MEDIASUM: A Large-scale Media Interview Dataset for Dialogue Summarization

Chenguang Zhu*, **Yang Liu***, **Jie Mei, Michael Zeng**
Microsoft Cognitive Services Research Group
{chezhu,yaliu10,jimei,nzeng}@microsoft.com

Abstract

This paper introduces MEDIASUM[1], a large-scale media interview dataset consisting of 463.6K transcripts with abstractive summaries. To create this dataset, we collect interview transcripts from NPR and CNN and employ the overview and topic descriptions as summaries. Compared with existing public corpora for dialogue summarization, our dataset is an order of magnitude larger and contains complex multi-party conversations from multiple domains. We conduct statistical analysis to demonstrate the unique positional bias exhibited in the transcripts of televised and radioed interviews. We also show that MEDIASUM can be used in transfer learning to improve a model's performance on other dialogue summarization tasks.

1 Introduction

Dialogue summarization can provide a succinct synopsis for conversations between two or more participants, based on human-transcribed or machine-generated transcripts. Dialogue summaries are useful for participants to recap salient information in the talk and for absentees to grasp the key points. As a result, several models have been recently proposed to summarize daily conversations (Gliwa et al., 2019; Chen and Yang, 2020), meeting transcripts (Zhu et al., 2020) and customer support conversations (Liu et al., 2019).

However, compared with the abundance of text summarization datasets, there are very few public datasets for dialogue summarization. And existing datasets are limited to their small sizes. For example, the benchmark datasets for meeting summarization, AMI (McCowan et al., 2005) and ICSI (Janin et al., 2003), only contain transcripts and abstractive summaries for 137 and 59 business meetings, respectively. While recently some larger dialogue summarization datasets have been proposed, they are either built from a narrow domain, e.g. the CRD3 dataset (Rameshkumar and Bailey) which is built from conversations in a live-streamed show for the Dungeons and Dragons game, or not publicized due to privacy reasons, e.g. the Didi dataset (Liu et al., 2019) from customer service conversations. This lack of large-scale dialogue summarization datasets is due to a higher labeling cost compared with news articles and privacy issues with many real daily dialogues and business meetings.

On the other hand, media interview transcripts and the associated summaries/topics can be a valuable source for dialogue summarization. In a broadcast interview, the host discusses various topics with one or more guests. As many interviews proceed with pre-defined topics, the accompanying summaries are of a relatively high quality. Also, the wide variety of topics, different backgrounds of speakers, and the colloquial form of chat make these interviews very close to daily conversations and business meetings.

Therefore, we collect public interview transcripts and the associated summaries/topics from NPR and CNN to build a large-scale dialogue summarization dataset, MEDIASUM.

In NPR, each transcript comes with an overview of the interview, which is used as the summary in our dataset. We leverage the INTERVIEW dataset (Majumder et al., 2020) to get transcripts and crawl the associated descriptions. We end up with 49.4K NPR transcripts with summaries.

We then collect 269.4K CNN interview transcripts from 2000 to 2020, each with a list of topic descriptions. As many CNN interviews contain multiple topics, we conduct segmentation at the boundary of commercial breaks to assign each topic to the most relevant interview segment via lexical matching. In this way, we not only obtain tran-

* Equal contribution
[1] https://github.com/zcgzcgzcg1/MediaSum/

scripts with a more concentrated topic but also enlarge the total number of instances. We end up with 414.2K CNN transcript segments with topic descriptions as summaries. Thus, in total, our MEDIASUM dataset contains 463.6K transcripts with summaries.

We show that compared to existing public dialogue summarization datasets, MEDIASUM contains more speakers, longer conversation and is an order of magnitude larger. Also, we demonstrate the unique positional bias in interview dialogues: while a televised interview often mentions keywords in the summary at the beginning of the program, a radio interview usually mentions these keywords at both the beginning and the end of the program.

In experiments, we evaluate several benchmark summarization models on our dataset. We then show that after fine-tuning on MEDIASUM, models' performance can be improved on other dialogue summarization tasks like AMI, ICSI and SAMSum, demonstrating the transfer learning capability of our dataset.

