Multi-document Summarization for News Articles Highlights Extraction

Matteo Berta
Politecnico di Torino
s295040@studenti.polito.it

Arcangelo Frigiola
Politecnico di Torino
s295406@studenti.polito.it

Francesco Marigioli
Politecnico di Torino
s296317@studenti.polito.it

Luca Varriale
Politecnico di Torino
s300795@studenti.polito.it

ABSTRACT

The extraction of highlights is a process which consists in selecting the salient sentences within the body of a test, which well summarize the meaning of the text under examination. This paper focuses on the problem of extracting highlights from news articles using transformer-based techniques.

The growing amount of news produced on a daily basis has led to information overload for consumers. Summarization techniques provide a solution to this problem by condensing long articles into a shorter and more comprehensive representation, in the form of highlights.

This paper reviews the state-of-the-art in the field of highlights extraction, and the evaluation metrics used to assess their performance. In particular, we exploit an already existent benchmark [6], with the goal of adapting it to the domain of multi-news. We also propose an ablation study, which aims to improve the baseline providing a generated context through the use of LED architecture [2]. The results obtained in the multi-document and in the ablation study, achieved respectively on two distinct datasets, have demonstrated the effectiveness of the model [6] in the field of multi-document summarization, and its adaptability to a different domain, as well as the possibility of its improvement by providing a broader context. Source code is available at shorturl.at/gzELZ.

I. Introduction

Text summarization is the process of automatically creating a shorter version of a longer text corpus, while retaining its most important information.

The growing amount of news available, and the vast choice that the user can have in selecting such information, has sparked and increased the interest in algorithms capable of extracting salient parts of such articles. Specifically, the availability of highlights allows the end user to select the news of interest to him more quickly, and discard the others.

Through extractive methods of summarization, we try to extrapolate the most salient sentences from the body of several newspaper articles relating to the same context or to the same news. This technique makes it possible to overcome typical problems of the abstractive approach, which instead could present the so-called phenomenon of hallucination, reporting qualitatively incorrect sentences.

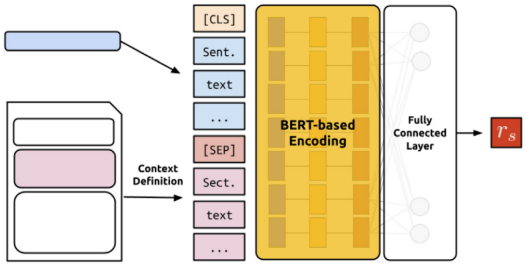

Fig. 1: THExt architecture relying on transformer-based encoder

To achieve this goal, we present an extension of the *Transformer-based Highlights Extraction* model (THExt in short), presented in [6]. Our task is to change the domain and provide a relatively short number of highlights from a corpus of news articles, clustered on the same topic. In order to overcome the limitations of BERT model in handling longer text corpus, we try to accomplish the task exploiting an alternative sentence encoder, also reported in [6], namely Longformer [2].

We conducted our experiments on a subset of the Multi-News dataset [1], as reported in Table I. The adaptation to a different domain and the results on a multi-document achieved valuable results in terms of ROUGE-1 and ROUGE-L F1-score, considering the evaluation on an abstractive ground truth. In addition, we present an ablation study on the benchmark model, which tries to put together the most performing context sources, according to [6] results. The scores obtained achieved interesting results on the same metrics cited above, if taking into account that the training step was operated on a tiny subset of the dataset (i.e. CSPubSumm [3], BIOPubSumm and AIPubSumm [4]), due to lack of resources.

