University of Manchester
School of Computer Science
Project Report 2023

Extractive Summarisation of UK Annual Reports

Author: Vladislav Yotkov

Supervisor: Dr. Jonathan Shapiro

Abstract

Extractive Summarisation of
UK Annual Reports

Author: Vladislav Yotkov

Although there has been considerable progress in Natural Language Processing (NLP) over the years, it has not fully reached the Accounting and Finance (AF) industry. In the meantime, companies worldwide produce vast amounts of textual data as part of their reporting packages to comply with regulations and inform shareholders of their financial performance. The glossy annual report is such an example, widely read by investors but it also tends to be quite long. Inspired by the Financial Narrative Summarisation (FNS) 2021 Task, we will design an Automatic Text Summarisation (ATS) system for the narrative parts of UK financial annual reports. With this goal in mind, we will implement and explore the following models for Extractive Text Summarisation (ETS): 1. custom Recurrent Neural Network (RNN), 2. fine-tuned FinBERT. In terms of evaluation, we will use the ROUGE metric to compare the performance of these models against standard ATS baselines: TextRank, and LexRank.

Supervisor: Dr. Jonathan Shapiro

Contents

List of Figures

List of Tables

0.1 Introduction

0.1.1 Financial Reports

Due to international regulations, companies are obliged to report their periodic performance (annual, bi-annual, quarterly) to various regulatory authorities[1] and other users (e.g., corporate stakeholders, investors, customers, suppliers, etc.). These reports contain essential information about the operations and finances of a business and are crucial for making informed decisions (from a user perspective), but are different in regulatory forms. For example,

1. 10-K reports filed to the SEC[2] and accessible through their Electronic Data Gathering, Analysis, and Retrieval[3] (EDGAR) system are only for US registered businesses. They follow a standardised template and are plain text, which makes them particularly easy for automated large-scale research ([EHAR+19]). Also, the contents of these reports are strict, requiring solely five information sections[4].

2. UK annual reports, regulated by the UK's Financial Reporting Council (FRC), are typically the primary annual reporting method (also provided as PDF files). Unlike the 10-K, they are glossy and more stakeholder-oriented and enjoy unlimited discretion over non-mandated content ([EHAR+19]) (e.g., photography and company brand material, non-mandatory narrative sections, etc.). However, these are more challenging for automated processing due to their variable section structure, formatting, and rich visual representations.

0.1.2 UK annual reports

The annual report is the primary corporate disclosure legally required for public companies by regulatory authorities. While it *does not have a rigid document structure* like the 10-K,

[1] Regulation authorities worldwide:

- Securities and Exchange Commission (SEC) in the USA
- European Securities and Markets Authority (ESMA) in Europe
- Financial Reporting Council (FRC) in the UK
- International Financial Reporting Standards (IFRS) in 167 jurisdictions worldwide

[2] `https://www.sec.gov`

[3] `https://www.sec.gov/edgar`

[4] (a) Business Overview (b) Risk Factors (c) Management's Discussion and Analysis of Financial Condition and Results of Operations (MD&A) (d) Financial Statements (e) Supplementary Disclosures

it typically has a *narrative component*[5] and the financial statements (at the rear).

As we outlined in Section 0.1.1, UK annual reports have the following inconvenient properties with regard to large-scale text understanding.

- They are very long documents. Throughout the years, their average length has been increasing significantly with the number of pages rising 57% for the median report from 2003 to 2016 (47 to 74 pages, respectively) ([LY19]), due to additional regulations between 2006 and 2008 ([EHAR$^+$19].
- They have variable nomenclature. From firm to firm, naming conventions vary "dramatically", with more than 20 unique titles for various sections (e.g., Chair's letter to shareholders, Management Commentary) ([LY19]).
- They incorporate embedded info-graphics. While domain experts hail the integration of highly interactive elements into corporate reporting ([KB16]), the compilation to PDF makes the task of analysing such unstructured documents automatically even harder ([LY19]).

These challenges motivate the work of [EHRY$^+$19] who (a) established a set of 8 generic section headers[6] and (b) built the CFIE-FRSE[7] extraction tool that converts a text-based PDF annual report to simple text.

