Extractive Financial Narrative Summarisation using SentenceBERT-Based Clustering

Tuba Gokhan, Phillip Smith and **Mark Lee**
School of Computer Science
University of Birmingham, United Kingdom
{txg857|smithpm|m.g.lee}@cs.bham.ac.uk

Abstract

We participate in the FNS-2021 Shared Task: "Financial Narrative Summarisation" organized by at 3rd Financial Narrative Processing Workshop (FNP-2021). We build an unsupervised extractive automatic financial summarisation system for the specific task. In our approach to the FNS-2021 shared task, the documents are first analyzed and an intermediate bespoke document set created containing the most salient sections of the reports. Next, vector representations are created for the intermediate document set based on SentenceBERT. Finally, the vectors are then clustered and sentences from each cluster are chosen for the final generated report summaries. The achieved results support the proposed method's effectiveness.

1 Introduction

With the growth of financial sector along with the economy's growth and development, there are quite a few new companies emerging and going public. Investors often find long-form annual reports of various companies difficult to deal with because their content may be tedious or redundant. Going through the reports can be arduous to filter out effective key information by human inspection alone, hence an automatic summarisation system would be useful to help investors effectively understand important company information.

The Financial Narrative summarisation Shared Task for 2021 (FNS-2021)(El-Haj et al., 2021) aims to evaluate the performance of automatic summarisation methods applied to annual reports from UK corporations listed on The London Stock Exchange (El-Haj et al., 2020). Compared to reports prepared by U.S companies, these reports have a notably less rigid structure that makes summarisation particularly challenging. These reports can be divided into two main sections. The first section is a "narrative" section which is also known as a "front-end" section containing textual information and reviews by the company's management and board of directors; the second section is the "back-end" section which contains financial statements that tend to consist of tables of numerical data. The FNS-2021 shared task entails determining which the most important narrative sections are and then summarise these to achieve a summary of approximately 1000 words.

In this paper, we will discuss the solution that we develope for the FNS-2021 shared task. The data set used is the annual reports in the financial field provided by the organizer. Since there is often a lot of redundant information in the annual reports, it is planned to choose the most useful parts first as the intermediate documents to be summarized. Thus, to identify the salience of the sentences in order to extract the most relevant ones from the original financial report our system uses a combination of sentence embedding and clustering algorithms.

2 Related Work

In text summarization, two methods are generally used as extractive and abstractive summarization methods. Extractive summarization methods try to find the most important topics of the introductory document and select sentences for these selected concepts to form the summary. Many approaches have been proposed for this type of summarization. We focus on an unsupervised extractive approach in our work.

The idea of clustering sentences in a high-dimensional area has also been used in the past for text summarization (Bookstein et al., 1995; McKeown and Radev, 1995; McKeown et al., 1999). However, these systems use TF-IDF representations of the sentence rather than sentence embeddings. Another class of vector-space-based methods uses Latent Semantic Indexing (Deerwester et al., 1990) to define sentences that describe hidden concepts in the document. In this paper, a new method for summarizing financial documents is presented by combining the traditional method of

Data Type	Training	Validation	Testing
Report Full Text	3000	363	500
Gold Summaries	9873	1250	1673

Table 1: FNS-2021 Shared Task Dataset

clustering algorithms with an innovative method of sentence embedding.

3 Data

For this shared task, the data that we used consisted of 3,863 United Kingdom annual reports for corporations listed on the London Stock Exchange (LSE) between 2002 and 2017. Annual reports in the UK are long papers that have average approximately 80 pages, with some exceeding 250 pages. For the FNS-2021 shared task, these annual reports are separated into three sections: training, testing, and validation.

The complete text of each yearly report, as well as the gold-standard summaries, are included in the training and validation sets. Each annual report has at least three gold-standard summaries on average, with some reports including up to seven gold-standard summaries. The task participants are only provided access to the full texts for the testing data set. Further details are shown in Table 1.

