PreSumm: Predicting Summarization Performance Without Summarizing

Anonymous ACL submission

Abstract

Summarization models have achieved impressive benchmark performance in recent years, sharing common strengths and weaknesses. In this work, we focus on source-based features that affect most summarization models. We show that documents have specific properties that influence summarization performance. Therefore, we ask the question: can we predict a document's summarization performance without actually generating a summary? We introduce PreSumm, a system designed to predict how well a general summarization system would perform on summarizing a certain document. Surprisingly, PreSumm demonstrates a high correlation with human evaluations with respect to automatic metrics, supporting the hypothesis that certain global document features consistently affect model performance across systems. We further demonstrate the model's utility to enable efficient hybrid systems and to filter outliers and noise from datasets. Overall, our findings underscore the importance of source-text-driven factors in summarization performance and offer insights into the limitations of current systems that could serve as the basis for future improvements.

1 Introduction

Recent years have witnessed a remarkable proliferation of summarization models, with many achieving impressive performance on widely used benchmarks. These models, often rooted in large-scale language modeling, represent a significant leap forward in natural language processing capabilities. Their high benchmark scores highlight advancements in capturing linguistic nuances, handling diverse textual structures, and generating coherent outputs.

Despite these advancements, many summarization models share inherent design principles and operational mechanisms, which contribute to both their successes and limitations. While they excel in producing grammatically correct and syntactically fluent summaries, they often struggle with challenges such as logical reasoning and factual accuracy, leading to issues like hallucinations. These recurring challenges indicate that such models may share intrinsic common strengths and weaknesses, or they are influenced by comparable factors within datasets. Identifying and analyzing these shared dynamics is essential to uncovering systemic patterns that can inform future innovations in summarization.

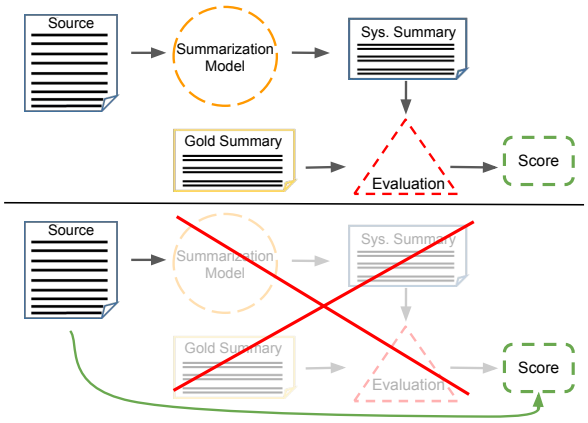

Figure 1: An illustration comparing the traditional evaluation process (top) with our approach (bottom).

In this work, we focus on identifying source-text-driven factors that influence the performance of summarization models. To that end, we introduce PreSumm, a system designed to predict the average performance of models summarizing a document without actually generating a summary, based on the document only. A success of PreSumm points out that it can distinguish between documents that most models summarize successfully, and those where performance falters. Analyzing such a model can reveal critical source-based features that challenge current systems, providing deeper insights into the limitations of existing approaches and paving the way for targeted improve-

ments in summarization technologies.

Additionally, PreSumm offers several practical benefits that could improve the efficiency and effectiveness of summarization workflows. One key advantage is its potential to enable hybrid systems where humans can focus their attention on difficult summary cases. For example, organizations often evaluate their output by relying on automatic metrics or manual human evaluation, which can reveal poor-quality summaries that require further manual summarization—an expensive and time-consuming process. By leveraging PreSumm to identify low-performing documents in advance, organizations can prioritize these cases for manual summarization or decide to opt-out before engaging in costly summarization workflows, thereby saving valuable resources and optimizing operational efficiency.

Furthermore, PreSumm may identify outliers and noisy documents in the dataset that eventually could enhance outcomes. An appealing example is multi-document summarization task, where one or more problematic documents might degrade the overall quality of the output. By filtering out such documents, PreSumm can help ensure more consistent, high-quality results across summarization tasks. These practical applications highlight PreSumm's potential as a valuable tool for improving both the cost-effectiveness and performance of summarization systems. An illustration of our approach is presented in Figure 1.

Surprisingly, PreSumm model showed a relatively high correlation with human evaluations with respect to automatic metrics, even though no summaries were generated in the process. This finding supports our hypothesis that certain global features of documents are consistently relevant across different systems.

Our results show that PreSumm outperforms comparable baselines in filtering out documents that require manual summarization (Sec. 6.1), or that hurt a multi document set (Sec. 6.2). Interestingly, we found that PreSumm assigns low scores to longer and more complex documents with a range of themes and perspectives (Sec. 7). By offering deeper insights into the limitations of current summarization systems, PreSumm lays a strong foundation for targeted advancements and future improvements in the field.

2 Related Work

Our work draws significant inspiration from PreQuEL (Don-Yehiya et al., 2022), which introduces a similar approach centered on machine translation. PreQuEL aims to predict translation system performance based solely on the source text. While our motivation and general methodology align with theirs, to the best of our knowledge, we are the first to apply this approach to text summarization. By leveraging PreSumm, we have enhanced summarization-specific applications and explored features more relevant to summarization, making our contributions distinct and novel.

While recent approaches (e.g., (Vig et al., 2022; Zhang et al., 2024) focused on embedding representation and did not leverage explicit source-based features, in the past, many summarization systems relied on explicit document-based features. These systems focused primarily on selecting key source sentences for inclusion in the summary—a task often framed as a classification problem. These features ranged from the presence of cue phrases (Gupta et al., 2011; Kulkarni and Prasad, 2010), the inclusion of numerical data (Prasad et al., 2012; Abuobieda et al., 2012), sentence length (Fattah and Ren, 2009; Abuobieda et al., 2012), and sentence position (Barrera and Verma, 2012; Fattah and Ren, 2009; Abuobieda et al., 2012; Li et al., 2016), to discourse structure (Louis et al., 2010), among others. While these studies employed such features to generate summaries, one may suggest a potential link between these features and the performance of summarization models. However, this property was not explicitly examined. In contrast, our work explores *system-independent document-level* features that impact the performance of modern summarization models in both *abstractive* and *extractive* modes.