2 Related Work

Due to the success of corpus-based methods, the past decade saw the emergence of many dialogue datasets on various domains (Budzianowski et al., 2018; Lowe et al., 2015). However, very few of these datasets contain corresponding summary text. As human dialogues have very different structures and language patterns from written articles, dialogue summarization models can only limitedly benefit from the largely available news summarization data (Zhu et al., 2020).

Current public datasets for dialogue summarization are either very small or in a specific domain. AMI (McCowan et al., 2005) and ICSI (Janin et al., 2003) contain 137 and 59 meeting transcripts with abstractive summaries. AMI meetings are recorded in an artificial environment with actors and ICSI contains meetings of a speech group. MultiWOZ (Budzianowski et al., 2018) is a multi-domain task-oriented dialogue dataset where the instructions have been used as summaries (Yuan and Yu, 2019). All dialogues are conducted between one user and one agent on the topic of booking and inquiry. SAMSum (Gliwa et al., 2019) hires linguists to write messenger-like daily conversations. Although the dialogues are open-domain, they are not from real human conversations. CRD3 (Rameshkumar and Bailey) contains 159 episodes from the Critical Role show with transcribed conversations between Dungeons and Dragon players. Additionally, there are non-public dialogue summarization datasets in th domains of customer support (Liu et al., 2019) and medical conversation (Krishna et al., 2020).

3 Media Interview Dataset: MEDIASUM

3.1 Data collection

We first collect interview transcriptions from National Public Radio (NPR, `www.npr.org`). The INTERVIEW dataset (Majumder et al., 2020) contains 105K transcripts from NPR but does not include interview summaries or the link to the transcript page. We find a majority of NPR interviews come with an overview description before the transcription text, which can be used as summaries. Thus, for each interview in the INTERVIEW dataset, we use the NPR searching service to get the link to the corresponding page and extract the description text if it exists. We filter out descriptions with more than 200 words and collect 49.4K transcripts with summaries.

The CNN transcription service provides transcripts of televised interviews and a list of discussed topics, which can be used as summaries (`transcripts.cnn.com`). We crawl CNN transcripts from 2014 to 2020, combined with the data from 2000 to 2014 (Sood, 2017), and end up with 269.4K transcripts with summaries.

Transcript segmentation for topic match. Interviews with multiple topics are often long, and the mixing of multiple topics makes it hard for models to generate accurate summaries. Among the collected CNN interviews, 157.9K transcripts, or 58.6%, have more than one topic. Thus, we try to partition multi-topic interviews into segments and match each topic to a segment. We find that the televised CNN interviews often contain several commercial breaks marked in the transcript. These ads usually come in between topics. Therefore, we partition the transcript at the boundaries of commercial breaks. Then, we assign each topic to the segment containing the most (at least one) non-stop words in the topic. We do not count the last 50 words in a segment where the host often reminds watchers of the next topic after the commercial break. Among the 157.9K multi-topic interviews, 330.4K segments are associated with at least one topic. To make sure that the summary contains enough information, we filter out summaries

Statistics	NPR	CNN
Dialogues	49,420	414,176
Avg. words in dialogue	906.3	1,630.9
Avg. words in summary	40.2	11.3
Turns	24.2	30.7
Speakers	4.0	6.8
Novel summary words	33.6%	24.9%

Table 1: Data statistics of NPR and CNN transcripts and summaries.

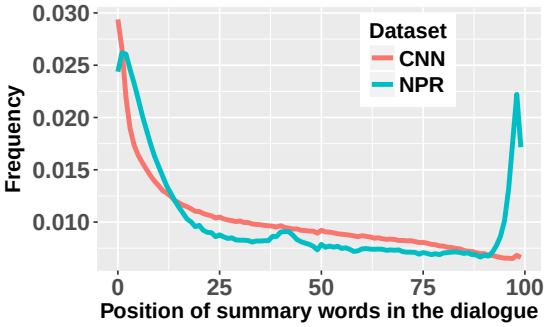

Figure 1: The frequency of the non-stop summary words appearing at different positions of the dialogue. The positions are normalized to [0, 100].

with fewer than 5 words. In the end, we construct 414.2K CNN interview transcripts with summaries.

