the exact and relevant answers, rather than the entire pool of answer documents to their question. One of the cardinal tasks in achieving robust consumer health question answering systems is the question summarization and multi-document answer summarization. This paper describes the participation of the U.S. National Library of Medicine (NLM) in Consumer Question and Multi-Answer Summarization tasks of the MEDIQA 2021 challenge at NAACL-BioNLP workshop. In this work, we exploited the capabilities of pre-trained transformer models and introduced a transfer learning approach for the abstractive Question Summarization and extractive Multi-Answer Summarization tasks by first pre-training our model on a task-specific summarization dataset followed by fine-tuning it for both the tasks via incorporating medical entities. We achieved the second, sixth and the fourth position for the Question Summarization task in terms ROUGE-1, ROUGE-2 and ROUGE-L scores respectively.

1 Introduction

Healthcare consumers often query over the web to find a quick and reliable answer to their healthcare information needs. On average, 6 million people only in the United States seek health-related information on the Internet every day (Fox and Rainie). One way to facilitate such information-seeking activities is to build a natural language question answering (QA) system that can extract precise answers from the myriad of health-related information sources (Sarrouti and Alaoui, 2020). Though existing search engines respond to the general health-related queries to some extent, users often reach out to specialized medical websites or online health communities for seeking personalized high-quality, and trustworthy answers for their complex health questions. Moreover, consumers while expressing their medical concern on these sources except the involvement of healthcare professionals (HPs) for a quality suggestion and virtual observation (Kummervold et al., 2002). However, the participation of HPs in large-scale discussion forums or medical websites is time-consuming and expensive.

Furthermore, the consumers' questions are very descriptive and contain several peripheral information (like patient's medical history), which contributes to the challenges of understanding natural language questions for automatic answer retrieval (Demner-Fushman et al., 2020). These elaborated details are often not required for providing the relevant answers. Hence, novel strategies should be devised for automatic question simplifications and answer retrieval.

Towards this, we study the tasks of Question Summarization (QS) and Multi-Answer Summarization (MAS) as a part of MEDIQA 2021 (Asma Ben Abacha, 2021) shared task challenge. For the task of Question Summarization (QS), we proposed the transfer learning approach by utilizing multiple pre-trained Transformer (Vaswani et al., 2017) models. In our best run, we fine-tuned the pre-trained models on a variety of question summarization datasets and proposed a medical entities coverage technique to select the best question summary from the pool of question summaries obtained

[*]All the authors contributed equally to this work.

Proceedings of the BioNLP 2021 workshop, pages 291–301

from the various transformer models.

We also explored the transfer learning approach for the Multi-Answer Summarization task. Specifically, the proposed method uses the Text-to-Text Transfer Transformer (T5) relevance-based re-ranking model (Raffel et al., 2020). In our best system, we first fine-tuned T5 on MSMARCO passage and then MEDIQA-QA 2019 datasets. It first ranks the sentences of the answers and then rejoins the top-k sentences as a summary.

2 Related Work

Existing works on the summarization can be broadly categorized into (i) extractive and (ii) abstractive approach which are discussed as follows:

Extractive Summarization: The recent development in the neural network and transformer based models has led to the significant progress in extractive document summarization. Majority of the models focus on the encoder-decoder model (Cheng and Lapata, 2016; Jadhav and Rajan, 2018; Nallapati et al., 2017), recurrent neural network (Nallapati et al., 2017; Zhou et al., 2018), and state-of-the-art Transformers encoders (Zhong et al., 2019b; Liu and Lapata, 2019). For instance, Cheng and Lapata (2016) and Nallapati et al. (2016b) proposed an encoder-decoder model as a binary classifier to decide whether the input sentence will be part of the summary or not. Chen and Bansal (2018) utilize a pointer generator network (Vinyals et al., 2015) to sequentially select sentences from the document for generating the extractive summary. Other decoding techniques, such as ranking (Narayan et al., 2018) has also been utilized for content selection. Recently several studies have explored pre-trained language models in summarization for contextual word representations (Zhong et al., 2019a; Liu and Lapata, 2019).

