UCSD-Adobe at MEDIQA 2021: Transfer Learning and Answer Sentence Selection for Medical Summarization

Khalil Mrini[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, efarcas, nnakashole}@ucsd.edu
[2]Adobe Research, San Jose, CA 95110
{franck.dernoncourt, syoon, bui, wachang}@adobe.com

Abstract

In this paper, we describe our approach to question summarization and multi-answer summarization in the context of the 2021 MEDIQA shared task (Ben Abacha et al., 2021). We propose two kinds of transfer learning for the abstractive summarization of medical questions. First, we train on HealthCareMagic, a large question summarization dataset collected from an online healthcare service platform. Second, we leverage the ability of the BART encoder-decoder architecture to model both generation and classification tasks to train on the task of Recognizing Question Entailment (RQE) in the medical domain. We show that both transfer learning methods combined achieve the highest ROUGE scores. Finally, we cast the question-driven extractive summarization of multiple relevant answer documents as an Answer Sentence Selection (AS2) problem. We show how we can preprocess the MEDIQA-AnS dataset such that it can be trained in an AS2 setting. Our AS2 model is able to generate extractive summaries achieving high ROUGE scores.

1 Introduction

The 2021 **Medi**cal NLP and **Q**uestion **A**nswering (MEDIQA) shared task (Ben Abacha et al., 2021) is comprised of three tasks, centered around summarization in the medical domain: Question Summarization, Multi-Answer Summarization, and Radiology Report Summarization. In this paper, we focus on the first two tasks. In Question Summarization, the goal is to generate a one-sentence formal question summary from a consumer health question – a relatively long question asked by a user. In Multi-Answer Summarization, we are given a one-sentence question and multiple relevant answer documents, and the aim is to compose a question-driven summary from the answer text.

In this paper, we first show that transfer learning from pre-trained language models can achieve very high results for question summarization. Sequence-to-sequence language model BART (Lewis et al., 2020) has achieved state-of-the-art results in various NLP benchmarks, including in the CNN-Dailymail news article summarization dataset (Hermann et al., 2015). We leverage this success and train BART on summarization datasets from the medical domain (Ben Abacha and Demner-Fushman, 2019; Zeng et al., 2020; Mrini et al., 2021). Moreover, we find that training on a different task in the medical domain – Recognizing Question Entailment (RQE) (Ben Abacha and Demner-Fushman, 2016) – can yield better improvements, especially in terms of ROUGE precision scores.

Second, we tackle the extractive track of the multi-answer summarization task, and we cast multi-answer extractive summarization as an Answer Sentence Selection (AS2) problem. A limitation of BART is that the input to its abstractive summarization cannot be as long as the multiple documents in this task. We therefore propose to mitigate this weakness by proposing to cut up the input into pairs of sentences, where the first sentence is the input question, and the second one is a candidate answer. We then train our BART model to score the relevance of each candidate answer with regards to its corresponding question. We also describe in this paper the algorithm used to extract an AS2 dataset from an multi-document extractive summarization dataset.

2 Question Summarization

Our approach to question summarization involves two kinds of transfer learning. First, we train our model to learn from medical summarization datasets. Second, we show that transfer learning from other tasks in the medical domain increases ROUGE scores.

Proceedings of the BioNLP 2021 workshop, pages 257–262

2.1 Training Details

We adopt the BART Large architecture (Lewis et al., 2020), as it set a state of the art in abstractive summarization benchmarks, and allows us to train a single model on generation and classification tasks.

We use a base model, which is trained on BART's language modeling tasks and the XSum abstractive summarization dataset (Narayan et al., 2018). We use a learning rate of $3 * 10^{-5}$ for summarization tasks and $1 * 10^{-5}$ for the recognizing question entailment task. We use 512 as the maximum number of token positions.

Following the MEDIQA instructions and leaderboard, we use precision, recall and F1 scores for the ROUGE-1, ROUGE-2 and ROUGE-L metrics (Lin, 2004).