3 Task Definition

Our task is to predict the average performance of summarization systems on a given document, using only the document as input. Averaging performance across multiple systems can reveal key properties of the document itself, while minimizing the influence of system-specific variability or noise. Formally, let $\mathcal{D}$ represent a corpus of N text passages, where each document is denoted as d_i. PreSumm aims to estimate the average quality of summaries generated by multiple systems for each document.

Specifically, consider M systems, denoted as $s_1, \ldots, s_M$, where system s_i produces a summary for document d_j that is scored as $S_{i,j}$. The goal of PreSumm is to first be able to generate a function $f : \mathcal{D} \rightarrow \mathbb{R}$ that predicts the average score assigned to the summaries of document d_j across all systems:

$$d_j^* = \frac{1}{M} \sum_{i}^{M} S_{i,j} \quad (1)$$

This score, d_j^*, reflects the average quality of the summaries generated by different systems for document d_j. We could leverage this score to rank the documents by their potential performance.

To evaluate the model's performance on ranking documents, we measure the correlation between the predicted scores and the gold-standard scores, following standard practices for assessing summarization metrics (Fabbri et al., 2021).

4 Dataset and Preliminary Analysis

4.1 The RoSE Dataset

The RoSE dataset (Liu et al., 2023) introduces a new method for annotating summarization datasets, which improves annotator agreement via *Atomic Content Units* (ACUs). The protocol tasks an annotator to convert a reference summary into atomic factual statements, then to compare the generated summary to these ACUs. The ACU score is defined as $f(s, A) = \frac{|A_s|}{|A|}$ where $|A|$ is the total number of ACUs for a given reference summary and $|A_s|$ is the number of matched ACUs of the generated summary with respect to the reference.

The authors manually evaluated over 22,000 summary-level annotations across 2,500 documents summarized by 28 top-performing systems on three datasets (CNN/DailyMail Nallapati et al. (2016), XSum Narayan et al. (2018), SamSum Gliwa et al. (2019)). This extensive manual evaluation of diverse systems and datasets aligns well with our task, and we adopt the ACU score as the primary metric to predict. Data statistics can be found in Table 1.

Dataset	Split	#Doc	#Sys.	#ACU
CNNDM	Test	500	12	5.6k
CNNDM	Valid	1,000	8	11.6k
XSum	Test	500	8	2.3k
SamSum	Test	500	8	2.3k

Table 1: Distribution of RoSE Dataset. Taken from Liu et al. (2023).

4.2 Preliminary Analysis - Do Systems Fail on The Same Documents?

In this section, we investigate whether different systems tend to consistently fail or succeed across the same documents. Since each system generates summaries for the same set of documents, their performance scores can be used to rank the documents. Our hypothesis is that different systems rank documents similarly, meaning the systems on certain documents consistently perform well or poorly across various systems.

To test this hypothesis, we measure the correlation between the ACU scores of documents on different systems. Specifically, we calculate the correlation between system scores l and k as $corr(S_{l,1..N}, S_{k,1..N})$. We then compute the average correlation across all systems using the formula:

$$\frac{2}{M^2 - M} \sum_{l=1}^{M} \sum_{k=1}^{l-1} corr(S_{l,1..N}, S_{k,1..N}) \quad (2)$$

We obtained a Kendall Tau correlation of 0.446 and a Spearman correlation of 0.565, indicating a moderate agreement in document rankings across systems. This suggests that many documents retain their relative ranking, regardless of the system used for summarization. These findings support our assumption that certain document-specific features significantly influence summarization performance. As a result, it might be possible to predict a document's average summarization performance based solely on its intrinsic characteristics.

5 Experiments

5.1 Models

We trained several models and examined their ability to rank the documents according to the average score across all systems, $d_j^* = \frac{1}{M} \sum_i^M S_{i,j}$. The RoSE dataset was divided into 80% train (2,000 documents) and 20% test (500 documents), where we trained all models with a fixed number of 5 epochs. More implementation details are elaborated in Appendix A.1.

Regression. In this model, we trained the model to predict the actual d_j^* scores. The input is a document j and the target is d_j^*. We leveraged the Longformer model (Beltagy et al., 2020) to allow a large input length of 4096 tokens required in many documents from our set. We added a regression head on top of the output of the 784 dimensional [CLS] token of the last layer of the Longformer. The regression head contains a feed forward component with an output of one dimension, which should predict the d_j^* score. Additionally, we used a standard MSE loss function for training.

Classification. Since some applications may only care about document rankings, we train an additional model to rank the documents rather than predict their exact d_j^* scores, we focused on ranking them directly based on pairwise comparisons. This approach aligns with our evaluation metric, which is based on correlation. Instead of predicting individual scores, we aim to determine which document in a pair should be ranked higher, and then aggregate these local decisions to form a global ranking.

In this model, the input consists of a pair of documents, and the task is to classify whether the first document should be ranked higher than the second. We represent each document using embeddings from a pre-trained Longformer model, concatenate the two embeddings, and feed the result into a RankNet model (Burges et al., 2005). The target label for each pair of documents (i, j) is $\delta_{ij} = \mathbb{1}_{d_i^* > d_j^*}$, where $\mathbb{1}$ is an indicator function that equals 1 if the condition $d_i^* > d_j^*$ is fulfilled, or 0 otherwise. The model is trained using a binary cross-entropy loss function.

