Towards Zero-Shot Conditional Summarization with Adaptive Multi-Task Fine-Tuning

Travis R. Goodwin, **Max E. Savery**, and **Dina Demner-Fushman**
U.S. National Library of Medicine
National Institutes of Health
{firstname.lastname}@nih.gov

Abstract

Automatic summarization research has traditionally focused on providing high quality general-purpose summaries of documents. However, there are many applications that require more specific summaries, such as supporting question answering or topic-based literature discovery. In this paper, we study the problem of conditional summarization in which content selection and surface realization are explicitly conditioned on an ad-hoc natural language question or topic description. Because of the difficulty in obtaining sufficient reference summaries to support arbitrary conditional summarization, we explore the use of multi-task fine-tuning (MTFT) on twenty-one natural language tasks to enable zero-shot conditional summarization on five tasks. We present four new summarization datasets, two novel "online" or adaptive task-mixing strategies, and report zero-shot performance using T5 and BART, demonstrating that MTFT can improve zero-shot summarization quality.

1 Introduction

Transfer learning, in which a model is first pretrained on one or more data-rich tasks before being fine-tuned on a downstream task of interest, has been repeatedly shown to obtain remarkable performance on many natural language processing tasks (Yang et al., 2019; Dong et al., 2019; Liu et al., 2019b). The most successful models resulting from this paradigm rely on self-supervised pretraining with prohibitively-large[1] datasets to facilitate adaptation to new tasks (i.e., fine-tuning) with less abundant data (Devlin et al., 2019; Lewis et al., 2020; Keskar et al., 2019; Raffel et al., 2019). Unfortunately, the benefits of pretraining are reduced for tasks in which there is little direct knowledge transfer, such as language generation for tasks and domains involving previously unseen lexical and semantic properties (as we demonstrate in this paper).

Document: Asthma is a condition in which your airways narrow and swell and produce extra mucus. This can make breathing difficult and trigger coughing, wheezing and shortness of breath. [...]

> **Question:** *What is the consensus of medical doctors as to whether asthma can be cured?*
> **Summary:** Asthma can't be cured, but its symptoms can be controlled. Because asthma often changes over time, it's important that you work with your doctor to track your signs and symptoms and adjust treatment as needed [...]

(a) Health Question (Savery et al., 2020)

Document: The United Nations Thursday set aside $1 million to assess environmental damage caused by this week's devastating tsunami, as reports of destroyed coral reefs and uprooted mangrove forests began trickling in [...]

> **Topic:** *Coral reefs*
> **Summary:** The waves of the tsunami in southeast Asia wreaked tremendous damage on coral reefs, but much more damage occurred when the waves receded, carrying [...]

> **Topic:** *Mangrove Forests*
> **Summary:** The recent 26 December 2004 tsunami in the Indian Ocean with destruction of mangrove forests has highlighted their environmental importance [...]

(b) TAC 2010 (Owczarzak and Dang, 2010)

Figure 1: Example conditional summaries for two tasks.

Transfer learning generalization failures are particularly problematic for a family of tasks we refer to as *conditional summarization*. Unlike traditional summarization, in which the goal is to produce an objective summary of the most salient information in a passage, in conditional summarization, the selection of the most salient points (i.e., content selection), as well as how those points are expressed (i.e., surface realization), are explicitly conditioned on an ad-hoc context, such as a question or topic of interest, as illustrated in Figure 1. In this setting, the same passage may have very

[1] As estimated by Strubell et al. (2019), the cost for training the 11 billion parameter variant of T5 (Raffel et al., 2019) can exceed $1.3 million USD for a single run.

Findings of the Association for Computational Linguistics: EMNLP 2020, pages 3215–3226
November 16 - 20, 2020. ©2020 Association for Computational Linguistics

different ideal summaries, depending on the summarization context, as shown in Figure 1b. Consequently, obtaining sufficient human-authored reference summaries for conditional summarization can be even more time- or cost-prohibitive than for traditional summarization – particularly when dealing with specialized domains such as healthcare.

In this paper, we explore the use of multi-task fine-tuning to enable zero-shot conditional summarization on previously unseen passages for previously unseen tasks. We report the impact of different tasks on zero-shot summary quality and the impact of different task mixing strategies for fine tuning when applied to T5 (Raffel et al., 2019) and BART (Lewis et al., 2020). The primary contributions of this work are:

1. An analysis of the role of 21 question answering, single- and multi-document summarization, causal reasoning, and argumentation tasks on zero-shot domain specific and general domain conditional summarization tasks;
2. Four new summarization datasets that can be used by the community; and
3. Two novel methods for "online" or adaptive task mixing.

2 Background

From its inception, automatic summarization aimed to condense documents either in a generic way – conveying the main points of the document to any user, or focusing on points tailored to specific users and applications, such as topic or query-driven summarization (Mani, 2009). Our aims are even more specific than topic-driven summarization: we are interested in summarizing documents in response to ad-hoc natural language health-related questions asked by the general public. Summarizing information to generate answers to such questions can only rarely be reduced to topic-driven summarization, e.g., if a person is looking for general information about a given health condition or treatment; in over 90% of cases, health questions are more specific and focus only on particular aspects of the topic (Demner-Fushman et al., 2019). For example, people may be looking for medications for a specific condition or asking how to store a drug. The summary, therefore, has to be tailored not only to the topic of the question and task but must also be restricted only to the aspects of the topic that directly address the question. Consider the following question from a user of a question answering system: *When your legs start cramping when you lay down & diabetic, what vitamin are you deficient in?* To answer this question, the summary must provide information about supplements (the topic), but only information about supplements indicating how the supplement can prevent or alleviate night leg cramps in diabetic patients. Selection of the content that needs to be extracted or generated in the summaries must be question-driven.

