Getting the Doc's Answers: Using Transfer Learning and Entailment to Answer Patient Questions

Steve Hall
hallsj10@.berkeley.edu

Angelina Wedemeyer
wedemeyea@berkeley.edu

Abstract

In healthcare, automated question-answering is challenging due to the lack of related examples. Questions tend to be specific to the patient and use medical terminology that is relatively rare and potentially 'unknown' in the latest word embedding models. One promising approach to domain-specific question-answering is to retrieve question-answer pairs that are most similar to new questions [1]. We extend this method by leveraging one of the latest NLP transformer models and a large medical question-answer dataset (MedQuAD) curated for this problem. We found that pretrainining a T5 transformer encoder-decoder model on MedQuAD with added context from similar questions produced a modest improvement on patient question-answering.

Keywords: question answering — patient health questions — embedding-based retrieval — question entailment

1 Introduction

A leading focus of research in natural language processing is automated question-answering. This task requires a computer program to understand the context and meaning of the question, identify relevant information from a structured or unstructured database (e.g. dictionary or collection of documents), and then extract (or generate) words or sentences that precisely answer the posed question. For generic question-answering, massive datasets and knowledge bases are available to search for the most probable answer [2]. For specific question-answering like in the healthcare domain, efficiency and accuracy are challenges due to the sparseness of publicly available data.

One survey of healthcare professionals found that respondents spent an average of 60 minutes per database and 3 minutes to examine the relevance of each document for 4 total hours of search time for a given task [3].

Automated healthcare question-answering, therefore, presents a tremendous opportunity for data scientists and healthcare information professionals to help close the gap between the latest in medical research and clinical practice. A faster and more accurate system could lead better to better health outcomes for patients, which is the ultimate motivation for our work.

2 Background

Most short form question answering systems rely on information retrieval and NLP models to extract the correct answer from the given context document(s). These extractive systems can simply rely on a context/type matching heuristic [5]. In other words, these systems aim to select answer spans in the context that match the expected answer type and are similar to the most important question words. These models have performed exceptionally well on clean datasets with the answer contained within the context, but they struggle to perform when the answer is not present. For example, a team of researchers at Stanford leveraged crowdworkers to create a new dataset of question-answer pairs with questions that are unanswerable [6]. The latest state-of-the-art model (at the time of writing) was only able to achieve a F1 score of 66.3% while human accuracy was 89.5%, or a difference of more than 23 points. Applying that same approach to SQuAD, a dataset of questions with answers contained within the provided context, the difference between the model and human accuracy was only 5 points. Clearly, these systems are still far from truly understanding our language.

Another approach to question-answering is to find similar questions that were already answered correctly, which goes all the way back to 2005 [7]. Question entailment is particularly useful when the answer cannot be extracted from the context and in fields like healthcare with higher specificity and heterogeneity. Dr. Asma B. Abacha (previously at the NIH) introduced an end-to-end QA approach that selected entailed questions and then ranked retrieved answers. Her system exceeded the best results of answering medical questions by a factor of 29.8% [1]. These encouraging results, however, were published in January 2019, nearly 18 months before the state-of-the-art T5 model was published in June 2020 [8]. In the conclusion of her paper, she recommended investigating transfer learning to improve model performance.

Unlike generic questions, context for consumer health questions is more difficult to retrieve due to the highly specific nature of the text. Instead, a model could respond using the knowledge it learned during pre-training (i.e. transfer learning). T5, developed to explore the limits of transfer learning, could then be fine-tuned on the downstream task of medical question-answering to learn the relationships and specifics of medical language and question types. Effectively, the model is forced to answer questions on this "knowledge" that it internalized during fine-tuning[11]. As suggested by Dr. Abacha, we also hypothesize that T5 and transfer learning would be beneficial to the downstream task of consumer health question answering.

We seek to extend Dr. Abacha's work by applying the latest in transformer architecture, leveraging the medical question answering dataset she curated, and utilizing an approach similar to Recognizing Question Entailment (RQE) she presents in her study.

3 Approach

In this paper, we explore the following:

1. Model 1. An evaluation of the latest transformer encoder-decoder architecture (T5) question-answering system on TREC 2017 LiveQA medical questions [4], our baseline model.

