Domain Adaptation with Pre-trained Transformers for Query-Focused Abstractive Text Summarization

Md Tahmid Rahman Laskar
Dialpad Canada Inc.
Information Retrieval and Knowledge
Management Research Lab
York University
tahmedge@cse.yorku.ca

Enamul Hoque
School of Information Technology
York University
enamulh@yorku.ca

Jimmy Xiangji Huang
Information Retrieval and Knowledge
Management Research Lab
York University
jhuang@yorku.ca

The Query-Focused Text Summarization (QFTS) task aims at building systems that generate the summary of the text document(s) based on the given query. A key challenge in addressing this task is the lack of large labeled data for training the summarization model. In this article, we address this challenge by exploring a series of domain adaptation techniques. Given the recent success of pre-trained transformer models in a wide range of natural language processing tasks, we utilize such models to generate abstractive summaries for the QFTS task for both single-document and multi-document scenarios. For domain adaptation, we apply a variety of techniques using pre-trained transformer-based summarization models including transfer learning, weakly supervised learning, and distant supervision. Extensive experiments on six datasets show that our proposed approach is very effective in generating abstractive summaries for the QFTS task while setting a new state-of-the-art result in several datasets across a set of automatic and human evaluation metrics.

1. Introduction

With the rapid growth of textual documents on the Internet, accessing information from the Web has become a challenging problem (Yao, Wan, and Xiao 2017). In a Web search, users may require the summary about a certain topic from various sources to fulfill their information needs (Xu and Lapata 2020b). Since the performance of the Web search engines largely depends on a system that possesses good question answering (QA) capabilities, many researchers are focusing on developing systems that can provide users with a summarized response to their queries (Deng et al. 2019). The Query-Focused Text Summarization (QFTS) task deals with such problems, where a query along with the source document(s) are given and the objective is to generate a summary from the source document(s) based on the given query (Yao, Wan, and Xiao 2017) (see Table 1).

The query-focused summarization task can be categorized depending on the type of the source document(s) and the generated summary. For instance, based on the type of the source document(s), we can consider two scenarios: (i) **Single-Document Scenario:** where the goal is to generate a summary from a single source document, and (ii) **Multi-Document Scenario:** where the goal is to generate a summary from a set of documents (Baumel, Eyal, and Elhadad 2018). Moreover, based on the type of generated summaries, this task can be either extractive or abstractive (Baumel, Eyal, and Elhadad 2018; Nema et al. 2017; Feigenblat et al. 2017; Yao, Wan, and Xiao 2017; Xie et al. 2020; Pasunuru et al. 2021). For the **Extractive Summarization Scenario**, relevant text spans are directly extracted from the source document(s). In contrast, for the **Abstractive Summarization Scenario**, the generated summaries can contain words that may not appear in the source document(s). Given the rise of conversational QA assistants such as Siri, Cortana, Alexa, and Google Assistant, researchers are interested in studying how to incorporate abstractive summarization capabilities in such systems for natural response generation (Nishida et al. 2019).

Due to the growing interest in QA systems with summarization capabilities, a number of methods have been proposed for the Query-Focused Abstractive Summarization (QFAS) task. More recent methods for such tasks adopted various state-of-the-art neural summarization models (Yao, Wan, and Xiao 2017; Qiu et al. 2020) by following the encoder-decoder architecture. However, there are some key challenges that are required to be addressed while building QFAS systems for both single and multi-document scenarios. For the single-document QFAS task, one major challenge is that the available datasets are very small in size compared with the generic abstractive summarization datasets (Nema et al. 2017; Baumel, Eyal, and Elhadad 2018; See, Liu, and Manning 2017). Thus, during training, the model needs to tackle the few-shot

Table 1
An example of the Query-Focused Text Summarization task to generate the abstractive summary from the given source document.

Query: What is the benefit of reality shows?
Document: Even if reality shows were not enlightening, they generate massive revenues that can be used for funding more sophisticated programs. Take BBC for example, it offers entertaining reality shows such as total wipeout as well as brilliant documentaries.
Summary: Reality show generates revenues.

learning problem. For the multi-document scenario, again the existing benchmark datasets are very small (Baumel, Eyal, and Elhadad 2018). On top of that, each gold reference summary in the available datasets are written for a given document set without including the reference summary of each individual document in that document set. The problem is that we cannot simply concatenate all the documents in a given document set and feed into the state-of-the-art neural architecture for text summarization—the transformer model (Liu and Lapata 2019b; Zhang et al. 2019a; Lewis et al. 2019; Raffel et al. 2019)—as the input sequence may become prohibitively long. This is because the transformer architecture has quadratic computational time and memory complexities, and these complexities worsen with the increasing length of the input sequence due to the matrix multiplication in self-attention blocks (Kitaev, Kaiser, and Levskaya 2019; Beltagy, Peters, and Cohan 2020; Zaheer et al. 2020; Wang et al. 2020; Choromanski et al. 2020).

To address the above challenges, in this article, we study how to utilize domain adaptation from pre-trained neural models for the QFAS task. Note that domain adaptation or transfer learning from pre-trained models is particularly suitable when the target dataset does not contain any labeled training data or the size of the training dataset is very small (Ramponi and Plank 2020). To leverage domain adaptation, we adopt a pre-trained transformer model. While the transformer-based models have been successfully applied for a wide range of natural language processing tasks (Vaswani et al. 2017), this has not been deeply studied for the QFAS task. To our knowledge, our work is among the first studies that explores domain adaptation for query-focused abstractive text summarization based on pre-trained transformer models. With extensive experiments in benchmark QFAS datasets, we show that transfer learning from pre-trained transformer-based generic summarization models can be effectively utilized to tackle the few-shot learning issue in both single-document and multi-document scenarios along with overcoming the computational complexity-related issues in long text sequences. More concretely, our contributions presented in this article are listed below.

- To address the lack of large training datasets, we propose a domain adaptation technique that utilizes transfer learning via leveraging the available large generic text summarization datasets by first pre-training a transformer-based model on such datasets and then fine-tuning the pre-trained model for the QFAS task by incorporating query relevance.

- To address the computational complexity problem while training neural models on multiple documents at once (Liu and Lapata 2019b; Beltagy, Peters, and Cohan 2020; Choromanski et al. 2020; Kitaev, Kaiser, and Levskaya 2019; Zaheer et al. 2020), we again utilize transfer learning from pre-trained transformers using the following two novel techniques:

  – First, we propose a weakly supervised learning model that generates the weak reference summary of each document in a document set. We then fine-tune the pre-trained transformer-based summarization model iteratively on each document for generating the query-focused abstractive summary.

  – Second, instead of generating the weak reference summary for each individual document, we propose a sentence filtering approach that selects the sentences in the document set that are most relevant to the query and feed them to the pre-trained abstractive

summarization model. For this approach, we also propose a novel sequential fine-tuning technique that effectively utilizes all the gold reference summaries to provide supervised training.

- We conduct comprehensive experiments with extensive ablation studies and case studies to validate our design choices on six datasets: three datasets for single-document scenarios and three datasets for multi-document scenarios. Experimental results show that our proposed approaches set new state-of-the-art results in terms of several automatic and human evaluation metrics across benchmark datasets.

In addition to demonstrating the effectiveness of our approach, our experimental findings reveal several important new insights: (i) most queries do not have any relations with the input documents in the existing single-document QFAS dataset Debatepedia, (ii) the type of attention mechanisms in the encoder can influence the performance for the QFAS task, (iii) the domain adaptation from generic abstractive summarization models can be effective on other related tasks (i.e., abstractive answer generation in the MS-MARCO dataset), and finally (iv) some recent transformer architectures (e.g., Raffel et al. 2019; Zhang et al. 2019a) provide superior performance over their counterparts while being utilized within our proposed approach. As a secondary contribution, we make our source code publicly available here: `https://github.com/tahmedge/PreQFAS`, so that other researchers can reproduce our experimental results and also use our codebase to push the state of the art in the future.

We organize the remaining sections of this article as follows: in Section 2, we discuss the prior work on the abstractive text summarization task; we first briefly review this task for the generic abstractive text summarization scenario then review prior work where the query relevance was also taken into account. In Section 3, we describe our proposed approaches for the QFAS task for both single-document and multi-document scenarios. In Section 4, we present the datasets used in our experiments and the details of our experimental settings. The analyses of the experimental results are then presented in Section 5. Finally, we summarize our contributions with future directions in Section 6.

2. Related Work

In this section, we first briefly introduce readers to the generic abstractive summarization task. Then, we discuss the QFAS task in single-document scenarios, followed by discussing this task in multi-document scenarios.

2.1 Generic Abstractive Text Summarization

In recent years, the impressive success of neural models for sequence-to-sequence modeling in different natural language generation tasks (Young et al. 2017) has inspired researchers to utilize the neural encoder-decoder architecture for the abstractive summary generation problem (Rush, Chopra, and Weston 2015; Nallapati et al. 2016; Chopra, Auli, and Rush 2016). However, one major issue with the neural models for abstractive summarization is that, while generating the summaries, such models tend to repeat the same word multiple times; this leads to the generation of non-cohesive summaries (See, Liu, and Manning 2017). To address this issue, See, Liu, and Manning (2017) proposed the Pointer Generation Network (PGN), which utilized a novel copy and coverage mechanism to discourage the repetition of the same words. More recently, the

BERTSUM (Liu and Lapata 2019b) model was proposed, which used the BERT model (Devlin et al. 2019) as the encoder and the decoder of the transformer model (Vaswani et al. 2017) as the decoder. The BERTSUM model, utilizing fine-tuning of pre-trained transformer encoders (Devlin et al. 2019; Liu et al. 2019a, 2019b; Lan et al. 2019; Clark et al. 2020; Fu et al. 2021), showed impressive performance for the abstractive summarization task and set new state-of-the-art results in several datasets by outperforming previous neural models that leveraged the recurrent neural network architecture (Sutskever, Vinyals, and Le 2014). The successful utilization of the transformer architecture (Liu and Lapata 2019b) for abstractive summarization has also led to the development of more new state-of-the-art neural models that utilized this architecture for such tasks (Zhang et al. 2019a; Dong et al. 2019; Lewis et al. 2019; Raffel et al. 2019; Kitaev, Kaiser, and Levskaya 2019; Song et al. 2019; Beltagy, Peters, and Cohan 2020; Zaheer et al. 2020; Qi et al. 2020; Fabbri et al. 2021). These findings have motivated us to adopt the transformer architecture in our query-focused summarization models.

2.2 Single-Document Query-Focused Abstractive Text Summarization

While significant research has utilized neural models for the generic abstractive summarization task, applying the neural network architecture for such tasks when the query relevance is also taken into account has been rare (Baumel, Eyal, and Elhadad 2018). One notable exception on utilizing neural models for such tasks is the Diversity-Driven Attention (DDA) model (Nema et al. 2017). This model generates query-focused abstractive summaries by focusing on different portions of a document based on the given query at different times. However, a key challenge in addressing the Single-Document Query-Focused Abstractive Summarization (SD-QFAS) task using neural models is that the number of datasets available for this task is quite small (Baumel, Eyal, and Elhadad 2018; Nema et al. 2017; Abdullah and Chali 2020). To the best of our knowledge, the only available dataset for this task is the Debatepedia dataset,[1] but the size of this dataset is very small compared with the datasets used for generic abstractive summarization (Baumel, Eyal, and Elhadad 2018; Liu and Lapata 2019b; See, Liu, and Manning 2017). Thus, the lack of large training data for the SD-QFAS task in the available dataset makes this task a few-shot learning problem. To address this issue, the Relevance Sensitive Attention (RSA) for Query-Focused Summarization (Baumel, Eyal, and Elhadad 2018) utilized transfer learning by first pre-training the PGN model (See, Liu, and Manning 2017) on a large generic abstractive summarization dataset and then utilized the pre-trained model for the QFAS task to generate the summaries in the Debatepedia dataset. They found utilizing transfer learning to be quite effective for the SD-QFAS task in that dataset. More recently, newer models based on the recurrent neural network architecture (Sutskever, Vinyals, and Le 2014) that did not utilize transfer learning failed to outperform the RSA model in terms of different ROUGE scores (Aryal and Chali 2020; Ishigaki et al. 2020). This may indicate that the utilization of transfer learning to tackle the few-shot learning problem has a strong effect on performance improvement in the Debatepedia dataset. However, one major limitation of the RSA model is that this model provided a poor Precision score by generating summaries much longer than the gold summaries (Baumel, Eyal, and Elhadad 2018). Also, the authors did not fine-tune the pre-trained RSA model on the target dataset. In contrast, we investigate the effectiveness of fine-tuning the transformer architecture for the

---

1 `http://www.debatepedia.org/`.

SD-QFAS task, motivated by the findings that fine-tuning pre-trained transformer models improve performance in a wide range of tasks including text summarization (Devlin et al. 2019; Liu and Lapata 2019b; Qiu et al. 2020).

2.3 Multi-document Query-Focused Abstractive Text Summarization

The topic of query-focused abstractive summarization has remained underexplored for the multi-document scenario as well (Kulkarni et al. 2020). More importantly, the currently available query focused multi-document abstractive summarization (MD-QFAS) datasets (e.g., DUC[2] 2005, 2006, 2007) do not contain any labeled training data, that is, these datasets only provide test data (Baumel, Eyal, and Elhadad 2018; Goodwin, Savery, and Demner-Fushman 2020; Su et al. 2020; Xu and Lapata 2021). To tackle the lack of training data for the MD-QFAS task, most previous work was based on various unsupervised approaches that could only generate extractive summaries (Wang et al. 2008; Haghighi and Vanderwende 2009; Wan and Xiao 2009; Yao, Wan, and Xiao 2015; Zhong et al. 2015; Wan and Zhang 2014; Ma, Deng, and Yang 2016; Feigenblat et al. 2017). To generate the abstractive summaries in such tasks, Baumel, Eyal, and Elhadad (2018) proposed a transfer learning technique that addressed the issue of no dedicated training data for the datasets available for such tasks. They adopted the PGN (See, Liu, and Manning 2017) pre-trained for the generic abstractive summarization task in a large dataset to predict the query-focused summaries in the target dataset by modifying the attention mechanism of the PGN model. However, their model failed to outperform the extractive approaches in terms of various ROUGE scores.

Here, utilizing the state-of-the-art neural summarization models (Liu and Lapata 2019b; Lewis et al. 2019; Raffel et al. 2019; Zhang et al. 2019a) that leveraged supervised training is not applicable in these datasets due to the unavailability of the training data. Although some recent studies utilized datasets similar to the target dataset as the training set to provide supervised training (Li and Zhuge 2019), some other studies used similar datasets as the development dataset for hyperparameter optimization (Xu and Lapata 2020a,b; Su, Yu, and Fung 2021). However, while using datasets similar to the target dataset as the training data (e.g., using two DUC datasets for training the other DUC dataset), we find that these datasets only contain multi-document gold reference summaries. Thus, the state-of-the-art neural summarization models cannot be trained on such datasets since these models cannot consider long text sequences (i.e., multiple documents) as input at once due to the computational complexities (Zaheer et al. 2020; Beltagy, Peters, and Cohan 2020). For this reason, we utilize distant supervision from pre-trained transformers to generate the weak reference summary of each document in a document set so that the computational complexities in the MD-QFAS task can be avoided by iteratively training our model on each individual document.

Another key challenge in the MD-QFAS task is that the model needs to identify sentences from multiple documents that are relevant to the query (Wang et al. 2018). There could be several irrelevant sentences in different documents that are semantically similar to the relevant ones as well as to the query (Baumel, Eyal, and Elhadad 2018; Feigenblat et al. 2017), however, making the task of finding relevant sentences more challenging. To identify the sentences that are relevant to the query, various approaches such as similar word count (Baumel, Eyal, and Elhadad 2018) or the Cross-Entropy Method (Feigenblat et al. 2017) were utilized. Though neural models based on

---

2 https://www-nlpir.nist.gov/projects/duc/data.html.

supervised training have significantly outperformed various non-neural models for the answer sentence selection task (Garg, Vu, and Moschitti 2019; Lai et al. 2019), because of the absence of labeled data for the relevant sentences in the MD-QFAS datasets, neural models have not been effectively utilized yet. Recently, Garg, Vu, and Moschitti (2019) showed that neural models such as BERT or RoBERTa pre-trained on a large question answering dataset could effectively select answers in other similar datasets without any supervised training. More recently, such pre-trained answer sentence selection models were used by Xu and Lapata (2020b) for the MD-QFAS task. In their work, they utilized distant supervision from various question answering datasets using the fine-tuned BERT (Devlin et al. 2019) model to filter out the irrelevant sentences from the documents. However, Baumel, Eyal, and Elhadad (2018) found that filtering sentences from the input document(s) as an early step to train recurrent neural network models for query-focused abstractive summarization could lead to performance deterioration. Thus, we also investigate how to effectively utilize sentence filtering with the pre-trained transformer models for the MD-QFAS task.

