

## Summary

- 278K 955K 203K 231K 120K 215K 133K 1.34M 21.3K 19.5K -

summ tokens 55 31 24 70 23 242 208 117 243 607 6.9x

doc tokens 774 767 438 501 444 6446 3143 3573 1686 9409 8.3x

summ sents 3.8 1.5 1 5.3 1.4 6.3 7.1 3.6 7.1 21.4 3.7x

doc sents 29 31 19 27 22 251 102 143 42 300 6.5x

Compression token 14.8 31.7 19.7 7.2 18.4 41.2 16.6 36.3 12.2 18.7 1.4x

Compression sent 8.3 22.4 18.9 3.3 14.5 44.3 15.6 58.7 9.7 18.1 2.2x

Coverage 0.890 0.855 0.675 0.610 0.728 0.920 0.893 0.861 0.913 0.942 1.2x

Density 3.6 9.8 1.1 1.1 1.4 3.7 5.6 2.1 6.6 7.7 1.5x

### - 3.4 Fine-grained Analysis on ArXiv

To perform fine-grained human analysis on the arXiv benchmark, this survey implements a stratified random sampling strategy based on the 6 different categories of scientific domains contained in the arXiv.org scientific repository: physics (ph), computer-science (cs), mathematics (math), quantitative-biology (q-bio), quantitative-finance (q-fin) and statistics (stat).

- Lastly, to summarize the limitations of a paper often requires more external knowledge outside of the content related to the source document and whether current summarization models are able to infer such knowledge from the benchmark dataset is an interesting study left for future works.

## 4 MODELS

### 4.1 Overview

The following describes the differences between the extractive, abstractive and hybrid summarization approaches and the general taxonomy of a summarization system.

- The works in automatic text summarization research are traditionally classified into three different summarization approaches: (i) the extractive approach that involves direct extraction of salient fragments such as sentences of the original documents into a summary [15,38], (ii) the abstractive

approach imitates human behavior of paraphrasing important parts of a document into a summary [101,103] and (iii) the hybrid approach that attempts to combine the best of both approaches

- January 2022. An Empirical Survey on Long Document Summarization: Datasets, Models and Metrics 1:11

Important sentence Less Important sentence Unimportant sentence Word/Sentence/Section Attention Mechanism R N N R N N R N N X1 X2 Xn Encoder... R N N R N N R N N Y1 Y2 Yn Decoder C1 Cn-1...

R N N R N N R N N Y1 Y2 Yn Decoder... Extractive Abstractive Context vector C

C0 Classifier Predict/generate A - Classic Graph (Extractive) B - RNN-based Model Architecture C

- Transformers Softmax Linear Add & Norm Feed Forward Add & Norm Multi-Head Attention Add &

Norm Masked Multi-Head Attention Output Embedding Outputs (Shifted right) Add & Norm Feed

Forward Add & Norm Multi-Head Attention Input Embedding Inputs N x N x Output Probabilities

Positional Encoding Positional Encoding

Fig.

- Im-

portantly, while there are other classical architectures [38,112], the graph architecture is worth a separate mentioning here due to the fact that (a) it remains as a strong baseline against other advanced architectures, (b) it can effectively incorporate external knowledge as an inductive bias to the calculation of the importance of a sentence and (c) it achieves state-of-the-art result in long document unsupervised extractive summarization setting when integrated with current state-of-the-art pre-trained models [25,69].

- Consequently, LoBART [77] proposes a hybrid summarization system that completes a summary generation in two separate steps, (i) content selection: using a multi-task RNN, select salient content from the original source document until the total text output reaches the limit of the sequence-to-sequence pre-trained BART model and (ii) abstractive summarization: summarize the carefully selected subset using a pre-trained BART model with efficient transformer mechanism.