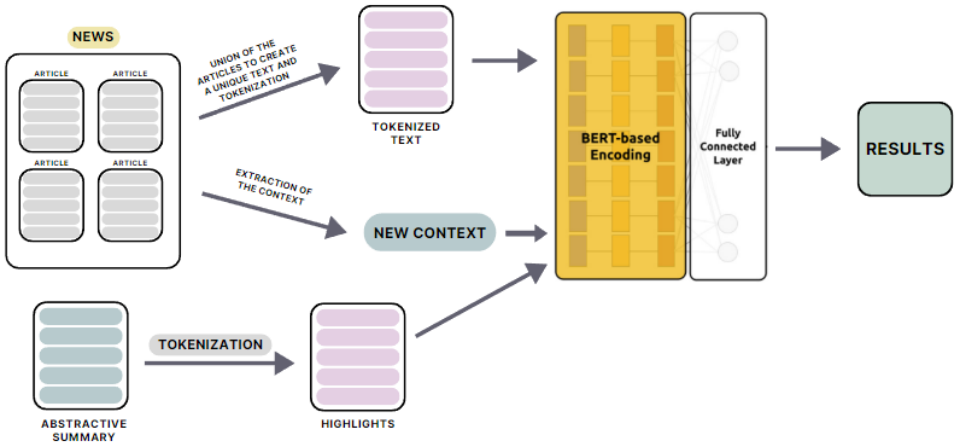

Fig. 2: How the adaptation of THExt on Multi-News is performed

II. Related Works

Highlights Extraction. The task of extracting highlights from text sources is a relatively unexplored and new category. We consider the work of La Quatra et al. [6], where they provide a new approach for highlights extraction, based on the BERT transformer model [5], a bidirectional encoder for text representation, that relies on the attention mechanism. As exposed in Fig.1, THExt architecture consists of a *context definition* block, a BERT based *sentence encoding*, a *regression* based on fully connected neural network. This latter component is responsible for the relevance score computation of each candidate sentence for the highlights extraction. We take this as benchmark to extend the work to the news domain.

Moreover, we exploit the Longformer architecture [2] reported in THExt to overcome the BERT limitations, and apply the model to a multi-document summarization task.

Multi-Document Summarization.

Multi-document summarization (MDS) is an essential task in natural language processing (NLP) that aims to generate a concise summary from multiple documents. Over the past few years, many neural network-based models have been proposed to solve this problem. Recent works utilize transformer models for multi-document summarization (MDS), like BERTSUM (Liu and Lapata, 2019) [7], which utilizes a pre-training strategy to improve the model's performance.

Another work that uses the transformers for MDS is the Multi-Document Longformers (MDLF) model proposed by Beltagy et al. (2020) [2]. The MDLF model extends the Longformer model to the multi-document setting by introducing a multi-level attention mechanism. These models have shown promising results in the MDS task, and the success of the Longformer model in several NLP tasks, such as question answering and text classification, further motivates its application to MDS.

This paper proposes a new Longformer-based model for MDS and evaluates its performance on MDS benchmark with the aim of using transformer-based models on these task, trying to expand this promising solution. It is possible to add a decoder to the regular Longformer Encoder by creating a new model. Beltagy, Peters, Cohan (2020) proposed a Bart-base model, called Longformer Decoder-Encoder (LED) [2] that is able to process 16k tokens in input. This model is used to conduct the ablation study discussed in further section.

News Articles Summarization. News summarization is the process of generating a shorter version of a news article or a collection of articles while retaining the essential information. Its goal is to provide readers with a quick and comprehensive overview of the news without having to read the entire article. See et al. (2017) [8] have proposed an approach based on abstractive summarization, using a hybrid pointer-generator network that can copy words from the source text and a coverage mechanism that keeps track of what has been summarized to discourage repetition. It has been shown to outperform previous abstractive models.

Zhong et al. (2020) [9] suggest formulating the summarization task as a semantic text matching problem. In this approach, the source document and candidate summaries are matched in a semantic space to create the summary. The same dataset we used for extractive summarization on multi-document is used for the evaluation of the proposed model. This paper represents the state of art regarding extractive summarization with this type of data, although the concept of extractive summarization is reinterpreted as a semantic text matching problem.

III. Methodology

In this study, we propose a pipeline for news summarization that consists of three main stages: preprocessing, training, and regression. Firstly, we preprocess the input news articles using techniques such as stopword removal, sentence segmentation, and separation of articles of the same cluster, which enhances

Fig. 3: An example from multi-document summarization dataset showing the input documents and their summary [1].

the performance of the summarization model. During this stage a tokenized unique text is created merging the texts of all the articles related to the same news. Also a new context is defined, merging the first 20% of sentences of each article for the same cluster of news. This is done because is general assumed that the lead or the headline in news articles is designed to capture the reader's attention and provide a brief summary of the story. It should convey the most significant and newsworthy aspects of the story, such as the who, what, when, where, why, and how.