We generate all possible n^2 pairs from the training set to fine-tune the model and use all m^2 pairs from the test set during evaluation. To derive the final global ranking, we follow the method outlined by Keswani and Jhamtani (2021), where the final score for document i is defined as $S(i) = \sum_j \hat{\delta} ij$, with $\hat{\delta} ij$ representing the predicted outcome for the document pair (i, j). We then sorted $S(i)$ scores for the final ranking.

Frozen Weights. Given the relatively small amount of training data, we explored a model with fewer trainable parameters to better handle the limited dataset. Specifically, we employed a variant of the regression model that demonstrated the best performance, freezing all weights except those in the regression head.

Model	Kendall τ	Pearson r	Spearman
Document length	-0.005	-0.048	-0.010
Count of Numbers	-0.016	-0.106	-0.023
# Unique Named Entities	-0.054	-0.134	-0.071
Flesch Reading Ease	-0.016	0.030	-0.021
Flesch Kincaid Grade	-0.0489	-0.0483	-0.0104
Regression (Ours)	**0.321**	**0.463**	**0.460**
Classification (Ours)	0.306	0.389	0.389
Frozen Weights (Ours)	0.279	0.406	0.403

Table 2: Test set correlations of different PreSumm models

We also compared our trained models with some simple baselines as follows.

Document Statistics. We explored several basic statistics about the documents, including document length by word, the number of numerical values, and the number of unique named entities identified using the NER module from the NLTK package. All of these features might be associated to the reading complexity of documents, where longer documents with more numeric details and name entities might be more complex to read.

Flesch–Kincaid Readability Tests. We applied the Flesch Reading Ease test (Flesch, 1948) to measure how easy the document is to read, with scores ranging from 1 to 100 where higher scores indicate easier readability. Similarly, we used the Flesch-Kincaid Grade Level to estimate the U.S. education grade level required to understand the text, with higher scores corresponding to a more advanced reading level

5.2 Main Results

All models are evaluated by measuring their correlation with the gold-standard ACU labels, which reflects how well they ranked the documents. The correlation scores are summarized in Table 2. Most baselines show near-zero correlation, except for the number of numeric values and named entities, which exhibit a small negative correlation. This suggests that an increased presence of numbers and entities may lead to lower summarization model performance. In contrast, most trained models achieve a moderate correlation, supporting our hypothesis that document ranking can, to some extent, be predicted based solely on the document content.

The table shows that the Regression PreSumm model outperformed the other models. Moreover, this model is also the most efficient one.

Model	Test	Kendall Tau	Pearson r	Spearman
PreSumm	w.o XSUM	0.294	0.433	0.423
$\text{PreSumm}_{\text{OOD}}$	w.o XSUM	0.296	0.434	0.420
PreSumm	XSUM only	0.246	0.362	0.352
$\text{PreSumm}_{\text{OOD}}$	XSUM only	0.116	0.240	0.177

Table 3: Results of in and out of distribution regression PreSumm models. Where OOD signifies that the PreSumm model was not trained on the XSUM portion of RoSE.

Unlike other supervised models, Regression processes each document only once. In contrast, the classification-based approach evaluates all possible pairs of documents, resulting in quadratic complexity. Additionally, the classification-based method requires an extra global aggregation step to consolidate local pairwise predictions. Given its superior performance and simplicity, we selected the Regression model for the next experiments as our PreSumm method of choice.

5.3 Out-of-Distribution Performance

All PreSumm models (Table 2) were trained and evaluated on our train and test sets that contain source documents from CNNDM, XSUM, and SAMSUM. In this section, we examine how well these models perform when applied to documents from a different dataset with distinct properties.

To estimate how the model would perform on out-of-distribution data, we trained a similar regression model without including the XSUM documents of RoSE and evaluated its performance on out-of-distribution data.[1]

We first compare this new model, $\text{PreSumm}_{\text{w.o_XSUM}}$, to the original PreSumm model in an "in-distribution" setting to ensure comparable performance. For this, we evaluate both models on a test set excluding XSUM, which mirrors the training data for $\text{PreSumm}_{\text{w.o_XSUM}}$. As shown in Table 3, the two models perform similarly, confirming their alignment. Details of the different training and test datasets used are provided in Table 4.

Next, we assess $\text{PreSumm}_{\text{w.o_XSUM}}$ on true out-of-distribution data, specifically the documents from XSUM. As expected, performance decreases. However, for most real-world applications and analyses, we are primarily interested in low-ranked documents for filtering or analysis rather than the entire ranked set. When measuring the ranking accuracy of only the n lowest-ranked documents, as detailed in Table 5, the performance gap between PreSumm and $\text{PreSumm}_{\text{w.o_XSUM}}$ diminishes and, in some cases, becomes negligible. Additional details are provided in Appendix C.

[1]We did not use the original PreSumm model itself, as we did not exclude any dataset with ACU annotations from its training due to the limited availability of such annotations.

Dataset	Train/Test	Datasets Used	#Doc	Model Trained On
All	Train	CNN, XSUM, SAM	2,000	PreSumm
All	Test	CNN, XSUM, SAM	500	N/A
w.o XSUM	Train	CNN, SAM	1,600	$\text{PreSumm}_{\text{w.o_XSUM}}$
w.o XSUM	Test	CNN, SAM	400	N/A
XSUM only	Test	XSUM	100	N/A

Table 4: Statistics on train and test sets used.

n	PreSumm	$\text{PreSum}_{\text{w.o_XSUM}}$
10	0.406	0.405
15	0.338	0.380
20	0.430	0.364
30	0.332	0.226

Table 5: Ranking accuracy (in scale [-1,1]) of the n low ranked documents over 'XSUM only' test dataset.