As transcripts from the NPR and CNN are from similar domains, we combine them into a unified summarization dataset, MEDIASUM, containing 463.6K pairs of transcripts and summaries. As far as we know, this is the largest public open-domain dialogue summarization dataset. We show an example dialogue with its summary in Table 5.

Here, we note that the summary styles of NPR and CNN are different. Table 1 shows that although the dialogue length and number of speakers are similar in NPR and CNN, the summaries from NPR are much longer and more abstractive, indicated by a higher ratio of novel words in summary that do not appear in the dialogue.

3.2 Data statistics

In this section, we investigate different aspects of the MEDIASUM dataset via statistics.

We leverage the Latent Dirichlet Allocation (Blei et al., 2003) tool in scikit-learn package (Pedregosa et al., 2011) to analyze the main dialogue topics. We manually name the topic clusters based on the returned top 10 words in each cluster. The top 5 topics are politics (26.3%), international news (13.3%), crime (12.7%), economy (12.5%) and US news (11.7%).

The dialogues in MEDIASUM have on average 30.0 turns, 6.5 speakers and 1,553.7 words, and the summaries have on average 14.4 words. This shows that most dialogues in our dataset are multi-party conversations of medium to long lengths.

Table 2 compares MEDIASUM with other public dialogue summarization datasets. As shown, MEDIASUM contains much longer dialogues and more speakers than MultiWOZ 2.0 and SAMSum. This makes it suitable for training models targeted for multi-party dialogue or meeting summarization. Also, while AMI, ICSI and MultiWOZ 2.0 contain dialogues either from limited domains or under artificial context, MEDIASUM is a much larger dataset containing radioed and televised interview transcripts covering much broader topics.

3.3 Positional Bias

It has been found that in many news articles, the most important information is often shown at the beginning, i.e. the inverted pyramid structure (Kedzie et al., 2018). In this section, we investigate whether a similar positional bias is present in multi-party dialogues.

We record the position of each non-stop word in the transcript that also appears in the summary. To normalize, we partition each transcript into 100 equal-length bins and count the frequency that summary words appear in each bin. As shown in Fig. 1, similar to news articles, the beginning of transcripts from both CNN and NPR contain more summary words on average. However, different from televised CNN interviews, NPR programs also contain many summary words near the end. To make sure that the trend in CNN is not caused by topic segmentation, we compute the frequency for original single-topic CNN transcripts and find that the trend is very similar to the overall distribution (Appendix C). Thus, we suggest that the difference in positional bias between televised and radioed programs may be because viewers watching interviews on TV are relatively more focused, diminishing the need to recapitulate the main points before the program ends.

4 Experiments

4.1 Results on MediaSum

We apply several benchmark summarization models to the MEDIASUM dataset and report the re-

Dataset	MEDIASUM	AMI	ICSI	DiDi	CRD3	MultiWOZ	SAMSum
Source	Transcribed Speech					Written	
Type	Interview	Meeting	Meeting	Customer	Game	Booking	Daily
Real dialogue	✓	✓	✓	✓	✓	✓	×
Open domain	✓	×	×	×	×	×	✓
Public	✓	✓	✓	×	✓	✓	✓
Dialogues	463,596	137	59	328,880	159	10,438	16,369
Dial. words	1,553.7	4,757	10,189	/	31,802.8	180.7	83.9
Summ. words	14.4	322	534	/	2062.3	91.9	20.3
Turns	30.0	289	464	/	2,507.4	13.7	9.9
Speakers	6.5	4	6.2	2	9.6	2	2.2

Table 2: Comparison of dialogue summarization datasets. The number of dialogue words, summary words, turns and speakers are all averaged across all dialogues in the dataset.

Model	R-1	R-2	R-L
LEAD-3	14.96	5.10	13.29
PTGen	28.77	12.24	24.18
UniLM	32.70	17.27	29.82
BART	**35.09**	**18.05**	**31.44**

Table 3: ROUGE-1, ROUGE-2 and ROUGE-L F1 scores for models on MEDIASUM test set.

sults, including PTGen (See et al., 2017), the pre-trained models UniLM-base-uncased (Dong et al., 2019) and BART-Large (Lewis et al., 2019). The input concatenates transcripts from all turns, each prepended with the speaker name. We also include the LEAD-3 baseline which takes the first three sentences of the transcript as the summary. More implementation details are shown in Appendix D.

