A Gradually Soft Multi-Task and Data-Augmented Approach to Medical Question Understanding

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

Users of medical question answering systems often submit long and detailed questions, making it hard to achieve high recall in answer retrieval. To alleviate this problem, we propose a novel Multi-Task Learning (MTL) method with data augmentation for medical question understanding. We first establish an equivalence between the tasks of question summarization and Recognizing Question Entailment (RQE) using their definitions in the medical domain. Based on this equivalence, we propose a data augmentation algorithm to use just one dataset to optimize for both tasks, with a weighted MTL loss. We introduce *gradually soft parameter-sharing*: a constraint for decoder parameters to be close, that is gradually loosened as we move to the highest layer. We show through ablation studies that our proposed novelties improve performance. Our method outperforms existing MTL methods across 4 datasets of medical question pairs, in ROUGE scores, RQE accuracy and human evaluation. Finally, we show that our method fares better than single-task learning under 4 low-resource settings.

1 Introduction

In order to retrieve relevant answers, one of the basic steps in Question Answering (QA) systems is understanding the intent of questions (Chen et al., 2012; Cai et al., 2017). This is particularly important for medical QA systems (Wu et al., 2020), as consumer health questions – questions asked by patients – may use a vocabulary distinct from doctors to describe similar health concepts (Ben Abacha and Demner-Fushman, 2019a). Consumer health questions may also contain peripheral information like patient history (Roberts and Demner-Fushman, 2016), that are not necessary to answer questions. There is a growing number of approaches to medical question understanding, including 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; McCreery et al., 2019).

Source User-written Question or *Consumer Health Question (CHQ)*:
SUBJECT: Morgellon Disease. MESSAGE: It appears as if I have had this horrible disease for many, many years and it is getting worst. I am trying to find a physician or specialist in the South Carolina area who can treat me for this medical/mental disease. It seems as if this disease has "NO" complete treatment and it is more least a disability!

Reference Summarized Question or *Frequently Asked Question (FAQ)*:
What are the treatments for Morgellon Disease, and how can I find physician(s) in South Carolina who specialize in it?

BART Trained on Summarization Loss Only (Baseline):
Where can I find physician(s) who specialize in morgellon disease?

Our Gradually Soft Multi-Task and Data-Augmented Model:
Where can I find a physician or specialist in South Carolina who can treat Morgellon Disease?

Figure 1: We highlight the main four aspects of the CHQ. Our method learns from the task of Recognizing Question Entailment to generate more informative summaries compared to the baseline.

Medical question summarization is the task of summarizing consumer health questions into short, single-sentence questions that capture essential information needed to give a correct answer. The task of Recognizing Question Entailment (RQE) is defined by Ben Abacha and Demner-Fushman (2016) in the medical domain as a binary classification task. For the purpose of this task, a first question is considered to entail a second one if and only if every answer to the second question is a correct, and either full or partial answer to the first question.

We find in initial experiments (Mrini et al., 2021b) that RQE can teach question summarizers to distinguish salient information from peripheral details, and likewise that question summarization can benefit RQE classifiers. In our setting, we cast the medical question understanding task as a Multi-

Task Learning (MTL) problem involving the two tasks of question summarization and Recognizing Question Entailment. We use a simple sum of learning objectives in Mrini et al. (2021b). In this paper, we introduce a novel, *gradually soft multi-task and data-augmented approach* to medical question understanding.[1]

Previous work on combining summarization and entailment uses at least 2 datasets – 1 from each task (Pasunuru et al., 2017; Guo et al., 2018). We first establish an equivalence between both tasks. This equivalence is the inspiration behind the data augmentation schemes introduced in our previous work (Mrini et al., 2021b). The goal of the data augmentation is to use a single dataset for Multi-Task Learning. We propose to use a weighted loss function to simultaneously optimize for both tasks. Then, we propose a gradually soft parameter-sharing MTL approach. We conduct ablation studies to show that our two novelties – data augmentation and gradually soft parameter-sharing – improve performance in both tasks.