Abstractive Summarization (AS): With the development of large-scale datasets on abstractive summarization, there has been a significant advancement in AS techniques in the open domain, from traditional sequence to sequence (seq2seq) models, pointer generator network to Transformer based models. Few earlier studies utilize the seq2seq learning approach, trained on the large corpus of news articles for AS (Takase et al., 2016; Rush et al., 2015; Chopra et al., 2016). Later, Li et al. (2018) exploited the seq2seq models on multi-sentence document summarization. However, it was observed that the seq2seq model often generates out-of-vocabulary (OOV) words, factually incorrect details, and repetitions. To mitigate the issues of the seq2seq model, the pointer generator network was introduced that has the capability of handling OOV words with the copy mechanism (Gu et al., 2016; Nallapati et al., 2016a). Further, to address the repetition problem, Chen et al. (2016) proposed Distraction-based attention model. The additional coverage mechanism (See et al., 2017) ensures the generation of non-hallucinated summaries. Although these methods are good at generating readable summaries to a certain extent, the problem of factual inconsistencies persists with them. To alleviate this issue, several new methods (Lebanoff et al., 2020; Huang et al., 2020) has been proposed to generate more factually correct summaries. Few other recent works (Falke et al., 2019; Kryściński et al., 2019; Wang et al., 2020a) have exploited question answering and natural language inference (NLI) models to identify factual coherence in the generated summary. Recently several new models (Gehrmann et al., 2019) have been proposed that investigates the use of the transfer learning approach. Most recently the pseudo-self attention method (Ziegler et al., 2019) has been developed, which enables transfer learning to be applied in abstractive summarization.

Recently, with the availability of benchmark clinical data sets (MIMIC-CXR, and OpenI), there have been some prominent advancements in abstractive summarization of radiology reports. Zhang et al. (2018) utilized the pointer-generator network to generate the summary of radiology impressions and observed very high overlap with the human summaries. MacAvaney et al. (2019) further advanced the performance of the pointer generator model by augmenting medical-ontologies. Ben Abacha and Demner-Fushman (2019) has focused on the consumer health question summarization task. They created the corpus of $1,000$ question summaries and exploited seq2seq and pointer generator model to generate the consumer-health question summaries.

This work advances the pre-trained models for the summarization of consumers' questions and introduces new approaches to preserve the intent and the salient medical entities of the original questions.

3 Methods

3.1 Question Summarization

We tackle the first task of MEDIQA 2021, consumer health questions (CHQ) summarization with the goal of generating summarized questions that contain the key focus and semantics of the original question. Formally, given a consumer health question Q having m words $q_1, q_2, \ldots, q_m$, the task is to generate the summary sentence $\hat{S}$ having a sequence of n words $\hat{S} = \{s_1, s_2, \ldots, s_n\}$ expressing the key focus and semantics of the original question Q. Mathematically,

$$\begin{aligned} \hat{S} &= \arg\max_{S} prob(S|Q;\phi) \\ &= \arg\max_{S} prob(S|q_1, q_2, \ldots q_m;\phi) \end{aligned} \tag{1}$$

where ϕ are network parameters.

Pre-trained Transformer Models: We utilized the following pre-trained models and uses the transfer learning-based approach to fine-tune them on the task of question summarization.

- **ProphetNet** (Qi et al., 2020): It is a sequence-to-sequence model which is pre-trained using the self-supervised objective called future n-gram prediction. The ProphetNet is pre-trained by predicting the next n tokens simultaneously based on previous context tokens at each time step thus optimizing n-step ahead predictions of the model. The n-step ahead predictions encourage the model to plan for the future tokens and prevent overfitting on strong local correlations. We chose ProphetNet because it is specifically designed for sequence-to-sequence training and it has shown near state-of-the-art results on natural language generation tasks.

- **PEGASUS** (Zhang et al., 2020a): It is a large Transformer-based encoder-decoder model which is pre-trained on massive text corpora with a novel self-supervised objective called Gap Sentences Generation. This object is specially designed to pre-trained the transformer model for abstractive summarization. The important sentences from the document are masked and are generated together as one output sequence from the remaining sentences of the document.

- **T5** (Raffel et al., 2020): This is another pre-trained model developed by exploring the transfer learning techniques for natural language processing (NLP) by introducing a unified framework that converts all text-based language problems into a text-to-text format. The T5 model is an Encoder-Decoder Transformer with some architectural changes as discussed in detail in Raffel et al. (2020).

Pre-processing: To summarize the test questions, we followed certain pre-processing steps to transform the input consumer health question into a well-formed question. We applied the following pre-processing steps to the input test questions.

1. **Spelling Correction:** As consumer health questions are often ill-formed and contain multiple misspelled words particularly the medical terms (entities), therefore, we performed spelling correction on the original consumer health questions. Specifically, we utilized the *CSpell*[1], that aims to correct spellings from consumer health text.

2. **Abbreviation Expansion:** In order to generate the factually complete summaries, we first detect the medical entities and later expand the abbreviated entities using the '*Another database of abbreviations in MEDLINE*' (ADAM[2]) (Zhou et al., 2006).

Post-processing: Our analysis on the generated summary from the validation dataset using the pre-trained model reveals the following: **(1)** The T5 model generates a long summary and ended up with better coverage of the key entities present in the original question; **(2)** For the longer and complex questions, the T5 model generates the extractive-type summary; **(3)** Unlike T5, PEGASUS generates the short and succinct summaries which are often abstractive in nature; **(4)** The ProphetNet model often generates the moderate length summaries but approximately cover the key information from the original questions.

