The Financial Narrative Summarisation Shared Task (FNS 2021)

Nadhem Zmandar[1], **Mahmoud El-Haj**[1], **Paul Rayson**[1], **Ahmed AbuRa'ed**[2], **Marina Litvak**[3], **Nikiforos Pittaras**[4], **Geroge Giannakopoulos** [4]

[1]Lancaster University, UK
[2]Universitat Pompeu Fabra, Spain
[3]Shamoon College of Engineering, Israel
[4]Demokritos, Greece

{n.zmandar, m.el-haj,p.rayson}@lancaster.ac.uk
ahmed.aburaed@upf.edu
marinal@ac.sce.ac.il
ggianna@iit.demokritos.gr, pittarasnikif@gmail.com

Abstract

This paper presents the results and findings of the Financial Narrative Summarisation Shared Task on summarising UK annual reports. The shared task was organised as part of the Financial Narrative Processing 2021 Workshop **(FNP 2021 Workshop)**. The shared task included one main task which is the use of either abstractive or extractive automatic summarisers to summarise long documents in terms of UK financial annual reports. This shared task is the second to target financial documents. The data for the shared task was created and collected from publicly available UK annual reports published by firms listed on the London Stock Exchange. A total number of 10 systems from 5 different teams participated in the shared task. In addition, we had two baseline and two topline summarisers to help evaluate the results of the participating teams and compare them to the state-of-the-art systems.

1 What are financial narratives

Companies produce a variety of reports containing both narrative and numerical information at various times during their financial year, including annual financial reports. This creates vast amounts of financial information which can be impossible to navigate, handle and keep track of. This shows the vital need for automatic summarisation systems in order to reduce the time and effort of both the shareholders and investors.

2 Related Work

The increased availability of financial reports data has been met with research interest for applying automatic summarisation methods. The task of automatic text summarisation aims to produce a condensed, informative and non-redundant summaries from a single or multiple input texts (Nenkova and McKeown, 2011). This is achieved by either identifying and ranking subsets of the input text (i.e. extractive approaches ((Gupta and Lehal, 2010)), or by generating the summary from scratch (i.e. abstractive methods (Moratanch and Chitrakala, 2016)). Extractive methods have been a popular venue for summarising text due to their relative simplicity and the comparatively high requirements of abstractive methods for computational resources and available data.

Extractive summarisation utilises scoring approaches to identify and reorder parts of the input (e.g. sentences, phrases and/or passages), using a variety of feature extraction and evaluation methods (Luhn, 1958; Baxendale, 1958; Edmundson, 1969; Mori, 2002; McCargar, 2004; Giannakopoulos et al., 2008). Where adequate data is available, machine learning methods have been employed, such as Hidden Markov Models (Fung and Ngai, 2006), topic-based modelling (Aries et al., 2015), genetic algorithms (Litvak et al., 2010) and clustering methods (Radev et al., 2000; Liu and Lindroos, 2006; Kruengkrai and Jaruskulchai, 2003).

The employment of summarisation and natural language processing techniques in general has promising applications in the financial domain (El-Haj et al., 2019b). The SummariserPort system (de Oliveira et al., 2002) has been used to produce summaries for financial news, where it utilized lexical cohesion (Flowerdew and Mahlberg, 2009), using sentence linkage heuristics to generate the output summary. A summarisation system for financial news was proposed in (Filippova et al., 2009) generating query-based company-tailored summaries. This was done through using unsupervised sentence ranking with simple frequency-based features. Recently, statistical features with heuristic approaches have been used to summarise financial textual disclosures (Cardinaels et al., 2019), generating summaries with reduced positive bias, leading to more conservative valuation judgements by investors that receive them. Further, the Financial Narrative Summarisation task (El-Haj, 2019) of the Multiling 2019 workshop (Giannakopoulos, 2019) involved the generation of structured summaries from financial narrative disclosures. Considering this body of work, the Financial Narrative Summarisation task (FNS 2020 (El-Haj et al., 2020a)) task resulted in the first large scale experimental results and state-of-the-art summarisation methods applied to financial data. The task focused on annual reports produced by UK firms listed on the London Stock Exchange (LSE). The shared task was held as part of the 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation (FNP-FNS 2020) (El-Haj et al., 2020b). The participating systems used a variety of techniques and methods ranging from rule based extraction methods (Litvak et al., 2020; Vhatkar et al., 2020; Arora and Radhakrishnan, 2020; Azzi and Kang, 2020) to traditional machine learning methods (Suarez et al., 2020; Vhatkar et al., 2020; Arora and Radhakrishnan, 2020) and high performing deep learning models (Agarwal et al., 2020; Singh, 2020; La Quatra and Cagliero, 2020; Vhatkar et al., 2020; Arora and Radhakrishnan, 2020; Azzi and Kang, 2020; Zheng et al., 2020).

