Medical Question Understanding and Answering with Knowledge Grounding and Semantic Self-Supervision

Khalil Mrini[1], Harpreet Singh[1], Franck Dernoncourt[2], Seunghyun Yoon[2], Trung Bui[2], Walter Chang[2], Emilia Farcas[1], and Ndapa Nakashole[1]

[1]University of California, San Diego, La Jolla, CA 92093
{khalil, h1singh, efarcas, nnakashole}@ucsd.edu
[2]Adobe Research, San Jose, CA 95110
{franck.dernoncourt, syoon, bui, wachang}@adobe.com

Abstract

Current medical question answering systems have difficulty processing long, detailed and informally worded questions submitted by patients, called Consumer Health Questions (CHQs). To address this issue, we introduce a medical question understanding and answering system with knowledge grounding and semantic self-supervision. Our system is a pipeline that first summarizes a long, medical, user-written question, using a supervised summarization loss. Then, our system performs a two-step retrieval to return answers. The system first matches the summarized user question with an FAQ from a trusted medical knowledge base, and then retrieves a fixed number of relevant sentences from the corresponding answer document. In the absence of labels for question matching or answer relevance, we design 3 novel, self-supervised and semantically-guided losses. We evaluate our model against two strong retrieval-based question answering baselines. Evaluators ask their own questions and rate the answers retrieved by our baselines and own system according to their relevance. They find that our system retrieves more relevant answers, while achieving speeds 20 times faster. Our self-supervised losses also help the summarizer achieve higher scores in ROUGE, as well as in human evaluation metrics. We release our code to encourage further research.[1]

1 Introduction

Motivation. Users of medical question answering systems often write long questions, called Consumer Health Questions (CHQs). Several aspects of CHQs hinder the capacity of current question answering (QA) systems to process them: long medical questions may contain peripheral information like patient history (Roberts and Demner-Fushman, 2016) that are not necessary to retrieve relevant answers. Consumer health questions may also use a distinct vocabulary from the one used by medical providers to describe the same health concepts (Ben Abacha and Demner-Fushman, 2019a).

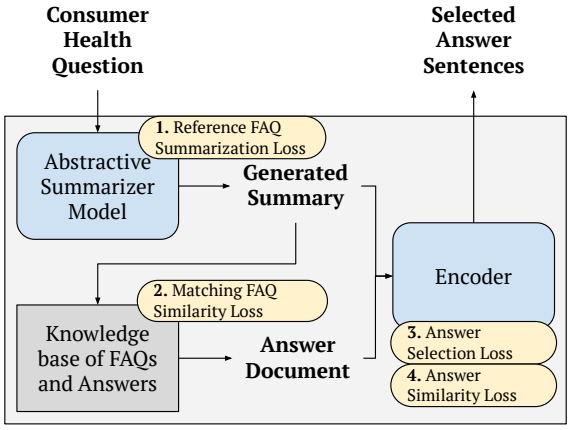

Figure 1: Overview of our proposed Consumer Health Question Understanding and Answering model. The input is a user question, called *Consumer Heath Question* (CHQ). The goal is to match the CHQ to relevant answer sentences associated with a *Frequently Asked Question* (FAQ) from a medical knowledge base.

A growing number of approaches attempt to enhance the processing of consumer health questions – or medical question understanding. These approaches include query relaxation (Ben Abacha and Zweigenbaum, 2015; Lei et al., 2020), question entailment (Ben Abacha and Demner-Fushman, 2016, 2019b; Agrawal et al., 2019), question summarization (Ben Abacha and Demner-Fushman, 2019a), and question similarity (Ben Abacha and Demner-Fushman, 2017; Yan and Li, 2018).

However, the above medical question understanding approaches stop short of retrieving answers after processing consumer health questions. The Medical Question Answering Task at TREC 2017 LiveQA (Ben Abacha et al., 2017) attempts to fill the gap by proposing the task of Consumer Health Question Answering. The goal is to retrieve relevant answers obtained using online search for

[1]Link: https://github.com/KhalilMrini/Medical-Question-Answering

Proceedings of the 29th International Conference on Computational Linguistics, pages 2734–2747
October 12–17, 2022.

the corresponding CHQ. As part of their participation in this task, Yang et al. (2017) find that online search engine queries introduce noise in performance, and that even collected and curated medical knowledge available offline can fare better.

Contributions. To enable the use of a curated medical knowledge base for answering long user questions, we introduce a novel, knowledge-grounded and semantically self-supervised system for Consumer Health Question Understanding and Answering (CHQUA). We tackle a challenging aspect of CHQUA: providing answers when no relevance labels are available. Our contributions are as follows:

(1) We propose an end-to-end pipeline, as shown in Figure 1, that takes as input a consumer health question, and trains a summarizer model to generate a short, formally worded question. We optimize a summarization training objective using the medical question summarization datasets.

(2) The medical knowledge base we use is separate from the question summarization datasets, and therefore we have no labels to indicate which knowledge base question matches a given consumer health question. We design a novel, semantically-guided self-supervised loss function to ground the generated summary with knowledge base FAQs, using semantic similarity as proxy to question matching. The Matching FAQ similarity loss helps the encoder pick the most semantically similar knowledge base question.

(3) The large medical knowledge base we use has no answer sentence relevance labels. We adapt to this scenario by designing two complementary self-supervised losses on the same encoder, and by considering semantic similarity as a proxy to relevance. The Answer Similarity loss pushes the model to distinguish between relevant and irrelevant answer sentences, whereas the Answer Selection loss works in a complementary way to push the model to select a given number of sentences.

Finally, we conduct an evaluation to compare the relevance of our system with two strong baselines of retrieval-based question answering. We ask evaluators to ask their own questions, and then perform a blind evaluation of the retrieved answers by each system. Seven evaluators find that our system retrieves more relevant answers compared to the two baselines, while achieving significantly faster processing speeds. We also find that the self-supervised losses help achieve better scores in ROUGE and human evaluation metrics. However, we find that the task remains challenging, with room for improvement. We release our code, model, and matched datasets to encourage further research in consumer health question understanding and answering.

