Enhancing News Highlights Extraction through Abstractive Context

Saeedeh Javadi
Politecnico di Torino
s301409
Torino, Italy
s301409@studenti.polito.it

Zahra Karimi
Politecnico di Torino
s302612
Torino, Italy
s302612@studenti.polito.it

Fatemeh Ahmadvand
Politecnico di Torino
s301384
Torino, Italy
s301384@studenti.polito.it

***Abstract*—This paper presents a comprehensive study on text summarization, focusing on the THExt method—a highlight extraction approach built upon the BERT model. The model is finetuned by using the XSum news and CNN Daily Email datasets to assess its performance in the news domain, with evaluation based on ROUGE scores. Additionally, we propose an innovative modification to the THExt approach, incorporating an abstractive summary by using Longformer and T5-Small models to enhance the context for the THExt model.**

Our experimental evaluations demonstrate the positive impact of these modifications. By extending and finetuning the THExt method on the XSum and CNN Daily Email datasets, along with introducing novel enhancements, this research significantly contributes to the advancement of highlight techniques. Moreover, we highlight the possibilities for further progress by exploring abstractive summary and enhanced contextual information. the code of our implementations at the following link: GitHub.

I. INTRODUCTIION

Text summarization plays a crucial role in the field of natural language processing (NLP) by distilling essential information from extensive text collections into concise and coherent summaries. As the digital landscape continues to produce an overwhelming amount of textual content, the need for efficient summarization methods becomes increasingly vital. News summarization, a specialized application of text summarization, specifically focuses on extracting key details from news articles, offering readers a condensed and comprehensive overview within limited time constraints. By automatically extracting crucial information, news summarization facilitates time-saving, enhance information retrieval, and supports quick decision-making processes.

This research presents a novel approach to text summarization by fine-tuning the THExt model using both the XSum and the CNN Daily Email datasets. In the second approach, We introduce a significant modification by utilizing abstractive summary as the model's context, enabling a more enriched representation of the source text. To enhance the context further, we incorporate techniques such as T5-small [9] and Longformer models.

These enhancements aim to improve the selection of final highlights during the summarization process, consequently enhancing the quality and relevance of generated summaries. Through extensive experimentation and evaluation, we demonstrate the effectiveness of our approach and its potential to advance text summarization techniques.

Our study focuses on the development of robust functionality for data preprocessing and organization in text summarization tasks. We employ various techniques, including dataset parsing, structural organization, text cleaning, and regression label computation using Rouge scores to achieve this. Leveraging parallel processing techniques and popular libraries such as NLTK, spaCy, and py-rouge, we ensure efficient management and preprocessing of datasets, guaranteeing high-quality data and facilitating accurate evaluation of text summarization algorithms.

Additionally, we tackle the challenges of redundancy management and highlight extraction from scientific papers. By employing advanced techniques like sentence transformers and fine-tuned models, we enhance extracted highlights' diversity, quality, and relevance by effectively identifying and eliminating similar sentences. Automating the highlight extraction process significantly reduces the time required for researchers to comprehend and extract key findings from vast collections of scientific literature.

By addressing these critical aspects, our research contributes to the advancement of text summarization techniques, providing valuable insights into the impact of context modification and additional linguistic features for enhancing the performance of highlight extraction models.

II. RELATED WORKS

Significant progress has been made in automatic highlight extraction from scientific papers. La Quatra, Moreno, and Cagliero [1] proposed a transformer-based method for highlight extraction, leveraging the power of transformer models to capture contextual information and generate accurate highlights from scientific papers. Their approach showed effectiveness in extracting informative summaries.

BERT, introduced by Devlin et al. [2], has been widely utilized for language understanding tasks, including highlight extraction. Collins et al. [3] explored a supervised approach for extractive summarization of scientific papers. Additionally, Beltagy, Peters, and Cohan [4] introduced the Longformer model, designed for handling long documents and addressing traditional transformer limitations by introducing a more efficient attention mechanism. Furthermore, By inputting the text to be summarized, T5-Small can generate a concise summary that captures the essence of the original document. This is useful for applications such as news articles, scientific papers, and legal documents [9]. These studies collectively contribute to the advancement of accurate highlight extraction from scientific papers.

