Joint Summarization-Entailment Optimization for Consumer Health Question Understanding

Khalil Mrini[1], Franck Dernoncourt[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, wachang}@adobe.com

Abstract

Understanding the intent of medical questions asked by patients, or Consumer Health Questions, is an essential skill for medical Conversational AI systems. We propose a novel data-augmented and simple joint learning approach combining question summarization and Recognizing Question Entailment (RQE) in the medical domain. Our data augmentation approach enables to use just one dataset for joint learning. We show improvements on both tasks across four biomedical datasets in accuracy (+8%), ROUGE-1 (+2.5%) and human evaluation scores. Human evaluation shows joint learning generates faithful and informative summaries. Finally, we release our code, the two question summarization datasets extracted from a large-scale medical dialogue dataset, as well as our augmented datasets[1].

1 Introduction

In order to answer questions, Conversational AI systems have to first understand the intent of questions (Chen et al., 2012; Cai et al., 2017). This is particularly important for medical conversational agents (Wu et al., 2020), as Consumer Health Questions (CHQ) are often long and contain peripheral information not needed to answer the question. Approaches to medical question understanding include query relaxation (Ben Abacha and Zweigenbaum, 2015; Lei et al., 2020), question entailment recognition (Ben Abacha and Demner-Fushman, 2016, 2019b; Agrawal et al., 2019) and summarization (Ben Abacha and Demner-Fushman, 2019a).

We approach the problem of medical question understanding using joint learning of medical question pairs in the two tasks of question summarization and Recognizing Question Entailment (RQE). Previous work on combining summarization and entailment uses at least two datasets – one for each task. We start from the observation that, given a pair of questions A and B, where A is the longer question, A entails B if and only if B is a summary of A. Using this observation, we propose a data augmentation scheme to use a single dataset for joint learning, instead of two. Then, we propose a simple, simultaneous joint learning approach with fully shared model parameters.

Our findings show that joint learning performs significantly better than single-task training. Our joint learning approach brings about an 8% increase in accuracy in the RQE task compared to single-task training, and shows an average of 2.5% increase in ROUGE-1 F1 scores across three medical question summarization datasets. Additionally, we perform human evaluation and find our approach generates more informative question summaries. Our results suggest the RQE objective makes our summaries more similar in style to the CHQ. Finally, we release the two consumer health question summarization datasets we extracted from an existing large-scale medical dialogue dataset, our augmented datasets and our code.

2 Background and Related Work

2.1 Recognizing Question Entailment (RQE)

The task of RQE was introduced by Ben Abacha and Demner-Fushman (2016) in the context of medical question answering. It is closely related to the task of Recognizing Textual Entailment (RTE) (Dagan et al., 2005, 2013), and early definitions of question entailment (Groenendijk and Stokhof, 1984; Roberts, 1996). Ben Abacha and Demner-Fushman (2016) define RQE as follows: given a pair of questions A and B, question A entails question B if every answer to B is a correct answer to A, and answers A either partially or fully.

2.2 Transfer Learning for Medical QA

Language models that use multi-task learning and transfer learning have become ubiquitous in various

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

Proceedings of the Second Workshop on Natural Language Processing for Medical Conversations, pages 58–65

NLP applications, including BioNLP. BERT (Devlin et al., 2019) has been fine-tuned using biomedical text from PubMed (Beltagy et al., 2019), PMC (Lee et al., 2020), and/or the MIMIC III dataset (Johnson et al., 2016; Huang et al., 2019; Alsentzer et al., 2019). In this paper, we use pre-trained BART models (Lewis et al., 2019).

Transfer learning was a popular approach at the 2019 MEDIQA shared task (Ben Abacha et al., 2019) on medical NLI, RQE and QA. The question answering task involved re-ranking answers, not generating them (Demner-Fushman et al., 2020). For the RQE task, the best-performing model (Zhu et al., 2019) uses transfer learning on NLI and ensemble methods.

In contemporaneous work of ours (Mrini et al., 2021), we participate in the question summarization task of the 2021 MEDIQA shared task (Ben Abacha et al., 2021). We show that transfer learning using medical RQE can improve performance on medical question summarization.

2.3 Summarization and Entailment

There is a growing body of work combining summarization and entailment (Lloret et al., 2008; Mehdad et al., 2013; Gupta et al., 2014).

Falke et al. (2019) use textual entailment predictions to detect factual errors in abstractive summaries generated by state-of-the-art models. Pasunuru and Bansal (2018) propose an entailment reward for their abstractive summarizer, where the entailment score is obtained from a pre-trained and frozen natural language inference model.