Our proposed gradually soft multi-task and data-augmented approach outperforms existing single-task and multi-task learning methods on architectures achieving state-of-the-art results in abstractive summarization. Compared to single-task learning, our approach achieves a 12% increase in accuracy on a medical RQE dataset, and an average increase of 3.5% in ROUGE-1 F1 scores across 3 medical question summarization datasets. Additionally, we perform human evaluation and find our approach generates more informative summarized questions. Finally, we find that our approach is more efficient at leveraging smaller amounts of data, and yields better performance under 4 low-resource settings.

2 Background and Related Work

Recognizing Question Entailment (RQE). Ben Abacha and Demner-Fushman (2016) introduce the task of RQE. It is closely related — but not exactly similar — 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).

The task of RQE is to predict, given two pairs of questions A and B, whether A entails B. RQE considers that question A entails question B if every answer to B is a correct answer to A, and answers A either partially or fully. It differs from traditional definitions of entailment, where we consider that the premise entails the hypothesis if and only if the hypothesis is true only if the premise is true.

Ben Abacha and Demner-Fushman (2016) define RQE within the context of Medical Question Answering. The goal is to match a Consumer Health Question (CHQ) to a Frequently Asked Question (FAQ), and ultimately match the CHQ to an expert-written answer.

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 reinforced 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. The authors optimize alternatively between the two tasks, and use separate Natural Language Inference (NLI) and abstractive summarization datasets. Only the decoder parameters are shared.

Li et al. (2018) closely follow the MTL setting of Pasunuru et al. (2017), and propose a model with a shared encoder, an NLI classifier and an NLI-rewarded summarization decoder.

Guo et al. (2018) introduce a pointer-generator summarization model with coverage loss (See et al., 2017). They build upon the work of Pasunuru et al. (2017), and add question generation on top of the two tasks of abstractive summarization and entailment generation. They also alternate between the three different objectives. The authors propose to share all parameters except the first layer of the encoder and the last layer of the decoder, and show that soft parameter-sharing improves over hard parameter-sharing. Their method outperforms the pointer-generator networks of See et al. (2017) on the CNN-Dailymail news summarization baseline. Here, the authors show performance increase in entailment on some batch sizes and decrease on other batch sizes, and they consider entailment as an auxiliary task.

[1] Our code is available at: https://github.com/KhalilMrini/Medical-Question-Understanding

Transfer Learning for Medical QA. BioNLP is one of many NLP applications to benefit from language models that use multi-task learning and transfer learning. There are pretrained language models that are geared towards BioNLP applications, that are based on BERT (Devlin et al., 2019). Those include SciBERT (Beltagy et al., 2019) which has been fine-tuned using biomedical text from PubMed. BioBERT (Lee et al., 2020) has been fine-tuned on the PMC dataset, whereas models named ClinicalBERT (Huang et al., 2019; Alsentzer et al., 2019) additionally use the MIMIC III dataset (Johnson et al., 2016).

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.

3 Methodology

We consider the multi-task learning of medical question summarization and medical RQE. The input to both tasks is a pair of medical questions. The first question is called a Consumer Health Question (CHQ), and the second question is called a Frequently Asked Question (FAQ). The CHQ is written by a patient and is usually longer and more informal, whereas the FAQ is usually a single-sentence question written by a medical expert. The purpose of both tasks is to match a CHQ to an FAQ, and ultimately to an expert-written answer that matches the FAQ. An example pair is shown in Figure 1.

Our novel gradually soft multi-task and data-augmented learning approach to medical question understanding has four main components. First, we establish the equivalence between medical question pairs in question summarization and RQE. Then, we use our equivalence observation to propose a scheme for data augmentation. Third, we show our simultaneous multi-task learning model architecture and learning objective. Finally, we describe our gradually soft parameter-sharing scheme.

3.1 Equivalence of Question Summarization and RQE

In the following, we evidence the equivalence between medical question summarization and medical RQE. We first consider a pair of medical questions C and F, where C is a CHQ and F and is an FAQ, such that C is longer than F.