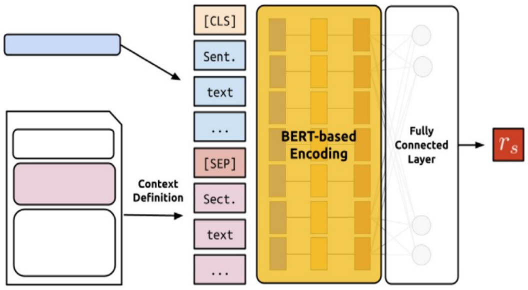


Fig.1: THExt architecture relying on transformer-based encoder

III. METHOD

A. Representing THExt Model Structure

The task of extracting highlights from text sources has recently gained significant attention. It aims to extend the benchmark methodology proposed by [1] to the news domain. Inspired by their THExt architecture (depicted in Fig. 1), this model incorporates a context definition block, BERT-based sentence encoding, and a regression module that utilizes fully connected neural networks. This regression module computes relevance scores for each candidate sentence, facilitating the extraction of highlights. To overcome the limitations of BERT, it adopts the Longformer architecture [4], as featured in THExt, and tailors it for the single document summarization task.

B. Fine-tuning with XSum and CNN Daily Email datasets

Moreno and Cagliero [1] employed three datasets for training: CSPubSumm (Computer Science), BIOPubSumm (Biology and Medicine), and AIPubSumm (Artificial Intelligence). This approach allowed for comprehensive training and evaluation across different knowledge domains. THExt is a method for extracting highlights from text, based on the BERT (Bidirectional Encoder Representations from Transformers) architecture. By fine-tuning the THExt model on the XSum and CNN Daily Email datasets, we intend to evaluate its performance in the news domain and demonstrate improvements compared to existing methods by using evaluation metrics such as ROUGE.

C. model

In our approach, we leveraged the power of T5-Small by providing context and finetuning the model on individual datasets. This process resulted in the extraction of the summary from the fine-tuned T5-Small model.

Next, we employed this extracted summary as contextual information for our THExt model, which was responsible for extracting highlights from the provided context. The inclusion of this additional contextual information enables the THExt model to benefit from the knowledge obtained during the T5-Small fine-tuning process.

With the new context in place, we evaluated the performance of the THExt model and assessed its ability to generate accurate and relevant highlights. Our experimental results shed light on the effectiveness of this approach and its potential in advancing the field of natural language processing and text summarization. This research provides valuable insights into the utilization of summaries as contextual information and serves as a foundation for further investigations into improved highlight extraction techniques using pre-trained language models like T5-Small.

- T5-Small (Text-to-Text Transfer Transformer): T5-Small is a versatile transformer-based model that can be fine-tuned for a wide range of text tasks, including summarization. It has shown promising results in various text-generation tasks. T5-small is a smaller, more efficient version of the T5 model with fewer parameters, suitable for limited resources. The full T5 model has higher performance but requires more computational power and memory. The choice depends on the use case and available resources.
- Longformer Decoder-Encoder (LED) : The Longformer Decoder-Encoder (LED) is a transformer-based model that combines the Longformer's attention mechanism with a decoder to enable bidirectional language modeling. It efficiently handles long documents by utilizing global and local attention strategies. LED is particularly effective for tasks requiring long-range context, such as document summarization and question answering. The LED model, which is an extension of the Longformer [4], was employed to create a new context. Specifically, the study performed an abstractive summarization of news text. This process resulted in a new context, which was then exploited by THExt architecture to generate new highlights.

IV. EXPERIMENTS

A. Datasets

- XSum: The Extreme Summarization (XSum) dataset is a dataset for evaluation of abstractive single-document summarization systems. The dataset consists of 226,711 news articles accompanied with a one-sentence summary. The articles are collected from BBC articles (2010 to 2017) and cover a wide variety of domains (e.g., News,