2.2 Transfer Learning from Medical Summarization

2.2.1 Summarization Datasets

In addition to the XSum base model, we train on two additional datasets. The first dataset is MeQSum (Ben Abacha and Demner-Fushman, 2019). It is an abstractive medical question summarization dataset, which consists of 1,000 consumer health questions (CHQs) and their corresponding one-sentence-long frequently asked questions (FAQs). It was released by the U.S. National Institutes of Health (NIH), and the FAQs are written by medical experts. Whereas Ben Abacha and Demner-Fushman (2019) use the first 500 datapoints for training and the last 500 for testing, participants in this shared task are encouraged to use the entire MeQSum dataset for training.

We also use the HealthCareMagic (HCM) dataset. It is also a medical question summarization dataset, but it is a large-scale dataset consisting of $181,122$ training instances. In contemporaneous work of ours (Mrini et al., 2021), we extract this dataset from the MedDialog dataset (Zeng et al., 2020), a medical dialog dataset collected from `HealthCareMagic.com` and `iCliniq.com`, two online platforms of healthcare service.

The dialogues in the MedDialog dataset consist of a question from a user, a response from a doctor or medical professional, and a summary of the question from the user. We form a question summarization dataset by taking the user question and its corresponding summary, and we discard the answers. We choose to work with HealthCareMagic as the questions are abstractive and resemble the formal style in the FAQs of the U.S. National Library of Medicine (NLM), whereas iCliniq question summaries are noisier and more extractive.

Given that MeQSum is 180 times smaller than HealthCareMagic, we train for 100 epochs on MeQSum, and 10 epochs for HealthCareMagic. We use the validation set of the MEDIQA question summarization task to select the best parameters.

2.2.2 Results and Discussion

We show the validation results in Table 1 and the test results in Table 2. In all test results, we follow approaches of 2019 MEDIQA participants (Zhu et al., 2019), and add the validation set to training for the leaderboard submissions only.

We notice that the validation results for the BART + XSum base model are significantly lower than other models. The corresponding test results are also the lowest-ranking, even though the difference is not as large as we trained on the validation set. These results show that training on an out-of-domain abstractive summarization dataset is not efficient for this task.

We consider now the training on the medical question summarization datasets. First, the validation results show that training on MeQSum achieves comparable F1 scores as training on HealthCareMagic. The main contrasting point is that training on HealthCareMagic yields higher precision, whereas training on MeQSum yields higher recall. This means that training on HealthCareMagic generates summaries with more relevant content, whereas training on MeQSum generates summaries with higher coverage of the content of the reference summaries. However, the corresponding test results show similar recall, but higher precision for HealthCareMagic. Accordingly, ROUGE F1 test scores are higher when training with HealthCareMagic compared to training with MeQSum.

Finally, we consider the results of training on HealthCareMagic followed by MeQSum (HCM + MeQSum). On the validation set, we notice this method generally scores lower precision than just training on HealthCareMagic, but significantly higher recall than any previous training method, therefore achieving higher F1 across all three ROUGE metrics. On the test set, scores are generally comparable with training on HealthCareMagic only.

Metric →	ROUGE-1			ROUGE-2			ROUGE-L		
Model ↓	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
BART + XSum	14.64	27.59	18.48	4.73	9.16	5.97	12.26	23.11	15.46
BART + XSum + MeQSum	27.08	37.05	30.46	10.66	14.43	11.92	25.03	34.37	28.20
BART + XSum + HCM	35.33	27.81	29.64	14.56	10.22	11.40	33.82	26.31	28.16
BART + XSum + HCM + MeQSum	32.14	**40.80**	**35.22**	14.84	18.01	15.92	28.94	**36.66**	31.66
BART + XSum + HCM + RQE	**38.86**	32.97	34.10	**20.31**	15.69	**16.88**	**37.89**	31.98	**33.15**
BART + XSum + HCM + RQE + MeQSum	31.81	40.22	34.52	14.60	**18.22**	15.78	28.82	36.57	31.29

Table 1: Validation results for Question Summarization. HCM is the HealthCareMagic dataset, and RQE is the Recognizing Question Entailment dataset.

