Does the structure of textual content have an impact on language models for automatic summarization?

Eve Sauvage[1,2], Sabrina Campano[1], Lydia Ould-Ouali[1], Cyril Grouin[2]

[1] EDF Lab Paris-Saclay, Palaiseau, France
[2]Université Paris-Saclay, CNRS, LISN, Orsay, France
Correspondence: eve.sauvage@lisn.upsaclay.fr

Abstract

Despite recent improvements, the processing of long sequences with *Transformers* models remains a subject in its own right, including automatic summary. In this work, we present experiments on the automatic summarization of scientific articles using BART models, considering textual information coming from distinct passages from long texts for summarization. We demonstrate that considering document structure improves the performance of state-of-the-art models and approaches the performance of LongFormer in English.

1 Introduction

Long texts are formatted with visual marking (such as paragraphs, sections, and so on) to help readers retrieve information quickly. These markers help skim the long documents and get a general idea of their content. Document skimming can be used to obtain an abstract of a document.

Automatic summarization has long suffered from the context limitation of Neural Networks (NN) models. Context limitation either restricts the possible size of the text given as input (with *Transformers* (Vaswani et al., 2017)) or the information retained during process (Hochreiter and Schmidhuber, 1997; Cho et al., 2014). *Transformers* consider the entire context to proceed with a given task. However, this processing of memory comes with a considerable calculation cost. That calculation cost is induced by the very mechanism that allows full information retention: a context memory that keeps the whole sequence in memory for processing one word. That computational cost is quadratic.

This computational complexity limits the first *transformers* models to sequences of 512 tokens. As calculation capacities improved, this limit was quickly pushed back from 512 to 1024 (Lewis et al., 2020) then 2048 and even going up to more than 200,000 tokens for the most recent LLM (*Large Language Model*)[1] (GPT-3, Mistral, Claude, *inter alia*). However, this progress comes at high costs in computing power and infrastructure. The cost of training basic models of the latest LLMs is estimated at around a million dollars (Chuan, 2020) and with a non-negligible carbon impact (Ludvigsen, 2023).

In parallel, approaches to reducing the computational complexity of the *transformers* architecture have been explored by research. In particular via alternatives to the *full attention* mechanism (the use of square matrices to model sequences) (Beltagy et al., 2020; Tay et al., 2020; Zaheer et al., 2021). Despite these improvements, the costs of training and inferring models remain high in terms of computing power.

These methods use textual data without the metadata that accompanies and structures them, but other solutions highlight the structured nature of long texts for their processing. Cohan et al. 2018; You et al. 2019 show an interest in taking into account the document structure (*i.e.* paragraphs and sections) in the processing of long texts and, in particular on the task of automatic summary.

From this hypothesis, we start to evaluate the impact of the document structure on the automatic summary of long text. We first present the context in which this study takes place. We will then discuss the methodology followed and the results obtained before concluding with the observations made.

The performance of our method approaches SOTA results for long contexts without modifying the structure of the models. A segmentation of tasks with a reflection on the construction of the writings could, therefore, allow a reduction in the costs necessary to obtain usable results.

[1] https://support.anthropic.com/en/articles/7996856-what-is-the-maximum-prompt-length

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop), pages 280–285
August 11-16, 2024 ©2024 Association for Computational Linguistics

2 Related Work

Summarization is a classic task of natural language processing. It is, therefore, particularly well documented and already has numerous methods and models.[2]

The first approaches to automatic summarization (Luhn, 1958) focus on so-called "extractive" methods. These methods consist of recovering the most important sentences from the text. However, they are criticized for their lack of readability.

The arrival of generative language models (Rush et al., 2015; Lewis et al., 2020; Raffel et al., 2020) has allowed "abstractive" methods to supplant extractive methods. These generative models arrive with the non-sequential text processing approach proposed by Vaswani et al. 2017 in the *Transformers* architecture. This methodology responds to one of text processing challenges using neural network architectures: information retention. The *transformers* models thus make it possible to improve context consideration.

Hybrid methods emerge to get the most out of extractive and abstractive methods. These methods, aimed at streamlining the result of extractive summaries thanks to the abstractivity of generative language models, are particularly effective for long texts due to the simplicity of their operation. They allow a reduction in calculation costs by only selecting the relevant sentences from the texts to be given to the generative model (Giarelis et al., 2023; Li and Gaussier, 2022).

Among hybrid methods, approaches based on text structure for processing long texts have been proposed (Cohan et al., 2018) using graph neural networks (GNN) to organize the hierarchy of sections. These methods make it possible to increase the performance of the models significantly. Other studies show the potential that the use of metadata can have in the processing of long texts (Xu et al., 2020; Ruan et al., 2022).