One of the main challenges and limitations reported by the participants was the average length of annual reports (around 60,000 words), which made the training process difficult as it requires powerful resources (e.g. GPUs) to avoid long training time. In addition, participants argued that extracting both text and structure from PDF files with numerous tables, charts, and numerical data resulted in noisy data being extracted. Such feedback highlights interesting aspects and challenging components of Financial Narrative Summarisation, which presents a high-difficulty task and an interesting research problem that is worth investigating. The 2021 Financial Narrative Summarization task (FNS 2021) promotes this effort by providing such a shared task in the FNP 2021 workshop[1].

3 Data Description

In the Financial Narrative Summarisation task we focus on annual reports produced by UK firms listed on The London Stock Exchange (LSE).

In the UK and elsewhere, annual report structure is much less rigid than those produced in the US. Companies produce glossy brochures with a much looser structure, which makes automatic summarisation of narratives in UK annual reports a challenging task.

For the FNS 2021 Shared task[2] we use approximately 4,000 UK annual reports for firms listed on LSE, covering the period between 2002 and 2017 (El-Haj et al., 2014, 2019a).

We divided the full text within annual reports into *training*, *testing* and *validation* sets providing both the full text of each annual report along with gold-standard summaries.

In total there are 3,863 annual reports divided into training, testing and validation sets. Table 1 shows the dataset details.

Data Type	Train	Validate	Test
Report full text	3,000	363	500
Gold summaries	9,873	1,250	1,673

Table 1: FNS 2021 Shared Task Dataset

4 Data Availability

For the shared task we first provide the training and validation sets, which include the full text of each annual report along with the gold-standard summaries. On average, there are at least three gold-standard summaries for each annual report with some reports containing up to seven gold-standard summaries. The full test set is available only to organisers who evaluate the participating systems. The gold-standard summaries for the test set were not provided to participants in advance.

5 Eval-AI Platform

This year we introduced a new feature of the shared task which is hosting the task on Eval-AI open source AI challenge platform[3].

Eval-AI (Yadav et al., 2019) is an open source platform for evaluating and comparing Machine Learning (ML) and Artificial Intelligence (AI)

[1] Main workshop: http://wp.lancs.ac.uk/cfie/fnp2021/

[2] http://wp.lancs.ac.uk/cfie/fns2021/

[3] https://eval.ai/web/challenges/challenge-page/1070/overview

algorithms. It is built to provide a scalable solution to the scientific research community and address the need to evaluate machine learning models by customisable metrics or through looping human evaluation. This will help researchers, students and data scientists to create, collaborate and participate in AI challenges organised around the world or by customising this platform and hosting it in a private cloud. This platform simplifies and standardises the process of bench-marking created models.

Using Eval-AI enabled us to automate the evaluation of the submissions and to use Custom evaluation phases and protocols.

6 Task Description

For the purpose of this task each team was asked to produce one summary for each annual report. The summary length should not exceed **1000** words. We advised that the summary is generated/extracted based on the narrative sections.

Only one summary was allowed for each report, but participating teams were welcome to participate with more than one system. The participants were asked to follow a standard file naming process to aid the automatic evaluation process. Also, for standardisation and consistency all output summary files were required to be in UTF-8 file format.

