REDTABS: A Collection of Report Document Datasets for Long Text and Multi-Table Summarization

Anonymous ACL submission

Abstract

Automatic document summarization aims to produce a concise summary covering the input document's salient content. Within a report document, both the textual and non-textual content (e.g., tables and figures) can be important information sources for the summary. However, most available document summarization datasets focus on the text and filter out the non-textual content. Missing tabular data can limit the informativeness of produced summaries, especially when target summaries require to cover quantitative descriptions of critical metrics, whose numerical information is usually kept in tables. In this paper, we address this issue by introducing REDTABS, the first collection of large-scale datasets for long text and multi-table summarization. Built on companies' annual reports, it includes three large-scale datasets for summarizing these companies' business, results of operations, and overall conditions, respectively. We also present the Segment-Alignment-based long Text and multi-Table summarization (SATT) method incorporating textual and tabular data into the summarization process. Besides, we propose a set of automatic evaluation metrics to assess the numerical information in summaries produced by summarization models. Dataset analyses and experimental results reveal the importance of incorporating textual and tabular data into the report document summarization. We will release our data and code to facilitate advances in summarization and text generation research.

1 Introduction

Automatic document summarization is the process of producing a concise summary covering the salient information within the input document. In recent years, both large-scale summarization datasets and the progress in neural summarization approaches boosted the progressive improvements in the quality of produced summaries.There have been various document summarization datasets collected from different domains, including the news articles (Hermann et al., 2015; Grusky et al., 2018; Fabbri et al., 2019), scientific literature (Cohan et al., 2018; Sharma et al., 2019; Lu et al., 2020), and law document (Eidelman, 2019).

These datasets usually regard input documents' non-textual content as noises and filter them out. When the target summaries only focus on the narratives and qualitative descriptions, removing non-textual content has little effect because textual content already contains most of the required information. However, when it comes to report documents, like companies' annual reports, their summaries should cover both the narrative content and quantitative descriptions of some critical metrics because their numerical information is crucial for readers' analysis and decision-making (SEC, 2021). As shown in Table 1, gold summaries in our collected financial reports usually contain more numerical information compared with that of previous summarization datasets. We also discover that about two-thirds of the numerical values in our gold summaries cannot be found in the corresponding input textual content. Missing tabular data can limit the informativeness of produced summaries, especially when target summaries require to cover quantitative descriptions of critical metrics, whose numerical information is usually kept in tables.

In this paper, we introduce REDTABS, the first collection of datasets for long text and multi-table summarization. To deal with the scarcity of available data, we develop a toolkit for extracting the textual and tabular data from numerous financial report documents and construct a corpus containing 21,125 annual reports from 3,794 companies. Based on this corpus, we build up three datasets named REDTABS-Overview, REDTABS-ROO, and REDTABS-MD&A for summarizing these companies' business, results of operations, and overall conditions, respectively. The average input text lengths of these three datasets range from 4,000 to 20,000 words. And the input also include

tens of tables' information in REDTABS-ROO and REDTABS-MD&A. The average gold summary lengths of these three datasets range from about 260 to nearly 1,500 words.

Because unstructured textual data and structured tabular data have different natures, it is not appropriate to directly treat tabular data as a part of the text sequence. To utilize tabular data in report documents, we solve a series of problems, including the data record extraction, salient content selection, incorporating both textual and tabular data into the summarization process. Meanwhile, we adopt the sparse attention mechanism and the segment-alignment-based training approach to deal with the long input and output text sequences.

We benchmark advanced extractive and abstractive summarization models as baselines on our three summarization datasets. To compare these models' performance, we conduct both automatic evaluation and human evaluation. In addition to the commonly used ROUGE scores (Lin, 2004), we also propose a set of automatic evaluation metrics to assess the numerical information in produced summaries. And experimental results show that our proposed method outperforms competitive baseline models when target summaries need to cover many quantitative descriptions.

Our contribution is threefold:

- We build REDTABS, the first collection of large-scale datasets for long text and multi-table summarization, and develop a report parsing toolkit to deal with the data scarcity.
- We propose a method incorporating textual and tabular data into summarization, and it outperforms other baselines on our datasets.
- We propose a set of evaluation metrics to evaluate the selection of financial data contained in produced summaries.

2 Related Work

2.1 Automatic Document Summarization

Previous text summarization methods can be generally classified into two categories: extractive and abstractive summarization methods. Extractive methods (Erkan and Radev, 2004; Mihalcea and Tarau, 2004; Liu and Lapata, 2019) select a subset of important sentences from input documents to form summaries. While abstractive methods (Rush et al., 2015; Nallapati et al., 2016; Chopra et al., 2016; Gehrmann et al., 2018; Zhang et al., 2020) capture and encode the salient content from input documents as the condition for generating novel sentences as summaries.

In addition to those datasets mentioned in the Introduction section, the Financial Narrative Summarisation (FNS) shared task in 2021 (Zmandar et al., 2021; El-Haj et al., 2020) delivered a dataset of annual reports from UK firms listed on the London Stock Exchange (LSE). But they still regard tabular data as noises and only focus on summarizing the narratives in companies' annual reports.