The correct summary of the consumer health questions must contain the key medical entities and question semantics of the original question. Motivated by the aforementioned observations, we obtained the generated summary from the pre-trained

[1] https://lsg3.nlm.nih.gov/LexSysGroup/Projects/cSpell/current/web/index.html

[2] http://arrowsmith.psych.uic.edu/arrowsmith_uic/index.html

Transformer models and performed the following steps to ensure the maximum coverage of medical entities so that it captures the key question-focus, and select the best question summary from the pool of generated summaries.

1. **Medical Entities Extraction:** We extracted the medical entities using the *Metamap*[3] (Aronson and Lang, 2010) and *Scispacy*[4] medical entity recognizer (en_ner_bionlp13cg_md). We removed some false entities ('*False Interventions*', '*False Anatomy*', '*False Problems*') using the Unified Medical Language System (UMLS) (Bodenreider, 2004) based filters[5]. Given a question Q, we obtained the list of medical entities as follows:

$$ent(Q) = MetaMap(Q) \cup Scispacy(Q)$$
$$entities(Q) = ent(Q) - False(ent(Q)) \quad (2)$$

where, $MetaMap(;)$ and $Scispacy(;)$ are the medical entities extracted using MetaMap and Scispacy respectively, $False(;)$ is a method which provided the list of *False* entities. The final entities of the question is obtained using the $entities(;)$ method, which filters the false entities from the union of the list of both the entities.

2. **Medical Entities Coverage:** Given the original question Q and candidate question summary C, we extracted the medical entities E_Q and E_C using the approach discussed in Eq 2. We computed the medical entities coverage as follows:

$$coverage(Q, C) = \frac{|E_Q \cap E_C|}{|E_Q|} \quad (3)$$

where $|x|$ is the cardinality of the set $x \in \{E_Q, E_Q \cap E_C\}$. We computed the coverage score for each candidate question summary generated using the different pre-trained Transformer models. We sort the candidate question summary based on the coverage score and passed the list to check the sanity of generated questions.

3. **Checking well-formed Question:** We check the list of generated questions against the well-formedness of the questions. Formally, we check:

   (a) Whether the generated questions starts with Wh words[6] or not.

   (b) Whether the generated question ends with the question word ('?').

   If the generated question having maximum coverage score is a well-formed question then we select the generated question as the final summary of the original question. Otherwise, we skip the non-well-formed candidate question and check against the next candidate question. In the case of the same coverage score among all three models, we selected the summary generated from PEGASUS, as it is more abstractive in nature.

3.2 Multi-Answer Summarization

To address the Multi-Answer Summarization (MAS) task at the MEDIQA 2021 challenge, we introduce an extractive method based on the T5 relevance-based re-ranking model (Raffel et al., 2020). The proposed method consists of extracting important and most relevant sentences from the answers and rejoining them to form a summary. To evaluate the importance of a sentence, we used T5 relevance-based ranking model. To do so, we first split the multiple answers of a given question into sentences using NLTK[7], and then ranked these sentences based on the relevance score that determines how relevant a candidate sentence is to a question. The sentences are ranked by a pointwise re-ranker (Nogueira et al., 2020) which uses T5, a sequence-to-sequence model that uses traditional transformer architecture, and BERT's masked language modeling (Devlin et al., 2019). We adopt the approach to sentence ranking by using the following input sequence:

$$Question : q \; Sentence : \; s \; Relevant : \quad (4)$$

The model is first fine-tuned to generate the tokens "true" when the sentence is relevant to the question and "false" when the sentence is not relevant to the question. It then applies softmax on

[3] https://metamap.nlm.nih.gov/

[4] https://allenai.github.io/scispacy/

[5] https://gist.github.com/h4ste/14b10d412d0d3c043c1d123c75c6ad29

[6] https://en.wikipedia.org/wiki/Interrogative_word

[7] https://www.nltk.org/

the logits of the "true" and "false" words and ranks the sentences using the probabilities of the "true" token. More details about this approach appear in (Nogueira et al., 2020).

The model is fine-tuned on (1) MS MARCO passage (Bajaj et al., 2018), (2) MS MARCO MED (MacAvaney et al., 2020), and (3) MEDIQA-QA 2019 dataset (Ben Abacha et al., 2019). We used the question-answer pairs in MEDIQA-QA with scores 1 and 2 (i.e., incorrect and related answers) as negative instances and the question-answer pairs with scores 3 and 4 (i.e., incomplete and excellent answers) as positive instances.

We form the summary by rejoining the selected top-k sentences. We also used Metamap[8] (Aronson and Lang, 2010) to replace the abbreviations by their definitions.

4 Experimental Results and Discussion

4.1 Evaluation Metrics

The performance of the question summarization and multi-answer summarization are evaluated against the ROUGE (Lin, 2004) score. We reported the results in terms of ROUGE-1 (R-1), ROUGE-2 (R-2), ROUGE-L (R-L). The organizer also release scores using the BERTScore (Zhang et al., 2020b) and HOLMS (Mrabet and Demner-Fushman, 2020).