Metric →	ROUGE-1			ROUGE-2			ROUGE-L		
Model ↓	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
BART + XSum	28.89	32.86	29.56	10.78	12.19	10.94	26.16	29.65	26.71
BART + XSum + MeQSum	29.88	34.73	30.70	11.69	13.16	11.87	26.71	30.82	27.38
BART + XSum + HCM	31.83	34.31	31.61	13.21	13.81	12.82	28.58	30.75	28.32
BART + XSum + HCM + MeQSum	31.85	35.58	32.00	12.77	13.59	12.51	28.41	31.68	28.53
BART + XSum + HCM + RQE	33.58	35.43	32.65	**14.23**	14.16	13.46	29.51	31.06	28.73
BART + XSum + HCM + RQE + MeQSum	**33.82**	**39.10**	**34.63**	13.91	**15.80**	**14.14**	**29.91**	**34.62**	**30.65**

Table 2: Test results for Question Summarization. All models are trained on the provided validation set as well.

2.3 Transfer Learning from Medical Question Entailment

We consider transfer learning using another task in the medical domain: Recognizing Question Entailment (RQE). Ben Abacha and Demner-Fushman (2016) introduce the RQE task as a binary classification problem, where the goal is to predict whether – given two questions A and B – A entails B. Ben Abacha and Demner-Fushman (2016) further define question entailment as the following: question A entails question B if every answer to B is a correct answer to A, whether partially or fully.

The BART architecture enables us to train on the RQE task using the checkpoint of the question summarization models. BART is an encoder-decoder model that can train, on top of generation tasks, classification tasks as well, such as RQE. We feed the entire RQE question pair as input to both the encoder and the decoder. We add a classification head to be able to predict the entailment score.

2.3.1 Entailment Dataset

For the RQE task, we use the RQE dataset from the 2019 MEDIQA shared task (Ben Abacha et al., 2019). The training set was introduced in Ben Abacha and Demner-Fushman (2016). Similarly to MeQSum, this dataset is released by the U.S. National Institutes of Health. The MEDIQA-RQE dataset contains 8,588 training question pairs. We train for 10 epochs and choose the best parameters using the validation set of the 2021 MEDIQA question summarization task.

2.3.2 Results and Discussion

Similarly to training on HealthCareMagic, we notice in Table 1 that the validation set for training on MEDIQA-RQE yields very high precision scores. This method produces the highest precision scores across all trialled methods, and achieves the highest F1 scores for ROUGE-2 and ROUGE-L. Adding MeQSum to the training (RQE + MeQSum) seems to decrease precision, increase recall, achieve similar ROUGE-1 F1, but lower ROUGE-2 and ROUGE-L F1 scores.

In Table 2, we notice that the test results that the RQE + MeQSum model is the clear winner, providing the highest scores across the board, with the exception of ROUGE-2 precision. Overall, it seems that pre-training on a similar task in the medical domain is beneficial for this medical question summarization task.

3 Multi-Answer Extractive Summarization

3.1 Dataset

The dataset for this task is the MEDIQA-AnS dataset (Savery et al., 2020). It contains 156 user-written medical questions, and answer articles to these questions, such that one question usually has more than one answer article. There are also manually-written abstractive and extractive summaries for the individual answer articles, as well as

for the overall question.

3.2 Casting as Answer Sentence Selection

Given that state-of-the-art summarizer BART can only take relatively short sequences of text as input, we cannot summarize directly from the long answer articles to generate the overall answer summary. We considered summarizing in stages: first training BART to generate summaries for individual answer articles, and then summarize the concatenation of those summaries to generate the answer summary for the user question. However, we only have reference summaries of individual answer articles in the training set of this task, not in the validation or test set. We notice that extractive answer summaries for questions consist of sentences extracted fully from the answer articles. Therefore, we decide to tackle the extractive track of this task, and cast multi-answer extractive summarization as an Answer Sentence Selection (AS2) problem. Similarly to RQE, AS2 is a binary classification task, and as such we are able to train it using BART.