2 Related Work

Consumer Health Question Answering. Ben Abacha et al. (2017) introduce the Medical QA shared task at TREC 2017 LiveQA, where the goal is to develop a consumer health question answering system. The training data is comprised of question-answer pairs. The questions are informally worded CHQs received by the U.S. National Library of Medicine (NLM). The answers are formally worded and come from websites of the U.S. National Institutes of Health or manually collected by librarians. The evaluation scores are given by humans, using a test set of CHQs and reference answers.

Many participating teams adopt a question matching approach, and train their models on question similarity datasets like the Quora question pair dataset (Iyer et al., 2017), or other datasets collected from community question answering websites. TODO (Mrini et al., 2021b)

In the MEDIQA 2019 Shared task, Ben Abacha and Demner-Fushman (2019a) introduce a differently defined consumer health question answering task. Here, the goal is to rank a given list of answers according to their relevance with regard to a CHQ. He et al. (2020) introduce a new disease knowledge infusion training procedure for BERT (Devlin et al., 2019) that scores well in this task.

Medical Question Answering. Medical QA approaches include translating questions to SPARQL queries (Ben Abacha and Zweigenbaum, 2012), semantic similarity between questions and candidate answers (Hao et al., 2019), knowledge representations (Terol et al., 2007; Goodwin and Harabagiu, 2017), ranking candidate answers (Ben Abacha et al., 2017, 2019), summarization of questions and/or answers (Ben Abacha et al., 2021; Mrini et al., 2021d,b,c), and medical entity linking (Basaldella et al., 2020; Mrini et al., 2022).

There is a variety of definitions for the task of medical QA and related sub-tasks in the literature. Hao et al. (2019) define medical QA as the task of finding the correct answer from a set of candidates and a body of evidence documents. They propose to work on two datasets: the National Medical Li-

censing Examination of China (NMLEC) (Shen et al., 2020), and Clinical Diagnosis based on Electronic Medical Records (CD-EMR), where the goal is to predict the correct diagnosis based on patient history.

Sharma et al. (2018) propose to tackle three kinds of medical questions found in the BioASQ challenge (Balikas et al., 2015): factoid questions where answers are single entities, list-type questions where answers are a set of entities, and yes/no questions.

Retrieval-based Question Answering. Recent methods for retrieval-based QA systems use contextual text embeddings to evaluate a candidate answer's relevance to a given question.

Tay et al. (2018) propose to use Multi-Cast Attention Networks (MCAN), a new attention mechanism, to model question-answer pairs.

Mrini et al. (2021e) introduce a recursive, tree-structured model that models sentences according to their syntactic tree. Their results show that tree structure sets a new state of the art in conventional, formally worded QA benchmarks like TrecQA and WikiQA (Yang et al., 2015), but does not fare well in informally worded, user-written datasets.

Karpukhin et al. (2020) introduce Dense Passage Retrieval (DPR): a dual-encoder based on BERT (Devlin et al., 2019), that predicts relevance scores of passages with regard to a question. DPR encoders are trained on the relevance of passages from datasets containing such labels, using a supervised negative log-likelihood loss based on the semantic similarity of questions and relevant passages.

Mao et al. (2021) modify the *query* part of retrieval-based QA: they propose to use language models to generate context for queries. They then feed the extended queries to retrieval systems, such as DPR or BM-25.

3 Problem Definition

We define knowledge-grounded Consumer Health Question Understanding and Answering (CHQUA) as the problem of retrieving a fixed number of answer sentences from a medical knowledge base that are the most relevant given a long and informal user question – called a Consumer Health Question (CHQ). There are three steps in CHQUA: question summarization, matching the summarized user question with a relevant FAQ from the knowledge base, and retrieval of the relevant answer sentences from the corresponding answer document.

Knowledge-grounded CHQUA is comprised of three elements used for training. First, the CHQ is the input of the task. Second, the Reference FAQ (Frequently Asked Question) is the golden or expert-written summary corresponding to the CHQ. Whereas the CHQ is a long and informally worded question, the reference FAQ is the corresponding short, one-sentence, formally worded question. At inference time, the reference FAQ is not available, and we will therefore use a summary generated by the model. Third, the medical knowledge base is comprised of FAQs, where each FAQ has a corresponding answer document with at least one sentence. FAQs in the knowledge base are also short, one-sentence, formally worded questions.

The goal of knowledge-grounded CHQUA is to find a set $\mathcal{R}$ of n relevant answer sentences, from a document comprised of answer sentences $\mathcal{A}_i$, such that $\mathcal{A}_i$ corresponds to question q_i from the knowledge base. We call q_i the retrieved or matching FAQ, such that q_i is the most similar question to the user's summarized question $\mathrm{q_u}$:

$$\mathrm{q}_i = \arg\max_{\mathrm{q} \in \mathcal{Q}} f(\mathrm{q}, \mathrm{q_u}) \tag{1}$$

where $\mathcal{Q}$ is the set of questions (FAQs) in the knowledge base, and f is a given similarity scoring function. $\mathrm{q_u}$ is the reference FAQ (during training) or a generated summary (during inference).

We find the set $\mathcal{R}$ of n relevant answer sentences such that it maximizes the relevance score with the user's summarized question $\mathrm{q_u}$:

$$\mathcal{R} = \arg\max_{\mathcal{R}' \subset \mathcal{A}_i} \sum_{\mathrm{a} \in \mathcal{R}'} g(\mathrm{a}, \mathrm{q_u}) \tag{2}$$

where a is an answer sentence, and g is a given relevance scoring function.

4 Our Pipeline

Our proposed pipeline for Consumer Health Question Understanding and Answering has three main components.

In the first step, our approach learns to *understand* the intent of user questions (CHQs) by summarizing them. We use an encoder-decoder-based summarization model for this step.