Pasunuru et al. (2017) propose an LSTM encoder-decoder model that incorporates entailment generation and abstractive summarization. They use separate natural language inference and summarization datasets, and train by optimizing the two objectives alternatively. Guo et al. (2018) build upon the work of Pasunuru et al. (2017), and add question generation as an auxiliary task.

Li et al. (2018) propose an encoder-decoder summarization model, with an entailment-aware encoder with a separate classification module, and an entailment-rewarded decoder. They follow closely the multi-task setting of Pasunuru et al. (2017).

3 Joint Learning for Consumer Health Question Understanding

We consider the joint learning of medical question summarization and Recognizing Question Entailment (RQE). In both tasks, a question pair includes a first medical question, written in an informal style by a patient – thus called a Consumer Health Question (CHQ). The second medical question is shorter, and often written in a formal style by medical experts: it is a Frequently Asked Question (FAQ). The inspiration for our joint learning scheme stems from the observation that a CHQ entails an FAQ, if and only if the FAQ is a summary of the CHQ.

Our data-augmented joint learning approach to consumer health question understanding has two main components. First, we use our equivalence observation to propose a scheme for data augmentation. Second, we show our joint learning model architecture and learning objective.

3.1 Data Augmentation

Instead of using separate datasets as in previous work, we propose to augment datasets to train jointly, such that we have the same amount of summarization and RQE pairs.

For summarization datasets, we create equivalent RQE pairs. For each existing summarization pair, we first choose with equal probability whether the equivalent RQE pair is labeled as entailment or not. If it is an entailment case, we create an RQE pair identical to the summarization pair. If it is not an entailment case, the CHQ of the RQE pair is identical to the CHQ of the summarization pair, and the FAQ of the RQE pair is a different, randomly selected from the FAQs of the same dataset split.

Inversely, for the RQE dataset, we create equivalent summarization pairs. For each existing RQE pair, we consider two cases. If the RQE pair is labeled as entailment, we create an identical summarization pair. If the RQE pair is labeled as not entailment, we create a summarization pair that is identical to a randomly selected entailment-labeled RQE pair from the same dataset split.

3.2 Joint Model

We adopt the architecture of BART Large (Lewis et al., 2019), a model that set a new state of the art in XSum (Narayan et al., 2018) and CNN-Dailymail (Hermann et al., 2015), two popular abstractive summarization benchmark datasets.

BART is an encoder-decoder seq2seq model, that can train generation as well as classification tasks, such as RQE. BART trains for abstractive summarization by feeding the source text (CHQ) to the encoder, and the negative log-likelihood loss is computed between the decoder output and the

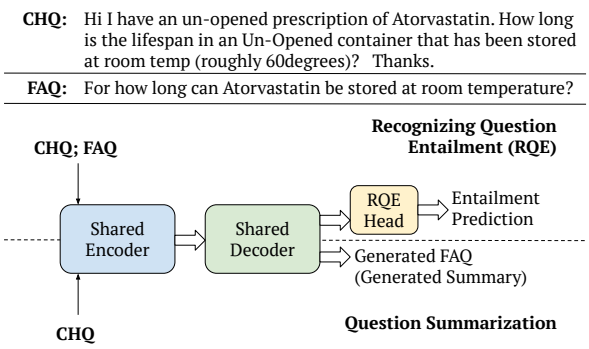

Figure 1: An example medical question pair. The first question is a Consumer Health Question (CHQ) and the second question is a Frequently Asked Question (FAQ). We use BART (Lewis et al., 2019) to jointly train question summarization (bottom) and RQE (top). We show how BART takes input differently for each task.

reference summary (FAQ). BART trains for classification by feeding the full input to the encoder – in the case of RQE, the full input is the concatenation of the CHQ and FAQ. An added classification head attached to the last decoder output then generates a prediction. We compute the binary cross-entropy loss based on the classification head's prediction and the RQE label. We show an overview of our joint training in Figure 1.