In the open domain, previous community efforts to focus on topic driven summarization include the Document Understanding Conference (DUC) and its successor, the Text Analysis Conference (TAC), both of which organized topic-based summarization tasks. In various iterations of these tasks, human assessors developed topic statements and documents cluster for those topics, and then manually authored summaries based on the topic statements. The tasks' participants were asked to develop automatic summarization approaches for generating single- or multi-document summaries that contained information relevant to the topic statement. Other community efforts involving summarization include the BioASQ[2], CL-SciSumm[3], and Scholarly Document Processing[4] challenges that involve summarization of scientific articles. However, despite the attention that summarization has received in the natural language processing community and the recent development of more sophisticated summarization algorithms, the task of automatically generating human-quality still poses many challenges.

A study of content selection across multiple domains, including medical articles, indicates that new forms of sentence representations and external knowledge sources are needed to identify the most suitable approaches to summarization (Kedzie et al., 2018). Recent work has shown models with transformer-based architectures, coupled with unsupervised pretraining approaches, to achieve state of the art results in many text generation tasks. Building on this, researchers have recently shown that these models can be conditioned on a prompt included in the input text. For example, this prompt can guide the content of the generated text towards either a desired topic (Keskar et al., 2019) or instruct the model to produce output for a specific task (Lewis et al., 2020; Raffel et al.,

[2] http://bioasq.org/

[3] https://github.com/WING-NUS/scisumm-corpus

[4] https://ornlcda.github.io/SDProc/

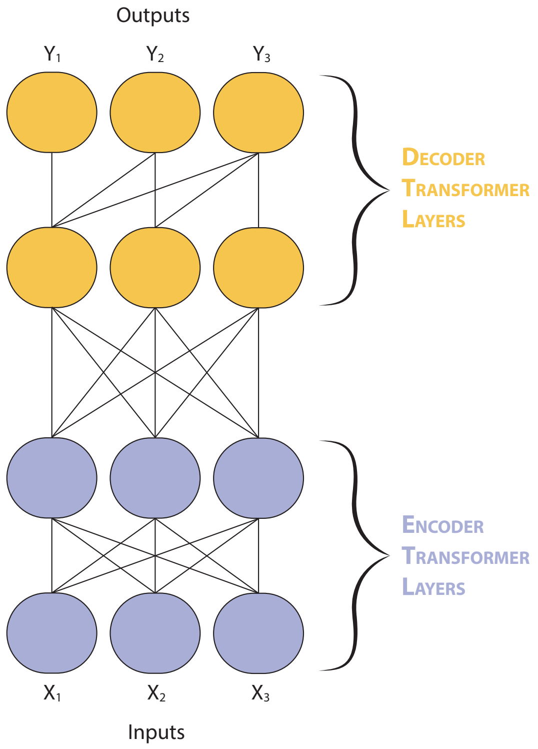

Figure 2: Simplified encoder-decoder transformer architectures used by BART and T5.

2019). Similar work on conditional generation includes Liu et al. (2020), in which the authors condition an extractive transformer on control codes specifying position, importance, and diversity of the sentences in the source text.

There have been relatively few publications focused on zero-shot learning specifically for summarization. Duan et al. (2019) experiment with zero-shot learning for cross-lingual sentence summarization, while Liu et al. (2019a) explored zero-shot abstractive summaries of five-sentence stories.

Prior work indicates that topic and question-driven summarization can be formulated as a text-to-text, conditional generation problem in which content selection and source realization are explicitly conditioned on a user-specified prompt. The formulation of summarization in this way intuitively dovetails with the desired goal described above: question-driven summarization of answers to user's health-related questions. In this study, we extend previous work done with BART and T5 using multi-task fine-tuning using a large body of tasks and exploring multiple mixing strategies to advance topic and question-driven summarization in the open and medical domains.

3 Models

Several transformer-based models have been shown to generate high quality natural language (Peters et al., 2018; Radford et al., 2018; Wang and Cho, 2019). The majority of these models cast summarization as language modeling wherein the input to the model is the sequence of words in the source document followed by a mask token for each word in the desired summary (Keskar et al., 2019; Radford et al., 2019). This substantially limits the length of summaries that can be generated due to the input sequence limits imposed during pretraining. Fortunately, more recent approaches use separate transformers for encoding and decoding, allowing the generation of potentially arbitrary length sequences. In this work, we explored the two most notable of these approaches: BART and T5.

BART (Bidirectional and Auto-Regressive Transformers) is pre-trained with sentence ordering and token in-filling tasks (Lewis et al., 2020). BART uses a separate bidirectional encoder and autoregressive decoder similar to BERT except that (1) BART's decoder incorporates cross attention over the final encoder layer and (2) BART's encoder does not use a feed-forward dense layer for word prediction. In our experiments, we used `BART-Large`, which includes 12 transformer layers in the encoder and decoder.