Ben Abacha and Demner-Fushman (2016) define question entailment as: question C entails question F ($C \Rightarrow F$) if and only if every answer to F is also a correct answer to C, whether partially or completely **(1)**.

According to the guidelines set in the data creation of a medical question summarization dataset by Ben Abacha and Demner-Fushman (2019a), doctors were told to grade manually written summarized questions (FAQs) as perfect, acceptable or incorrect. The two conditions for a perfect FAQ are: first, an FAQ should enable to retrieve "*complete and correct answers*" to the original CHQ, and second, the summarized question should not be so short that it violates the first condition. The resulting medical question summarization dataset includes perfect and acceptable FAQs. We assume that a perfect FAQ provides complete and correct answers to the corresponding CHQ, and that an acceptable FAQ provides correct answers to the corresponding CHQ, whether partially or completely. We therefore conclude that: F is a good summary of C, if and only if F enables to retrieve correct answers to C, whether partially or completely **(2)**.

We have: F enables to retrieve correct answers to C, if and only if answers to F are correct answers to C. Therefore, F enables to retrieve correct answers to C, if and only if every answer to F is also a correct answer to C, whether partially or completely. Given the equivalences **(1)** and **(2)** above, it follows that: question F is a good summary of question C, if and only if question C entails question F **(3)**.

3.2 Data Augmentation

Medical question understanding datasets are scarce, and new high-quality datasets are complex and costly to create. We propose in Mrini et al. (2021b) to augment existing datasets in one of the two tasks to create a synthetic dataset of the same size for the other task. Our two-way data augmentation algorithm is inspired by the equivalence shown in the previous subsection, and enables us to train in a simultaneous multi-task setting. Our data augmentation method also addresses a weakness in previous work in multi-task learning, where each task involves a distinct dataset, often from a different domain. Our data augmentation will enable us to use datasets in the same domain, and we hypoth-

esize this can benefit performance in both tasks.

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 use the equivalence in **(3)** and create an RQE pair identical to the summarization pair. If it is not an entailment case, then we have: **(3)** $\Leftrightarrow$ question F is not a summary of question C if and only if question C does not entail question F **(4)**. Therefore, to create an equivalent RQE pair labeled as not entailement, the RQE CHQ is identical to the CHQ of the summarization pair, and the RQE FAQ is randomly selected from a distinct question pair from 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, then following **(4)**, we create a summarization pair that is identical to a randomly selected and distinct RQE pair labeled as entailment from the same dataset split.

3.3 Simultaneous Multi-Task Learning

Previous work on multi-task learning with summarization and entailment (Pasunuru et al., 2017; Guo et al., 2018) optimize for the objectives of the different tasks by alternating between them. This alternating multi-task training follows a ratio between the different tasks, that depends on the size of the dataset of each task (e.g. a ratio of 10:1 means training for 10 batches on one task, and then for 1 batch on the other task). In our approach, we propose to optimize simultaneously for the objectives of both tasks. We do not use ratios, as we are not alternating between objectives and the resulting datasets from our data augmentation algorithm are of equal size.

Whereas many previous multi-task settings chose generation tasks (entailment generation and question generation), we choose the BART Large architecture (Lewis et al., 2019) as it enables to optimize for a classification task (RQE) and a generation task (summarization) using the same architecture. In addition, BART is adequate as it achieves very strong results in benchmark datasets of recognizing textual entailment and abstractive summarization. The input works differently between both tasks. For summarization, the encoder takes the CHQ as input and the decoder takes the FAQ as input. For RQE, both the encoder and decoder take the entire RQE pair as input. We add a classification head for RQE, to which we feed the last decoder output, as it attends over all decoder and encoder positions. We show an overview of our architecture in Figure 2.