We propose to optimize a single loss function that is the sum of the objectives of both tasks. At each training step, we have a summarization question pair that is used for the negative log-likelihood loss, and an RQE question pair that is used for the Binary Cross-Entropy (BCE) loss. Given a CHQ embedding $\mathbf{x}$, the corresponding FAQ embedding $\mathbf{y}$, and the entailment label $l_{entail} \in \{0, 1\}$, we optimize the following loss function:

$$\mathcal{L}_{joint} = -\log p(\mathbf{y}|\mathbf{x};\theta) + \text{BCE}(\mathbf{x}, \mathbf{y}, l_{entail};\theta) \quad (1)$$

For RQE, we consider two loss alternatives, in which we create summarization pairs that are identical to the RQE pairs, regardless of entailment. In the first alternative we simply remove the negative log-likelihood loss for pairs labeled as not entailment. In the second alternative, we flip the negative log-likelihood loss for pairs labeled as not entailment, such that we try to maximize the summarization loss instead of minimizing it.

Dataset	Train	Dev	Test
MeQSum	400	100	500
HealthCareMagic	181,122	22,641	22,642
iCliniq	24,851	3,105	3,106
MEDIQA RQE	8,588	302	230

Table 1: Statistics of the medical dataset splits.

4 Experiment Setup

4.1 Datasets

We consider three medical question summarization datasets and one medical RQE dataset, all in English. Table 1 shows dataset statistics.

(1) 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. The authors used the first 500 datapoints as training and the last 500 as testing. We use a randomly selected 100 datapoints from the training set as our dev set.

We extract the **(2) HealthCareMagic and (3) iCliniq** question summarization datasets from MedDialog (Zeng et al., 2020), a large-scale medical dialogue dataset collected from two online healthcare service platforms: `HealthCareMagic.com` and `iCliniq.com`.

These two datasets include first a one-sentence question describing the medical condition of the patient, followed by two long utterances: one from the patient that includes a description of the problem and a question, and then one from the doctor that includes the response. To form medical question summarization datasets, we consider the single-sentence descriptions as summaries of the patient utterances. HealthCareMagic's summaries are more abstractive and are written in a formal style, unlike iCliniq's patient-written summaries. We create a 80/10/10 split for train/dev/test sets.

(4) MEDIQA RQE is the RQE dataset of the 2019 MEDIQA shared task (Ben Abacha et al., 2019). The test set comprises manually written question pairs, whereas the train and dev sets (Ben Abacha and Demner-Fushman, 2016) are automatically collected. This difference explains the higher dev set results in Ben Abacha et al. (2019). Similarly to MeQSum, the question pairs match a longer CHQ received by the US National Library of Medicine (NLM) and a FAQ from the NIH.

Dataset	MeQSum			HealthCareMagic			iCliniq		
Metric	**R1**	**R2**	**RL**	**R1**	**R2**	**RL**	**R1**	**R2**	**RL**
Seq2seq Attentional Model (Nallapati et al., 2016)	24.8	13.8	24.3	-	-	-	-	-	-
Pointer-Generator Networks (PG) (See et al., 2017)	35.8	20.2	34.8	-	-	-	-	-	-
PG + Data Augmentation (Ben Abacha and Demner-Fushman, 2019a)	44.2	27.6	42.8	-	-	-	-	-	-
PG + Coverage Loss (See et al., 2017)	39.6	23.1	38.5	-	-	-	-	-	-
PG + Coverage Loss + Data Augmentation (Ben Abacha and Demner-Fushman, 2019a)	41.8	24.8	40.5	-	-	-	-	-	-
BART (Lewis et al., 2019)	45.7	26.8	40.8	**44.5**	**22.3**	**39.7**	48.7	28.0	43.5
BART + Data-Augmented Joint Learning	**48.5**	**29.7**	**44.9**	42.1	20.7	37.9	**53.5**	**36.5**	**48.6**

Table 2: Results on the test set comparing BART with and without joint learning of question summarization. The R1, R2 and RL metrics refer to the F1 scores of ROUGE-1, ROUGE-2 and ROUGE-L (Lin, 2004).

4.2 Setup

All of our models use the BART large architecture, with different pre-trained models for transfer learning. For the question summarization experiments, we use the BART Large model pre-trained on the XSum dataset (Narayan et al., 2018). For the RQE experiments, we pre-train a BART Large model on the RTE dataset (Dagan et al., 2005; Haim et al., 2006; Giampiccolo et al., 2007; Bentivogli et al., 2009) from the GLUE benchmark (Wang et al., 2018), and re-use the same classification head for RQE.

4.3 Training Settings

We train for 100 epochs for the MeQSum dataset, and for 10 epochs for all other datasets. We report ROUGE F1 scores for the question summarization datasets, and accuracy for the RQE dataset, as it is a binary classification task with two labels: entailment and not entailment.