T5 (Text-to-Text Transfer Transformer) uses several pretraining objectives, including unsupervised fill-in-the-blank as well as supervised translation, summarization, classification, and reading comprehension tasks where each task is represented as a language generation task (Raffel et al., 2019). T5 closely follows the originally-proposed Transformer architecture (Vaswani et al., 2017) except using relative positional embeddings rather than sinusoidal encoding. In this work, we used `T5-Base`, which includes 12 transformer layers.

4 Adaptive Multi-task Fine-Tuning

We adapt the text-to-text setting used to pre-train T5 (Raffel et al., 2019) to enable fine-tuning on a large body of tasks with the intent of injecting knowledge from related natural language processing tasks to enable improved zero-shot conditional summarization. In this section, we describe (1) the fine-tuning tasks used in our experiments, (2) how these tasks are encoded as text-generation, and (3) approaches for task mixing.

4.1 Fine-Tuning Tasks

We considered a total of 21 tasks related to summarization, question answering, commonsense reasoning, and argumentation; new summarization datasets or new extensions of previous datasets are denoted with an '*'.

BioASQ is a challenge for medical semantic indexing and question answering (QA) (Tsatsaronis et al., 2015). The QA challenges provide participants with questions, PubMed articles, snippets extracted from those articles, and human-generated answers to the questions. For single-document summarization, we used each extracted snippet as a summary of the corresponding article. For multi-document summarization, we used each human-generated answer as a summary of the corresponding set of articles. The single-document summarization dataset contains 27.1 K examples, and the multi-document summarization dataset contains 3.2 K examples.

CNN/DailyMail includes 287.1 K news articles, as well as highlights of the articles which are used as summaries (See et al., 2017; Hermann et al., 2015).

CoPA The Choice of Plausible Alternatives dataset (Roemmele et al., 2011) presents 400 training sets of questions involving choosing the most plausible cause or effect entailed by a given premise; questions were drawn from (1) personal blog stories (Gordon and Swanson, 2009), and (2) subject terms from the Library of Congress Thesaurus for Graphic Materials.

Cochrane* contains 5.0 K reviews and plain language summaries from the Cochrane Database of Systematic Reviews; we use only the main body of the review as the source document for single-document summarization.

Cosmos QA includes 287.1 K multiple-choice reading comprehension questions requiring commonsense causal reasoning; it focuses on cause and effect in everyday narratives (Huang et al., 2019).

CQaD-S* is based on a collection of consumer questions about drugs and answers to those questions manually extracted from reliable web pages (Ben Abacha et al., 2019); we adapted the 272 manually selected sections as question-driven summaries of their source documents.

EBM is a collection of Evidence-Based Medicine summaries, including questions, answers, justifications of those answers, and the references for those justifications (Molla and Santiago-Martinez, 2011). We adapted it for two multi-document summarization tasks: **EBM Answers***, using the answers as the summary and the abstracts from the reference articles as the set of source documents and **EBM Justifications***, using the reference articles and the answer as the source text and the justification for the answer as the summary. This produced 1.2 K and 2.8 K examples, respectively.

IBM Evidence 4.3 K examples of questions with pairs of evidence, annotated for which evidence in the pair is the most convincing evidence for answering the question; the training set includes 48 topics (Shnarch et al., 2018).

MC-TACO is a set of 13 K question-answer pairs requiring temporal commonsense comprehension; questions pertain to various temporal aspects of events, such as duration, frequency, and temporal order (Zhou et al., 2019).

MedlinePlus Summaries* contains summaries of health topics obtained from MedlinePlus, a service of the U.S. National Library of Medicine providing human-curated, reliable, and easy-to-understand articles about over 1 K health topics. Each article contains a summary of the topic and links relevant web pages; we used the summary and the content of linked pages[5] to generate a multi-document summarization collection consisting of 969 examples.

Movie Rationales contains 1.6 K human annotated rationales for movie reviews; used as multi-document summarization (Zaidan et al., 2008; DeYoung et al., 2020).

PubMed PubSum* contains publisher-submitted summaries of PubMed articles written in consumer-friendly language; we collected 240 articles with accompanying summaries as single-document summarization.

[5] We considered links provided in the *Start Here* and *Learn More* sections of MedlinePlus.

QA4MRE was created for the CLEF shared tasks to promote research in question answering and reading comprehension; we used the English questions provided for training in 2011, 2012, and 2013 including 120, 160, and 284 examples, respectively, as well as the Alzheimer's questions provided in 2012 and 2013 which each provide 40 examples (Peñas et al., 2013).

Scientific Papers contains two sets of long documents and their abstracts, including 203.0 K articles from arXiv.org and 119.9 K articles from the Open Access Subset of PubMed Central® (Cohan et al., 2018).

SQuAD the Stanford Question Answering Dataset is a reading comprehension dataset consisting of 87.6 K questions over Wikipedia articles where the question is considered unanswerable if the answer cannot be extracted from the corresponding passage (Rajpurkar et al., 2016).

4.2 Conditional Generation

As in Raffel et al. (2019), we used a Text-to-Text setting to train BART and T5 such that the model inputs and targets are both encoded as sequences of tokens. For summarization tasks, the input was provided as `<task-name> [question:<question>] summarize:<document>` and the target was the target summary, where the conditional summarization context (if applicable) is provided as the question portion of the input. For question answering and reading comprehension tasks, the input was provided as `<task-name> question:<question> [choice:<choice>...] context:<document>` and the target was either (a) `True` or `False` for (binary choice questions), or (b) the text of the correct choice for n-ary choice questions.

