EMS: Ensemble Model for Summarization

Giovanni Mantegna
Politecnico di Torino
Turin, Italy
s296555@studenti.polito.it

Domenico Mereu
Politecnico di Torino
Turin, Italy
s302151@studenti.polito.it

Chiara Scagliola
Politecnico di Torino
Turin, Italy
s303037@studenti.polito.it

***Abstract*—This paper aims to investigate the extraction of highlights from text utilizing an extractive sentence-based summarization approach. The proposed methodology involves the application of the Transformer-based Highlights Extractor (THExt) model that utilizes contextualized embeddings and transformer models to extract significant paper highlights. The THExt model is fine-tuned using the CNN Daily Email dataset to assess the generalization capability of the proposed method. Additionally, our study proposes an ensemble model that combines non-transformer-based techniques with the THExt model to enhance its effectiveness. The non-transformer-based summarization approaches include Latent Semantic Analysis (LSA), Relevance score, TF-IDF, and Text Rank. Results show that the ensemble model outperforms THExt alone by leveraging the benefits of each approach and compensating for their limitations. Non-transformer-based methods are better suited for capturing local relevance and co-occurrence relationships between words, while transformer-based models are better at capturing more complex semantic relationships and global context. The combination of these techniques leads to a more effective summarization approach with better overall performance. It is possible to find the code of our implementations at the following link: GitHub.**

I. Introduction

In the world of Natural Language Processing (NLP), Text Summarization has emerged as one of the most crucial and beneficial tasks since the inception of the field. Single-Document Summarization refers to the automatic creation of a condensed version of a document that retains its essential information. Through summarization, it becomes possible to extract the most important information and present it in a compact form, thereby improving the efficiency of information retrieval and making long and complex documents more accessible. In this paper, we focus on a specific aspect of summarization - the extraction of highlights. Our proposed method employs an extractive sentence-based summarization strategy, with the objective of selecting three sentences from the text that can serve as candidate highlights. Inspired by the work of M. La Quatra and L. Cagliero [1], we adopt the Transformer-based Highlights Extractor (THExt) model, which views highlights extraction as a regression task. Additionally, we fine-tune the THExt model using the CNN Daily Mail dataset in paragraph III-A and we build an ensemble model incorporating other several non-transformer based summarization methods (paragraph III-B) to further improve the selection of the final three highlights.

II. Related works

Over the years there have been various approaches to extractive text summarization, the main division is between extractive and abstractive text summarization. Extractive summarization is the task of identifying and selecting a subset of sentences or phrases from the input text to create a summary. There are statistical approaches like LSA [6], that is widely used for extractive text summarization. It offers a solution to the high dimensionality problem by transforming the original text into a lower dimensional space, where related words and concepts are clustered together. By doing so, LSA makes it possible to identify the most important words and concepts in a text, which can be used to generate a summary.

The frequency and positioning approach is another statistical method for extractive text summarization. It is based on the observation that important sentences tend to contain the most frequent and most important words, and that these words tend to appear near the beginning of the sentence [5].

Another type of approach is the graph-based.TextRank [12] is an unsupervised graph-based approach for text summarization that ranks sentences by their position in the graph and the similarity of their content to other sentences. It uses a graph to represent the text, with sentences as nodes and edges based on similarity, and an iterative algorithm to assign each sentence a score based on linked sentences. Finally, it selects the top-ranked sentences to generate a summary. The algorithm was inspired by Google's PageRank [8] algorithm.

With the rapid development of natural language processing (NLP) in recent years, transformer-based models have emerged as a powerful tool for text summarization.

In recent years, there has been a growing interest in using BERT [9], a transformer-based language model, for text summarization. BERT, which stands for Bidirectional Encoder Representations from Transformers, was originally designed for a range of natural language processing tasks, including question answering, named entity recognition, and sentiment analysis. However, researchers have found that BERT can also be used for summarization tasks.

One of the earliest attempts to use BERT for summarization was by Liu et al, who proposed the BERTSUM [10] model for extractive summarization. BERTSUM uses a two-stage approach, where BERT is first used to encode the input text, and then a summarization module selects the most important sentences based on the encoded representations. BERTSUM

achieved state-of-the-art results on several benchmark datasets. The main work we were inspired by is THExt, which will be discussed in the next section.

III. Method

A. THExt model structure

Our proposed model utilizes THExt architecture which incorporates contextualized embeddings and transformer models to extract relevant paper highlights. Specifically, it employs BERT bidirectional encoder representation that uses the attention mechanism to understand contextual relationships between words. THExt structure comprises of three main parts: context definition, sentence encoding using BERT, and regression based on a fully connected network. Context definition plays a crucial role in the architecture as it allows the model to capture global context dependencies, enabling accurate evaluation of candidate sentences. Fig.1 shows architecture structure of THExt.