The second step is question matching, or the retrieval of the relevant FAQ from the knowledge base: we *ground* the generated summary to a medical knowledge base of FAQs and corresponding

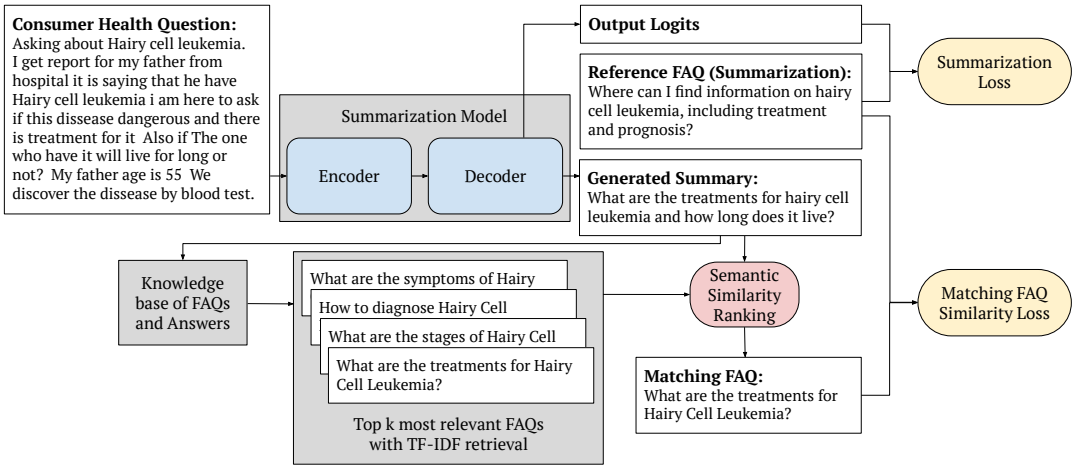

Figure 2: The Consumer Health Question (user question) is first summarized, and we then retrieve a relevant question from the knowledge base using the generated summary. The **top** half of the figure illustrates the first step: question understanding through summarization (§4.1). The **bottom** half of the figure illustrates the second step: question matching through self-supervised knowledge grounding (§4.2).

answer documents. As there are no question matching labels, we consider semantic similarity as a proxy to question matching, and we optimize a self-supervised similarity loss.

The third step is the retrieval of the relevant answer sentences: our model learns to *select* the top-k most relevant answer sentences from the matching answer document. To achieve this task in the absence of answer relevance labels, we consider semantic similarity as a proxy for relevance, and we optimize two novel, semantically-guided, and self-supervised loss functions. The first pushes the model to discriminate between relevant and irrelevant sentences, and the other pushes the model to consider only a fixed number of sentences as relevant.

We show an overview of the model and learning objectives in Figure 1. The entire pipeline is trained together, as the summarizer encoder is re-used to encode the questions and answer sentences.

4.1 Question Understanding through Summarization

Our work aims to flip the burden of question understanding on the question answering model. Instead of asking the user to shorten or reformulate their question, we train an encoder-decoder abstractive summarizer to shorten user questions. Figure 2 illustrates this part of the model.

At training time, we input a Consumer Health Question (CHQ) to the summarization model. The reference Frequently Asked Question (FAQ) is the corresponding shorter and formal question. Given a CHQ embedding $\mathbf{x}$ and the corresponding reference FAQ embedding $\mathbf{y_{ref}}$, the summarization loss is defined as the following negative log-likelihood objective:

$$\mathcal{L}_{\text{sum}} = -\log p(\mathbf{y}_{\text{ref}}|\mathbf{x};\theta) \quad (3)$$

4.2 Question Matching through Self-Supervised Knowledge Grounding

In the next step, we match the summarized user question with the most relevant FAQ from the medical knowledge base. We use semantic similarity as a proxy for question matching, in the absence of such labels.

The knowledge-grounding process is comprised of two steps. First, we use TF-IDF-weighted bag-of-word and n-gram vectors to get the top k most relevant FAQs from the knowledge base. This first step acts as a fast filter to extract a small subset of candidate FAQs. Our retrieval approach follows the retrieval methods commonly used in question answering systems (Chen et al., 2017; Dinan et al., 2018). Dinan et al. (2018) note that the retriever is a potentially learnable part of the model. In our case, using TF-IDF retrieval is computationally optimal and scalable given a large knowledge base with thousands of FAQs. We use a TF-IDF embedder fitted on all the FAQs of the knowledge base, as well as reference FAQs from the training set of the question summarization dataset.

The second step of knowledge-grounding is to

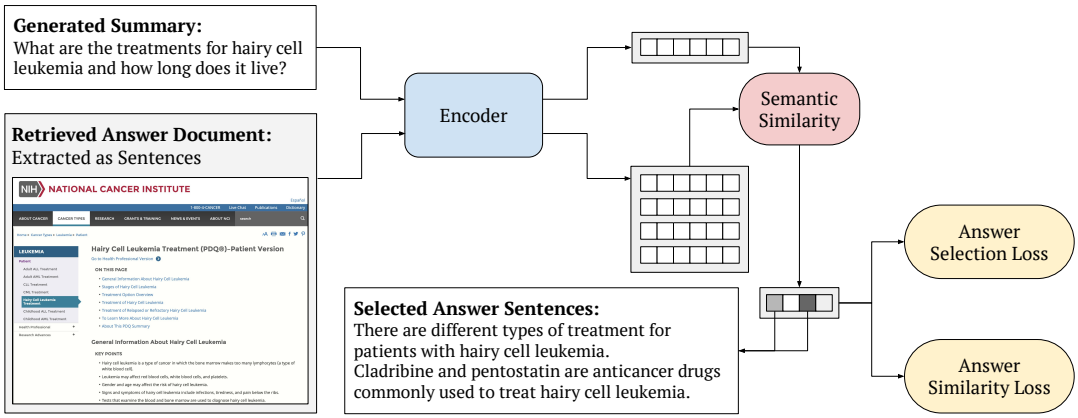

Figure 3: Illustration of the third step of our pipeline: answer retrieval through self-supervised similarity and selection losses (§4.3). Following the same example as Figure 2, our model encodes sentences from the retrieved answer document from the knowledge base, and compares them to the FAQ generated by the summarization model. We use the encoder of the summarization model to embed sentences.

rank the top k FAQs using semantic similarity. To get semantic embeddings of the generated summary and the corresponding top k most relevant FAQs from the knowledge base, we use the encoder of the summarization model. We take inspiration from the precision formula of BERTSCORE (Zhang et al., 2019), and compute the weighted semantic similarity score as follows:

$$\text{Sim}(\text{q}_\text{u}, \text{q}_i) = \sum_{w \in \mathcal{W}_\text{u}} \max_{w' \in \mathcal{W}_i} \frac{\text{idf}(w) \cdot \text{CosSim}(\mathbf{x}_w, \mathbf{x}_{w'})}{\sum_{w'' \in \mathcal{W}_\text{u}} \text{idf}(w'')} \quad (4)$$

where q_u is the reference FAQ (during training) or the generated summary (during inference), q_i is the i-th question from the top k most relevant FAQs, $\mathcal{W}_\text{u}$ and $\mathcal{W}_i$ are the corresponding sets of words, CosSim is the cosine similarity function, and $\text{idf}(w)$ is the inverse document frequency of the word w.

