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

Alexander R. Fabbri[1], **Faiaz Rahman[1]**, Imad Rizvi[1], Borui Wang[1],
Haoran Li[2], Yashar Mehdad[2], Dragomir Radev[1]

[1]Department of Computer Science, Yale University [2]Facebook AI

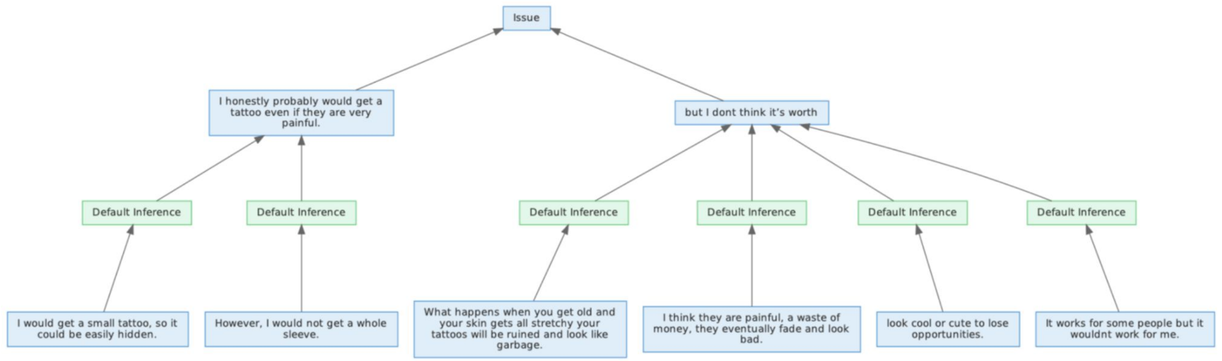
LILY Lab

Introduction

Automatic text summarization has seen recent progress due to the release of large-scale datasets such as the CNN–Daily Mail dataset (Nallapati et al., 2016) and the use of large self-supervised pretrained models such as BART (Lewis et al., 2020) and Pegasus (Zhang et al., 2019). However, less work has focused on the abstractive summarization of online conversations, particularly beyond the domain of news articles. This research gap is due, in part, to the lack of standardized datasets for summarizing online discussions. To address this gap, we introduce **ConvoSumm**, a suite of four datasets that can evaluate a model's performance on a broad spectrum of conversation dialogue data, specifically in the subdomains 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, to unify modeling across these conversational domains, we incorporate recent work in **end-to-end argument mining** to instantiate the Barker and Gaizauskas (2016a) graph framework for conversation summarization, using an argument graph construction method involving entailment relations, graph linearization, and graph-to-text training. We apply such argument mining to model the structure of our conversation data better as well as reduce noise in long-text input, showing improved results in both automatic and human evaluations.

Data Selection

For the news comments subdomain, we use the NYT Comments dataset. For discussion forums and debate, we select Reddit data from CoarseDiscourse (Zhang et al., 2017), which contains discourse structure annotations. For community question answering, we use StackExchange, which has been used in answer relevance modeling and question deduplication (Hoogeveen et al., 2015). For emails, we use the W3C corpus (Craswell et al., 2005).



Quality-Controlled Crowdsourcing

Data Filtering and Context Generation We filtered out examples based on conversation length, maximum number of tokens in an individual post, total number of tokens across the conversation. For NYT, rather than use the entire articles, we generated extractive BERT–based summaries (Miller, 2019) to provide context for annotators.

Annotation Protocol We designed annotation instructions for crowdsourced workers to write abstractive summaries for each of the four datasets. We present the source threads (or BERT-based summaries in the case of NYT) with the additional metadata, and had a summary protocol based on the issues–viewpoints–assertions framework of Barker and Gaizauskas (2016b) and abstractive properties.

Quality Control Our generated summaries were crowd-sourced using a screened set of Amazon Mechanical Turkers and then reviewed manually internally for rewrites.

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

Table 1: Statistics across datasets in ConvoSumm, including novel uni/bi/trigrams and ROUGE-1/2/L Extractive Oracle scores.

Argument Graph Modeling

Argument Graph Formulation We build on the argument graph formulation of Lenz et al. (2020), with claims and premises represented as information I-nodes, and their relations represented as scheme S-nodes. Let $V = I \cup S$ such that $E \subset V \times V$ is the set of edges describing support relationships. We initialize the argument graph $G = (V, E)$.

Argument Extraction We train a three-way classifier for the extraction of claims, premises, and non-argumentative units. This creates a less noisy version of the input, **-arg-filtered**.

Relation Type Classification We use entailment to determine the relationships between argumentative units with RoBERTa fine-tuned on the MNLI entailment dataset. We create an edge between any premise and the claim it most entails if the entailment score from RoBERTa is greater than our threshold of 0.33.

Graph Construction For each of the documents in an example, we greedily add edges according to the entailment support score, leaving nodes which do not entail any other nodes; these are considered viewpoints. We then identify differing viewpoints by greedily joining edges into Issue nodes, which in turn point to the Conversation root node.

Graph Linearization The graph is linearized depth-first from the Conversation node; seq2seq models are then trained on our linearized graph input, **-arg-graph**.

Results

We provide results for baseline, unsupervised extractive models in Table 2. We train BART on 200 examples from our validation set and test with few-shot, and also train on our argument mining–based approaches, which benefited all data subdomains except email, which was likely due to its linear structure not benefiting from argument structure modeling inherent in conversations with shorter, more frequent utterances; both are shown in Table 3. Additionally, we benchmark results on prior datasets for dialgoue, community question answering, email, forum, and news comments summarization, shown in Table 4.

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 2: ROUGE-1/2/L results for extractive models.

Dataset/Method	BART	BART-arg
NYT	35.91/9.22/31.28	**36.02/9.60/32.34**
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**	40.91/13.67/37.10

Table 3: ROUGE-1/2/L results for vanilla and argument mining BART trained on 200 ConvoSumm points.

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 4: Benchmarking results for five subdomains.

Conclusion

We introduce ConvoSumm, a benchmark of four new, crowdsourced conversation datasets and SOTA baselines on widely-used datasets that promote more unified progress in conversation summarization. We also apply argument graph modeling, showing that such structure helps better quantify viewpoints in nonlinear input.