In the AS2 setting, we train BART to predict the relevance score of a candidate answer given a question. To obtain the pairs of questions and candidate answers from the MEDIQA-AnS dataset, we proceed as follows. First, we concatenate for each question the text data of its corresponding answer articles. Then, we use the NLTK sentence tokenizer (Loper and Bird, 2002) to split this text data into individual sentences. Finally, we form question-sentence pairs for AS2 by pairing the user question with each sentence from the corresponding answer article text data.

In this training context, AS2 is a binary classification task, where each pair of question and candidate answer is labeled as relevant (1) or irrelevant (0). We use cross-entropy as the loss function. We label sentences contained in the reference extractive summary as relevant. We notice that some sentences in the reference summary may appear slightly changed in the answer articles, or in exceptional cases may not appear at all. We decide to allow a margin of difference between a reference summary sentence and an answer article sentence, such that if the max-normalized Levenshtein distance between both sentences is 25% or less, we consider the answer article sentence to be relevant. In the rare cases when the reference summary sentence does not appear at all in the answer articles, we add it to our training set and label the sentence as relevant. We show the statistics of the resulting dataset in Table 3.

Set	# sentences	# relevant	% relevant
Train	48,317	3,995	8.27
Dev	2,494	692	27.8

Table 3: Statistics for MEDIQA-AnS cast as an Answer Sentence Selection dataset.

Metric → / **Model ↓**	**Acc.**	**MAP**	**MRR**
BART + XSum + MEDIQA-AnS	71.52	58.63	68.61
BART + XSum + HCM + RQE + MeQSum + MEDIQA-AnS	72.09	57.08	68.52

Table 4: Validation results for Multi-Answer Extractive Summarization, cast as an Answer Sentence Selection problem. We use accuracy and Information Retrieval metrics like Mean Average Precision (MAP) and Mean Reciprocal Rank (MRR).

3.3 Results and Discussion

In Answer Sentence Selection, we use two Information Retrieval metrics for evaluation: Mean Average Precision (MAP) and Mean Reciprocal Rank (MRR). MAP measures how many of the top-ranked answers are relevant, whereas MRR measures how highly a first relevant answer is ranked. We compute the scores as follows, given a set Q of questions:

$$\text{MAP}(Q) = \frac{\sum_{q \in Q} \text{average\_precision}(q)}{|Q|} \quad (1)$$

$$\text{MRR}(Q) = \frac{\sum_{q \in Q} \frac{1}{\text{rank}(q)}}{|Q|} \quad (2)$$

We take as base models the BART + XSum model, as well as the best-performing model in the test set of the question summarization task, as shown in Table 2. We train for 10 epochs on the AS2 version of the MEDIQA-AnS dataset. We show classification and AS2 validation results in Table 4. We notice that both models perform somewhat similarly. Accuracy, MAP and MRR scores are independent of the extractive summary.

We now evaluate the same two models on Multi-Answer Summarization. To form an extractive summary of k sentences, we concatenate the top k most relevant sentences, in the order in which they appeared in the answer articles. We consider two options. First, we generate extractive summaries of

Metric →		ROUGE-1			ROUGE-2			ROUGE-L		
Model ↓	**# sentences ↓**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
BART + XSum + MEDIQA-AnS	Same as ref.	**70.89**	61.48	**65.17**	**53.82**	47.43	**49.99**	**40.28**	34.86	37.00
BART + XSum + MEDIQA-AnS	11	65.13	66.65	61.10	50.45	54.37	48.49	36.57	39.26	35.00
BART + XSum + HCM + RQE + MeQSum + MEDIQA-AnS	Same as ref.	68.53	63.28	65.06	52.09	48.41	49.65	40.10	36.40	**37.77**
BART + XSum + HCM + RQE + MeQSum + MEDIQA-AnS	11	61.84	**67.83**	60.52	46.72	**54.57**	47.08	35.64	**40.53**	35.36

Table 5: Validation results for Multi-Answer Extractive Summarization.