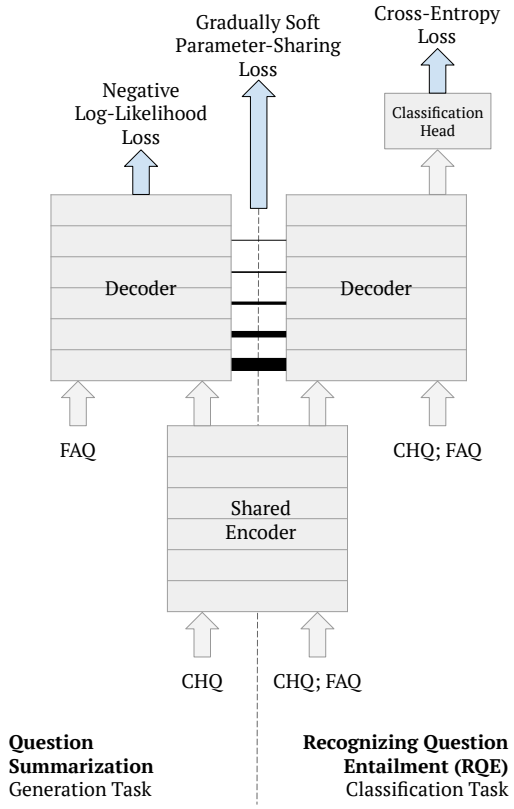

Figure 2: Overview of the architecture of our proposed gradually soft multi-task and data-augmented model. The gradually thinning links between decoder layers represent the loosening parameter-sharing constraint.

We propose to optimize a single loss function that combines objectives of both tasks. Our loss function is the weighted sum of the negative log-likelihood summarization objective, and the binary cross-entropy classification objective of RQE.

More formally, 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 multi-task learning loss function:

$$\begin{aligned}\mathcal{L}_{\text{MTL}}(\theta) = &- \lambda * \log p(\mathbf{y}|\mathbf{x}; \theta) \\ &+ (1 - \lambda) * \text{BCE}\left([\mathbf{x}; \mathbf{y}], l_{entail}; \theta\right)\end{aligned} \tag{1}$$

where BCE is binary cross entropy, and λ is a hyperparameter between 0 and 1.

3.4 Gradually Soft Parameter-Sharing

In multi-task learning, there are two widely used approaches: hard parameter-sharing and soft parameter-sharing. Guo et al. (2018) propose soft parameter-sharing for all parameters except the first layer of the encoder and last layer of the decoder. Liu et al. (2019) introduce MT-DNN and show that hard parameter-sharing of all of the transformer encoder layers, and only having task-specific classification heads produces results that set a new state of the art for the GLUE benchmark (Wang et al., 2018).

We propose a hybrid approach, where we apply hard parameter-sharing for the encoder, and a novel *gradually soft parameter-sharing* approach for the decoder layers. We define gradually soft parameter-sharing as a smooth transition from hard parameter-sharing to task-specific layers. It is a soft parameter-sharing approach that is gradually toned down from the first layer of the decoder to the last layer, which is entirely task-specific.

In gradually soft parameter-sharing, we constrain decoder parameters to be close by penalizing their l_2 distances, and the higher the layer the looser the constraint. Given a decoder with N layers, the gradually soft parameter-sharing loss term is as follows:

$$\mathcal{L}_{\text{GS}}(\theta) = \gamma * \sum_{n=1}^{N-1} \left(e^{\frac{N-n}{N}} - 1 \right) \left\| \theta^{\text{QS}}_{\text{dec},n} - \theta^{\text{RQE}}_{\text{dec},n} \right\|^2 \tag{2}$$

where γ is a hyperparameter, $\theta^{\text{QS}}_{\text{dec},n}$ represents the decoder parameters for the question summarization at the n-th layer, and likewise $\theta^{\text{RQE}}_{\text{dec},n}$ represents the decoder parameters for the RQE task at the n-th layer. We iterate from the 1st to the $(N-1)$-th layer, as the N-th layer is entirely task-specific and unconstrained. We show a high-level representation in Figure 2.

4 Experiments

4.1 Datasets

We consider 3 medical question summarization datasets and 1 medical RQE dataset. We show dataset statistics in Table 1. MeQSum and MEDIQA RQE can be considered low-resource, whereas the other two are far larger. Our datasets are in the English language. Due to space constraints, we briefly introduce the datasets and leave additional details in the appendix.