Next, we fine-tune the chosen models, BERT and Longformer, on the Multi-News dataset, which helps obtain the optimized weights for the models. Fine-tuning is an essential step, as our proposed summarization model, THExt, relies on the end-to-end training of a deep regression model.

To train an encoder model, we use a tokenizer to tokenize the abstractive summaries of the Multi-News dataset and use them as "proxy" targets for the encoder. We then fine-tune the encoder on the Multi-News dataset to extract similar sentences as those found in the abstractive summaries.

Finally, we use a fully connected layer-based regression to produce the resulting highlights for the cluster of articles. The complete architecture is showed in Fig. 2.

IV. Experiments

A. Datasets

To test our solution for Multi-Document Domain Adaptation we used a subset of Multi-News dataset [1], the first large-scale Multi-Documents news collection. From the original size of 56,216 articles-summary pairs we extracted a subset of 3600 articles-summary pairs, for computational reasons. The structure of the dataset is showed in Fig. 3.

Dataset	Train/Test/Val	#Chars/Cluster	#Words/Cluster	#Sentences/Cluster
Multi-News	2000/800/800	11565.617	2155.964	83.68

TABLE I: Dimensions of multi-news

From analysis on our subset of the original dataset emerge that the average number of articles for each cluster is 2.89, with a distribution showed in Table II. In Table I we can evaluate properly how the input file is big in dimension.

B. Experimental Setup

The training of the extraction, abstraction, and reinforcement modules is carried out in the Google Colab© environment, using a 12GB Test P100 GPU. In order to match the resources offered by Google Colab© we used subsets of the original datasets, in order to respect the constraints imposed by computational resources. We use NLTK library and AST library to perform preprocessing and tokenization. The number of highlights K extracted for each news was set to K=5. The number of epochs during the training was set to 2, the batch size was set to 8 and the number of jobs was set to 1.

C. Performance metrics

Rouge score is a way to measure the quality of text summaries by comparing them to reference texts. It uses precision and recall to calculate F1 score, which is a measure of performance.

Rouge-1, Rouge-2, and Rouge-L are variations that measure overlap between unigrams, bigrams, and longest common subsequences. This score can be helpful in evaluating text summarization algorithms and comparing different summaries. In our case the most significative metric is Rouge-1.

D. Results

The results of our study, visible in Table III, were impacted by several limitations and challenges. First, due to limited computational resources, we had to reduce the size of the dataset used in our experiments. This reduction in dataset size may have had a negative impact on the quality of our results, as we were not able to train and evaluate our model on the full dataset.

# Of Source	Frequency
2	1783
3	1016
4	428
5	165
6	90
7	33
8	16
9	11
10	4

TABLE II: Frequencies of number of sources

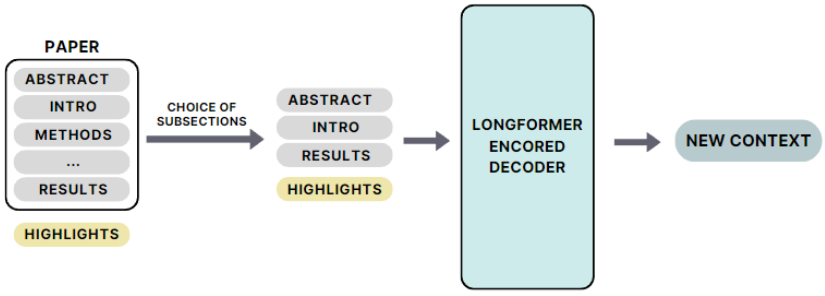

Fig. 4: Steps for new context generation

Encoder	BERT			Longformer		
ROUGE	R1	R2	RL	R1	R2	RL
Precision	0.08	0.018	0.050	0.074	0.014	0.045
Recall	0.424	0.089	0.262	0.325	0.061	0.199
F-Measure	0.135	0.029	0.082	0.117	0.022	0.071

TABLE III: Results obtained on Multi-News with K=5

Dataset	BIO			A1			CS		
ROUGE	R1	R2	RL	R1	R2	RL	R1	R2	RL
Precision	0.398	0.124	0.252	0.466	0.152	0.293	0.443	0.157	0.283
Recall	0.266	0.079	0.166	0.271	0.091	0.172	0.302	0.107	0.192
F-Measure	**0.308**	0.094	0.193	**0.332**	0.110	0.209	**0.348**	0.129	0.222

TABLE IV: Results obtained on CS, AI and BIO using $3 \leq K \leq 5$

Second, our summarization approach was effective at capturing key information from the source documents, we observed lower quality results when compared to the abstractive gold summaries. This disparity in results may be attributed to the fact that our approach was extractive, while the gold summaries were abstractive.