4.2 Datasets

Question Summarization: For the task of question summarization, we use the following dataset to fine-tuned the pre-trained Transformer models.

1. **MeQSum** (Ben Abacha and Demner-Fushman, 2019): We use the $1,000$ consumer health question summarization dataset created by the medical experts. The questions are selected from a collection distributed by the U.S. National Library of Medicine (Kilicoglu et al., 2018).

2. **Clinical Questions** (Ely et al., 2000): We also utilized the $4,655$ clinical questions dataset, which contains the clinical questions and their short summaries.

3. **MEDIQA-RQE** (Ben Abacha et al., 2019): This dataset is released in the BioNLP 2019 shared task. The dataset is derived from consumer health questions (CHQs) and frequently asked questions (FAQs) from the U.S. National Library of Medicine and National Institute of Health respectively. We use the MEDIQA-RQE training dataset and choose only the entailed question pairs to form the silver-standard training dataset. We choose the longer question as the source question and the other as the target question. With this process, we formulated the $4,655$ additional training question pairs to train the question summarization model.

4. **MedNLI** (Romanov and Shivade, 2018): We also used the MedNLI - a dataset annotated by doctors, performing a natural language inference task, grounded in the medical history of patients. We augment training, validation, and test datasets and choose only the entailed question pairs to form the silver-standard training dataset. Similar to MEDIQA-RQE, we choose the longer question as the source question and the other as the target question. We obtained the $4,683$ question pairs from this dataset to include in the question summarization training dataset.

5. **LiveQA17** (Ben Abacha et al., 2017): We also utilized the 104 questions and their summary from the LiveQA17 test dataset as it contains the gold summaries of the source questions.

Multi-Answer Summarization: We used the following datasets to fine-tuned the T5 model for the multi-answer summarization task:

1. **MS MARCO Passage** (Bajaj et al., 2018): It is a large dataset for passage ranking. It contains 8.8M passages retrieved by Bing search engine for around 1M natural language questions.

2. **MSMARCO MED** (MacAvaney et al., 2020): This dataset contains the medical subset of MS MARCO. It includes only medical-related queries.

3. **MEDIQA-QA 2019** (Ben Abacha et al., 2019): It is a dataset for medical question answering obtained by submitting medical questions to the consumer health QA system CHiQA. The answers for the questions were manually ranked by medical experts.

[8] https://metamap.nlm.nih.gov/

4.3 Implementation Details

For question summarization task, we used the T5-large[9], ProphetNet-large-uncased[10] and pegasus-large[11] pre-trained models. The models are fine-tuned with maximum source question length of 120 and target summary length of 20. We train the model for 10 epochs and choose the best model based on the model performance (in terms of ROUGE-2) on the MEDIQA 2021 validation dataset. In our MAS experiments, we used the T5-base implementations provided in HuggingFace's Transformers package version 2.10 (Wolf et al., 2020). All models were trained with a batch size of 8 and a maximum sequence length of 512 tokens for 20 epochs using single P100 GPUs (16 GB VRAM) on a shared cluster. We use the beam search method to generate the summarized questions. For both the task Adam optimizer (Kingma and Ba, 2015) with a learning rate of 1e-5 was used for the parameters updates.

4.4 Results and Discussion

We devise multiple runs to assess **(1)** the ability of pre-trained Transformer model to summarize consumer health questions, **(2)** the role of additional datasets to improve the performance of CHQs summarization systems, and **(3)** the effect of the medical entities coverage to effectively select the best summarized questions from the pool of multiple summarized questions generated by pre-trained Transformer models. For the Question Summarization task, we submitted multiple runs which are described below:

1. **Run-1:** In this run, we fine-tuned the MiniLM (Wang et al., 2020b) model on the MeQSum (only 500 question-summary pairs) and Clinical Questions datasets. The summaries are generated using a beam of size 4.
2. **Run-2:** This run is similar to the **Run-1**, except we generated the summaries with the beam of size 6.
3. **Run-3:** For this run, we fine-tuned the ProphetNet model on the MeQSum and Clinical Questions datasets. The summaries are generated using a beam of size 4.
4. **Run-4:** We fine-tuned the T5 model on the MeQSum and Clinical Questions datasets. The summaries are generated using a beam of size 4.
5. **Run-5:** The PEGASUS model is fine-tuned on the MeQSum and Clinical Questions datasets. The summaries are generated using a beam of size 4.
6. **Run-6:** The T5 model is fine-tuned on the MeQSum, Clinical Questions, and MEDIQA-RQE datasets. The summaries are generated using a beam of size 4.
7. **Run-7:** We fine-tuned the T5, PEGASUS, ProphetNet models on the MeQSum, Clinical Questions, MEDIQA-RQE, LiveQA17, and MedNLI datasets. We also performed the pre-processing and post-processing steps (without well-formed questions) discussed in Section 3.1. The summaries are generated using a beam of size 4.
8. **Run-8:** The PEGASUS model is fine-tuned on the MeQSum, Clinical Questions, MEDIQA-RQE, LiveQA17, and MedNLI datasets. We also performed the pre-processing step discussed in Section 3.1. The summaries are generated using a beam of size 4, Top-K Sampling (Fan et al., 2018) with $K = 50$ and Top-p (nucleus) Sampling (Holtzman et al., 2019) with $p = 0.97$.
9. **Run-9:** The run is similar to **Run-7** however, we performed both the pre-processing and post-processing steps as described in Section 3.1 and the beam of size 5 is used to generate the summaries.
10. **Run-10:** This is final run similar to **Run-9**, however, we also included a subset $(10,324)$ of questions from Quora duplicate question detection dataset[12] to fine-tuned the pre-trained models. We choose only those questions from the Quora dataset which are duplicates. We consider the question having more than 2 sentences and longer than the associated duplicate question as the source question and other duplicate question as target summary question.