3. Our Proposed Approach

Let us assume that we have a query $Q = q_1, q_2, \ldots, q_k$ containing k words. For the QFAS task in single-document scenarios, a source document $D_S = d_1, d_2, \ldots d_n$ containing n words is given where the objective is to utilize the given query Q to generate an abstractive summary $S = s_1, s_2, \ldots s_m$ containing m words from D_S. For the multi-document scenario, a set of N documents $D_M = D_{S_1}, D_{S_2}, \ldots, D_{S_N}$ are given where the goal is to generate the summary $S = s_1, s_2, \ldots s_m$ containing m words from the document set D_M based on the given query Q.

Recall that in this article, we aim to develop a QFAS system that can leverage the effectiveness of the transformer model (Vaswani et al. 2017; Liu and Lapata 2019b) to generate high-quality summaries. To achieve this goal, we need to address issues such as the lack of large training datasets for the QFAS task in both single and multi-document scenarios (Nema et al. 2017; Baumel, Eyal, and Elhadad 2018; Feigenblat et al. 2017), as well as the computational complexity–related problems that occur while training transformer models in long text sequences (Kitaev, Kaiser, and Levskaya 2019; Beltagy, Peters, and Cohan 2020; Zaheer et al. 2020; Choromanski et al. 2020). In our proposed method, we utilize transfer learning from generic abstractive summarization models to address these issues. We choose such models for transfer learning because the available generic text summarization datasets are much larger in size compared with the QFAS datasets (Baumel, Eyal, and Elhadad 2018; Nema et al. 2017). Thus, we hypothesize that once the transformer-based models are pre-trained on large generic summarization datasets, utilizing domain adaptation from such pre-trained models on QFAS datasets would be beneficial for few-shot learning. Later on, we again utilize transfer learning from pre-trained transformers to handle the computational complexities in long sequences. Below, we describe our proposed model, denoted as **PreQFAS**, that utilizes **Pre**-Trained Transformers for the **Q**uery-**F**ocused **A**bstractive Text **S**ummarization task, in detail.

3.1 The PreQFAS Model for the SD-QFAS Task

For our proposed PreQFAS model, we first adopt a transformer-based (Vaswani et al. 2017) model that has been pre-trained on a large generic abstractive text summarization dataset. For that purpose, we adopt the pre-trained BERTSUM (Liu and Lapata

2019b) model as our base model. We choose BERTSUM for three main reasons: (i) this model achieves impressive performance for abstractive summary generation (Liu and Lapata 2019b), (ii) the transformer architecture used by this model is conceptually much simpler than other recently proposed transformer-based summarization models (e.g., requires fewer number of parameters) (Lewis et al. 2019; Raffel et al. 2019; Zhang et al. 2019a), and (iii) this model also does not require the tuning of too many hyperparameters to achieve the optimized performance (Liu and Lapata 2019b).

Note that the BERTSUM model follows an encoder-decoder architecture that uses the BERT model as its encoder and the decoder of Transformer as its decoder. Because the BERTSUM model was designed for the generic text summarization task without considering any query relevance (Liu and Lapata 2019b), we incorporate the query relevance by concatenating the query with the input document and feed into the pre-trained BERTSUM. Then, we fine-tune the pre-trained BERTSUM model to generate the summaries in the target QFAS dataset. More specifically, our proposed PreQFAS model performs the QFAS task in the following two steps (see Figure 1). In the first step, we pre-train the BERTSUM model on a large training corpus of generic abstractive summarization. Then, we fine-tune the pre-trained model for the QFAS task by incorporating the query relevance. Below, we describe these two steps in detail.

(i) Pre-training the BERTSUM Model: In this step, we pre-train the BERTSUM model on a large generic abstractive summarization dataset. During this pre-training stage, the model utilizes the pre-trained BERT model (Devlin et al. 2019) as the encoder and the randomly initialized transformer decoder (Vaswani et al. 2017) as the decoder. Note that this model is first trained for extractive summarization and then it is re-trained for abstractive summarization. However, unlike the original BERT model, which inserts the special token [CLS] at the beginning of only the first sentence, the BERTSUM model (Liu and Lapata 2019b) inserts the [CLS] token at the beginning of each sentence. BERTSUM does so to calculate the probability of each sentence to identify the most relevant sentences. Moreover, each sentence-pair in BERTSUM is separated by the [SEP] token.
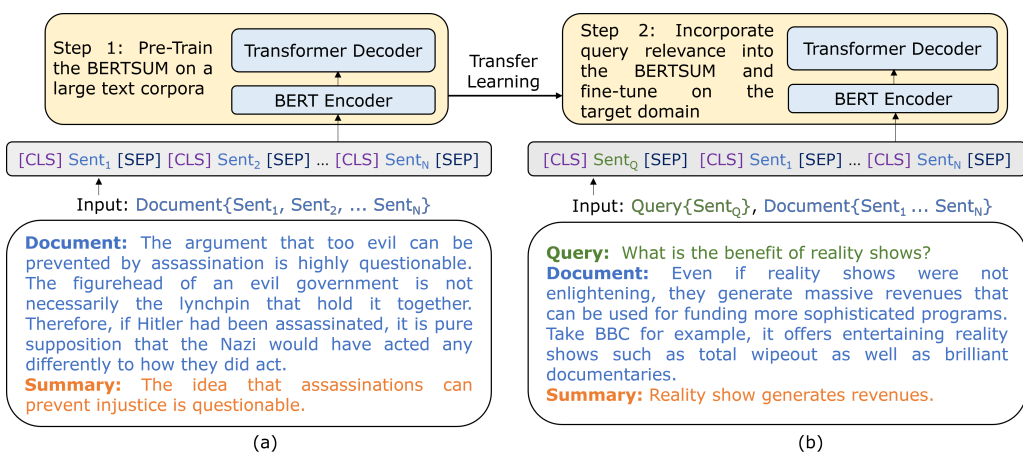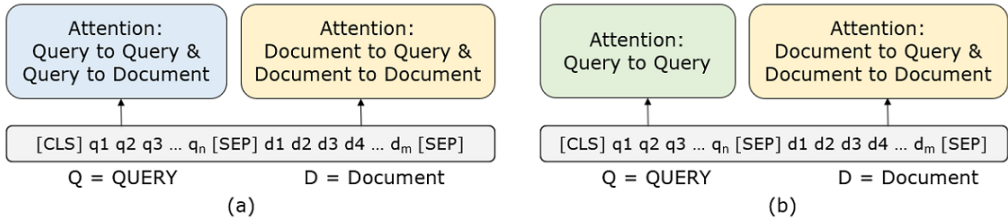

Figure 1
Our proposed PreQFAS model works in two steps: (a) Pre-train the BERTSUM model on a generic abstractive summarization corpus and (b) Fine-tune the pre-trained model for the QFAS task on the target domain.

Figure 2
An overview of various attention models. (a) The Bidirectional Self-Attention Mechanism.
(b) The Query-Document Attention Mechanism.

(ii) Incorporating Query Relevance and Fine-tuning BERTSUM: In this step, we fine-tune the BERTSUM model on the target QFAS dataset that was pre-trained on a generic abstractive summarization dataset in the previous step. During fine-tuning, we incorporate the query relevance via concatenating the query with the document as the input to the encoder (see Figure 1b). We do this because we find that a similar approach of concatenating the question with the document works well with neural models for different question-answering tasks (Lewis et al. 2019). In this way, we fine-tune a pre-trained generic abstractive text summarization model for query-focused abstractive summary generation to tackle the few-shot learning problem.

Attention Mechanisms: To utilize the query relevance in the pre-trained BERTSUM model for summary generation, we use two types of attention mechanisms (as shown in Figure 2). They are: (i) the bidirectional self-attention mechanism, and (ii) the query-document attention mechanism. Below, we describe these two attention mechanisms.

(i) The Bidirectional Self-Attention Mechanism: In the original BERTSUM architecture, the bidirectional self-attention mechanism (Devlin et al. 2019) is utilized by the BERT encoder to generate the encoded representation of the input text. In the bidirectional self-attention mechanism, when a pair of sentences are combined together and given as input to the BERT encoder, both sentences will give attention to each other. Thus, when we utilize the bidirectional self-attention mechanism (see Figure 2a) in the PreQFAS model, both the query and the document will not only give attention to themselves, but also they will give attention to each other to provide the encoded representation of the concatenated input.

(ii) The Query-Document Attention Mechanism: Dong et al. (2019) proposed the sequence-to-sequence language modeling objective for text sequences consisting of two segments. In such text sequences, each token in the first segment can only attend to the tokens in both directions within the same segment but cannot attend to any tokens in the second segment, while the tokens in the second segment can attend to the leftward tokens in their own segment as well as to all tokens in the first segment. Following this approach, we propose the Query-Document (QD) attention mechanism, where each token in the query can only attend to the tokens that are within the query while the tokens in the document can attend to all tokens in both the query and the document bidirectionally. The intuition here is that in the original PreQFAS model, the bidirectional self-attention allows the query to also attend to the document and thus the query segment might be influenced by the document segment. As a consequence, the final encoded representation of the concatenated input may lose some query-related information and the decoder may produce summaries that may not be fully relevant to the query. To avoid such scenarios, we allow the query segment to only attend to itself

whereas the document segment is allowed to provide a query-focused representation by attending to both the query and to itself. Given the query, key, and value vectors **Q**, **K**, and **V**, respectively, with $\mathbf{d}_k$ as the square root of the dimension of **K**, we calculate the encoded representation **Z** using QD attention by adding the mask matrix M in the self-attention formula of the transformer encoder (Vaswani et al. 2017):

$$Z = \textit{softmax} \left(\frac{Q \times K^{T}}{\sqrt{d_k}} + M \right) V \tag{1}$$

In equation (1), $M_{ij} = 0$ allows attention from token i to token j, whereas $M_{ij} = -\infty$ prevents attention from token i to token j.

3.2 Extending PreQFAS for Long Sequences in the MD-QFAS Task

In this section, we discuss how we utilize our proposed PreQFAS model to address the computational complexity issue that occurs while training transformer models in long text sequences (e.g., multiple documents[3]). Because the available MD-QFAS datasets only contain the gold reference summaries written for the whole document-set by human experts without containing the gold reference summary of each individual document (Baumel, Eyal, and Elhadad 2018; Ma, Deng, and Yang 2016; Feigenblat et al. 2017), neural models are ideally required to be trained on all documents in a multi-document set at once to leverage supervised training. Nonetheless, forcing neural models to be trained on all documents at once will result in computational complexity–related problems (Wang et al. 2020; Choromanski et al. 2020; Kitaev, Kaiser, and Levskaya 2019; Zaheer et al. 2020; Tay et al. 2020).

To address these issues, we propose two approaches that leverage pre-trained transformer-based models. In one approach, we propose a weakly supervised learning technique that first generates the weak reference summary of each individual document in a document set. Then, we fine-tune the pre-trained transformer-based summarization model on each individual document using the weak reference summaries. In this way, we generate the summary of each individual document and then select the most relevant sentences as the final summary using a transformer-based answer selection model (Laskar, Huang, and Hoque 2020; Laskar, Hoque, and Huang 2020b). In another approach, instead of training our model on each document, we again utilize a transformer-based answer selection model and construct a filtered input document via selecting the sentences (up to n tokens) in the document set that are most relevant to the query. Afterward, we fine-tune the summarization model on the filtered input document. Note that we study the sentence filtering technique by applying it differently in these two approaches: For the weakly supervised learning approach, we apply it at the final stage to select the relevant sentences from the generated summary; whereas for the other approach, we apply it at the beginning to select the relevant sentences from the multi-document set. In the following, we describe these two approaches in detail.

---

3 We address the computational complexity issue in multi-document scenarios only because we could not find any available single-document query-focused summarization datasets that have long text sequences and lead to computational complexities.
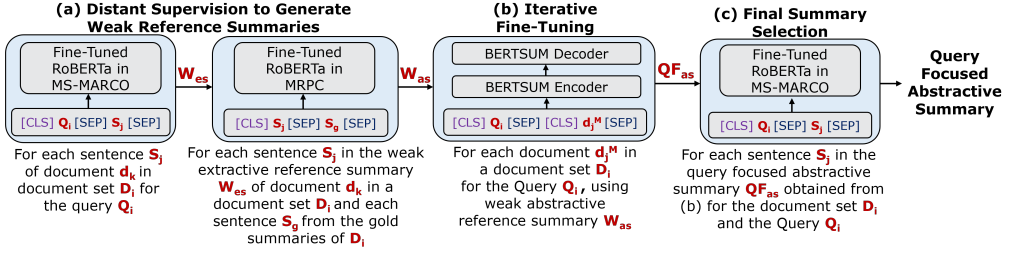
Figure 3
An overview of our proposed PreQFAS model for long text sequences (i.e., multi-document scenarios) that uses the fine-tuned **RoBERTa$_{MS-MARCO}$** model to (a) generate the initial weak extractive reference summary of each document followed by utilizing the **RoBERTa$_{MRPC}$** model for distant supervision to generate the weak abstractive reference summary. Then, (b) the pre-trained **BERTSUM** model is fine-tuned to iteratively generate the query focused abstractive summary of each document. Finally, all the generated query focused abstractive summaries are (c) ranked by the **RoBERTa$_{MS-MARCO}$** model to select the final summary.

3.2.1 Approach 1: Weakly Supervised Learning with Distant Supervision. Figure 3 shows an overview of our proposed approach that leverages weakly supervised learning. At first we generate the weak reference summary of each document in a document set by leveraging distant supervision from the multi-document gold reference summaries. Then, we propose an iterative approach that generates the query-focused abstractive summary of each document by fine-tuning a pre-trained single-document generic abstractive summarization model. Finally, we select the sentences (up to n tokens) that are most relevant to the query from the generated query-focused summary of the multi-document set by utilizing a pre-trained answer selection model. Note that contrary to the prior work where sentence filtering was applied as an early step to filter the input document (Baumel, Eyal, and Elhadad 2018; Xu and Lapata 2020b), in this approach we apply sentence filtering during the final step to filter the generated summary. In the following, we describe our proposed weakly supervised learning approach that tackles the computational complexity issue in the MD-QFAS task. First, we discuss how we utilize distant supervision to generate the weak reference summary of each individual document in a document set. Then, we discuss our proposed iterative approach that generates the query-focused abstractive summary of each document in a document set. Finally, we describe how we select the most relevant sentences from the generated query-focused summary as the final summary.

(a) Weak Reference Summary Generation. We generate the weakly supervised reference summary of each document in a document set in two steps (see Figure 3a). In the first step, we utilize a pre-trained model to generate the initial weak reference summary of each document. In the second step, we replace each sentence in the generated weak reference summary by each sentence in the multi-document gold reference summaries by utilizing the RoBERTa model (Liu et al. 2019b) fine-tuned for sentence similarity modeling. For that purpose, we measure the similarity between each sentence in the multi-document gold reference summaries with each sentence in the generated weak reference summary. Then, based on the similarity score, we select the most relevant sentences from the gold reference summaries as the final weak reference summary for each document. We generate the initial weak extractive reference summaries instead of directly generating the weak abstractive reference summaries since this additional step

allows us to only compare the similarity between each sentence in the multi-document gold reference summaries with each sentence in the initial weak extractive reference summary. Thus, it helps our model to be more efficient during the weakly supervised reference summary generation stage by avoiding the comparison between each sentence in the multi-document set with each sentence in the gold reference summaries (see Appendix E for details). Below, we describe these two steps in detail:

- **Initial Weak Reference Summary Generator:** To generate the initial weak reference summary of each document in a document set, we utilize a pre-trained transformer encoder model to generate the extractive summary of each document. To achieve our goal, we first adopt the pre-trained RoBERTa model (Liu et al. 2019b) and fine-tune it on the QA-ALL dataset of MS-MARCO (Wang et al. 2018) for the passage ranking (i.e., answer sentence selection) task. We choose RoBERTa in this regard because of its impressive performance on similar tasks in different answer selection datasets (Laskar, Huang, and Hoque 2020). Afterward, we utilize the fine-tuned RoBERTa model to measure the similarity score C between the given query Q_i and each sentence S_j in each document d_k. Based on the similarity score, we select the top 3 most relevant sentences as the weak extractive reference summary, because extracting only 3 sentences was found effective in different extractive summarizers such as the LEAD-3 baseline as well as the BERTSUM$_{\text{EXT}}$ model (Liu and Lapata 2019b).