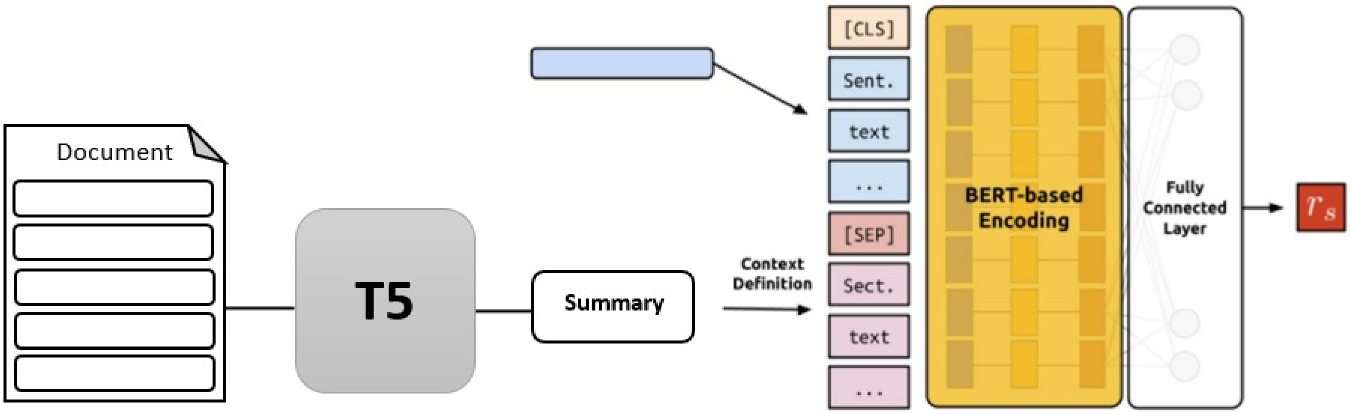
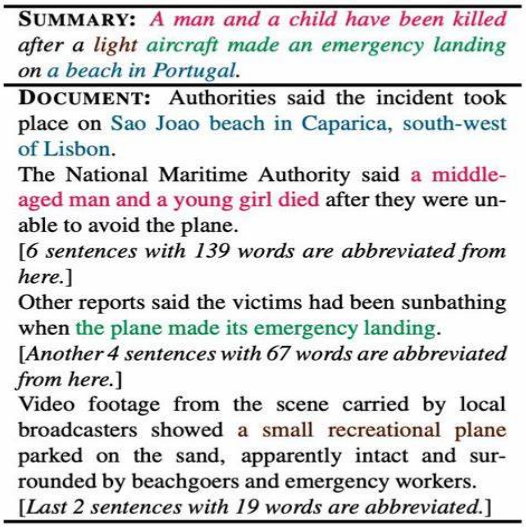


Fig. 1. T5 Context-summary builder

Politics, Sports, Weather, Business, Technology, Science, Health, Family, Education, Entertainment, and Arts). The document-summary pairs from XSum are typical of that in the news domain, where the source document represents a news article while the summary represents either a human-curated summary [7].



SUMMARY: *A man and a child have been killed after a light aircraft made an emergency landing on a beach in Portugal.*

DOCUMENT: Authorities said the incident took place on Sao Joao beach in Caparica, south-west of Lisbon.
The National Maritime Authority said a middle-aged man and a young girl died after they were unable to avoid the plane.
[*6 sentences with 139 words are abbreviated from here.*]
Other reports said the victims had been sunbathing when the plane made its emergency landing.
[*Another 4 sentences with 67 words are abbreviated from here.*]
Video footage from the scene carried by local broadcasters showed a small recreational plane parked on the sand, apparently intact and surrounded by beachgoers and emergency workers.
[*Last 2 sentences with 19 words are abbreviated.*]

Fig.3: An abridged example from an extreme summarization dataset showing the document and its one-line summary which one-line summary is the input of our model [7].

- CNN Daily Email: The CNN Daily Mail dataset is a collection of news articles from CNN and the Daily Mail, annotated with human-generated summaries. It is widely used for abstractive text summarization research and evaluation. The dataset contains pairs of news articles and highlights, providing valuable resources for training and evaluating summarization models.

B. Experimental Setup

The experimental setup involved training the extraction, abstraction, and reinforcement modules on a 500 GPU within the Google Colab pro+ environment. To overcome the resource limitations of Google Colab pro+, a subset of the original datasets was utilized. Each news article was subjected to highlight extraction, resulting in a fixed number of H=3 highlights. The training process encompassed 2 epochs, employing specific batch size=32 configurations to ensure comprehensive coverage and accurate model learning.

We wanted to use T5-long, which could use long sentence, but due to the lack of resources, it was not possible even with google colab pro+.

We performed separate fine-tuning processes for each model, T5-Small and Longformer, for 20 epochs and 8000 data from each of those datasets and archived them on the Hugging Face platform for accessibility and convenience for future use [10].

- saeedehj/t5-small-finetune-cnn
- saeedehj/led-base-finetune-cnn
- saeedehj/t5-small-finetune-xsum
- saeedehj/led-base-finetune-xsum

the extracted summary is given to the THExt model as the context to extract highlights. then check the performance of the model with the new context.

C. Evaluation metrics

we employ the Recall-Oriented Understudy for Gisting Evaluation (ROUGE) metric as an intrinsic evaluation measure based on syntax. ROUGE enables analysts to compare automatically generated summaries with a set of reference summaries, typically generated by humans. This metric assesses the degree of overlap between the text of the reference summary, which serves as the ground truth, and the text present in the automatically generated summary. By quantifying the unit overlaps between these texts, we can effectively evaluate the quality and effectiveness of our summarization system in capturing the essential information contained in the reference summaries. ROUGE provides a valuable means for assessing the performance and coherence of our summarization approach in relation to the desired summaries, aiding in the analysis and comparison of the automatically generated summaries [8]. In this article, we used Rouge-N and Rouge-L to evaluate the quality of the obtained results.