0.1.3 NLP in Accounting and Finance

The relevance of this project should also be understood from the perspective of the development of Natural Language Processing (NLP) in the Accounting and Finance (AF) domain. As outlined in [Ell98], investors' trust in the accountability of businesses would be based no longer as much on just the financial statements, but also on more descriptive narratives that define strategy and planning of resource use. While some recognise the importance of understanding in-domain textual information ([L$^+$10]), others like [EHRW$^+$19] report that the industry is still doubtful and cynical about the NLP applications in the analysis of financial market disclosures. Furthermore, the latter also observe that AF researchers rely extensively on bag-of-words models, which are *not sufficient to encode complex contextual and semantic meaning* (especially in a domain with such *specialized language*).

[5]The narrative component of a UK annual report typically consists of 1. Management's Commentary 2. Letter to Shareholders 3. Corporate Governance Statement 4. Auditor's Report 5. Remuneration Report 6. Business Review 7. Environmental, Social, and Governance (ESG) Report 8. Risk Management Report

[6](a) Chairman Statement (b) CEO Review (c) Corporate Governance Report (d) Directors Remuneration Report (e) Business Review (f) Financial Review (g) Operating Review (h) Highlights

[7]The CFIE-FRSE stands for Corporate Financial Information Environment - Final Report Structure Extractor. It is publicly available at `https://github.com/drelhaj/CFIE-FRSE` and it can be used to convert English, Spanish and Portuguese annual reports.

As for ATS [CHW19] is said to be the single AF study into disclosure summarisation. It demonstrates that machine-generated summaries are less likely to bias positively investor decisions compared to managerial ones. Therefore, this only confirms the existence of a wide gap in NLP applications in Accounting research, which further motivates our work.

0.1.4 Financial Narrative Summarisation 2021 (FNS21) Task

The FNS Task is part of the annual Financial Narrative Processing (FNP) Workshop [8] organised by Lancaster University since 2018, which aims to:

- encourage the advancement of financial text mining & narrative processing
- examine methods of structured content retrieval from financial reports
- explore causes and consequences of corporate disclosure

as stated in their inaugural proceedings [9].

For that purpose, they produce datasets of extracted narratives (with the help of the CFIE-FRSE tool) from annual reports of UK companies listed on the London Stock Exchange (LSE).

In their FNS21 Task, there were $3,863$ such reports (Table 1), while the average length was reported at 80 pages, and the maximum of more than 250 pages ([LV21]).

Additionally, for every report, there were at least two gold summaries situated in the annual report itself [10] The workshop's goal was to build ATS systems that generate a single summary for an annual report, no longer than $1,000$ words (almost just as long as the gold summaries on average).

We acknowledge that due to the scarcity of publicly available financial data this third-year project could not have been possible without the kind permission of the FNP organisers to use the training and validation datasets from their FNS21 Task ([EHRZ21]).

[8] https://wp.lancs.ac.uk/cfie/

[9] https://wp.lancs.ac.uk/cfie/fnp2018/

[10] The gold summaries being already in the annual report is not problematic because these reports are already written by domain experts who know how to summarise the financial state of a company. Hence, multiple sections/paragraphs could achieve this thoroughly, and the organisers have identified & extracted them manually with the help of the professional writers of the individual reports. At this moment, one can begin to doubt the point of applying ATS techniques, but due to the *lack of rigid document structure*, *it is not trivial to automatically find these text excerpts with heuristic methods*. Furthermore, we can formulate this challenge as finding the latent features of a summarising (i.e., "to-be-in-the-summary") sentence, highlighted as one of the fundamental advantages of NLP in AF research ([LY19], [EHRW$^+$19]).