- **Final Weak Reference Summary Generator:** The weak reference summaries generated in the previous step are extractive, while our goal is to generate abstractive summaries. Thus, we further provide distant supervision to manipulate the weak extractive reference summary generated in the previous step by replacing each sentence in the weak extractive reference summary with the most similar sentence found in the multi-document gold reference summaries written by humans. For this purpose, at first we adopt the RoBERTa model fine-tuned for the sentence similarity modeling task in the MRPC dataset (Liu et al. 2019b). Then, for each document d_k in a document set D_i, we utilize the fine-tuned RoBERTa$_{\text{MRPC}}$ model to measure the similarity between each sentence S_j in the weak extractive reference summary and each sentence S_g in the gold reference summaries. Based on the similarity score, each sentence in the weak extractive reference summary of a document is replaced with the most relevant sentence found in the multi-document abstractive gold reference summaries. Note that for a document d_k when a sentence S_g from the gold reference summaries is already used to replace a sentence S_j in the weak extractive reference summary, then for the same document d_k we do not consider the sentence S_g again for replacement. Instead, we use the next most relevant sentence from the multi-document gold reference summaries for replacement. The resulting summaries generated in this step can be considered as weak abstractive reference summaries because they are constructed from the gold reference summaries written by human annotators. In the following, we discuss how we train our model using these weak abstractive reference summaries.

(b) Iterative Fine-Tuning. In the MD-QFAS task, because the available datasets are also small in size (Feigenblat et al. 2017; Xu and Lapata 2020b; Roitman et al. 2020), we again utilize the PreQFAS model proposed in Section 3.1 to address the few-shot learning problem. However, the PreQFAS model is based on the BERTSUM model, pre-trained for the single-document generic summarization task by considering at most 512 tokens (Liu and Lapata 2019b). In reality, the total number of tokens in a document set in multi-document scenarios could be much larger than 512 tokens (Baumel, Eyal, and Elhadad 2018; Feigenblat et al. 2017). Thus, to avoid the computational complexities of training transformer-based models in such long sequences at once (Kitaev, Kaiser, and Levskaya 2019; Beltagy, Peters, and Cohan 2020; Zaheer et al. 2020; Choromanski et al. 2020), we take an iterative approach where we fine-tune the pre-trained summarization model on each individual document in a multi-document set (see Figure 3b). In this approach, similar to the PreQFAS model proposed in Section 3.1 for the SD-QFAS task, we first adopt the pre-trained BERTSUM model. Then, we incorporate the query relevance into the pre-trained BERTSUM and fine-tune it using the weak abstractive reference summary to generate the query-focused abstractive summary of each document in the given document set. Finally, we select the top N most relevant sentences from the generated summaries as the final summary. We describe the final summary selection procedure in detail next.

(c) Summary Sentence Selection. In this stage, for each document set, all the sentences in the query focused abstractive summaries generated in the previous step are ranked using a fine-tuned RoBERTa model. For this purpose, we adopt the RoBERTa model fine-tuned for the answer selection task in the MS-MARCO dataset, which we also utilized for initial weak reference summary generation. The fine-tuned RoBERTa$_{\text{MS-MARCO}}$ model is then utilized to measure the relevance between each sentence S_i in the generated summary and the query Q_j for the document set D_j to select the sentences that are most relevant to the query as the final summary. In this way, we utilize sentence filtering in the final step such that the total length of the selected sentences in the final summary does not exceed n tokens (see Figure 3c). To reduce redundancy in the final summary, we use Trigram Blocking (Paulus, Xiong, and Socher 2018).

3.2.2 Approach 2: Sequential Fine-Tuning with Sentence Filtering. In our weakly supervised learning approach demonstrated earlier, we apply sentence filtering in the final stage to identify the most relevant sentences in the generated summary. While most summarization models have attempted sentence filtering in the early stage, where the irrelevant sentences or paragraphs were filtered out from the source document(s) prior to generating the abstractive summaries, the results obtained by these models were conflicting (Liu and Lapata 2019a; Baumel, Eyal, and Elhadad 2018). For instance, Liu and Lapata (2019a) found that sentence filtering as an early step did not deteriorate the performance in the generic abstractive summarization task, whereas Baumel, Eyal, and Elhadad (2018) found that such a step deteriorated the performance in query-based multi-document abstractive summarization. To investigate the performance of sentence filtering as an early step, we develop the following approach. First, we select the sentences from the multi-document set that are most relevant to the query to construct a filtered input document. Then, we give the filtered input document as input to a PreQFAS model for fine-tuning.

More specifically, we first adopt a transformer-based answer selection model to identify the sentences in a document set that are most relevant to the query. Then, based on the relevance score, we rank the sentences in the document set. Next, we keep
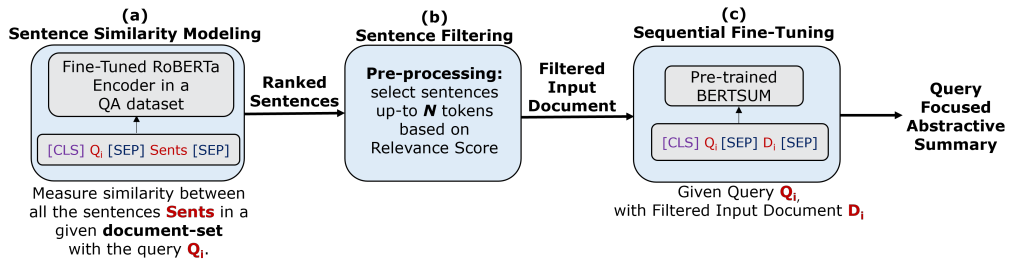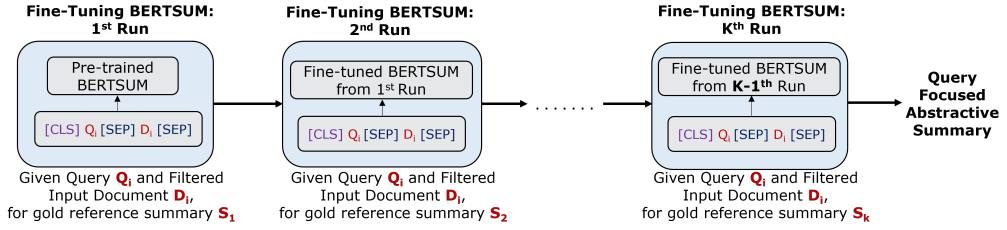
Figure 4
The proposed PreQFAS model for long sequences (i.e., multi-document scenarios) based on Sentence Filtering and Sequential Fine-Tuning: (a) First, all sentences in a document set are ranked by measuring their similarity score with the query. (b) Then, these ranked sentences are combined together to create a filtered input (up-to n tokens). (c) Finally, the query relevance is incorporated into the filtered input document and then given as input to the pre-trained BERTSUM model for sequential fine-tuning.

selecting the sentences until the total length of the selected sentences along with the query does not exceed n tokens. In this way, we create a filtered input document. Then, we utilize our proposed PreQFAS architecture to combine the query and the filtered input document together in order to give them as input to the BERTSUM model pre-trained for generic abstractive summarization. We then propose a sequential technique to fine-tune the BERTSUM model to provide supervised training via leveraging all gold reference summaries written by different human annotators for a given document set to generate the query focused abstractive summary. The overall approach is shown in Figure 4. Below, we describe our input document filtering process followed by the summary generation process in detail.

(a) Sentence Filtering. In this step, for each document set we measure the relevance of all sentences to the given query Q_i. For that purpose, we adopt the RoBERTa model and fine-tune it for the answer ranking task in the QA-ALL dataset of MS-MARCO (Liu et al. 2019b; Wang et al. 2018; Laskar, Huang, and Hoque 2020). Based on the relevance score, we then rank all the sentences in a document set. Afterward, we concatenate the query and the ranked sentences and consider the first n tokens as our input document for the summarization model. Note that this sentence filtering approach not only allows us to leverage the state-of-the-art neural summarization models to provide supervised training for the MD-QFAS task, but also allows us to overcome the computational complexity issue that occurs while training neural models on long documents.

(b) Sequential Fine-Tuning. In the previous step, we select the most relevant sentences to the query Q_i to construct the input document containing n tokens. In this way, the multiple documents are converted into a single document that consists of only those sentences that are most relevant to the query. Therefore, the filtered document allows us to leverage the effectiveness of fine-tuning pre-trained single-document generic abstractive summarization models. Thus, we adopt the BERTSUM model that was pre-trained for single-document abstractive summarization and fine-tune it to generate the query-focused abstractive summary for the given input document (i.e., the filtered input document constructed from a given document set).
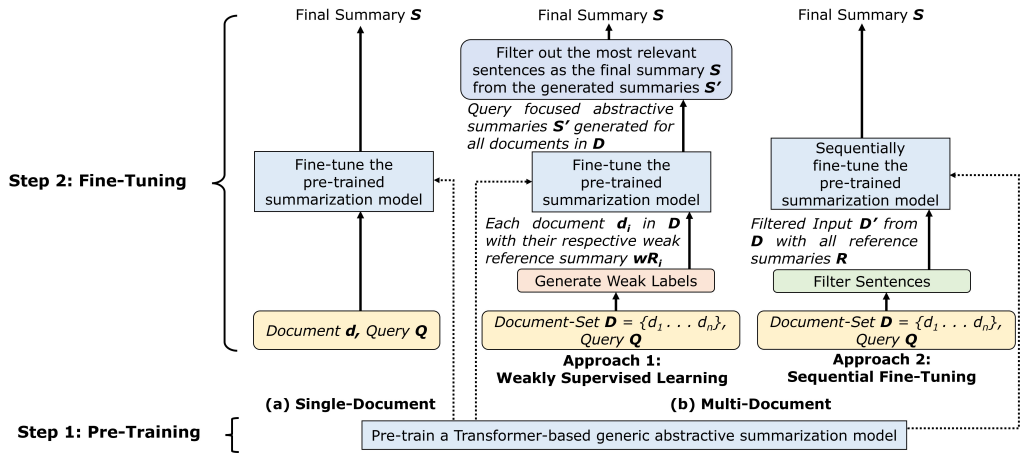
Figure 5
Sequential fine-tuning of the BERTSUM model using *K* different gold summaries. For *K* gold summaries, the model will be fine-tuned *K* times (i.e., *K* fine-tuning runs).

Because the available MD-QFAS datasets contain multiple gold reference summaries written by different human experts for the same query (Feigenblat et al. 2017), training neural models using multiple gold reference summaries will allow an encoder-decoder model to enhance its vocabulary in the decoder (Rush, Chopra, and Weston 2015; Nallapati et al. 2016). In order to leverage the advantage of multiple gold summaries, we propose a sequential fine-tuning model. In our proposed approach, if there are *K* gold summaries for a given training document set, then we fine-tune the model *K* times where each fine-tuning run[4] will have gold summaries different than the other runs for the same filtered input document. Thus, in Figure 4(c), the BERTSUM model will be fine-tuned *K* times. Note that for the first fine-tuning run, we adopt the model for fine-tuning that is pre-trained on a generic abstractive summarization task. For the subsequent runs, we fine-tune the model that is fine-tuned in the immediate previous run. We show the sequential fine-tuning process in Figure 5.

3.3 Summary of the Proposed Models

So far, we have presented three different approaches for the QFAS task, one for the single-document scenario and two for the multi-document scenario (see Figure 6). In all three approaches, we first pre-trained a transformer-based summarization model (e.g., BERTSUM) on a large generic abstraction summization dataset. For the single-document scenario (see Figure 6a), we fine-tune the pre-trained model on the target query-focused summarization dataset by incorporating query relevance. For the multi-document scenario (see Figure 6b), we propose two approaches: (i) PreQFAS$_{WSL}$: a weakly supervised approach that generates weak labels, that is, the weak reference summary of each document in a document set to fine-tune a pre-trained transformer-based summarization model by avoiding the computational issues; and (ii) PreQFAS$_{SFT}$: a sequential fine-tuning technique that first selects the most relevant sentences from the multi-document set and sends them to a pre-trained transformer-based summarization model to fine-tune the model sequentially using multiple gold summaries. In the first approach, we perform the sentence filtering at the last step of summary generation, while in the second approach we perform the sentence filtering as an early step.

---

4 Each fine-tuning run may consist of *X* epochs or *Y* steps used for training a neural network model.

Final Summary **S**   Final Summary **S**   Final Summary **S**

Step 2: Fine-Tuning

| Filter out the most relevant sentences as the final summary **S** from the generated summaries **S'** |

*Query focused abstractive summaries **S'** generated for all documents in **D***

| Fine-tune the pre-trained summarization model | | Fine-tune the pre-trained summarization model | | Sequentially fine-tune the pre-trained summarization model |

*Each document **d_i** in **D** with their respective weak reference summary **wR_i***   *Filtered Input **D'** from **D** with all reference summaries **R***

| | | Generate Weak Labels | | Filter Sentences |

| Document **d,** Query **Q** | | Document-Set **D** = {d_1 . . . d_n}, Query **Q** | | Document-Set **D** = {d_1 . . . d_n}, Query **Q** |

Approach 1: Weakly Supervised Learning   **Approach 2: Sequential Fine-Tuning**

(a) Single-Document   **(b) Multi-Document**

Step 1: Pre-Training | Pre-train a Transformer-based generic abstractive summarization model |

Figure 6
An overview summary of our proposed approaches: (a) one approach for single-document scenarios, (b) two approaches for multi-document scenarios.

4. Experimental Setup

In this section, we describe the datasets that we use to evaluate the effectiveness of our approach, followed by the evaluation metrics, the training parameters that have been used in our experiments, and finally the implementation details of our proposed models.

4.1 Datasets

For the QFAS task in single-document scenarios, we primarily use the Debatepedia (Nema et al. 2017) dataset to evaluate our proposed approach. Additionally, due to the lack of available datasets for this task, we modify the QA-NLG dataset from MS-MARCO (Wang et al. 2018) and utilize it for the SD-QFAS task in order to investigate the generalized effectiveness of our proposed approach across different datasets in a related domain. For the QFAS task in multi-document scenarios, we use three datasets from DUC (2005, 2006, 2007) because these datasets are widely used for such tasks. Below, we discuss all datasets in detail.

Debatepedia: Debatepedia is an encyclopedia of pro and con arguments and quotes on debate topics. Nema et al. (2017) utilized Debatepedia to create a dataset containing 13,573 instances for the SD-QFAS task. The average number of words per document, summary, and query in the Debatepedia dataset is 66.4, 11.16, and 9.97, respectively. They used 10-fold cross-validation in their experiments, where each fold has 80% data for training, 10% for validation, and 10% for testing, which resulted in average instances of 10,859 for training, 1,357 for testing, and 1,357 for validation, respectively. We pre-processed the dataset by removing the start token <s> and the end token <eos>.

MS-MARCO: As mentioned earlier, due to the lack of datasets for the SD-QFAS task, we also utilize the QA-NLG dataset from MS-MARCO (Wang et al. 2018), which was designed for the abstractive answer generation task. With 153,725 training samples, this dataset is much larger than the Debatepedia dataset. Therefore, this dataset can give useful insights to investigate the generalized effectiveness of our model rather than

only evaluating its performance for few-shot learning. In the original task setup, a set of passages along with a query are given and the goal is to generate an abstractive answer from the most relevant passage among them. To treat this dataset as an SD-QFAS dataset, we follow the work of Nishida et al. (2019), in which they utilized only the gold passages in the training set as well as in the development set in one of their experiments. We use this dataset similarly by utilizing only the gold passage as the single source document along with the associated query. We use the development set of this dataset that contains 12,467 queries for evaluation. During experiments, we used 10% data from the training set for validation.

