A Appendices

A.1 Detailed Dataset introduction

CNN/DailyMail The CNN/DailyMail question answering dataset (Hermann et al., 2015) modified by Nallapati et al. (2016) is commonly used for summarization. The dataset consists of online news articles with paired human-generated summaries. For the data preprocessing, we use the non-anonymized data as See et al. (2017), which doesn't replace named entities.

XSUM XSUM (Narayan et al., 2018) is a dataset consists of the articles and the single-sentence answers of the question "What is the article about?" as summary. It is more abstractive compared with CNN/DailyMail.

PUBMED PUBMED (Cohan et al., 2018) is drawn from scientific papers specifically medical journal articles from the PubMed Open Access Subset. We use the introduction as source document and the abstract as summary here.

BIGPATENT BIGPATENT (Sharma et al., 2019) consists of 1.3 million records of U.S. patent documents and the corresponding summaries are created by human. According to Cooperative Patent Classification (CPC), the dataset is divided to nine categories. One of the nine categories is chosen as a dataset in difference domain in our experiment (Category B: Performing Operations; Transporting).

REDDIT TIFU REDDIT TIFU (Kim et al., 2019) is a dataset with less formal posts compared with datasets mentioned above which mostly use formal documents as source. It is collected from the online discussion forum Reddit. They regard the body text as source, the title as short summary, and the TL;DR summary as long summary, thus making two sets of datasets: TIFU-short and TIFU-long. TIFU-long is used in this paper.

A.2 Dataset statistics

The detailed dataset statistics are presented in table 1

Datasets	Statistics	Topics	Oracle	Lead-k
CNNDM	2,764/123/107M	News	55.21	40.32
Xsum	1126/60/59M	News	30.41	16.38
Pubmed	644/36/38M	Scientific	46.21	37.52
BigPatent B	4,812/265/262M	Patents	51.53	31.85
Reddit	206/3.3/3.6M	Posts	36.47	11.09

Table 1: Detailed statistics of five datasets. Lead-k indicates ROUGE-1 F1 score of the first k sentences in the document and Oracle indicates the globally optimal combination of sentences in terms of ROUGE-1 F1 scores with ground truth, the latter represents the upper bound of extractive models.

A.3 Experimental setup

A.3.1 Extractive Summarizers

We use the same training setup in (Zhong et al., 2019). We use cross entropy as loss function to train $LSTM_{non}$ and $Trans_{auto}$. The hidden state dimension of LSTM in $LSTM_{non}$ is set to 512 and the hidden state dimension of Transformer in $Trans_{auto}$ is 2048. We use Transformer with 8 heads.

$BERT_{non}$ and $Trans_{non}$ is constructed according to Liu and Lapata (2019). All documents and summaries are truncated to 512 tokens when training. $BERT_{non}$ and $Trans_{non}$ are trained for 50000 steps, the gradient is accumulated every two steps. We use Adam as optimizer and the learning rate is set to 2e-3.

$BERT_{match}$ is trained as in Zhong et al. (2020). It uses the base version of BERT as base model. We use Adam optimizer with warming up. The learning rate schedule follows Vaswani et al. (2017).

A.3.2 Abstractive Summarizers

L2L, $L2L_{ptr}$ and $L2L_{ptr}^{cov}$ are trained using the pytorch reproduced version code of See et al. (2017). We use the same size of vocabulary(50k), hidden state dimension (256) and word embedding dimension (128) as in the paper. All of three models are trained with 650000 maximum training steps, We use Adagrad to train the models with learning rate of 0.15.

BE2T and *T2T* is constructed according to Liu and Lapata (2019). We use two separate optimizers for the decoder and encoder regarding *BE2T* to offset the mismatch of encoder and decoder, since the former is pre-trained while the latter is not. Learning rates for the optimizers of encoder and decoder are 0.002 and 0.2 respectively. On the other hand, *BE2T* and *T2T* are trained with gradient accumulation every five steps, training step for which is 200000.

BART uses the large pre-trained sequence to sequence model in Lewis et al. (2019). The total learning step when fine-tuning is set to 20000 with 500 steps warming up. We use Adam as optimizer and learning rate is 3e-05.