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.

The medical question summarization datasets are MeQSum (Ben Abacha and Demner-Fushman, 2019a), HealthCareMagic and iCliniq. We extract in Mrini et al. (2021b) and in Mrini et al. (2021c) the HealthCareMagic and iCliniq datasets from the large-scale MedDialog dataset (Chen et al., 2020). Whereas MeQSum is a high-quality dataset from the U.S. National Institutes of Health (NIH), HealthCareMagic and iCliniq are from online healthcare service platforms. HealthCareMagic's summaries are more abstractive and are written in a formal style, unlike iCliniq's patient-written summaries.

The medical RQE dataset is the MEDIQA RQE dataset from the 2019 MEDIQA shared task (Ben Abacha et al., 2019). Similarly to MeQSum, the question pairs match a longer CHQ received by the U.S. National Library of Medicine (NLM) and a FAQ from NIH institutes. Whereas the train and dev sets have automatically generated CHQs, the test set has manually written CHQs. This results in significantly higher dev set results than for test sets, as has been observed during the 2019 MEDIQA shared task.

In addition, we use two pretraining datasets. We use the XSum dataset (Narayan et al., 2018), an abstractive summarization benchmark, for question summarization. For the RQE task, we use the Recognizing Textual Entailment (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).

4.2 Setup and Training Settings

All of our models use the BART large architecture. Unless otherwise noted, all experiments on the 3 question summarization datasets are made using a checkpoint pre-trained on the XSum dataset using only the summarization objective, and all experiments on the RQE dataset are made using a checkpoint pre-trained on the RTE dataset, only optimizing the cross-entropy loss.

We report ROUGE F1 scores for the question

Dataset	MeQSum			HealthCareMagic			iCliniq			RQE
Metric	R1	R2	RL	R1	R2	RL	R1	R2	RL	Accuracy
Ablation of Data Augmentation										
Gradually Soft MTL + Existing Dataset	51.3	32.3	47.5	45.1	22.9	40.3	59.4	46.0	54.5	81.1%
Ablation of Gradually Soft Parameter-Sharing										
Hard-shared Decoder + Data Aug.	52.0	34.0	47.9	44.3	23.3	41.5	60.1	47.0	56.3	77.5%
Soft-shared Decoder + Data Aug.	53.2	35.6	48.9	44.8	22.8	40.9	60.7	48.3	57.8	79.4%
Task-specific Decoder + Data Aug.	50.8	31.7	45.4	46.0	25.1	43.4	61.8	47.5	56.9	81.8%
Our Model										
Gradually Soft MTL + Data Aug.	54.5	37.9	50.2	46.9	24.8	43.2	62.3	48.7	58.5	82.1%

Table 2: Dev set results for the ablation studies on our two main novelties: our data augmentation algorithm, and our gradually soft parameter-sharing method. The R1, R2 and RL metrics refer to the F1 scores of ROUGE-1, ROUGE-2 and ROUGE-L (Lin, 2004).

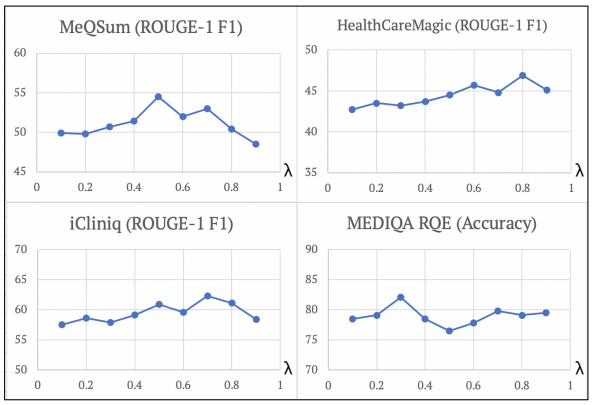

Figure 3: Dev set performance of multi-task learning as a function of the loss hyperparameter λ. The closer λ is to 0, the more the loss focuses on the RQE objective, and vice-versa for the question summarization objective.

summarization datasets, and accuracy for the RQE dataset, as it is a binary classification task with two labels: entailment and not entailment.