   DUC: We use the DUC 2005, 2006, and 2007 datasets for the MD-QFAS task. The number of multi-document sets were 50, 50, and 45 and the average number of documents in each multi-document set were 32, 25, and 25 in DUC 2005, 2006, and 2007 datasets, respectively (Feigenblat et al. 2017). Each document set is associated with a topic statement (considered as the query) and the goal is to generate a summary containing at most 250 words from the document set based on that query. Given the absence of the training data, to evaluate our model in each year's dataset we use the datasets from the other two years for training. From each year's training data, we randomly selected 20% of the document sets for validation while the rest were used for training.

4.2 Evaluation Metrics

To evaluate the performance of our models in different datasets, we select the evaluation metrics by following the prior studies (Nema et al. 2017; Baumel, Eyal, and Elhadad 2018; Nishida et al. 2019). For the Debatepedia dataset, we report the results based on the Recall-Oriented Understudy for Gisting Evaluation (ROUGE) (Lin 2004) metric in terms of the ROUGE-1, ROUGE-2, and ROUGE-L scores.[5] Though the prior studies that used the Debatepedia dataset reported the ROUGE scores only in terms of the Recall metric (Nema et al. 2017; Baumel, Eyal, and Elhadad 2018), we additionally include the Precision and the F1 metrics. We calculate the result based on the average across 10 folds. For the MS-MARCO dataset, the prior work used the Bilingual Evaluation Understudy (BLEU) (Papineni et al. 2002; Reiter 2018) metric based on unigrams in addition to the ROUGE-L metric for performance evaluation (Nishida et al. 2019). We also use these two metrics in terms of the F1 score to evaluate our proposed models. For the DUC datasets, we report the results based on both Recall and F1 metrics in terms of ROUGE-1, ROUGE-2, and ROUGE-SU4 scores (Lin 2004) using the standard parameter setting[6] as used in prior work (Feigenblat et al. 2017; Roitman et al. 2020).

4.3 Training and Parameter Settings

In order to pre-train the BERTSUM model on a generic abstractive summarization dataset, we adopt the BERTSUM models that are pre-trained either on the CNN/DailyMail (CNN-DM) dataset or the XSUM dataset, as used by Liu and Lapata (2019b) for generic abstractive summarization.

   For pre-training, we kept the parameters similar to the original work (Liu and Lapata 2019b): dropout = 0.1, label smoothing with smoothing factor = 0.1, hidden units in the transformer decoder = 768 and hidden size for all feed-forward layers = 2,048,

---

5 We used the following package for calculation: `https://pypi.org/project/pyrouge/`.
6 ROUGE-1.5.5.pl −a −c 95 −m −n 2 −2 4 −u −p 0.5 −l 250.

warmup_steps for the encoder = 20,000 and for the decoder = 10,000, learning_rate for the encoder = 0.002 and for the decoder = 0.1. The batch size was also set to 140. When the CNN-DM dataset was used for pre-training, the total pre-training step was 148,000, while for the XSUM dataset the total pre-training step was 30,000.

To fine-tune the BERTSUM model on the target SD-QFAS datasets, we set new values to the following parameters: batch size = 500, warmup_steps_encoder = 6,000, and warmup_steps_decoder = 2,000. We ran an additional 12,000 training steps for fine-tuning to set the total training steps = 160,000 when the CNN-DM dataset was initially used for pre-training. When we used the XSUM dataset for pre-training, we ran an additional 30,000 training steps (in total 60,000) to do the fine-tuning. Moreover, for Debatepedia, we truncated each input document to 100 tokens and at most 25 tokens for each generated summary. For MS-MARCO, each input document was truncated to 256 tokens and each generated summary had 100 tokens. As used in the original BERTSUM model (Liu and Lapata 2019b), we also utilized the beam search decoding mechanism with size = 5. To fine-tune the BERTSUM model for the MD-QFAS task, we kept most parameters similar to what we used for the SD-QFAS task. However, in this case, we only ran 50 additional steps from the pre-trained model for fine-tuning with batch size equal to 250. For the RoBERTa sentence similarity model (Liu et al. 2019b), we fine-tuned its pre-trained model for the pair-wise sentence classification task using the same parameters that were utilized by Laskar, Huang, and Hoque (2020). For all tasks, we used the models for evaluation on the test dataset that performed the best on the validation dataset.

4.4 Implementation

For the RoBERTa model, we use its Large version (Liu et al. 2019b) for all cases for the MD-QFAS task: when we generate the initial weak reference summaries and while ranking the generated query-focused abstractive summaries in the final step of the PreQFAS$_{WSL}$ model, as well as when we create the filtered input document for the PreQFAS$_{SFT}$ model. To implement this model, we use the Transformer library of HuggingFace (Wolf et al. 2019). For the BERTSUM model, we utilize the BERTSUM$_{EXT-ABS}$ architecture that was used by Liu and Lapata (2019b). For implementation, we use the official source code of the BERTSUM[7] model (Liu and Lapata 2019b). All of our experiments were run using NVIDIA V100 with 4 GPUs.

5. Results and Discussions

In this section, we discuss the performance of our proposed PreQFAS architecture in different datasets. We first demonstrate our findings in the SD-QFAS task, followed by discussing our findings in the MD-QFAS task.

5.1 Performance on the SD-QFAS Task

In the following, we first discuss the performance of the PreQFAS model[8] for few-shot learning presented in Section 3.1 on the Debatepedia dataset, followed by performance

---

7 https://github.com/nlpyang/PreSumm.
8 To utilize PreQFAS for SD-QFAS, we adopt the BERTSUM model pre-trained on the XSUM dataset.

Table 2
Performance of different models for the SD-QFAS task on the Debatepedia dataset. Here, 'R', 'P', and 'F' denote 'Recall', 'Precision', and 'F1', respectively, while 'QD' denotes 'Query-Document Attention' and 'BSA' denotes 'Bidirectional Self-Attention'. The results for the DDA, the Selection Driven, the Overlap-Wind, and the RSA model are collected from Nema et al. (2017), Aryal and Chali (2020), Ishigaki et al. (2020), and Baumel, Eyal, and Elhadad (2018), respectively.

MODEL	ROUGE-1			ROUGE-2			ROUGE-L		
	R	**P**	**F**	**R**	**P**	**F**	**R**	**P**	**F**
QR-BERTSUM$_{Vanilla}$	22.3	35.7	26.4	9.9	16.7	11.9	21.2	33.9	25.1
DDA	41.3	–	–	18.8	–	–	40.4	–	–
Selection Driven	43.2	–	–	27.4	–	–	42.7	–	–
Overlap-Wind	44.4	–	–	30.5	–	–	44.2	–	–
RSA	53.1	–	–	16.1	–	–	46.2	–	–
PreQFAS (QD)	**58.0**	60.3	**58.7**	**45.2**	**46.1**	**45.5**	**57.1**	59.2	**57.7**
PreQFAS (BSA)	**58.0**	60.4	58.5	**45.2**	**46.1**	**45.5**	**57.1**	59.3	**57.7**

on the MS-MARCO dataset. Finally, we present a set of case studies as well as ablation studies to provide a deeper understanding of the effectiveness of our approach.

5.1.1 Performance on the Debatepedia Dataset. In order to compare the performance of our proposed model, we adopt the original BERTSUM model (Liu and Lapata 2019b) as a baseline (denoted as QR-BERTSUM$_{Vanilla}$) by concatenating the query with the document as input and train it end-to-end only on the target Debatepedia dataset. In addition to this baseline, we also compare our model with some other models that were evaluated on this dataset: the DDA model by Nema et al. (2017), which was the first model proposed for this dataset; the RSA model (Baumel, Eyal, and Elhadad 2018), which provided the state-of-the-art performance among the recurrent neural network models (in terms of ROUGE-1 and ROUGE-L); the Overlap-Wind model (Ishigaki et al. 2020) (set a new state-of-the-art based on ROUGE-2); and the recently proposed Selection Driven model (Aryal and Chali 2020). For comparison, we evaluate our PreQFAS model for both the query-document attention and the bidirectional self-attention.

Table 2 shows the results for our proposed model compared with other models. We find that the PreQFAS model with both attentions significantly improved the performance over the QR-BERTSUM$_{Vanilla}$ model, which did not leverage any transfer learning from generic abstractive summarization datasets. This improvement suggests the effectiveness of domain adaptation from pre-trained generic abstractive summarization models for the SD-QFAS task in the Debatepedia dataset.

When we compare the performance between different attentions in the PreQFAS model, we observe that both attentions provide the exact same result in terms of most ROUGE scores, with only a few exceptions. Based on the result, we find that the PreQFAS model with the bidirectional self-attention outperforms its QD attention counterpart in two cases in terms of the Precision metric, with an improvement of 0.17% for both ROUGE-1 and ROUGE-L scores. The only case when the PreQFAS model with the QD attention outperforms the PreQFAS with the bidirectional self-attention is one based on the F1 metric in terms of the ROUGE-1 score, with an improvement of 0.34%. The overall result in the Debatepedia dataset suggests that introducing the QD attention is not more effective than the original bidirectional self-attention used by the BERT encoder.

Table 3
Performance of different models for the SD-QFAS task on the MS-MARCO dataset in terms of
ROUGE-L and BLEU-1 based on the F1 metric.

MODEL	ROUGE-L	BLEU-1
QR-BERTSUM$_{Vanilla}$	71.6	70.2
MASQUE (Nishida et al. 2019)	**78.7**	78.1
PreQFAS (QD Attention)	72.3	72.1
PreQFAS (Bidirectional Self-Attention)	78.4	**80.4**

In comparison to prior work, we observe that the proposed PreQFAS model sets a
new state-of-the-art result in all three ROUGE scores for both attentions. More specif-
ically, in terms of Recall, we find that the PreQFAS model (for both attentions) has an
improvement of 9.23% and 23.59% in terms of ROUGE-1 and ROUGE-L, respectively,
over the RSA model (Baumel, Eyal, and Elhadad 2018). As mentioned by Baumel, Eyal,
and Elhadad (2018), the RSA model provided a very low ROUGE Precision score (the
paper does not report the exact score) by generating very long summaries that are 10
times longer than the required length. In contrast, our proposed model shows a high
Precision score by effectively generating summaries according to the required length.
We also observe a huge gain over other models based on the ROUGE-2 score, with an
improvement of 140.43%, 180.75%, 64.96%, and 48.20% over the DDA (Nema et al. 2017),
RSA (Baumel, Eyal, and Elhadad 2018), Selection Driven (Aryal and Chali 2020), and
Overlap-Wind (Ishigaki et al. 2020) models, respectively, in terms of the Recall metric.

5.1.2 Performance on the MS-MARCO Dataset. For the MS-MARCO dataset, in addition
to the baseline,[9] we compare our proposed model with the MASQUE model (Nishida
et al. 2019), the current state of the art in this dataset. We observe from Table 3 that our
proposed PreQFAS model (for both attentions) again outperforms the baseline model.
More specifically, our best performing PreQFAS using the bidirectional self-attention
outperforms the baseline QR-BERTSUM$_{Vanilla}$ with an improvement of 9.50% in terms
of ROUGE-L and 14.53% in terms of BLEU-1. These improvements demonstrate the
effectiveness of our proposed approach that utilizes transfer learning by fine-tuning
pre-trained generic abstractive summarization models. Our model with bidirectional
self-attention also outperforms the MASQUE (Nishida et al. 2019) model by 2.94% in
terms of BLEU-1, while the result was almost identical in terms of the ROUGE-L score.

A possible explanation for why our model could not outperform the MASQUE
model in terms of ROUGE-L is that the size of the training data used for the MASQUE
model was much larger than the training set that we used to fine-tune our model. In
our case, we utilize the QA-NLG dataset from MS-MARCO, where the training data
contains 153,725 instances, whereas the MASQUE model used the QA-ALL dataset,
where the training data contains 808,731 instances (Wang et al. 2018). Despite these dif-
ferences, with less training data our proposed model outperforms the MASQUE model
in terms of BLEU-1, while also achieving a similar result in terms of ROUGE-L.

When we compare between different attentions in the MS-MARCO dataset, we
find that the QD attention is much less effective than the bidirectional self-attention
mechanism. More specifically, we find that when the QD attention is used instead of

---

9 The baseline QR-BERTSUM$_{Vanilla}$ model in Table 3 was trained end-to-end on the MS-MARCO dataset.

the bidirectional self-attention, the performance deteriorates by 7.78% in terms of ROUGE-L and 10.32% in terms of BLEU-1. This can be explained based on the findings of Peters, Ruder, and Smith (2019), as they suggest that the performance in downstream tasks depends on the similarity between the pre-training stage and the fine-tuning stage. Since the BERT encoder was pre-trained by using the bidirectional self-attention, the utilization of the QD attention only during fine-tuning could possibly be the reason behind poorer performance.

Interestingly, when comparing the performance of these attentions in different datasets, we observe a very surprising trend. Based on our experiments, we find that both the QD attention and the bidirectional self-attention perform similarly in the Debatepedia dataset (see Table 2), whereas the QD attention performs more poorly than the bidirectional self-attention in the MS-MARCO dataset (see Table 3). Furthermore, we find that when domain adaptation from pre-trained summarization models is not utilized, the performance of the baseline QR-BERTSUM$_{Vanilla}$ model in the MS-MARCO dataset (see Table 2) is much better than its performance in the Debatepedia dataset (see Table 3). This could be due to the fact that the total number of training instances (153,725 examples) in the MS-MARCO dataset is almost 15 times higher than the number of total training instances (10,859 examples) in the Debatepedia dataset. Nonetheless, the performance improvement in our PreQFAS model from the baseline in all datasets shows that our proposed model is not only effective in handling the few-shot learning problem, but also it can achieve a huge performance gain when the size of the training dataset is large.

5.1.3 Ablation Study. In this section, we conduct ablation tests to demonstrate the effectiveness of using different components in our proposed model. For the ablation test, our key questions are

- **Why does *fine-tuning* help to improve the performance?** To answer this question, we simply use the pre-trained model for inference without fine-tuning it on the target dataset.

- **To what extent is utilizing the *query relevance* useful?** To answer this question, we remove the query as input to our model.

We show the result of our ablation test in Table 4. We can readily see that removing fine-tuning degrades the performance in both the MS-MARCO and the Debatepedia dataset significantly (based on a paired t-test with $p \leq .05$ on both datasets). Removal of query relevance also leads to huge performance deterioration in the MS-MARCO dataset, which is statistically significant based on a paired t-test ($p \leq .05$). Surprisingly, we find that the performance deterioration in Debatepedia is very small (less than 1%) and this difference was not statistically significant according to the paired t-test ($p > .05$). Such a striking difference in performance between MS-MARCO and Debatepedia suggests that the queries in the Debatepedia dataset may not be effective for summarization.

5.1.4 Case Studies. While we found fine-tuning a pre-trained generic summarization model on the target domain to be effective, we now investigate if this approach is still effective for the zero-shot learning scenario where the training dataset for the target domain is not available. To answer this question, we create a zero-shot learning setup, where we first adopt the BERTSUM model (pre-trained on the XSUM dataset) and then instead of fine-tuning the model on the target dataset, we fine-tune it on the

Table 4
Ablation test results in terms of Recall on Debatepedia and in terms of F1 on MS-MARCO. For this ablation test, all models were pre-trained on the XSUM dataset. Because removing the query relevance as well as fine-tuning makes both PreQFAS models (based on attention) the same, we only mention the result once for the Bidirectional Self-Attention. Here, we denote 'Bidirectional Self-Attention' as 'BSA' and 'QD attention' as 'QD', and 'w/o' denotes 'without'.

| | Datasets | | | | |
| | Debatepedia | | | MS-MARCO | |
MODEL	ROUGE-1	ROUGE-2	ROUGE-L	BLEU-1	ROUGE-L
PreQFAS (BSA)	57.96	45.20	57.05	80.39	78.39
PreQFAS (QD)	57.97	45.21	57.06	72.10	72.25
w/o Query Relevance	56.82	44.66	56.07	66.21	61.50
w/o Fine-Tuning	17.36	11.48	13.32	20.14	21.52

MS-MARCO dataset. We choose the MS-MARCO dataset because it is much larger in size and so we hypothesize that fine-tuning on it will provide better generalization.

We study the zero-shot learning scenario with two different target datasets: (i) Debatepedia, and (ii) MEDIQA-Answer Summarization (MEDIQA-AnS) dataset (Savery et al. 2020). MEDIQA-AnS is a question answering dataset in the healthcare domain, where, given a question and a long answer, the goal is to summarize the answer. The MEDIQA-AnS dataset is particularly suitable for the zero-shot setup because it does not contain any training data and so the model needs to generate the abstractive summaries for a given question without any in-domain knowledge. The MEDIQA-AnS dataset has two versions based on the type of the input document: (i) Pages Version: the input is composed of some Web pages that are relevant to the question, and (ii) Passages Version: the input only contains some passages from the relevant Web pages. We use both versions in our study.