Overall, our results indicate that for low-ranked documents, the performance of $\text{PreSumm}_{\text{w.o_XSUM}}$ on out-of-distribution data is nearly equivalent to that of the original PreSumm model. Furthermore, in Section 6, we present two downstream applications where our model is used to filter low-ranked documents and demonstrate its effectiveness with out-of-distribution data.

6 Extrinsic Evaluation through Downstream Tasks

In this section, we explore practical applications that benefit from predicting in advance the summarization model performance over documents. Additionally, these applications serve as extrinsic evaluations of our model. Specifically, we examine two use cases: (1) identifying in advance the documents where models perform poorly to enable manual summarization in hybrid systems, and (2) filtering out noisy documents in a multi-document summarization setting.

6.1 Selecting Documents for Manual Summarization in Hybrid Systems

Here, we focus on a use case within a hybrid summarization system, where a fixed percentage of documents can be manually summarized within the available budget. Accordingly, we aimed to assess whether PreSumm can effectively identify,

Based on	Selection Type	BART		Pegasus	
		Replace 10%	Replace 20%	Replace 10%	Replace 20%
Source	Random	0.331^{+*}	0.375^{+*}	0.331^{+*}	0.377^{+*}
	Num. Entities	0.339^{+*}	0.421^{*}	0.333^{+*}	0.411^{+}
	Flesch	0.335^{+*}	0.407^{+*}	0.322^{*}	0.409^{+}
	PreSumm	**0.358**	**0.452**	**0.347**	**0.427**
	$\text{PreSumm}_{\text{w.o_XSUM}}$	**0.359**	**0.440**	**0.346**	**0.432**
Source + Sys.	Blanc	0.331^{+*}	0.416^{*}	0.326^{*}	0.396^{+*}
	SummQa	0.339^{+*}	0.421^{*}	0.334	0.401^{+}
Source + Sys.+ Ref.	Rouge	0.365	0.447	0.353	0.443
	Bert	0.365	0.459	0.354	0.446
	Meteor	0.357	0.453	0.345	0.435

Table 6: Averaged ACU scores of system summaries (from the 'XSUM only' test set) after replacing summaries generated from the lowest-scoring documents and summaries (based on several metrics) with manual summaries. Scores significantly worse than PreSumm or $\text{PreSumm}_{\text{w.o_XSUM}}$ are marked with * and $^{+}$, respectively.

in advance, documents that models are likely to fail on, allowing these documents to be prioritized for manual summarization. The goal is to maximize the overall score of the entire document set.

For this experiment, we used the generated summaries of two systems from the RoSE dataset, Pegasus Zhang et al. (2020) and BART Lewis et al. (2020). For evaluation, we used the average of the human ACU scores, where instead of manually summarizing a selected document, we assigned it an ACU score of 1. For the manual summarization budget, we selected either 10% or 20% of the test set documents. To support the conclusion from Section 5.3—that out-of-distribution performance should be similar when focusing on low-ranked documents—we conducted the experiment on the XSUM-only test set. The same experiment was conducted on the All test set, yielding equivalent results, as shown in Table 9 in the Appendix. For statistical significance testing, we used the paired bootstrap test (Efron and Tibshirani, 1994) as explained in (Berg-Kirkpatrick et al., 2012). The detailed algorithm is provided in Algorithm 1 in the Appendix.

In addition to PreSumm, we evaluated several baseline methods. Baselines included Random selection and ranking methods based solely on the source document features, such as Flesch Reading Ease and the number of unique named entities. We also tested reference-free metrics such as Blanc (Vasilyev et al., 2020) and SummQA (Scialom et al., 2019). However, a limitation of these reference-free metrics is that they require system-generated summaries, unlike PreSumm. For comparison, we included reference-based metrics such as ROUGE-2 F1 (Lin, 2004), BERTScore F1 (Zhang et al., 2019), and METEOR (Banerjee and Lavie, 2005), which represent an upper bound since they rely on both system and reference summaries—resources unavailable in real-world scenarios.

As shown in Table 6, both PreSumm and $\text{PreSumm}_{\text{w.o_XSUM}}$ significantly outperform all source-only and reference-free baselines in most cases, and approach the upper bound set by reference-based metrics. Overall, these experiments demonstrate that the PreSumm model can effectively identify in advance documents that models are likely to fail on, optimizing the summarization process by saving time and resources.

6.2 Multi-Document Summarization

In a Multi-Document Summarization (MDS) task, a set of documents on the same topic needs to be summarized. However, these sets often include noisy documents that can negatively impact model performance Giorgi et al. (2023). Additionally, summarizing a large number of documents poses challenges due to high input length, which can be costly or constrained by the token limits of certain models. Conventional MDS approaches typically concatenate all documents and truncate the input to meet the model's token limit or user budget, leading to the exclusion of some documents. This raises the question: can we achieve better results by using PreSumm to identify and exclude noisy documents from the set?

To test this, we adopted the MultiNews dataset (Fabbri et al., 2019) as our MDS test set and used the Pegasus summarization model (Zhang et al., 2020), to generate summaries. We conducted experiments with token limits of 256, 512, and 1024.

Token Lim.	Order	R-1	R-2	R-L
1024	original	0.458	0.185	0.243
	PreSumm	**0.461**	**0.189**	**0.246**
512	original	0.439	0.172	0.235
	PreSumm	**0.444**	**0.179**	**0.240**
256	original	0.398	0.147	0.215
	PreSumm	**0.402**	**0.153**	**0.219**

Table 7: Multi-Document Summarization Table. Comparing rouge scores of summarizations generated from the original order of the dataset, to PreSumms order.

In each experiment, we tested two variations: one where the documents were kept in their original order and truncated once the token limit was exceeded, and another where the documents were reordered based on their PreSumm scores, with the lowest-scored documents placed at the end. It is important to note that reordering the documents has an additional benefit—better document sequencing can enhance summarization quality even without excluding documents, as suggested by Zhao et al. (2022). Thus, we aimed to determine whether combining PreSumm-based document exclusion with PreSumm-based reordering would lead to improved summarization outcomes.