Data Type	Training	Validation	Testing	Total
Report full text	3,000	363	500	3,863
Gold summaries	9,873	1,250	1,673	12,796

Table 1: FNS21 Data Split

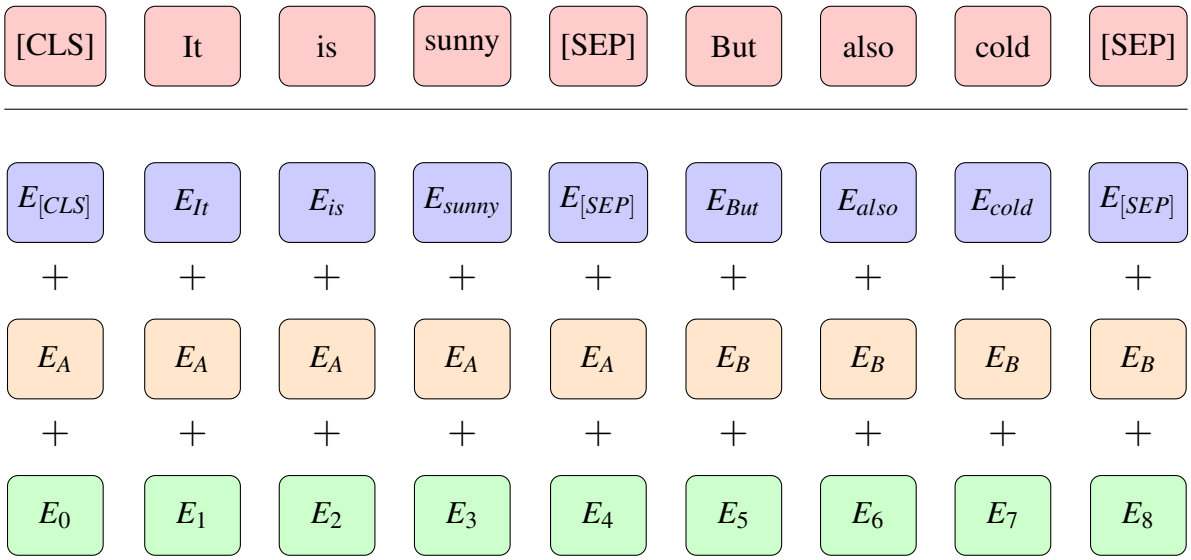


Figure 1: BERT: Input Embeddings

0.2 Background

0.2.1 Supervised Learning

0.2.2 TFIDF

0.2.3 Word Embeddings

0.2.4 Attention

0.2.5 RNN

0.2.6 Transformers

0.2.7 Text Summarisation

Text summarisation is the task of transforming a piece of text into a shorter version that retains the most important information. There are two overarching categories: extractive and abstractive text summarisation. The former formulates the problem as a subset selection problem by returning only the most salient text excerpts from the original document ([ZLC+20]), while the latter aims to generate content anew, similar to how humans would do.

We will outline some key models that inspired our work below:

- **Gokhan**: The authors employ an unsupervised summariser based on K-Means clustering of sentences encoded with SentenceBERT ([RG19]). However, their embeddings are pre-trained on general text, and they suggest that employing in-domain language models would result in a better performance.
- **AMUSE** ([LV21]): The authors design an ETS system comprised of the following steps 1. shortening of report with an existing Genetic Algorithm [LL13], 2. encoding sentences with BERT vectors, and 3. performing binary classification with LSTMs for salient sentence extraction. They suggest that further work should incorporate 1. efficient preliminary sentence removal, and 2. additional neural modelling stages for the representation and detection of relevant input text parts.
- **Hybrid model with RL** ([ZSEHR21]): The authors train a joint extractive-abstractive summarisation model with reinforcement learning optimised for the ROUGE-2 F1 metric. Their networks are based on attentive LSTMs augmented with an additional copy mechanism ([VFJ15]) achieving the second highest F1 score in the FNS21 competition.
- **T5** Hybrid ([Orz21]): The author used T5 ([RSR$^+$20]) for a hybrid model fine-tuned to generate the beginning of an abstractive summary and find the closest match of the output in the report's full text. This is the best performing algorithm in the FNS21 competition but also the first to consider transformer models from an abstractive summarisation perspective in the FNP workshops so far.

In this work we will be solely exploring the extractive method, and more specifically - the *supervised neural-based* (i.e., RNN, Transformer) type and the *unsupervised graph-based* (i.e., TextRank, LexRank) type.

