ConvoSumm: Conversation Summarization Benchmark and Improved Abstractive Summarization with Argument Mining

Alexander R. Fabbri[†] **Faiaz Rahman**[†] **Imad Rizvi**[†] **Borui Wang**[†]
Haoran Li [‡] **Yashar Mehdad**[‡] **Dragomir Radev**[†]
[†] Yale University [‡] Facebook AI
{alexander.fabbri, faiaz.rahman, imad.rizvi,
borui.wang, dragomir.radev}@yale.edu
{aimeeli, mehdad}@fb.com

Abstract

While online conversations can cover a vast amount of information in many different formats, abstractive text summarization has primarily focused on modeling solely news articles. This research gap is due, in part, to the lack of standardized datasets for summarizing online discussions. To address this gap, we design annotation protocols motivated by an issues–viewpoints–assertions framework to crowdsource four new datasets on diverse online conversation forms of news comments, discussion forums, community question answering forums, and email threads. We benchmark state-of-the-art models on our datasets and analyze characteristics associated with the data. To create a comprehensive benchmark, we also evaluate these models on widely-used conversation summarization datasets to establish strong baselines in this domain. Furthermore, we incorporate argument mining through graph construction to directly model the issues, viewpoints, and assertions present in a conversation and filter noisy input, showing comparable or improved results according to automatic and human evaluations.

Headline: SuperBowl
Snippet: Whether you're a football fan or not, what do you like about Super Bowl Sunday?
Comment: ... In my opinion I think the Falcons will stomp the patriots. I think Tom Brady will choke the Super Bowl. ...
Comment: I am big Arizona Cardinals fan so when they didn't even make the playoffs i was upset. ...
Comment: I'm not a very big football fan at all. So when it comes to Superbowl Sunday, I'm in it for the commercials and the half time show. ...
Comment: I am not exactly a football fan, but I enjoy watching the Super Bowl....
...
Summary: Several commenters list their favorite things about the Super Bowl, including half-time shows, the funny commercials, the Puppy Bowl, eating food, and spending time with family. A couple of commenters admit to not being football fans but still enjoying the Super Bowl. Some commenters discuss whether they thought the Falcons or the Patriots were going to win, while others list teams they wish were in the game.

Table 1: Example summary of comments from a New York Times article discussing people's favorite parts of the Super Bowl. The summary is an analysis of the comments and quantifies the viewpoints present.

1 Introduction

Automatic text summarization is the process of outputting the most salient parts of an input in a concise and readable form. Recent work in summarization has made significant progress due to introducing large-scale datasets such as the CNN-DailyMail dataset (Nallapati et al., 2016) and the New York Times dataset (Sandhaus, 2008). Furthermore, the use of large self-supervised pretrained models such as BART (Lewis et al., 2020) and Pegasus (Zhang et al., 2019) has achieved state-of-the-art performance across summarization tasks and strong performance in zero and few-shot settings (Fabbri et al., 2020a). However, less work has focused on summarizing online conversations. Unlike documents, articles, and scientific papers, which contain specific linguistic structures and conventions such as topic sentences and abstracts, conversational text scatters main points across multiple utterances and between numerous writers. As a result, the text summarization task in the conversational data domain offers a challenging research field to test newly-developed models (Chen and Yang, 2020).

Recently, Gliwa et al. (2019a) introduced a dataset for chat-dialogue conversation summarization consisting of 16k examples, the first large-scale dataset of its kind. Previous work in conversation summarization was limited by the data available and focused primarily on meeting summarization, such as the AMI (Kraaij et al., 2005) and ICSI (Janin et al., 2003) datasets. The datasets

Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, pages 6866–6880
August 1–6, 2021.

used in recent conversation papers are often not uniform, ranging from visual dialogue data (Goo and Chen, 2018a) to customer-service dialogues (Yuan and Yu, 2019), not initially intended for summarization. The availability of benchmark datasets for comparing methods has limited work in other conversation summarization domains and thus likely inhibited progress (Kryscinski et al., 2019; Fabbri et al., 2020b).

We aim to address this research gap by crowdsourcing a suite of four datasets, which we call **ConvoSumm**, that can evaluate a model's performance on a broad spectrum of conversation data. In determining the domains of data to collect, we use the general definition of conversation as "any discourse produced by more than one person" (Ford, 1991). We identify several key categories of data for which standard human-created development and testing datasets do not exist, namely (1) news article comments, (2) discussion forums and debate, (3) community question answering, and (4) email threads. We design annotation protocols motivated by work in quantifying viewpoints present in news comment data (Barker and Gaizauskas, 2016a) to crowdsource 250 development and 250 test examples for each of the above domains. We provide an example of comments to a New York Times news article, and our crowdsourced summary in Table 1.

In addition to introducing manually-curated datasets for conversation summarization, we also aim to unify previous work in conversation summarization. Namely, we benchmark a state-of-the-art abstractive model on several conversation datasets: dialogue summarization from SAMSum (Gliwa et al., 2019b), heuristic-generated community question answering from CQASumm (Chowdhury and Chakraborty, 2018), meeting summarization data from AMI and ICSI, and smaller test sets in the news comments, discussion forum, and email domains. We believe that such benchmarking will facilitate a more straightforward comparison of conversation summarization models across domains.

To unify modeling across these conversational domains, we propose to use recent work in end-to-end argument mining (Lenz et al., 2020; Stab and Gurevych, 2014; Chakrabarty et al., 2019) to instantiate the theoretical graph framework which motivated our annotation protocol, proposed by Barker and Gaizauskas (2016a) for conversation summarization. This protocol is employed to both identify and use the "issues–viewpoints–assertions" argument structure (discussed in Related Work) for summarizing news comments. We construct this argument graph using entailment relations, linearize the graph, train a graph-to-text model (Ribeiro et al., 2020), and experiment with argument mining as a way to reduce noise in long-text input.

Our contributions are the following: (1) we crowdsource datasets for four domains of conversational data and analyze the characteristics of our proposed datasets; (2) we benchmark state-of-the-art models on these datasets as well as previous widely-used conversation summarization datasets to provide a clear baseline for future work; and (3) we apply argument mining to model the structure of our conversational data better as well as reduce noise in long-text input, showing comparable or improved results in both automatic and human evaluations.[1]

2 Related Work

Modeling Conversation Summarization Early approaches to conversation summarization consisted of feature engineering (Shasha Xie et al., 2008), template selection methods (Oya et al., 2014), and statistical machine learning approaches (Galley, 2006; Wang and Cardie, 2013). More recent modeling approaches for dialogue summarization have attempted to take advantage of conversation structures found within the data through dialogue act classification (Goo and Chen, 2018b), discourse labeling (Ganesh and Dingliwal, 2019), topic segmentation (Liu et al., 2019c), and keypoint analysis (Liu et al., 2019a). Chen and Yang (2020) utilize multiple conversational structures from different perspectives in its sequence-to-sequence model. However, such approaches focus exclusively on dialogue summarization, and it is not trivial to extend such methods to longer conversations with many more participants. We thus introduce a method to model the structure of the discourse over the many-party conversation.