	Models	CNNDM			XSUM			PubMed			Bigpatent b			Reddit		
		R1	R2	RL	R1	R2	RL	R1	R2	RL	R1	R2	RL	R1	R2	RL
Ext.	LSTM_{non}	41.36	18.81	37.73	19.51	3.10	14.50	42.98	16.59	38.28	39.29	13.07	32.61	20.46	5.05	16.33
	Trans_{non}	40.84	18.23	37.09	15.74	1.67	11.58	38.45	13.28	34.16	34.41	10.05	28.75	16.25	2.60	12.57
	Trans_{auto}	41.35	18.77	37.75	19.29	2.80	14.21	42.74	16.34	38.05	38.76	12.60	32.17	18.55	3.44	14.62
	BERT_{non}	42.69	19.88	38.99	21.76	4.24	16.00	38.74	13.62	34.48	35.85	11.05	29.97	21.84	5.21	17.15
	BERT_{match}	44.26	20.58	40.40	24.97	4.76	18.48	41.19	14.91	36.73	38.89	12.82	32.48	25.32	6.16	20.17
Abs.	L2L	32.80	12.84	30.34	28.31	8.71	22.30	27.84	7.45	25.69	30.46	9.76	27.61	16.89	1.24	13.63
	L2L_{ptr}	37.06	15.96	33.74	29.67	9.58	23.40	32.04	10.38	28.97	31.03	9.92	25.35	21.32	4.46	17.14
	L2L_{ptr}^{cov}	39.95	17.54	36.25	28.83	8.83	22.62	35.27	11.89	31.92	35.90	12.31	32.78	21.28	4.39	17.22
	T2T	39.90	17.66	37.08	29.01	9.13	22.77	30.71	8.10	27.97	42.94	16.75	37.06	19.96	3.36	15.60
	BE2T	41.34	18.98	38.41	38.99	16.64	31.23	37.11	13.38	33.72	43.10	17.11	37.34	26.66	7.00	21.21
	BART	44.75	21.69	41.46	44.73	21.99	37.02	45.02	16.94	41.17	45.78	18.31	38.98	34.00	11.88	26.91

Table 2: Representative summarizers we have studied in this paper and their correspond performance (ROUGE-1 F1, ROUGE-2 F1, ROUGE-L F1) on different datasets.

A.4 In-dataset ROUGE results for all models

Tab. 2 displays in-dataset ROUGE-1 F1 ,ROUGE-2 F1 ,ROUGE-L F1 scores.

A.5 The ROUGE-1 F1 score difference of all model pairs which are meaningful to compare

The holistic and fine-grained results of pair-wise comparison are displayed in Tab. 5.

A.6 Cross-dataset factuality results of all models

The cross-dataset factcc results for abstractive models are shown in Fig. 3 and the factcc results of extractive models are demonstrated in Fig. 4.

A.7 Code urls

A.7.1 Training code urls

The models and their training code urls are listed below:

$LSTM_{non}$ and *$Trans_{auto}$* are trained from the code in Zhong et al. (2019), the code url is https://github.com/maszhongming/Effective_Extractive_Summarization.

We use the code from Liu and Lapata (2019) for *$BERT_{non}$*, *$Trans_{non}$*, *BE2T* and *T2T*. Code url is https://github.com/nlpyang/PreSumm.

$BERT_{match}$ uses the code from Zhong et al. (2020) and the code url is https://github.com/maszhongming/MatchSum.

L2L, *$L2L_{ptr}$* and *$L2L_{ptr}^{cov}$* are trained from the code of See et al. (2017), code url is https://github.com/atulkum/pointer_summarizer.

We use code in fairseq (Ott et al., 2019) to fine-tune *BART*, the code url is https://github.com/pytorch/fairseq/tree/master/examples/bart.

A.7.2 Evaluation code urls

The evaluation metrics code urls are listed below:

We use pyrouge (https://github.com/bheinzerling/pyrouge) to evaluate the ROUGE performance of models.

The url for Factcc (Kryściński et al., 2019) is https://github.com/salesforce/factCC.

The url for other metrics for dataset bias is https://github.com/zide05/Data-bias-metrics.

References

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A discourse-aware attention model for abstractive summarization of long documents. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, volume 2, pages 615–621.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems*, pages 1684–1692.

Byeongchang Kim, Hyunwoo Kim, and Gunhee Kim. 2019. Abstractive summarization of reddit posts with multi-level memory networks. 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 2519–2531.

Wojciech Kryściński, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Evaluating the factual consistency of abstractive text summarization. *arXiv*, pages arXiv–1910.