0.2.8 LexRank

LexRank ([ER04]) is an unsupervised extractive summarisation method consistently used as a baseline in the FNS21 and previous challenges. It retrieves the most salient document sentences by computing their importance based on *eigenvector centrality*. To do that the algorithm creates a graph where each sentence represents a node and each edge is a weight between two nodes ([SGWM20]). The sentences are encoded as bag-of-words vectors of size N - the vocabulary size, and the weight metric is a combination of tf-idf (Eq.1,2) and cosine similarity - Eq.3.

$$\text{idf}(t,D) = \log \frac{|D|}{|\{d \in D : t \in d\}|} \tag{1}$$

$$\text{tf-idf}(t,d,D) = \text{tf}(t,d) \cdot \text{idf}(t,D) \tag{2}$$

$$\text{tf\_idf\_cosine\_similarity}(s_1, s_2) = \frac{\sum_{t \in T} \text{tf-idf}(t, s_1, D) \cdot \text{tf-idf}(t, s_2, D)}{\sqrt{\sum_{t \in T} \text{tf-idf}(t, s_1, D)^2} \cdot \sqrt{\sum_{t \in T} \text{tf-idf}(t, s_2, D)^2}} \quad (3)$$

where t is a term, d is a document within a collection of documents/sentences D.

Also, s_1 and s_2 are two sentences and T represents the set of all terms in both of them while $tf(t,d)$ denotes the term frequency of t in d, and $idf(t,D)$ is the inverse document frequency of t in the collection D.

The authors further propose finding the most important sentences by 1. applying a threshold for the creation of edges with Eq.3, 2. building an adjacency matrix and normalizing it to produce *transition probabilities*, 3. computing in an iterative fashion the *eigenvector centrality* until convergence, and finally 4. ranking sentences based on their *lexical* PageRank ([BP98]) score.

0.3 Design & Development

0.3.1 Methodology

We approach the annual report summarisation problem from a supervised perspective - we cast the task of Extractive Text Summarisation (ETS) as a binary classification problem defined on the sentence level. More formally, we can describe the annual report as $d = \{s_1, s_2, \ldots, s_n\}$, where d is a document, represented in terms of sentences $s_i,\ 1 \leq i \leq n$ ([Liu19]).

Then, a candidate summary can be $c = \{s_1, s_2, \ldots, s_k | s_i \in d\},\ 0 \leq k \leq n$.

We further need to define the *gold summary*, c^* for a document d.

In the case of the FNS21 task, there are at least two summaries per report, hence we will use the following notation for the set of all gold summaries for each document $C^* = \{c_1^*, c_2^*, \ldots, c_p^*\}$. Furthermore, the supervised learning labels are $y_i \in \{1, 0\}$ for each sentence s_i in d if the sentence is or is not in ***any***[11] of the gold summaries c_j^* for that document.

In general, to assess the quality of a candidate summary c, we measure its similarity with the gold summary c^* based on their n-gram overlap $R = (c, c^*)$, where R is the ROUGE-F_1[12] metric([Lin04]).

For the FNS21 task due to the extractive nature of our approach we will evaluate our models based on the ROUGE-maximising c_i^* gold summary, i.e.,