Several existing works have focused on conceptualizing conversation structure for summarization and how to present this structure to end-users. Barker et al. (2016a) propose a conversation overview summary that aims to capture the key argumentative content of a reader comment conversation. Misra et al. (2017) use summarization

[1]For reproducibility of our findings, we will make our data and code publicly available at https://github.com/Yale-LILY/ConvoSumm.

as a means of probing online debates to discover central propositions, which they cluster to identify argument facets. Barker and Gaizauskas (2016b) identify three key components of conversational dialogue: *issues* (that individuals discuss), *viewpoints* (that they hold about these issues), and *assertions* (that they make to support their viewpoints). We build on this framework and advances in argument mining for end-to-end training for summarization.

Argument Mining Work in argument mining (Stab and Gurevych, 2014) has aimed to identify these argumentative units and classify them into claims, premises, and major claims, or claims describing the key concept in a text. More recently, Chakrabarty et al. (2019) propose to finetune BERT (Devlin et al., 2019) for identifying argumentative units and relationships between them within a text and across texts. Lenz et al. (2020) are the first to propose an end-to-end approach for constructing an *argument graph* (Stede et al., 2016), a structured representation of claims and premises in an argumentative text; the graph is built by connecting claim and premise argumentative discourse units. We build on this framework for modeling discourse in conversational data.

Few-Shot Summarization As the datasets we introduce are not on a scale with larger datasets, we focus on few-shot and domain transfer summarization techniques. Wang et al. (2019) examine domain adaptation in extractive summarization, while Hua and Wang (2017) examine domain adaptation between opinion and news summarization. Within unsupervised abstractive summarization, several approaches have made use of variational autoencoders (Baziotis et al., 2019; Chu and Liu, 2019; Bražinskas et al., 2020) and pretrained language models (Zhou and Rush, 2019; Laban et al., 2020).

Recent work in abstractive (Zhang et al., 2019; Fabbri et al., 2020a) and extractive-compressive summarization (Desai et al., 2020) has shown the power of pretrained models for a few-shot transfer. The quality of models trained on several hundred examples in these papers is comparable to that of models trained on the equivalent full datasets. Thus, we believe that introducing curated validation and testing datasets consisting of a few hundred examples is a valuable contribution within the current paradigm, which was confirmed by the poor performance of models transferred from other domains compared to that trained on this validation data.

3 ConvoSumm

In this section, we introduce our dataset selection, our annotation protocol, and the characteristics of our crowdsourced dataset.

Data Selection For the news comments subdomain, we use the NYT Comments dataset, which consists of 2 million comments made on 9,000 New York Times articles published between 2017 and 2018. It is publicly available and has been used in work for news-comment relevance modeling (Kolhatkar and Taboada, 2017); it also contains metadata that may be of use in summarization modeling. For the discussion forums and debate subdomain, we select Reddit data from CoarseDiscourse (Zhang et al., 2017), which contains annotations about the discourse structure of the threads. For the community question answering subdomain, we use StackExchange (Stack), which provides access to all forums and has been used in modeling for answer relevance and question deduplication (Hoogeveen et al., 2015). We chose StackExchange over the commonly-used Yahoo! Answers data due to licensing reasons. For the email threads subdomain, we use the publicly-available W3C corpus (Craswell et al., 2005). Previous work also made use of this dataset for email summarization (Ulrich et al., 2008) but provided only a small sample of 40 email threads, for which we provide transfer testing results.

We generally follow the guidance of Tomasoni and Huang (2010), from summarizing community question answering forums, for determining which subsets of data to select from the above datasets. We remove an example if (1) there were less than five posts (four in the case of email threads; "post" refers to any answer, comment, or email); (2) the longest post was over 400 words; (3) the sum of all post lengths was outside of $[100, 1400]$ words (although we extended this maximum length for NYT comments); or (4) the average length of the posts was outside of the $[50, 300]$ words interval. For Stack data, we first filtered answers which received a negative community rating, as defined by the number of user upvotes minus the number of user downvotes. While real-world settings may contain much longer threads, we later show that this setting is already challenging.

Annotation Protocol We designed annotation instructions for crowdsourced workers to write abstractive summaries for each of the four

Dataset	% novel n-grams	Extractive Oracle	Summary Length	Input Length	# Docs/Example
NYT	36.11/79.72/94.52	36.26/10.21/31.23	79	1624	16.95
Reddit	43.84/84.98/95.65	35.74/10.45/30.74	65	641	7.88
Stack	35.12/77.91/93.56	37.30/10.70/31.93	73	1207	9.72
Email	42.09/83.27/93.98	40.98/15.50/35.22	74	917	4.95

Table 2: Statistics across dataset sources in ConvoSumm, showing novel uni/bi/tri-grams, ROUGE-1/2/L extractive oracle scores, the average input and summary lengths (number of tokens), as well as the number of documents per example, where each comment/post/answer/email is considered a document.

Dataset/Method	Inter-document Similarity	Redundancy	Layout Bias
NYT	-11.71	-0.23	0.2/0.5/0.3
Reddit	-7.56	-0.49	0.2/0.5/0.2
Stack	-9.59	-0.27	0.2/0.3/0.4
Email	-1.76	-0.18	0.3/0.4/0.3

Table 3: Multi-document summarization-specific dataset analysis on our proposed datasets with metrics introduced in Dey et al. (2020a): inter-document similarity (father from zero is less similarity), redundancy (father from zero is less overall redundancy of semantic units), and start/middle/end layout bias.

datasets, motivated by work in summarizing viewpoints present in online conversation (Barker and Gaizauskas, 2016a). We present the crowdsource workers with the data threads, along with any available metadata. For NYT, we presented the workers with the article headline, keywords, and, rather than providing the entire article as context, an extractive BERT-based summary (Miller, 2019) of the article. We use a BERT summary to give the annotators an idea of the topic of the article. We avoided having annotators read the entire article since the focus of their summaries was solely the content of the comments as per the annotation protocols, and reading the entire article could end up introducing information in the summaries that was not necessarily representative of the comments' main points. We found that these summaries were useful in initial in-house annotations, and allowed us to better understand the context of the comments being summarized. For Reddit and Stack, question tags and information about the subforum were provided; the Stack data includes both answers and answer comments. Reddit data was filtered simply on word limits due to the unavailability of up/down votes from the Coarse Discourse data. Stack data includes the prompt/title as well. Whenever possible, we included username information and the scores of all comments, posts, and answers.