4.3 Task Mixing

Neural models are notorious for overfitting data – particularly in the case of natural language text for which transformer-based models have been shown to memorize spurious cues (Niven and Kao, 2019). A major factor in overfitting is the size of data used for training, and, as documented in Section 4.1, the available training data for each of our fine-tuning tasks vary by orders of magnitude. In order to avoid overfitting (and to avoid overcorrecting and underfitting) small datasets, for each fine-tuning step, we sample a batch of data from a single task assuming a Multinomial distribution θ over fine-tuning tasks. We refer to this distribution over tasks as the *mixing rate*, such that θ_t indicates the probability that a batch will be drawn from fine-tuning task $t \in \{1, \cdots, K\}$. We explored four approaches to defining the mixing rate: proportional and temperature-scaled task mixing as in Raffel et al. (2019) and two novel "online" approaches, i.e., adaptive and self-adaptive mixing.

Proportional Mixing The most intuitive way to avoiding overfitting is to define the mixing rate based on the proportion of data in each task compared to the total amount of data over all tasks. Formally, let N_t be the size of the training set for task t. In proportional mixing, we define:

$$\theta_t^{(\mathrm{PM})} = \min(\eta, N_t) \div \sum_{t'} \min(\eta, N_{t'}) \qquad (1)$$

where η is a maximum data size constant used to prevent large datasets from dominating θ. In our experiments we used $\eta = 2^{14}$.

Temperature-scaled Mixing Another way to handle disparity between the data available for each task is to use temperature-scaling. Formally, for temperature T, we take the T^{th}-root of the mixing rate for each task θ_t, and then renormalize i.e.:

$$\theta_t^{(\mathrm{TS})} = \sqrt[T]{\theta_t^{(\mathrm{PM})}} \div \sum_{t'} \sqrt[T]{\theta_{t'}^{(\mathrm{PM})}} \qquad (2)$$

When $T = 1$, temperature scaling reduces to proportional mixing, and as T is increased, the mixing rates approach a uniform distribution. We considered temperature scaling as a means to reduce the ability of tasks with large datasets to eclipse tasks with significantly fewer examples.

Adaptive Mixing In addition to data size, the task's difficulty can have a strong impact on whether the model underfits or overfits a dataset. Even with temperature-scaling, we observed that the model spent the majority of training steps on data-rich tasks and that the performance of the model on a task was not always proportional to the amount of data available for that task – some tasks were inherently harder for the model to adapt to. Consequently, we wanted to develop a mixing strategy that would decrease the time the model spent training on tasks it had already learned and more time on tasks it was still struggling with. Thus, to capture and account for task difficulty, we implemented an adaptive mixing strategy: after a certain number of warm-up epochs, the mixing rate is updated after each epoch proportionally to the average validation cross-entropy loss for each task and re-normalized. Formally:

$$\theta_t^{(\mathrm{AM})} = L(t)^\gamma \div \sum_{t'} L(t')^\gamma \tag{3}$$

where γ is a scaling constant akin to the focus parameter reported in Lin et al. (2020).

Self-adaptive Mixing While adaptive mixing can account for the difficulty of a task in terms of generalizability, it does not consider the degree to which the model has fit the training dataset – i.e., it does not account for bias in the fine-tuning data. Moreover, adaptive mixing requires the availability of validation data for each task used in fine-tuning, which may not always be available. For these reasons, we explored a second form of adaptive mixing in which the mixing rate is determined based on the training loss for each task. Unlike the validation loss setting above, using training loss is sensitive to epoch size – if the model has not explored a sufficient percent of the training data for a task, the loss for that task may not accurately reflect the model's mastery of the task. Consequently, we needed to balance the *exploration ratio*, e_t, of task t – i.e., the percent of all training data for a task that has been seen by the model during fine-tuning – with the training loss on that task. Formally:

$$\theta_t^{(\mathrm{SAM})} = \frac{(1 - e_t)\widehat{L} + (e_t)L(t)}{\sum_{t'} (1 - e_{t'})\widehat{L} + (e_t)L(t')} \tag{4}$$

where $\widehat{L}$ is the macro-average cross entropy training loss over all tasks. In this way, the model begins with a close-to-uniform mixing strategy and begins to favor tasks proportionally to the task's loss and exploration rate. As with adaptive mixing, we wait a certain number of warm-up epochs before computing the exploration rate or updating θ.

5 Experiments

In our experiments, we trained on the datasets described in Section 4.1 and evaluated on five tasks originating from three datasets previously unseen by the model. All models were trained with a batch size of 8, maximum sequence length of 512 tokens, 3 warm-up epochs followed by 10 training epochs, and 1,000 batches-per-epoch, using single V100X GPUs (32 GB VRAM) on a shared cluster. Training took between four-to-six hours, depending on cluster load. Additional implementation details are provided in Appendix A. To reduce variance between runs, we report results with greedy decoding (i.e., no beam search).

