COLING 2022 Shared Task: LED Finteuning and Recursive Summary Generation for Automatic Summarization of Chapters from Novels

Prerna Kashyap
AWS New Apps Initiatives
Amazon, Seattle, USA
kaprerna@amazon.com

Abstract

We present the results of the Workshop on Automatic Summarization for Creative Writing 2022 Shared Task[1] on summarization of chapters from novels. In this task, we finetune a pretrained transformer model for long documents called LongformerEncoderDecoder which supports seq2seq tasks for long inputs which can be up to 16k tokens in length. We use the Booksum dataset for longform narrative summarization for training and validation, which maps chapters from novels, plays and stories to highly abstractive human written summaries. We use a summary of summaries approach to generate the final summaries for the blind test set, in which we recursively divide the text into paragraphs, summarize them, concatenate all resultant summaries and repeat this process until either a specified summary length is reached or there is no significant change in summary length in consecutive iterations. Our best model achieves a ROUGE-1 F-1 score of 29.75, a ROUGE-2 F-1 score of 7.89 and a BERT F-1 score of 54.10 on the shared task blind test dataset.

1 Introduction

Condensing long novel chapters into succinct and easy to digest summaries could be helpful as an informative bookmark to serve as a reminder of what happened in the last read chapter. This is much harder than other summarization tasks like summarizing news articles (Nallapati et al., 2016; Grusky et al., 2018; Narayan et al., 2018) or legal (Sharma et al., 2019) and scientific documents (Cohan et al., 2018). The reason for this is two fold. Firstly, the importance of automatic summarization systems for these tasks is diminished by the presence of article headlines, highlights or abstracts, as well as the length of the text to be summarized which is limited to a few hundred words to a few pages. Secondly, due to shorter text length and fact heavy content, there is no scope for extensive paraphrasing in the summaries. This is also due to short ranged causal and temporal dependencies and absence of convoluted plot lines. On the other hand, the task of summarizing chapters from novels (Ladhak et al., 2020; Kryściński et al., 2021) introduces all these additional challenges, including processing of long texts, abrupt changes in plot lines, dialogue and narration, and generation of highly abstractive summaries.

We present a recursive summary of summaries approach inspired by (Wu et al., 2021), where we decompose the long novel chapters into paragraphs and summarize them separately, thereby reducing computational complexity and noise in the target summaries. This is also similar to the divide and conquer approach used in (Gidiotis and Tsoumakas, 2020). These partial summaries are then combined to obtain an intermediate summary. This intermediate summary is then treated as the long text to be summarized and this process is repeated until either a final summary of a specified length is obtained or there is no significant change in summary length between consecutive intermediate summaries. The model used to generate these summaries is a pretained LongformerEncoderDecoder model[2] (Beltagy et al., 2020) finetuned on paragraph alignments obtained from the novel chapters. The datasets used for finetuning are described in Section 2 and the models are presented in Section 3. We present our results and analysis in Section 4.

2 Dataset

Some key challenges in long form summarization are computational constraints and limits on input length of pretrained models used for finetuning. To address these challenges, instead of using entire chapter to summary mappings, we use paragraph level alignments obtained for the novel chapters.

[1] https://creativesumm.github.io/sharedtask

[2] https://github.com/allenai/longformer

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 19–23

https://doi.org/10.26615/978-954-452-056-4_003

The paragraph level alignments are computed between paragraphs extracted from chapters and individual sentences of chapter-level summaries, by leveraging paragraph-sentence similarity scores using a SentenceTransformer (Reimers and Gurevych, 2019) and a stable matching algorithm as mentioned in the Booksum paper (Kryściński et al., 2021).

We use two datasets for finetuning, one containing just the paragraph alignments (we will refer to this as Dataset 1) and another containing paragraph alignments along with a subset of the chapter to summary data with maximum chapter length constrained to 500 words (we will refer to this as Dataset 2). The maximum chapter length of 500 words is chosen because the maximum encoder and decoder length for the models is set to 512 and this ensures that a very small percentage of the total number of examples exceeds the maximum token length of 512 after tokenization. Before training, all chapter and summary text is cleaned by stripping away hyperlinks, multiple consecutive whitespaces and non ASCII characters. The number of examples for training and validation splits for both datasets can be seen in Table 1. Some statistics for the train and validation splits of both datasets after tokenization using the LED tokenizer can be found in Table 2 and Table 3 respectively. All lengths presented in the table are number of words in the text.

3 Models

The reference summaries for the novel chapters are highly abstractive with high semantic and low lexical overlap. The novel chapters have long range causal and temporal dependencies that can be effectively captured by the self attention component in transformers (Vaswani et al., 2017), which enables the network to capture contextual information. However, the memory and computational requirements of self-attention grow quadratically with sequence length, making it very expensive for longer texts like novel chapters. Longformer (Beltagy et al., 2020) is a modified transformer with a self-attention operation that scales linearly with the sequence length, making it a lucrative option for processing long sequences.