2.2 Table Summarization

There have been some existing datasets for table summarization or table-to-text generation, like the WEATHERGOV (Liang et al., 2009), WikiBio (Lebret et al., 2016), ROTOWIRE (Wiseman et al., 2017), and SBNATION (Wiseman et al., 2017). But they are usually limited to generating a short description for a single table with fixed schema, and they mainly focus on cell-level or row-level content selection from a single table. However, there could be tens of tables in annual report documents, and the gold summaries can combine the information from different tables, which makes the table-level content selection and integration very important. In addition to multi-table summarization, we observe that human-written summaries can combine the information from the input text content and tabular data from multiple tables within the report document. To fill in the gap between the properties of existing table summarization datasets and the actual requirements of report document summarization, we build a collection of new datasets for long text and multi-table summarization.

3 REDTABS Datasets

In this section, we first present our data sources and procedures of data collaboration and preprocessing. And then, we will introduce our three summarization datasets in REDTABS. We also conduct the descriptive statistics and in-depth analysis of these datasets and compare them with other commonly used document summarization datasets.

3.1 Data Source

We collected companies' annual reports on Form 10-K from the Electronic Data Gathering, Analysis, and Retrieval (EDGAR) system maintained by the U.S. Securities and Exchange Commission

Dataset	Pairs	Words (Doc)	Sents (Doc)	Words (Sum)	Sents (Sum)	Numbers (Sum)	% Covered Numbers	Cov.	Comp.
CNN/DM	312,085	810.6	39.8	56.2	3.7	0.57	78.65	0.85	13.0
PubMed	133,215	3049.0	87.47	202.4	6.8	3.29	68.21	0.79	16.2
arXiv	215,913	6029.9	205.67	272.7	9.6	0.71	53.93	0.87	39.8
REDTABS-Overview	21,125	4,000.03	125.45	264.99	8.13	3.27	36.75	0.88	20.31
REDTABS-ROO	21,024	20107.80	584.80	660.74	16.26	**24.32**	**26.25**	0.94	31.24
REDTABS-MD&A	21,125	19382.08	564.22	1495.44	42.70	**27.26**	**25.97**	0.94	12.97

Table 1: Statistical information of our three summarization datasets. "Pairs" denotes the number of examples. "Words" and "Sents" indicate the average number of words and sentences in input text or gold summary. "Numbers" represents the average number of numerical values included in the gold summary, and "Covered Numbers" is the proportion of the gold summary's numerical value that can be found in the input text. "Cov." is the extractive fragment coverage, and "Comp." is the compression ratio of gold summaries (Grusky et al., 2018).

(SEC)[1]. The SEC makes companies' regular filings available to the public through the EDGAR system. Among these filings, Form 10-K is the most commonly filed form. And it is the annual report that comprehensively describes a company's financial performance in the prior fiscal year (SEC, 2021).

The SEC stipulates the format and required content of the Form 10-K, which usually contains four parts and sixteen items (SEC). And the seventh item "Management's Discussion and Analysis of Financial Condition and Results of Operations" (MD&A) is the management's narrative of the results of operations, liquidity, and capital resources of a company, according to item 303 of Regulation S-K (NARA). The MD&A includes management's identification of the important information for the accurate understanding of the company's financial position and operating results (SEC, 2002). And we noticed that some parts of the MD&A can be regarded as summaries of the company's business, results of operations, and the overall financial and operations results in the annual report.

3.2 Data Collection and Pre-processing

In this subsection, we discuss our pipeline for data collection, pre-processing, and selection. Firstly, we collect the HTML files of 10-K forms from the EDGAR system and remove duplicate files. To parse these HTML files, we develop a rule-based parsing toolkit, which can extract the text and tables in each item of 10-K forms and remove noises (e.g., the cover pages before the first item and special characters used to compose a style). Compared with text extraction, tabular data extraction needs more steps, including tables' position identification, table content parsing, data format transformation, and alignment with text. After identifying a table in the report document, the toolkit can remove blank cells and convert each cell in the extracted table into a fixed format record, including the cell value and the name and index of the row and column to which it belongs. And we will extract and concatenate nested row names or column names. After separating each table from the text, we replace it with a special token containing the table's index in the original document to support the alignment of the text and table content. These extracted items' text and tables will be stored in separate JSON files. Some of them will be combined as inputs for our three summarization datasets. And we also extract some sub-sections in the MD&A as the gold summaries for these datasets.

We also conduct some operations to select samples, including removing outliers with the too-short input text, truncating the input text and gold summaries, and splitting the training (80%), validation (10%), and test (10%) sets. Considering that annual reports of the same company in different years have great overlaps, we divide these sets by company to minimize overlaps among these three sets.