Although the instructions differed slightly with the specific nuances of each dataset, they had standard overall rules: (1) summaries should be an analysis of the given input rather than another response or utterance; (2) summaries should be abstractive, i.e., annotators were required to paraphrase and could not repeat more than five words in a row from the source; and (3) summary lengths should contain $[40, 90]$ tokens. Following the issues–viewpoints–assertions framework presented in Barker and Gaizauskas (2016b), we also instructed annotators that summaries should summarize all viewpoints in the input and should try to include specific details from assertions and anecdotes (unless this made the summary too lengthy). Summarizing based on similar viewpoints is analogous to clustering then summarizing, similar to the comment label grouping procedure before summarization in Barker et al. (2016b). To help with this, we recommended wording such as "Most commenters suggest that..." and "Some commenters think that..." to group responses with similar viewpoints.

However, the email dataset was unique among the selected datasets given that it contained more back-and-forth dialogue than clusters of viewpoints, and thus identifying the speakers was essential to creating summaries that still retained meaning from the original email dialogue. Since the email threads contained fewer individual speakers than the other datasets, this sort of summarization remained feasible. Thus, for this dataset, annotators were instructed to specify the speakers when summarizing the conversation.

Quality-Controlled Crowdsourcing We crowdsourced our data using Amazon Mechanical Turk. We required that our workers be native English speakers and pass a qualifying exam for each domain to be summarized. We worked with a select group of about 15 workers who formed a community of high-quality annotators. Example summaries were provided to the workers. The workers submitted the qualifying exam, and then one of the authors of this paper provided feedback. If the worker was not sure of the quality of the summaries

written, at any point, they could enlist the input of one of the authors.

Additionally, after the workers wrote all summaries, we manually reviewed every summary and made corrections to grammar, wording, and overall structure. Summaries we could not fix ourselves, either because they were poorly written or did not follow the annotation protocols, were flagged to be re-written. They were then sent to our approved group of workers to be re-written, excluding any workers who had written a flagged summary. While data crowdsourced from non-experts may contain noise (Gillick and Liu, 2010), we believe that our setup of working closely with a small group of workers, providing feedback to individual workers, and manually reviewing all final summaries mitigates these issues.

Dataset Statistics We provide statistics in Table 2. The percentage of novel n-grams in our summaries is higher than that of the very abstractive XSum dataset (Narayan et al., 2018) (35.76/83.45/95.50 -% novel uni/bi/tri-grams). This level of abstraction is likely due to the instructions to perform abstractive summarization and the summaries being an analysis of the input, which results in the insertion of new words (e.g. "commenters" likely isn't seen in the input). The influence of this abstraction is further seen by an analysis of the Extractive Oracle, for which we show ROUGE-1/2/L (Lin, 2004). We see that the performance of an extractive model is above the Extractive Oracle on the very abstractive XSum (Narayan et al., 2018) (29.79 ROUGE-1), but much lower than the Extractive Oracle on the CNN-DailyMail (CNNDM) dataset (Nallapati et al., 2016) (>50 ROUGE-1). The summary lengths are fairly consistent, while the input lengths are the longest for NYT and Stack data. We include the title and additional meta-data such as the headline and snippet in NYT data in input length calculations.

We analyze multi-document summarization–specific characteristics of our datasets, as proposed by Dey et al. (2020a). In particular, inter-document similarity measures the degree of overlap of semantic units in the candidate documents, with scores further from zero signifying less overlap. The notion introduced for redundancy measures the overall distribution of semantic units; the farther the score is from zero, the more uniform semantic units are across the entire input, with the maximum when each unit is present only once. Layout bias measures the similarity of multi-sentential documents with the reference. For more precise definitions, we refer the reader to Dey et al. (2020a). We provide results for our data in Table 3. Email data exhibits the most inter-document similarity, which follows the intuition that an email thread consists of a focused discussion typically on a single topic. For redundancy, we see Reddit shows the most uniform distribution of semantic units, perhaps due to Reddit threads' less focused nature compared to the remaining datasets. We do not see a particularly strong layout bias across any parts of the input documents. Our datasets exhibit greater or comparable levels of novel-ngrams compared to multi-document summarization datasets such as MultiNews (Fabbri et al., 2019) and CQASUMM (Chowdhury and Chakraborty, 2018). Our Stack subset has lower inter-document similarity, which presents challenges for models which rely strictly on redundancy in the input, and our datasets generally exhibit less layout bias, when compared to the analysis done in Dey et al. (2020b).

Comparison to Existing Datasets Although previous work on conversation summarization, before the introduction of SAMSum (Gliwa et al., 2019b), has largely featured unsupervised or few-shot methods, there exist several datasets with reference summaries. These include SENSEI (Barker et al., 2016b) for news comments, the Argumentative Dialogue Summary Corpus (ADS) (Misra et al., 2015) for discussion forums, and the BC3 (Ulrich et al., 2009) dataset for email data. However, much of the existing datasets are not wide in scope. For example, SENSEI only covers six topics and the ADS Corpus covers one topic and only has 45 dialogues. Furthermore, they each pertain to one subdomain of conversation. Our dataset avoids these issues by covering four diverse subdomains of conversation and having approximately 500 annotated summaries for each subdomain. Additionally, since neural abstractive summarization baselines do not exist for these datasets, we benchmark our models on these datasets to further their use as test sets. We similarly include the AMI and ICSI meeting datasets within our benchmark.

Within community question answering, the WikiHowQA dataset (Deng et al., 2020) consists of user response threads to non-factoid questions starting with "how to," including labels for the answer selection task and reference summaries. The CQASUMM dataset (Chowdhury and Chakraborty,

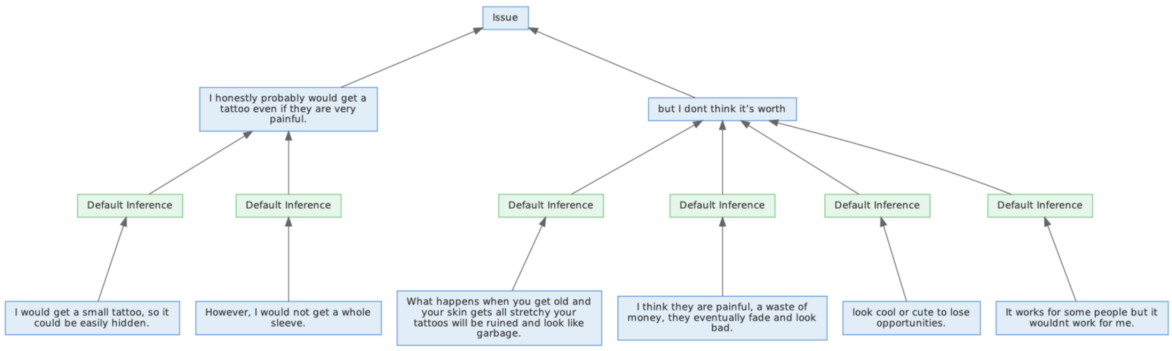

Figure 1: Sample argument subgraph construct from NYT news comments illustrating varying viewpoints. Claims "I honestly..." and "but I dont.." are entailed by premises, connected through `Default Inference` nodes, and opposing claims are connected through `Issue` nodes.