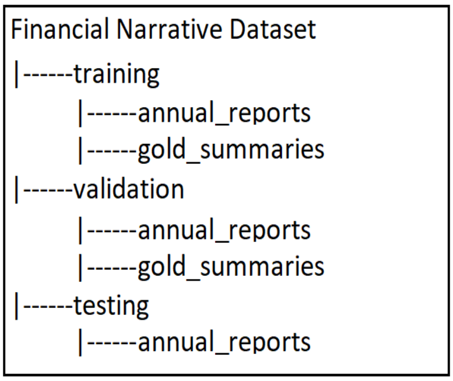
Regarding generated outputs from a participant system, the following criteria were requested:

- Each team should produce a no more than 1000 words summary for each annual report in the testing set.
- One summary should be provided for each report.
- Each summary should be named following the pattern **ID_summary.txt**. Example: 25082_summary.txt.
- All outputs should be in UTF-8 file format.
- All output summaries should be compressed following the pattern *<TeamName>_Summaries.tar.gz.*

6.1 Evaluation

To evaluate the generated system summaries against the human gold-standard summaries we used the Java Rouge (JRouge)[4] package for ROUGE, using multiple variants (i.e. ROUGE-1, ROUGE-2, ROUGE-L and ROUGE-SU4). (Ganesan, 2018)

7 Data Sample



```
Financial Narrative Dataset
|------training
        |------annual_reports
        |------gold_summaries
|------validation
        |------annual_reports
        |------gold_summaries
|------testing
        |------annual_reports
```

Figure 1: Dataset Structure

Figure 1 shows the structure of the Financial Narrative Summarisation dataset. At the beginning of the shared task we provided the participants with two directories, corresponding to “training” and “validation” sets. Each contained the full text of the annual reports and the gold standard summaries.

The data was provided in plain text format in a directory structure as in Figure 1. Each annual report has a unique ID and it is used across in order to link the full text from an annual report to its gold-standard summaries.

For example, the gold standard summaries for the file called **19.txt** in the *training/annual_reports* directory can be located in the *training_gold_summaries* as files with the same ID (19) as a prefix: **19_1.txt** to **19_3.txt**.

8 Baseline and Topline Summarisers

We compared the results of participating systems to four topline and baseline summarisers—MUSE (Litvak and Last, 2013), POLY (Litvak and Vanetik, 2013), TextRank (Mihalcea and Tarau, 2004), and LexRank (Erkan and Radev, 2004). See (El-Haj et al., 2020a) for more details on the topline and baseline summaries.

9 Participants and Systems

In total, 10 summarisation systems by five different teams have participated and submitted their system summaries to FNS 2021, which are presented in Table 2.

[4] https://github.com/kavgan/ROUGE-2.0

Team	Affiliation
Orzhan	Independent researcher
SRIB-lancs	Samsung + Lancaster university
UoBNLP	University of Birmingham
SCE	Shamoon college of engineering
CILab_KIT	Kumoh National Institute of Technology, Korea .

Table 2: FNS 2021 Participating Teams

10 Results and Discussion

The participating systems used a variety of techniques and methods ranging from fine tuning pre-trained transformers to using high performing deep learning models and word embeddings.

In addition, the participating teams used methods to investigate the hierarchy of the annual reports to try and detect structure and extract the narrative sections, in order to identify the parts in the report from which the gold summaries were extracted.

The majority of the applied techniques were extractive, since the dataset is highly structured with discrete sections. We report the use of T-5 (Test-to-text transfer Transformer)(Raffel et al., 2019) and BERT-based (Devlin et al., 2018) extractive models. Some extractive summarisers used word embeddings such word2vec (Mikolov et al., 2013). An end-to-end hybrid extractive-abstractive training method using pointer network generators have also been reported.

The results are reported in Table 3. Overall, the best model outperforms results compared to the baselines with $ROUGE1$: 0.54, $ROUGE$-2 : 0.38, $ROUGE$-L : 0.52 and $ROUGE$-SU4 : 0.43. The results are sorted in descending order of Rouge-2 F1-score. The results show that all participating systems outperformed TextRank baseline and most systems (eight) systems performed better than the LexRank and POLY baselines. On the other hand, results from our topline MUSE system indicate that it is a challenging opponent, but we are happy to see that two participating systems have managed to outperform it. Such results will be used as a comparison line in the future, by incorporating them into a venue of results, techniques and approaches, which we hope will be useful to researchers working on Financial Text Summarisation.