For all our runs, we kept the maximum length of generated summary is 20. We have shown the detailed performance evaluation based on different metrics in Table 1. Our best submission (Run-9) achieved the maximum of ROUGE-1 (35.58), ROUGE-2 (15.14), HOLMS (56.59) and

[9] https://huggingface.co/t5-large

[10] https://huggingface.co/microsoft/prophetnet-large-uncased

[11] https://huggingface.co/google/pegasus-large

[12] http://qim.fs.quoracdn.net/quora_duplicate_questions.tsv

Run#	ROUGE-1	ROUGE-2	ROUGE-L	HOLMS	BERTScore
1	26.24	9.06	23.68	53.74	63.07
2	25.88	8.76	23.23	53.27	63.25
3	30.38	11.25	26.58	54.54	65.62
4	33.01	12.91	27.61	52.26	65.58
5	33.24	13.87	28.77	55.69	67.35
6	34.10	13.71	29.65	55.17	68.54
7	35.58	15.12	31.16	56.51	68.90
8	33.73	14.38	29.79	56.21	68.22
9	35.56	15.14	31.10	56.49	68.92
10	35.28	15.08	30.79	56.59	68.94
Our Best Run	**35.58**	**15.14**	**31.16**	**56.59**	**68.94**
Best Participants	35.80	16.08	31.49	57.87	70.27
Average Participants	29.55	11.59	26.60	53.25	64.93

Table 1: Official results of MEDIQA 2021: NLM runs for the Question Summarization task.

Run#	ROUGE-1	ROUGE-2	ROUGE-L	HOLMS	BERTScore
1	0.524	0.410	0.322	0.674	0.758
2	0.504	0.414	0.302	0.640	0.772
3	0.507	0.417	0.303	0.643	0.773
4	0.547	0.468	0.328	0.657	0.764
5	0.524	0.446	0.309	0.633	0.786
Our Best Run	**0.547**	**0.468**	**0.328**	**0.657**	**0.764**
Best Participants	0.585	0.508	0.435	0.704	0.803
Average Participants	0.524	0.422	0.353	0.668	0.751

Table 2: Official results of MEDIQA 2021: NLM runs for the Multi-Answer Summarization task.

BERTScore (68.94). Run-7 achieves the maximum ROUGE-L score of 31.16. Our best run achieved the ROUGE-2 score of 15.14, which is slightly (0.94) lower than the best run submitted for the Question Summarization task in MEDIQA 2021. Similarly, our best run obtained the improvement of 3.55 ROUGE-2 points over the average ROUGE-2 score obtained by all the participant's runs. We achieved the second-best result (35.58) in terms of the ROUGE-1 score over all the submitted runs for the Question Summarization task in MEDIQA 2021. We also show the best and average results among all the participants against various evaluation metrics in Table 1.

Qualitative Analysis: We carried out an in-depth analysis of the generated summaries of the models (Run 3,4,5,7,9) as shown in Table-3 for the question summarization task. We randomly selected 20 summaries from the test set and manually evaluated the summaries generated by the models. Table-3 shows that for question #1 and #2, our best run (#9) generates the readable summaries with the correct question focus and type. However, for the question #3, our best run is only able to capture partial question type and therefore generated the partially correct summary. We also observed that though T5 and PEGASUS generate factually correct summaries, sometimes it fails to generate a fully correct summary. Overall, the pre-trained models generate readable and succinct summaries which can be further enhanced by integrating the information about question focus and types.