2018) sampled threads from Yahoo! Answers in which the best answer could be used as a reference summary. However, this heuristic is not guaranteed to cover all the user answers' perspectives, so we believe our dataset is a more principled benchmark for community question answering.

It is also noted that several large-scale MDS datasets have been introduced in the news domain (Fabbri et al., 2019; Gu et al., 2020; Gholipour Ghalandari et al., 2020), for creating Wikipedia lead-paragraphs (Liu et al., 2018), and for long-form question answering (Fan et al., 2019). However, these do not focus on the conversational domain.

4 Argument Graph Summarization

As our annotation protocol is motivated by the issues-viewpoints-assertions framework proposed in Barker and Gaizauskas (2016a), we propose to instantiate a modified version of that work's theoretical, proposed graph model.

Argument Graph Construction We build on the argument graph formulation of Lenz et al. (2020), a variant of Argument Interchange Format (Chesnevar et al., 2006). Claims and premises are represented as information nodes (I-nodes), with the relations between them represented as scheme nodes (S-nodes). Let $V = I \cup S$ be the set of nodes, and $E \subset V \times V$ the set of edges describing support relationships among the nodes. We then define the argument graph $G = (V, E)$.

Lenz et al. (2020) breaks the construction of the argument graph down into four steps: (1) *argument extraction*, or the identification of argumentative discourse units; (2) *relationship type classification*, or the classification of edges between nodes; (3) *major claim detection*; and (4) *graph construction*, or the construction of the final graph based on the identified nodes and edges. To adapt this formulation to our multi-document setting, we first perform *argument extraction* and *relationship type classification* for each individual input document and finally *graph construction* to determine relationships among claims from all documents.

Argument Extraction For extracting arguments from a single document, we build on work in argument mining with pretrained models (Chakrabarty et al., 2019). As in Lenz et al. (2020), our argumentative units are sentences, from which we identify *claims*, which are assertions that something is true, and *premises*, which are propositions from which a conclusion is drawn. Additionally, we identify and remove non-argumentative units. We train a three-way classifier for the task of argument extraction, following Chakrabarty et al. (2019) and making use of data for argument mining from that paper and from Stab and Gurevych (2014). The output of this step can also simply be used without further graph construction as a less noisy version of the input, which we call **-arg-filtered**.

Relationship Type Classification We follow the procedure in Lenz et al. (2020) and use entailment to determine the relationship between argumentative units within a document. However, rather than using the classifier provided, we make use of RoBERTa (Liu et al., 2019b) fine-tuned on the MNLI entailment dataset (Williams et al., 2018). Rather than using both support and contradiction edges between claims and premises, we make the simplification that all relationships can be captured with support edges, as we are dealing with a single document in this step. Within a single text, the

Dataset/Method	Lexrank	Textrank	BERT-ext
NYT	22.30/3.87/19.14	25.11/3.75/20.61	25.88/3.81/22.00
Reddit	22.71/4.52/19.38	24.38/4.54/19.84	24.51/4.18/20.95
Stack	26.30/5.62/22.27	25.43/4.40/20.58	26.84/4.63/22.85
Email	16.04/3.68/13.38	19.50/3.90/16.18	25.46/6.17/21.73

Table 4: ROUGE-1/2/L results for extractive LexRank (Erkan and Radev, 2004), TextRank (Mihalcea and Tarau, 2004), and BERT-based (Miller, 2019) models.

Data/Method	BART	BART-arg
NYT	35.91/9.22/31.28	**36.60/9.83/32.61**
Reddit	35.50/10.64/32.57	**36.39/11.38/33.57**
Stack	39.61/10.98/35.35	**39.73/11.17/35.52**
Email	**41.46/13.76/37.70**	40.32/12.97/36.90

Table 5: ROUGE-1/2/L results for vanilla BART as well as one trained on argument-mining input. Both are trained on 200 points from ConvoSumm.

premise can be tied as following from one of the claims. We create an edge between any premise and the claim it most entails if the entailment score from RoBERTa is greater than 0.33, based on manual analysis of the scores. If a premise is not labeled as supporting a claim, then we heuristically create an edge between that premise and the closest claim preceding it in the text.

Since not all texts in the benchmark datasets may be argumentative or may be too short to contain major claims, we use some heuristics in our graph creation. If none of the argumentative sentences are labeled as claims (i.e., all are labeled as premises) in argument extraction, the text's first sentence is labeled as the claim. Furthermore, we do not identify a single claim as the major claim since there may be multiple major points of discussion.

Graph Construction For the final graph, for each of the documents in an example, we run the above procedure and obtain a set of claims and associated premises. We then identify support edges between claims, which may be across documents. One claim may make a larger assertion, which is supported by other claims. We run our entailment model over all potential edges (in both directions) among claims in the document and greedily add edges according to the entailment support score while no cycles are made. After this step, we are left with a set of claims which do not entail any other nodes or, stated otherwise, do not have parent nodes. Following the terminology of Barker and Gaizauskas (2016b), these nodes can be considered viewpoints.

We then identify issues or topics on which the viewpoints differ. We run our entailment model for all parent claim nodes again in both directions over these claims and identify nodes that contradict each other with probability over 0.33, based on manual analysis of the resulting graphs. We greedily add edges to maintain a tree structure, joining these nodes to a special node, which we call the Issue node. All Issue nodes, as well as claims which are not connected to any Issue node, are connected to a dummy 'Conversation Node' which serves as the root of the argument graph. We show an example Issue subgraph for NYT data in Figure 1.

Argument Graphs to Summaries Recent work has shown the strength of text-based pretrained models on graph-to-text problems (Ribeiro et al., 2020). Following that work, we linearize the graph by following a depth-first approach starting from the Conversation Node. We found that inserting special tokens to signify edge types did not improve performance, likely due to the size of our data, and simply make use of an arrow $\rightarrow$ to signify the relationship between sentences. We train a sequence-to-sequence model on our linearized graph input, which we call **-arg-graph**.

5 Experimental Settings

We use the fairseq codebase (Ott et al., 2019) for our experiments. Our base abstractive text summarization model is BART-large (Lewis et al., 2020), a pretrained denoising autoencoder with 336M parameters that builds on the sequence-to-sequence transformer of Vaswani et al. (2017). We fine-tune BART using a polynomial decay learning rate scheduler with Adam optimizer (Kingma and Ba, 2015). We used a learning rate of 3e-5 and warmup and total updates of 20 and 200, following previous few-shot transfer work (Fabbri et al., 2020a). We could have equally fine-tuned other pretrained models such as Pegasus (Zhang et al., 2019) or T5 (Raffel et al., 2019), but Fabbri et al. (2020a) find that BART largely performs equally well in few-shot settings when compared to Pegasus.