The matching FAQ is the knowledge base FAQ with the highest similarity score with q_u, as shown in the example in Figure 2. During training, the summarization model may produce low-quality or degenerate FAQs. For this reason, at training time, we choose to use the reference FAQ instead to compute the semantic similarity scores and find the matching FAQ. At test time, we only use the generated summary.

Since we are using different datasets for the question summarization and for the knowledge base, we have to reconcile the questions from the knowledge base and the reference questions. We propose to force the model to learn a representation space that does not distinguish between the reference FAQ and the most similar knowledge base FAQ. To accomplish this, we compute the matching FAQ similarity loss. Given the embedding of a summarization reference FAQ q_sum and the embedding of a matching FAQ q_mat, the matching FAQ similarity loss is defined as:

$$\mathcal{L}_\text{mat} = 1 - \text{ReLU}\left(\text{Sim}\left(\text{q}_\text{sum}, \text{q}_\text{mat}; \theta\right)\right) \quad (5)$$

4.3 Answer Retrieval through Self-Supervised Similarity and Selection Losses

After summarizing the user question and retrieving a relevant FAQ from the knowledge base, the next step is to retrieve relevant sentences from the corresponding answer document. In our setting, we need to retrieve a fixed number of sentences relevant to the user question. However, we have no labels for the answer sentences indicating relevance to the user question. We propose two complementary self-supervised learning objectives, that use semantic similarity as a proxy to relevance scoring, and satisfy the constraint of selecting a fixed number of answer sentences.

We show an overview of our answer retrieval approach in Figure 3. In the example of the figure, we show for simplicity a relatively short answer document with four sentences, from which the model chooses the two most relevant ones. In practice, there are close to ten sentences in answer documents.

We compute semantic similarity scores between the generated summary (for inference) or the reference FAQ (for training), and each of the sentences

of the retrieved answer document. We obtain the semantic embeddings of each sentence using the encoder of the summarization model. We then compute semantic similarity scores as shown in equation 4. Cosine similarity scores have values in the $[-1; 1]$ range. For a pair of sentences, a cosine similarity value closer to -1 means that the corresponding sentence embeddings are negatively correlated, or that the sentences have opposite meanings. A value closer to 0 means that the embeddings are not correlated, and that there is no particular semantic relation between the sentences. A value closer to 1 means that the sentence embeddings are positively correlated, and the sentences are close semantically. We consider that a sentence is relevant when the values are closer to 1, and irrelevant otherwise. For this reason, we apply a ReLU activation on the cosine similarity scores before feeding them to the loss functions.

We propose two learning objectives to achieve the self-supervised selection of relevant answer sentences. The semantic similarity loss pushes the model to increase its confidence in the relevance of answer sentences, whereas the answer selection loss pushes the model to select only a fixed number of sentences. The intuition for sharing the encoder with the summarization model, is that these two losses will enable the summarizer to absorb notions of relevance and semantic similarity.

Given the summarization reference FAQ $\mathrm{q_{sum}}$ and the i-th sentence of the retrieved answer document a_i, we compute the ReLU-activated semantic similarity score as follows:

$$S(\mathrm{q_{sum}}, \mathrm{a}_i; \theta) = \mathrm{ReLU}\left(\mathrm{Sim}\left(\mathrm{q_{sum}}, \mathrm{a}_i; \theta\right)\right) \quad (6)$$

We then define the semantic similarity loss $\mathcal{L}_{\mathrm{sim}}$ and the answer selection loss $\mathcal{L}_{\mathrm{sel}}$ as follows:

$$\mathcal{L}_{\mathrm{sim}} = \sum_{i=1}^{|\mathcal{A}|} S(\mathrm{q_{sum}}, \mathrm{a}_i; \theta) * (1 - S(\mathrm{q_{sum}}, \mathrm{a}_i; \theta)) \quad (7)$$

$$\mathcal{L}_{\mathrm{sel}} = \left| \min(n, |\mathcal{A}|) - \sum_{i=1}^{|\mathcal{A}|} S(\mathrm{q_{sum}}, \mathrm{a}_i; \theta) \right| \quad (8)$$

where $\mathcal{A}$ is the set of sentences in the retrieved answer document, and n is the fixed number of sentences to be retrieved.

DATASET SPLIT	TRAIN	DEV	TEST
MeQSum	405	50	50
HealthCareMagic	1,314	164	165

Table 1: Statistics of the medical dataset splits.

The semantic similarity loss $\mathcal{L}_{\mathrm{sim}}$ pushes the semantic similarity values to be either 1 (relevant) or 0 (irrelevant). In combination with $\mathcal{L}_{\mathrm{sim}}$, the answer selection loss pushes the model to only select up to n sentences to have semantic similarity values close to 1. Our system then outputs the sentences with the highest semantic similarity values in the order in which they appear in the answer document. Therefore, the particular semantic similarity ranking of the relevant sentences does not matter – it only matters that relevant sentences have the n highest values.

Finally, the learning objective $\mathcal{L}$ is as follows:

$$\mathcal{L} = \mathcal{L}_{\mathrm{sum}} + \lambda * \mathcal{L}_{\mathrm{mat}} + \gamma * (\mathcal{L}_{\mathrm{sim}} + \mathcal{L}_{\mathrm{sel}}) \quad (9)$$

where λ and γ are hyperparameters. We use only one weight for $\mathcal{L}_{\mathrm{sim}}$ and $\mathcal{L}_{\mathrm{sel}}$ as these two losses are complementary.