As shown in Table 7, the PreSumm-ordered documents consistently achieve higher ROUGE scores compared to the original document order across all input length limits. PreSumm is significantly better across all metrics according to the Wilcoxon Rank Test (Wilcoxon, 1945), except for 1024-token limit with R-1 and R-L. This demonstrates that using PreSumm to identify independent documents that models are likely to summarize unsuccessfully is also effective in enhancing multi-document summarization settings, as such independent documents contribute noise to the entire document set.

7 Analysis

In this section, we aim to investigate PreSumm to better understand the properties that make a document less likely to be successfully summarized by various summarization systems. To that end, we examine the influence of document-based features over PreSumm (Section 7.1). Then, we conducted a manual analysis to reveal more insights and explain the automatic results (Section 7.2). Additional analysis examines how PreSumm deals with different corruptions can be found in Appendix B.1.

Feature	Correlation
Document length	-0.0956
Count of Numbers	-0.0576
# of Unique Named Entities	-0.120
Flesch Reading Ease	0.182
Flesch Kincaid Grade	-0.166
Avg Loc of Salient Sent (top 10)	-0.266
Avg Loc of Salient Sent (top 5)	-0.305

Table 8: Pearson R Correlations of document features to regression PreSumm predicted scores.

7.1 Document Feature Correlations

To gain deeper insights into which document-based features most strongly influence the performance of summarization systems over a certain document, we analyzed the correlations between various document characteristics and the PreSumm score at the document level. For features, we leveraged the baseline methods from Section 5.1, including the document statistics and the reading ease tests. In addition, we added the salient sentence location feature that uses the reference summary, and therefore it could not be used as method in Section 5.1.

Salient Sentence Location. Previous work pointed that the salient information in news documents tend to be at the beginning of the document Lebanoff et al. (2019). As in this work we focus on the news domain, we hypothesis that news document that their main theme appear in the middle of the document, or when there are several themes in the paper, models might struggle to summarize it successfully. Therefore, we would like to set the location of the salient sentences in the document as a feature. To determine the location of key sentences within a document, we adopted a method similar to Nallapati et al. (2017); Chen and Bansal (2018), where each sentence in the document is ranked by its similarity to the reference summary, using ROUGE scores as the metric. We selected the top-5 or top-10 most salient sentences, and their positions were normalized by the total number of sentences in the document. The average of these normalized indices represents the typical location of the most important sentences.

The correlation results are shown in Table 8. The most strongly correlated feature is the location of salient information. The negative correlation indicates that the earlier the key information appears in the document, the easier it is for the model to summarize. This observation aligns with expectations,

as this is a well-known characteristic of the news domain, where essential information is typically presented at the beginning.

The Flesch Reading Ease and Flesch-Kincaid Grade Level scores show positive and negative correlations, respectively, indicating that more complex documents tend to result in poorer summarization model performance. In the same way, all basic document statistics show a relatively weak negative correlation, suggesting that greater document complexity—whether in length, the number of numerical values, or named entities—negatively impacts summarization performance.

Overall, while the correlation scores and their signs are consistent with our expectations, none of the features exhibited strong correlations. Consequently, we conducted a manual analysis to shed light on new document-based insights.

7.2 Manual Analysis

To gain deeper insights into document performance, we conducted a manual analysis. Specifically, one of the authors read the 15 best and worst-ranked documents according to PreSumm predictions. In general, this analysis found a significant difference between the top-ranked documents and the bottom-ranked ones. More specifically, it revealed three distinct types of bottom-ranked documents:

Content Complexity. The main and most common characteristic of low-ranked documents is that they often cover complex topics, such as science or politics, which include numerous numbers, intricate details, and long, difficult-to-follow sentences and documents. In contrast, the top-ranked documents were typically much shorter (some with only a single sentence), with simple words and topics.

Coherence. Some bottom-ranked documents exhibited weak sentence-to-sentence connections or lacked sufficient background information, starting abruptly in the middle of a story, counting on specific terms and knowledge of a specific unique field. Sometimes it happens because of the crawling process of documents, that includes the image caption or sub-headers in the middle of the text. Such crawling issues were also seen in the top-ranked documents, but less frequently.

Theme Change: Some low-ranked documents contain multiple, almost disjointed themes, making it difficult to determine which theme should be prioritized in the summary. This issue was especially pronounced in cases where the main theme was in the middle, requiring the model to go beyond its usual focus at the beginning of the document.

Overall, the low-ranked documents were notably more difficult to read, often requiring re-reading of certain sentences for comprehension, whereas the top-ranked documents were much more fluent and easier to follow. Examples of documents with these challenging types are provided in Appendix D.

As discussed in Section 7.1, we identified some correlations between specific features and the properties described here, such as results from Reading Ease tests and the location of salient information. However, these correlations were weak. Upon closer analysis, we hypothesize that the underlying reason may be the orthogonal nature of these factors in many cases. For instance, some low-ranked documents are short and cover straightforward, day-to-day topics, yet they suffer from coherence issues. Such documents weaken the correlation with Reading Ease scores. A comprehensive model would likely need to integrate multiple parameters to capture these nuances effectively.

We also examined low-ranked documents based on the gold labels of the averaged ACU scores. Interestingly, we observed similar patterns. However, many of the low human-scored documents were penalized for misalignment between the reference summary and the document content. Since our model does not rely on the reference summary, it was unaffected by this factor and, therefore, did not predict these documents as low-ranked.

These findings suggest that models tend to struggle with cases where humans also face challenges in reading. This observation is noteworthy because factors like sentence length, which significantly impact human readability, might not necessarily challenge models with a large input context length.