For performance comparisons, we use the BERTSUM model as a baseline that was pre-trained only on the XSUM dataset and did not leverage any fine-tuning. The result of our case study is shown in Table 5. We observe that in all zero-shot learning scenarios, fine-tuning the model on a dataset that is not from the target domain is more effective than no fine-tuning at all. For instance, despite the fact that the PreQFAS model was fine-tuned on the MS-MARCO dataset, which is different than the target

Table 5
Case study results to investigate the zero-shot learning performance based on the F1 metric. In this experiment, the BERTSUM$_{XSUM}$ is used as a baseline model that was pre-trained only on the XSUM dataset without any fine-tuning while the PreQFAS$_{MS-MARCO}$ model was first pre-trained on XSUM and then fine-tuned on MS-MARCO. Here, we denote 'ROUGE' as 'R'.

| | Datasets | | | | | | | | |
| **MODEL** | MediQA-AnS (Pages) | | | MediQA-AnS (Passages) | | | Debatepedia | | |
	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L
BERTSUM$_{XSUM}$	19.87	3.59	13.49	21.36	4.02	13.95	13.3	2.8	11.5
PreQFAS$_{MS-MARCO}$	**23.07**	**5.41**	**15.35**	**29.89**	**11.29**	**21.05**	**22.2**	**6.0**	**19.7**

Table 6
Debatepedia dataset analysis based on a randomly sampled 100 examples.

Analysis Type	Result
Queries having no relevance with the documents or the summaries	52%
Queries are Yes/No type close-ended questions	70%
Queries are relevant to the documents but the summaries are more generic	81%

MEDIQA/Debatepedia dataset, it still demonstrates superior performance over the baseline BERTSUM model.

In addition to this case study, we conduct another case study (which can be found in Appendix A) where we investigate the effects of using different datasets for pre-training.

5.1.5 Analyzing the Debatepedia Dataset. Because of the surprising performance in the Debatepedia dataset that we observe in our ablation study in Section 5.1.3 after removing the query relevance, we manually analyze the dataset to identify the possible reasons. For our analysis, we randomly sampled 100 query-document-summary tuples. The result of our analysis is shown in Table 6. We observe that many queries in this dataset are not relevant to the source document or to the reference summary (about 52%). Table 7(a) shows such an example from this dataset where the query has no relevance with the source document and the gold summary. We also find many examples where the query contains only one word, which partially explains the lack of effectiveness of incorporating the query. Furthermore, we find that most queries are just yes/no type questions (see Table 7(b) for an example) that do not necessarily require the

Table 7
Some examples from the Debatepedia dataset.

(a) Query having no relevance with the document or the summary

Query: Does an MBA enhance leadership skills?
Document: Business schools might improve your quantitative presentation and communication skills. It might but get you thinking about ethical and strategy. But two years of case studies aren't go to turn you into a leader if you weren't died one. There's no learning charisma persuasiveness elegance or gut instinct.
Gold Summary: PhD will not improve cm factors of leaders.

(b) Query is a Yes/No type close-ended question

Query: Is investing in new technologies desirable?
Document: Student will neglect their thought skill and rely too much on technology for everything.
Gold Summary: Spending cash on technologies is a waste.

(c) Query is relevant to the document but the summary is quite generic

Query: Is merit-based pay fair?
Document: Merit pay creates an incentive for teachers to cheat by improving student test scores so that they can appear to be doing better as a result of the teacher's work resulting in bonuses and higher pay. Obviously, the resulting differences in pay would not be fair.
Gold Summary: Merit pay motivates teachers to cheat on test-scoring.

generation of a query-focused summary (about 70%). Additionally, we observe that among the documents where the generated summaries are relevant to the query, excluding queries from most of these documents will not have any negative effects on generating the relevant summaries. The possible reason behind this is that because the average document length in Debatepedia is very small (66.4 words on average per document), the gold summaries for most documents tend to reflect the overall generic summaries of these documents (81% according to the result in Table 6), where the queries that are used for such documents do not influence the summary (see Table 7(c) for such an example). These findings strongly indicate that most queries in Debatepedia are not relevant to the generated summaries and as such this dataset can be considered more of a generic summarization dataset (as opposed to a query-focused summarization dataset).

5.2 Performance on the MD-QFAS Task

We now analyze the effectiveness of our approach[10] in the multi-document query focused summarization scenario. Recall that we proposed two model variations for this scenario in Section 3.2.1 and Section 3.2.2, respectively. We denote our proposed approach that utilizes **W**eakly **S**upervised **L**earning as **PreQFAS$_{WSL}$** (see Section 3.2.1), and the one that utilizes **S**equential **F**ine-**T**uning as **PreQFAS$_{SFT}$** (see Section 3.2.2).

We compare our models with two baselines that utilize the pre-trained BERT-SUM model for zero-shot transfer learning without leveraging any supervised signals and fine-tuning. For each document, one baseline generates an extractive (EXT) summary (**BERTSUM$_{EXT}$**), while the other generates an abstractive (ABS) summary (**BERTSUM$_{ABS}$**). Similar to the **PreQFAS$_{WSL}$** model, the generated summaries in both baselines are also ranked using the RoBERTa model. In addition, we compare our models with four recent works: (i) CES-50 (Feigenblat et al. 2017), (ii) RSA (Baumel, Eyal, and Elhadad 2018), (iii) Dual-CES (Roitman et al. 2020), and (iv) QUERYSUM (Xu and Lapata 2020b).

5.2.1 Performance on DUC Datasets. The results of our experiments on the DUC 2005, DUC 2006, and DUC 2007 datasets are shown in Table 8. In all three datasets, both variations of our model outperform all the prior work in terms of the F1 metric. More specifically, in the DUC 2005 dataset, the **PreQFAS$_{SFT}$** sets a new state-of-the-art in all ROUGE scores (outperforms the previous state-of-the-art DUAL-CES by 6.85%, 22.94%, and 14.81% in terms of ROUGE-1, ROUGE-2, and ROUGE-SU4 scores, respectively). In the other two datasets, **PreQFAS$_{WSL}$** model also provides state-of-the-art performance across all ROUGE scores. In DUC 2006, it beats the QUERYSUM model by 4.54%, 13.47%, and 7.52% in terms of ROUGE-1, ROUGE-2, and ROUGE-SU4, respectively. Finally, in DUC 2007, it made an improvement of 3.28% over QUERYSUM based on ROUGE-1, and 5.60% and 5.29% over Dual-CES based on ROUGE-2 and ROUGE-SU4, respectively.

In terms of the Recall metric, the proposed **PreQFAS$_{WSL}$** model outperforms the prior state-of-the-art (Roitman et al. 2020) in ROUGE-2 in DUC 2005 and 2006 with an improvement of 13.63% and 6.05%, respectively. For the other two metrics (i.e., ROUGE-1 and ROUGE-3) based on Recall, none of our models could outperform the prior state-of-the-art models in any datasets. However, the results are still comparable.

---

10 To utilize PreQFAS for MD-QFAS, we adopt the BERTSUM model pre-trained on the CNN-DM dataset.

Table 8
Performance comparisons in terms of **F1** and **Recall**. Here, '*' denotes an extractive model. Moreover, the results for CES-50, QUERYSUM, DUAL-CES, and RSA are taken from Feigenblat et al. (2017), Xu and Lapata (2020b), Roitman et al. (2020), and Baumel, Eyal, and Elhadad (2018), respectively. Here, we denote 'ROUGE' as 'R'.

				Metric: F1 Score					
					Datasets				
MODEL		DUC 2005			DUC 2006			DUC 2007	
	R-1	R-2	R-SU4	R-1	R-2	R-SU4	R-1	R-2	R-SU4
CES-50 *	37.78	7.45	13.02	40.47	9.13	14.73	42.86	11.34	16.53
QUERYSUM *	–	–	–	41.6	9.5	15.3	43.3	11.6	16.8
DUAL-CES *	38.08	7.54	13.17	41.23	9.47	14.97	43.24	11.78	16.83
BERTSUM$_{EXT}$ *	37.52	7.84	13.29	40.68	9.29	14.66	42.57	11.20	15.98
BERTSUM$_{ABS}$	38.35	7.94	13.44	40.87	9.43	14.83	42.17	10.82	15.98
PreQFAS$_{WSL}$	40.32	9.17	14.71	**43.49**	**10.78**	**16.45**	**44.72**	**12.44**	**17.72**
PreQFAS$_{SFT}$	**40.69**	**9.27**	**15.12**	43.01	10.51	16.40	42.71	10.87	16.45

				Metric: Recall					
					Datasets				
MODEL		DUC 2005			DUC 2006			DUC 2007	
	R-1	R-2	R-SU4	R-1	R-2	R-SU4	R-1	R-2	R-SU4
CES-50 *	40.35	7.94	13.91	43.01	9.69	15.65	45.45	12.02	17.54
RSA	39.82	6.98	**15.73**	42.89	8.73	**17.75**	43.92	10.13	**18.54**
DUAL-CES *	**40.82**	8.07	14.13	**43.94**	10.09	15.96	**46.02**	**12.53**	17.91
BERTSUM$_{EXT}$ *	37.55	7.84	13.31	40.41	9.22	14.56	42.41	11.08	15.92
BERTSUM$_{ABS}$	38.36	7.92	13.43	40.59	9.39	14.73	42.05	10.79	15.91
PreQFAS$_{WSL}$	40.36	**9.17**	14.74	43.22	**10.70**	16.35	44.61	12.40	17.66
PreQFAS$_{SFT}$	39.61	9.01	14.71	41.47	10.08	15.77	41.33	10.52	15.92

When comparing the zero-shot baselines, we find that in both the DUC 2005 and DUC 2006 datasets, our abstractive baseline outperforms its extractive counterpart. However, in the DUC 2007 dataset, we find that the extractive baseline performs better than the abstractive one. This may indicate that the gold reference summaries in the DUC 2007 dataset are more extractive in nature. Moreover, when comparing these two baselines with the proposed **PreQFAS$_{WSL}$** model, we find that for all ROUGE scores, the performance improvement in our proposed model is statistically significant based on a paired t-test ($p \leq .05$).

5.2.2 Ablation Study. We conduct four ablation tests for the multi-document scenario. The first three tests examine how the following components of our weakly supervised approach impact the performance of the **PreQFAS$_{WSL}$** model: (i) *Distant Supervision*, (ii) *Trigram Blocking*, and (iii) *Weakly Supervised Learning*.

Our ablation study results in Table 9 suggest that instead of leveraging *Weakly Supervised Learning* to fine-tune the BERTSUM model, if we directly rank the sentences in the source documents using the RoBERTa$_{MS-MARCO}$ model and select the first 250

Table 9
Ablation test result in terms of ROUGE-1 based on the average across all three datasets. Here, 'without' is denoted by 'w/o'.

Model	Recall	F1	Statistically Significant
PreQFAS$_{WSL}$	**42.73**	**42.84**	
w/o Distant Supervision	41.77 $_{(-2.25\%)}$	41.88 $_{(-2.24\%)}$	No (paired t-test, $p > .05$)
w/o Trigram Blocking	40.92 $_{(-4.24\%)}$	41.01 $_{(-4.27\%)}$	No (paired t-test, $p > .05$)
w/o Weakly Supervised Learning	40.01 $_{(-6.37\%)}$	40.12 $_{(-6.35\%)}$	**Yes** (paired t-test, $p \le .05$)
PreQFAS$_{SFT}$	40.80	42.14	
w/o Fine-Tuning Sequentially	32.87 $_{(-19.44\%)}$	38.63 $_{(-8.33\%)}$	**Yes** (paired t-test, $p \le .05$)

tokens as the summary, the performance is significantly degraded (based on a paired t-test with $p \le .05$). The performance is also deteriorated if we exclude *Distant Supervision* by removing the RoBERTa$_{MRPC}$ model as well as if the *Trigram Blocking* is not utilized. However, in these two cases, the performance deterioration is not statistically significant based on a paired t-test ($p > .05$).

In our final ablation test, we study the effect of sequential fine-tuning in the **PreQFAS$_{SFT}$** model. For this ablation test, instead of sequential fine-tuning where we vary gold reference summaries in multiple runs, we fine-tune only once by varying the gold reference summaries when the same filtered document is given as input to the summarization model in different batches. We find that when the fine-tuning is done only once, the performance deterioration from our proposed sequential fine-tuning approach is statistically significant. This indicates the effectiveness of sequential fine-tuning with multiple runs where we vary the gold reference summaries in each run.

5.2.3 Case Studies. In this section, we perform case studies to investigate how modifying different stages of PreQFAS impact its performance. For these case studies, we investigate the following questions for the PreQFAS$_{WSL}$ model and the PreQFAS$_{SFT}$ model:

    i.   **For PreQFAS$_{WSL}$**, we investigate what happens if we fine-tune other pre-trained transformer-based generic abstractive summarization models such as BART (Lewis et al. 2019), PEGASUS (Zhang et al. 2019a), and T5 (Raffel et al. 2019) instead of BERTSUM.

    ii.   **For PreQFAS$_{SFT}$**, we investigate how the total number of gold reference summaries K used for fine-tuning impacts the performance.

Below, we present our findings for each of the above questions.

(i) Fine-Tuning Other Models for Summary Generation using PreQFAS$_{WSL}$. Recall that we fine-tune the pre-trained BERTSUM model (Liu and Lapata 2019b) to generate the query-focused abstractive summaries since it uses a simple transformer architecture that has fewer complexities (see Section 3.1 for more details). Thus, it allows us to evaluate the effectiveness of our proposed approach while utilizing a conceptually simple model. Now, we investigate how replacing BERTSUM with other newly proposed transformer-based summarization models affects the performance. More specifically,

we use the following pre-trained models for fine-tuning to generate query-focused abstractive summaries.

BART: BART (Lewis et al. 2019) is a sequence-to-sequence model based on the transformer architecture. It was pre-trained based on denoising objectives to map a corrupted document to its original form. To pre-train BART for the original document reconstruction, the following objectives were utilized: *document rotation, sentence permutation, text-infilling, token masking,* and *token deletion*. We choose the pre-trained BART model since fine-tuning this model was found to be effective for the text generation task. Moreover, it utilizes a bidirectional encoder similar to the BERT encoder, while using a left-to-right autoregressive decoder.

PEGASUS: This is another transformer-based encoder-decoder model that we choose for analysis because it is particularly designed for abstractive summarization (Zhang et al. 2019a). For its pre-training objective, it resembles the downstream abstractive summarization task, which involves generating summary-like text from an input document. To do so, it first selects and masks some sentences from the input document(s). Then it concatenates these selected sentences together to use them as a pseudo-summary. To select these sentences, the PEGASUS model investigates different approaches, such as: (i) randomly selecting m sentences from the input document, (ii) selecting the first m sentences in the input document, and (iii) computing the ROUGE-1 score between each sentence and the rest of the document to select the top m scored sentences. Using one of these approaches,[11] it identifies the sentences that are more important to the document and utilizes them as a pseudo reference summary for self-supervised learning. This way of self-supervised pre-training on large datasets leads to better and faster fine-tuning performance on various downstream abstractive summarization datasets.

T5: The T5 model (Raffel et al. 2019) is also a transformer model based on the BERT architecture. However, contrary to the traditional BERT-based models (Devlin et al. 2019; Liu et al. 2019b; Lan et al. 2019) that classify the given input text to a class label, the T5 model treats all tasks, such as *neural machine translation*, *text classification*, *question answering*, or *text summarization*, as a sequence-to-sequence problem. The model is pre-trained on a large dataset with different training and masking objectives to identify the best pre-training objective. The pre-trained model is then fine-tuned to generate the correct output for a given input sequence for the required task.

To use these models for our case study, we use the *HuggingFace Transformer* (Wolf et al. 2019) for implementation. We compare the results of these new variations with our originally proposed BERTSUM-based PreQFAS$_{WSL}$ model as well as the current state of the art in different evaluation metrics.

We show the results of our experiments in Table 10, where we find that these new transformer-based models are also effective when used with our proposed PreQFAS$_{WSL}$ architecture. More importantly, some of these models even obtain new state-of-the-art results in different datasets. More specifically, we find that the PreQFAS$_{WSL}$ model with T5 sets a new state of the art in the DUC 2005 dataset in terms of both F1 and Recall in all ROUGE scores.