We use the led-base-16384[3] LongformerEncoderDecoder model for finetuning, which is initialized from bart-base[4] (Lewis et al., 2019) since both models share the exact same architecture. We finetune the pretrained LED base model on the two datasets mentioned in Section 2 for 10 epochs and evaluate on the validation split after every 3000 steps. Model outputs are decoded using beam search with 2 beams and n-gram repetition blocking for n > 3. The LED config min and max length is set to 100 and 512 respectively, with a length penalty of 2.0, early stopping set to False and a batch size of 1 due to computational constraints. The maximum encoder and decoder length is set to 512.

In addition to the usual attention mask, LED can make use of an additional global attention mask defining which input tokens are attended globally and which are attended only locally, just as in the case of Longformer. We follow recommendations of the paper (Beltagy et al., 2020) and use global attention only for the very first token and we ensure that no loss is computed on padded tokens by setting their index to -100. We also disable gradient checkpointing and the caching mechanism to save memory. We use the ROUGE metric (Lin, 2004) for evaluation during model training and validation.

4 Experiments and Results

We train two models on Dataset 1 and 2 (referred to as Model 1 and Model 2 respectively) and choose the model with the best overall validation score for final submission for the shared task. The mid ROUGE F-1 scores on the validation set of Dataset 1 and 2 for both models can be found in Table 4.

For the final submission for the shared task, the final summaries for input novel chapters are generated using the recursive summary of summaries method described in the previous sections. The novel chapters in the blind test dataset are divided into paragraphs not exceeding 400 words in length, with an overlap of one sentence per chunk. This means that the last sentence from the previous paragraph chunk becomes the first sentence of the new paragraph chunk. If addition of any sentence to a chunk exceeds the chunk size of 400 words, that sentence becomes a part of the next chunk. These chunks are then summarized separately and concatenated in a recursive fashion to get the final summary. We observe that the training data has a mean summary to chapter length ratio of 0.15 and a standard deviation of 0.20 (where length is considered

[3]https://huggingface.co/allenai/led-base-16384

[4]https://huggingface.co/facebook/bart-base

Dataset	Train split	Val split	# of unique train books	# of unique val books
Dataset 1	13720	2334	53	11
Dataset 2	14208	2486	82	15

Table 1: Training and validation splits for datasets used.

Dataset	Mean-article-len	Mean-summary-len	%-article-len > 512	%-summary-len > 512
Dataset 1	204.77	40.00	0.03	0.00
Dataset 2	213.04	45.31	0.04	0.00

Table 2: Train split stats after tokenization for datasets used.

Dataset	Mean-article-len	Mean-summary-len	%-article-len > 512	%-summary-len > 512
Dataset 1	189.99	38.88	0.02	0.00
Dataset 2	205.65	43.53	0.04	0.00

Table 3: Validation split stats after tokenization for datasets used.

to be number of words in the text). So, during the generation of summaries by the finetuned model, we keep the maximum predicted summary length to be 35% of the input chapter length i.e. mean plus standard deviation of the summary to chapter length ratio of training datasets. This means that the input text is decomposed into paragraphs and intermediate summaries are created by generating individual summaries of these paragraphs and concatenating them until summary length of atmost 35% of the input text is reached or consecutive intermediate summary lengths are within 200 words of each other. The final evaluation metrics of the best performing model i.e. Model 1 on the shared task's blind test set can be found in Table 5.

4.1 Qualitative Analysis

A brief qualitative analysis of the predicted summaries in comparison to the reference summaries yields a few important observations. Examples of reference and predicted summaries from the Booksum validation dataset using Model 1 and recursive summary generation can be seen in Table 6. The highlighted portions of both summaries indicate the semantically relevant parts and it is evident that the predicted summary manages to capture most of important information from the chapter accurately. The text presented in red color in the predicted summary section indicates grammatically or factually inaccurate sentences in the summary, which accounts for a small percentage of the overall predicted summary. One problem that the model generated summaries frequently suffer from is repetition (which often results in nonsensical sentences) as seen in Example 3 in Table 6. The model generated summaries also lack coherence due to generation of summaries of independent paragraphs.

4.2 Model Limitations and Future Work

Due to computational constraints, the full power of the LED model, in which input length up to 16k tokens can be used, could not be leveraged and the encoder decoder maximum length was limited to 512 tokens. Also, the abstractive nature of the reference summaries makes lexical overlap measured by ROUGE (Lin, 2004) an inadequate metric for model evaluation and can be substituted with BERTScore (Zhang* et al., 2020) or SummaQA (Scialom et al., 2019) which leverage pretrained neural models.

Deciding the maximum length of the final summary for input text during the recursive summary generation method is also a problem, because the current value of 35% of input text length might throw away important information or incorporate information that isn't too relevant. Other methods apart from plain concatenation to generate a final summary using intermediate summaries can also be considered a part of future work.