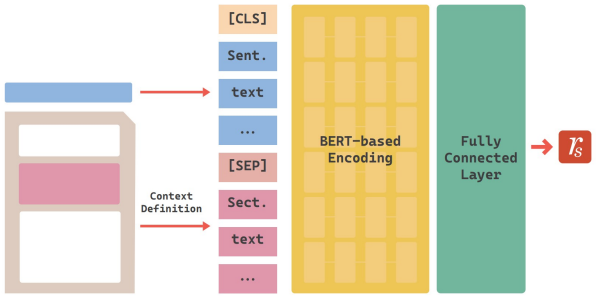

Fig. 1. Architecture structure of THExt model

B. Fine-tuning with CNN Daily Mail dataset

In the article by Moreno and Cagliero [1], three datasets were utilized in the training phase, including *CSPubSumm* which encompasses scientific papers in the Computer Science domain, *BIOPubSumm* which consists of papers in the biology and medicine field, and *AIPubSumm* which encompasses papers in the Artificial Intelligence domain. The present study aims to expand on the work of Moreno and Cagliero by evaluating the generalization capability of the proposed method. To achieve this, we fine-tuned the THExt model using the CNN Daily Email dataset, which includes complete articles and corresponding highlights utilized as ground truth during the training phase. To avoid retraining the model from scratch, we fixed the weights of the BERT structure used in THExt and performed training only on the part of the fully connected network. BERT was used exclusively as an encoder, while the fully connected network was used to perform the regression and return a score for each sentence. We took the first part of the particular article as context for each sentence and concatenated it with the sentence until we reached BERT's maximum input size. Since the computational resources at our disposal were not enough to fine-tune THExt using the CNN Daily Mail training set containing 287,113 articles. We considered only 11,490 items to fine tune and finally we tested on the validation set of 13,368 items.

C. Ensemble model

To improve the effectiveness of the summarization method introduced in this paper, we suggest an ensemble model that combines non-transformer-based approaches with the Thext model. Our rationale is that different summarization methods might excel in different aspects, and a combination of these techniques could result in superior performance.

1) Methods for summarization: Among non-transformer based summarization methods we use: *Latent Semantic Analysis (LSA)*, *Relevance score*, *TF-IDF*, *Text Rank*.

- Latent Semantic Analysis (LSA) is a statistical technique that has found extensive use in NLP. This technique is based on Singular Value Decomposition (SVD), which helps compress the term-document matrix and identify latent semantic relationships within data. Notably, LSA can also be utilized in summarization tasks. In our study, we implemented an algorithm proposed by Ozsoy et al. [6]. In this approach processing steps are taken to remove less-related sentences by setting scores below the average score for each topic to zero and removing noise from the SVD factorization by setting singular values below a threshold to zero. Then salience scores are calculated for each sentence based on the dot product of the square of the singular values and the square of the right singular matrix. The resulting salience scores reflect the most important and relevant sentences in the document.
- Relevance score: is calculated as described in the paper Gong et al. [11]. First the authors break the text into sentences to create a candidate sentence set S. Then, for each sentence $i \in S$, they create a weighted term-frequency vector A_i and for the whole document, a weighted term-frequency vector D. Finally, they compute the relevance score between A_i and D as the inner product. The sentence k with the highest relevance score is add to the summary. Furthermore they remove k from S and eliminate all the terms in k from the document, to avoid redundancy. This process is repeated until the desired number of sentences is reached.
- TF-IDF: is a classic machine learning method for quantifying how important a word is to a document in a collection or corpus. Among other things, this method can also be used for summarization. In our work the score of a sentence is computed by adding the TF-IDF values of all words in the sentence and dividing by the total number of words
- Text-Rank: is an algorithm first proposed in 2004 by Mihalcea et al. [12] the idea is to consider a fully connected graph structure where the nodes are sentences and the edges are weighted according to the semantic similarity between sentences. The sematic similarity in our case is given by the encoding of the pretrained BERT model. Once that graph has been constructed, PageRank