Discussion: Our results confirms the role of additional datasets in fine-tuning the pre-trained Transformer models to improve the performance of CHQs summarization task. Run-1 and Run-2 which are based on fine-tuning the pre-trained Transformer models with the MeQSum dataset achieve low R-2 scores (8.76 and 9.06). The additional Clinical Questions dataset helps to improve the performance of the pre-trained model in Run-3, 4, and 5. The additional datasets (MEDIQA-RQE,

Question #1	Hi to whom this may concern: I have a non alcoholic fatty liver and when I seen my consultant he say I don't have to see Him again for 5 years and advised me to drink 3 cups of coffee a day, But I can't drink coffee at all, So I email a consultant asking if he new of a coffee tablet, So he email me back saying it properly glutathione S-transferases (GSTs) in the coffee so am looking for a cheap tablet which contains.these enzymes to help repair my liver I also only have one kidney which am on long term antibiotics for I do hope you can help. I look forward to your reply
Run-3	how can i find physician(s) who specialize in non alcoholic fatty liver disease?
Run-4	Where can I buy a tablet containing glutathione-s-transferase enzymes to
Run-5	Is there a coffee tablet which contains glutathione-s-transferase enzymes?
Run-7	Where can I buy cheap tablets containing glutathione-s-transferase enzymes?
Run-9	Where can I buy cheap tablets containing glutathione-s-transferase enzymes?
Question #2	Can Irritable bowel syndrome (IBS) be cured with the patent Medicines say Mebverine or others as prescribed the Doctor. Does Probiotics are helpful in controlling the Irritable bowel syndrome (IBS). Some says Peeper Mint can also be suitable. These are few general questions as regards Irritable bowel syndrome (IBS) and not personal.
Run-3	can ibs be cured with the patent medicines mebverine or others as prescribed?
Run-4	Can IBS be cured with the patent Medicines Mebverine or others as prescribed the Doctor. Does Pro
Run-5	Can IBS be cured with the patent Medicines say Mebverine or others?
Run-7	Can Irritable bowel syndrome (IBS) be cured with the patent Medicines say Mebverine or others?
Run-9	Can Irritable Bowel Syndrome (IBS) be cured with the Mebverine ?
Question #3	SUBJECT: I NEED PARENT TRAINING. MESSAGE: My 9 year old son has oppositional defiant disorder (ODD) and attention deficit disorder (ADHD). I'm in desperate need of an extensive parent training class. I don't know what to do anymore and really need help. My son says I do nothing for him after spending the entire day taking him places. He yells and screams at me and tells me I'm a terrible mother. He refuses to listen and I have to ask 10 times before he reacts. He will not do homework.. I have tried positive reinforcement..... Please I need someone that know their stuff..
Run-3	what is the treatment for a child with odd and adhd?
Run-4	Is there a parent training program for ODD and ADHD that I can take with my 9-year-old?
Run-5	What are the treatments for obsessive-compulsive disorder and attention-deficit-hyperactivity disorder?
Run-7	what are the treatments for oppositional defiant disorder (odd) and attention deficit disorder (adhd)?
Run-9	what are the treatments for oppositional defiant disorder (odd) and attention deficit disorder (adhd)?

Table 3: Qualitative Analysis: Examples showing models generated summaries.

LiveQA17, and MedNLI) with the pre-processing and post-processing steps further boost the performance of the question summarization as shown in Run-7 and Run-9. We also fine-tuned the Transformers model with the Quora duplicate question detection dataset in Run-10, in order to generate more diverse summaries. However, it could not improve the question summarization performance compare to the Run-9. It is because Quora dataset is a open domain dataset, which may not be well suited for the medical summarization task.

Multi-answer Summarization Task: We submitted the following runs for the multi-answer summarization task at MEDIQA 2021:

- **Run-1**: We fine-tuned the T5 model on the MSMARCO passage. We ranked the sentences of the answers based on the T5 relevance score and rejoined the top-10 sentences as a summary. We also identified the long-form of abbreviations in the test set.
- **Run-2**: We fine-tuned the T5 model on the MSMARCO passage. We ranked the sentences of the answers based on the T5 relevance score and then concatenated the top-10 sentences to form the summary.
- **Run-3**: We fine-tuned the T5 model on the MSMARCO passage. We ranked the sentences of the answers based on the T5 relevance score and rejoined the top-20 sentences as a summary.
- **Run 4**: We fine-tuned the T5 model on MSMARCO passage and then MEDIQA-QA 2019 dataset. The top-20 sentences are concatenated to form the summary.
- **Run-5**: We fine-tuned the T5 model on MEDMSMARCO and then MEDIQA-QA 2019 dataset. The top-20 sentences are concatenated to form the summary.

Table 2 presents the official results of our systems in the multi-answer summarization task of the MEDIQA 2021 challenge. Out of the five runs, our best result was obtained by the run #4, achieving 0.547, 0.468, and 0.328 in terms of ROUGE-1, ROUGE-2, and ROUGE-L respectively. In terms of BERTScore, our run #5 achieved the best results among our runs. On the other hand, run #1 achieved the highest HOLMS. The obtained results also showed that our T5-based system is more competitive in terms of various evaluation metrics over the other participant's systems.