2. Model 2. A study of transfer learning by training our baseline T5 model on MedQUAD before answering the TREC 2017 Live QA medical questions

3. Model 3. An extension and exploration of Dr. Abacha's question-entailment approach by clustering MedQUAD questions, finding the most similar cluster and questions within that cluster that entail the newly posed question, and then using the answer(s) of those entailed questions as context for the question-answering system

4 Methods

4.1 Data

For model fine-tuning and evaluation, we used two datasets medical question-answer pairs. The statistics of the datasets are as follows:

Task	Data	Train	Valid	Test
Fine-Tuning	MedQuAD	11481	4921	
Fine-Tuning, Testing	TREC-2017 LiveQA	204	102	103

Table 1: Dataset statistics

MedQuAD: To build on Dr. Abacha's research, we leveraged the MedQuAD collection she developed, which consists of 47,457 question-answer pairs extracted from 12 trusted medical sources [1]. MedQuAD was the dataset used in the paper A Question-Entailment Approach to Question Answering in 2019 to retrieve entailed questions to the test questions in the TREC-2017 LiveQA challenge [9]. A subset of this data was removed from the publicly released dataset for proprietary reasons. This data set is used to pretrain the model, and as such, is only split into train and validation sets.

Data	Question
MedQuAD	how is rabies diagnosed?
TREC-2017 LiveQA	A street dog bit me five years ago, I take all the vaccine from very next day from biting, now on that spot where bite there was etching problem since few day, please guide me is there any problem will create in future if create is there any treatment for rabies.

Table 2: Example questions

TREC-2017 LiveQA: As we also aim to focus on consumer health question-answering, we used the TREC-2017 LiveQA medical dataset for pretraining and final evaluation. These questions vary in length, type, and often the question needs to be inferred from the text, making this task more realistic and difficult. This data is split into train and validation for fine-tuning, as well as a test set to perform the final ROUGE evaluation.

From the examples in Table 1, the questions from each dataset have different characteristics. In reviewing the distribution of each dataset, the questions of the TREC-2017 Live QA dataset were 43 words on average, much longer than the 8 word average of the questions in the MedQuAD dataset.

4.2 Models

All models were built on T5 pre-trained models provided by Raffel et al[9] and hosted by huggingface. We fine-tuned using the T5-base model with T5ForConditionalGeneration. We are unable to use the "question:" prompt from T5-base since it was fine-tuned for extractive question-answering, which is not how we will be able to answer these consumer health questions. As a result, we will create a new prompt when fine-tuning each of the following three models:

Model 1: T5-Base TREC-2017 LiveQA To set a baseline for assessing the impact of fine-tuning on a large set of medical question-answer pairs, we fit a T5 encoder-decoder model first on a small list of training questions from the TREC-2017 LiveQA dataset. As the dataset contains only 409 questions we reserved 25% for final testing so that the ROUGE score had a sufficient amount of data to evaluate on. This left us with 204 and 102 question-answer pairs for the training and validation steps, respectively. This training was done with the prefix "'patient medical question: '". Finally,

the model is evaluated by taking the average ROUGE score using the TREC-2017 LiveQA test data.

Model 2: T5-Base TREC-2017 LiveQA, pre-training on MedQuAD To assess the benefits of transfer learning, we fine-tuned T5-base on medical questions using the MedQuAD dataset. First, the model is trained on 11,481 question-answer pairs and validated on 4,921 additional question-answer pairs using the prefix "'research medical question: '". Subsequently, the model is fine-tuned again on the set of TREC-2017 LiveQA data with prefix "'patient medical question: '". Again, the model is evaluated by taking the average ROUGE score using the TREC-2017 LiveQA test data.

Model 3: T5-Base plus additional pre-training on MedQuAD plus similar question-answer context Lastly, we explored the potential advantages of providing the model with additional context to answer new questions. To select similar questions, we used an embedding based retrieval method. The steps of our question retrieval method are outlined below:

1. Remove stop words of all questions to mitigate noise these might create when identifying nearest neighbors in step 7
2. Generate embeddings for each question in the MedQuAD dataset using SentenceTransformers[10]
3. Fit k clusters of questions using k-Means clustering
4. Encode the new premise question using the same encoder model as in step 1
5. Calculate cosine similarity between the premise question and the centroids of k clusters
6. Sort clusters by cosine similarity and select the most similar cluster
7. Find most similar N questions using Nearest Neighbors model
8. Select and concatenate answer(s) for N questions to use as context in question-answering model

After generating context for each question in the MedQuAD and TREC datasets, we went through the same fine-tuning process as we did for Model 2.

4.3 Training

In reviewing the results of the generated questions, we often found the question repeated at the beginning of the answer. We discovered that 18% of the answers begin with a question starting with "What", and 6% of those questions were exact repeats of the original question. This led us to believe that a tactic of answering a patient's question can begin with summarizing the question, before giving the answer. For this reason, we left all answers beginning with questions. Examples of these embedded questions in answers are shown in Table 4.