System/Metric	R-1	R-2	R-L	R-SU4
orzhan	0.54	0.38	0.52	0.43
SRIB-lancs	0.52	0.30	0.46	0.32
MUSE	0.5	0.28	0.45	0.32
SCE-1	0.5	0.27	0.44	0.30
UoBNLP-2	0.48	0.26	0.4	0.29
UoBNLP-3	0.47	0.25	0.4	0.29
UoBNLP-1	0.47	0.25	0.4	0.29
CILab_KIT	0.38	0.17	0.32	0.21
CILab_KIT-B	0.35	0.16	0.29	0.20
POLY	0.37	0.12	0.26	0.18
LEXRANK	0.26	0.12	0.22	0.14
SCE-3	0.33	0.12	0.27	0.17
SCE-2	0.35	0.11	0.26	0.18
TEXTRANK	0.17	0.07	0.21	0.08

Table 3: ROUGE-1 and ROUGE-2 and ROUGE-L and ROUGE-SU4 F-measure scores.

References

Raksha Agarwal, Ishaan Verma, and Niladri Chatterjee. 2020. Langresearchlab_nc at fincausal 2020, task 1: A knowledge induced neural net for causality detection. In *Proceedings of the 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation*, pages 33–39.

Abdelkrime Aries, Djamel Eddine Zegour, and Khaled Walid Hidouci. 2015. Allsummarizer system at multiling 2015: Multilingual single and multi-document summarization. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 237–244. The Association for Computer Linguistics.

Piyush Arora and Priya Radhakrishnan. 2020. Amex ai-labs: An investigative study on extractive summarization of financial documents. In *Proceedings of the 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation*, pages 137–142.

Abderrahim Ait Azzi and Juyeon Kang. 2020. Extractive summarization system for annual reports. In *Proceedings of the 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation*, pages 143–147.

Phyllis B Baxendale. 1958. Machine-made index for technical literature—an experiment. *IBM Journal of research and development*, 2(4):354–361.

Eddy Cardinaels, Stephan Hollander, and Brian J White. 2019. Automatic summarization of earnings releases: attributes and effects on investors' judgments. *Review of Accounting Studies*, 24(3):860–890.

Paulo Cesar Fernandes de Oliveira, Khurshid Ahmad, and Lee Gillam. 2002. A financial news summarization system based on lexical cohesion. In *Proceedings of the International Conference on Terminology and Knowledge Engineering, Nancy, France*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Harold P Edmundson. 1969. New methods in automatic extracting. *Journal of the ACM (JACM)*, 16(2):264–285.

Mahmoud El-Haj. 2019. Multiling 2019: Financial narrative summarisation. In *Proceedings of the Workshop MultiLing 2019: Summarization Across Languages, Genres and Sources*, pages 6–10.

Mahmoud El-Haj, Ahmed AbuRa'ed, Marina Litvak, Nikiforos Pittaras, and George Giannakopoulos. 2020a. 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.

Mahmoud El-Haj, Vasiliki Athanasakou, Sira Ferradans, Catherine Salzedo, Ans Elhag, Houda Bouamor, Marina Litvak, Paul Rayson, George Giannakopoulos, and Nikiforos Pittaras. 2020b. Proceedings of the 1st joint workshop on financial narrative processing and multiling financial summarisation. In *Proceedings of the 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation*.

Mahmoud El-Haj, Paul Rayson, Paulo Alves, Carlos Herrero-Zorita, and Steven Young. 2019a. Multilingual financial narrative processing: Analysing annual reports in english, spanish and portuguese. *World Scientific Publishing*.

Mahmoud El-Haj, Paul Rayson, Martin Walker, Steven Young, and Vasiliki Simaki. 2019b. 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.

Mahmoud El-Haj, Paul Rayson, Steven Young, and Martin Walker. 2014. Detecting document structure in a very large corpus of uk financial reports. In *European Language Resources Association (ELRA)*.

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.

Katja Filippova, Mihai Surdeanu, Massimiliano Ciaramita, and Hugo Zaragoza. 2009. Company-oriented extractive summarization of financial news. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 246–254.

John Flowerdew and Michaela Mahlberg. 2009. *Lexical cohesion and corpus linguistics*, volume 17. John Benjamins Publishing.