In the DUC 2006 dataset, we find that the variation that uses the PEGASUS model with PreQFAS$_{WSL}$ sets a new state of the art in terms of ROUGE-1 and ROUGE-2 metrics based on both Recall and F1. In terms of ROUGE-SU4, though our PreQFAS$_{WSL}$

---

11 It was empirically found that the approach that computed the ROUGE-1 score to select the top m scored sentences was more effective (Zhang et al. 2019a).

Table 10
Case study results for the PreQFAS$_{WSL-DS}$ architecture in terms of F1 and Recall on the MD-QFAS datasets based on fine-tuning different models for summary generation. Here, we denote 'ROUGE' as 'R'. For each dataset, the State-Of-The-Art (SOTA) result is taken from the following: for DUC 2005, all results are taken from Roitman et al. (2020); for DUC 2006, all results are taken from Xu and Lapata (2020b); for DUC 2007, R-1 is taken from Xu and Lapata (2020b) while R-2 and R-SU4 are taken from Roitman et al. (2020).

	Metric: F1 Score								
	Datasets								
PreQFAS$_{WSL}$ model	**DUC 2005**			**DUC 2006**			**DUC 2007**		
	R-1	**R-2**	**R-SU4**	**R-1**	**R-2**	**R-SU4**	**R-1**	**R-2**	**R-SU4**
with **BERTSUM**	40.3	9.3	14.7	43.5	10.8	16.5	44.7	12.4	17.7
with **BART**	40.3	8.7	14.8	43.1	10.2	16.1	44.3	11.4	17.0
with **PEGASUS**	39.8	9.2	14.7	**44.3**	**11.5**	16.9	44.8	12.7	17.8
with **T5**	**41.3**	**9.9**	**15.8**	44.0	11.2	**17.0**	**45.4**	**13.0**	**18.3**
SOTA	38.1	7.5	13.2	41.2	9.5	15.3	43.3	11.8	16.8

	Metric: Recall								
	Datasets								
PreQFAS$_{WSL}$ model	**DUC 2005**			**DUC 2006**			**DUC 2007**		
	R-1	**R-2**	**R-SU4**	**R-1**	**R-2**	**R-SU4**	**R-1**	**R-2**	**R-SU4**
with **BERTSUM**	40.4	9.2	14.7	43.2	10.7	16.4	44.6	12.4	17.7
with **BART**	40.3	8.7	14.8	42.9	10.2	16.0	44.2	11.4	17.0
with **PEGASUS**	39.8	9.2	14.7	**44.0**	**11.4**	16.7	44.7	12.6	17.7
with **T5**	**41.1**	**9.8**	**15.8**	43.5	11.1	16.8	45.2	**12.9**	18.2
SOTA	40.8	8.1	15.7	43.9	10.1	**17.8**	**46.0**	12.5	**18.5**

model with T5 sets a new state of the art in the DUC 2006 dataset based on the F1 metric, it fails to outperform the current state-of-the-art RSA model (Baumel, Eyal, and Elhadad 2018) based on the Recall metric.

In the DUC 2007 dataset, we again find that the PreQFAS$_{WSL}$ model with T5 sets a new state of the art in terms of all ROUGE scores based on the F1 metric. However, in terms of Recall, none of our models could outperform the current state-of-the-art models, Roitman et al. (2020) and Baumel, Eyal, and Elhadad (2018), for both ROUGE-1 and ROUGE-SU4, respectively. Based on Recall for ROUGE-2, though both the T5 model and the PEGASUS model with PreQFAS$_{WSL}$ outperform all prior work, the PreQFAS$_{WSL}$ model with T5 performs better than its counterpart the PreQFAS$_{WSL}$ model with PEGA-SUS to set the new state-of-the-art result.

From this case study, we find that fine-tuning different pre-trained transformers in the proposed PreQFAS$_{WSL}$ model is also useful for the QFAS task. This demonstrates the effectiveness of our proposed approach across various transformer-based summarization models. We conduct some additional case studies for the PreQFAS$_{WSL}$ model to investigate (i) *how weak supervision by different pre-trained transformer models* and (ii) *how ranking the summary sentences in the final stage by different answer selection models* may impact the overall performance. The results from these case studies can be found in Appendix B and Appendix C, respectively.

(ii) Varying the Number of Gold Reference Summaries K in PreQFAS$_{SFT}$. In this case study, contrary to the previous analysis where we investigate the performance of the PreQFAS$_{WSL}$ model, here we study the performance of the other variant of the PreQFAS architecture: the PreQFAS$_{SFT}$ model. To do so, we run different experiments with different numbers of gold reference summaries K. In other words, we vary the total number of fine-tuning runs in the PreQFAS$_{SFT}$ model where each fine-tuning run contains a different gold reference summary than other runs. Moreover, we use the following baseline for this case study where batchwise[12] fine-tuning is used: BERTSUM. For the batchwise fine-tuning, instead of utilizing sequential fine-tuning by varying the gold reference summaries in different fine-tuning runs, we run the fine-tuning only once by using different gold reference summaries for the same input document in different batches.

From Figure 7, we find that in all datasets based on F1 and Recall, using multiple gold reference summaries improves the ROUGE-1 score in both PreQFAS$_{SFT}$ (i.e., sequential fine-tuning) and BERTSUM (i.e., batchwise fine-tuning) models. However, the performance gain via increasing the number of gold reference summaries in our proposed PreQFAS$_{SFT}$ model is greater than the baseline BERTSUM. More specifically, the maximum improvement from $k = 1$ to $k = 4$ in our proposed PreQFAS$_{SFT}$ model is 33.23% (obtained in DUC 2007 in terms of the Recall metric), while in the baseline BERTSUM model it is 5.71% (obtained in DUC 2006 in terms of the F1 metric). Furthermore, in terms of F1, the average improvement from $k = 1$ to $k = 4$ in the PreQFAS$_{SFT}$ model is 14.44% while in the BERTSUM model it is 4.99%. The improvement in terms of Recall is even greater as we find that the average improvement from $k = 1$ to $k = 4$ in the PreQFAS$_{SFT}$ model is 30.90% while in the BERTSUM model it is 7.18%. This case study demonstrates the effectiveness of our proposed sequential fine-tuning technique with multiple gold reference summaries.

5.2.4 Human Evaluation. So far, we primarily use the ROUGE scores (Lin 2004) to evaluate our proposed models, which are computed based on exact matches between the tokens of the generated summary and the gold reference summaries. As a consequence, ROUGE scores become lower when the generated summary contains tokens that are semantically similar to the gold reference summaries but not an exact match. Moreover, this fails to consider other important factors such as how much informative or how fluent the generated summary is, as well as whether the generated summary maintains coherence. Therefore, we also conduct human evaluation on Amazon Mechanical Turk[13] for a qualitative analysis of our proposed models. For this purpose, we randomly selected 10 document sets from each of the three DUC datasets (2005–2007). Thus, a total of 30 document sets were selected with each dataset containing 10 document sets. For each document set, we selected 3 human annotators who were asked to rate the summaries of different models with a score between 1 and 5 (inclusive) based on the following three metrics:

(i)   *Informativeness: It measures how much informative is the generated summary.*

---

12 It refers to the traditional training procedure of neural network models where the input is given to the
    model in different batches.
13 https://www.mturk.com/.
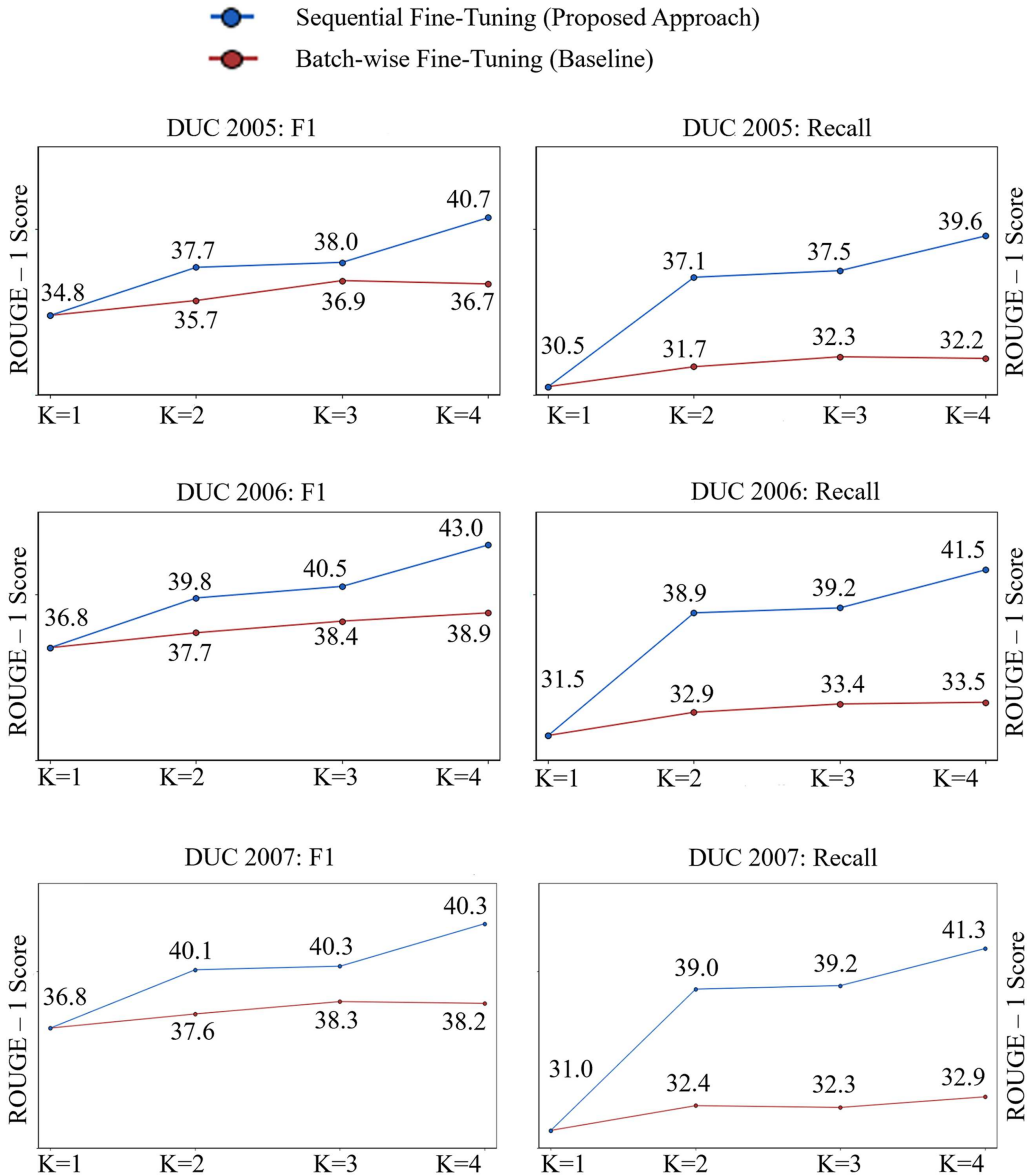
Figure 7
Case study results in terms of ROUGE-1 on various MD-QFAS datasets by varying the total number of gold reference summaries K used for sequential fine-tuning of the proposed PreQFAS$_{SFT}$ model and batchwise fine-tuning of the baseline BERTSUM model.

(ii)   *Coherence: A coherent summary generates a meaningful text where different sentences in the generated summary maintain a consistent connection between them.*

(iii)   *Fluency: A fluent summary contains sentences that are grammatically correct.*

For the human evaluation, we use the summaries generated by all variations of the PreQFAS$_{WSL}$ model: BERTSUM, BART, PEGASUS, and T5. In addition, we use the

Table 11
Human evaluation results in terms of Coherence (C), Fluency (F), and Informativeness (I).

Models	Datasets								
	DUC 2005			DUC 2006			DUC 2007		
	C	F	I	C	F	I	C	F	I
PreQFAS$_{SFT}$ - BERTSUM	3.42	3.34	3.61	3.70	3.73	4.07	3.37	3.33	3.70
PreQFAS$_{WSL}$ - BERTSUM	3.63	3.73	3.63	3.93	3.70	3.97	3.87	3.77	3.83
PreQFAS$_{WSL}$ - BART	**4.23**	**4.20**	**4.49**	**4.50**	**4.43**	**4.57**	4.01	**4.17**	4.37
PreQFAS$_{WSL}$ - PEGASUS	3.87	4.13	4.11	4.23	4.17	4.40	**4.23**	4.07	4.43
PreQFAS$_{WSL}$ - T5	3.90	4.17	4.31	4.11	4.20	4.23	4.13	**4.17**	**4.53**

summaries generated by the PreQFAS$_{SFT}$ model that utilizes the BERTSUM model. We show the results of the human evaluation[14] in Table 11. We find from this table that even though the **PreQFAS$_{WSL}$ - BART** could not set a new state-of-the-art result in terms of different ROUGE scores (see Table 10), it performs better than all other models based on all human evaluation metrics in the DUC 2005 and DUC 2006 datasets. In DUC 2007, we find that in terms of Coherence, **PreQFAS$_{WSL}$ - PEGASUS** performs the best while in terms of Informativeness, the **PreQFAS$_{WSL}$ - T5** performs the best. In terms of Fluency in DUC 2007, we find that both **PreQFAS$_{WSL}$ - T5** and **PreQFAS$_{WSL}$ - BART** perform the best, as they obtain the highest fluency score of 4.17.

Furthermore, we observe that the PreQFAS$_{SFT}$ model performs worse in most cases than all other models that are based on PreQFAS$_{WSL}$. This may suggest that the utilization of weakly supervised learning to fine-tune the pre-trained model on each individual document provides more human-readable summaries than its counterpart that applies filtering on the input document set for summary generation. Moreover, the superior performance of the PreQFAS$_{WSL}$ model that utilizes BART over other models gives a strong indication that the quality of the generated summaries of this model is better in terms of informativeness, fluency, and coherence.

6. Conclusions and Future Work

In this article, we have presented a series of domain adaptation techniques from pre-trained transformer-based models to address the challenge of lack of training data for the query-focused abstractive text summarization task. For the single-document scenario, we perform domain adaptation by pre-training a transformer-based model on a large dataset for generic abstractive summarization followed by fine-tuning it via incorporating the query relevance. For the multi-document scenario, we have presented two domain adaptation techniques that tackle the computational complexity problem for long text sequences. The first approach generates the weak reference summary of each document in the document set using distant supervision to fine-tune the pre-trained summarization model on each document in order to generate the query-focused abstractive summary. The second approach filters out the sentences in the multi-document set to feed only the sentences that are most relevant to the query as input to the

---

14 The cases when at least two out of three annotators agreed on their ratings are 61%, 44%, and 60% for coherence, fluency, and informativeness, respectively.

pre-trained transformer model. Then, we sequentially fine-tune the pre-trained transformer model using all gold reference summaries for a given multi-document set.

We conducted extensive experiments with different variants of transformer-based models and different types of attention mechanisms for incorporating query relevance in the summarization models. Moreover, we conducted a series of ablation studies as well as case studies to carefully investigate the advantages of our proposed architecture. Additionally, we conducted human evaluations in all query-focused multi-document summarization datasets to get a better understanding of the quality of the generated summaries of our proposed models.

To the best of our knowledge, the work presented in this article is the first to give a comprehensive overview of how domain adaptation from pre-trained transformers can be effectively utilized to tackle the few-shot learning problem in both single-document and multi-document query-focused summarization datasets. Our experiments show that utilizing transfer learning from a transformer model pre-trained on a large dataset for generic abstractive summarization and then fine-tuning on the target query-focused summarization dataset results in significant performance gains, setting new state-of-the-art results. In addition, our analysis reveals several new insights including the limitations of the Debatepedia dataset, the superior performance of recent transformer-based models such as T5 (Raffel et al. 2019) and PEGASUS (Zhang et al. 2019a) over their counterparts in the fine-tuning stage, and the weakness of the query-document attention mechanism compared to the bidirectional self-attention mechanism in some datasets. We also analyze the memory requirements of our proposed architecture (see Appendix D for details) and find that it does not require huge computational resources for real-world production deployments.