3.3 Summarization Datasets

Based on the collected annual report corpus, we define three summarization datasets: REDTABS-Overview, REDTABS-ROO, and REDTABS-MD&A. And they aim to produce summaries of different lengths to summarize the report documents' different aspects of information. Their statistical information is shown in Table 1. We count the numerical information in gold summaries of these three datasets and find that a large proportion of them cannot be found in the text of input documents. Besides, gold summaries in REDTABS-ROO and REDTABS-MD&A contain much more numerical

[1]https://www.sec.gov/edgar/searchedgar/companysearch.html

information compared with that of previous summarization datasets. The following subsections will introduce the definitions and properties of these three datasets.

3.3.1 REDTABS-Overview

This dataset aims to produce the first three hundred words in the overview part of the MD&A, given the first four thousand words in Form 10-K (start from the first item named Business). We calculate input sections' recall of n-grams in the overview part of the MD&A, and find the overview is highly correlated with the first item named Business in Form 10-K. In our case study, we find that the first three hundred words in the overview usually summarize the company's business.

As shown in Table 1, this dataset's average input length and target output length are similar to previous academic literature summarization datasets. We further analyze gold summaries' content and find out they usually cover the narrative information and qualitative descriptions. We also count the numerical information in gold summaries and find few quantitative descriptions, which is also similar to previous summarization datasets. Considering these similarities, previous text summarization model should be able to adapt to this dataset.

3.3.2 REDTABS-ROO

The results of operations is a required part in MD&A, in which the company's management usually compares and explains the revenue and expense items in the current period and that of the prior period (SEC). As our statistical results reveal, the results of operations part usually contains a lot of numerical information, which is necessary for readers to analyze the company and make investment decisions. Compared with the previous summarization datasets, the REDTABS-ROO requires summarization systems to produce more quantitative descriptions of some critical metrics. However, nearly three-quarters of the numerical information cannot be found in the input textual content, as shown in Table 1. Because some of the critical numerical information is kept in the tabular data, it is appropriate to incorporate them into summarization models as a part of the input.

We define the summarization task on the REDTABS-ROO dataset as generating the results of operations part in MD&A, given the text content from the rest parts and all tables in the report. To prepare the input data, we truncate the text before and after the results of operations part and concatenate the two truncated parts as the input text. Text and table selection will be discussed in 4.1. Considering unstructured textual data and structured tabular data have different natures, it brings new challenges for text summarization models to utilize the tabular data.

3.3.3 REDTABS-MD&A

Gold summaries in previous summarization datasets usually contain no more than a few hundred words, which can be enough for documents that are not too long. For the extreme long documents (e.g., books or financial reports), their summaries should be longer than that of short documents. And the extended summary should comprehensively cover different aspects of information in long documents. In our REDTABS-MD&A dataset, gold summaries are the first 1,500 words of the original MD&A item, and the inputs include the textual content from other items and all the tables in the Form 10-K. Long inputs and outputs will bring a series of challenging issues: selecting salient content from the long text and multiple tables, ensuring the informativeness, fluency, and non-redundancy of the produced summaries, and improving the efficiency of training and inference.

3.4 Evaluation Metrics

In addition to the commonly used evaluation metrics, like ROUGE scores (Lin, 2004), we propose a set of evaluation metrics to evaluate the selection of numerical information in produced summaries. We use D, S, and H to denote the input document, human-written gold summary, and the summary produced by a summarization model. D_n, S_n, and H_n represent sets of numbers contained in them, and $|D_n|$, $|S_n|$, $|H_n|$ denote these number sets' size. For a produced summary H, we first extract the number set H_n from it.[2] An then, we check if these numbers are contained in the gold summary S and $M(H_n, S_n)$ count numbers contained in both the produced summary H and the gold summary S. $M(D_n, S_n)$ count numbers contained in both the input document D and the gold summary S.

We mainly consider three metrics: Number Precision (NP), Number Coverage (NC), and Number Selection (NS). Number precision is the ratio of numbers in the produced summary that also appears in the gold summary. Calculated by Equation

[2]We do not count the date and numbers belonging to a word, like covid-19.

(1), it measures how well the produced summary matches the gold summary in terms of contained numbers. Number coverage measures how well the produced summary covers the numbers in the gold summary that also appear in the input document. Some of the numbers in the gold summary cannot be directly found in the input document (including textual and tabular data) and need numerical reasoning, and some of them may be lost when preparing the summarization model's inputs, which can limit produced summary's number recall computed by Equation (2b). To evaluate the summarization model's coverage capability, we divide the produced summary's number recall by the input document's number recall in Equation (2c). Number selection calculates the harmonic mean of NP and NC in Equation (3) and reflects the quality of number selection in the produced summary.

$$\mathrm{NP}(H_n, S_n) = \frac{M(H_n, S_n)}{|H_n|} \tag{1}$$

$$\mathrm{NC}(D_n, H_n, S_n) = \frac{\mathrm{NR}(H_n, S_n) * |S_n|}{M(D_n, S_n)} \tag{2a}$$

$$\mathrm{NR}(H_n, S_n) = \frac{M(H_n, S_n)}{|S_n|} \tag{2b}$$

$$\mathrm{NC}(D_n, H_n, S_n) = \frac{M(H_n, S_n)}{M(D_n, S_n)} \tag{2c}$$

$$\mathrm{NS}(D_n, H_n, S_n) = \frac{2 * \mathrm{NP} * \mathrm{NC}}{\mathrm{NP} + \mathrm{NC}} \tag{3}$$