Lastly, we are confident that improving computational resources can effectively impact our pipeline performance. Exploiting it to its fullest potential, by training on the whole dataset, could increase significantly the quality of the summaries extracted. We believe our approach provides valuable insights into the summarization task, and offers a strong foundation for future work to build upon.

E. Ablation Study

In addition to the work carried out so far, we also tried to improve the benchmark model for highlights extraction [6] by rethinking the step of context generation. The basic idea is that some salient information of a paper context could be lost by providing the abstract only to the encoder. The architecture used in the study is depicted in Figure 4. The LED model, which is an extension of the Longformer [2], was employed to create a new context. Specifically, the study performed an abstractive summarization of different sections of the papers, namely the *abstract*, *introduction*, and *results*. This process resulted in a new context, which was then exploited by THExt architecture to generate new highlights.

We ran the experiments on a subset of the one use by La Quatra et al. in [6]. More precisely, we exploited 150 rows for the training and 80 for the test. As mentioned in the sec. I, the dataset is a collection of three documents: CSPubSumm, a collection of scientific papers related to the Computer Science area, BIOPubSumm, papers from the biology and medicine domain and AIPubSumm, articles from the Artificial Intelligence field. In order to match computational needs as well as achieving good performances, the cut of the original dataset was accompained by the usage of Google Colab Pro+© hardware resources. For the experiment, the number of highlights extracted is included between 3 and 5, depending on the number of highlights in the gold summaries of the dataset, so that a coherent number of sentences could be compared. The results of the study, exposed in Table IV, present interesting scores leading to the idea that this type of context generation can improve the baseline model. However, there is still potential for further improvement by exploring alternative combinations in context generation.

V. Conclusion

A. Multi-News Domain Adaptation of THExt

By the domain adaptation of La Quatra et al. [6] and the combination of the Longformer model [2], this work presented a valid solution to the task of highlights extraction in the field of news articles. The proposed approach lead to intresenting results, taking into account the diversity of the task between our work and the one presented in related works. Finally, a look at the extracted sentences shows that THExt is able to extract successfully salient highlights that could be used as key phrases for indexing or as subtitles.

B. Future Works

As future works could be interesting to enhance the preprocessing on news datasets, adding Named Entity Recognition and Part-of-Speech Tagging.

Regarding the context generation extension, therefore, it could be interesting to try to resolve the problem of propagation of false information that could be generated using an Abstractive Encoder-Decoder to generate the new context.

REFERENCES

[1] Tianwei She Suyi Li Dragomir R. Radev Alexander R. Fabbri, Irene Li. Multi-news: a large-scale multi-document summarization dataset and abstractive hierarchical model. *arXiv:1906.01749v3*, 2019.

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

[3] Luca Cagliero and Moreno La Quatra. Extracting highlights of scientific articles: A supervised summarization approach. *Expert Systems with Applications*, 160:113659, 2020.

[4] Ed Collins, Isabelle Augenstein, and Sebastian Riedel. A supervised approach to extractive summarisation of scientific papers. *arXiv preprint arXiv:1706.03946*, 2017.

[5] 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.

[6] Moreno La Quatra and Luca Cagliero. Transformer-based highlights extraction from scientific papers. *Knowledge-Based Systems*, page 109382, 2022.

[7] Yang Liu and Mirella Lapata. Text summarization with pretrained encoders. *arXiv preprint arXiv:1908.08345*, 2019.

[8] Abigail See, Peter J Liu, and Christopher D Manning. Get to the point: Summarization with pointer-generator networks. *arXiv preprint arXiv:1704.04368*, 2017.

[9] Ming Zhong, Pengfei Liu, Yiran Chen, Danqing Wang, Xipeng Qiu, and Xuanjing Huang. Extractive summarization as text matching. *arXiv preprint arXiv:2004.08795*, 2020.