For the question summarization datasets, the negative log likelihood on the dev set is used to select the best model. For the RQE dataset, the RQE accuracy on the dev set is the metric used to select the best model.

For single-task training, we use binary cross entropy for RQE, and negative log-likelihood for question summarization.

The learning rate for RQE experiments is 10^{-5} and for the question summarization experiments, it is $3 * 10^{-5}$. We use an Adam optimizer where the betas are 0.9 and 0.999 for summarization, and 0.9 and 0.98 for RQE. In all experiments, the Adam epsilon is 10^{-8}, and the dropout is 0.1.

4.4 Inference

At test time, we evaluate each task completely separately. For RQE, we feed the concatenation of the CHQ and FAQ as input to the model. For question summarization, we only feed the CHQ as input to the model. This way, we ensure that the model never sees the reference FAQs when being evaluated for question summarization.

5 Results and Discussion

5.1 Summarization Results

In their introduction of MeQSum, Ben Abacha and Demner-Fushman (2019a) show results with seq2seq models and pointer-generated networks. They additionally propose to augment MeQSum using semantically selected relevant pairs from the Quora Question Pairs dataset (Iyer et al., 2017). We report these baselines as well as our BART baseline results.

We show our summarization results in Table 2. On MeQSum and iCliniq, our joint learning objective achieves increases between 3 and 8 points across all three metrics – a significant improvement despite MeQSum being extremely low-resource. On the more abstractive and larger HealthCareMagic dataset, there is a decrease of 2 points compared to the BART baseline.

5.2 Human Evaluation

Given that ROUGE is notoriously unreliable, we hire 2 volunteer annotators, and we pick 40 generated summaries from each model in each summarization dataset, resulting in 240 generated summaries (FAQs). We collect 960 evaluations using best-worst scaling. The annotators could also choose to judge both generated FAQs as equal with regards to the given criteria. We show the annotators the generated FAQs in a random order, so that they do not know which model generated which FAQ. We evaluate the generated summaries on 4 criteria:

Datasets	Fluency	Coherence	Informative	Correct
MeQSum	+21.25%	+12.50%	+5.00%	–1.25%
HealthCareMagic	+3.75%	+8.75%	+11.25%	+2.50%
iCliniq	0%	–1.25%	–2.50%	0%

Table 3: Human Evaluation results on 120 samples from the question summarization datasets. The percentages indicate the added value of our joint learning.

Loss Function	Accuracy
Joint Learning	**78.1%**
Removing NLL if not entailment	73.1%
Maximizing NLL if not entailment	72.8%

Table 4: RQE accuracy results on the dev set of our joint loss compared to the two loss alternatives. NLL is Negative Log-Likelihood, the summarization loss.

- 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?

Our human evaluation results are in Table 3. Scores are generally in favor of our approach in MeQSum and HealthCareMagic. There is a high increase in informativeness for HealthCareMagic, and the results for iCliniq show that our approach gives summaries of roughly similar quality as the BART baseline. The ROUGE score increases in the extractive iCliniq and decreases in the abstractive HealthCareMagic indicate that our approach's summaries are more faithful to patient writing styles, suggesting a stronger influence from entailment.

5.3 RQE Results

We compare the joint loss function of equation 1 with the two loss alternatives in section 3.2. We show the results on the dev set in Table 4. Our joint loss function fares the best, exceeding the alternatives by 5%. The results suggest that optimizing RQE jointly with question summarization does help improve performance on the RQE side as well. The difference with the alternative where we remove NLL for not-entailment pairs shows that optimizing our joint learning objective is more efficient than alternating single-task objectives.

We show our RQE results in Table 5. We see an 8% increase on the test set compared to optimizing only on the RQE objective. Our findings show that joint learning helps both tasks equally.

Method	Accuracy
BART (Lewis et al., 2019)	52.1%
Feature-based SVM (Ben Abacha and Demner-Fushman, 2016)	54.1%
BART + Data-Augmented Joint Learning	**60.0%**

Table 5: Accuracy results on MEDIQA RQE test set.

6 Conclusions

We propose a novel data-augmented joint learning approach for the tasks of RQE and question summarization. Our data augmentation method extends a dataset such that it can be used for both tasks. Our results show improvements in both tasks, across three question summarization datasets (+2.5% in ROUGE-1 F1) and one RQE dataset (+8% accuracy). We perform a human evaluation for our generated summaries: we find that our approach generates more informative summaries for formally written FAQs, and summaries that are faithful to patient writing styles in the more extractive iCliniq dataset. Finally, we make our datasets, code and training details publicly available.

