Summaformers @ LaySumm 20, LongSumm 20

Scientific Document Summarization for LaySumm '20 and LongSumm '20

Sayar Ghosh Roy[1], **Nikhil Pinnaparaju**[1], **Risubh Jain**[1], **Manish Gupta**[1, 2] **& Vasudeva Varma**[1]
iREL, IIIT Hyderabad[1], Microsoft, India[2]

Abstract

Automatic text summarization has been widely studied as an important task in natural language processing. Summarization is a cognitively challenging task – extracting summary worthy sentences is laborious, and expressing semantics in brief when doing abstractive summarization is complicated. In this paper, we specifically look at the problem of summarizing scientific research papers from multiple domains. We differentiate between two types of summaries, namely, (a) LaySumm: A very short summary that captures the essence of the research paper in layman terms restricting overly specific technical jargon and (b) LongSumm: A much longer detailed summary aimed at providing specific insights into various ideas touched upon in the paper. While leveraging latest Transformer-based models, our systems are simple, intuitive and based on how specific paper sections contribute to human summaries of the two types described above. Evaluations against gold standard summaries using ROUGE [3] metrics prove the effectiveness of our approach. On blind test corpora, our systems rank first and third for the LongSumm and LaySumm tasks respectively.

Introduction

Text Summarization is cognitively challenging, extracting summary worthy sentences and expressing semantics in brief is laborious. Scientific papers are large, complex documents that tend to be geared towards a particular audience which is often a very small percentage of the population. Length of such documents often spanning several pages demand a great deal of time and attention. Considering scientific research papers from multiple domains, we differentiate between two types of summaries:

- LaySumm: A very short summary that captures the essence of the research paper in layperson terms restricting overly specific technical jargon
- LongSumm: A much longer detailed summary aimed at providing specific insights into various ideas touched upon in the paper

Dataset

LaySumm Dataset

A dataset of 572 research papers and corresponding gold standard lay-summaries were available for training, 84 tokens being the average length of a summary. A set of 37 research papers were provided as the blind test data. The LaySumm dataset contains papers from a variety of domains (epilepsy, archeology, and materials engineering). Elsevier made available a set of lay summaries of papers from a multidisciplinary collection of journals, as well as their abstracts and full-texts.[1].

LongSumm Dataset

The corpus for this task includes a training set that consists of 1705 extractive summaries, and 531 abstractive summaries of scientific papers in the domains of Natural Language Processing and Machine Learning. The extractive summaries are based on video talks from associated conferences while the abstractive summaries are blog posts created by NLP and ML researchers. The average gold summary length was 767 tokens. The research papers were parsed using the science-parse[2] library. A collection of PDFs of 22 research papers served as the blind test set.[3].

Proposed Models

Scientific research papers are fairly structured documents containing standard sections like: abstract, introduction, background, related work, experiments, results, conclusion and acknowledgments. Models meant for processing such documents should therefore be aware of such sectional structure [1]. A simple way of achieving that would be to pick a few sentences from each of the sections to be a part of the summary. However the following questions would need to be addressed:

- How do we decide how many sentences to pick from each section?
- Which sentences to pick from a particular section?
- Can we rewrite sentences so as to obtain a concise and coherent abstractive summary?

LaySumm

From section-contribution studies, we see that 'abstract' was the most significant section followed by 'conclusion' (see Figure 1). High ROUGE-L overlap indicates some degree of verbatim copying from 'abstract' onto lay-summary. With the 'conclusion' section, we see high ROUGE overlap + relatively shorter section length indicating that it contains a great degree of useful information in a more condensed fashion.

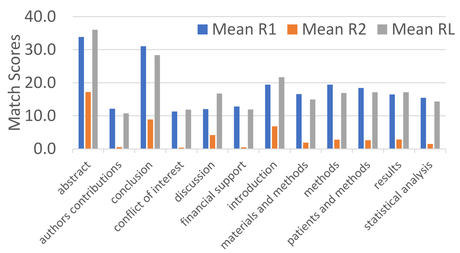

Figure 1: ROUGE-1, ROUGE-2 and ROUGE-L overlaps between paper sections and LaySumm summary

For LaySumm, we leveraged pre-trained Transformer models for conditional generation given a set of individual sections. Our results indicate that using 'abstract' as the only sequence for conditional generation is a better choice as compared to utilizing more sections. Therefore, the problem at hand is one of capturing salient information as one would expect from a summarization task, with the additional flavor of text style transfer.

