ECTQA: Expanding Earnings Call Transcript Summarization

Andrew Abrahamian
abrahaa@berkeley.edu

Theresa Azinge
ashioma@berkeley.edu

Abstract

There has been significant progress in automatic text summarization, yet financial document summarization is still an emerging field of study, particularly for public companies' earnings calls. Techniques to efficiently and effectively summarize company earnings calls in a comprehensive yet consumable manner are still in their infancy. In this paper, we present **ECTQA**, a novel dataset of the Q&A sections of *earnings call transcripts* (ECTs) hosted by publicly-traded companies and reference summaries generated by an unsupervised salient sentence extractive model. These Q&A sections are typically 90% of any given earnings call and are largely under-analyzed by academic research and financial news organizations. We also present methods to fine-tune the **ECT-BPS** modeling framework that generates both extractive and abstractive paraphrased summaries of the *Prepared Remarks* of ECTs with the intent to apply it to our novel **ECTQA** dataset.

1 Introduction

Earnings Calls are hosted by publicly-traded companies to discuss important aspects of their quarterly or annual earnings reports. These calls typically include a *Prepared Remarks* section, where a representative from the company (often its Chief Executive Officer or CEO) shares relevant financial and operational metrics with an audience of Wall Street analysts, and a *Question and Answer* section, where the CEO answers questions from those same analysts.

Information from these calls is valuable for individual investors, corporations, and governments. The application of this information includes, but is not limited to, investment analysis, mergers and acquisitions, and auditing. Corresponding earnings call transcripts (abbreviated to **ECTs**) are typically long, unstructured documents. Coupled with the large number of public corporations providing these quarterly and annual reports, identifying the relevant and valuable information within these ECTs is a tedious and resource-intensive task.

Text summarization models can be useful for creating summaries from these lengthy ECTs while preserving the valuable information contained in the original document. However, both the domain-specific nature of ECTs and the fact that they are long, unstructured documents makes off-the-shelf pre-trained summarization models a poor fit for this task.

The ECTSum paper (Mukherjee et al., 2022) presents a valuable benchmark dataset, **ECTSum**, that contains **2,425 document-summary pairs** of ECTs and accompanying reference summaries sourced from *Reuters*. However, the dataset does not include the *Question and Answers* (Q&A) section of ECTs, focusing only on the *Prepared Remarks* of the public company's management. The paper also presents **ECT-BPS**, an effective approach to summarize earnings call transcripts with an extractive model followed by an abstractive model, termed the *Extractive Module* and *Paraphrasing Module* respectively.

We expand on the ECTSum paper by collecting and processing the Q&A sections of 182 unique ECTs downloaded from the Thomson Reuters Eikon service (Roozen and Lelli, 2021), a dataset originally intended for financial sentiment analysis. We then introduce an unsupervised extractive summarization approach (Gao et al., 2020) on these Q&A documents to generate reference summaries, resulting in the **182 document-summary pairs of the ECTQA dataset**.

We then focus on improving the ECT-BPS model framework for our task by fine-tuning different pre-trained summarization models to produce concise, informative, representative, and accurate summaries of ECTs inclusive of the Q&A sections.

Our contributions can be summarized as follows:

- We present **ECTQA**, a long document summarization dataset in the financial domain that includes Q&A sections of earnings call transcripts, extremely long and unstructured documents that require models to process and summarize them while maintaining relevance and factual consistency.

- We present an unsupervised approach to generate ground truth reference summaries using a salient sentences extractive model and is intended to evaluate our model-generated summaries.

- We introduce fine-tuning approaches of the **ECT-BPS** model framework on the original **ECTSum** dataset to generate concise and valuable summaries of the documents in the **ECTQA** dataset.

2 Background and Related Work

Various methods of automatic text summarization, including extractive (Zhong et al., 2020), abstractive (Zhang et al., 2020a; Lewis et al., 2020), and long document summarization (Beltagy et al., 2020) have seen significant progress in recent years. The field of financial data summarization has seen some advancements recently, with the Financial Narrative Summary (FNS) (El-Haj et al., 2020) released in 2020 containing extractive narratives as summaries of UK company annual reports.