4 System

4.1 Pre-processing

Before development of our system, the gold standard summaries are examined in detail. In the provided training set, we find that the main narrative sections are mostly under four headings: "Chief Executive's review", "At a glance", "Highlights" and "Chairman's Statement". These sections typically include summarized financial topics. The other sections contain a significant amount of statistical data, tables, graphs, and diagrams, therefore they are not included in the organizer's gold summary. The focus of the study is to determine narrative segments to build our summarisation system. For this reason, a condensed intermediate document covers only the sections we wish to appear in the final summaries is required. As a result, both the processing speed and the success percentage of the generated summaries is considerably improved.

We also note that these sections are generally in the first 10% of each report. We therefore extract these and a new data set contains only these sections are created. On this new condensed dataset, sentences are tokenized using the '`tokenize`' package from the NLTK library (Loper and Bird, 2002).

4.2 Sentence Embedding

One of the main problems of natural language processing is the encoding and symbolising of words or characters in a text so that the machine can, to a degree, understand them. Embedding is a mathematical method of mapping an object in one domain to another object in different domain. Sentence embeddings converts a sentence into a vector. The BERT architecture is chosen due to its high performance over other NLP algorithms for Sentence Embedding. BERT is built on transformer architecture (Vaswani et al., 2017). Two BERT models are published by Google for public use, one containing 110 million parameters and the other 340 million parameters. (Devlin et al., 2019).

Sentence-BERT (SBERT) is a modification of the pre-trained BERT network that uses Siamese and triplet network structures to derive semantically meaningful sentence embeddings that can be compared using cosine-similarity. This reduces the effort for finding the most similar pair from 65 hours with BERT / RoBERTa to about 5 seconds with SBERT, while maintaining the accuracy from BERT (Reimers and Gurevych, 2019). Due to its performance in the larger pre-trained BERT model, SBERT is ultimately chosen for our experiments. In our studies, the sentences are vectorised with SBERT. Different models are used in the vectorisation phase. The highest performing models and their results are shown in Table 2.

4.3 Clustering

Since the BERT model produces sentence embeddings, these sentences can be clustered to form dynamic summaries of the FNS-2021 shared task dataset. A maximum of 1000 words is expected in the output summaries. When the data set is examined, the number of created clusters must be less than 25 in order to create 1000-word summaries.

During our experiments, the Scikit-Learn Clus-

	Model Name	R-1/F	R-2/F	R-L/F	R-SU4/F
Our System 1	nli-mpnet-base-v2	0.48	0.25	0.41	0.29
Our System 2	distiluse-base-multilingual-cased-v2	0.48	0.25	0.40	0.29
Our System 3	nli-distilroberta-base-v2	0.47	0.25	0.40	0.29

Table 2: Results on validation set.

System	R-1/ F	R-2/ F	R-L /F	R-SU4 /F
TEXTRANK (baseline)	0.17	0.07	0.21	0.08
LEXRANK (baseline)	0.26	0.12	0.22	0.14
PointT-5	0.46	0.28	0.45	0.28
SumTO	0.42	0.24	0.39	0.26
HULAT	0.44	0.26	0.38	0.26
MUSE (topline)	0.5	0.28	0.45	0.32
Our System 3-1	0.47	0.25	0.4	0.29
Our System 2	0.48	0.26	0.4	0.29

Table 3: Results measured by the organizers for the test set.

tering library [1] is examined in detail. The K-means algorithm (MacQueen et al., 1967) is an unsupervised clustering algorithm which partitions a set of data, usually termed dataset into a certain number of clusters. K-Means is chosen as the appropriate model for clustering sentence embeddings from the BERT model because the algorithm is scalable of clustering amount and is applied on large data. For the final summary, sentences closest to the centroid are selected from the clusters. Euclidean distance is used to measure the distances to centroids.

5 Results

Following the development of the aforementioned system we evaluate as follows. The FNS-2021 Shared Task contest decides to use the ROUGE2 package[2] to evaluate the system outputs. The ROUGE2 package is a multilingual tool that implements ROUGE (Lin, 2004) metrics.

The method we develope to measure system performance in our experiments is applied on the validation dataset. The summaries we created are evaluated using Rouge-1, Rouge-2, Rouge-SU4 and Rouge-L metrics. The results obtain as a result of our measurements of the three systems with the highest performance among the results are shown in Table 2. The results of our systems' summaries produce scores very similar to each other. Since we develop an unsupervised approach, we only use validation data sets in our study.