For the NYT and Stack datasets, which contain sequences over the typical 1024 max encoder length with which BART is trained, we copied the encoder positional embeddings to allow sequences up to length 2048. To address the input-length of meeting summaries, which range from 6k to 12k tokens, we use the Longformer (Beltagy et al., 2020), which allows for sequences up to length 16k to-

Method/Dataset	AMI	ICSI
HMNet	53.02/18.57/-	**46.28**/10.60/-
DDA-GCN	53.15/**22.32**/-	-
Longformer-BART	54.20/20.72/51.36	43.03/**12.14**/40.26
Longformer-BART-arg	**54.47**/20.83/**51.74**	44.17/11.69/**41.33**

Table 6: ROUGE-1/2/L results for DDA-GCN (Feng et al., 2020) and HMNet (Zhu et al., 2020) on the AMI and ICSI meeting summarization dataset along with our Longformer and Longformer-arg models.

kens. We initialize the Longformer model with BART parameters trained on the CNN-DailyMail dataset, as the meeting summarization datasets contain fewer than 100 data points. We otherwise fine-tune models from vanilla BART, following intuition in few-shot summarization (Fabbri et al., 2020a) and based on initial experiments. In the tables which follow, "-arg" refers to any model trained with argument-mining-based input, and we specify which -arg-graph or -arg-filtered settings were used for each dataset below.

6 Results

We provide results for baseline, unsupervised extractive models in Table 4. Lexrank (Erkan and Radev, 2004) and Textrank (Mihalcea and Tarau, 2004), and BERT-ext (Miller, 2019), which makes use of BERT (Devlin et al., 2019). The unsupervised extractive models perform well below the extractive oracle performance, suggesting the difficulty of content selection in this setting.

We train BART on 200 examples from our validation set for abstractive models, using the remaining 50 as validation and test on the final test set of 250 examples. We tested zero-shot transfer from CNNDM and SAMSum in zero-shot settings, although these resulted in a much lower performance of about 28 ROUGE-1. Few-shot model performance is shown in Table 5. The abstractive model performs at or above the Extractive Oracle, suggesting the need for better abstractive models.

We also train on our argument mining-based approaches and show results in Table 5. We see ROUGE improvements when applying BART-arg-graph for Reddit, and Stack data. The -arg-filtered variation (which, as defined in Section 4, is the less noisy version of the input produced by the argument extraction step) outperformed the -arg-graph variation on both email and NYT data. For email data, however, this did not improve upon the BART baseline, likely due to the dataset's characteristics; email data is shorter and more linear, not benefiting from modeling the argument structure or removing non-argumentative units. We provide full results for both variations in the Appendix.

Dataset/Method	Our results	Previous SOTA
SAMSum	**52.27/27.82/47.92**	49.30/25.60/47.70
CQASUMM	**32.79/6.68/28.83**	31.00/5.00/15.20
BC3	39.59/13.98/21.20	-
ADS	37.18/11.42/21.27	-
SENSEI	34.57/7.08/16.80	-

Table 7: Benchmarking results on conversational datasets such as SAMSum (Gliwa et al., 2019b) and CQASUMM (Chowdhury and Chakraborty, 2018) and initial neural abstractive summarization results for email (BC3) (Ulrich et al., 2008), debate discussion forums (ADS) (Misra et al., 2015), and news comments (SENSEI) (Barker et al., 2016b).

Benchmarking Other Conversation Summarization Datasets We benchmark our models on widely used meeting summarization datasets. Due to the input's linear nature and the size of the meeting transcripts, we found improved results using -arg-filtered to filter non-argumentative units rather than incorporating the graph structure. Results are shown in Table 6. The Longformer model performs as well or better than previous state-of-the-art results on these datasets, despite not making use of more complex modeling structures, and we generally see improvement with argument-mining.

As noted above, there exist prior datasets for dialogue, community question answering, email, forum, and news comments summarization. We benchmark results on these datasets in Table 7. We outperform prior work on SAMSum (Gliwa et al., 2019b), and CQASUMM (Chowdhury and Chakraborty, 2018) with our BART and BART-arg-graph models, respectively. We did not find improvement on SAMSum with the BART-arg model due to the extremely short and focused nature of the dialogues, analogous to email data performance. We also provide transfer results of BART and BART-arg-graph models from our email and news-comment data to BC3 (Ulrich et al., 2009), ADS (Misra et al., 2015), and SENSEI data (Barker et al., 2016b), for which no prior neural abstractive summarization results existed.

Human Evaluations We collect human judgment annotations for two of the four quality dimensions studied in Kryscinski et al. (2019) and Fabbri et al. (2020b), namely consistency and relevance. Consistency is defined as the factual alignment be-

Target Dataset	BART		BART-arg	
	Relevance	Consistency	Relevance	Consistency
Reddit	3.39 (0.13)	3.40 (0.12)	3.47 (0.12)	3.41 (0.10)
AMI	4.07 (0.16)	3.67 (0.16)	4.13 (0.17)	3.70 (0.17)

Table 8: Mean relevance and factual consistency annotations for BART and BART-arg outputs on Reddit and AMI. Standard errors are reported in parentheses.

tween the summary and the summarized source text, while relevance is defined as the summary's ability to select important content; only relevant information and viewpoints should be included. We did not include fluency as an initial inspection of the data found fluency to be of very high quality, as has shown to be the case for pretrained models in news summarization (Fabbri et al., 2020b). We did not include coherence as this was generally not an issue of concern in the initial analysis.

We randomly select 25 random examples from the Reddit corpus and ten examples from the AMI corpus, and output from the BART and BART-arg-graph models. These data points were chosen to demonstrate what characteristics are realized in differences across ROUGE for argument-graph and argument-noise-reduction approaches. Ten examples were chosen from AMI due to the size of the input and annotation constraints. The annotator sees the source article and randomly-ordered output from the model and then rates the summaries for relevance and consistency on a Likert from 1 to 5, with 5 being the best score. We averaged the score of three native English-speaking annotators on each example and then across examples. Results are shown in Table 8. We find that the annotators prefer our argument mining-based approaches in both dimensions. However, the results are close. Furthermore, the scores for relevance and consistency are rather low, especially on the Reddit dataset and when compared to results on the CNN-DailyMail Dataset from Fabbri et al. (2020b). These results demonstrate the difficulty of modeling such conversational data. Examples are included in the appendix.

7 Conclusion

We propose ConvoSumm, a benchmark of four new, crowdsourced conversation datasets and state-of-the-art baselines on widely-used datasets that promote more unified progress in summarization beyond the news domain. Our benchmark consists of high-quality, human-written summaries that call for abstractive summaries and a deeper understanding of the input texts' structure. We provide results for baseline models and propose to model the text's argument structure, showing that such structure helps better quantify viewpoints in non-linear input in both automatic and human evaluations. Our analysis notes challenges in modeling relevance and consistency in abstractive conversation summarization when compared to news summarization.