ECTSum (Mukherjee et al., 2022) is by far the most comprehensive summarization effort aimed at company earnings call transcripts. The paper presents **ECTSum**, a dataset with transcripts of earnings calls (ECTs) and expert-written bullet-point summaries sourced from corresponding Reuters articles. The paper also presents **ECT-BPS**, a model with an extractive module and paraphrasing module that respectively use FinBERT[1] and T5 parameters initialized from HuggingFace. We have chosen to build on the advances of this paper by the methods described in subsequent paragraphs.

First, the ECTSum benchmark dataset only consists of the *Prepared Remarks* section of the earnings call. Typically, this is the shortest section of any public company's earnings call, with executives often spending the vast majority of their time on the investor *Question and Answer* section (Q&A). We have collected statistics that show earnings call Q&A has approximately 9x more tokens than prepared remarks. Some public companies, like Amazon or Tesla, have abandoned prepared remarks on their earnings calls altogether, opting to spend all of the time on investor Q&A.

We introduce the **ECTQA** dataset by collecting 182 unique ECTs from Thomson Reuters Eikon service with the specific intent to summarize the Q&A sections of these transcripts. Focusing on summarizing Q&A, the longest part of every single earnings call, maximizes our model's usefulness for stakeholders.

Second, the ECTSum dataset relies on reference summaries collected from *Reuters* which exclusively focus on the *Prepared Remarks* of the earnings call. Such reference summaries do not exist for the Q&A section of these calls, as the financial press typically focus on key metrics like Revenue, Net Income, and Earnings Per Share, rather than the more qualitative insights that emerge during Q&A sessions.

We introduce the use of unsupervised document summarization (Wu et al., 2020; Gao et al., 2020) on ECTs to generate reference summaries to evaluate our candidate summaries of the **ECTQA** documents.

Lastly, we experiment with fine-tuning the ECT-BPS *Paraphrasing Module* by initializing BART[2], T5[3], and PEGASUS[4] pre-trained models. These transformer models are meant for long document summarization and have been pre-trained on financial data.

We introduce pre-trained model and hyperparameter fine-tuning with the **2,425 document-summary pairs in the ECTSum dataset** to generate useful summaries from the **182 document-summary pairs in the ECTQA dataset**.

3 Methods

3.1 Structure of the ECT-BPS Model Framework

For developing ECT summaries similar to those developed by Reuters analysts, the ECT-BPS framework consists of two separate modules: (1) an *extractive* module to select the most relevant sentences from the ECT document, and (2) a *paraphrasing* module to rephrase the extracted sentences into a telegram-style document similar to that developed by the analysts at media houses like *Reuters*.

[1] https://huggingface.co/ProsusAI/finbert

[2] https://huggingface.co/eugenesiow/bart-paraphrase

[3] https://huggingface.co/ramsrigouthamg/t5_paraphraser

[4] https://huggingface.co/tuner007/pegasus_qa

The *extractive* module is based on the architecture of SummaRuNNer (Nallapati et al., 2016) which uses a two-layer bi-directional GRU-based RNN. The first layer is FinBERT (Araci, 2019), a BERT model trained on large financial communication texts, and is run at the word-level to get the hidden state representations. The second layer takes the average pool of the hidden states from the first layer as its sentence-level input. The document is now represented as a non-linear transformation of the concatenation of these average-pooled hidden states of sentences. At the classification layer, each sentence is passed through a binary decision to determine its inclusion in the summary. This binary decision is based on the richness of the content of the sentence, its salience in the context of the document, and its novelty based on the sentences already included in the summary.

The *paraphrasing* module takes the output of the extractive module and paraphrases it to the Reuters format with telegram-style bullet points while ensuring that numerical values are correctly rephrased to minimize value hallucination.

3.2 Training the Extractive and Paraphrasing Modules

We trained the extractive module by using a greedy search to select all sentences in the source ECT document that contained numerical values mentioned in each target sentence in the reference Reuters summary. In cases where there were no exact matches, the closest match (using cosine similarity) in the source document was selected. The sentences selected from this search were used as the target summary for the extractive module, and the model was trained by minimizing the binary cross-entropy loss between the predicted summary and the target summary.