The learning rate for RQE experiments is 1×10^{-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. We set the γ hyperparameter to 1×10^{-7}.

4.3 Balancing between the Objectives

Our loss function as defined in Eq.1 has a hyperparameter λ to balance between the question summarization objective and the RQE objective. We run experiments where λ varies from 0.1 to 0.9 in 0.1 increments. The results are in Figure 3. The best λ values are 0.5 for MeQSum, 0.7 for iCliniq, 0.8 for HealthCareMagic and 0.3 for MEDIQA RQE. For the question summarization datasets, we notice that the smaller the dataset, the more it benefits from data-augmented MTL with RQE.

4.4 Ablation Studies

We perform two ablation studies to show the added value of our main novelties: our equivalence-inspired data augmentation algorithm and our gradually soft parameter-sharing algorithm.

Data Augmentation. We compare our data augmentation algorithm against the following alternative: instead of training using a synthetic dataset for the auxiliary task, we choose a separate, existing dataset for abstractive summarization or recognizing textual entailment. This follows the approach taken by most MTL models. For the question summarization task, we optimize the cross-entropy objective using the RTE dataset. For the RQE task, we optimize the summarization objective using the XSum dataset. For the sake of fair comparison, we use the simultaneous MTL objective and the same architecture. Results in Table 2 show a consistent increase in performance across all datasets when using our data augmentation method, suggesting that in-domain MTL is more efficient.

Comparing Parameter-Sharing Configurations. We compare our gradually soft parameter-sharing method with 3 other parameter-sharing configurations. For all configurations, we keep using our data augmentation method, and sharing encoder parameters entirely.

1. Hard-shared decoder: decoder parameters are shared using hard parameter-sharing.

2. Soft-shared decoder: we apply soft parameter-sharing on decoder parameters across all N layers

using the following, unweighted loss term:

$$\mathcal{L}_{\text{S}}(\theta) = \gamma * \sum_{n=1}^{N} \left\| \theta^{\text{sum}}_{\text{dec},n} - \theta^{\text{ent}}_{\text{dec},n} \right\|^2 \quad (3)$$

3. Task-specific decoder: we train two task-specific decoders.

Our ablation study results in Table 2 show that our gradually soft parameter-sharing method exceeds all 3 of the other parameter-sharing configurations in RQE accuracy, and in the sum of ROUGE F1 scores. These results show our proposed smoother parameter-sharing transition between encoder and decoder layers brings about higher performance.

4.5 Results and Discussion

4.5.1 Summarization Results

Baselines. We consider three main baselines. The first one is BART (Lewis et al., 2019), where we only train on the summarization task. The second baseline trains BART on the same MTL settings as Pasunuru et al. (2017), using alternative training with entailment generation on the Stanford Natural Language Inference (SNLI) corpus (Bowman et al., 2015) and having a shared decoder and task-specific encoders. The third baseline trains BART on the same MTL settings as Guo et al. (2018), where, on top of the entailment generation task, we add the question generation task using the Stanford Question Answering Dataset (SQuAD) (Rajpurkar et al., 2016), and all parameters are soft-shared, except for the task-specific first encoder layer and last decoder layer.

In addition, we also report the baselines assessed by Ben Abacha and Demner-Fushman (2019a) for MeQSum. For data augmentation, they use semantically-selected relevant question pairs from the Quora Question Pairs dataset (Iyer et al., 2017). Their results show that coverage loss (See et al., 2017) diminishes the added value of data augmentation in pointer-generator networks. Our summarization-only BART baseline exceeds all of the reported MeQSum baselines in ROUGE-1 F1.

Summarization Results. We report our summarization results in Table 3. Compared to the single-task BART baseline, our gradually soft multi-task and data-augmented method performs better across all three ROUGE metrics, and achieves increases ranging from 1.4 to 5.5 points in ROUGE-1 F1. This differences shows that our method is consistently more efficient compared to training only on summarization.