4 Summarization Method

Our datasets bring several challenging issues for summarization methods, including selecting the salient content from the extremely long input text sequence and multiple tables, incorporating the unstructured textual and structured tabular data into the summarization model, and efficiently processing the long input and output sequences. To demonstrate the usage of REDTABS datasets, we propose the Segment Alignment based long Text and multi-Table summarization (SATT) method. And we will present our approaches to deal with above challenging issues in following subsections.

4.1 Content Selection

In long document summarization, adding a content selection step to prepare the inputs for the abstractive summarization model is important. Accurately selecting the salient content can be challenging, especially when the salient content is scattered in different sections of text and multiple tables.

After these preprocessing steps discussed in subsection 3.2, we can get the long text from multiple sections and numerous data records from tens of tables in a report document. In our experiments, we truncate and segment the input long text to a limited length and select important data records that should be mentioned in produced summaries. Since the important financial data can be scattered in different tables, finding them from numerous extracted records can be challenging. We select important records by using pre-defined rules. For the multiple tables in the report document, our rule preference tables appearing near the target summary text since they are most likely to be mentioned in the target summary. Within one table, we select columns describing data for the most recent two fiscal years because the management is required to discuss the material changes of financial condition and results of operations from the preceding year in the MD&A (SEC).

4.2 Incorporating Textual and Tabular Data

To incorporate unstructured textual data and structured tabular data into the summarization process, we combine the template-based data-to-text generation method and the neural summarization method. We discover that most table cells usually describe values or changes of accounting elements, including asset, liability, equity, revenue, and expense. And there are limited types of relations between the cell value and corresponding name of row or column, which can be classified and handled by predefined rules. Therefore, we utilize a data-to-text generation method based on predefined templates to covert each selected data record to a sentence, and these sentences will be concatenated as the result of data-to-text generation. After that, we will concatenate the selected input text and results of data-to-text generation as the input of the neural summarization model.

4.3 Dealing with Long Inputs and Outputs

Previous abstractive text summarization methods usually focus on generating short summaries containing no more than a few hundred words for the input document containing hundreds or thousands of words. On the contrary, input documents in our REDTABS-ROO and REDTABS-MD&A datasets contain tens of thousands of words, and the average length of gold summaries in REDTABS-MD&A

is close to 1,500 words. Extremely long inputs and outputs will bring a series of problems: 1) The widely used self-attention mechanism in the transformer model (Vaswani et al., 2017) scales quadratically with the number of tokens in the input sequence, which is prohibitively expensive for long input (Choromanski et al., 2020) and precludes the usage of large pre-trained models with limited computational resources. 2) The salient content can be scattered in different parts of the long input documents, making it more challenging to identify and cover them. 3) Long target summaries can contain multiple sections written in different ways and focus on various aspects of information. Previous abstractive summarization methods usually adopt autoregressive decoding, which has difficulty in generating well-organized long summaries.

To deal with the first problem, we adopt the sparse attention mechanisms, including the random attention, window local attention, and global attention in (Zaheer et al., 2020) in our encoder part. Although sparse attention mechanisms can reduce the complexity of transformer-based summarization models, it is still difficult to directly train large models on extremely long inputs with limited GPU memory resources. Therefore, we further decompose the problem of long summary generation into multiple sub-problems of summary segment generation and adopt the segment-alignment-based training approach to train multiple summarization models generating different summary segments in parallel. During inference, the output summary segments of these models will be concatenated to form the final summary.

After dividing the long input document and summary into several segments, we need to find multiple input segments that should be aligned with each summary segment to form a sample pair. In our experiments, we first calculate the proportion of each summary segment's n-grams [3] that can be covered by each input segment. And we discover that all summary segments have strong connections with the first input segment, which usually briefly introduces the company's business. The reason is that all discussions on the company's financial condition should be based on the company's business. In addition to the first input segment, we match another input segment for each summary segment with a greedy method and avoid reusing other input segments to minimize repetitions and maximize the informativeness of the merged summary.

[3]We use the average recall of unigram, bigram, trigram, and 5-gram

Method	R-1	R-2	R-L
LexRank	36.01	10.69	18.11
TextRank	36.07	10.20	18.40
BertExt	25.53	11.45	21.18
CopyTransformer	47.01	26.05	32.30
BertAbs	45.71	**33.56**	**42.84**
BART	**49.71**	28.27	34.24
PEGASUS	48.00	24.70	30.78
BigBird-PEGASUS	48.65	25.59	31.70

Table 2: Evaluation results on the REDTABS-Overview test set.