We measured the impact of (1) multi-task fine-tuning (MTFT), (2) different task mixing strategies, and (3) excluding various tasks from fine-tuning on zero-shot summary quality. We report traditional summarization and generation metrics such as BLEU (Papineni et al., 2002), ROUGE (Lin, 2004) and METEOR (Banerjee and Lavie, 2005). Because the reference summaries for many tasks are highly abstractive, we adopt the embedding-based metrics proposed in Sharma et al. (2017), i.e., GloVe (Pennington et al., 2014) cosine similarity using Embedding Averages (EACS), Vector Extrema (VECS, Forgues et al., 2014), and greedy matching (GMS, Rus and Lintean, 2012).

5.1 Evaluation Tasks

MEDIQA-AnS The MEDIQA-AnS collection contains consumer health questions, articles from reliable websites, passages extracted from those web pages, and single- and multi-document summaries of the passages intended to provide consumer-friendly answers for the questions (Savery et al., 2020). We used the 552 extractive single-document question-driven summaries.

DUC The Document Understanding Conference (DUC) was hosted by NIST from 2001-2007, to promote summarization research. In 2004, there were 50 questions each associated with very short single-document summaries (limited to 75 bytes), while in 2007, there were 45 questions, each associated with long 10-document summaries (between 230 and 250 words). Documents were from the AQUAINT English news corpus (Graff, 2002).

TAC The Text Analysis Conference (TAC) is the successor to DUC with ongoing public challenges on summarization. In this work, we considered the 2009 and 2010 tracks. Both tracks explored summarizing sets of 10 newswire articles into 100-word reference summaries. In 2009, the track had 44 topics, each associated with a natural language topic description and four reference summaries (Dang and Owczarzak, 2009).

In 2010, track explored 46 topics, each associated with a natural language topic description, four reference summaries, and, unlike 2009, one of five pre-defined categories. TAC 2010 summaries were expected to cover all aspects associated with that category (e.g., for *Health and Safety*, summaries

(a) **MEDIQA Summarization**

System	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	METEOR	EACS	VECS	GMS
BART	37.59	31.35	28.35	26.43	31.53	19.92	93.34	54.70	84.19
BART +MTFT	46.92	41.96	39.24	37.33	42.73	25.43	95.04	65.21	88.51
T5	14.82	12.68	11.53	10.71	31.18	14.01	93.83	58.85	83.35
T5 +MTFT	43.17	38.10	35.52	33.77	40.21	22.72	94.96	62.97	87.06

(b) **Document Understanding Conference (DUC) Summarization**

	DUC 2004					DUC 2007				
System	BLEU-1	ROUGE-L	EACS	VECS	GMS	BLEU-1	ROUGE-L	EACS	VECS	GMS
BART	10.34	14.18	67.88	41.59	70.81	30.30	13.96	96.22	38.93	82.06
BART +MTFT	8.74	12.08	66.74	39.32	69.49	29.95	14.25	96.17	39.94	82.19
T5	7.23	9.60	65.80	37.99	65.34	6.32	9.86	93.55	42.14	75.31
T5 +MTFT	7.84	10.68	65.33	38.26	68.18	26.85	3.37	95.94	38.23	81.22
Human	N\A	N\A	N\A	N\A	N\A	83.18	20.52	98.36	52.31	86.30

(c) **Text Analysis Conference (TAC) Summarization**

	TAC 2009					TAC 2010				
System	BLEU-1	ROUGE-L	EACS	VECS	GMS	BLEU-1	ROUGE-L	EACS	VECS	GMS
BART	2.71	4.76	44.90	21.59	64.11	20.32	10.79	88.16	36.89	75.63
BART +MTFT	29.73	16.22	95.18	42.67	79.99	27.67	14.59	95.06	44.92	78.96
T5	14.30	13.55	93.95	42.35	76.80	12.25	11.91	93.55	42.14	75.31
T5 +MTFT	29.45	16.33	95.30	42.46	79.73	27.49	15.06	95.05	45.28	79.04
Human	53.25	23.20	97.08	52.70	83.57	50.62	20.75	97.03	55.03	82.88

Table 1: Impact of multi-task fine-tuning (MTFT) on zero-shot summarization quality; "Human" refers to cross-evaluation of human-authored references summaries.

Mixing strategy	BLEU-4	ROUGE-L
Proportional	33.658	40.219
Temperature-scaled, $T = 2$	34.944	41.999
Temperature-scaled, $T = 4$	34.752	41.352
Adaptive, $\gamma = 1$	30.241	38.619
Adaptive, $\gamma = 2$	32.853	40.464
Adaptive, $\gamma = 4$	30.809	37.659
Self-Adaptive, $\gamma = 1$	33.226	40.339
Self-Adaptive, $\gamma = 2$	35.315	42.102
Self-Adaptive, $\gamma = 4$	**36.434**	**43.465**

Table 2: MEDIQA performance with different mixing strategies.

		DUC		TAC	
Model	MEDIQA	2004	2007	2009	2010
T5	0.43	0.24	1.70	0.39	0.61
BART	1.11	0.37	1.38	0.46	0.62

Table 3: Standard deviation of ROUGE-L over 10 runs.

should cover (a) what happened, (b) who was affected, (c) how they were affected, (d) why the health or safety issue occurred, and (e) any countermeasures or prevention efforts) (Owczarzak and Dang, 2010).

