The Effect of Pretraining on Extractive Summarization for Scientific Documents

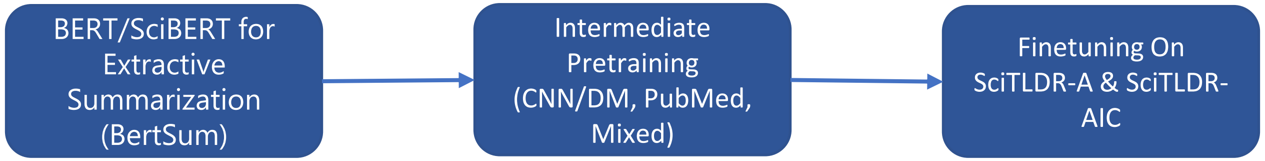
Yash Gupta[1] Pawan Sasanka Ammanamanchi[2] Shikha Bordia[3] Arjun Manoharan[3] Deepak Mittal[3] Ramakanth Pasunuru[4]
Manish Shrivastava[2] Maneesh Singh[3] Mohit Bansal[4] Preethi Jyothi[1]

[1]Indian Institute of Technology, Bombay, [2]International Institute of Information Technology, Hyderabad, [3]Verisk Analytics, [4]University of North Carolina, Chapel Hill

Problem Formulation

- Base model: BERT-based extractive summarization system for scientific articles.
- Investigating the influence of intermediate pretraining using existing summarization tasks across three criteria:
  - ❑ Domain of Intermediate Pretraining Corpus
  - ❑ Size of Corpus
  - ❑ Input Length

Methodology



BERT/SciBERT for Extractive Summarization (BertSum)

Intermediate Pretraining (CNN/DM, PubMed, Mixed)

Finetuning On SciTLDR-A & SciTLDR-AIC

- **Domain of Intermediate Pretraining Corpus:** Every intermediate pretraining corpus set to the same size using sampling. We use different domain corpus - CNN/DailyMail, Pubmed and a Mixed dataset. Mixed dataset is composed of documents from different domains that are semantically closest to the target domain. We select 83K articles with the least averaged L2 distance between BERT-base embeddings, derived using [CLS] tokens, of the intermediate pretraining corpus (Pubmed/ CNN/Daily Mail) and target corpus (SciTLDR)
- **Size of Corpus:** We study the effect of varying the size of intermediate training corpus size (CNN/Daily Mail)
- **Input Length**: We vary the input length of target data (SciTLDR-AIC) in the finetuning stage

Experiments and Results

Dataset Size	R1	R2	RL
83K articles	41.93	20.1	33.95
176K articles	**42.27**	**20.37**	**34.32**
286K articles	42.21	20.24	34.19

Size of Corpus Finetuning results on SCITLDR-AIC for different size of the pretraining dataset (CNN/Daily Mail)

Pretraining Corpus	R1	R2	RL
BERT			
Finetuning	36.99	16.14	29.64
Pubmed (83K)	40.82	18.98	32.84
CNN/DM (83K)	41.93	20.1	33.95
MIXED (83K)	**42.78**	**21.06**	**34.83**
SCIBERT			
Finetuning	37.16	15.94	29.65
Pubmed (83K)	40.61	18.69	32.68
CNN/DM (83K)	40.74	19.09	32.95

Domain of Intermediate Pretraining Corpus Max ROUGE scores for SCITLDR-AIC

Input Length	R1	R2	RL
512 tokens	42.21	20.24	34.19
1024 tokens	42.21	20.34	34.35
1500 tokens	**42.23**	**20.65**	**34.41**

Input Length Finetuning results on SCITLDR-AIC for different input sequence length.

Conclusion

- Intermediate task training benefits domain adaptation
- Additional benefits can be observed by filtering the filtering the intermediate training corpus to best match the target domain/task
- Using a scientific domain Pretrained Language Model (SciBERT) does not result in additional gains. In fact, it performs slightly worse on the SciTLDR dataset, with or without pretraining compared to BERT

Future Work

Explore different criteria for selective intermediate pretraining. Examine its benefits on both abstractive and extractive summarization.

References

- Yang Liu and Mirella Lapata, 2019. Text Summarization with Pretrained Encoders. In EMNLP-IJCNLP 2019
- Isabel Cachola, Kyle Lo, Arman Cohan, Daniel Weld, 2020. {TLDR}: Extreme Summarization of Scientific Documents. In Findings of the Association for Computational Linguistics: EMNLP 2020