The segment-alignment-based training approach parallelizes model training and inference. It brings several benefits, including reducing the requirements for memory size, improving the efficiency of the summarization method, and making full use of advanced pre-trained models, which can effectively model hundreds or thousands of tokens.

5 Experiments

5.1 Baselines

In our experiments, we compare different types of baseline models, including extractive and abstractive summarization models.

LexRank and TextRank (Erkan and Radev, 2004; Mihalcea and Tarau, 2004) are two graph-based ranking methods that can be used for unsupervised extractive summarization.

BertExt (Liu and Lapata, 2019) stacks additional transformer layers on top of the BERT model (Devlin et al., 2019) to capture document-level features for extractive summarization.

CopyTransformer (Gehrmann et al., 2018; Fabbri et al., 2019) adds the copy mechanism (See et al., 2017) to the transformer model (Vaswani et al., 2017) for abstractive summarization.

BertAbs (Liu and Lapata, 2019) utilizes the BERT model as the encoder and a randomly initialized decoder for abstractive summarization.

BART (Lewis et al., 2020) is a denoising autoencoder built with a sequence-to-sequence model that is pretrained to reconstruct the original input text from the corrupted text.

PEGASUS (Zhang et al., 2020) is a transformer-based model pretrained with the Gap Sentences Generation (GSG) and Masked Language Model (MLM) objectives.

REDTABS-ROO Dataset									
Method	**Segment 1**			**Segment 2**			**Combined**		
	R-1	**R-2**	**R-L**	**R-1**	**R-2**	**R-L**	**R-1**	**R-2**	**R-L**
LexRank	34.64	8.88	16.42	35.73	9.76	17.20	34.43	7.73	14.92
TextRank	35.15	9.06	16.65	36.00	9.79	17.20	35.93	7.74	15.08
BART	43.13	13.82	21.10	40.99	11.74	18.38	49.00	16.88	19.14
PEGASUS	44.79	15.17	21.53	44.46	14.21	19.70	51.92	19.31	21.47
BigBird-PEGASUS	46.25	16.78	22.67	45.34	15.28	20.23	53.08	20.85	20.94
SATT (Our Method)	**47.29**	**17.84**	**23.20**	**46.56**	**16.40**	**20.79**	**54.31**	**22.18**	**23.02**
REDTABS-MD&A Dataset									
Method	**Segment 1**			**Segment 2**			**Segment 3**		
	R-1	**R-2**	**R-L**	**R-1**	**R-2**	**R-L**	**R-1**	**R-2**	**R-L**
LexRank	38.87	13.57	19.82	33.53	8.12	16.35	31.10	6.29	2.70
TextRank	39.93	14.78	21.63	34.13	8.62	17.00	31.49	6.09	15.05
CopyTransformer	45.99	23.39	29.59	25.40	6.10	13.44	20.86	3.88	5.57
BART	**47.08**	**23.93**	**30.23**	37.10	10.88	18.56	34.68	8.03	16.73
PEGASUS	46.59	22.11	28.34	37.09	10.17	17.84	35.49	8.09	16.36
BigBird-PEGASUS	46.97	22.53	28.85	**38.38**	**11.30**	**18.97**	**35.96**	**8.47**	**16.82**
Method	**Segment 4**			**Segment 5**			**Combined**		
	R-1	**R-2**	**R-L**	**R-1**	**R-2**	**R-L**	**R-1**	**R-2**	**R-L**
LexRank	29.25	4.99	14.15	29.08	4.51	13.88	44.34	13.90	17.79
TextRank	29.93	5.10	14.46	30.05	4.75	14.19	44.33	12.91	18.16
CopyTransformer	22.00	3.91	11.71	17.79	3.44	10.20	29.52	13.54	16.26
BART	33.11	6.29	15.34	33.36	6.14	15.30	50.91	19.14	21.85
PEGASUS	34.80	7.13	15.74	35.80	7.36	16.03	53.64	19.98	20.88
BigBird-PEGASUS	35.19	7.61	16.14	35.83	7.37	16.02	54.72	21.28	21.84
SATT (Our Method)	**35.61**	**7.79**	**16.21**	**36.25**	**7.68**	**16.19**	**55.24**	**21.67**	**21.91**

Table 3: Automatic evaluation results of each output summary segment and final combined summary on test sets of REDTABS-Result and REDTABS-MD&A.

	NP	**NC**	**NS**
REDTABS-ROO			
SATT (Our Method)	**15.72**	**30.81**	**20.82**
BigBird-PEGASUS	13.15	23.82	16.95
PEGASUS	10.90	21.89	14.55
LexRank	14.68	10.96	12.55
TextRank	14.77	9.73	11.73
Last two summary segments in REDTABS-MD&A			
SATT (Our Method)	**10.76**	**17.21**	**13.24**
BigBird-PEGASUS	9.74	15.90	12.08
PEGASUS	9.37	14.88	11.50
LexRank	12.15	11.18	11.64
TextRank	11.31	11.63	11.47

Table 4: Evaluation results of numbers in produced summaries. Columns indicate: Number Precision (NP), Number Coverage (NC), and Number Selection (NS).