6 Results

Table 1 provides summarization results using T5 and BART with and without multi-task fine-tuning (MTFT) for zero-shot summarization. Clearly, MTFT had a strong impact on MEDIQA and TAC summary quality. DUC results, however, were more varied. Interestingly, we can observe that MTFT had a greater impact on BART than T5 summarization quality, despite structuring fine-tuning tasks with the same prompts and configuration as those used to train T5.

Table 2 illustrates the zero-shot Rouge-L achieved on each testing task when using various mixing strategies described in Section 4.3. Self-adaptive attention ($\gamma = 4$) obtains the highest performance, at the cost of implementation complexity; temperature-scaled mixing ($T = 2$) obtains reasonable performance as well.

Table 4 shows the impact of removing each task during fine-tuning on zero-shot summary quality. The most impactful tasks for MEDIQA are BioASQ (single- and multi-document), MedlinePlus, and IBM Evidence; BioASQ (multi-document only), MedlinePlus, ArXiv, and Cosmos QA were the most consistent for DUC; while

	MEDIQA		DUC 2004		DUC 2007		TAC 2009		TAC 2010	
Ablation	ROUGE	Δ	ROUGE	Δ	ROUGE	Δ	ROUGE	Δ	ROUGE	Δ
All Tasks	39.30		10.17		13.25		16.12		14.85	
– QA4MRE 2013 Alz.	40.16	+0.86	10.61	+0.45	13.41	+0.16	15.95	+1.11	15.17	+0.32
– QA4MRE 2012 Alz.	40.18	+0.03	10.54	−0.08	13.35	−0.06	15.81	−0.14	15.02	−0.15
– QA4MRE 2013 Main	40.21	+0.02	10.17	−0.37	13.58	+0.23	15.79	−0.03	14.96	−0.06
– QA4MRE 2012 Main	39.57	−0.64	10.27	+0.11	13.37	−0.21	16.07	+0.28	14.79	−0.17
– QA4MRE 2011 Main	38.97	−0.60	10.65	+0.37	13.35	−0.02	15.83	−0.24	15.09	+0.30
– MC-TACO	39.81	+0.84	10.76	+0.11	13.50	+0.15	15.79	−0.04	14.79	−0.30
– Cosmos QA	39.46	−0.35	10.18	−0.58	13.50	+0.00	15.86	+0.08	15.03	+0.24
– IBM Evidence	38.42	−1.04	10.46	+0.28	13.31	−0.20	16.12	+0.26	14.95	−0.08
– Movie Rationales	38.56	+0.14	10.50	+0.04	13.36	+0.06	15.81	−0.32	14.69	−0.26
– SQuAD	38.36	−0.20	10.43	−0.07	13.40	+0.03	16.58	+0.78	14.72	+0.03
– EBM Justifications	39.52	+1.16	10.37	−0.06	13.44	+0.05	16.14	−0.44	15.03	+0.30
– EBM Answers	41.46	+1.94	10.73	+0.36	13.26	−0.18	16.46	+0.33	14.94	−0.09
– CNN/DailyMail	40.64	−0.82	12.71	+1.99	13.53	+0.26	15.92	−0.54	14.85	−0.09
– Cochrane	42.17	+1.52	12.52	−0.19	13.62	+0.10	15.99	+0.07	14.75	−0.10
– PubMed	41.70	−0.47	12.37	−0.15	13.61	−0.01	15.44	−0.56	14.63	−0.12
– ArXiv	43.14	+1.44	10.97	−1.40	13.50	−0.10	15.59	+0.15	14.78	+0.15
– CoPA	44.01	+0.87	10.97	−0.01	13.74	+0.24	15.94	+0.35	15.10	+0.33
– MedlinePlus	42.98	−1.03	10.64	−0.33	13.60	−0.14	16.04	+0.10	14.89	−0.21
– PubMed PubSum	43.40	+0.42	10.69	+0.05	13.57	−0.03	16.06	+0.03	14.93	+0.04
– BioASQ (multi-doc)	42.31	−1.09	9.45	−1.23	13.47	−0.10	16.09	+0.03	14.34	−0.59
– BioASQ (single-doc)	37.64	−4.67	12.49	+3.04	12.67	−0.80	16.27	+0.18	14.97	+0.62
– CQaD-S	14.01	−23.63	9.60	−2.89	9.86	−2.89	13.55	−2.72	14.97	−3.06

Table 4: Multi-task fine tuning ablation on summarization quality; ROUGE refers to ROUGE-L.

PubMed, CNN/DailyMail, and Movie Rationales had the highest impact on TAC.

Finally, Table 3 reports the standard deviation of T5 and BART for all evaluation tasks; as in Raffel et al. (2019), we assume the standard deviation can be applied to all reported experiments.

7 Discussion

Table 1 indicates that multi-task fine-tuning (MTFT) provides improved zero-shot summarization quality on domains with clear knowledge transfer (e.g., news documents) as well as new domains with less-direct knowledge transfer such as consumer health (i.e., MEDIQA). We note that for highly abstractive summarization, e.g., DUC and TAC, surface-level metrics such as BLEU and ROUGE are poor summarization quality indicators. Embedding-based measures that are capable of capturing semantic similarity show a strong improvement when MTFT is used. DUC results are more perplexing, likely due to the extreme disparity between MTFT summarization tasks and the DUC evaluation: in 2004, DUC summaries were between 4 and 20 tokens long and highly abstractive (as indicated by human performance), making automatic measures less effective. For DUC 2007, all summaries were between 140 and 250 words long, much longer than most summaries seen during MTFT.