The other two MTL baselines are generally performing better than the single-task BART baseline, except for the larger HealthCareMagic dataset. We observe that the different parameter-sharing configurations and tasks used in the MTL baselines are scoring about 1 to 4 points below our method in terms of ROUGE-1 F1 scores. This shows that our choice of tasks, simultaneous MTL loss, data augmentation and gradually soft parameter-sharing method work consistently better than existing MTL methods.

Human Evaluation. Given that ROUGE is notoriously unreliable, we hire 2 annotators to judge 120 randomly selected summaries from the summarization test sets, generated from the single-task BART baseline and our own method in Table 3. We ask the annotators to judge the Fluency, Coherence, Informativeness and Correctness of each generated summary, using Best-Worst scaling, with the possibility of ranking both summaries equally. The annotators are presented with 2 generated summaries, in a randomized order at each evaluation, such that they cannot identify which method generated which summary.

Our human evaluation results are in Table 4. Scores generally favor our method, more strongly so in the abstractive datasets – HealthCareMagic and MeQSum. However, we note an increase in correctness for the more extractive iCliniq dataset. On average, our gradually soft multi-task and data-augmented method outputs summarized questions that are more fluent and more informative than the single-task BART baseline.

4.5.2 RQE Results and Discussion

Baselines. We compare our method to three baselines. The first one trains a single-task BART on RQE, with a classification head pre-trained on RTE. The second baseline is a feature-based SVM from Ben Abacha and Demner-Fushman (2016) who introduced the MEDIQA RQE dataset. The third baseline (Zhou et al., 2019) is an adversarial MTL method combining medical question answering and RQE. The architecture consists of a shared transformer encoder using BioBERT embeddings (Lee et al., 2020), separate classification heads for RQE and medical QA, and a task discriminator for adversarial training. A separate dataset is used for medical QA (Ben Abacha et al., 2019).

Dataset	MeQSum			HealthCareMagic			iCliniq		
Metric	R1	R2	RL	R1	R2	RL	R1	R2	RL
Baselines									
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	-	-	-	-	-	-
Models using BART									
BART (Lewis et al., 2019)	45.7	26.8	40.8	44.5	22.3	39.7	48.7	28.0	43.5
BART + Entailment Generation + MTL of Pasunuru et al. (2017)	46.5	27.7	42.3	42.2	20.6	38.1	49.6	29.3	43.8
BART + Entailment Generation & Question Generation + MTL of Guo et al. (2018)	47.2	28.1	42.0	44.7	23.5	41.9	51.4	32.3	46.5
BART + Recognizing Question Entailment + Gradually Soft MTL + Data Augmentation (Ours)	49.2	29.5	44.8	45.9	24.3	42.9	54.2	36.9	49.1

Table 3: Test set results on the 3 question summarization datasets.

Datasets	Fluency	Coherence	Informative	Correct
MeQSum	+11.25%	+2.50%	+7.50%	0%
HealthCareMagic	+6.25%	-2.50%	+12.50%	+1.25%
iCliniq	+2.50%	0%	+3.75%	+5.00%

Table 4: Human Evaluation results on 120 samples from the question summarization datasets. The percentages indicate the added value of our method.

Method	Accuracy
BART (Lewis et al., 2019)	52.1%
Feature-based SVM (Ben Abacha and Demner-Fushman, 2016)	54.1%
BioBERT + Adversarial MTL with Medical QA (Zhou et al., 2019)	63.6%
BART + Summarization + Gradually Soft MTL + Data Aug. (Ours)	64.3%

Table 5: Accuracy results on MEDIQA RQE test set.

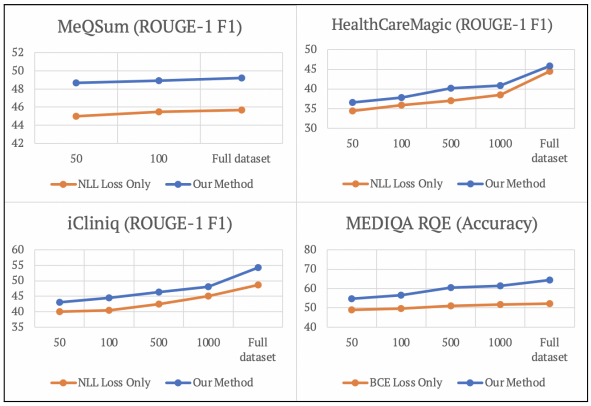

Figure 4: Test set 4-run average performance of our method compared to single-task BART in low-resource settings. Full dataset results are shown for comparison.