5 Experiments and Results

In this section, we evaluate our proposed pipeline for Consumer Health Question Understanding and Answering, and we propose to compare our proposed pipeline against two strong baselines. Seven medical experts judge the performance of our system and baselines by asking their own questions, and rating the relevance of the answers retrieved. Then, we analyze the results through the lens of summarization metrics, human evaluation, and computational speed.

5.1 Datasets

We use one medical knowledge base, MedQuAD (Ben Abacha and Demner-Fushman, 2019b), and two medical question summarization datasets: MeQSum (Ben Abacha and Demner-Fushman, 2019a) and HealthCareMagic (Zeng et al., 2020). All datasets are in English. We show dataset statistics in Table 1.

5.1.1 Dataset Details

MedQuAD is a large-scale Medical Question Answering Dataset. Ben Abacha and Demner-Fushman (2019b) collect trusted medical question-answer pairs by crawling them from 12 websites

of the U.S. National Institutes of Health (NIH). Each web page contains information about a health-related topic, like a disease or a drug. The authors automatically collect the question-answer pairs by composing handcrafted patterns adapted to each website based on document structure and section titles. They manually evaluate 1,721 CHQs to come up with automatic wording patterns for each of 36 question types. Therefore, even though answers are curated and written by medical experts, questions are automatically formulated and may have some noise.

We collect the publicly available (e.g. not copyrighted) question-answer pairs from the MedQuAD dataset[2]. We then use the NLTK sentence tokenizer (Bird, 2006) to split answer documents into sentences. We get 16,423 questions and 157,592 answer sentences, making for an average of 9.6 answer sentences for each question.

MeQSum (Ben Abacha and Demner-Fushman, 2019a) is a medical question summarization dataset released by the U.S. National Institutes of Health (NIH). It contains 1,000 consumer health questions summarized into FAQ-style single-sentence questions by medical experts.

HealthCareMagic is a medical dialogue dataset issued as part of the MedDialog dataset (Zeng et al., 2020)[3]. It is crawled from `HealthCareMagic.com`, an online healthcare service platform. This dataset includes first a formally worded, one-sentence question describing the intent of the patient question, followed by 2 long utterances: a CHQ from the patient that includes a description of the problem and a question, and then an answer from the doctor. To form a medical question summarization dataset, we consider the single-sentence descriptions as summaries of the patient's CHQ. We collect 226,405 question pairs.

5.1.2 Knowledge-based Filtering of Datasets

We conduct experiments for each of the two question summarization datasets, and we use MedQuAD as the underlying knowledge base in all experiments. For this reason, we decide to filter each of the question summarization datasets to reconcile their differences with MedQuAD.

We first fit a TF-IDF embedding model, similar to the one of (Dinan et al., 2018), on the reference FAQs of each question summarization dataset and the questions of MedQuAD. We then compute the dot products of the TF-IDF-weighted vectors for all possible pairs of summarization FAQs and MedQuAD questions. We assign a matching score $m(\mathrm{q_{sum}})$ to each summarization reference FAQ:

$$m(\mathrm{q_{sum}}) = \max_{\mathrm{q}' \in \mathcal{Q}_{\mathrm{MedQuAD}}} \mathrm{tfidf}(\mathrm{q_{sum}}) \cdot \mathrm{tfidf}(\mathrm{q}') \tag{10}$$

We manually evaluate the matching scores for each summarization dataset to set a cutoff matching score of filtering. This way, we obtain question summarization datasets where reference FAQs have matches in the medical knowledge base. Finally, we perform a random and rough 80/10/10 split for the train/dev/test sets. The dataset statistics are in the main paper.

5.2 Training Settings

We adopt the BART encoder-decoder model (Lewis et al., 2020), as it set a state of the art in abstractive summarization benchmarks. We train our model using the HuggingFace implementation (Wolf et al., 2020), on a learning rate of $2 \cdot 10^{-6}$. The question matching pool retrieved by TF-IDF is comprised of $k = 32$ knowledge base FAQs. Our answer selection loss $\mathcal{L}_{\mathrm{sel}}$ is optimized to select up to $n = 3$ sentences. We use $\lambda = 0.01$ and $\gamma = 0.01$ as weights for the self-supervised losses. The BART encoder is used for embedding sentences for question matching and answer selection.

We train for 50 epochs for MeQSum, and 20 epochs for HealthCareMagic. Each training epoch takes about 10 minutes for MeQSum, and about 35 minutes for HealthCareMagic. Inference takes 1 minute for the MeQSum test set and 3 minutes for the HealthCareMagic test set. The best checkpoint is selected based on the lowest loss value $\mathcal{L}$ on the dev set.

We use BART Large pre-trained on the CNN-Dailymail dataset, and each BART Large model contains 406 million parameters, as per the HuggingFace implementation.

5.3 Baselines

We propose the two following baselines in retrieval-based question answering: Dense Passage Retrieval (DPR) (Karpukhin et al., 2020), and Generation-Augmented Retrieval (GAR) (Mao et al., 2021). We adapt these two baselines to our case, and adopt BART-based pre-trained encoders.

[2] https://github.com/abachaa/MedQuAD

[3] https://github.com/UCSD-AI4H/Medical-Dialogue-System

SYSTEM	MeQSum	HealthCareMagic	Time/Query
DPR (Karpukhin et al., 2020)	1.42	1.73	47 seconds
GAR (Mao et al., 2021)	1.40	1.64	48 seconds
Ours	2.13	2.35	2 seconds

Table 2: Evaluation of the relevance (out of 5) of answers retrieved by our proposed system and two strong baselines for questions asked by seven evaluators. The systems trained on MeQSum are evaluated on 60 questions by 3 evaluators, and the ones trained on the larger HealthCareMagic dataset are evaluated on 80 questions by 4 evaluators. The column on the right shows the number of seconds it takes for a loaded system to retrieve the answer to a query.