5 Conclusion and Future Work

In this paper, we describe our submissions for the tasks of Question Summarization and Multi Answer Summarization at MEDIQA 2021 shared task. For the Question Summarization task, our best run achieved the second-best ROUGE-1 score among all the submitted runs in the shared task. We also obtained the competitive scores in terms of various evaluation metrics over the other participant's runs. For the Multi-Answer Summarization task, our T5-based approach achieved good performances compared to participants' systems. In the future, we will explore the techniques to integrate the medical entities and semantics in the pre-trained transformer models for the task of question summarization. Further, we will also explore the abstractive approaches for multi-answer summarization.

References

Alan R Aronson and François-Michel Lang. 2010. An overview of metamap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17(3):229–236.

Yuhao Zhang Chaitanya Shivade Curtis Langlotz Dina Demner-Fushman Asma Ben Abacha, Yassine Mrabet. 2021. Overview of the mediqa 2021 shared task on summarization in the medical domain. In *Proceedings of the 20th SIGBioMed Workshop on Biomedical Language Processing, NAACL-BioNLP 2021*. Association for Computational Linguistics.

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. 2018. Ms marco: A human generated machine reading comprehension dataset.

Asma Ben Abacha, Eugene Agichtein, Yuval Pinter, and Dina Demner-Fushman. 2017. Overview of the medical question answering task at trec 2017 liveqa. In *TREC 2017*.

Asma Ben Abacha and Dina Demner-Fushman. 2019. On the role of question summarization and information source restriction in consumer health question answering. *AMIA Summits on Translational Science Proceedings*, 2019:117.

Asma Ben Abacha and Dina Demner-Fushman. 2019. On the summarization of consumer health questions. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 2228–2234. Association for Computational Linguistics.

Asma Ben Abacha, Chaitanya Shivade, and Dina Demner-Fushman. 2019. Overview of the MEDIQA 2019 shared task on textual inference, question entailment and question answering. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 370–379, Florence, Italy. Association for Computational Linguistics.

Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270.

Qian Chen, Xiaodan Zhu, Zhenhua Ling, Si Wei, and Hui Jiang. 2016. Distraction-based neural networks for document summarization. *arXiv preprint arXiv:1610.08462*.

Yen-Chun Chen and Mohit Bansal. 2018. Fast abstractive summarization with reinforce-selected sentence rewriting. *arXiv preprint arXiv:1805.11080*.

Jianpeng Cheng and Mirella Lapata. 2016. Neural summarization by extracting sentences and words. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 484–494.

Sumit Chopra, Michael Auli, and Alexander M Rush. 2016. Abstractive sentence summarization with attentive recurrent neural networks. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 93–98.

Dina Demner-Fushman, Yassine Mrabet, and Asma Ben Abacha. 2020. Consumer health information and question answering: helping consumers find answers to their health-related information needs. *Journal of the American Medical Informatics Association*, 27(2):194–201.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

John W Ely, Jerome A Osheroff, Paul N Gorman, Mark H Ebell, M Lee Chambliss, Eric A Pifer, and P Zoe Stavri. 2000. A taxonomy of generic clinical questions: classification study. *Bmj*, 321(7258):429–432.

Tobias Falke, Leonardo FR Ribeiro, Prasetya Ajie Utama, Ido Dagan, and Iryna Gurevych. 2019. 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*, pages 2214–2220.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. *arXiv preprint arXiv:1805.04833*.

Susannah Fox and Lee Rainie. Main report: The search for online medical help.

Sebastian Gehrmann, Zachary Ziegler, and Alexander M Rush. 2019. Generating abstractive summaries with finetuned language models. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 516–522.

Jiatao Gu, Zhengdong Lu, Hang Li, and Victor OK Li. 2016. Incorporating copying mechanism in sequence-to-sequence learning. *arXiv preprint arXiv:1603.06393*.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*.

Luyang Huang, Lingfei Wu, and Lu Wang. 2020. Knowledge graph-augmented abstractive summarization with semantic-driven cloze reward. *arXiv preprint arXiv:2005.01159*.

Aishwarya Jadhav and Vaibhav Rajan. 2018. 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)*, pages 142–151.

Halil Kilicoglu, Asma Ben Abacha, Yassine Mrabet, Sonya E Shooshan, Laritza Rodriguez, Kate Masterton, and Dina Demner-Fushman. 2018. Semantic annotation of consumer health questions. *BMC bioinformatics*, 19(1):34.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Wojciech Kryściński, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Evaluating the factual consistency of abstractive text summarization. *arXiv preprint arXiv:1910.12840*.

Per E Kummervold, Deede Gammon, Svein Bergvik, Jan-Are K Johnsen, Toralf Hasvold, and Jan H Rosenvinge. 2002. Social support in a wired world: use of online mental health forums in norway. *Nordic journal of psychiatry*, 56(1):59–65.

