A Implementation Details

For pretraining, we use a dropout rate of 0.3 for all inputs to transformer layers. We use RAdam (Liu et al., 2019a) as the optimizer, with a learning rate of 10^{-4}. Also, due to the different numerical scales of the positional embedding and initialized sentence piece embeddings, we divide the positional embedding by 100 before feeding it into the transformer. We pretrain one model for 10 epochs. After each epoch, the model is evaluated on validation data. We pick the check points with the highest ROUGE L.

For unsupervised finetuning on specific datasets, the learning rate is set to 2×10^{-4} and dropout ratio stays the same as in pretraining. The batch size is 16, and the vocabulary embeddings are also updated in the training process. During the test phase, we generate the summarization from trained encoder and decoder by beam search. The ROUGE version we use for evaluation is ROUGE-1.5.5. This is consistent with benchmark models whose version of ROUGE are available in open-sourced codes and original papers.

At test time, we limit the longest length of generated summaries, which is set based on validation dataset. For instance, the maximum generation length for CNN/DM dataset is 175.

B Datasets Information

For a better understanding of the evaluation protocols, the statistical information of evaluation datasets is summarized in Table 4.

Table 4: Average document and summary length in number of words and sentences on NYT, CNN/DM, and English Gigaword datasets (test set).

Dataset	# docs	avg. document		avg. summ.	
		words	sen.	words	sen.
CNN/DM	11,490	641.9	28.0	54.6	3.9
NYT	4,375	1,290.5	50.7	79.8	3.5
Gigaword	1,937	29	1	8	1