Metric →		ROUGE-1			ROUGE-2			ROUGE-L		
Model ↓	**# sentences ↓**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
BART + XSum + MEDIQA-AnS	11	**61.57**	**67.19**	**60.74**	**47.33**	**53.09**	**47.20**	**43.27**	**48.07**	**42.90**
BART + XSum + HCM + RQE + MeQSum + MEDIQA-AnS	11	59.74	66.34	59.22	45.87	52.21	45.95	42.08	46.98	41.70

Table 6: Test results for Multi-Answer Extractive Summarization.

the same number of sentences as the corresponding reference extractive summary. Second, we generate extractive summaries of 11 sentences, as the average number of sentences in the reference extractive summaries is 10.66. We show validation results in Table 5 and test results in Table 6. For the test results, we are not able to match the number of sentences since we do not have access to the reference summaries. In addition, we train on the validation set as well to report test results, following the approach of MEDIQA 2019 participants (Zhu et al., 2019).

The summarization results on the validation set show that extractive summaries with the same number of sentences as the corresponding reference summaries have higher precision, whereas the 11-sentence extractive summaries have higher recall. Overall, the model trained on BART + XSum fares better than the one fine-tuned on top of question summarization. The test results in Table 6 display the same trend, as the model trained on BART + XSum achieves higher scores across the board. It seems that for this task, transfer learning from other medical datasets was not as useful as for medical question summarization.

4 Conclusions

This paper describes the approach taken by our team, UCSD-Adobe, at the 2021 MEDIQA shared task. We tackle the tasks of question summarization and multi-answer summarization.

For question summarization, we propose two kinds of transfer learning. First, we propose to pre-train on a large-scale dataset of abstractive summarization of medical questions, HealthCareMagic. Our results show that training on this dataset enhances performance in both validation and test sets. Then, we propose to transfer from another medical question-based task: recognizing question entailment. This binary classification task increases performance, and precision scores in particular. In the test results, the highest ROUGE scores are achieved by a model trained on both transfer learning methods.

We tackle the extractive track of the multi-answer summarization task. We propose to cast the question-driven extractive summarization of multiple answer documents as an answer sentence selection problem. We show how we can transform the MEDIQA-AnS dataset into an AS2 dataset. We show that we achieve good ROUGE scores with and without transfer learning from question summarization on the validation set. In the test results, the model without question summarization training achieves the highest ROUGE scores.

Acknowledgments

We gratefully acknowledge the award from NIH/NIA grant R56AG067393. Khalil Mrini is additionally supported by unrestricted gifts from Adobe Research. We thank the anonymous reviewers for their feedback.

References

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. 2019. 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, 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.

Karl Moritz Hermann, Tomás Kociský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *NIPS*, pages 1693–1701.

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.

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

Edward Loper and Steven Bird. 2002. Nltk: The natural language toolkit. In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*, pages 63–70.

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

Shashi Narayan, Shay B Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807.

Max Savery, Asma Ben Abacha, Soumya Gayen, and Dina Demner-Fushman. 2020. Question-driven summarization of answers to consumer health questions. *Scientific Data*, 7(1):1–9.

Guangtao Zeng, Wenmian Yang, Zeqian Ju, Yue Yang, Sicheng Wang, Ruisi Zhang, Meng Zhou, Jiaqi Zeng, Xiangyu Dong, Ruoyu Zhang, et al. 2020. Meddialog: A large-scale medical dialogue dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9241–9250.

Wei Zhu, Xiaofeng Zhou, Keqiang Wang, Xun Luo, Xiepeng Li, Yuan Ni, and Guotong Xie. 2019. Panlp at mediqa 2019: Pre-trained language models, transfer learning and knowledge distillation. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 380–388.