Acknowledgments

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

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.

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3606–3611.

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. 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. 2015. MEANS: A medical question-answering system combining NLP techniques and semantic web technologies. *Information processing & management*, 51(5):570–594.

Luisa Bentivogli, Peter Clark, Ido Dagan, and Danilo Giampiccolo. 2009. The fifth PASCAL recognizing textual entailment challenge. In *TAC*.

Ruichu Cai, Binjun Zhu, Lei Ji, Tianyong Hao, Jun Yan, and Wenyin Liu. 2017. A CNN-LSTM attention approach to understanding user query intent from online health communities. In *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 430–437. IEEE.

Long Chen, Dell Zhang, and Levene Mark. 2012. Understanding user intent in community question answering. In *Proceedings of the 21st international conference on world wide web*, pages 823–828.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The PASCAL recognising textual entailment challenge. In *Machine Learning Challenges Workshop*, pages 177–190. Springer.

Ido Dagan, Dan Roth, Mark Sammons, and Fabio Massimo Zanzotto. 2013. Recognizing textual entailment: Models and applications. *Synthesis Lectures on Human Language Technologies*, 6(4):1–220.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Tobias Falke, Leonardo FR Ribeiro, Prasetya Ajie Utama, Ido Dagan, and Iryna Gurevych. 2019. Ranking generated summaries by correctness: An interesting but challenging application for natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2214–2220.

Danilo Giampiccolo, Bernardo Magnini, Ido Dagan, and Bill Dolan. 2007. The third PASCAL recognizing textual entailment challenge. In *Proceedings of the ACL-PASCAL workshop on textual entailment and paraphrasing*, pages 1–9.

Jeroen Antonius Gerardus Groenendijk and Martin Johan Bastiaan Stokhof. 1984. *Studies on the Semantics of Questions and the Pragmatics of Answers*. Ph.D. thesis, Univ. Amsterdam.

Han Guo, Ramakanth Pasunuru, and Mohit Bansal. 2018. Soft layer-specific multi-task summarization with entailment and question generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 687–697.

Anand Gupta, Manpreet Kaur, Shachar Mirkin, Adarsh Singh, and Aseem Goyal. 2014. Text summarization through entailment-based minimum vertex cover. In *Proceedings of the Third Joint Conference on Lexical and Computational Semantics (* SEM 2014)*, pages 75–80.

R Bar Haim, Ido Dagan, Bill Dolan, Lisa Ferro, Danilo Giampiccolo, Bernardo Magnini, and Idan Szpektor. 2006. The second PASCAL recognising textual entailment challenge. In *Proceedings of the Second PASCAL Challenges Workshop on Recognising Textual Entailment*.

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

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2019. ClinicalBERT: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*.

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

Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-Wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. MIMIC-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

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, Ves Stoyanov, and Luke Zettlemoyer. 2019. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Haoran Li, Junnan Zhu, Jiajun Zhang, and Chengqing Zong. 2018. Ensure the correctness of the summary: Incorporate entailment knowledge into abstractive sentence summarization. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1430–1441.

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

Elena Lloret, Oscar Ferrández, Rafael Munoz, and Manuel Palomar. 2008. A text summarization approach under the influence of textual entailment. In *NLPCS*, pages 22–31.

Yashar Mehdad, Giuseppe Carenini, Frank Tompa, and Raymond Ng. 2013. Abstractive meeting summarization with entailment and fusion. In *Proceedings of the 14th European Workshop on Natural Language Generation*, pages 136–146.

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

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar GuÌ‡lçehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, Berlin, Germany. 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*, Brussels, Belgium.

Ramakanth Pasunuru and Mohit Bansal. 2018. Multi-reward reinforced summarization with saliency and entailment. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 646–653.

Ramakanth Pasunuru, Han Guo, and Mohit Bansal. 2017. Towards improving abstractive summarization via entailment generation. In *Proceedings of the Workshop on New Frontiers in Summarization*, pages 27–32.

Craige Roberts. 1996. Information structure: Towards an integrated formal theory of pragmatics. *Semantics and Pragmatics*, 5:6–1.

Abigail See, Peter J Liu, and Christopher D Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355.

Chaochen Wu, Guan Luo, Chao Guo, Yin Ren, Anni Zheng, and Cheng Yang. 2020. An attention-based multi-task model for named entity recognition and intent analysis of chinese online medical questions. *Journal of Biomedical Informatics*, 108:103511.

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.

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.