For the paraphrasing module, we use each sentence from the extractive module target summary as the input and the corresponding sentence in the reference summary as the target. We generate the **ECTQA** reference summary using the salient sentence extractor in the **SUPERT** benchmark paper (Gao et al., 2020). The SUPERT extractor creates contextualized embeddings for each sentence in the ECTQA input document with SentenceBERT (Reimers and Gurevych, 2019) and then clusters them with an affinity propagation algorithm (Frey and Dueck, 2007). We use the center of each cluster to build our annotations and create the dataset of document-summary pairs. Then, we train the model by minimizing the cross entropy loss between the predicted sentence and the target reference sentence.

3.3 Generalizing to a Novel Dataset: ECTQA

This subsection describes our novel dataset, **ECTQA**, including data sources and steps taken to clean and process the data.

3.3.1 Data Collection

ECTs of listed companies are publicly available on websites like The Motley Fool[5] or SeekingAlpha[6], but require resources invested in web scraping. We sourced 182 unique ECTs from a public dataset for financial sentiment analysis (Roozen and Lelli, 2021) created via the Thomson Reuters Eikon service. These ECTs consist of two sections: *Prepared Remarks* where the company presents its financial results for the reporting period, and *Question and Answers*, where representatives from Wall Street investment banks ask questions regarding presented results and other dynamics affecting the company's financial performance.

3.3.2 Data Cleaning

The source documents are formatted with headings for each section as well as labels indicating the speaker and their organizational affiliation. We remove these headings, identifying labels and the *Prepared Remarks* section from each ECT.

3.3.3 Unsupervised Summarization for Ground Truth

The candidate ECT documents did not have accompanying reference summaries, which are necessary to evaluate our model-generated summaries. We first attempted to find summaries of these ECTs online but were met with two challenges that prevented this approach:

- The search, scraping, and cleaning of this data from the web would prove very resource-intensive for our ECTs.
- Online summaries of ECTs on popular finance websites like CNBC and Reuters do not explicitly cover the Q&A portion of earnings calls.

[5] https://www.fool.com/earnings-call-transcripts/

[6] https://seekingalpha.com/earnings/earnings-call-transcripts

Dataset	# Docs.	Coverage	Density	Comp. Ratio	# Tokens	
					Doc.	Summary
ARXIV/PUBMED (Cohan et al., 2018)*	346,187	0.87	3.94	31.17	5179.22	257.44
BILLSUM (Kornilova and Eidelman, 2019)*	23,455	-	4.12	13.64	1813.0	207.7
BIGPATENT (Sharma et al., 2019)*	1,341,362	0.86	2.38	36.84	3629.04	116.67
GOVREPORT (Huang et al., 2021)*	19,466	-	7.60	19.01	9409.4	553.4
BOOKSUM (Kryściński et al., 2022)*	12,293	0.78	1.69	15.97	5101.88	505.32
ECTSum (Mukherjee et al., 2022)*	2,425	0.85	2.43	**103.67**	2916.44	49.23
ECTQA	182	0.22	4.82	87.50	**25607.20**	**3163.80**

Table 1: Comparing Statistics of ECTQA with existing long document summarization datasets. The numbers marked with * are copied from (Mukherjee et al., 2022). Unreported numbers are blank. ECTQA has the most tokens across its documents and reference summaries while having a similar *compression ratio* to ECTSum.

We then considered naive approaches like taking the first 15 sentences of our input document as its corresponding reference summary. Because our input documents are lengthy, this approach was insufficient to summarize the document. The first 15 sentences of any given Q&A would not cover the breadth of questions asked throughout the source document.

Through our literature review, we identified two unsupervised summarization approaches, a contrastive learning approach proposed by Wu et al (Wu et al., 2020) and a salient sentences extractive model from the SUPERT benchmark paper (Gao et al., 2020). We chose the latter, as the SUPERT workflow explicitly contains a salient sentences extractive model that generates a pseudo reference summary.