Similarly to our own pipeline, we create a two-stage retrieval to get answers. The first stage encodes questions from the knowledge base, and retrieves the question that is most relevant to the query. The second stage encodes the corresponding answer document, and retrieves the three sentences that are most relevant to the query.

For DPR, the query is simply the user question. For GAR, we need to generate a context to add to the user question: we choose to add the summary of the user question as the context. We train a BART encoder to summarize user question, using the question summarization datasets.

Whereas our system's retrieval encoder is trained on our proposed self-supervised objectives, the retrieval encoders of the baselines are trained on Wikipedia for the task of retrieval-based question answering.

5.4 Do we retrieve relevant answers?

5.4.1 Evaluation Strategy

We hire seven annotators: four of which are medical doctors, and the remaining three hold degrees related to healthcare or immunology.

We ask the evaluators to first write user questions, and then evaluate the answers retrieved by our system and the two existing systems. Given that our medical knowledge base has limited questions, we ask the evaluators to limit their questions to the topics covered by the nine sources from which the knowledge base was extracted.

Then, we ask the evaluators to rate the relevance of the answers retrieved by each system independently, on a scale of 1 (not relevant) to 5 (relevant). The full description of scores given to the annotators is in the Appendix.

Each of the seven annotators wrote 20 questions, and each question gets three answers (one per system). We assign three annotators to the models trained on MeQSum, and four to the models trained on HealthCareMagic. The annotators rate answers only for the questions that they wrote themselves.

5.4.2 Results and Discussion

We show the results of the evaluations in Table 2. The first three columns show the averages of relevance scores that were given by annotators for all systems.

The results show that the evaluators have preferred our system's answers over the answers retrieved by the two baselines. Our system gets relevance scores that are 0.6 to 0.7 points higher, out of 5 on the relevance scale. An annotator commented that they find our system to be "*more organized and to-the-point than the rest of systems*."[4]

The two baselines seem to perform similarly to each other. This is likely due to the fact that the main difference between them is that the query is generation-augmented for GAR, whereas the query is simply the user question for DPR.

Overall, the relevance scores are on the lower side, as no system exceeds an average score of 2.5/5. This shows that consumer health question answering and understanding is a challenging task, especially since there are no labels to indicate whether an answer is relevant to a particular question, or which FAQ matches the user's intent.

In addition, the challenges of the task are also due to the limitations of the knowledge base. Some annotators noted that the retrieved answers were often not appropriate, or close to the topic but not answering the question. This is due to the fact that MedQuAD does not cover all possible illnesses and medical conditions that the users could ask about. Whereas a larger database would potentially solve coverage problems, it could be at the expense of the quality or verifiability of the answers. The MedQuAD dataset is at times noisy, and contains generic sentences that may not answer any question, or generic templates related to percentages of symptoms and how frequent they are.

[4] Annotators were not told that either system was ours or not. The systems were simply numbered for a blind evaluation.

CRITERIA	Fluency			Coherence			Informativeness			Correctness		
EVALUATION	Win	Lose	Tie	Win	Lose	Tie	Win	Lose	Tie	Win	Lose	Tie
MeQSum	11	5	28	10	6	28	12	3	29	12	4	28
HealthCareMagic	45	17	42	44	19	41	46	18	40	44	18	42

Table 3: Question Understanding evaluation: blind evaluation by 2 annotators of the generated summaries for the test set CHQs. A "Win" evaluation means that our model generates a better summary than the baseline summarizer.

DATASET	MeQSum			HealthCareMagic		
METRIC	R1	R2	RL	R1	R2	RL
GAR (Mao et al., 2021)	45.72	30.43	42.02	31.04	13.68	27.90
Ours	46.74	30.10	42.81	33.13	14.71	30.18

Table 4: Question Understanding evaluation: summarization results on test set (reference FAQs). The R1, R2 and RL metrics refer to the F1 scores of ROUGE-1, ROUGE-2 and ROUGE-L.

5.5 Computational Speed

We run our system on a single 11GB GPU, whereas the two baselines are each run on four 16GB GPUs. We show the average duration required to retrieve answers for a single query in the right column of Table 2.

We notice that, in addition to the higher relevance scores, the advantage of our system is that it is significantly (more than 20 times) faster compared to the two baselines. This is largely due to the fact that we limit to 32 the number of knowledge base questions that we encode and compare the query embedding to. In contrast, DPR and GAR encode all questions in the knowledge base. This is done at the beginning when loading the models, but the query similarity computation is done at each run, thereby lengthening the processing time.

5.6 Analysis of Question Understanding

An additional way that our system outperforms the two baselines could be through summarization. We evaluate the summarization of consumer health questions using the ROUGE metric (Lin, 2004). Our GAR baseline uses a BART model trained on the summarization loss only. We show the results in Table 4. We notice that sharing encoder parameters between the summarization loss and our proposed self-supervised losses generally increases ROUGE F1 scores across both datasets. For HealthCareMagic, score increases exceed 2 points in ROUGE-1 and ROUGE-L.

Given that ROUGE is notoriously unreliable, we hire two additional annotators on Upwork who are healthcare workers to judge the fluency, coherence, informativeness and correctness of generated summaries. We show the annotators the consumer health question (source text), the reference FAQ (target text) and two generated summaries. The annotators do not know which system generated which summary. We show the evaluation scores in Table 3. We remove repetitions of reference FAQs in the test sets put up for evaluation. The results confirm that our self-supervised losses increase the quality of generated summaries. Summaries generated with our model score more wins more often than losses on all four metrics, and score more wins than ties with the summarization-only baseline for HealthCareMagic.

6 Conclusions

We introduce an end-to-end pipeline for knowledge-grounded consumer health question answering and understanding (CHQUA). Our challenge is that we have no labels for question matching or answer relevance. We propose to use semantic similarity as a proxy for those labels, and we design three novel self-supervised losses: one works to match the user's summarized question to a knowledge base question, and the other two losses work complementarily to teach our model to select a fixed number of relevant answer sentences. We compare our proposed system against two strong baselines of retrieval-based question answering. We hire seven medical experts to ask their questions, and they find that our system provides more relevant answers. Our system also achieves processing times that are more than 20 times faster. Finally, we find that our proposed self-supervised losses enable the summarizer model to achieve higher scores in ROUGE and human evaluation metrics, compared to a summarization-only baseline. However, we find that this task remains challenging and that there is still room for improvement. We release our code and model to encourage further research.