BigBird-PEGASUS (Zaheer et al., 2020) combines the BigBird encoder with the decoder from the PEGASUS model (Zhang et al., 2020).

5.2 Results and Discussion

In this section, we present and analyze our experimental results. To compare the quality of summaries generated by different models, we first conduct automatic evaluation and report the ROUGE F_1 scores (Lin, 2004), including the overlap of unigrams (R-1), bigrams (R-2), and longest common subsequence (R-L).

We train and evaluate various advanced summarization methods on the REDTABS-Overview dataset and do not introduce the tabular data because the gold summaries in this dataset have few quantitative descriptions, as discussed in subsection 3.3.1. Table 2 shows that these abstractive summarization models significantly outperform these extractive models on the REDTABS-Overview dataset. And pretrained models modeling longer context do not show obvious advantages since gold summaries in this dataset usually focus more on the beginning of input documents.

We adopt our SATT method on the REDTABS-Result dataset and the last two summary segments of REDTABS-MD&A dataset since their target summaries usually contain a lot of numerical information and quantitative descriptions. We report ROUGE scores of produced summary segments and the final combined summary for the REDTABS-Result dataset and the REDTABS-MD&A dataset. As shown in Table 3, our model significantly outperforms other baseline models thanks to the incorporation of tabular information. And the baseline model with the sparse attention

	Win	Lose	Tie	Kappa
REDTABS-ROO				
Informativeness	43.5%	21.0%	35.5%	0.657
Fluency	28.0%	26.5%	45.5%	0.617
Non-Redundancy	30.0%	25.5%	44.5%	0.634
REDTABS-MD&A				
Informativeness	40.5%	22.0%	37.5%	0.660
Fluency	27.0%	25.5%	47.5%	0.623
Non-Redundancy	29.5%	27.5%	43.0%	0.621

Table 5: Human evaluation results. "Win" represents the generated summary of our proposed method SATT is better than that of BigBird-PEGASUS in one aspect.

	R-1	R-2	R-L
REDTABS-ROO			
SATT	**54.31**	**22.18**	**23.02**
w/o tabular data	53.08	20.85	20.94
w/o sparse attn	51.92	19.31	21.47
w/o input text	25.54	6.11	12.34
REDTABS-MD&A			
SATT	**55.24**	**21.67**	**21.91**
w/o tabular data	54.72	21.28	21.84
w/o sparse attn	53.64	19.98	20.88

Table 6: Ablation study on the test sets of REDTABS-ROO and REDTABS-MD&A. We report the ROUGE scores of merged summaries here.

mechanism (Zaheer et al., 2020), which enables the modeling of longer context, also shows its advantage on these two datasets since the target summaries need to cover more content scattered in different parts of input segments. The only exception is the first summary segment in the REDTABS-MD&A dataset, whose gold summaries' properties are similar to that of gold summaries in the REDTABS-Overview dataset.

We also calculate the three metrics presented in subsection 3.4, for assessing selection of numerical information in summaries produced by these summarization models. The results shows that our proposed method outperforms other baseline models, and introducing the tabular data can improve the produced summaries' coverage of numbers in gold summaries of the REDTABS-Result dataset and the REDTABS-MD&A dataset.

In addition to automatic evaluation, we performed a human evaluation to compare the generated summaries in terms of informativeness (the coverage of information from input documents), fluency (content organization and grammatical correctness), and non-redundancy (less repetitive information). We randomly selected 50 samples from the test set of the REDTABS-ROO dataset and the REDTABS-MD&A dataset, respectively. Four annotators are required to compare two models' generated summaries that are presented anonymously. We also assess their agreements by Fleiss' kappa (Fleiss, 1971). The human evaluation results in Table 5 exhibit that our method SATT significantly outperforms the BigBird-PEGASUS in terms of informativeness and is comparative in terms of fluency and non-redundancy.

We also conduct the ablation study to validate the effectiveness of individual components in our proposed method. In Table 6, "w/o tabular data" refers to the BigBird-PEGASUS model (Zaheer et al., 2020). The results confirm that incorporating the tabular data is beneficial for report document summarization, and the sparse attention in the encoder also benefits our model's performance. Besides, we tried only using the data-to-text generation result as the produced summary, which is represented by "w/o input text". The performance degradation reveals that it is important to incorporate the textual and tabular data into the summarization model.

In our experiments, we also found current neural abstractive summarization models still have flaw in the fidelity to the input content, which can cause errors in produced summaries. The segment-alignment-based training approach can make it possible to generate long summaries with existing summarization models. But it will also bring some challenging issues, including minimizing the loss of salient information, repetition avoidance, and ensuring the cohesion between adjacent paragraphs. We will do further research and improve our method in future work.

6 Conclusion

In this paper, we introduce REDTABS, the first collection of datasets for long text and multi-table summarization, which are built on companies' annual reports. REDTABS includes three datasets for summarizing these companies' business, results of operations, and overall conditions. We present the Segment-Alignment-based long Text and multi-Table summarization (SATT) method and a set of evaluation metrics for assessing the numerical information in produced summaries. Dataset analyses and experimental results reveal the importance of incorporating textual and tabular data into the report document summarization.