8 Ethical Considerations

As we propose novel conversation summarization datasets and modeling components, this section is divided into the following two parts.

8.1 New Dataset

Intellectual Properties and Privacy Rights All data for our newly-introduced datasets are available online; please see the following for New York Times comment data[2], StackExchange data[3], and W3C email data[4]. Reddit data is available via the Google BigQuery tool[5].

Compensation for Annotators We compensated the Turkers approximately $12–$15 per hour. We first annotated examples in-house to determine the required annotation speed. Typically, the summarization task took around 10 minutes, and we compensated the workers from $2.25 to $3.00 per task, depending on the domain and deadline requirements.

Steps Taken to Avoid Potential Problems We interacted closely with the Turkers to ensure that compensation was fair and that the instructions were clear. To maintain the quality of the dataset, we manually reviewed the crowdsourced summaries for language use. Initial investigation into Reddit data showed certain inappropriate language usage, so we filtered these examples automatically.

8.2 NLP Application

Bias Biases may exist in the datasets, such as political bias in the news datasets and gender bias in potentially all of the datasets. Thus, models trained on these datasets may propagate these biases. We

[2] https://www.kaggle.com/aashita/nyt-comments

[3] https://archive.org/download/stackexchange

[4] https://tides.umiacs.umd.edu/webtrec/trecent/parsed_w3c_corpus.html

[5] https://console.cloud.google.com/bigquery

removed data with offensive language when possible.

Misuse Potential and Failure Mode When used as intended, applying the summarization models described in this paper can save people much time. However, the current models are still prone to producing hallucinated summaries, and in such a case, they may contribute to misinformation on the internet. Further research is needed to ensure the faithfulness of abstractive summaries to address this issue, as this issue is present among all current abstractive summarization models.

Environmental Cost The experiments described in the paper make use of V100 GPUs. We used up to 8 GPUs per experiment (depending on the experiment; sometimes, a single GPU was used to run the maximum number of experiments in parallel). The experiments may take up to a couple of hours for the larger datasets. Several dozen experiments were run due to parameter search, and future work should experiment with distilled models for more light-weight training. We note that while our work required extensive experiments to draw sound conclusions, future work will be able to draw on these insights and need not run as many large-scale comparisons. Models in production may be trained once for use using the most promising settings.

References

Emma Barker and Robert Gaizauskas. 2016a. Summarizing multi-party argumentative conversations in reader comment on news. In *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, pages 12–20, Berlin, Germany. Association for Computational Linguistics.

Emma Barker and Robert Gaizauskas. 2016b. Summarizing multi-party argumentative conversations in reader comment on news. In *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, pages 12–20, Berlin, Germany. Association for Computational Linguistics.

Emma Barker, Monica Lestari Paramita, Ahmet Aker, Emina Kurtic, Mark Hepple, and Robert Gaizauskas. 2016a. The SENSEI annotated corpus: Human summaries of reader comment conversations in on-line news. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 42–52, Los Angeles. Association for Computational Linguistics.

Emma Barker, Monica Lestari Paramita, Ahmet Aker, Emina Kurtic, Mark Hepple, and Robert Gaizauskas. 2016b. The SENSEI annotated corpus: Human summaries of reader comment conversations in on-line news. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 42–52, Los Angeles. Association for Computational Linguistics.

Christos Baziotis, Ion Androutsopoulos, Ioannis Konstas, and Alexandros Potamianos. 2019. SEQˆ3: Differentiable sequence-to-sequence-to-sequence autoencoder for unsupervised abstractive sentence compression. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 673–681, Minneapolis, Minnesota. Association for Computational Linguistics.

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

Arthur Bražinskas, Mirella Lapata, and Ivan Titov. 2020. Unsupervised opinion summarization as copycat-review generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5151–5169, Online. Association for Computational Linguistics.

Tuhin Chakrabarty, Christopher Hidey, Smaranda Muresan, Kathy McKeown, and Alyssa Hwang. 2019. AMPERSAND: Argument mining for PERSuAsive oNline discussions. 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 2933–2943, Hong Kong, China. Association for Computational Linguistics.

Jiaao Chen and Diyi Yang. 2020. Multi-view sequence-to-sequence models with conversational structure for abstractive dialogue summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4106–4118, Online. Association for Computational Linguistics.

Carlos Chesnevar, Sanjay Modgil, Iyad Rahwan, Chris Reed, Guillermo Simari, Matthew South, Gerard Vreeswijk, Steven Willmott, et al. 2006. Towards an argument interchange format. *The knowledge engineering review*, 21(4):293–316.

Tanya Chowdhury and Tanmoy Chakraborty. 2018. Cqasumm: Building references for community question answering summarization corpora.

Eric Chu and Peter J. Liu. 2019. Meansum: A neural model for unsupervised multi-document abstractive summarization. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 1223–1232. PMLR.

Nick Craswell, Arjen P de Vries, and Ian Soboroff. 2005. Overview of the trec 2005 enterprise track. In *TREC*, volume 5, pages 199–205.

Yang Deng, Wai Lam, Yuexiang Xie, Daoyuan Chen, Yaliang Li, Min Yang, and Ying Shen. 2020. Joint learning of answer selection and answer summary generation in community question answering. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 7651–7658. AAAI Press.

Shrey Desai, Jiacheng Xu, and Greg Durrett. 2020. Compressive summarization with plausibility and salience modeling. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6259–6274, Online. Association for Computational Linguistics.

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

Alvin Dey, Tanya Chowdhury, Yash Kumar, and Tanmoy Chakraborty. 2020a. Corpora evaluation and system bias detection in multi-document summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2830–2840, Online. Association for Computational Linguistics.

Alvin Dey, Tanya Chowdhury, Yash Kumar, and Tanmoy Chakraborty. 2020b. Corpora evaluation and system bias detection in multi document summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 2830–2840, Online. Association for Computational Linguistics.

Günes Erkan and Dragomir R Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of artificial intelligence research*, 22:457–479.

Alexander Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir Radev. 2019. Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1074–1084, Florence, Italy. Association for Computational Linguistics.

Alexander R Fabbri, Simeng Han, Haoyuan Li, Haoran Li, Marjan Ghazvininejad, Shafiq Joty, Dragomir Radev, and Yashar Mehdad. 2020a. Improving zero and few-shot abstractive summarization with intermediate fine-tuning and data augmentation. *arXiv preprint arXiv:2010.12836*.

Alexander R Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2020b. Summeval: Re-evaluating summarization evaluation. *arXiv preprint arXiv:2007.12626*.

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. ELI5: Long form question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3558–3567, Florence, Italy. Association for Computational Linguistics.