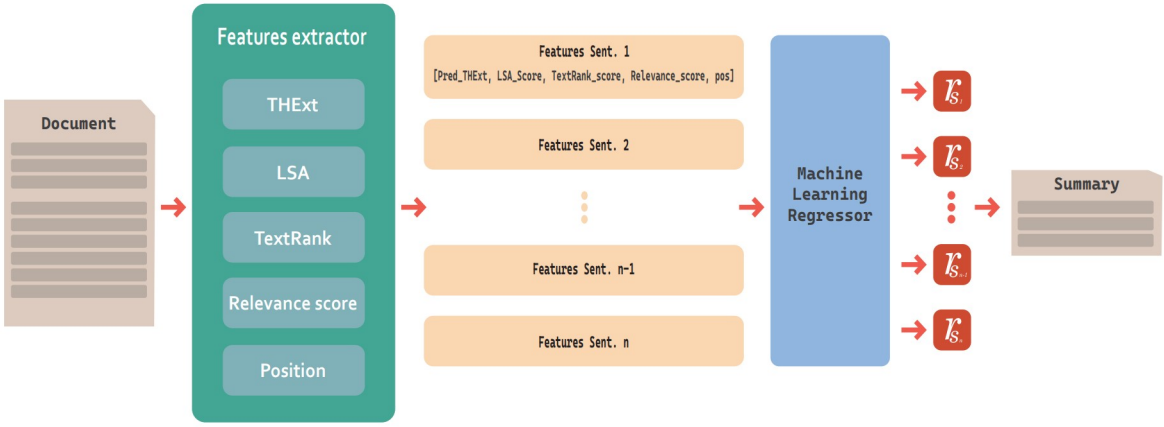

Fig. 2. Pipeline of the ensamble model

centrality is considered. The sentences with a higher centrality will have a higher score.

2) Models: Once the scores were calculated considering each of the methods seen in the previous paragraph, we used these scores as features of a new dataset described in paragraph IV-A. In our work we have tested some regression models: Random Forest Regressor, Stochastic Gradient Descent Regressor (SGDRegressor), Lasso. We tested more than one model because every model has pros and cons. The Random Forest Regressor is a tree-based model that can handle non-linear relationships between features and has low bias, but it can suffer from overfitting if the number of trees is too large. The SGDRegressor is a linear model that can handle large datasets and can be trained quickly, but it may not perform well if the data is not well-scaled. Lasso is a linear model that can perform feature selection by setting coefficients of irrelevant features to zero, but it can suffer from high variance if the number of features is much larger than the number of samples.

IV. Experiments

A. Experiment setup

We constructed a dataset with different features. The first set of features includes *text_rank*, *lsa_score*, *tf-idf*, and *relevance_score*, which are all from non-transformer based summarization methods. Additionally, we incorporated the *thext_score*, which is based on the approach described in [1]. We also included another feature *pos_i* which indicates the position of a word within a sentence. Currently, there is no single solution available for the automatic evaluation of summarization systems that can effectively assess all of their properties. However, the most commonly used evaluation framework at present is ROUGE (Recall-Oriented Understudy for Gisting Evaluation) [13]. ROUGE includes a set of scores that measure the degree of overlap between system-generated summaries and their corresponding human references. These scores are:

- Rouge-N: measures the degree of overlap between the N-grams of the generated highlights and the ground truth.
- Rouge-L: identifies the longest common subsequence of words between the generated highlights and the ground truth.

For each of this scores, it is possibile to consider recall, precision and F1 measures. In order to handle ground truth during the training, we followed the approach used in [1], whereby we determined the maximum value of rouge-2 F-measure between each sentence and its corresponding sentences in the highlights. A graphical illustration of the pipeline is shown in Fig 2.

B. Maximum independent set (MIS) for redundancy reduction

With the aim of preventing too similar sentences from being used to create the summary a further experiment was conducted using the results of the random forest regressor, we used an algorithm that considered the similarity of the sentences. This algorithm first extracts the sentences and their corresponding scores from the input list. It then creates a sentence vector for each sentence using the TfidfVectorizer method. The method then creates a graph where each sentence is represented as a node and the edges represent the similarity between the sentences. The weights of the edges are the cosine similarities between the sentence vectors. The method filters out edges that have weights above a certain threshold (median weight) and removes them from the graph. It then identifies the node n with the highest score and applies a maximal independent set algorithm to obtain a set of independent and non-redundant sentences containing n. The maximal independent set (MIS) approach filters out similar sentences based on a threshold, potentially excluding important information, whereas the baseline ensemble approach considers various factors for determining sentence importance. It should be noted that the performance of the MIS approach heavily relies on the choice of threshold for filtering out similar sentences. Setting the threshold too high may lead to important information being excluded from the summary, while setting it too low may result in redundant information being included. Finding the right balance can be difficult, and it may require manual tuning or experimenting with different values to determine the optimal

TABLE I
TABLE OF THE RESULTS

Method	Rouge-1 F1 score (%)	Rouge-2 F1 score (%)	Rouge-l F1 score (%)
THExt (baseline)	22.28	6.33	18.06
SGD	24.57	7.81	20.20
Lasso	28.50	10.70	23.43
Random Forest + MIS	27.77	9.09	22.40
Random Forest	**29.96**	**11.72**	**24.53**

threshold for a given dataset. Although the MIS approach has benefits, it did not improve summarization performance when included in the baseline ensemble model based on ROUGE scores.