Abstractive methods remain the most used because they avoid going through a pipeline (while hybrid methods need the choice of the extractive method and an appropriate generative model).

3 Method

To show the impact of document structure on summarization, we select different specific parts of the text as input for abstractive models. We then compare the summaries produced by a model with a sub-selection of the document with the reference summary produced by a human editor. This approach is a hybrid method combining extractivity in the selection of relevant parts of texts and abstraction using generative language models.

Model	Input Data
BART BARTxiv	first 1024 tokens of the article
LongFormer	first 16,000 tokens of the article
Bart BARTxiv	first sentences of each section
	last sentences of each section
	introduction and conclusion

Table 1: Configuration of the experiments carried out. The results with the *baseline* models are carried out with the first three experiments and compared with the results obtained when taking into account the context as input to the models for Bart, BARTxiv.

We select several fragments of the text (see Table 1) based on the visualization of human writer usage of document structure (*cf.* Figure 1). This visualization corroborates the hypothesis of Dong et al. 2021 that information is mainly contained in textual units (paragraph or text) borders. We evaluate several configurations based on our observation to compare the different results.

Models selection We want to compare the performances of a model adapted to long sequences with those of a "classic" model with a shorter context window. As the LongFormer[3] model of Beltagy et al. 2020 is based on the smaller model Bart, it is particularly suitable for this comparison.

Bart is an auto-encoder *transformer* model built according to the architecture proposed by Lewis et al. 2020. Its context window is limited to 1024 tokens, much smaller than the scientific articles in the corpus treated here. The BART model for automatic summarization was trained on the CNN/Daily Mail corpus, bringing together English-written press articles and their summaries.

The LongFormer model modifies the attention of Bart to obtain a linear complexity on the size of the sequences and thus allows the processing of texts beyond 1024 tokens while maintaining achievable calculation costs by current infrastructure.

[2] Approximately 1500 models for the task of automatic summary on huggingface

[3] https://huggingface.co/docs/transformers/model_doc/led

		ROUGE Score			BERTSCORE		
		Rouge 1	Rouge 2	Rouge L	Precision	Recall	F-score
LONGFORMER		**46.32±10**	**19.87±11**	**27.46±10**	**86.49±2**	**85.4±2**	**85.92±2**
COHAN ET AL. 2018		35.80	11.05	31.80	n/a	n/a	n/a
BART		28.61±8	8.51±6	17.17±6	85.65±2	80.63±2	83.05±2
BART (ours)	$1^{st} sentences$	28.67±8	8.67±6	17.32±6	85.58±2	80.93±2	83.17±2
	last sentences	**29.74±9**	**9.31±7**	**17.95±6**	**85.65±2**	**81.2±2**	**83.35±2**
	intro.&conclu.	**29.43±8**	**8.96±6**	**17.57±6**	**85.75±2**	**80.83±2**	**83.2±2**
BARTXIV		41.17±8	14.8±7	22.89±6	85.55±2	84.44±2	84.97±2
BARTXIV (ours)	$1^{st} sentences$	28.01±8	11.74±7	18.99±7	82.67±2	86.35±2	84.31±2
	last sentences	37.83±9	11.86±7	20.16±6	84.4±2	84.05±2	84.21±2
	intro.&conclu.	**42.16±8**	**15.45±8**	**23.06±6**	**85.42±2**	**84.99±2**	**85.18±2**

Table 2: Mean results of the scores ROUGE and BERTSCORE and their standard deviation (showed after the ± sign) per entry on the different experiments summarized in the table 1.

In order not to penalize the smallest models, we also used a BART model adapted to scientific texts, BARTXIV,[4] trained as LONGFORMER on the SCIENTIFIC-PAPERS corpus with 9 epochs and a learning rate of $1e^{-}6$.

Dataset used Despite the interest in using document structure for the processing of long texts (Wu et al., 2023), the number of corpora available is small. Here, we use the corpus of scientific texts SCIENTIFIC-PAPERS[5] made available by Cohan et al. 2018 for their study of the impact of document structure on automatic summary using LSTM.

This corpus combines articles in English from the article repository platforms ARXIV and PUBMED. The texts of the articles thus obtained are divided by section and cleaned of their summary. Having been used to train numerous models adapted to long sequences due to its specificity, this corpus is particularly suited to our task. A sub-selection of 2000 texts seemed sufficient to obtain representative results while limiting the impact of the calculations carried out for the experiment. We extracted these texts from the *test* of the ARXIV part of the corpus to avoid data contamination during the experiments and to maximize thematic coverage (the PUBMED articles being focused on research in medicine). This sub-corpus includes articles with an average of 32,600 subtokens and abstracts with approximately 969 subtokens, i.e., an input size 30 times larger than the context window available for models of type BART (Lewis et al., 2020).