$$r = \underset{c^* \in C^*}{\operatorname{argmax}} R(c, c_i^*) \tag{4}$$

Figure 2: Candidate summary evaluation as a gold summary ROUGE-maximisation

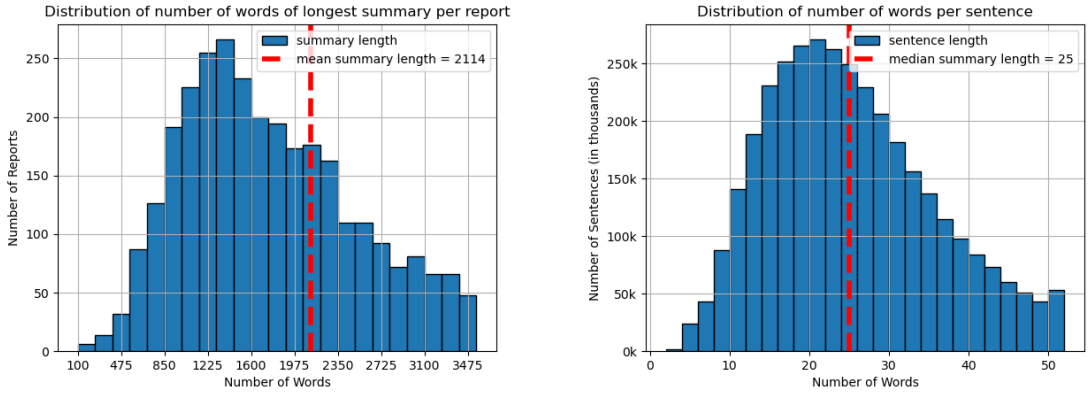
While some authors ([ZSEHR21]) follow the greedy ROUGE-maximisation method of matching summary sentences to document sentences (established in [NZZ17]), we approach the problem in a more practical and faster fashion. After manual observation of the

[11] To increase the positive samples (i.e., the summarizing sentences) we do not restrict ourselves to just one gold summary in the training process unlike [Orz21]. Our goal is to achieve better latent feature extraction of summaries through the employment of all existing data. However, we are aware that this approach is more likely to encounter standard ETS issues, specifically - extracted summary sentences could be retrieved from unrelated paragraphs in the report. This causes the "dangling anaphora" phenomenon, i.e. decontextualised extracts are stitched together and could mislead the reader due to out-of-context references as specified in [Lin09].

[12] We use a slightly different but faster version of ROUGE compared to the official metric [Lin04]. It can be accessed at: `https://github.com/pltrdy/rouge`

reports against their gold summaries, it became clear that for almost all sentences of c_i^*, there was an exact match with a sentence in the whole annual report d.

This hypothesis was proven correct by one of the FNS21 contestants ([Orz21]) who reported that 99.4% of the summaries were included in the report as whole subsequences. Hence, after having pre-processed the text documents we iteratively match the sentences and generate the binary classification labels ($\{1,0\}$ representing *summary* and *non-summary*, respectively) for both the training and testing datasets.



(a) Number of words in longest report summary

(b) Number of words in training sentences

Figure 3: Distribution of number of words in training sentences and report summaries

0.4 Evaluation

0.4.1 Confusion Matrix

The confusion matrix is an essential tool to visualise and help assessing the performance of trained classifiers against the true labels y_i.

For the problem of binary classification, it is a square matrix (Table 2) that displays the following key elements:

- True Positives (TP): Correct predictions of the positive class.
- True Negatives (TN): Correct predictions of the negative class.
- False Positives (FP): Incorrect predictions of the positive class (Type I error).
- False Negatives (FN): Incorrect predictions of the negative class (Type II error).

where for the problem of extractive text summarisation, the positive and negative classes correspond to *summary* and *non-summary* sentences, respectively. These matrix elements can be further combined into informative classification metrics:

- Accuracy: Proportion of correctly classified instances out of the total instances. Formulated as $\frac{TP+TN}{TP+TN+FP+FN}$.
- Precision (or Positive Predictive Value): Proportion of true positive instances out of all instances predicted as positive. Formulated as $\frac{TP}{TP+FP}$.
- Recall (or Sensitivity): Proportion of true positive instances out of all actual positive instances. Formulated as $\frac{TP}{TP+FN}$.
- F1-score: Harmonic mean of precision and recall (i.e., the trade-off between the two). Formulated as $2 \cdot \frac{Precision \cdot Recall}{Precision+Recall}$.

	Actual	
	Positive	**Negative**
Predicted Positive	TP	FP
Predicted Negative	FN	TN

Table 2: Confusion Matrix

$$ROUGE-N = \frac{\sum_{S \in R} \sum_{n-gram \in S} count_{match}(n-gram)}{\sum_{S \in R} \sum_{n-gram \in S} count(n-gram)} \quad (5)$$

Figure 4: ROUGE-N: N-gram Co-Occurrence Statistics

Bibliography

[BP98] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. *Computer Networks*, 30:107–117, 1998.

[CHW19] Eddy Cardinaels, Stephan Hollander, and Brian White. Automatic summarization of earnings releases: attributes and effects on investors' judgments. *Review of Accounting Studies*, 24, 09 2019.