8 Conclusion

In this work, we introduced PreSumm, a novel approach that opens up new research avenues in understanding the structural features that make a document less likely to be summarized successfully. Our findings suggest that documents that are more complicated to read for humans are also ranked low by PreSumm, implying the centrality of this feature for summarization models. We hope these insights will contribute to the design of more robust and effective models in the future.

9 Limitations

Our study covers the RoSE dataset extensively and focuses on summarization of the news domain only. Therefore, we cannot explore the complete space of summarization systems and we are limited to both the datasets and summarization systems that RoSE provides due to our extensive use of the manual ACU score. Because of this, some of the results could be due to other factors that relate to the dataset and would not generalize strongly outside of this study or to other domains.

Although we showed in our work that our model works relatively well on out-of-distribution data, we did not examine dataset out of the news domain. Therefore our conclusions are limited to this domain.

In future works, we would seek to train a similar model on a larger dataset. However, using ACU scores would be difficult because of the human labor involved, which could be avoided by using an annotated metric to train on. However, will make us biased towards the chosen metric, which is another limitation.

Out of distribution data is a major factor when it comes to model performance. To mitigate this would require greatly increasing the scope of the experiment and to train on a broader dataset for more accurate predictions.

Overall, overcoming these limitations would necessitate a much larger corpus with either a large set of automated or human annotated metrics to perform a similar study on a much larger set of the space of documents and summarization system.

References

Albaraa Abuobieda, Naomie Salim, Ameer Tawfik Albaham, Ahmed Hamza Osman, and Yogan Jaya Kumar. 2012. Text summarization features selection method using pseudo genetic-based model. *2012 International Conference on Information Retrieval & Knowledge Management*, pages 193–197.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Araly Barrera and Rakesh Verma. 2012. Combining syntax and semantics for automatic extractive single-document summarization. In *Computational Linguistics and Intelligent Text Processing*, pages 366–377, Berlin, Heidelberg. Springer Berlin Heidelberg.

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

Taylor Berg-Kirkpatrick, David Burkett, and Dan Klein. 2012. An empirical investigation of statistical significance in NLP. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 995–1005, Jeju Island, Korea. Association for Computational Linguistics.

Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. 2005. Learning to rank using gradient descent. In *Proceedings of the 22nd international conference on Machine learning*, pages 89–96.

Yen-Chun Chen and Mohit Bansal. 2018. Fast abstractive summarization with reinforce-selected sentence rewriting. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 675–686, Melbourne, Australia. Association for Computational Linguistics.

Shachar Don-Yehiya, Leshem Choshen, and Omri Abend. 2022. PreQuEL: Quality estimation of machine translation outputs in advance. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11170–11183, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Bradley Efron and Robert J Tibshirani. 1994. *An introduction to the bootstrap*. Chapman and Hall/CRC.

Alexander Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir Radev. 2019. Multi-News: A Large-Scale Multi-Document Summarization Dataset and Abstractive Hierarchical Model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1074–1084, Florence, Italy. Association for Computational Linguistics.

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. SummEval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Mohamed Abdel Fattah and Fuji Ren. 2009. Ga, mr, ffnn, pnn and gmm based models for automatic text summarization. *Computer Speech & Language*, 23(1):126–144.

Rudolph Flesch. 1948. A new readability yardstick. *Journal of applied psychology*, 32(3):221.

John Giorgi, Luca Soldaini, Bo Wang, Gary Bader, Kyle Lo, Lucy Wang, and Arman Cohan. 2023. Open domain multi-document summarization: A comprehensive study of model brittleness under retrieval. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8177–8199, Singapore. Association for Computational Linguistics.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 70–79, Hong Kong, China. Association for Computational Linguistics.

Pankaj Gupta, Vijay Shankar Pendluri, and Ishant Vats. 2011. Summarizing text by ranking text units according to shallow linguistic features. *13th International Conference on Advanced Communication Technology (ICACT2011)*, pages 1620–1625.

Vishal Keswani and Harsh Jhamtani. 2021. Formulating neural sentence ordering as the asymmetric traveling salesman problem. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 128–139, Aberdeen, Scotland, UK. Association for Computational Linguistics.

Uday Kulkarni and Rajesh Shardanand Prasad. 2010. Implementation and evaluation of evolutionary connectionist approaches to automated text summarization. *Journal of Computer Science*, 6(11):1366–1376.

Logan Lebanoff, Kaiqiang Song, Franck Dernoncourt, Doo Soon Kim, Seokhwan Kim, Walter Chang, and Fei Liu. 2019. Scoring Sentence Singletons and Pairs for Abstractive Summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2175–2189, Florence, Italy. Association for Computational Linguistics.

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.

Junyi Jessy Li, Kapil Thadani, and Amanda Stent. 2016. The role of discourse units in near-extractive summarization. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 137–147, Los Angeles. Association for Computational Linguistics.

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.

Yixin Liu, Alex Fabbri, Pengfei Liu, Yilun Zhao, Linyong Nan, Ruilin Han, Simeng Han, Shafiq Joty, Chien-Sheng Wu, Caiming Xiong, and Dragomir Radev. 2023. Revisiting the gold standard: Grounding summarization evaluation with robust human evaluation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4140–4170, Toronto, Canada. Association for Computational Linguistics.

Annie Louis, Aravind Joshi, and Ani Nenkova. 2010. Discourse indicators for content selection in summarization. In *Proceedings of the SIGDIAL 2010 Conference*, pages 147–156, Tokyo, Japan. Association for Computational Linguistics.

Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. 2017. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In *Thirty-First AAAI Conference on Artificial Intelligence*.

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.

Rajesh Shardanand Prasad, Nitish Milind Uplavikar, Sanket Shantilals Wakhare, VY Jain, Tejas Avinash, et al. 2012. Feature based text summarization. *International journal of advances in computing and information researches*, 1(2):15–18.