We randomly select 10K instances for validation and another 10K for test. We use the ROUGE (Lin, 2004) metrics and hyper-parameters are chosen based on the highest ROUGE-L score on the validation set.

As shown in Table 3, the LEAD-3 baseline has a relatively weak performance, indicating that media dialogues exhibit less lead bias than news articles. This aligns with the general guideline to avoid inverted pyramid structure in digital programs (Macadam, 2017). Moreover, pre-trained models such as BART and UniLM outperform the non-pre-trained PTGen model, showing the effectiveness of pre-training.

4.2 Transfer Learning

In this section, we evaluate the transfer capability of MEDIASUM by employing it for further training to improve the performance on other dialogue summarization tasks of different domains and styles. Specifically, we take the pre-trained model UniLM (Dong et al., 2019), fine-tune it on MEDIASUM, and then train it on datasets for meeting and dialogue summarization: AMI (McCowan et al., 2005), ICSI (Janin et al., 2003) and SAMSum (Gliwa et al., 2019).

Model	R-1	R-2	R-L
AMI			
UniLM	50.61	**19.33**	25.06
UniLM+MEDIASUM	**51.90**	**19.33**	**25.58**
ICSI			
UniLM	42.91	9.78	17.72
UniLM+MEDIASUM	**43.65**	**10.13**	**18.59**
SAMSum			
UniLM	50.00	26.03	42.34
UniLM+MEDIASUM	**50.55**	**26.39**	**42.68**

Table 4: Results on AMI, ICSI and SAMSum by using MEDIASUM as a dataset for transfer learning.

As shown in Table 4, on all three datasets, training on MEDIASUM leads to improvement on the target dataset. This shows the potential of using MEDIASUM as a transfer learning dataset for other dialogue summarization tasks.

5 Conclusion

We introduce MEDIASUM, a large-scale media interview dataset for dialogue summarization, consisting of 463.6K transcripts and summaries from NPR and CNN. We conduct transcript segmentation to align topic descriptions to segments for CNN interviews. The MEDIASUM dataset is an order of magnitude larger than existing corpora and

```
{
  "id": "NPR-11",
  "program": "Day to Day",
  "date": "2008-06-10",
  "url": "https://www.npr.org/templates/story/story.php?storyId=91356794",
  "title": "Researchers Find Discriminating Plants",
   "summary": "The 'sea rocket' shows preferential treatment to plants that are its kin. Evolutionary plant
            ecologist Susan Dudley of McMaster University in Ontario discusses her discovery.",
  "utt": [
    "This is Day to Day. I'm Madeleine Brand.",
    "And I'm Alex Cohen.",
    "Coming up, the question of who wrote a famous religious poem turns into a very unchristian battle.",
       "First, remember the 1970s? People talked to their houseplants, played them classical music. They
    were convinced plants were sensuous beings and there was that 1979 movie, 'The Secret Life of Plants.'",
       "Only a few daring individuals, from the scientific establishment, have come forward with offers to
    replicate his experiments, or test his results. The great majority are content simply to condemn his efforts
                without taking the trouble to investigate their validity.",
    ...
    "OK. Thank you.",
      "That's Susan Dudley. She's an associate professor of biology at McMaster University in Hamilt on
                Ontario. She discovered that there is a social life of plants."
  ],
  "speaker": [
    "MADELEINE BRAND, host",
    "ALEX COHEN, host",
    "ALEX COHEN, host",
    "MADELEINE BRAND, host",
    "Unidentified Male",
    ...
    "Professor SUSAN DUDLEY (Biology, McMaster University)",
    "MADELEINE BRAND, host"
  ]
}
```


Table 5: Example dialogue and summary from MEDIASUM. The number of strings in *utt* and *speaker* fields are the same.

contains complex multi-party conversations from multiple domains. We also show that MEDIASUM can be used as a dataset for transfer learning to improve a model's performance on other dialogue summarization tasks.

Ethics

We have used only the publicly available transcripts data from the media sources and adhere to their only-for-research-purpose guideline.

As media and guests may have biased views, the transcripts and summaries will likely contain them. The content of the transcripts and summaries only reflect the views of the media and guests, and should be viewed with discretion.