Ethical Considerations

Our model is for medical question answering, but should be used with caution as it does not claim

to provide medical advice. Potential users of our system should be warned to not blindly trust the answers given to their medical questions. Potential users should always consult their physician for medical advice.

Each of our annotators spent between two and four hours on the task we gave them. Each annotator was compensated fairly for their work. We answered all of the annotators' questions about the task before they started. Hiring platform Upwork guarantees the payment, fair treatment and informed consent of our nine hired annotators through a mutually agreed-upon contract. The platform fee for Upwork was paid by us, and not deducted from the compensation of the annotators.

Acknowledgements

Khalil Mrini is supported by Adobe Research Unrestricted Gifts, and by an Amazon Research Award under his AWS AI proposal "*Learning Representations for Voice-Based Conversational Agents for Older Adults*". This work is part of the VOLI project (Mrini et al., 2021a; Johnson et al., 2020), and we gratefully acknowledge the award from NIH/NIA grant R56AG067393.

References

Anumeha Agrawal, Rosa Anil George, Selvan Suntiha Ravi, Sowmya Kamath, and Anand Kumar. 2019. Ars_nitk at mediqa 2019: analysing various methods for natural language inference, recognising question entailment and medical question answering system. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 533–540.

Georgios Balikas, Anastasia Krithara, Ioannis Partalas, and George Paliouras. 2015. Bioasq: A challenge on large-scale biomedical semantic indexing and question answering. In *Multimodal Retrieval in the Medical Domain*, pages 26–39, Cham. Springer International Publishing.

Marco Basaldella, Fangyu Liu, Ehsan Shareghi, and Nigel Collier. 2020. Cometa: A corpus for medical entity linking in the social media. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3122–3137.

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*.

Asma Ben Abacha and Dina Demner-Fushman. 2016. Recognizing question entailment for medical question answering. In *AMIA Annual Symposium Proceedings*, volume 2016, page 310. American Medical Informatics Association.

Asma Ben Abacha and Dina Demner-Fushman. 2017. Nlm_nih at semeval-2017 task 3: from question entailment to question similarity for community question answering. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 349–352.

Asma Ben Abacha and Dina Demner-Fushman. 2019a. On the summarization of consumer health questions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2228–2234.

Asma Ben Abacha and Dina Demner-Fushman. 2019b. A question-entailment approach to question answering. *BMC bioinformatics*, 20(1):511.

Asma Ben Abacha, Yassine Mrabet, Yuhao Zhang, Chaitanya Shivade, Curtis Langlotz, and Dina Demner-Fushman. 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.

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.

Asma Ben Abacha and Pierre Zweigenbaum. 2012. Medical question answering: Translating medical questions into sparql queries. In *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium*, IHI '12, page 41–50, New York, NY, USA. Association for Computing Machinery.

Asma Ben Abacha and Pierre Zweigenbaum. 2015. Means: A medical question-answering system combining nlp techniques and semantic web technologies. *Information processing & management*, 51(5):570–594.

Steven Bird. 2006. Nltk: the natural language toolkit. In *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*, pages 69–72.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1870–1879.

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.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2018. Wizard of wikipedia: Knowledge-powered conversational agents. In *International Conference on Learning Representations*.

Travis R. Goodwin and Sanda M. Harabagiu. 2017. Knowledge representations and inference techniques for medical question answering. *ACM Trans. Intell. Syst. Technol.*, 9(2).

Yu Hao, Xien Liu, Ji Wu, and Ping Lv. 2019. Exploiting sentence embedding for medical question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 938–945.

Yun He, Ziwei Zhu, Yin Zhang, Qin Chen, and James Caverlee. 2020. Infusing disease knowledge into bert for health question answering, medical inference and disease name recognition. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4604–4614.

Shankar Iyer, Nikhil Dandekar, and Kornél Csernai. 2017. First quora dataset release: Question pairs.

Janet Johnson, Khalil Mrini, Allison Moore, Emilia Farkas, Ndapa Nkashole, Michael Hogarth, and Nadir Weibel. 2020. Voice-based conversational agents for older adults. In *Proceedings of the CHI 2020 Workshop on Conversational Agents for Health and Wellbeing, Honolulu, Hawaii*.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Chuan Lei, Vasilis Efthymiou, Rebecca Geis, and Fatma Ozcan. 2020. Expanding query answers on medical knowledge bases. In *EDBT*, pages 567–578.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Yuning Mao, Pengcheng He, Xiaodong Liu, Yelong Shen, Jianfeng Gao, Jiawei Han, and Weizhu Chen. 2021. Generation-augmented retrieval for open-domain question answering. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4089–4100.

Khalil Mrini, Chen Chen, Ndapa Nakashole, Nadir Weibel, and Emilia Farcas. 2021a. Medical question understanding and answering for older adults. *The 3rd Southern California (SoCal) NLP Symposium*.

Khalil Mrini, Franck Dernoncourt, Walter Chang, Emilia Farcas, and Ndapa Nakashole. 2021b. Joint summarization-entailment optimization for consumer health question understanding. In *Proceedings of the Second Workshop on Natural Language Processing for Medical Conversations*, pages 58–65, Online. Association for Computational Linguistics.

Khalil Mrini, Franck Dernoncourt, Seunghyun Yoon, Trung Bui, Walter Chang, Emilia Farcas, and Ndapa Nakashole. 2021c. A gradually soft multi-task and data-augmented approach to medical question understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1505–1515, Online. Association for Computational Linguistics.

Khalil Mrini, Franck Dernoncourt, Seunghyun Yoon, Trung Bui, Walter Chang, Emilia Farcas, and Ndapa Nakashole. 2021d. UCSD-adobe at MEDIQA 2021: Transfer learning and answer sentence selection for medical summarization. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 257–262, Online. Association for Computational Linguistics.

Khalil Mrini, Emilia Farcas, and Ndapa Nakashole. 2021e. Recursive tree-structured self-attention for answer sentence selection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4651–4661, Online. Association for Computational Linguistics.