LongSumm

For our summary generation architecture, we considered one section at a time without the global context based on existing scientific evidence. [5] We found that 'introduction' was the most important section followed by 'related work' and 'results' (see Figure 2). We utilized SummaRunner, a neural extractive text summarization architecture [4] as our section-level summarizer. We pre-trained SummaRuNNer on the PubMed [1] dataset to generate paper abstracts as closely as possible from various paper sections. Based on section-contribution evaluations, we constructed a budget module to calculate how much weight to assign each section for the purpose of combining section summaries into the final long-summary. Figure 3 illustrates the broad architecture of our proposed system.

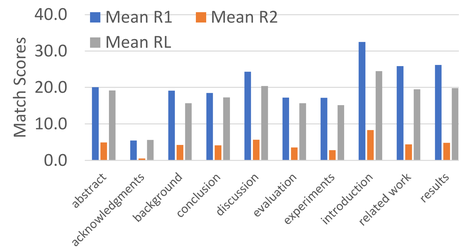

Figure 2: ROUGE-1, ROUGE-2 and ROUGE-L overlaps between paper sections and LongSumm summary

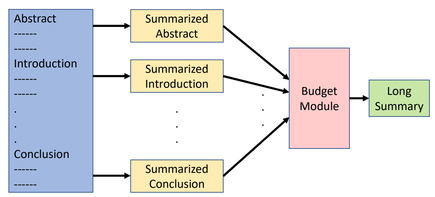

Figure 3: System Architecture for LongSumm

Results

On blind test corpora of 22 and 37 papers, our proposed systems achieve ROUGE R1 scores of 49.46 and 45.94, ranking first and third for LongSumm and LaySumm tasks respectively.

Method	R-1 F1	R-2 F1	R-L F1
Lead-150 baseline	40.85	17.40	25.01
(abs)+SummaRuNNer	39.89	16.30	24.44
(abs)+T5-base	40.74	15.29	24.13
(abs+conc)+T5-base	40.99	15.21	23.72
(abs+conc+intro)+T5-base	40.94	15.32	23.49
(SummaRuNNer)+BART$_L$	40.87	14.72	24.31
(small abs)+BART$_L$	44.81	18.76	26.71
(abs+conc)+BART$_L$	45.45	**19.22**	27.24
(abs+conc+intro)+BART$_L$	45.69	19.07	27.17
(abs+conc+intro+methods)+BART$_L$	45.61	18.95	27.05
(abs)+BART$_L$	**45.94**	19.01	**27.43**

Table 1: LaySumm Results (Best results are highlighted in bold)

Method	R-1 F1	R-2 F1	R-L F1
Section cutoff at R-1=10.0	47.18	14.10	18.37
Section cutoff at R-1=17.5	49.20	16.49	21.03
Section cutoff at R-1=20.0	48.93	16.57	21.07
Section cutoff at R-1=20.0 + Post-Proc	**49.46**	**16.86**	**21.42**

Table 2: LongSumm Results (Best results are highlighted in bold)

Conclusion

In this paper, we studied two scientific document summarization tasks, namely, LaySumm and LongSumm. We experimented with popular text neural models in a section-aware manner. Our results indicate that modeling of the document structure with strong focus on which parts of a research paper to attend to while composing a summary gives a significant boost to the quality of the resultant output. On blind test corpora, our systems rank first and third for the LongSumm and LaySumm tasks respectively.

References

[1] Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. A discourse-aware attention model for abstractive summarization of long documents. *arXiv preprint arXiv:1804.05685*, 2018.

[2] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, 2019.

[3] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. page 10, 01 2004.

[4] Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. *arXiv preprint arXiv:1611.04230*, 2016.

[5] Wen Xiao and Giuseppe Carenini. Extractive summarization of long documents by combining global and local context. *arXiv preprint arXiv:1909.08089*, 2019.

[1] https://github.com/WING-NUS/scisumm-corpus/blob/master/README$_{Laysumm.md}$
[2] https://github.com/allenai/science-parse
[3] https://github.com/guyfe/LongSumm