In our implementation of SUPERT's salient sentences extractive model, we generate contextualized embeddings for each sentence in the input document using SBERT (Reimers and Gurevych, 2019). We then implement a clustering algorithm described in the paper that first measures the similarity of sentence pairs and clusters sentences using the affinity propagation algorithm (Frey and Dueck, 2007), with the center of each cluster building the pseudo reference. We implemented the global graph version of this model, which builds the graph considering all sentences across all source documents.

3.3.4 Statistics and Analysis

The data cleaning and processing process resulted in a total of 182 document pairs, with an average document length of 25.6K words and an average pseudo reference summary length of 3.1K words. *Coverage* measures the extent a summary is derivative of the input text while *Density* measures how well the word sequence can be described as a series of extractions. While our Density score of 4.82 is similar to other benchmark datasets, our Coverage score of 0.22 is much lower. We believe this difference is because the average length of our documents is up to 10 times the length of documents in other benchmark datasets. Our *compression ratio* of 87.50 is similar in scale to ECTSum.

3.3.5 Dataset Limitations

ECTQA presents a novel approach to developing document-summary pairs using an unsupervised extractive model to generate ground truth reference summaries. Alternative approaches are too resource-intensive, so we assume this is the best method to evaluate the performance of our models.

4 Results and Discussion

4.1 Baselines

We evaluate and compare the summarization performance across several algorithms. We chose our baseline to be our unsupervised approach against the entirety of the input documents. We then evaluated the performance of several fine-tuned ECT-BPS models as well as the Longformer Encoder Decoder (LED) (Beltagy et al., 2020) long document summarization model.

4.2 Evaluation Metrics

We consider ROUGE (Lin, 2004) and BERTScore (Zhang et al., 2020b) to evaluate the content quality of model-generated summaries. We report the F-1 scores corresponding to ROUGE-1, ROUGE-2, ROUGE-L, ROUGE-LSUM and BERTScore.

4.3 Fine-Tuning the ECT-BPS Paraphrasing Modules (BART, T5, PEGASUS)

We evaluated the impact of the following decoders on the paraphrasing module:

Model	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-LSUM	BERTScore
Baseline Approach vs Input Documents					
SUPERT (Gao et al., 2020)	0.222	0.214	0.222	0.223	0.844
Summarization Approaches vs Reference Summaries					
LED (Beltagy et al., 2020)	0.081	0.011	0.047	0.071	0.789
ECT-BPS Extractive-Only (Mukherjee et al., 2022)	**0.434**	**0.236**	0.189	**0.412**	**0.844**
ECT-BPS w/ PEGASUS	**0.264**	0.083	0.121	**0.251**	0.829
ECT-BPS w/ BART	**0.539**	0.208	0.195	**0.507**	**0.846**
ECT-BPS w/ T5	**0.384**	0.128	0.154	**0.363**	0.837

Table 2: Comparison of representative summarizers against automatic evaluation metrics. SUPERT extractive summaries are evaluated against input documents. ECT-BPS-generated summaries are evaluated against SUPERT extractive sumnmaries. Scores that match or beat our baseline are **bolded**.

- BART: This is an open-source large BART seq2seq (text2text generation) model fine-tuned on three paraphrase datasets.

- PEGASUS: Open-source PEGASUS paraphrasing model fine-tuned for Question and Answering using text2text approach.

- T5: Open-source paraphrasing model trained on a custom dataset and T5 large model. This is the baseline paraphrasing model used by the ECT-BPS paper.

We recognized the lack of labeled data for fine-tuning the models as the **ECTQA** sample size (n=182 document-summary pairs) is too small for a useful train-validation-test split. So, we opted to fine-tune our models on the **ECTSum** dataset itself (n=2,425 document-summary pairs), and then evaluate our models' results against the baseline.

For the three paraphrasing models evaluated, the model parameters were initialized with pre-trained weights from Huggingface and then trained end-to-end on the **ECTSum** dataset to fine-tune the parameters. The other hyperparameters were as specified in the **ECTSum** paper.