4 Results

Although the results do not allow us to exceed the values of the ROUGE scores obtained with LONGFORMER (0.46 for LONGFORMER against 0.42 for ROUGE-1 for the BARTXIV model in the best configuration see table 2), the use of certain parts of the text improves the results compared to the simple truncation to the first tokens of the texts. The results obtained (see Table 2) with ROUGE and BERTSCORE show the usefulness of targeting the processing of models according to the structure of the texts for long texts.

This improvement is noticeable when the model is not adapted to the type of text processed (like BART, see Table 2). In this case, selecting only the last sentences of each paragraph or the introduction and conclusion seems to be an exciting configuration to improve the automatic summary results of these models (improvement by 1 points for ROUGE-1 in the case of BART).

In Figure 1, we can observe the parts of the document that are the most used in the abstract depending on the abstract given. The yellow concentration shows that the most overlapping parts for human redactors are the end of the introduction and the end of the conclusion. This confirms the hypothesis of Dong et al. 2021 and advocates for prior text reduction based on text structure for summarization.

To obtain the best ROUGE or BERTSCORE scores, the distribution profiles of the automatically generated summaries must come as close as possible to those obtained by human summaries. That is to say, obtaining a maximum n-gram overlap at the end of the introduction and the end of the conclusion (see figure 1-reference summary). This

[4] https://huggingface.co/kworts/BARTxiv

[5] https://github.com/armancohan/long-summarization/tree/master

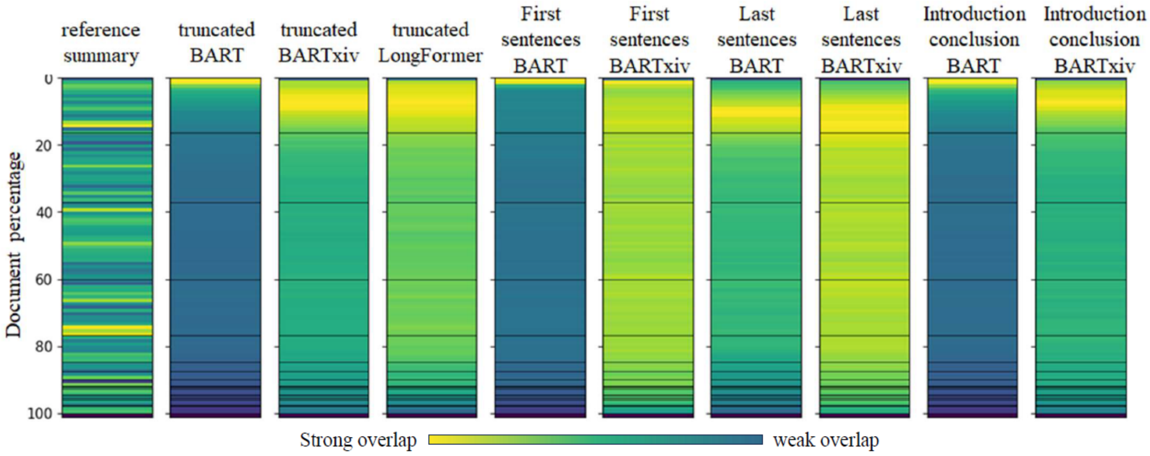

Figure 1: Position in the article of the n-grams also used in the summary (in terms of percentage). Strong overlap is showed in shade of yellow. Overlapping of 1-2-3-grams between the article and the summaries are summed at their position in terms of percentage of the text (0% corresponds to the start of the text). The average position of the section boundaries is represented by black lines.

representation also makes it possible to evaluate the impact of fine-tuning on selecting information in texts for automatic summarization. Having been trained on a corpus of press articles, BART tends to concentrate on the first sentences of the texts (see Figure 1-BART except for BART-last sentences)

This profile differs from the distribution of information retrieved by the human editor for the summary of scientific articles. It seems better modeled by BARTXIV and LONGFORMER. Indeed, in the profiles obtained with BARTXIV, we can see that the maximum overlap is shifted downwards compared to the profiles of BART (figure 1-BARTxiv).

5 Conclusion

We were able to show here that the use of structure by human authors in writing a summary is poorly imitated by the models even when they have access to most of the text to select information. Humanly produced summaries remain highly abstract compared to the language models targeted here. This particularity reinforces our hypothesis that taking structure into account could allow the creation of better summaries by the models.