RQE Results. We show our RQE results in Table 5. We see a 12% increase on the test set compared to optimizing only on the RQE objective, and 10% increase. Without a separate dataset or embeddings trained on large-scale biomedical data, our method is able to exceed the performance of Zhou et al. (2019) by 0.7%. This confirms the strength of our method, and shows our method can increase performance in both RQE and Question Summarization in the medical domain.

4.6 Performance in low-resource settings

We compare our gradually soft MTL and data-augmented method with the single-task BART baseline on four low-resource settings. For each dataset, we limit the training data to a subset of 50, 100, 500 or 1000 datapoints, and keep the same training settings. To avoid selection bias, we select four random and distinct subsets per low-resource setting, and show average ROUGE-1 F1 scores in Figure 4.

The results show that our approach is able to perform much better in low-resource settings. We notice in particular that, on all 4 datasets, the scores of the single-task BART baseline for 100 and 1000 datapoints are lower than or roughly equal to the scores of our method for a training subset of half the size (50 and 500 datapoints respectively). This suggests that our method's performance increase is not only related to additional datapoints, but also its gradually soft MTL setting.

5 Conclusions

We propose a novel multi-task learning approach for medical question understanding. Our approach trains on the tasks of RQE and question summarization in a simultaneous, weighted MTL loss function, where we add a loss term to constrain the decoder layers to be close, and we loosen the constraint gradually as we move higher up the layers. We show using the definitions of both tasks in the medical domain that we can augment datasets, such that we only need one dataset for MTL. Our two ablation studies show that our gradually soft parameter-sharing and our data augmentation algorithm each increase performance individually. We compare our method to single-task learning and existing MTL work, and show improvements across 3 medical question summarization datasets and 1 medical RQE dataset. Finally, we test our approach under low-resource settings: we find that it is able to efficiently leverage small quantities of data, and that these performance increases do not only depend on additional data from augmentation.

Acknowledgements

We gratefully acknowledge the award from NIH/NIA grant R56AG067393. Khalil Mrini is additionally supported by Adobe Research Unrestricted Gifts. This work is part of the VOLI project (Mrini et al., 2021a; Johnson et al., 2020). 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. 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, 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*.

Samuel Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642.

Ruichu Cai, Binjun Zhu, Lei Ji, Tianyong Hao, Jun Yan, and Wenyin Liu. 2017. An 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.

Shu Chen, Zeqian Ju, Xiangyu Dong, Hongchao Fang, Sicheng Wang, Yue Yang, Jiaqi Zeng, Ruisi Zhang, Ruoyu Zhang, Meng Zhou, Penghui Zhu, and Pengtao Xie. 2020. Meddialog: a large-scale medical dialogue dataset. *arXiv preprint arXiv:2004.03329*.

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*.

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.

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*.

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.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019. Multi-task deep neural networks for natural language understanding. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4487–4496.

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.

Clara McCreery, Namit Katariya, Anitha Kannan, Manish Chablani, and Xavier Amatriain. 2019. Domain-relevant embeddings for medical question similarity. *arXiv preprint arXiv:1910.04192*.

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, 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, Emilias Farcas, and Ndapa Nakashole. 2021c. 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.

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.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392.

Craige Roberts. 1996. Information structure in discourse: Towards an integrated formal theory of pragmatics.

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.

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.

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.

Huiwei Zhou, Xuefei Li, Weihong Yao, Chengkun Lang, and Shixian Ning. 2019. Dut-nlp at mediqa 2019: an adversarial multi-task network to jointly model recognizing question entailment and question answering. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 437–445.

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.