Acknowledgement

We thank William Hinthorn for proof-reading the paper and thank the anonymous reviewers for their insightful comments.

References

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Inigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. Multiwoz-a

large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. *arXiv preprint arXiv:1810.00278*.

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.

Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified language model pre-training for natural language understanding and generation. *arXiv preprint arXiv:1905.03197*.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. Samsum corpus: A human-annotated dialogue dataset for abstractive summarization. *arXiv preprint arXiv:1911.12237*.

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. *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03)*, 1:I–I.

Chris Kedzie, Kathleen McKeown, and Hal Daume III. 2018. Content selection in deep learning models of summarization. *arXiv preprint arXiv:1810.12343*.

Kundan Krishna, Sopan Khosla, Jeffrey P Bigham, and Zachary C Lipton. 2020. Generating soap notes from doctor-patient conversations. *arXiv preprint arXiv:2005.01795*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out*.

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

Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. 2015. The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. *arXiv preprint arXiv:1506.08909*.

Alison Macadam. 2017. The journey from print to radio storytelling: A guide for navigating a new landscape. https://training.npr.org/2017/12/06/the-journey-from-print-to-radio-storytelling-a-guide-for-navigating-a-new-landscape/#section4.

Bodhisattwa Prasad Majumder, Shuyang Li, Jianmo Ni, and Julian McAuley. 2020. Interview: A large-scale open-source corpus of media dialog. arXiv:2004.03090.

Iain McCowan, Jean Carletta, Wessel Kraaij, Simone Ashby, S Bourban, M Flynn, M Guillemot, Thomas Hain, J Kadlec, Vasilis Karaiskos, et al. 2005. The ami meeting corpus. *Proceedings of the 5th International Conference on Methods and Techniques in Behavioral Research*, 88:100.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Revanth Rameshkumar and Peter Bailey. Storytelling with dialogue: A Critical Role Dungeons and Dragons Dataset. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Abigail See, Peter J Liu, and Christopher D Manning. 2017. Get to the point: Summarization with pointer-generator networks. *arXiv preprint arXiv:1704.04368*.

Gaurav Sood. 2017. CNN Transcripts 2000–2014.

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

Chenguang Zhu, Ruochen Xu, Michael Zeng, and Xuedong Huang. 2020. A hierarchical network for abstractive meeting summarization with cross-domain pretraining. arXiv:2004.02016.

A Data statistics

Fig. 2 shows the distribution of the number of turns, speakers, dialogue words and summary words in the dialogues of MEDIASUM dataset. As shown, most dialogues have more than 500 words and 2 to 5 speakers.

B Topic analysis

Table 6 shows the top 10 words in each cluster of MEDIASUM dialogues computed by the Latent Dirichlet Allocation tool in scikit-learn package.

C Positional bias

Fig. 3 shows the frequency of non-stop topic words appearing in different positions of the dialogue. The dialogues are from the original CNN transcripts with one topic. The trend is mostly similar to that in Fig. 1, except for a slight increase near the end. Thus, it shows that in televised programs, most topic keywords are mentioned at the beginning.

D Implementation Details

For BART (Lewis et al., 2019), we use a learning rate of 2×10^{-5}, a batch size of 24 and train for 10 epochs. During beam search, we use a beam width of 3, and limits the minimum/maximum length of generated summary to be 3 and 80 tokens, respectively. The result on validation set of MEDIASUM is: 35.01 in ROUGE-1, 17.92 in ROUGE-2 and 31.15 in ROUGE-L.

For PTGen (See et al., 2017), we use a vocabulary of 50,000 words. The model is a LSTM-based encoder-decoder model with a hidden size of 512. We train the model with Adagrad optimizer for 10 epochs and a learning rate of 0.1. The result on validation set of MEDIASUM is: 28.07 in ROUGE-1, 12.11 in ROUGE-2 and 23.40 in ROUGE-L.

For UniLM (Dong et al., 2019), we train the model with Adam optimizer for 100,000 steps with 2,000 warmup steps and learning rate is set to 1.5×10^{-5}. The result on validation set of MEDIASUM is: 32.27 in ROUGE-1, 16.99 in ROUGE-2 and 29.06 in ROUGE-L.