Acknowledgements

We would like to thank our colleagues Stephen Vilke, Mark Andress, Bobby Williams and Monica Hu, who helped in reviewing and assisting in the writing of this paper and facilitating more research opportunities within the team.

Model	Dataset	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-LSUM
Model 1	Dataset 1	19.13	4.83	13.51	13.52
Model 2	Dataset 1	18.41	4.60	13.20	13.20
Model 1	Dataset 2	19.46	4.85	13.60	13.97
Model 2	Dataset 2	18.69	4.58	13.25	13.25

Table 4: Validation scores for finetuned models on validation splits of all datasets used.

ROUGE-1	ROUGE-2	ROUGE-L	BERTScore-F1	SummaCZS	LitePyramid-l3c
29.75	7.89	28.33	54.10	15.62	6.91

Table 5: Final evaluation metrics for Model 1 on shared task blind test set.

References

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv:2004.05150*.

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)*, pages 615–621, New Orleans, Louisiana. Association for Computational Linguistics.

Alexios Gidiotis and Grigorios Tsoumakas. 2020. A divide-and-conquer approach to the summarization of long documents. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:3029–3040.

Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 708–719, New Orleans, Louisiana. Association for Computational Linguistics.

Wojciech Kryściński, Nazneen Rajani, Divyansh Agarwal, Caiming Xiong, and Dragomir Radev. 2021. Booksum: A collection of datasets for long-form narrative summarization.

Faisal Ladhak, Bryan Li, Yaser Al-Onaizan, and Kathleen R. McKeown. 2020. Exploring content selection in summarization of novel chapters. *CoRR*, abs/2005.01840.

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

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gulçehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, Berlin, Germany. Association for Computational Linguistics.

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, Brussels, Belgium. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Thomas Scialom, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. 2019. Answers unite! unsupervised metrics for reinforced summarization models. *CoRR*, abs/1909.01610.

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

Jeff Wu, Long Ouyang, Daniel M. Ziegler, Nissan Stiennon, Ryan Lowe, Jan Leike, and Paul Francis Christiano. 2021. Recursively summarizing books with human feedback. *ArXiv*, abs/2109.10862.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Chapter and Book	Reference summary	Predicted Summary
Antony and Cleopatra, Act 2, Scene 1	Messina. Pompey's house. Pompey discusses strategy with his men, Menas and Menecrates, confidently assessing his fortunes. When Menas reports that Lepidus and Octavius Caesar are in the field against him, Pompey dismisses it: they wait in Rome, helpless without Antony. Pompey is counting on Antony to stay in Egypt with Cleopatra. Varrius enters with bad news: Antony has returned. Pompey is distressed, as Antony is by far the best soldier of the triumvirate. Menas hopes that Antony and Octavius will not be able to work together, due to the fact that Antony's brother and late wife warred against Octavius, but Pompey points out that natural enemies may band together against a common threat.	Messina. Pompey's house. Menecrates and Menas discuss the fact that Antony and Cleopatra are in Egypt together. Menas tells Pompey that he cannot hope Caesar and Lepidus to meet in Rome together, as Antony is already in Rome. Menass tells him that he has a message for Antony, and he tells him he will deliver it in a day or two. He tells Menas that he will give Antony a room for further travel.
Much Ado About Nothing, Act 5, Scene 3	Claudio, Don Pedro and several other men visit Hero's gravesite and perform a short memorial service. Claudio has written an epitaph for Hero, after which he sings a song and then promises to perform the same ritual every year. Don Pedro bids the other men good night and takes Claudio with him to Leonato's house.	Claudio, Don Pedro, and the Musicians all go to Hero's tomb and sing a hymn to honor Hero's memory. Claudio and Don Pedro leave the tomb to bid farewell to Hero. The crowd sings a song to Hero, and then they all leave to go to Leonato to bid goodbye to the Hero. Don Pedro and Claudio leave to say a prayer to Hero to be buried in the morning. The rest of the group leaves to go and bid farewell Hero to Claudio.
Antony and Cleopatra, Act 4, Scene 4	The following day, Eros arms Antony for battle, and Cleopatra insists on helping. Antony feels confident about the coming fight, promising Cleopatra that anyone who attempts to undo his armor before he is ready to remove it and rest will confront his rage. An armed soldier enters and reports that a thousand others stand ready for Antony's command. Antony bids Cleopatra adieu, kisses her, and leads his men into battle.	Alexandria. Cleopatra's palace. Antony and Cleopatra arrive and Eros kisses her goodbye. The officer then kisses her and tells her to go to the port and come back. The soldier then leaves and the rest of the soldiers arrive. The soldiers then leave, and Antony kisses Cleopratra goodbye. He kisses her farewell and leaves. The officers then go off to battle, and he kisses her. The guards then leave and the soldiers go to battle.

Table 6: Examples of reference and predicted summaries from the Booksum validation dataset.