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

ABS models	L2L						$L2L_{ptr}$						$L2L^{cov}_{ptr}$						T2T						BE2T						BART					
	CNN.	XSUM	Pubm.	Patent B	Red.	avg	CNN.	XSUM	Pubm.	Patent B	Red.	avg	CNN.	XSUM	Pubm.	Patent B	Red.	avg	CNN.	XSUM	Pubm.	Patent B	Red.	avg	CNN.	XSUM	Pubm.	Patent B	Red.	avg	CNN.	XSUM	Pubm.	Patent B	Red.	avg
CNN	68.6	71.1	73.3	69.9	53.9	67.4	89.4	91.3	92.2	91.7	83.5	89.6	95.9	94.5	90.9	96.9	94.6	94.6	72.4	75.7	71.5	71.8	70.5	72.4	78.7	83.9	87.7	92.1	78.7	84.2	69.9	77.9	87.4	84.1	90.2	81.9
XSUM	13.4	23.5	18.1	13.2	31.0	19.8	6.3	17.8	9.0	8.2	23.2	12.9	7.4	18.1	11.0	7.6	6.5	10.1	9.7	22.6	10.8	9.9	19.1	14.4	14.5	21.1	29.8	8.7	31.3	21.1	35.5	24.7	36.1	50.1	50.7	39.4
Pubm.	61.0	70.0	62.8	78.6	46.6	63.8	77.6	80.7	81.5	75.1	85.9	80.2	70.7	75.6	76.6	67.9	75.4	73.2	58.5	59.3	56.2	72.3	34.9	56.2	55.4	58.7	70.8	71.7	56.4	62.6	69.5	61.5	58.4	61.3	94.1	69.0
Patent B	94.4	94.3	89.0	71.9	91.0	88.1	65.2	60.3	70.9	62.8	71.0	66.0	67.0	63.3	64.6	61.6	77.4	66.8	79.2	81.2	84.4	68.7	73.9	77.5	85.4	88.4	80.3	66.5	82.0	80.6	52.1	53.8	69.0	67.4	76.8	63.8
Red.	20.9	40.2	11.1	13.2	50.9	27.3	37.2	21.5	55.2	62.6	61.1	47.5	27.4	23.5	42.9	49.7	62.2	41.1	34.8	35.7	50.6	44.6	52.5	43.6	17.2	25.7	25.1	30.0	50.3	29.6	59.6	50.3	69.1	49.3	44.2	54.5
avg	51.7	59.8	50.9	49.4	54.7	53.3	55.2	54.3	61.8	60.1	65.0	59.2	53.7	55.0	57.2	56.7	63.2	57.2	50.9	54.9	54.7	53.5	50.2	52.8	50.2	55.6	58.7	53.8	59.8	55.6	57.3	53.6	64.0	62.4	71.2	61.7

Table 3: factcc result for Abstractive models

EXT models	$LSTM_{non}$						$Trans_{non}$						$Trans_{auto}$						$BERT_{non}$						$BERT_{match}$					
	CNN.	XSUM	Pubm.	Patent B	Red.	avg	CNN.	XSUM	Pubm.	Patent B	Red.	avg	CNN.	XSUM	Pubm.	Patent B	Red.	avg	CNN.	XSUM	Pubm.	Patent B	Red.	avg	CNN.	XSUM	Pubm.	Patent B	Red.	avg
CNN	99.2	99.9	96.0	99.1	95.2	97.9	100.0	100.0	98.0	99.1	100.0	99.4	98.1	100.0	91.3	93.5	100.0	96.6	99.6	99.9	97.3	98.2	98.6	98.7	99.8	99.4	92.9	95.7	99.1	97.4
XSUM	84.1	94.3	90.3	81.4	94.1	88.9	99.8	100.0	97.4	98.2	100.0	99.1	86.8	99.3	82.9	69.9	100.0	87.8	98.4	99.7	96.6	95.7	99.9	98.1	99.7	99.5	93.2	95.1	98.8	97.3
Pubm.	70.5	84.3	80.8	65.1	89.0	77.9	97.7	98.8	95.1	94.7	100.0	97.3	87.5	99.6	79.0	64.4	99.7	86.1	95.3	99.3	95.1	94.3	99.5	96.7	99.7	99.2	93.1	95.2	99.3	97.3
Patent B	86.1	96.0	90.9	74.1	96.0	88.6	98.3	99.8	96.3	97.4	99.5	98.3	90.7	99.8	85.5	68.8	99.7	88.9	97.0	99.0	96.0	94.8	99.1	97.2	99.7	99.0	93.0	94.5	98.4	96.9
Red.	81.0	92.1	86.9	64.6	90.2	83.0	90.3	94.1	94.1	86.7	96.3	92.3	79.4	98.7	79.6	56.4	98.1	82.5	97.0	98.9	95.3	91.9	98.8	96.4	99.7	99.3	93.1	96.1	99.3	97.5
avg	84.2	93.3	89.0	76.8	92.9	87.2	97.2	98.6	96.2	95.2	99.2	97.3	88.5	99.5	83.7	70.6	99.5	88.4	97.5	99.4	96.1	95.0	99.2	97.4	99.7	99.3	93.0	95.3	99.0	97.3