Xiachong Feng, Xiaocheng Feng, Bing Qin, Xinwei Geng, and Ting Liu. 2020. Dialogue discourse-aware graph convolutional networks for abstractive meeting summarization. *arXiv preprint arXiv:2012.03502*.

Cecilia E Ford. 1991. Linguistics: The cambridge survey: Volume 4. language: The socio-cultural context. frederick h. newmeyer (ed.). *Studies in Second Language Acquisition*, 13(3):412–413.

Michel Galley. 2006. A skip-chain conditional random field for ranking meeting utterances by importance. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 364–372, Sydney, Australia. Association for Computational Linguistics.

Prakhar Ganesh and Saket Dingliwal. 2019. Abstractive summarization of spoken and written conversation. *CoRR*, abs/1902.01615.

Demian Gholipour Ghalandari, Chris Hokamp, Nghia The Pham, John Glover, and Georgiana Ifrim. 2020. A large-scale multi-document summarization dataset from the Wikipedia current events portal. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1302–1308, Online. Association for Computational Linguistics.

Dan Gillick and Yang Liu. 2010. Non-expert evaluation of summarization systems is risky. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 148–151, Los Angeles. Association for Computational Linguistics.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019a. SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 70–79, Hong Kong, China. Association for Computational Linguistics.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019b. SAMSum corpus: A human-annotated dialogue dataset for abstractive

summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 70–79, Hong Kong, China. Association for Computational Linguistics.

Chih-Wen Goo and Yun-Nung Chen. 2018a. Abstractive dialogue summarization with sentence-gated modeling optimized by dialogue acts. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 735–742. IEEE.

Chih-Wen Goo and Yun-Nung Chen. 2018b. Abstractive dialogue summarization with sentence-gated modeling optimized by dialogue acts. *CoRR*, abs/1809.05715.

Xiaotao Gu, Yuning Mao, Jiawei Han, Jialu Liu, You Wu, Cong Yu, Daniel Finnie, Hongkun Yu, Jiaqi Zhai, and Nicholas Zukoski. 2020. Generating representative headlines for news stories. In *WWW '20: The Web Conference 2020, Taipei, Taiwan, April 20-24, 2020*, pages 1773–1784. ACM / IW3C2.

Doris Hoogeveen, Karin M Verspoor, and Timothy Baldwin. 2015. Cqadupstack: A benchmark data set for community question-answering research. In *Proceedings of the 20th Australasian document computing symposium*, pages 1–8.

Xinyu Hua and Lu Wang. 2017. A pilot study of domain adaptation effect for neural abstractive summarization. In *Proceedings of the Workshop on New Frontiers in Summarization*, pages 100–106, Copenhagen, Denmark. Association for Computational Linguistics.

Adam Janin, Don Baron, Jane Edwards, Dan Ellis, David Gelbart, Nelson Morgan, Barbara Peskin, Thilo Pfau, Elizabeth Shriberg, Andreas Stolcke, et al. 2003. The icsi meeting corpus. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).*, volume 1, pages I–I. IEEE.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Varada Kolhatkar and Maite Taboada. 2017. Using New York Times picks to identify constructive comments. In *Proceedings of the 2017 EMNLP Workshop: Natural Language Processing meets Journalism*, pages 100–105, Copenhagen, Denmark. Association for Computational Linguistics.

Wessel Kraaij, Thomas Hain, Mike Lincoln, and Wilfried Post. 2005. The ami meeting corpus.

Wojciech Kryscinski, Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Neural text summarization: A critical evaluation. 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 540–551, Hong Kong, China. Association for Computational Linguistics.

Philippe Laban, Andrew Hsi, John Canny, and Marti A. Hearst. 2020. The summary loop: Learning to write abstractive summaries without examples. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5135–5150, Online. Association for Computational Linguistics.

Mirko Lenz, Premtim Sahitaj, Sean Kallenberg, Christopher Coors, Lorik Dumani, Ralf Schenkel, and Ralph Bergmann. 2020. Towards an argument mining pipeline transforming texts to argument graphs. *arXiv preprint arXiv:2006.04562*.

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

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

Chunyi Liu, Peng Wang, Jiang Xu, Zang Li, and Jieping Ye. 2019a. Automatic dialogue summary generation for customer service. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2019, Anchorage, AK, USA, August 4-8, 2019*, pages 1957–1965. ACM.

Peter J. Liu, Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer. 2018. Generating wikipedia by summarizing long sequences. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Yinhan Liu, 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*.

Zhengyuan Liu, Angela Ng, Sheldon Lee Shao Guang, Ai Ti Aw, and Nancy F. Chen. 2019c. Topic-aware pointer-generator networks for summarizing spoken conversations. *CoRR*, abs/1910.01335.

Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing order into text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 404–411, Barcelona, Spain. Association for Computational Linguistics.

Derek Miller. 2019. Leveraging bert for extractive text summarization on lectures. *arXiv preprint arXiv:1906.04165*.

Amita Misra, Pranav Anand, Jean E. Fox Tree, and Marilyn Walker. 2015. Using summarization to discover argument facets in online idealogical dialog. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 430–440, Denver, Colorado. Association for Computational Linguistics.

Amita Misra, Pranav Anand, Jean E Fox Tree, and Marilyn Walker. 2017. Using summarization to discover argument facets in online ideological dialog. *arXiv preprint arXiv:1709.00662*.

Ramesh Nallapati, 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, Berlin, Germany. Association for Computational Linguistics.

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

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Tatsuro Oya, Yashar Mehdad, Giuseppe Carenini, and Raymond Ng. 2014. A template-based abstractive meeting summarization: Leveraging summary and source text relationships. In *Proceedings of the 8th International Natural Language Generation Conference (INLG)*, pages 45–53, Philadelphia, Pennsylvania, U.S.A. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

Leonardo FR Ribeiro, Martin Schmitt, Hinrich Schütze, and Iryna Gurevych. 2020. Investigating pretrained language models for graph-to-text generation. *arXiv preprint arXiv:2007.08426*.

Evan Sandhaus. 2008. The new york times annotated corpus. *Linguistic Data Consortium, Philadelphia*, 6(12):e26752.

Shasha Xie, Yang Liu, and Hui Lin. 2008. Evaluating the effectiveness of features and sampling in extractive meeting summarization. In *2008 IEEE Spoken Language Technology Workshop*, pages 157–160.

Christian Stab and Iryna Gurevych. 2014. Identifying argumentative discourse structures in persuasive essays. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 46–56, Doha, Qatar. Association for Computational Linguistics.

Manfred Stede, Stergos Afantenos, Andreas Peldszus, Nicholas Asher, and Jérémy Perret. 2016. Parallel discourse annotations on a corpus of short texts. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1051–1058, Portorož, Slovenia. European Language Resources Association (ELRA).

Mattia Tomasoni and Minlie Huang. 2010. Metadata-aware measures for answer summarization in community question answering. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 760–769, Uppsala, Sweden. Association for Computational Linguistics.