We set padding to the longest sequence in the batch and truncation to a maximum length of 60 tokens. The number of beams was set at 5, the number of highest scoring beams returned was set at 1 and the maximum length of the output limited to 60 tokens, for the three paraphrasing models evaluated.

4.4 Main Results

Table 2 presents the performance of all the evaluated methods on the full **ECTQA** dataset. We found that LED's abstractive approach performs poorly relative to our extractive SUPERT baseline. The ECT-BPS extractive-only module (without paraphrasing) performs better than the baseline approach, with a 44% improvement on the average of the *ROUGE* scores and a similar *BERTScore*.

For the paraphrasing models, only BART beats the *BERTScore* of the baseline SUPERT approach, while also showing a 64% improvement in the average *ROUGE* score. In addition to outperforming the other two paraphrasing models (T5 and PEGASUS), BART also outperforms the ECT-BPS extractive-only module. This result indicates that the ECT-BPS framework with a BART paraphrasing module ensures the best performance for summarizing ECTQA over typical baselines like LED and SUPERT.

4.5 Human Evaluation of Summaries

To manually evaluate our model-generated summaries, we used a similar process described in the ECTSum paper (Mukherjee et al., 2022). We randomly selected a sample of 10 model-generated summaries from our ECT-BPS model with a fine-tuned T5 paraphrasing module. We selected the T5 paraphrasing module outputs as it originally performed the best relative to our baseline. However, after adjusting the BART beam search hyperparameter to match its value in our T5 and PEGASUS experiments, BART performed the best. Given the absence of time to change which summaries we would manually evaluate, we opted to keep our T5 model-generated summaries in the human evaluation analysis.

We evaluated these model-generated summaries on three metrics: **factual correctness**, **relevance**, and **coverage**.

- **Factual Correctness**: Count of sentences represented in the summary supported by the ECTQA input document.

- **Relevance**: Count of sentences among the most important in the summary supported by the ECTQA input document.

The final scores for both were mapped to a 1-5 scale determined by the percentage of sentences that were factually correct/relevant: 5 (>80%), 4 (>60% & $\leq$ 80%), 3 (>40% & $\leq$ 60%), 2 (>20% & $\leq$ 40%), 1 ($\leq$20%).

Coverage is an overall score for the summary based on our impression about the coverage of relevant content in the input document.

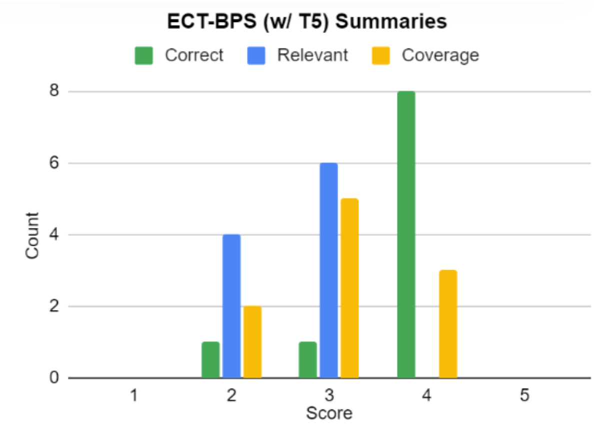

Figure 1: Histogram for human evaluation scores assigned to model-generated summaries

The summary results from this analysis are presented in Figure 1. We found that, on average, 64% of the sentences in our ECT-BPS model-generated summaries were *factually correct* while only 42% were *relevant*. We believe the significant sentence compression ratio of our summaries is causing this low relevance score. Our average input document contains 305 sentences while our average summary contains only 20. Our model compresses the **ECTQA** input documents by over 15x.

We also qualitatively assessed the model output across this random sample. We found that the LED model often hallucinates different time and monetary values, while ECT-BPS does not. This is likely due to our strategy of masking these values when fine-tuning the different paraphrasing modules.