Our findings in this article lead to several new directions for future work. First, our work demonstrates the pressing need for constructing a new query-focused single-document abstractive summarization dataset, as we find that the existing benchmark dataset for the single-document query-focused summarization task (i.e., the Debate-pedia dataset) is more of a generic summarization dataset and many queries in this dataset have no relation with the reference summaries. Second, we will investigate the performance of our proposed approach with other domain adaptation mechanisms such as adversarial losses and reweighting (Ramponi and Plank 2020). We also aim to investigate the performance of our approach in additional domains (Zhong et al. 2021; Sanh et al. 2021), and with other evaluation metrics (Louis and Nenkova 2013; Xenouleas et al. 2019; Zhang et al. 2019b; Yuan, Neubig, and Liu 2021). Third, we will study our proposed approach using other transformer models (Radford et al. 2019; Tay et al. 2020; Zaheer et al. 2020; Beltagy, Peters, and Cohan 2020; Brown et al. 2020) and explore the performance in other related tasks, such as visual question answering (Antol et al. 2015), chart question answering (Kim, Hoque, and Agrawala 2020), knowledge base question answering (Zhou et al. 2021; Kwiatkowski et al. 2019), sentiment analysis (Liu et al. 2007; Yu et al. 2012), and biomedical information retrieval (Huang and Hu 2009; Huang, Zhong, and Si 2005). Finally, we hope that our source code that we make publicly available will facilitate other researchers in reproducing our work and help to push the state of the art in the future research of query-focused summarization.

Bibliographic Note

Portions of this work have been published as short papers at the COLING 2020 (Laskar, Hoque, and Huang 2020c) and the Canadian AI 2020 (Laskar, Hoque, and Huang 2020a) conference proceedings. However, this work substantially extends the published

papers in several ways, most notably: (i) we investigate the performance of different attentions in query-focused summary generation (Section 3.1); (ii) we propose a novel sequential fine-tuning approach to utilize all the available multi-document reference summaries for supervised training (Section 3.2); (iii) for the query-focused abstractive summarization task in single-document scenarios, we conduct several ablation tests to investigate the effectiveness of different components used in our model (Section 5.1.3) as well as case studies to analyze the effectiveness of our model in the zero-shot learning setup (Section 5.1.4), as well as summarize the key limitations of the Debatepedia dataset (Section 5.1.5); (v) for the multi-document scenario, we study how incorporating recent transformer-based pre-trained summarizers in our proposed model impact performance (Section 5.2.3); and finally, (vi) in addition to extensive experiments on benchmark datasets, we also conduct human evaluation to qualitatively compare among different models proposed for query-focused multi-document summarization (Section 5.2.4). Besides these extensions, the Related Work section was updated and a significant portion of this article was rewritten to adapt to a journal-style publication.

Appendices

Appendix A: Using Different Datasets for Pre-training (Case Study)

Here, we investigate whether using different datasets to pre-train the proposed Pre-QFAS model can lead to different results. For that purpose, we use two pre-training datasets: (i) XSUM, and (ii) CNNDM. After pre-training two BERTSUM models in these two datasets, we fine-tune both pre-trained models on the MS-MARCO dataset (the fine-tuned models are then used to generate the summaries in the MediQA-AnS datasets), as well as fine-tune these models on the Debatepedia dataset (the fine-tuned models are then used to generate the summaries in the target Debatepedia dataset).

Table A1 shows that in all datasets used for evaluation, using the CNN-DM dataset for pre-training is more effective than using the XSUM dataset in terms of all ROUGE scores. This may indicate that using a larger-sized dataset for pre-training is more helpful (the CNN-DM dataset has 287,227 training instances whereas the XSUM dataset contains 204,045 training instances).

Appendix B: Utilizing Different Models for Weak Supervision (Case Study)

As mentioned earlier (see Section 3.2.1), in our PreQFAS$_{WSL}$ model, we generate the initial weak reference summaries using the RoBERTa model (Liu et al. 2019b) fine-tuned

---

Table A1
Case study results based on the F1 metric on MediQA-AnS and Debatepedia while using different datasets for pre-training. Here, we denote 'ROUGE' as 'R'.

	Datasets								
MODEL	MediQA-AnS (Pages)			MediQA-AnS (Passages)			Debatepedia		
	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L
PreQFAS$_{XSUM}$	23.07	5.41	15.35	29.89	11.29	21.05	58.5	45.5	57.7
PreQFAS$_{CNN-DM}$	**25.30**	**7.47**	**17.53**	**33.19**	**15.49**	**24.80**	**59.3**	**45.6**	**58.2**

on the MS-MARCO dataset (Wang et al. 2018) for the answer selection task. To examine the effect of using other models for weak supervision, we use the following variations:

BERTSUM$_{EXT}$: For this variation, we adopt the pre-trained BERTSUM model from (Liu and Lapata 2019b) that was trained for the generic extractive summarization task in the CNN-DM dataset (Hermann et al. 2015). For each dataset, the initial weak reference summaries are first generated using this model. Then, distant supervision from the multi-document gold reference summaries is applied using the RoBERTa$_{MRPC}$ model to generate the weak abstractive reference summary of each training document.

BERTSUM$_{ABS-EXT}$: For this variation, we adopt the pre-trained BERTSUM model from Liu and Lapata (2019b) that was trained for the generic abstractive summarization task (after being initially trained for extractive summarization) in the CNN-DM dataset (Hermann et al. 2015). Then, for each dataset, the initial weak reference summaries are generated using this model. Afterward, we applied distant supervision from the multi-document gold reference summaries using the RoBERTa$_{MRPC}$ model to generate the weak abstractive reference summary of each training document.

We show the result of our experiments in Table B1 and find that for ROUGE-1 and ROUGE-SU4 (in terms of both F1 and Recall), the new variants that utilized the BERTSUM$_{ABS-EXT}$ and the BERTSUM$_{EXT}$ models for initial weak reference summary generation could not outperform the original architecture that utilized RoBERTa. Based on ROUGE-2, we find that in both DUC 2005 and 2007 datasets, the RoBERTa model outperforms the BERTSUM-based weak reference summary generators (for both F1 and Recall). The only scenario in which RoBERTa does not perform the best in this metric is in the DUC 2006 dataset, where the BERTSUM$_{EXT}$ model outperforms other variants in terms of both F1 and Recall metrics. Though the RoBERTa model performs better than other variations in most scenarios for the initial weak reference summary generation

Table B1
Case study results in terms of F1 and Recall on the MD-QFAS datasets based on utilizing various models for weak supervision. Here, we denote 'ROUGE' as 'R'.

Metric: F1 Score									
	Datasets								
PreQFAS$_{WSL}$ model	**DUC 2005**			**DUC 2006**			**DUC 2007**		
	R-1	**R-2**	**R-SU4**	**R-1**	**R-2**	**R-SU4**	**R-1**	**R-2**	**R-SU4**
with RoBERTa	**40.3**	**9.2**	**14.7**	**43.5**	10.8	**16.5**	**44.7**	**12.4**	**17.7**
with BERTSUM$_{EXT}$	40.1	8.9	14.5	43.4	**10.9**	**16.5**	44.3	11.9	17.2
with BERTSUM$_{ABS-EXT}$	40.0	8.7	14.5	42.5	10.6	16.0	44.2	11.8	17.1

Metric: Recall									
	Datasets								
PreQFAS$_{WSL}$ model	**DUC 2005**			**DUC 2006**			**DUC 2007**		
	R-1	**R-2**	**R-SU4**	**R-1**	**R-2**	**R-SU4**	**R-1**	**R-2**	**R-SU4**
with RoBERTa	**40.4**	**9.2**	**14.7**	**43.2**	10.7	**16.4**	**44.6**	**12.4**	**17.7**
with BERTSUM$_{EXT}$	40.2	8.9	14.5	43.1	**10.9**	16.3	44.1	11.8	17.2
with BERTSUM$_{ABS-EXT}$	40.1	8.7	14.5	42.0	10.4	15.8	43.7	11.5	16.9

Table C1
Case study results in terms of F1 and Recall on the MD-QFAS datasets based on using different datasets to fine-tune (FT) the answer selection model. Here, we denote 'ROUGE' as 'R'.

	Metric: F1 Score								
	Datasets								
PreQFAS$_{WSL}$ model	DUC 2005			DUC 2006			DUC 2007		
	R-1	R-2	R-SU4	R-1	R-2	R-SU4	R-1	R-2	R-SU4
FT on MS-MARCO	40.3	9.2	14.7	43.5	10.8	16.5	44.7	12.4	17.7
FT on TREC-QA	39.9	8.9	14.5	42.3	10.3	15.8	43.9	11.5	16.9
FT on Wiki-QA	39.6	8.7	14.2	42.5	10.3	15.9	43.4	11.4	16.7
FT on SemEval (2015)	39.6	8.6	14.2	42.8	10.2	15.9	43.9	11.5	16.7
FT on SemEval (2016–17)	40.0	8.7	14.4	42.9	10.3	15.9	44.4	11.8	17.0

	Metric: Recall								
	Datasets								
PreQFAS$_{WSL}$ model	DUC 2005			DUC 2006			DUC 2007		
	R-1	R-2	R-SU4	R-1	R-2	R-SU4	R-1	R-2	R-SU4
FT on MS-MARCO	40.4	9.2	14.7	43.2	10.7	16.4	44.6	12.4	17.7
FT on TREC-QA	39.9	8.9	14.5	42.1	10.2	15.7	43.8	11.5	16.8
FT on Wiki-QA	39.6	8.7	14.2	42.3	10.3	15.8	43.3	11.4	16.6
FT on SemEval (2015)	39.6	8.6	14.2	42.6	10.1	15.8	43.7	11.4	16.6
FT on SemEval (2016–17)	40.0	8.8	14.4	42.9	10.2	15.8	44.3	11.7	17.0

(based on different datasets and evaluation metrics), the difference between this model and other variants is not statistically significant based on a paired t-test ($p > .05$).

Appendix C: Using Different Datasets for Fine-tuning the Answer Selection Model (Case Study)

During the final stage of the PreQFAS$_{WSL}$ model, we select the most relevant sentences as the query-focused abstractive summary using the RoBERTa$_{Large}$ model fine-tuned for the answer selection task on the MS-MARCO dataset. Now, we examine how varying the dataset for fine-tuning the RoBERTa$_{Large}$ model affects the performance. For this purpose, we utilize the following question answering datasets.

TREC-QA: This dataset is created from the Text REtrieval Conference (Wang, Smith, and Mitamura 2007). It contains 1,229 questions with 53,417 candidate answers.

WikiQA: This is an open domain QA dataset (Yang, Yih, and Meek 2015) that contains 2,118 questions with 20,360 candidate answers from Wikipedia.

SemEvalCQA (2015): This is a Community Question Answering (CQA) dataset that has been created from *Qatar Living Forums*.[15] It contains 2,600 questions and 16,541 candidate answers in the training set.

SemEvalCQA (2016 & 2017): The SemEvalCQA-2016 and the SemEvalCQA-2017 datasets are also CQA datasets created from *Qatar Living Forums*. Both datasets have the same training data containing 4,879 questions and 36,198 candidate answers.

---

15 https://www.qatarliving.com/forum.

We show the results of our experiments in Table C1 to find that in terms of both Recall and F1, the original model that was fine-tuned on the MS-MARCO dataset for the final summary selection performs better than the variations that were find-tuned on other datasets. This could be because the MS-MARCO dataset consists of 153,725 queries and 1,537,250 candidate answers that are much larger than other datasets.

Appendix D: Model Requirements

In this work, we keep our models reasonably lightweight by demonstrating the effectiveness of our proposed approach via utilizing a simple transformer-based summarization architecture, the BERTSUM model, which utilizes the BERT model as encoder and the decoder of Transformer as decoder. The size of the trained model is only 2.32GB, which makes it feasible for the single-document query-focused abstractive summarization task in industrial production scenarios in a computationally limited resource environment.

For our proposed approaches for multi-document scenarios, we additionally use a RoBERTa model that only contains 1.43GB of additional memory. In total, our proposed system can be used for production use-cases in a very lightweight machine that does not require more than 4GB of RAM to host our models (1.43GB RoBERTa model for sentence filtering or final summary selection, and 2.32GB BERTSUM model for summarization). In more powerful computing environments, other models that provide superior performance (e.g., BART, PEGASUS, T5) can be used too.

Appendix E: The Rationale Behind Generating the Initial Weak Extractive Summaries

To generate the weak abstractive reference summary, we apply the RoBERTa sentence similarity model to measure the similarity between each sentence in the weak extractive reference summary and each sentence in the gold reference summaries. However, instead of generating the weak extractive reference summary, if we compare the similarity between each sentence in the multi-document set with each sentence in the gold reference summaries, the model will take a considerable amount of time. In our experiments for multi-document scenarios, each DUC dataset consists of up to 37,925 sentences on average, whereas the average number of sentences in the gold reference summaries per dataset may contain only about 3,089 sentences. Thus, if there are M sentences in the document set and N sentences in the gold reference summaries, the total time complexity of the similarity modeling will be $O(M * N)$. This will make the weak reference summary generation process very time-consuming, which may not be acceptable in real-world scenarios where models are required to be trained regularly.

Therefore, to avoid this huge computation, for each document in a document set we first generate the initial weak reference summary (3 sentences long) using a pre-trained model. And then, based on the similarity score, we replace each sentence in the weak extractive reference summary with the most relevant sentence in the gold summaries.

References

Abdullah, Deen Mohammad and Yllias Chali. 2020. Towards generating query to perform query focused abstractive summarization using pre-trained model. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 80–85.

Antol, Stanislaw, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*. pages 2425–2433. `https://doi.org/10.1109/ICCV.2015.279`

Aryal, Chudamani and Yllias Chali. 2020. Selection driven query focused abstractive document summarization. In *Canadian Conference on Artificial Intelligence*, pages 118–124. `https://doi.org/10.1007/978-3-030-47358-7_11`

Baumel, Tal, Matan Eyal, and Michael Elhadad. 2018. Query focused abstractive summarization: Incorporating query relevance, multi-document coverage, and summary length constraints into seq2seq models. *arXiv preprint arXiv:1801.07704*.

Beltagy, Iz, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Chopra, Sumit, Michael Auli, and Alexander M. Rush. 2016. Abstractive sentence summarization with attentive recurrent neural networks. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 93–98. `https://doi.org/10.18653/v1/N16-1012`

Choromanski, Krzysztof, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Lukasz Kaiser, et al. 2020. Rethinking attention with performers. *arXiv preprint arXiv:2009.14794*.

Clark, Kevin, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

Deng, Yang, Wai Lam, Yuexiang Xie, Daoyuan Chen, Yaliang Li, Min Yang, and Ying Shen. 2019. Joint learning of answer selection and answer summary generation in community question answering. *arXiv preprint arXiv:1911.09801*. `https://doi.org/10.1609/aaai.v34i05.6266`

Devlin, Jacob, 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*, pages 4171–4186.

Dong, Li, 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 *Advances in Neural Information Processing Systems*, pages 13063–13075.

Fabbri, Alexander Richard, Simeng Han, Haoyuan Li, Haoran Li, Marjan Ghazvininejad, Shafiq Joty, Dragomir Radev, and Yashar Mehdad. 2021. Improving zero and few-shot abstractive summarization with intermediate fine-tuning and data augmentation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 704–717. `https://doi.org/10.18653/v1/2021.naacl-main.57`

Feigenblat, Guy, Haggai Roitman, Odellia Boni, and David Konopnicki. 2017. Unsupervised query-focused multi-document summarization using the cross entropy method. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 961–964. `https://doi.org/10.1145/3077136.3080690`

Fu, Xue Yong, Cheng Chen, Md Tahmid Rahman Laskar, Shashi Bhushan, and Simon Corston-Oliver. 2021. Improving punctuation restoration for speech transcripts via external data. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*,

pages 168–174. https://doi.org/10
.18653/v1/2021.wnut-1.19

Garg, Siddhant, Thuy Vu, and Alessandro
Moschitti. 2019. TANDA: Transfer and
adapt pre-trained transformer models for
answer sentence selection. *arXiv preprint
arXiv:1911.04118*. https://doi.org/10
.1609/aaai.v34i05.6282

Goodwin, Travis, Max Savery, and Dina
Demner-Fushman. 2020. Flight of the
PEGASUS? Comparing transformers on
few-shot and zero-shot multi-document
abstractive summarization. In *Proceedings
of the 28th International Conference on
Computational Linguistics*, pages 5640–5646.
https://doi.org/10.18653/v1/2020
.coling-main.494

Haghighi, Aria and Lucy Vanderwende.
2009. Exploring content models for
multi-document summarization. In
*Proceedings of Human Language Technologies:
The 2009 Annual Conference of the North
American Chapter of the Association for
Computational Linguistics*, pages 362–370.
https://doi.org/10.3115/1620754
.1620807

Hermann, Karl Moritz, Tomas Kocisky,
Edward Grefenstette, Lasse Espeholt,
Will Kay, Mustafa Suleyman, and
Phil Blunsom. 2015. Teaching machines to
read and comprehend. In *Advances in
Neural Information Processing Systems*,
pages 1693–1701.

