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.

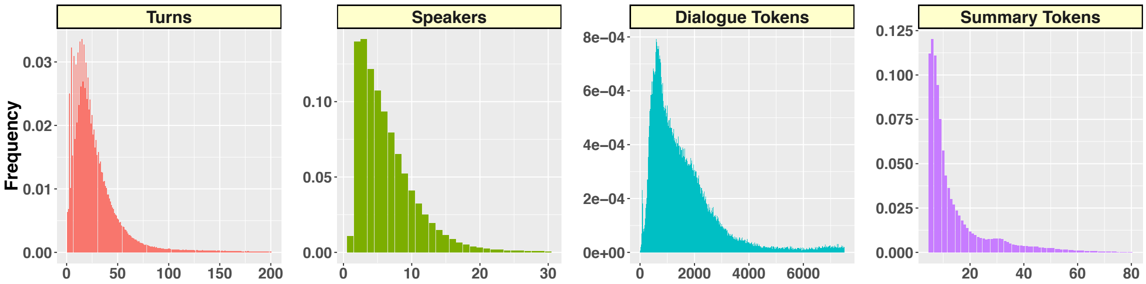

Figure 2: Distribution of the number of turns, speakers, dialogue words and summary words in the dialogues of MEDIASUM dataset.

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.

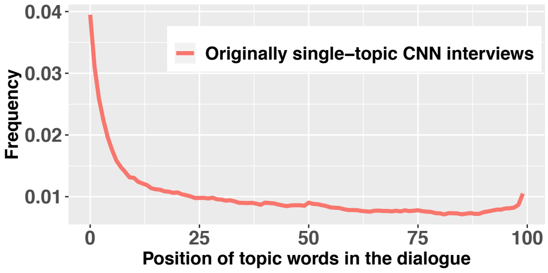

Figure 3: The frequency of non-stop topic words appearing in different positions of the dialogue. The dialogues are from the original CNN transcripts with one topic. The positions are normalized to [0, 100].

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.