We found that our model-generated summaries contain relevant questions from Wall Street analysts but not the answers from the public company representative. This necessitates further research into both our data collection strategy and unsupervised approach to generate ground truth reference summaries. One potential solution is using a multi-document summarization approach, treating each question and answer response as a *chapter* within a single **ECTQA** input document. This could improve our model's *relevance* and *coverage* scores as it would summarize *each* question and answer exchange individually within a single ECT instead of *all* question and answer exchanges in the entire ECT.

5 Conclusion

In this work, we develop **ECTQA** - a dataset consisting of the Question and Answers section of ECTs and the corresponding reference summaries generated by an unsupervised extractive approach found in the **SUPERT** benchmark paper. We then extend the **ECT-BPS** modeling framework to the **ECTQA dataset**, and show that it is possible to derive concise, valuable, and accurate insights from the long and unstructured Q&A sections of ECTs in an efficient manner without losing salient information.

We evaluate the performance of three paraphrasing models (T5, BART, and PEGASUS) in the **ECT-BPS** framework, an extractive-only ECT-BPS approach, and a long document summarization model (LED) in summarizing **ECTQA** input documents. We determine that **ECT-BPS** with a BART paraphrasing model fine-tuned on the **ECTSum** dataset performs significantly better on **ECTQA** than other leading summarization approaches. We believe our novel contributions to this dataset and summarization methodology will be valuable for future research in the finance domain.

References

Dogu Araci. 2019. Finbert: Financial sentiment analysis with pre-trained language models.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer.

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 the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 615–621, New Orleans, Louisiana. Association for Computational Linguistics.

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.

Brendan J. Frey and Delbert Dueck. 2007. Clustering by passing messages between data points. *Science*, 315(5814):972–976.

Yang Gao, Wei Zhao, and Steffen Eger. 2020. Supert: Towards new frontiers in unsupervised evaluation metrics for multi-document summarization.

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.

Mandy Guo, Joshua Ainslie, David Uthus, Santiago Ontanon, Jianmo Ni, Yun-Hsuan Sung, and Yinfei Yang. 2022. Longt5: Efficient text-to-text transformer for long sequences.

Luyang Huang, Shuyang Cao, Nikolaus Parulian, Heng Ji, and Lu Wang. 2021. Efficient attentions for long document summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1419–1436, Online. Association for Computational Linguistics.

Anastassia Kornilova and 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, Hong Kong, China. Association for Computational Linguistics.

Wojciech Kryściński, Nazneen Rajani, Divyansh Agarwal, Caiming Xiong, and Dragomir Radev. 2022. Booksum: A collection of datasets for long-form narrative summarization.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2021. Summac: Re-visiting nli-based models for inconsistency detection in summarization.

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.

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.

Rajdeep Mukherjee, Abhinav Bohra, Akash Banerjee, Soumya Sharma, Manjunath Hegde, Afreen Shaikh, Shivani Shrivastava, Koustuv Dasgupta, Niloy Ganguly, Saptarshi Ghosh, and Pawan Goyal. 2022. ECTSum: A new benchmark dataset for bullet point summarization of long earnings call transcripts. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10893–10906, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. 2016. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents.

Tatiana Passali, Alexios Gidiotis, Efstathios Chatzikyriakidis, and Grigorios Tsoumakas. 2021. Towards human-centered summarization: A case study on financial news. In *Proceedings of the First Workshop on Bridging Human–Computer Interaction and Natural Language Processing*, pages 21–27, Online. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Dexter Roozen and Francesco Lelli. 2021. Stock values and earnings call transcripts: a dataset suitable for sentiment analysis. *Preprints.org*.

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, Florence, Italy. Association for Computational Linguistics.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. *arXiv preprint arXiv:2004.09297*.

Hanlu Wu, Tengfei Ma, Lingfei Wu, Tariro Manyumwa, and Shouling Ji. 2020. Unsupervised reference-free summary quality evaluation via contrastive learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3612–3621, Online. Association for Computational Linguistics.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2020a. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020b. Bertscore: Evaluating text generation with bert.

Ming Zhong, Pengfei Liu, Yiran Chen, Danqing Wang, Xipeng Qiu, and Xuanjing Huang. 2020. Extractive summarization as text matching. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6197–6208, Online. Association for Computational Linguistics.