- Rouge-N: Overlap of n-grams between the system and reference summaries.
  - ROUGE-1 refers to the overlap of unigrams (each word) between the system and reference summaries.
  - ROUGE-2 refers to the overlap of bigrams between the system and reference summaries.

$$\text{ROUGE-N} = \frac{\sum_{S\in\{RefrenceSummeries\}} \sum_{gram_n\in S} Count_{match}(gram_n)}{\sum_{S\in\{RefrenceSummeries\}} \sum_{gram_n\in S} Count(gram_n)} \tag{1}$$

- ROUGE-L: Longest Common Subsequence (LCS) based statistics. Longest common subsequence problem takes into account sentence-level structure similarity naturally and identifies longest co-occurring in sequence n-grams automatically.

V. Result

The results of our study, as presented in the following tables I-III, were influenced by several limitations and challenges that need to be acknowledged. Firstly, due to limited computational resources, we were compelled to downsize the dataset used in our experiments. Instead of employing the Long-T5 model, we opted for T5-Small. This reduction in dataset size may have adversely affected the quality of our results, as we were unable to train and evaluate our model on the complete dataset.

Secondly, our abstractive context builder approach proved enhancement in capturing essential information from the documents and lead to improving results on highlight extraction. Despite the limitations, our method demonstrated competence in extracting key details.

Lastly, we firmly believe that enhancing our computational resources would significantly enhance the performance of our pipeline. Utilizing the full dataset for training purposes has the potential to substantially elevate the quality of the generated summaries. We consider our approach to be a valuable contribution to the field of summarization and a solid foundation for future research.

TABLE I
CNN Daily Email RESULTS

	THExt (baseline)			CNN Daily Email Finetuned		
ROUGE	R1	R2	RL	R1	R2	RL
Precision	0.371	0.100	0.245	**0.372**	0.061	**0.245**
Recall	0.171	0.045	0.110	**0.138**	0.022	0.091
F-Measure	0.225	0.060	0.146	0.197	0.032	0.130

TABLE II
XSum RESULTS

	THExt (baseline)			XSum Finetuned		
ROUGE	R1	R2	RL	R1	R2	RL
Precision	0.367	0.057	0.242	0.341	0.055	0.228
Recall	0.119	0.018	0.078	**0.148**	**0.023**	**0.098**
F-Measure	0.175	0.026	0.120	**0.200**	**0.032**	**0.133**

TABLE III
FINAL RESULTS ON USING ABSTRACTIVE CONTEXT ON XSUM AND CNN Daily Email

Dataset	XSum			CNN Daily Email		
ROUGE	R1	R2	RL	R1	R2	RL
Precision	0.352	**0.062**	0.237	**0.404**	**0.147**	**0.279**
Recall	**0.151**	**0.026**	**0.102**	**0.243**	**0.083**	**0.167**
F-Measure	**0.205**	**0.035**	**0.138**	**0.292**	**0.103**	**0.206**

VI. CONCLUSIONS

By the domain adaptation of La Quatra et al. [1] and integrating the Longformer model [4] and T5-Small [9] for extracting new context, this research presents a compelling approach to extract highlights from news articles. The results are particularly notable, especially when considering the abstractive summary as context. Furthermore, a careful examination of the extracted sentences reveals THExt's proficiency in capturing important highlights, making them valuable candidates for indexing or subtitles.

References

[1] La Quatra, Moreno, and Luca Cagliero. ”Transformer-based highlights extraction from scientific papers.” Knowledge-Based Systems 252 (2022): 109382

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018

[3] LE. Collins, I. Augenstein, S. Riedel, A supervised approach to extractive summarisation of scientific papers, in: Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017), 2017, pp. 195–205.

[4] Iz Beltagy, Matthew E Peters, and Arman Cohan. Longformer: The long document transformer. arXiv preprint arXiv:2004.05150, 2020

[5] Huan Yee Koh, Jiaxin Ju, Ming Liu, Shirui Pan.An Empirical Survey on Long Document Summarization: Datasets, Models and Metrics. arXiv:2207.00939v1, 2022

[6] Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In Text summarization branches out. 74–81

[7] Shashi Narayan, Shay B. Cohen, Mirella Lapata. Don't Give Me the Details, Just the Summary! Topic-Aware Convolutional Neural Networks for Extreme Summarization. arXiv:1808.08745v1, 2018.

[8] Lin, Chin-Yew. ”Rouge: A package for automatic evaluation of summaries.” Text summarization branches out. 2

[9] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P. J. (Year of publication). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer.arXiv:1910.10683v3, 2020.

[10] https://huggingface.co/saeedehj