Huang, Xiangji and Qinmin Hu. 2009. A
Bayesian learning approach to promoting
diversity in ranking for biomedical
information retrieval. In *Proceedings of the
32nd International ACM SIGIR Conference on
Research and Development in Information
Retrieval*, pages 307–314. https://
doi.org/10.1145/1571941.1571995

Huang, Xiangji, Ming Zhong, and Luo Si.
2005. York University at TREC 2005:
Genomics track. In *Proceedings of the
Fourteenth Text REtrieval Conference, TREC*,
pages 56–59.

Ishigaki, Tatsuya, Hen-Hsen Huang, Hiroya
Takamura, Hsin-Hsi Chen, and Manabu
Okumura. 2020. Neural query-biased
abstractive summarization using copying
mechanism. In *European Conference on
Information Retrieval*, pages 174–181.
https://doi.org/10.1007/978-3
-030-45442-5_22

Kim, Dae Hyun, Enamul Hoque, and
Maneesh Agrawala. 2020. Answering
questions about charts and generating
visual explanations. In *Proceedings of the
2020 CHI Conference on Human Factors in

Computing Systems*, pages 1–13. https://
doi.org/10.1145/3313831.3376467

Kitaev, Nikita, Lukasz Kaiser, and Anselm
Levskaya. 2019. Reformer: The efficient
transformer. In *International Conference on
Learning Representations*.

Kulkarni, Sayali, Sheide Chammas, Wan
Zhu, Fei Sha, and Eugene Ie. 2020.
Aquamuse: Automatically generating
datasets for query-based multi-document
summarization. *arXiv preprint
arXiv:2010.12694*. https://doi.org/10
.1007/978-3-030-86331-9_6

Kwiatkowski, Tom, Jennimaria Palomaki,
Olivia Redfield, Michael Collins, Ankur
Parikh, Chris Alberti, Danielle Epstein,
Illia Polosukhin, Jacob Devlin, Kenton Lee,
et al. 2019. Natural questions: A
benchmark for question answering
research. *Transactions of the Association for
Computational Linguistics*, 7:453–466.

Lai, Tuan, Quan Hung Tran, Trung Bui,
and Daisuke Kihara. 2019. A gated
self-attention memory network for answer
selection. In *Proceedings of the 2019
Conference on Empirical Methods in Natural
Language Processing and the 9th International
Joint Conference on Natural Language
Processing*, pages 5955–5961.
https://doi.org/10.18653/v1/D19
-1610

Lan, Zhenzhong, Mingda Chen, Sebastian
Goodman, Kevin Gimpel, Piyush Sharma,
and Radu Soricut. 2019. AlBERT: A lite
BERT for self-supervised learning of
language representations. *arXiv preprint
arXiv:1909.11942*.

Laskar, Md Tahmid Rahman, Enamul
Hoque, and Jimmy Huang. 2020a. Query
focused abstractive summarization via
incorporating query relevance and transfer
learning with transformer models. In
Canadian Conference on Artificial Intelligence,
pages 342–348. https://doi.org/10
.1007/978-3-030-47358-7_35

Laskar, Md Tahmid Rahman, Enamul
Hoque, and Jimmy Xiangji Huang. 2020b.
Utilizing bidirectional encoder
representations from transformers for
answer selection. *arXiv preprint
arXiv:2011.07208*. https://doi.org/10
.1007/978-3-030-63591-6_63

Laskar, Md Tahmid Rahman, Enamul
Hoque, and Xiangji Huang. 2020c.
WSL-DS: Weakly supervised learning with
distant supervision for query focused
multi-document abstractive
summarization. In *Proceedings of the 28th
International Conference on Computational

Linguistics, pages 5647–5654.
https://doi.org/10.18653/v1/2020
.coling-main.495

Laskar, Md Tahmid Rahman, Xiangji Huang,
and Enamul Hoque. 2020. Contextualized
embeddings based transformer encoder
for sentence similarity modeling in answer
selection task. In *Proceedings of the 12th
Language Resources and Evaluation
Conference*, pages 5505–5514.

Lewis, Mike, 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*. https://doi.org/10
.18653/v1/2020.acl-main.703

Li, Wei and Hai Zhuge. 2019. Abstractive
multi-document summarization based on
semantic link network. *IEEE Transactions
on Knowledge and Data Engineering*,
33(1):43–54.

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

Liu, Xiaodong, Pengcheng He, Weizhu Chen,
and Jianfeng Gao. 2019a. Improving
multi-task deep neural networks via
knowledge distillation for natural
language understanding. *arXiv preprint
arXiv:1904.09482*.

Liu, Yang, Xiangji Huang, Aijun An,
and Xiaohui Yu. 2007. ARSA: A
sentiment-aware model for predicting
sales performance using blogs. In
*Proceedings of the 30th International ACM
SIGIR Conference on Research and
Development in Information Retrieval*,
pages 607–614. https://doi.org/10
.1145/1277741.1277845

Liu, Yang and Mirella Lapata. 2019a.
Hierarchical transformers for
multi-document summarization. In
*Proceedings of the 57th Annual Meeting of the
Association for Computational Linguistics*,
pages 5070–5081. https://doi.org/10
.18653/v1/P19-1500

Liu, Yang and Mirella Lapata. 2019b. Text
summarization with pretrained encoders.
In *Proceedings of the 2019 Conference on
Empirical Methods in Natural Language
Processing and the 9th International Joint
Conference on Natural Language Processing*,
pages 3721–3731. https://doi.org/10
.18653/v1/D19-1387

Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei
Du, Mandar Joshi, Danqi Chen, Omer

Levy, Mike Lewis, Luke Zettlemoyer, and
Veselin Stoyanov. 2019b. RoBERTa: A
robustly optimized BERT pretraining
approach. *arXiv preprint arXiv:1907.11692*.

Louis, Annie and Ani Nenkova. 2013.
Automatically assessing machine
summary content without a gold standard.
Computational Linguistics, 39(2):267–300.
https://doi.org/10.1162/COLI_a_00123

Ma, Shulei, Zhi-Hong Deng, and Yunlun
Yang. 2016. An unsupervised
multi-document summarization
framework based on neural document
model. In *Proceedings of COLING 2016,
the 26th International Conference on
Computational Linguistics: Technical Papers*,
pages 1514–1523.

Nallapati, Ramesh, Bowen Zhou, Cicero dos
Santos, Çağlar Gulçehre, 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.
https://doi.org/10.18653/v1/K16-1028

Nema, Preksha, Mitesh M Khapra, Anirban
Laha, and Balaraman Ravindran. 2017.
Diversity driven attention model for
query-based abstractive summarization. In
*Proceedings of the 55th Annual Meeting of the
Association for Computational Linguistics
(Volume 1: Long Papers)*, pages 1063–1072.
https://doi.org/10.18653/v1/P17-1098

Nishida, Kyosuke, Itsumi Saito, Kosuke
Nishida, Kazutoshi Shinoda, Atsushi
Otsuka, Hisako Asano, and Junji Tomita.
2019. Multi-style generative reading
comprehension. In *Proceedings of the 57th
Annual Meeting of the Association for
Computational Linguistics*, pages 2273–2284.
https://doi.org/10.18653/v1/P19-1220

Papineni, Kishore, 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.
https://doi.org/10.3115/1073083
.1073135

Pasunuru, Ramakanth, Asli Celikyilmaz,
Michel Galley, Chenyan Xiong, Yizhe
Zhang, Mohit Bansal, and Jianfeng Gao.
2021. Data augmentation for abstractive
query-focused multi-document
summarization. *arXiv preprint
arXiv:2103.01863*.

Paulus, Romain, Caiming Xiong, and
Richard Socher. 2018. A deep reinforced
model for abstractive summarization. In

International Conference on Learning Representations.

Peters, Matthew E., Sebastian Ruder, and Noah A. Smith. 2019. To tune or not to tune? Adapting pretrained representations to diverse tasks. In *Proceedings of the 4th Workshop on Representation Learning for NLP*, pages 7–14. https://doi.org/10.18653/v1/W19-4302

Qi, Weizhen, Yu Yan, Yeyun Gong, Dayiheng Liu, Nan Duan, Jiusheng Chen, Ruofei Zhang, and Ming Zhou. 2020. ProphetNet: Predicting future N-gram for sequence-to-sequence pre-training. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 2401–2410. https://doi.org/10.18653/v1/2020.findings-emnlp.217

Qiu, Xipeng, Tianxiang Sun, Yige Xu, Yunfan Shao, Ning Dai, and Xuanjing Huang. 2020. Pre-trained models for natural language processing: A survey. *arXiv preprint arXiv:2003.08271.* https://doi.org/10.1007/s11431-020-1647-3

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

Raffel, Colin, 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. *arXiv preprint arXiv:1910.10683.*

Ramponi, Alan and Barbara Plank. 2020. Neural unsupervised domain adaptation in NLP—a survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6838–6855. https://doi.org/10.18653/v1/2020.coling-main.603

Reiter, Ehud. 2018. A structured review of the validity of BLEU. *Computational Linguistics*, 44(3):393–401. https://doi.org/10.1162/coli_a_00322

Roitman, Haggai, Guy Feigenblat, Doron Cohen, Odellia Boni, and David Konopnicki. 2020. Unsupervised dual-cascade learning with pseudo-feedback distillation for query-focused extractive summarization. In *Proceedings of The Web Conference 2020*, pages 2577–2584. https://doi.org/10.1145/3366423.3380009

Rush, Alexander M., Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 379–389. https://doi.org/10.18653/v1/D15-1044

Sanh, Victor, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al. 2021. Multitask prompted training enables zero-shot task generalization. *arXiv preprint arXiv:2110.08207.*

Savery, Max, Asma Ben Abacha, Soumya Gayen, and Dina Demner-Fushman. 2020. Question-driven summarization of answers to consumer health questions. *Scientific Data*, 7(1):1–9. https://doi.org/10.1038/s41597-020-00667-z

See, Abigail, 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. https://doi.org/10.18653/v1/P17-1099

Song, Kaitao, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2019. MASS: Masked sequence to sequence pre-training for language generation. In *International Conference on Machine Learning*, pages 5926–5936.

Su, Dan, Yan Xu, Tiezheng Yu, Farhad Bin Siddique, Elham J. Barezi, and Pascale Fung. 2020. CAiRE-COVID: A question answering and query-focused multi-document summarization system for COVID-19 scholarly information management. *arXiv preprint arXiv:2005.03975.* https://doi.org/10.18653/v1/2020.nlpcovid19-2.14

Su, Dan, Tiezheng Yu, and Pascale Fung. 2021. Improve query focused abstractive summarization by incorporating answer relevance. *arXiv preprint arXiv:2105.12969.* https://doi.org/10.18653/v1/2021.findings-acl.275

Sutskever, Ilya, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*, pages 3104–3112.

Tay, Yi, Mostafa Dehghani, Dara Bahri, and Donald Metzler. 2020. Efficient transformers: A survey. *arXiv preprint arXiv:2009.06732.*

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones,

Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Wan, Xiaojun and Jianguo Xiao. 2009. Graph-based multi-modality learning for topic-focused multi-document summarization. In *Twenty-First International Joint Conference on Artificial Intelligence*.

Wan, Xiaojun and Jianmin Zhang. 2014. CTSUM: extracting more certain summaries for news articles. In *Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 787–796. https:// doi.org/10.1145/2600428.2609559

Wang, Dingding, Shenghuo Zhu, Tao Li, Yun Chi, and Yihong Gong. 2008. Integrating clustering and multi-document summarization to improve document understanding. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, pages 1435–1436. https://doi.org/10.1145/1458082 .1458319

Wang, Mengqiu, Noah A. Smith, and Teruko Mitamura. 2007. What is the jeopardy model? A quasi-synchronous grammar for QA. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 23–32.

Wang, Sinong, Belinda Z. Li, Madian Khabsa, Han Fang, and Hao Ma. 2020. Linformer: Self-attention with linear complexity. *arXiv preprint arXiv:2006.04768*.

Wang, Yizhong, Kai Liu, Jing Liu, Wei He, Yajuan Lyu, Hua Wu, Sujian Li, and Haifeng Wang. 2018. Multi-passage machine reading comprehension with cross-passage answer verification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1918–1927. https://doi.org/10.18653/v1 /P18-1178

Wolf, Thomas, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. HuggingFace's Transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771. https://doi.org/10.18653/v1/2020 .emnlp-demos.6

Xenouleas, Stratos, Prodromos Malakasiotis, Marianna Apidianaki, and Ion Androutsopoulos. 2019. Sum-QE: A

BERT-based summary quality estimation model. 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 6005–6011. https://doi.org/10.18653/v1/D19-1618

Xie, Yujia, Tianyi Zhou, Yi Mao, and Weizhu Chen. 2020. Conditional self-attention for query-based summarization. *arXiv preprint arXiv:2002.07338*.

Xu, Yumo and Mirella Lapata. 2020a. Abstractive query focused summarization with query-free resources. *arXiv preprint arXiv:2012.14774*. https://doi.org/10 .18653/v1/2021.acl-long.475

Xu, Yumo and Mirella Lapata. 2020b. Coarse-to-fine query focused multi-document summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3632–3645. https://doi.org/10.18653/v1 /2020.emnlp-main.296

Xu, Yumo and Mirella Lapata. 2021. Text summarization with latent queries. *arXiv preprint arXiv:2106.00104*.

Yang, Yi, Wen-tau Yih, and Christopher Meek. 2015. WikiQA: A challenge dataset for open-domain question answering. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2013–2018. https://doi.org/10.18653 /v1/D15-1237

Yao, Jin ge, Xiaojun Wan, and Jianguo Xiao. 2015. Compressive document summarization via sparse optimization. *Twenty-Fourth International Joint Conference on Artificial Intelligence*, pages 1376–1382.

Yao, Jin Ge, Xiaojun Wan, and Jianguo Xiao. 2017. Recent advances in document summarization. *Knowledge and Information Systems*, 53(2):297–336. https://doi.org /10.1007/s10115-017-1042-4

Young, Tom, Devamanyu Hazarika, Soujanya Poria, and Erik Cambria. 2017. Recent trends in deep learning based natural language processing. *arXiv preprint arXiv:1708.02709*.

Yu, Xiaohui, Yang Liu, Xiangji Huang, and Aijun An. 2012. Mining online reviews for predicting sales performance: A case study in the movie domain. *IEEE Transactions on Knowledge and Data Engineering*, 24(4):720–734. https://doi.org/10 .1109/TKDE.2010.269

Yuan, Weizhe, Graham Neubig, and Pengfei Liu. 2021. BARTscore: Evaluating

generated text as text generation. *arXiv preprint arXiv:2106.11520*.

Zaheer, Manzil, Guru Guruganesh, Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. 2020. Big Bird: Transformers for longer sequences. *arXiv preprint arXiv:2007.14062*.

Zhang, Jingqing, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2019a. PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization. *arXiv preprint arXiv:1912.08777*.

Zhang, Tianyi, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019b. BERTscore: Evaluating text generation with BERT. In *International Conference on Learning Representations*.

Zhong, Ming, Da Yin, Tao Yu, Ahmad Zaidi, Mutethia Mutuma, Rahul Jha, Ahmed Hassan, et al. 2021. QMSum: A new benchmark for query-based multi-domain meeting summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5905–5921. `https://doi.org/10.18653/v1/2021.naacl-main.472`

Zhong, Sheng hua, Yan Liu, Bin Li, and Jing Long. 2015. Query-oriented unsupervised multi-document summarization via deep learning model. *Expert Systems with Applications*, 42(21):8146–8155. `https://doi.org/10.1016/j.eswa.2015.05.034`

Zhou, Guangyou, Zhiwen Xie, Zongfu Yu, and Jimmy Xiangji Huang. 2021. DFM: A parameter-shared deep fused model for knowledge base question answering. *Information Sciences*, 547:103–118. `https://doi.org/10.1016/j.ins.2020.08.037`