C. Results

The result of the experiments performed are resumed in Table I. The first row shows the result of THExt considered with the BERT weights fixed and used as a baseline for the different ensemble methods considered. The values of the various Rougue (Rougue-1, Rougue-2, and Rougue-l) were obtained by averaging all the rougue values of the validation set and are used to evaluate the performance of the regressors. Based on the table, it can be seen that all the regressors perform better than the baseline THExt across all three metrics. In particular, the random forest regressor showed the largest improvement over the baseline, with gains of 7.68, 3.5, and 6.9 points in Rougue-1, Rougue-2, and Rougue-l, respectively. The Lasso and random forest with MIS also showed significant improvement over the baseline, with gains ranging from 3.22 to 6.85 points in all metrics. The SGD method as well outperformed the THExt baseline across all three Rouge metrics. However, the F1 scores obtained using the SGD method were lower than the scores obtained using the other three ensemble methods. Based on the results in the table, it can be seen that all four ensemble methods outperformed the THExt baseline across all three Rouge metrics. Ensemble methods may have outperformed the baseline due to their ability to capture complex patterns, handle data variability, and incorporate different techniques. The specific strengths of each method may have also contributed to their respective performance gains.

V. CONCLUSIONS

In conclusion, our proposed approach for extracting highlights from text using an extractive sentence-based summarization strategy, employing the Transformer-based Highlights Extractor (THExt) model, has yielded promising results. Through fine-tuning the model using the CNN Daily Email dataset, we have demonstrated the effectiveness of the proposed method and its ability to generalize well across different text datasets. To further improve the effectiveness of the THExt model, we proposed an ensemble model that combines non-transformer-based summarization techniques with THExt. These techniques include Latent Semantic Analysis (LSA), Relevance score, TF-IDF, and Text Rank. Our experiments show that all the regressors outperform the baseline THExt across all three metrics, with the random forest regressor achieving the largest improvement. Our proposed ensemble model achieves superior results compared to THExt alone, indicating that the combination of these techniques has the potential to further enhance performance. Further work could involve testing the proposed method on larger datasets to evaluate its generalization to other text datasets. In addition, it would be interesting to investigate the efficacy of combining other transformer-based and non-transformer-based approaches to improve text summarization performance.

REFERENCES

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

[2] Erkan, G., Radev, D. R. LexRank: Graph-based centrality as salience in text summarization. 2004

[3] Ani Nenkova, Lucy Vanderwende, and Kathleen McKeown. 2006. A compositional context sensitive multi-document summarizer: exploring the factors that influence summarization.

[4] Landauer, T. K., Foltz, P. W., Laham, D. An introduction to Latent Semantic Analysis.1998.

[5] Luhn, H. P. (1958). The automatic creation of literature abstracts. IBM Journal of Research and Development, 2(2), 159-165.

[6] Ozsoy, Makbule Gulcin, Ferda Nur Alpaslan, and Ilyas Cicekli. "Text summarization using latent semantic analysis." 2011

[7] P. C. Karmaker and M. S. Hossen, "Performance analysis of frequencyand graph theoretic based text summarization," .2019

[8] . Mihalcea, P. Tarau, and E.Figa. PageRank on semantic networks, with application to word sense disambiguation. 2004.

[9] Liu, Y., Li, S., Ouyang, L., Ma, F. (2019). Fine-tune BERT for extractive summarization. arXiv preprint arXiv:1903.10318.

[10] Yuxiang Wu and Pascale Fung. "BertSum: Text Summarization Using BERT". 2019.

[11] Gong, Yihong, and Xin Liu. "Generic text summarization using relevance measure and latent semantic analysis." Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval. 2001.

[12] Mihalcea, Rada, and Paul Tarau. "Textrank: Bringing order into text." Proceedings of the 2004 conference on empirical methods in natural language processing. 2004.

[13] Lin, Chin-Yew. "Rouge: A package for automatic evaluation of summaries." Text summarization branches out. 2004.