Thomas Scialom, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. 2019. Answers unite! unsupervised metrics for reinforced summarization models. *arXiv preprint arXiv:1909.01610*.

Oleg Vasilyev, Vedant Dharnidharka, and John Bohannon. 2020. Fill in the blanc: Human-free quality estimation of document summaries. *arXiv preprint arXiv:2002.09836*.

Jesse Vig, Alexander Fabbri, Wojciech Kryscinski, Chien-Sheng Wu, and Wenhao Liu. 2022. Exploring neural models for query-focused summarization. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1455–1468, Seattle, United States. Association for Computational Linguistics.

Frank. Wilcoxon. 1945. Individual comparisons by ranking methods. *Biometrics*, 1:196–202.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International conference on machine learning*, pages 11328–11339. PMLR.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B. Hashimoto. 2024. Benchmarking large language models for news summarization. *Transactions of the Association for Computational Linguistics*, 12:39–57.

Chao Zhao, Tenghao Huang, Somnath Basu Roy Chowdhury, Muthu Kumar Chandrasekaran, Kathleen McKeown, and Snigdha Chaturvedi. 2022. Read top news first: A document reordering approach for multi-document news summarization. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 613–621, Dublin, Ireland. Association for Computational Linguistics.

A Appendix

A.1 Implementation Details

The SummEval models were built using the Longformer architecture, each with distinct output heads and training methodologies tailored to the task.

All models were trained for 5 epochs with a learning rate of 1e-6, using an 80%-20% train-validation split of the dataset. The base model used was the "allenai/longformer-base-4096" (Beltagy et al., 2020), configured to handle the maximum sequence length.

A.1.1 Regression Head

For regression tasks, a single linear layer was added to map the 768-dimensional CLS token embedding to a one-dimensional output, representing the predicted score.

A.1.2 Classification Head

For classification tasks, the Longformer backbone was paired with a more complex classification head. This head comprised a feedforward neural network with the following structure: a linear layer mapping 768 dimensions to 512, a ReLU activation, and a second linear layer reducing 512 dimensions to a single scalar output. During the forward pass, the model computed individual scores from two heads, denoted as $s1$ and $s2$, and generated the final probability by subtracting these scores and applying the sigmoid function.

A.1.3 Frozen Weights

The Frozen Weights model followed the same training procedure as the regression model but kept the weights of the Longformer backbone fixed, allowing only the output layer to be updated during training.

B Analysis

B.1 Document-Based Transformations

To further investigate the influence of document-based features on model performance, we explored how the predicted score of a document changes when specific features are perturbed. By comparing the predicted scores before and after these transformations, we can gauge the importance PreSumm places on each feature. Intuitively, the transformations causing the largest change in predicted scores should indicate which features have the greatest impact on document performance in summarization. We applied these transformations to our test set. Below, we detail the transformations applied.

Removing content. We tested several removal strategies: removing the first sentence (often critical for summarization), removing 5 or 10 salient sentences (as defined in Section 7.1), and randomly removing 30% of the words or sentences to disrupt fluency and coherence. Additionally, we removed all sentences except the first three or last three to assess the importance of content location.

Moving content. Given that salient information location was identified as a key feature in Section 7.1, we moved the salient sentences to the end of the document. We also randomly shuffled all sentences to disrupt coherence.

Replacing content. We replaced named entities to test the impact on consistency. We also introduced contradictions by adding negation sentences and corrupted the grammar by converting all verbs to their lemma forms.

As expected, the most impactful corruptions, with the highest changes in predicted scores, were removing the 10 most salient sentences and removing all content except for the last three sentences. Other significant transformations included deleting 30% of the words, removing 5 salient sentences, and shuffling sentences randomly.

Surprisingly, moving the salient sentences to the end of the document had little effect, despite the location of salient information being one of the most influential features in Section 7.1.

Interestingly, content replacement, such as grammar corruption or adding contradictions, did not significantly affect PreSumm's performance. It also appears that strong perturbations, such as deleting

Based on	Selection Type	BART		Pegasus	
		Replace 10%	Replace 20%	Replace 10%	Replace 20%
Source	Random	0.425*	0.478*	0.394*	0.452*
	Num. Entities	0.437*	0.508*	0.416	0.486*
	Flesch	0.427*	0.494*	0.403*	0.472*
	PreSumm	**0.451**	**0.525**	**0.423**	**0.499**
Source + Sys.	Blanc	0.435*	0.512*	0.416	0.496
	SummQa	0.434*	0.502*	0.414	0.481*
Source + Sys.+ Ref.	Rouge	0.459	0.540	0.432	0.515
	Bert	0.454	0.527	0.429	0.507
	Meteor	0.459	0.538	0.433	0.516

Table 9: Averaged ACU scores of system summaries (from the 'All' test set) after replacing summaries generated from the lowest-scoring documents and summaries (based on several metrics) with manual summaries. Scores significantly worse than PreSumm are marked with *.

Transformation	Src.	Trans.	Delta
Remove 10 salient sentences	0.528	0.428	-0.100
Keep last 3 sentences	0.528	0.441	-0.0877
Delete 30% of words	0.528	0.470	-0.0584
Remove 5 salient sentences	0.528	0.476	-0.0523
Randomly Shuffle of Sentences	0.528	0.486	-0.0427
Move 10 salient sentences to end	0.528	0.502	-0.0269
Keep first 3 sentences	0.528	0.553	0.0246
Remove first sentence	0.528	0.506	-0.0225
Move 5 salient sentences to end	0.528	0.509	-0.0199
Replace names w/ from bank	0.528	0.512	-0.0163
Replace names w/ spacy name	0.528	0.518	-0.0102
Corrupt Grammar	0.528	0.520	-0.008
Append contradictions	0.528	0.533	0.005
Delete 30% of sentences	0.528	0.530	0.00150

Table 10: Predicted scores of documents before a transformation (Src.) and after (Trans.)