When analyzing the impact of different tasks on down-stream performance as indicated by Table 4, it is clear that each final summarization task benefits from different fine-tuning task combinations. While it may appear that CQaD-S had a strong impact on all tasks, additional experiments suggest that fine-tuning on any single summarization provides similar zero-shot improvements compared to using `T5-Base` or `BART-Large` and that CQaD-S and BioASQ had similar impacts on MEDIQA. Our results suggest that picking the optimal combination of fine-tuning tasks is non-trivial, and more work is needed to improve the robustness of training and task-mixing strategies and that in-depth analysis or principled guidelines for task selection would benefit the community. In a zero-shot setting, it is difficult to determine the optimal combination of fine-tuning tasks. However, in future work, we plan to explore feature selection techniques such as additive or recurrent feature elimination to determine an efficient way to select optimal tasks in a few-shot learning environment.

Table 2 suggests that for the case of zero-shot learning, self-adaptive training was most effective at exploiting fine-tuning tasks. However, taken with Table 4, it is clear that adaptive mixing can be further improved to be more resilient against sub-optimal fine-tuning task combinations. We note that temperature-scaling with $T = 2$ offers a strong competitor to self-adaptive task mixing with the additional advantage of a simpler implementation.

While an in-depth manual assessment of all tasks is beyond the scope of this work, a shallow manual review suggests that conditional summarization would benefit from new metrics that emphasize the role of the conditional context (i.e., question or topic description) in the summary to ensure that summaries are not too generic.

8 Conclusions

In this paper, we explored the impact of multi-task fine-tuning (MTFT) on zero-shot conditional summarization for consumer health questions (MEDIQA, Savery et al., 2020) as well as topic-driven news article summarization (i.e., the TAC and DUC summarization challenges). We introduced four new summarization datasets and proposed two online or adaptive methods for task mixing during fine-tuning. Our experimental results indicate that MTFT enables BART to produce higher quality summaries than T5, and that MTFT improved summary quality on unseen tasks in terms of ROUGE-L by 35.50 % (relative; 11.20 % absolute) for consumer health and 35 %–241 % (relative; 3.80 %–11.46 % absolute) for TAC. DUC results were inconclusive, with MTFT improving T5 results but hindering BART. Ablation analysis indicates that all tasks are not created equal and careful consideration must be taken to ensure each task has transferable characteristics (even subtle semantic properties such as argumentation properties) to the down-stream zero-shot application. Our proposed self-adaptive task mixing strategy was able to lessen the impact of irrelevant tasks on zero-shot performance by 8.25 % (relative; 2.75 % absolute) BLEU-4 and 7.57 % (relative; 3.04 % absolute) ROUGE-L. In future work, we plan to explore automatic approaches for determining the optimal set of fine-tuning tasks, improving the robustness of task mixing strategies to accommodate sub-optimal task combinations, and exploring new evaluation metrics that better reflect the role of the summarization context (i.e., question or topic description).

Reproducibility

Experiments used TensorFlow version 2.1, PyTorch version 1.4, and the T5 and BART implementations provided in HuggingFace's Transformers package, version 2.10 (Wolf et al., 2019). Evaluation metrics were computed using NLG Eval (Sharma et al., 2017), existing datasets were obtained using the TensorFlow DataSets catalogue, version 3.1. The source code for this paper is available at `https://github.com/h4ste/mtft_zsl`.

Acknowledgments

This work was supported by the intramural research program at the U.S. National Library of Medicine, National Institutes of Health, and utilized the computational resources of the NIH HPC Biowulf cluster (`http://hpc.nih.gov`).

References

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Asma Ben Abacha, Yassine Mrabet, Mark Sharp, Travis Goodwin, Sonya E. Shooshan, and Dina Demner-Fushman. 2019. Bridging the gap between consumers' medication questions and trusted answers. In *Proceedings of the 17th World Congress of Medical and Health Informatics (MEDINFO)*.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A discourse-aware attention model for abstractive summarization of long documents. 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 615–621, New Orleans, Louisiana. Association for Computational Linguistics.

Hoa Trang Dang and Karolina Owczarzak. 2009. Overview of the TAC 2009 summarization track. In *Proceedings of the Second Text Analysis Conference*, volume 2, Gaithersburg, Maryland, USA. National Institute of Standards and Technology.

Dina Demner-Fushman, Yassine Mrabet, and Asma Ben Abacha. 2019. 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, Minneapolis, Minnesota. Association for Computational Linguistics.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. 2020. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4443–4458, Online. Association for Computational Linguistics.

Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified language model pre-training for natural language understanding and generation. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 13063–13075. Curran Associates, Inc.

Xiangyu Duan, Mingming Yin, Min Zhang, Boxing Chen, and Weihua Luo. 2019. Zero-shot cross-lingual abstractive sentence summarization through teaching generation and attention. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3162–3172, Florence, Italy. Association for Computational Linguistics.

Gabriel Forgues, Joelle Pineau, Jean-Marie Larchevêque, and Réal Tremblay. 2014. Bootstrapping dialog systems with word embeddings. In *NIPS* Modern Machine Learning and Natural Language Processing Workshop*, volume 2.

Andrew Gordon and Reid Swanson. 2009. Identifying personal stories in millions of weblog entries. In *Third International Conference on Weblogs and Social Media, Data Challenge Workshop, San Jose, CA*, volume 46.