Question	Answer
Who is at risk for Breast Cancer?	What Is Cancer Prevention? Cancer prevention i..
What is (are) Kidney Disease ?	What the Kidneys Do You have two kidneys. They...

Table 3: Questions in answers

To implement each of these models, we started with the default parameters discussed in the papers. When choosing the optimizer, we started with AdaFactor with learning rate 0.001 as from the original implementation [9]. After observing the loss over several epochs, we tested the potential impact of using the AdamOptimizer, as it had been shown to yield better training loss and generalize well [12]. As you can see in Figure 1, although the AdaFactor does converge faster, the AdamW optimizer is able to achieve the best loss. Given the constraints of this project, we identify this as an area for further research, and use AdamW for all experiments.

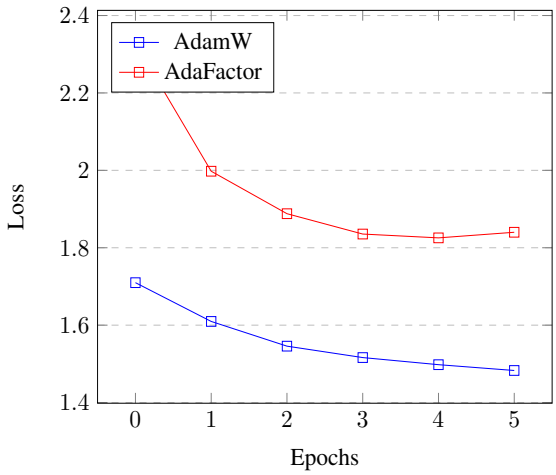

Figure 1: Optimizer AdamW loss compared to AdaFactor over epochs.

5 Results

The results for our baseline and two advanced models are shown in Table 5. We share the ROUGE f1 scores for ROUGE-1, ROUGE-2 and ROUGE-L as well as examples from each model.

Model 1: The baseline model had the lowest ROUGE 1 and ROUGE L scores, but surprisingly produced the best ROUGE-2 score. We expected that this model would exhibit the worst performance, but found some answers to be reasonable as shown in the example in Table 5. Many times even when intelligible, a predicted answer did little more than repeat the question.

	Rouge 1	Rouge 2	Rouge L
Model 1: Fine-tune on TREC-2017	9.40	1.87	7.58
Model 2: Fine-tune on MedQuAD and TREC-2017	10.84	1.52	8.54
Model 3: Model 2 + additional context via question similarity	15.17	1.21	9.29

Table 4: Results

Question	Predicted Answer
I need to lose fat	Diet and exercise can help you lose weight.

Table 5: Model 1: Example predictions

Model 2: Fine-tuning T5 on MedQuAD before attempting to answer TREC-2017 test questions improved the ROUGE-1 and ROUGE-L, but resulted in a lower ROUGE-2 score. Unexpectedly there are more cases of predicted repetition as shown in the example in Table 6. The model does produce correct coherent answers at about the same frequency of the Model 1. One interesting find was a new type of answer that occurred for about 10% of submissions where the model would recommend the patient to go see their doctor or ask their doctor for help.

Question	Predicted Answer
I need to lose fat	- Lose weight. - Maintain a healthy weight. - Maintain a healthy weight.

Table 6: Model 2: Example predictions

Model 3: Adding context via embedding-based question retrieval before fine-tuning T5 on MedQuAD and attempting to answer TREC-2017 test questions improved the ROUGE- 1 and ROUGE- L scores relative to Model 2, but resulted in a lower ROUGE-2 score. Many of the predicted answers, however, are irrelevant, unreadable, or inadequate. For example, the predicted answer to the question in table 8 is related to "I need to lose fat" but does not provide a coherent recommendation before considering a fat-loss program.

Question	Predicted Answer
I need to lose fat	the most common way to find out whether youre overweight or obese is to figure out your body mass index bmi bmi is an estimate of body fat and its a good gauge of your risk for diseases that occur with more body fat...

Table 7: Model 3: Example predictions

6 Analysis of Predicted Answers

Encouragingly, the ROUGE scores did improve on each evolution of the models, yet we find these scores to be inadequate to assess a good answer to a question. To be helpful, the answer must be relevant, readable, and adequate.

Relevancy As shown below, some answers were not relevant to the question asked. We were able to curb the frequency of this type of answer in the dataset by penalizing repetition when generating answers to questions. Although in prior research a repetition penalty of 1.2 was determined as optimal[13], we found 1.5 to provide us with better results. Despite this fix, many of the answers were still not relevant. We hypothesize that this is due to the minimal exposure the model has to medical information, which could help it create an informed answer. To address this in the future, we believe a massive pretraining step on medical information would help to build domain knowledge.