Table 4: factcc result for Extractive models

analysis aspect	Architecture						
model type	ABS						
compare models	$L2L_{ptr}$ vs. L2L	$L2L^{cov}_{ptr}$ vs. $L2L_{ptr}$	T2T vs. L2L	BE2T vs. T2T	BART vs. BE2T	BART vs. L2L	BART vs. T2T
holistic analysis	stiff. : 20.74 vs. 18.03 stable. : 68.63 vs. 66.93	stiff. : 22.81 vs. 20.74 stable. : 70.71 vs. 68.63	stiff. : 19.79 vs. 18.03 stable. : 62.12 vs. 66.93	stiff. : 23.49 vs 19.79 stable. : 62.93 vs. 62.12	stiff. : 31.66 vs. 23.49 stable. : 73.83 vs. 62.93	stiff. : 31.66 vs. 18.03 stable. : 73.83 vs. 66.93	stiff. : 31.66 vs. 19.79 stable. : 73.83 vs. 62.12

fine-grain analysis		$L2L_{ptr}$ vs. L2L						$L2L^{cov}_{ptr}$ vs. $L2L_{ptr}$						T2T vs. L2L					
		CNN.	Xsum	Pubm.	Patent b	Red.	avg	CNN.	Xsum	Pubm.	Patent b	Red.	avg	CNN.	Xsum	Pubm.	Patent b	Red.	avg
ROUGE origin	CNN.	4.3	0.5	5.3	3.2	1.5	3.0	2.9	1.8	6.4	3.4	1.7	3.2	7.1	2.4	7.2	5.2	5.4	5.4
	Xsum	3.4	1.4	3.4	4.2	0.1	2.5	-0.8	-0.8	-4.5	-2.4	-0.1	-1.7	0.4	0.7	-5.4	-4.5	-0.6	-1.9
	Pubm.	10.3	2.3	4.2	3.0	2.6	4.5	4.5	1.7	3.2	3.4	2.7	3.1	1.6	-1.5	2.9	2.9	2.1	1.6
	Patent b	1.1	-1.1	2.5	0.6	-0.3	0.5	1.0	2.0	2.2	4.9	0.8	2.2	1.2	-2.5	3.0	12.5	-0.0	2.9
	Red.	2.2	3.1	2.6	2.9	4.4	3.0	3.3	1.0	6.5	6.9	-0.0	3.5	2.0	0.1	-0.9	-0.7	3.1	0.7
	avg	4.2	1.2	3.6	2.8	1.7	2.7	2.2	1.1	2.8	3.2	1.0	2.1	2.5	-0.1	1.3	3.1	2.0	1.8
ROUGE normali.	CNN.	0.0	-1.0	4.8	8.7	-9.9	0.5	0.0	8.1	9.6	-4.1	8.0	4.3	0.0	6.8	15.0	-14.3	14.5	4.4
	Xsum	1.8	0.0	0.7	12.2	-13.8	0.2	-6.7	0.0	-19.7	-18.0	-0.2	-8.9	-10.7	0.0	-24.9	-31.2	-13.4	-16.0
	Pubm.	23.3	5.6	0.0	8.4	1.6	7.8	6.7	7.4	0.0	-1.2	12.6	5.1	-2.9	-6.1	0.0	-13.8	2.7	-4.0
	Patent b	-1.6	-5.8	-0.4	0.0	-14.5	-4.4	-0.1	8.1	0.8	0.0	3.7	2.5	-3.8	-9.6	4.2	0.0	-9.9	-3.8
	Red.	1.9	8.7	3.4	8.4	0.0	4.5	5.6	4.9	14.8	11.7	0.0	7.4	-1.0	-0.4	-6.4	-15.2	0.0	-4.6
	avg	5.1	1.5	1.7	7.5	-7.3	1.7	1.1	5.7	1.1	-2.3	4.8	2.1	-3.7	-1.9	-2.4	-14.9	-1.2	-4.8