David Graff. 2002. *The AQUAINT corpus of English news text: Portions © 1998-2000 New York Times, Inc., © 1998-2000 Associated Press, Inc., © 1996-2000 Xinhua News Service*. Linguistic Data Consortium.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 1693–1701. Curran Associates, Inc.

Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Cosmos QA: Machine reading comprehension with contextual commonsense reasoning. 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 2391–2401, Hong Kong, China. Association for Computational Linguistics.

Chris Kedzie, Kathleen McKeown, and Hal Daumé III. 2018. Content selection in deep learning models of summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1818–1828, Brussels, Belgium. Association for Computational Linguistics.

Nitish Shirish Keskar, Bryan McCann, Lav R. Varshney, Caiming Xiong, and Richard Socher. 2019. CTRL: A conditional transformer language model for controllable generation.

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

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

T. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. 2020. Focal loss for dense object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(2):318–327.

Peter J. Liu, Yu-An Chung, and Jie Ren. 2019a. Summae: Zero-shot abstractive text summarization using length-agnostic auto-encoders.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019b. 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, Florence, Italy. Association for Computational Linguistics.

Zhengyuan Liu, Ke Shi, and Nancy F. Chen. 2020. Conditional neural generation using sub-aspect functions for extractive news summarization.

Inderjeet Mani. 2009. Summarization evaluation: an overview. In *Proceedings of the NTCIR Workshop*, volume 2.

Diego Molla and Maria Elena Santiago-Martinez. 2011. Development of a corpus for evidence based medicine summarisation. In *Proceedings of the Australasian Language Technology Association Workshop 2011*, pages 86–94, Canberra, Australia.

Timothy Niven and Hung-Yu Kao. 2019. Probing neural network comprehension of natural language arguments. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4658–4664, Florence, Italy. Association for Computational Linguistics.

Karolina Owczarzak and Hoa Trang Dang. 2010. Overview of the TAC 2010 summarization track. In *Proceedings of the Third Text Analysis Conference*, volume 3, Gaithersburg, Maryland, USA. National Institute of Standards and Technology.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Anselmo Peñas, Eduard Hovy, Pamela Forner, Álvaro Rodrigo, Richard Sutcliffe, and Roser Morante. 2013. QA4MRE 2011-2013: Overview of question answering for machine reading evaluation. In *Information Access Evaluation. Multilinguality, Multimodality, and Visualization*, pages 303–320, Berlin, Heidelberg. Springer Berlin Heidelberg.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Alec Radford, Karthik Narasimhan, Time Salimans, and Ilya Sutskever. 2018. Improving language understanding with unsupervised learning. *Technical report, OpenAI*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer.

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, Austin, Texas. Association for Computational Linguistics.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S Gordon. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *2011 AAAI Spring Symposium Series*.

Vasile Rus and Mihai Lintean. 2012. A comparison of greedy and optimal assessment of natural language student input using word-to-word similarity metrics. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, pages 157–162, Montréal, Canada. Association for Computational Linguistics.

Max Savery, Asma Ben Abacha, Soumya Gayen, and Demner-Fushman. 2020. Question-driven summarization of answers to consumer health questions. https://doi.org/10.17605/OSF.IO/FYG46.

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, Vancouver, Canada. Association for Computational Linguistics.

Shikhar Sharma, Layla El Asri, Hannes Schulz, and Jeremie Zumer. 2017. Relevance of unsupervised metrics in task-oriented dialogue for evaluating natural language generation.

Eyal Shnarch, Carlos Alzate, Lena Dankin, Martin Gleize, Yufang Hou, Leshem Choshen, Ranit Aharonov, and Noam Slonim. 2018. Will it blend? blending weak and strong labeled data in a neural network for argumentation mining. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 599–605, Melbourne, Australia. Association for Computational Linguistics.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy. Association for Computational Linguistics.

TensorFlow Datasets. TensorFlow Datasets, a collection of ready-to-use datasets. `https://www.tensorflow.org/datasets`.

George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, Yannis Almirantis, John Pavlopoulos, Nicolas Baskiotis, Patrick Gallinari, Thierry Artieres, Axel Ngonga, Norman Heino, Eric Gaussier, Liliana Barrio-Alvers, Michael Schroeder, Ion Androutsopoulos, and Georgios Paliouras. 2015. An overview of the bioasq large-scale biomedical semantic indexing and question answering competition. *BMC Bioinformatics*, 16:138.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio,

H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Alex Wang and Kyunghyun Cho. 2019. BERT has a mouth, and it must speak: BERT as a Markov random field language model. In *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*, pages 30–36, Minneapolis, Minnesota. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. XLNet: Generalized autoregressive pretraining for language understanding. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 5753–5763. Curran Associates, Inc.

Omar F Zaidan, Jason Eisner, and Christine Piatko. 2008. Machine learning with annotator rationales to reduce annotation cost. In *Proceedings of the NIPS* 2008 Workshop on Cost Sensitive Learning*, pages 260–267.

Ben Zhou, Daniel Khashabi, Qiang Ning, and Dan Roth. 2019. "Going on a vacation" takes longer than "going for a walk": A study of temporal commonsense understanding. 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 3363–3369, Hong Kong, China. Association for Computational Linguistics.