Readability Even when the answers did have relevant content, they were not immediately readable. To address this we increased the beam size to 2 and found the results improved. To further improve readability in the future the model could be trained to unscramble the medical answers to anticipate the expected readability of a response.

Adequacy Finally even when answers were on topic and readable, they often did not answer the patients questions. This could potentially be due to the model not being able to detect the actual question within the source text. As mentioned previously in the data section, all medical pretraining was conducted on questions that averaged 8 words in len, whereas the patients questions on average are more than 5 times as long. A way to address this would be to use an intermediary step to summarize the patient's question. Another way that answers can be inadequate, is due to the many sub questions within a question that are left unaddressed. Next steps that can be taken to train the model on this would be to provide it with other question answer datasets, where the answers have sub questions. Even though the dataset will not be in the health domain, we believe that with a large enough training set there is potential for transfer learning could help the health case question answering.

7 Conclusion

Dr. Asma B. Abacha found that question-entailment is a powerful tool for question-answering in a highly domain-specific field, like healthcare and medicine. We attempted to extend her body of work by using the transfer learning inherent in T5, and the robust dataset she developed to further investigate this problem. Through our research, we are encouraged by the results as the ROUGE-1 f1 score did improve by nearly 6 points compared to our baseline model, but the answers were not consistently relevant, readable, and adequate. Going forward, we would recommend exploring question summarization to reduce the complexity and excessively descriptive information that can be distracting for a NLP model.

8 Acknowledgements

We thank Joachim Rahmfeld from the UC Berkeley School of Information who both introduced the topics and also helped to steer our project into a manageable approach.

9 References

[1] Abacha Ben Asma and Dina Demner-Fushman. A Question-Entailment Approach to Question Answering. BMC Bioinformatics. January 25, 2019.

[2] Z. Dai, L. Li, and W. Xu. CFO: conditional focused neural question answering with large-scale knowledge bases. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers, 2016. URL http://aclweb.org/anthology/P/P16/P16-1076.pdf.

[3] T. Russell-Rose and J. Chamberlain. Expert search strategies: The information retrieval practices of healthcare information professionals. JMIR Med Inform, 5(4):e33, Oct 2017. URL http://medinform.jmir.org/2017/4/e33/.

[4] A. Ben Abacha, E. Agichtein, Y. Pinter, and D. Demner-Fushman. Overview of the medical question answering task at TREC 2017 LiveQA. In Proceedings of The Twenty-Sixth Text REtrieval Conference, TREC 2017, Gaithersburg, Maryland, USA, November 15-17, 2017, 2017. URL https://trec.nist.gov/pubs/trec26/papers/Overview-QA.pdf.

[5] Dirk Weissenborn, Georg Wiese, and Laura Seiffe. Making Neural QA as Simple as Possible but not Simpler. Language Technology Lab, DFKI. Berlin, Germany. June 2017.

[6] Pranav Rajpurkar, Robin Jia, and Percy Liang. Know What You Don't Know: Unanswerable Questions for SQuAD. Computer Science Department, Stanford University. July 2018.

[7] Valentin Jijkoun and Maarten de Rijke. 2005. Retrieving answers from frequently asked questions pages on the web. In Proceedings of the 14th ACM international conference on Information and knowledge management (CIKM '05). Association for Computing Machinery, New York, NY, USA, 76–83. https://doi.org/10.1145/1099554.1099571

[8] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J. Liu. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. Journal of Machine Learning Research 21. June 2020.

[9] Asma Ben Abacha, Eugene Agichtein, Yuval Pinter Dina Demner-Fushman. Overview of the Medical Question Answering Task at TREC 2017 LiveQA. TREC, Gaithersburg, MD, 2017.

[10] Reimers Nils and Iryna Gurevych. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. EMNLP 2019.

[11] Adam Roberts, Colin Raffel, and Noam Shazeer. How much knowledge can you pack into the parameters of a language model? EMNLP 2020

[12] Rohan Anil, Vineet Gupta, Tomer Koren, and Yoram Singer. Memory-Efficient Adaptive Optimization for Large-Scale Learning. In arXiv. 2019

[13] Keskar, N. S., McCann, B., Varshney, L. R., Xiong, C., and Socher, R. CTRL: a conditional transformer language model for controllable generation. arXiv cs.CL, 1909.05858. 2019