fine-grain analysis		BE2T vs. T2T						BART vs. BE2T						BART vs. L2L						BART vs. T2T					
		CNN.	Xsum	Pubm.	Patent b	Red.	avg	CNN.	Xsum	Pubm.	Patent b	Red.	avg	CNN.	Xsum	Pubm.	Patent b	Red.	avg	CNN.	Xsum	Pubm.	Patent b	Red.	avg
ROUGE origin	CNN.	1.4	0.2	1.1	1.4	1.1	1.0	3.7	2.0	1.3	1.8	0.8	1.9	12.3	4.5	9.6	8.3	7.3	8.4	5.2	2.1	2.4	3.1	1.9	3.0
	Xsum	2.1	10.0	3.2	4.3	5.1	4.9	4.5	6.0	7.4	9.7	5.8	6.7	7.0	16.7	5.1	9.6	10.3	9.7	6.6	16.0	10.6	14.1	10.9	11.6
	Pubm.	6.7	3.1	6.4	8.4	1.4	5.2	14.1	9.2	7.4	4.1	6.1	8.2	22.4	10.8	16.7	15.4	9.7	15.0	20.8	12.3	13.8	12.5	7.5	13.4
	Patent b	1.1	0.4	2.0	0.2	2.4	1.2	17.6	12.9	12.1	2.7	7.9	10.6	19.9	10.9	17.1	15.4	10.2	14.7	18.7	13.3	14.0	2.9	10.3	11.8
	Red.	5.6	3.5	7.9	7.0	6.7	6.1	17.7	10.0	18.5	14.3	6.6	13.4	25.3	13.7	25.5	20.7	16.4	20.3	23.3	13.5	26.4	21.3	13.3	19.6
	avg	3.4	3.4	4.1	4.3	3.3	3.7	11.5	8.0	9.3	6.5	5.4	8.2	17.4	11.3	14.8	13.9	10.8	13.6	14.9	11.5	13.4	10.8	8.8	11.9
ROUGE normali.	CNN.	0.0	-17.1	-14.9	2.9	-20.1	-9.8	0.0	-2.5	-11.9	-0.9	-12.9	-5.6	0.0	-12.8	-11.9	-12.2	-18.4	-11.1	0.0	-19.6	-26.9	2.1	-32.9	-15.5
	Xsum	3.3	0.0	-0.4	9.9	5.5	3.7	5.2	0.0	8.1	18.2	5.5	7.4	-2.2	0.0	-17.1	-3.1	-2.4	-5.0	8.4	0.0	7.7	28.1	11.0	11.1
	Pubm.	14.9	-1.6	0.0	19.4	-8.5	4.8	27.1	15.6	0.0	4.3	9.1	11.2	39.2	7.9	0.0	10.0	3.4	12.1	42.1	14.0	0.0	23.7	0.6	16.1
	Patent b	1.4	-8.4	-6.0	0.0	-4.6	-3.5	36.1	24.8	17.2	0.0	13.8	18.4	33.7	6.8	15.4	0.0	-0.7	11.0	37.4	16.4	11.2	0.0	9.2	14.9
	Red.	12.3	0.2	16.0	16.2	0.0	8.9	35.5	17.7	33.9	28.4	0.0	23.1	46.8	17.4	43.4	29.3	0.0	27.4	47.8	17.9	49.8	44.6	0.0	32.0
	avg	6.4	-5.4	-1.1	9.7	-5.5	0.8	20.8	11.1	9.4	10.0	3.1	10.9	23.5	3.9	6.0	4.8	-3.6	6.9	27.1	5.7	8.4	19.7	-2.4	11.7