Logan Lebanoff, John Muchovej, Franck Dernoncourt, Doo Soon Kim, Lidan Wang, Walter Chang, and Fei Liu. 2020. Understanding points of correspondence between sentences for abstractive summarization. *arXiv preprint arXiv:2006.05621*.

Wei Li, Xinyan Xiao, Yajuan Lyu, and Yuanzhuo Wang. 2018. Improving neural abstractive document summarization with explicit information selection modeling. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 1787–1796.

Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 workshop*, volume 8. Barcelona, Spain.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. *arXiv preprint arXiv:1908.08345*.

Sean MacAvaney, Arman Cohan, and Nazli Goharian. 2020. Sledge: a simple yet effective baseline for covid-19 scientific knowledge search. *arXiv e-prints*, pages arXiv–2005.

Sean MacAvaney, Sajad Sotudeh, Arman Cohan, Nazli Goharian, Ish Talati, and Ross W Filice. 2019. Ontology-aware clinical abstractive summarization. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1013–1016.

Yassine Mrabet and Dina Demner-Fushman. 2020. HOLMS: Alternative summary evaluation with large language models. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5679–5688, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. 2017. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

Ramesh Nallapati, Bowen Zhou, Caglar Gulcehre, Bing Xiang, et al. 2016a. Abstractive text summarization using sequence-to-sequence rnns and beyond. *arXiv preprint arXiv:1602.06023*.

Ramesh Nallapati, Bowen Zhou, and Mingbo Ma. 2016b. Classify or select: Neural architectures for extractive document summarization. *arXiv preprint arXiv:1611.04244*.

Shashi Narayan, Shay B Cohen, and Mirella Lapata. 2018. Ranking sentences for extractive summarization with reinforcement learning. *arXiv preprint arXiv:1802.08636*.

Rodrigo Nogueira, Zhiying Jiang, Ronak Pradeep, and Jimmy Lin. 2020. Document ranking with a pre-trained sequence-to-sequence model. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 708–718, Online. Association for Computational Linguistics.

Weizhen Qi, Yu Yan, Yeyun Gong, Dayiheng Liu, Nan Duan, Jiusheng Chen, Ruofei Zhang, and Ming Zhou. 2020. ProphetNet: Predicting future n-gram for sequence-to-SequencePre-training. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2401–2410, Online. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Alexey Romanov and Chaitanya Shivade. 2018. Lessons from natural language inference in the clinical domain. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1586–1596, Brussels, Belgium. Association for Computational Linguistics.

Alexander M Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. *arXiv preprint arXiv:1509.00685*.

Mourad Sarrouti and Said Ouatik El Alaoui. 2020. SemBioNLQA: A semantic biomedical question answering system for retrieving exact and ideal answers to natural language questions. *Artificial Intelligence in Medicine*, 102:101767.

Abigail See, Peter J Liu, and Christopher D Manning. 2017. Get to the point: Summarization with pointer-generator networks. *arXiv preprint arXiv:1704.04368*.

Sho Takase, Jun Suzuki, Naoaki Okazaki, Tsutomu Hirao, and Masaaki Nagata. 2016. Neural headline generation on abstract meaning representation. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 1054–1059.

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.

Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. 2015. Pointer networks. *arXiv preprint arXiv:1506.03134*.

Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020a. Asking and answering questions to evaluate the factual consistency of summaries. *arXiv preprint arXiv:2004.04228*.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020b. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *arXiv preprint arXiv:2002.10957*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Huggingface's transformers: State-of-the-art natural language processing.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020a. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pages 11328–11339. PMLR.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020b. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Yuhao Zhang, Daisy Yi Ding, Tianpei Qian, Christopher D Manning, and Curtis P Langlotz. 2018. Learning to summarize radiology findings. *arXiv preprint arXiv:1809.04698*.

Ming Zhong, Pengfei Liu, Danqing Wang, Xipeng Qiu, and Xuanjing Huang. 2019a. Searching for effective neural extractive summarization: What works and what's next. *arXiv preprint arXiv:1907.03491*.

Ming Zhong, Danqing Wang, Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. 2019b. A closer look at data bias in neural extractive summarization models. *arXiv preprint arXiv:1909.13705*.

Qingyu Zhou, Nan Yang, Furu Wei, Shaohan Huang, Ming Zhou, and Tiejun Zhao. 2018. Neural document summarization by jointly learning to score and select sentences. *arXiv preprint arXiv:1807.02305*.

Wei Zhou, Vetle I Torvik, and Neil R Smalheiser. 2006. Adam: another database of abbreviations in medline. *Bioinformatics*, 22(22):2813–2818.

Zachary M Ziegler, Luke Melas-Kyriazi, Sebastian Gehrmann, and Alexander M Rush. 2019. Encoder-agnostic adaptation for conditional language generation. *arXiv preprint arXiv:1908.06938*.