In all experiments, we truncate the input after 1,024 tokens. We use 8 v100 GPUs for the computation.

We follow Zhu et al. (2020) to adopt 100/17/20 and 43/10/6 for train/dev/test split on AMI and ICSI respectively. We employ the split for SAMSum following Gliwa et al. (2019).

E Results on partitions

Table 7 shows the results of models on the CNN and NPR partitions of the test data. All models are trained on the corresponding partition of the training data, except UniLM$_{Com}$, which is trained on the entire MEDIASUM.

First, we notice that the result on NPR partition are better than that on CNN partition. Secondly, training on MEDIASUM can improve the ROUGE-L score by 0.6% on NPR partition, compared with using NPR partition only for training.

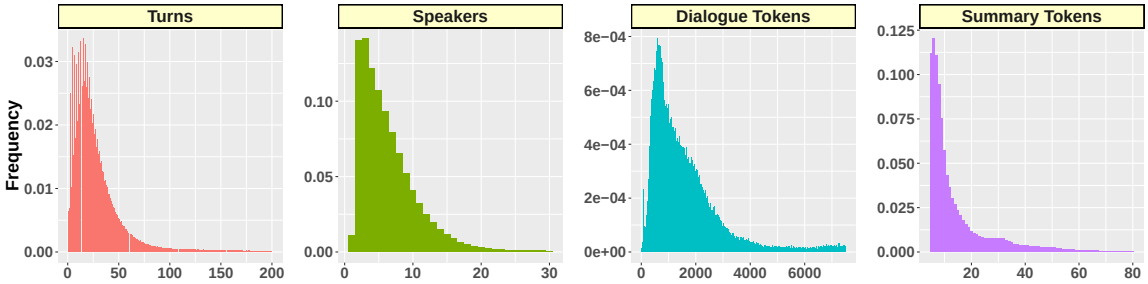

Figure 2: Distribution of the number of turns, speakers, dialogue words and summary words in the dialogues of MEDIASUM dataset.

Cluster	Top 10 words
1	prime, gop, iraq, bush, president, secretary, clinton, south, minister, white
2	plane, gop, today, look, report, rep, libya, crash, flight, continues
3	obama, coronavirus, attack, school, big, toll, saudi, gas, war, prices
4	forces, war, qaeda, crisis, syria, attack, middle, new, east, iraq
5	jobs, campaign, russian, news, white, tax, interview, old, president, iran
6	virginia, dead, new, suspect, day, case, covid, murder, 19, death
7	election, police, supreme, democrats, vote, house, impeachment, new, china, care
8	report, york, cnn, sanders, candidates, race, biden, democratic, president, presidential

Table 6: Top 10 topics words in each cluster of MEDIASUM dialogues computed by the Latent Dirichlet Allocation tool in scikit-learn package.

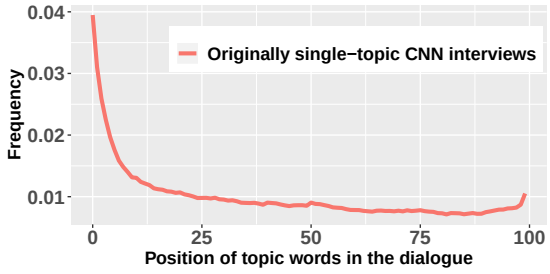

Figure 3: The frequency of non-stop topic words appearing in different positions of the dialogue. The dialogues are from the original CNN transcripts with one topic. The positions are normalized to [0, 100].

Model	R-1	R-2	R-L
CNN			
LEAD-3	13.36	4.37	11.10
PTGen	27.54	11.47	23.45
BART	34.07	17.57	31.36
UniLM	31.97	16.97	29.88
UniLM$_{Com}$	31.88	16.97	29.79
NPR			
LEAD-3	28.39	11.21	19.90
PTGen	35.86	16.01	24.46
BART	43.55	21.99	32.03
UniLM	41.42	20.73	30.65
UniLM$_{Com}$	41.58	21.25	31.24

Table 7: ROUGE-1, ROUGE-2 and ROUGE-L F1 scores on the CNN and NPR partitions of the test data. All models are trained on the corresponding partition of the training data, except UniLM$_{Com}$, which is trained on the entire MEDIASUM.