Using hybrid models improves the results ROUGE or BERTSCORE of un-fine-tuned limited context window models and allows an alternative to more attention-expensive models. However, this limitation of inputs loses its interest with fine-tuned models whose learning conditions the position of the information retrieved. In addition to adapting the models' vocabulary, fine-tuning allows the models' attention to be focused on certain parts of the texts.

This observation is specific to the automatic summary task and requires additional analysis to verify its generalization to other tasks.

Using hybrid models based on the document structure of texts is an interesting approach when using a limited context window model that is not fine-tuned to the target data. However, access to the entire context allowed by the LONGFORMER architecture remains more efficient for automatic text summarization. These observations confirm the importance of the search for an alternative to the full attention mechanism of *transformer* architectures, which are costly in computational terms. To this end, we plan to implement a new representation of texts.

Limitations

The part of the documents used by human redactors to write the summary may strongly be linked with the document type. In this study, we only review scientific articles which might bias our conclusion. Distinct text parts may be used for other kinds of documents, such as press articles or books. Furthermore, scientific articles often follow a strongly constrained writing style, which may influence our results.

Most of the time, the same authors write the document and the abstract for scientific papers. This particularity is not shared over all document types and can lead our conclusion to a certain angle.

A qualitative review of the generated summaries may be conducted to determine whether the difference in scores fits with human appreciation of the summaries.

The model used for this study were fine-tuned for the summarization task. Using LLMs which have more general capacities in term of language generation may show different results. A comparison with SoTA LLMs should be conducted to assess the contribution of our experiments. However, models fine-tuned on a specific tasks often show better results than general LLMs on the same task. Preliminary work were done on this topic which seem to confirm this statement.

References

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The Long-Document Transformer. *arXiv preprint*. ArXiv:2004.05150 [cs].

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. *arXiv preprint*. Version Number: 3.

Li Chuan. 2020. OpenAI's GPT-3 Language Model: A Technical Overview. https://lambdalabs.com/blog/demystifying-gpt-3.

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.

Yue Dong, Andrei Mircea, and Jackie Chi Kit Cheung. 2021. Discourse-Aware Unsupervised Summarization for Long Scientific Documents. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1089–1102, Online. Association for Computational Linguistics.

Nikolaos Giarelis, Charalampos Mastrokostas, and Nikos Karacapilidis. 2023. Abstractive vs. Extractive Summarization: An Experimental Review. *Applied Sciences*, 13(13):7620. Number: 13 Publisher: Multidisciplinary Digital Publishing Institute.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780. Place: US Publisher: MIT Press.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Minghan Li and Eric Gaussier. 2022. BERT-based Dense Intra-ranking and Contextualized Late Interaction via Multi-task Learning for Long Document Retrieval. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, pages 2347–2352, New York, NY, USA. Association for Computing Machinery.

Kasper Groes Albin Ludvigsen. 2023. The Carbon Footprint of ChatGPT. https://towardsdatascience.com/the-carbon-footprint-of-chatgpt-66932314627d.

Hans. Peter. Luhn. 1958. The Automatic Creation of Literature Abstracts. *IBM Journal of Research and Development*, 2(2):159–165.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *arXiv preprint*. ArXiv:1910.10683 [cs, stat].

Qian Ruan, Malte Ostendorff, and Georg Rehm. 2022. HiStruct+: Improving Extractive Text Summarization with Hierarchical Structure Information. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1292–1308, Dublin, Ireland. Association for Computational Linguistics.

Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 379–389, Lisbon, Portugal. Association for Computational Linguistics.

Yi Tay, Mostafa Dehghani, Samira Abnar, Yikang Shen, Dara Bahri, Philip Pham, Jinfeng Rao, Liu Yang, Sebastian Ruder, and Donald Metzler. 2020. Long Range Arena : A Benchmark for Efficient Transformers.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Lingfei Wu, Yu Chen, and Kai Shen. 2023. *Graph Neural Networks for Natural Language Processing: A Survey*. Foundations and Trends in Machine Learning Series. Now Publishers.

Jiacheng Xu, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Discourse-Aware Neural Extractive Text Summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5021–5031, Online. Association for Computational Linguistics.

Yongjian You, Weijia Jia, Tianyi Liu, and Wenmian Yang. 2019. Improving Abstractive Document Summarization with Salient Information Modeling. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2132–2141, Florence, Italy. Association for Computational Linguistics.

Manzil Zaheer, Guru Guruganesh, Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. 2021. Big Bird: Transformers for Longer Sequences. *arXiv preprint*. ArXiv:2007.14062 [cs, stat].