analysis aspect	Architecture				Generation way		
model type	EXT				LSTM	BERTSUM	Transformer
compare models	$Trans_{non}$ vs. $LSTM_{non}$	$Trans_{auto}$ vs. $Trans_{non}$	$BERT_{match}$ vs. $BERT_{non}$	$BERT_{non}$ vs. $Trans_{non}$	$LSTM_{non}$ vs. L2L	$Trans_{non}$ vs. T2T	$BERT_{non}$ vs. BE2T
holistic analysis	stiff. : 28.02 vs. 28.51 stable. : 99.05 vs. 87.00	stiff. : 28.51 vs. 28.02 stable. : 88.71 vs. 99.05	stiff. : 32.27 vs. 28.98 stable. : 91.98 vs. 88.93	stiff. : 28.98 vs. 28.02 stable. : 88.93 vs. 99.05	stiff. : 28.51 vs. 18.03 stable. : 87.00 vs. 66.93	stiff. : 28.02 vs. 19.79 stable. : 99.05 vs. 62.12	stiff. : 28.98 vs. 23.49 stable. : 88.93 vs. 62.93

fine-grain analysis		$Trans_{non}$ vs. $LSTM_{non}$						$Trans_{auto}$ vs. $Trans_{non}$						$BERT_{match}$ vs. $BERT_{non}$					
		CNN.	Xsum	Pubm.	Patent b	Red.	avg	CNN.	Xsum	Pubm.	Patent b	Red.	avg	CNN.	Xsum	Pubm.	Patent b	Red.	avg
ROUGE origin	CNN.	-0.5	-0.8	-1.8	-0.8	13.8	2.0	0.5	0.4	3.1	0.9	-12.9	-1.6	1.6	4.1	4.5	3.0	4.7	3.6
	Xsum	3.2	-3.8	-2.5	4.2	2.9	0.8	-3.2	3.5	4.6	-5.7	-2.3	-0.6	2.9	3.2	3.5	1.6	5.7	3.4
	Pubm.	4.5	-1.6	-4.5	-0.7	-3.0	-1.0	-0.7	1.2	4.3	2.2	0.1	1.4	0.9	4.0	2.4	0.2	8.7	3.3
	Patent b	3.9	0.9	-2.6	-4.9	-2.2	-1.0	-2.9	0.1	3.9	4.4	1.4	1.4	4.6	3.1	3.5	3.0	3.7	3.6
	Red.	-4.4	-2.0	-3.4	-1.9	-4.2	-3.2	1.2	1.7	4.5	-0.5	2.3	1.8	3.3	4.2	3.5	-1.4	3.5	2.6
	avg	1.3	-1.5	-2.9	-0.8	1.5	-0.5	-1.0	1.4	4.1	0.3	-2.3	0.5	2.6	3.7	3.5	1.3	5.3	3.3
ROUGE normali.	CNN.	0.0	16.7	5.8	9.2	104.9	27.3	0.0	-17.7	-2.1	-7.9	-91.9	-23.9	0.0	5.8	5.3	0.7	6.9	3.7
	Xsum	8.9	0.0	4.5	22.8	37.1	14.6	-9.0	0.0	1.0	-25.6	-26.0	-11.9	3.4	0.0	2.8	-2.7	11.5	3.0
	Pubm.	12.0	11.0	0.0	9.9	4.4	7.4	-2.8	-12.2	0.0	-4.7	-10.9	-6.1	-1.2	6.1	0.0	-6.5	26.5	5.0
	Patent b	10.5	25.2	4.2	0.0	7.4	9.5	-8.0	-19.4	-0.6	0.0	-3.6	-6.3	7.3	1.8	2.8	0.0	3.3	3.0
	Red.	-9.7	8.0	2.4	6.3	0.0	1.4	2.0	-8.8	0.7	-11.2	0.0	-3.4	4.4	6.2	2.9	-10.5	0.0	0.6
	avg	4.3	12.2	3.4	9.6	30.7	12.1	-3.6	-11.6	-0.2	-9.9	-26.5	-10.3	2.8	4.0	2.7	-3.8	9.6	3.1