References

S. Chopra, M. Auli, and Alexander M. Rush. 2016. Abstractive sentence summarization with attentive recurrent neural networks. In *HLT-NAACL*.

Krzysztof Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Lukasz Kaiser, et al. 2020. Rethinking attention with performers. *arXiv preprint arXiv:2009.14794*.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A discourse-aware attention model for abstractive summarization of long documents. In *Proceedings of NAACL-HLT*, pages 615–621.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Vladimir Eidelman. 2019. Billsum: A corpus for automatic summarization of us legislation. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 48–56.

Mahmoud El-Haj, Ahmed AbuRa'ed, Marina Litvak, Nikiforos Pittaras, and George Giannakopoulos. 2020. The financial narrative summarisation shared task (FNS 2020). In *Proceedings of the 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation*, pages 1–12, Barcelona, Spain (Online). COLING.

Günes Erkan and Dragomir R Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of artificial intelligence research*, 22:457–479.

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.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Sebastian Gehrmann, Yuntian Deng, and Alexander Rush. 2018. Bottom-up abstractive summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4098–4109, Brussels, Belgium. Association for Computational Linguistics.

Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 708–719, New Orleans, Louisiana. Association for Computational Linguistics.

Karl Moritz Hermann, Tomáš Kočiskỳ, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Proceedings of the 28th International Conference on Neural Information Processing Systems-Volume 1*, pages 1693–1701.

Rémi Lebret, David Grangier, and Michael Auli. 2016. Neural text generation from structured data with application to the biography domain. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1203–1213, Austin, Texas. 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.

Percy Liang, Michael Jordan, and Dan Klein. 2009. Learning semantic correspondences with less supervision. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 91–99, Suntec, Singapore. Association for Computational Linguistics.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3721–3731.

Yao Lu, Yue Dong, and Laurent Charlin. 2020. Multi-XScience: A large-scale dataset for extreme multi-document summarization of scientific articles. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8068–8074, Online. Association for Computational Linguistics.

Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 404–411.

Ramesh Nallapati, Bowen Zhou, C. D. Santos, Çaglar Gülçehre, and B. Xiang. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. In *CoNLL*.

NARA. Regulation s-k item 303 management's discussion and analysis of financial condition and results of operations [online].

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

United States SEC. Form 10-k general instructions [online].

United States SEC. Commission statement about management's discussion and analysis of financial condition and results of operations [online]. 2002.

United States SEC. How to read a 10-k/10-q [online]. 2021.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

Eva Sharma, Chen Li, and Lu Wang. 2019. Bigpatent: A large-scale dataset for abstractive and coherent summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2204–2213.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Sam Wiseman, Stuart M Shieber, and Alexander M Rush. 2017. Challenges in data-to-document generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2253–2263.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. 2020. Big bird: Transformers for longer sequences. In *NeurIPS*.

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.

Nadhem Zmandar, Mahmoud El-Haj, Paul Rayson, Ahmed Abura'Ed, Marina Litvak, Geroge Giannakopoulos, and Nikiforos Pittaras. 2021. The financial narrative summarisation shared task FNS 2021. In *Proceedings of the 3rd Financial Narrative Processing Workshop*, pages 120–125, Lancaster, United Kingdom. Association for Computational Linguistics.

A Appendix

A.1 Corpus Statistics

We do statistics on the length of the parsed text and the number of tables in our collected annual report corpus. Table 1 summarizes the statistical information of our corpus.

Item	Words	Sents	Tables
MD&A	11,108.46	299.94	11.10
Overview in MD&A	1,726.96	46.68	1.26
ROO in MD&A	1,038.21	25.48	2.00
Before MD&A	21,855.03	620.94	5.14
After MD&A	23,711.04	629.55	42.5

Table 7: Statistical information of our corpus. "Words" and "Sent" denote the average number of words and sentences in the different parts of parsed text. "Tables" represents the average number of parsed tables. "ROO" is the results of operations section

A.2 Diversity Analysis of Datasets

To measure how abstractive our summaries are, we report the percentage of gold summaries' n-grams, which do not appear in input documents. Table 8 shows that our three datasets are similar to the abstractiveness of previous document summarization datasets.

Dataset	% of novel n-grams in gold summary			
	unigrams	bigrams	trigrams	4-grams
CNN/DM	19.50	56.88	74.41	82.83
PubMed	18.38	49.97	69.21	78.42
arXiv	15.04	48.21	71.66	83.26
Overview	21.69	46.25	58.53	60.67
ROO	17.79	50.59	72.13	81.66
MD&A	14.59	44.64	62.42	69.49

Table 8: The proportion of novel n-grams in gold summaries across different summarization datasets.