J. Ulrich, G. Murray, and G. Carenini. 2008. A publicly available annotated corpus for supervised email summarization. In *AAAI08 EMAIL Workshop*, Chicago, USA. AAAI.

Jan Ulrich, Giuseppe Carenini, Gabriel Murray, and Raymond Ng. 2009. Regression-based summarization of email conversations. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 3.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Danqing Wang, Pengfei Liu, Ming Zhong, Jie Fu, Xipeng Qiu, and Xuanjing Huang. 2019. Exploring domain shift in extractive text summarization.

Lu Wang and Claire Cardie. 2013. Domain-independent abstract generation for focused meeting summarization. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1395–1405, Sofia, Bulgaria. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans,

Louisiana. Association for Computational Linguistics.

Lin Yuan and Zhou Yu. 2019. Abstractive dialog summarization with semantic scaffolds. *arXiv preprint arXiv:1910.00825*.

Amy Zhang, Bryan Culbertson, and Praveen Paritosh. 2017. Characterizing online discussion using coarse discourse sequences.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2019. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization.

Jiawei Zhou and Alexander Rush. 2019. Simple unsupervised summarization by contextual matching. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5101–5106, Florence, Italy. Association for Computational Linguistics.

Chenguang Zhu, Ruochen Xu, Michael Zeng, and Xuedong Huang. 2020. A hierarchical network for abstractive meeting summarization with cross-domain pretraining. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 194–203, Online. Association for Computational Linguistics.

A Full Results

We present the results of BART and -arg variations on our four crowdsourced datasets in Table 9.

Data/Method	BART	BART-arg-graph	BART-arg-filtered
NYT	35.91/9.22/31.28	36.02/9.60/32.34	**36.60/9.83/32.61**
Reddit	35.50/10.64/32.57	36.39/**11.38/33.57**	**36.51**/11.02/33.14
Stack	39.61/10.98/35.35	**39.73/11.17/35.52**	39.40/10.98/35.51
Email	**41.46/13.76/ 37.70**	39.05/12.14/35.99	40.32/12.97/36.90

Table 9: Full ROUGE-1/2/L results for vanilla BART, -arg-graph, and -arg-filtered input. All are trained on 200 points from ConvoSumm.

Title: What should I draw next? **Subreddit**: pokemon
A FB user said they wanted to see a Zapdos, but depending on what people say on here, if any, I may do one that you all choose! So help out if you want! Or don't. :)
Source Documents: 1. I'd like to see any of the Legendary Beasts, Raikou, Suicune, Entei. 2. Ooo I'm kinda thinking of Suicune now. Good choice! I'll consider it for sure! If not, it will definitely be very soon! 3. Pancham ← Oooo good choice. That'll be one of the next ones too. 4. If you have the time, I'd love to see Reshiram! ← I'll definitely do that as one of the next ones! Thank you very much! 5. Seel? ← With your username, I may just do that soon. ← Cool! 6. MAGIKARP ← Dear Lord. 7. Aron lairon! 8. You should girafarig because it's the best pokemon ever created no matter what ANYONE SAYS. ← While I admire your steadfast courage in how awesome Girafig is, I may wait on that one. :P BUT. Maybe. :)
BART: Most commenters say that they would like to see a Zapdos or Suicune. A few other suggestions are Girafig, Arron, lairon, Magikarp, Raikou, Entei, Reshiram, Seel, and Pancham. One commenter says that they want to see Girafarig, but another says that he may wait on that one for a while.
BART-arg: Most commenters say that they would like to see a Zapdos, Raikou, Suicune, Entei, Pancham Aron, and lairon. A few commenters say they would love to see Girafarig. One commenter says that he would like Girafig because it's the best pokemon ever created no matter what anyone says about it. Another commenter says they want to see Reshiram.
Gold Summary: Commenters suggest ideas of the next Pokemon for the original poster to draw next. Each commenter has a different suggestion including Raikou, Reshiram, Seel, Magikarp and Lairon, and the original poster comments their thoughts on each of the suggestions with either approval or a joking disapproval.

Table 10: Example source documents and summaries from Reddit data in which the models are faithful to the source and largely relevant.

B Sample Output

We provide examples of model outputs to offer more insight into the datasets and models. An example of Reddit input and outputs for which the models remain faithful to the source is found in Table 10. The gold summary balances being a meta-analysis of the input documents with providing sufficient details. We provide an additional example of outputs that struggle with consistency and relevance in Table 11. In the BART output, the model mistakes the suggestion in the input to pay debt before starting a business. In BART-arg, the model incorrectly determines relevance, as the suggestion that one should invest in pumpkins was sarcastic and not emphasized in the input. This output points to a need to better model interactions and salience in the conversation data.

C Additional Details

For reproducibility purposes, we provide (smallest NLL loss, largest NLL loss) bounds on the validation loss from non-baseline models encountered when training on our newly-proposed datasets. NYT (5.77, 5.83); Reddit: (5.62, 5.69); Stack (5.60, 5.63); Email: (5.40, 5.45).

Title: Complete noob here. Any advice welcome! **Subreddit**: personalfinance
I've always been responsible when it comes to my money. In the sense that I dont spend what I dont have, I save my money, and I watch what I spend. But now I've recently graduated (with 28k in debt) and want to do better. I've got, what I think, is a decent plan to pay off my debt. ... But what else should I be doing? ...
Source Documents:
1. Please read the information found in the [FAQ](http://www.reddit.com/r/personalfinance/wiki/faq), particularly ... 2. Make sure any extra money is going towards the principal. 3. If it makes you feel better there was never a class where this was explained. ... The best way to learn is to check out the FAQs on the sidebar ... 4. ... You should also save up some cash in a bank account and consider that your emergency fund,... 5. The two common retirement accounts are 401ks and IRAs 6. Invest in pumpkins.
BART: Most commenters agree that the best way to learn about investing is to read the FAQ. A few commenters say that if you have debt, then you should start making contributions to a 401k retirement plan. One commenter says that you need to start a small business to pay off your debt, and another says that it would be a good idea to have some cash in a bank account. One commentator says that they think the best thing to do is start a business.
BART-arg: Most commenters agree that the best way to start a small business is to pay off your debt. Some commenters say that you should also invest in pumpkins. One commenter says that if you don't have a 401k, then you should start making contributions to a Roth retirement plan. Another commenter suggests that you invest in a bank account to build up an emergency fund. One commentator says that you need to be responsible when it comes to your money.
Gold Summary: Commenters offer their advice for settling a debt. Several commenters agree that utilizing the avalanche model is the correct way to pay the debt. Another also suggests having an emergency fund for any unexpected bills or necessary expenses. In understanding finance, commenters stress the need to check the sidebar for FAQs.

Table 11: Example source documents and summaries from Reddit data in which the models fail to completely capture salience while remaining faithful to the input.