Multi-XScience: A Large-scale Dataset for Extreme Multi-document Summarization of Scientific Articles

Yao Lu
Mila
University of Waterloo
lu.yao@ucl.ac.uk

Yue Dong
Mila / McGill University
yue.dong2
@mail.mcgill.ca

Laurent Charlin
Mila / HEC Montréal
Canada CIFAR AI Chair
lcharlin@gmail.com

A Model Implementation Details

All the models used in our paper are based on the open-sourced code released by authors. For all the models, we use the default configuration (model size, optimizer learning rate, etc) provided by the original implementation.

Here are the code bases we rely on to run all experiments. Unless otherwise specified, all the systems are developed based on PyTorch.

- LexRank (Python): https://github.com/crabcamp/lexrank
- TextRank (Python): https://github.com/summanlp/textrank
- HierSumm: https://github.com/nlpyang/hiersumm
- HiMAP: https://github.com/Alex-Fabbri/Multi-News
- BertAbs: https://github.com/nlpyang/PreSumm
- BART: https://github.com/pytorch/fairseq
- SciBertAbs: https://github.com/allenai/scibert
- Pointer-Generator (TensorFlow): https://github.com/abisee/pointer-generator

For all the models except BART, we truncate the total input length to a maximum of 1024 tokens. For BART, we take at most 512 tokens as input due to the model length restriction.

During the decoding process, we use beam search (beam size=4) with forbidding repeated trigrams setting, which is widely used in sequence-to-sequence models. We also set the minimal generation length to 110 tokens according to the dataset statistics.

B Computation Infrastructure

We use 4 NVIDIA V100 GPUs to run HierSumm, HiMAP, BertABS, SciBertAbs, and BART experiments. The experiment time varies from a few hours to at most 2 days. For the pointer-generator model, we use 1 NVIDIA V100 GPU to run for 1 day. For the other models including lead baseline, extractive oracle, LexRank, and TextRank, we run on CPUs for no more than half an hour.

C Computation of ROUGE score

We use the files2rouge[1] python package for the automatic ROUGE evaluation in our paper. Similar to CNN/Dailymail dataset, we adopt the anonymized setting of citation symbols for the evaluation. In our dataset, the target related work contains citation reference to specific papers with special symbols (e.g. cite_2). We normalize all these symbols to a standard symbol (e.g. cite) for the evaluation process.

[1] https://github.com/pltrdy/files2rouge