30% of sentences, did not lead to large differences in scores. This might be because these artificial corruptions are not natural and therefore deviate too much from the patterns seen during model training, causing the model to mispredict their impact on summarization performance.

C Compare Accuracy on Low-ranked Documents

In our main evaluation (Table 2 we used Kendall-tau correlation. In this correlation we compare a pair (x_i,y_i) to (x_j,y_j) for all i, j. If $x_i > x_j$ and $y_i > y_j$, or $x_i < x_j$ and $y_i < y_j$ these pairs are concordant, and the final correlation score is increased.

In order to measure the ranking of the low-ranked documents only, we applied the same

Algorithm 1 PreSumm vs method X Paired Bootstrap Significance Test

1: **Input:** Test set of documents, PreSumm scores, X scores, averaged ACU scores
2: **Output:** p-value
3: Extract the current difference in performance between PreSumm filtering and X filtering, denoted as `original_diff`.
4: Initialize $s = 0$
5: **for** each document, with 3 scores: averaged ACU, PreSumm score, X score **do**
6: Sample n instances with replacement from the test set (same number of instances, each document can appear more than once)
7: Filter 10% or 20% of documents according to PreSumm, and average the ACU scores to get `PreSumm_filter_score`
8: Filter 10% or 20% of documents according to X, and average the ACU scores to get `X_filter_score`
9: **if** PreSumm_filter_score - X_filter_score > 2 × `original_diff` **then**
10: $s = s + 1$
11: **end if**
12: **end for**
13: Repeat steps 2-5 for $b = 10000$ iterations.
14: Compute $p_val = \frac{s}{b}$

Kendall-tau process, but considered only pairs that one of them is one of the n low-ranked documents. This way, we compare each of the low-ranked documents to all other documents, and therefore measuring the accuracy of its ranking (in a [-1,1] scale). In Table 5 we showed the the performance over low-ranked documents is higher, and $\text{PreSumm}_{\text{w.o_XSUM}}$ preforms almost equivalently to PreSumm.

D Example Documents

In tables 11 and 12 are examples of documents from the data that were annotated with different challenges.

Challenging Characteristics	Document
Coherence	Judge Thokozile Masipa did the same for the lawyers on Thursday, urging them to make good use of the upcoming fortnight break for the Easter holidays. In that spirit, here are a few questions that have been niggling me in recent days. Tweet your thoughts and suggestions to @BBCAndrewH. I will be taking a week off and then focusing on South Africa's general election before returning to the hard benches of Courtroom GD on 5 May.
Theme Change	The images, taken by Syd Shelton, from Pontefract, include pictures of The Clash, Misty in Roots and The Specials. The collection also features photos taken at the Rock Against Racism Carnival at Victoria Park, Hackney, which attracted a crowd of 100,000. The show runs from Friday to 3 September at the Impression Gallery. The Rock Against Racism (RAR) movement formed in response to controversial remarks made by Eric Clapton in 1976. In the following years, RAR staged marches, festivals and more than 500 concerts in the UK in a bid to fight racism through music. Shelton, who studied Fine Art in Leeds and Wakefield, said he became involved with the movement after returning to the UK from America in 1976. He said: Ï was appalled at the state of race relations in Britain, in particular things like the Black and White Minstrel Show and the signs I saw in some windows saying 'No Blacks, No Dogs, No Irish'. Ït was a pretty serious situation and I always loved music and very quickly hooked up with the people that had set up RAR. Ït was a bizarre mixture of people, photographers, graphic designers, writers, actors and, of course, musicians. Ẅe were very lucky in the sense that we tuned in to that explosion of punk and UK reggae and brought the two together. That said more about what RAR was about than any of the slogans we may have shouted from the stage.¨ He added: Ï hope the exhibition shows that you can change things and you can actually take a stand, even in the most difficult of situations. Ïf it inspires people to be photographers that would be great but I hope it will also inspire people to fight against racism and inequality.

Table 11: Examples of documents from the set of 15 with the lowest predicted scores by PreSumm, accompanied by their main characteristics that made them challenging for annotators to read.

Challenging Characteristics	Document
Content	Welsh language minister Alun Davies told AMs it would help efforts to reach that goal stay on the right track. Targets to meet growing demand for Welsh-speaking teachers and public sector workers will also be set. Culture committee chairwoman Bethan Jenkins said AMs had been told 70% more Welsh-medium teachers were needed. Mr Davies responded that around a third of teachers in Wales could speak Welsh, and that the challenge was to see if more of them would be willing to teach through the medium of Welsh. Earlier this month, Welsh language commissioner Meri Huws called for r̈adical changeïn the education system to ensure all children under the age of seven were ïmmersedïn Welsh.":
Content	He said new forests would slow flooding by trapping water with their roots. The idea of r̈ewildingẗhe uplands is catching on fast as parts of Britain face repeated flooding, with more rainfall on the way. Environment Secretary Owen Paterson said he would seriously consider innovative solutions like rewilding. The government has been criticised for being slow to capitalise on the benefits of capturing rain where it falls. Lord Rooker, a Labour peer, said too much emphasis had been attached to the look of the countryside rather than practical considerations like trapping water. Ẅe pay the farmers to grub up the trees and hedges; we pay them to plant the hills with pretty grass and sheep to maintain the chocolate box image, and then wonder why we've got floods,ḧe said. The idea of reintroducing forests into catchments has been strongly supported by several leading scientists. The government is sponsoring a handful of catchment trials to assess the potential of the upstream areas to catch water and send it slowly downhill. ...

Table 12: Examples of documents from the set of 15 with the lowest predicted scores by PreSumm, accompanied by their main characteristics that made them challenging for annotators to read.