[EHAR+19] Mahmoud El-Haj, Paulo Alves, Paul Rayson, Martin Walker, and Steven Young. Retrieving, classifying and analysing narrative commentary in unstructured (glossy) annual reports published as pdf files. *Accounting and Business Research*, 2019. Forthcoming.

[EHRW+19] Mahmoud El-Haj, Paul Rayson, Martin Walker, Steven Young, and Vasiliki Simaki. In search of meaning: Lessons, resources and next steps for computational analysis of financial discourse. *Journal of Business Finance & Accounting*, 46(3-4):265–306, 2019.

[EHRY+19] Mahmoud El Haj, Paul Edward Rayson, Steven Eric Young, Paulo Alves, and Carlos Herrero Zorita. *Multilingual Financial Narrative Processing: Analysing Annual Reports in English, Spanish and Portuguese*. World Scientific Publishing, 2019.

[EHRZ21] Mahmoud El-Haj, Paul Rayson, and Nadhem Zmandar, editors. *Proceedings of the 3rd Financial Narrative Processing Workshop*, Lancaster, United Kingdom, 15-16 September 2021. Association for Computational Linguistics.

[Ell98] Robert K Elliott. Accounting in the 21st century. 1998.

[ER04] Gunes Erkan and Dragomir R. Radev. Lexrank: Graph-based centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, 2004.

[KB16] P Kriz and H Blomme. The future of corporate reporting–creating the dynamics for change. *International Federation of Accountants (IFAC), available at: www. ifac. org/global-knowledgegateway/viewpoints/future-corporate-reporting-creating-dynamics-change (accessed 29 May 2016)*, 2016.

[L+10] Feng Li et al. Textual analysis of corporate disclosures: A survey of the literature. *Journal of accounting literature*, 29(1):143–165, 2010.

[Lin04] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

[Lin09] Jimmy Lin. Summarization. In *Encyclopedia of Database Systems*, pages 2906–2910. Springer, Heidelberg, Germany, 2009.

[Liu19] Yang Liu. Fine-tune bert for extractive summarization. *arXiv preprint arXiv:1903.10318*, 2019.

[LL13] Marina Litvak and Mark Last. Multilingual single-document summarization with MUSE. In *Proceedings of the MultiLing 2013 Workshop on Multilingual Multi-document Summarization*, pages 77–81, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.

[LV21] Marina Litvak and Natalia Vanetik. Summarization of financial reports with AMUSE. In *Proceedings of the 3rd Financial Narrative Processing Workshop*, pages 31–36, Lancaster, United Kingdom, 15-16 September 2021. Association for Computational Linguistics.

[LY19] Craig Lewis and Steven Young. Fad or future? automated analysis of financial text and its implications for corporate reporting. *Accounting and Business Research*, 49(5):587–615, 2019.

[NZZ17] Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31, 2017.

[Orz21] Mikhail Orzhenovskii. T5-LONG-EXTRACT at FNS-2021 shared task. In *Proceedings of the 3rd Financial Narrative Processing Workshop*, pages 67–69, Lancaster, United Kingdom, 15-16 September 2021. Association for Computational Linguistics.

[RG19] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.

[RSR+20] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. 21(1), 2020.

[SGWM20] Steven Shearing, Abigail S. Gertner, Ben Wellner, and Liz Merkhofer. Automated text summarization: A review and recommendations. 2020.

[VFJ15] Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. Pointer networks. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

[ZLC+20] Ming Zhong, Pengfei Liu, Yiran Chen, Danqing Wang, Xipeng Qiu, and Xuanjing Huang. Extractive summarization as text matching. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6197–6208, Online, July 2020. Association for Computational Linguistics.

[ZSEHR21] Nadhem Zmandar, Abhishek Singh, Mahmoud El-Haj, and Paul Rayson. Joint abstractive and extractive method for long financial document summarization. In *Proceedings of the 3rd Financial Narrative Processing Workshop*, pages 99–105, Lancaster, United Kingdom, 15-16 September 2021. Association for Computational Linguistics.