On the other hand, we also adopt three measures defined by Grusky et al. (2018) for assessing the extractive nature of specific dataset. In Equation (4a), extractive fragment coverage measures the percentage of words in the summary that are part of an extractive fragment from the input document. And Equation (4b) calculates the extractive fragment density for assessing the average length of the extractive fragment to which each word in the summary belongs. Besides, the compression ratio is the word ratio between the articles and its summaries, as shown in Equation (4c).

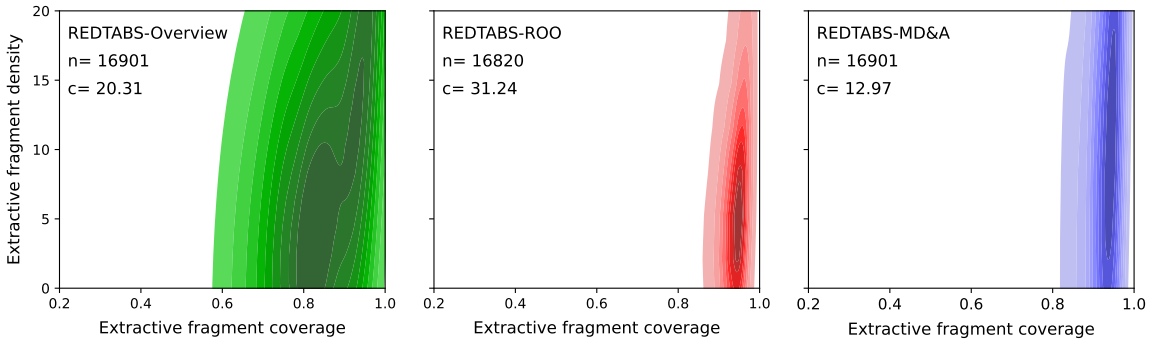

Figure 1: Density and coverage distributions of three datasets.

$$\text{COVERAGE}(D,S) = \frac{1}{|S|} \sum_{f \in F(D,S)} |f| \quad \text{(4a)}$$

$$\text{DENSITY}(D,S) = \frac{1}{|S|} \sum_{f \in F(D,S)} |f|^2 \quad \text{(4b)}$$

$$\text{COMPRESSION}(D,S) = \frac{|D|}{|S|} \quad \text{(4c)}$$

Results of these three measures are visualized using kernel density estimation. Figure 1 shows that the coverage and density of our datasets are higher than that of previous datasets. And the variability along the y-axis (extractive fragment density) suggests varying styles of word sequence arrangement in gold summaries.

A.3 An Example in REDTABS

Figure 2 shows an example in the REDTABS-ROO dataset, in which the content of gold summary come from two different paragraphs and three tables in the same report document. And it is collected from Apple's Form 10-K in 2017.

Gold Summary

Net sales increased 6% or $13.6 billion during 2017 compared to 2016, primarily driven by growth in Services, iPhone, and Mac. The year-over-year increase in net sales reflected growth in each of the geographic operating segments, with the exception of Greater China. The weakness in foreign currencies relative to the U.S. dollar had an unfavorable impact on net sales during 2017 compared to 2016. In May 2017, the Company announced an increase to its capital return program by raising the expected total size of the program from $250 billion to $300 billion through March 2019. This included increasing its share repurchase authorization from $175 billion to $210 billion and raising its quarterly dividend from $0.57 to $0.63 per share beginning in May 2017. During 2017, the Company spent $33.0 billion to repurchase shares of its common stock and paid dividends and dividend equivalents of $12.8 billion. Additionally, the Company issued $24.0 billion of U.S. dollar-denominated term debt, €2.5 billion of euro-denominated term debt and C$2.5 billion of Canadian dollar-denominated term debt during 2017.

Inputs

Text fragments

Weakening of foreign currencies relative to the U.S. dollar adversely affects the U.S. dollar value of the Company's foreign currency-denominated sales and earnings ..... In May 2017, the Company's Board of Directors increased the total capital return program from $250 billion to $300 billion, which included an increase in the share repurchase authorization from $175 billion to $210 billion of the Company's common stock. Additionally, the Company announced that the Board of Directors raised the Company's quarterly cash dividend from $0.57 to $0.63 per share, beginning with the dividend paid during the third quarter of 2017.

Table 1

Sales Data

	2017	Change	2016
Net Sales by Operating Segment:			
Greater China	44,764	(8)%	48,492
Total net sales	229,234	6%	215,639

Table 2

CONSOLIDATED STATEMENTS OF CASH FLOWS

	Years ended September 30, 2017
Financing activities:	
Payments for dividends and dividend equivalents	(12,769)
Repurchases of common stock	(32,900)

Table 3

Term Debt

	2017 Amount (in millions)
Third quarter 2017 euro-denominated debt issuance of €2.5 billion:	
Fixed-rate 0.875% notes	1,469
Fourth quarter 2017 Canadian dollar-denominated debt issuance of C$2.5 billion:	
Fixed-rate 2.513% notes	2,017

Figure 2: One example in the REDTABS-ROO dataset. Part of AAPL 2017 Form 10-K