Pascale Fung and Grace Ngai. 2006. One story, one flow: Hidden markov story models for multilingual multidocument summarization. *ACM Transactions on Speech and Language Processing (TSLP)*, 3(2):1–16.

Kavita Ganesan. 2018. Rouge 2.0: Updated and improved measures for evaluation of summarization tasks. *arXiv preprint arXiv:1803.01937*.

George Giannakopoulos. 2019. Proceedings of the workshop multiling 2019: Summarization across languages, genres and sources. In *Proceedings of the Workshop MultiLing 2019: Summarization Across Languages, Genres and Sources*.

George Giannakopoulos, Vangelis Karkaletsis, George Vouros, and Panagiotis Stamatopoulos. 2008. Summarization system evaluation revisited: N-gram graphs. *ACM Transactions on Speech and Language Processing (TSLP)*, 5(3):1–39.

Vishal Gupta and Gurpreet Singh Lehal. 2010. A survey of text summarization extractive techniques. *Journal of emerging technologies in web intelligence*, 2(3):258–268.

Canasai Kruengkrai and Chuleerat Jaruskulchai. 2003. Generic text summarization using local and global properties of sentences. In *Proceedings IEEE/WIC International Conference on Web Intelligence (WI 2003)*, pages 201–206. IEEE.

Moreno La Quatra and Luca Cagliero. 2020. End-to-end training for financial report summarization. In *Proceedings of the 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation*, pages 118–123.

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

Marina Litvak, Mark Last, and Menahem Friedman. 2010. A new approach to improving multilingual summarization using a genetic algorithm. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 927–936.

Marina Litvak and Natalia Vanetik. 2013. Mining the gaps: Towards polynomial summarization. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 655–660.

Marina Litvak, Natalia Vanetik, and Zvi Puchinsky. 2020. Sce-summary at the fns 2020 shared task. In *Proceedings of the 1st Joint Workshop* *on Financial Narrative Processing and MultiLing Financial Summarisation*, pages 124–129.

Shuhua Liu and Johnny Lindroos. 2006. Experiences from automatic summarization of imf staff reports. *Practical Data Mining: Applications, Experiences and Challenges*, page 43.

Hans Peter Luhn. 1958. The automatic creation of literature abstracts. *IBM Journal of research and development*, 2(2):159–165.

Victoria McCargar. 2004. Statistical approaches to automatic text summarization. *Bulletin of the American Society for Information Science and Technology*, 30(4):21–25.

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.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

N Moratanch and S Chitrakala. 2016. A survey on abstractive text summarization. In *2016 International Conference on Circuit, power and computing technologies (ICCPCT)*, pages 1–7. IEEE.

Tatsunori Mori. 2002. Information gain ratio as term weight: the case of summarization of ir results. In *COLING 2002: The 19th International Conference on Computational Linguistics*.

Ani Nenkova and Kathleen McKeown. 2011. *Automatic summarization*. Now Publishers Inc.

D Radev, H Jing, and M Budzikowska. 2000. Centroid-based summarization of multiple documents: Clustering, sentence extraction, and evaluation. In *Proceedings of the ANLP/NAACL-2000 Workshop on Summarization*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

Abhishek Singh. 2020. Point-5: Pointer network and t-5 based financial narrative summarisation. In *Proceedings of the 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation*, pages 105–111.

Jaime Baldeon Suarez, Paloma Martínez, and Jose Luis Martínez. 2020. Combining financial word embeddings and knowledge-based features for financial text summarization uc3m-mc system at fns-2020. In *Proceedings of the 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation*, pages 112–117.

Amit Vhatkar, Pushpak Bhattacharyya, and Kavi Arya. 2020. Knowledge graph and deep neural network for extractive text summarization by utilizing triples. In *Proceedings of the 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation*, pages 130–136.

Deshraj Yadav, Rishabh Jain, Harsh Agrawal, Prithvijit Chattopadhyay, Taranjeet Singh, Akash Jain, Shiv Baran Singh, Stefan Lee, and Dhruv Batra. 2019. Evalai: Towards better evaluation systems for ai agents.

Siyan Zheng, Anneliese Lu, and Claire Cardie. 2020. Sumsum@ fns-2020 shared task. In *Proceedings of the 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation*, pages 148–152.