fine-grain analysis		$BERT_{non}$ vs. $Trans_{non}$						$LSTM_{non}$ vs. L2L						$Trans_{non}$ vs. T2T						$BERT_{non}$ vs. BE2T					
		CNN.	Xsum	Pubm.	Patent b	Red.	avg	CNN.	Xsum	Pubm.	Patent b	Red.	avg	CNN.	Xsum	Pubm.	Patent b	Red.	avg	CNN.	Xsum	Pubm.	Patent b	Red.	avg
ROUGE origin	CNN.	1.8	1.2	0.3	0.8	-10.9	-1.3	8.6	0.1	13.2	4.9	2.0	5.7	0.9	-3.1	4.3	-1.1	10.3	2.3	1.3	-2.0	3.5	-1.8	-1.7	-0.1
	Xsum	-0.9	6.0	0.1	-1.6	-0.7	0.6	13.1	-8.8	18.3	7.1	3.8	6.7	15.9	-13.3	21.3	15.8	7.3	9.4	12.9	-17.2	18.3	9.9	1.5	5.1
	Pubm.	2.5	1.4	0.3	0.6	-2.2	0.5	18.6	4.8	15.1	11.1	9.0	11.7	21.4	4.7	7.7	7.6	3.9	9.1	17.2	2.9	1.6	-0.3	0.3	4.3
	Patent b	0.5	1.1	0.2	1.4	3.8	1.4	19.7	2.8	22.8	8.8	5.9	12.0	22.3	6.1	17.2	-8.5	3.7	8.2	21.8	6.7	15.4	-7.2	5.1	8.4
	Red.	8.3	3.0	-0.1	1.6	5.6	3.7	21.4	7.3	30.7	18.0	3.6	16.2	15.0	5.1	28.2	16.8	-3.7	12.3	17.8	4.6	20.2	11.4	-4.8	9.8
	avg	2.4	2.5	0.2	0.6	-0.9	1.0	16.3	1.2	20.0	10.0	4.9	10.5	15.1	-0.1	15.7	6.1	4.3	8.2	14.2	-1.0	11.8	2.4	0.1	5.5
ROUGE normali.	CNN.	0.0	-23.9	0.1	-1.5	-96.6	-24.4	0.0	28.4	-0.7	-7.9	-4.8	3.0	0.0	38.3	-9.8	15.6	85.6	25.9	0.0	31.5	5.2	11.1	9.0	11.4
	Xsum	-6.1	0.0	-0.5	-8.3	-31.8	-9.3	18.3	0.0	15.8	2.0	6.6	8.5	37.8	0.0	45.1	56.1	57.2	39.2	28.4	0.0	45.0	37.8	19.9	26.2
	Pubm.	2.0	-21.0	0.0	-2.2	-33.7	-11.0	36.8	44.3	0.0	12.5	35.1	25.7	51.6	61.4	0.0	36.1	36.7	37.2	38.7	42.0	0.0	14.5	11.5	21.3
	Patent b	-2.6	-24.8	-0.2	0.0	-5.5	-6.6	39.6	35.2	31.4	0.0	17.8	24.8	53.9	70.1	31.5	0.0	35.0	38.1	49.9	53.7	37.2	0.0	34.1	35.0
	Red.	16.3	-12.8	-1.0	1.0	0.0	0.7	44.7	52.9	58.4	35.1	0.0	38.2	36.1	61.4	67.2	56.6	0.0	44.3	40.1	48.4	50.2	41.5	0.0	36.1
	avg	1.9	-16.5	-0.3	-2.2	-33.5	-10.1	27.9	32.2	21.0	8.3	10.9	20.1	35.9	46.2	26.8	32.9	42.9	36.9	31.4	35.1	27.5	21.0	14.9	26.0

Table 5: The difference of ROUGE-1 F1 scores between different models pairs. Every column of the table represents the compared result of one pair of models. The line of holistic analysis displays the overall stiffness and stableness of compared models. The rest of the table is the fine-grained results, the first and third lines of which are the origin compared result ($\mathbf{U_A} - \mathbf{U_B}$ for models pairs A and B) and the second and fourth lines are the normalized compared result ($\hat{\mathbf{U}}_\mathbf{A} - \hat{\mathbf{U}}_\mathbf{B}$ for models pairs A and B). For all heatmap, 'grey' represents positive, 'red' represents negative and 'white' represents approximately zero.

for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. 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 3721–3731.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Ça glar Gulçehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. *CoNLL 2016*, page 280.

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.

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 NAACL-HLT 2019: Demonstrations*.

Abigail See, Peter J Liu, and Christopher D Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1073–1083.

Eva Sharma, Chen Li, and Lu Wang. 2019. Bigpatent: A large-scale dataset for abstractive and coherent summarization. *arXiv preprint arXiv:1906.03741*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Ming Zhong, Pengfei Liu, Yiran Chen, Danqing Wang, Xipeng Qiu, and Xuanjing Huang. 2020. Extractive summarization as text matching.

Ming Zhong, Pengfei Liu, Danqing Wang, Xipeng Qiu, and Xuan-Jing Huang. 2019. Searching for effective neural extractive summarization: What works and what's next. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1049–1058.