Khalil Mrini, Shaoliang Nie, Jiatao Gu, Sinong Wang, Maziar Sanjabi, and Hamed Firooz. 2022. Detection, disambiguation, re-ranking: Autoregressive entity linking as a multi-task problem. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1972–1983, Dublin, Ireland. Association for Computational Linguistics.

Kirk Roberts and Dina Demner-Fushman. 2016. Interactive use of online health resources: a comparison of consumer and professional questions. *Journal of the American Medical Informatics Association*, 23(4):802–811.

Vasu Sharma, Nitish Kulkarni, Srividya Pranavi, Gabriel Bayomi, Eric Nyberg, and Teruko Mitamura. 2018. Bioama: towards an end to end biomedical question answering system. In *Proceedings of the BioNLP 2018 workshop*, pages 109–117.

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 *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8822–8829.

Yi Tay, Luu Anh Tuan, and Siu Cheung Hui. 2018. Multi-cast attention networks for retrieval-based question answering and response prediction. *arXiv preprint arXiv:1806.00778*.

Rafael M Terol, Patricio Martínez-Barco, and Manuel Palomar. 2007. A knowledge based method for the medical question answering problem. *Computers in biology and medicine*, 37(10):1511–1521.

Thomas Wolf, Julien Chaumond, Lysandre Debut, Victor Sanh, Clement Delangue, Anthony Moi, Pierric Cistac, Morgan Funtowicz, Joe Davison, Sam Shleifer, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45.

Guokai Yan and Jianqiang Li. 2018. Medical question similarity calculation based on weighted domain dictionary. In *Proceedings of the 2018 International Conference on Big Data and Computing*, pages 104–107.

Yi Yang, Wen-tau Yih, and Christopher Meek. 2015. Wikiqa: A challenge dataset for open-domain question answering. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 2013–2018.

Yuan Yang, Jingcheng Yu, Ye Hu, Xiaoyao Xu, and Eric Nyberg. 2017. Cmu livemedqa at trec 2017 liveqa: A consumer health question answering system. *arXiv preprint arXiv:1711.05789*.

Guangtao Zeng, Wenmian Yang, Zeqian Ju, Yue Yang, Sicheng Wang, Ruisi Zhang, Meng Zhou, Jiaqi Zeng, Xiangyu Dong, Ruoyu Zhang, Hongchao Fang, Penghui Zhu, Shu Chen, and Pengtao Xie. 2020. MedDialog: Large-scale medical dialogue datasets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9241–9250, Online. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

A Annotation Details

A.1 Topics Covered by the Knowledge Base

We ask the annotators to limit their questions to the nine sources of MedQUAD. The nine sources from which questions and answer documents are extracted are as follows:

- National Cancer Institute
- Genetic and Rare Diseases Information Center: various aspects of genetic/rare diseases
- Genetics Home Reference (GHR): consumer-oriented information about the effects of genetic variation on human health
- MedlinePlus Health Topics: information on symptoms, causes, treatment and prevention for diseases, health conditions, and wellness issues
- National Institute of Diabetes and Digestive and Kidney Diseases
- National Institute of Neurological Disorders and Stroke: neurological and stroke-related diseases
- NIHSeniorHealth: health and wellness information for older adults
- National Heart, Lung, and Blood Institute (NHLBI): diseases, tests, procedures, and other relevant topics on disorders of heart, lung, blood, and sleep
- Centers for Disease Control and Prevention (CDC)

A.2 Answer Relevance Scoring

We ask annotators to rate answers retrieved by our system and the two baselines according to the following criteria:

- Score of 1/5: The system's answer is completely irrelevant to the question, and does not even contain any concept related to the question.
- Score of 2/5: The system's answer mentions notions that are related to the question, but does not contain a word or concept mentioned in the question.

- Score of 3/5: The system's answer mentions one or more words or concepts from the question, but does not actually answer the question.
- Score of 4/5: The system's answer partially answers the question, mentions one or more words or concepts from the question, but does not fully answer the question.
- Score of 5/5: The system's answer fully answers the question.

A.3 Question Understanding

For question summarization, we evaluate the generated summaries on 4 criteria. We define these criteria for the two healthcare worker annotators as follows:

- Fluency: which generated FAQ is more grammatically correct, and easier to read and to understand?
- Coherence: which generated FAQ is better structured and more organized?
- Informativeness: which generated FAQ captures the most out of the concern of the patient who wrote the CHQ?
- Correctness: which generated FAQ is more factually correct given the CHQ?

A.4 Upwork

We ask annotators to work on Google docs that we share with them. We show in Figure 4 an example of a Google doc that we shared with an annotator (medical doctor) to ask their own question, and the answers we pasted for them to evaluate.

10. TB

    My grandmother was suffering from fever and cough for one months and I was taking care of her. Yesterday she was confirmed as tuberculosis positive and treatment was started. I am worried if I have any risk of being infected by her. However, I have no symptoms at all. Should I get tested for TB?

    System 1: The syndrome usually occurs in children younger than five years; although it has been reported in children up to 13 years. The syndrome is sporadic and non-hereditary. The course of PFAPA can be persistent for years before spontaneous, full resolution.

    System 2: How might periodic fever, aphthous stomatitis, pharyngitis and cervical adenitis be treated? Treatment options that have been successful in improving symptoms of this condition include: oral steroids (prednisone or prednisolone), tonsillectomy with adenoidectomy and cimetidine

    System 3: Tuberculosis (TB) is a disease that is spread through the air from one person to another. Tuberculin skin test: The TB skin test (also called the Mantoux tuberculin skin test) is performed by injecting a small amount of fluid (called tuberculin) into the skin in the lower part of the arm. IGRAs, unlike the TB skin tests, are not affected by prior BCG vaccination and are not expected to give a false-positive result in people who have received BCG.

Relevance Score for System 1: 1/5
Relevance Score for System 2: 1/5
Relevance Score for System 3: 3/5

Figure 4: Example of a Google document, where a hired annotator (medical doctor) asks a question, and rates the answers that we pasted once retrieved by our system and the two baselines.