In the FNS-2021 Shared Task contest, the gold summaries of the test set in which the results will be evaluated are not shared with the participants. The performance of the generated summaries is measured by the organizers over the test set. Table 3 shows the organizers' calculated scores for the three systems we provide. In Table 3, the results of our systems, the results of the baseline TEXRANK (Mihalcea and Tarau, 2004) and LEXRANK (Erkan and Radev, 2004) algorithms, the results of PointT-5 (Singh, 2020), SumTO (La Quatra and Cagliero, 2020) and HULAT (Baldeon Suarez et al., 2020), which are the systems with the highest performance in the FNS-2020 Shared task, and the results of the topline MUSE (Litvak et al., 2010) algorithm are presented.

The overall ranking of systems varies depending on the evaluation metric considered. When our results are compared to baseline algorithms, we achieve relatively successful performance in all metrics. Furthermore, when compared to FNS-2020 Shared Task, our results indicate good outcomes in the Rouge-1 and Rouge-SU4 metrics. And, when we compare it to the MUSE method, which is based on topline, we see that our results are slightly lower.

6 Discussion

In this study, a SentenceBERT-based clustering approach is proposed as an unsupervised method for the FNS Shared task. As a result of this approach, extractive summaries of less than 1000 words are

[1] https://scikit-learn.org/stable/modules/clustering.html
[2] https://github.com/kavgan/ROUGE-2.0

created. In order to create high quality summaries in this dataset, first and foremost, it is necessary to define the "Chief Executive's review", "At a glance", "Highlights" and "Chairman's Statement" sections that form the basis of gold summaries. Since complex text documents are produced as a result of converting the dataset from PDF, it makes this definition difficult. For this reason, the preprocessing phase is extended in our work.

Another challenge in this task is producing 1000-word summaries.The basis of our proposed approach is clustering. In order to create summaries of 1000 words, we need to limit the number of cluster sets to a maximum number of 25. However, in clustering approaches, it is necessary to determine the ideal number of clusters according to the data distribution. This number of clusters varies depending on the documents, and the restriction of the number of clusters causes sentences that do not have similar meanings to be included in the same cluster.

In addition, when creating sentence vectors, our method employs pre-trained language models. These models are created using various datasets. We believe that the use of fine-tuned language models using financial documents and terms to improve the performance of the study helps improving the performance of the summary system.

7 Conclusion

The paper describes an extractive summarisation approach to summarizing textual financial reports for the Financial Narrative Summarisation Shared Task (FNS-2021). The proposed approach relies on clustering sentence vectors created with sentence embedding. First, an intermediate document dataset covering the most important parts of the documents is prepared. Then, pre-trained language representation model Bidirectional Encoder Representations from Transformers (BERT) is utilized to generate sentence embeddings. Finally, the K-means clustering algorithm is applied to find similar sentences and a sentence vector representing the set is selected from each cluster for the final summary. Three systems are created using different sentence embedding models are submitted. The performance of the obtained summaries is measured with the ROUGE metric. Our approach outperforms the baseline algorithms in terms of performance when is compared to the literature, whereas the topline algorithm produce partially near results.

References

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, Barcelona, Spain (Online). COLING.

Abraham Bookstein, Shmuel T Klein, and Timo Raita. 1995. Detecting content-bearing words by serial clustering. In *Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 319–327.

Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407.

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, Minneapolis, Minnesota. 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.

Mahmoud El-Haj, Nadhem Zmandar, Paul Rayson, Ahmed AbuRa'ed, Marina Litvak, Nikiforos Pittaras, and George Giannakopoulos. 2021. The Financial Narrative Summarisation Shared Task (FNS 2021). In *The Third Financial Narrative Processing Workshop (FNP 2021)*, Lancaster, UK.

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.

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, Barcelona, Spain (Online). COLING.

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

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.

Edward Loper and Steven Bird. 2002. Nltk: The natural language toolkit. *arXiv preprint cs/0205028*.

James MacQueen et al. 1967. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA.

Kathleen McKeown, Judith L Klavans, Vasileios Hatzivassiloglou, Regina Barzilay, and Eleazar Eskin. 1999. Towards multidocument summarization by reformulation: Progress and prospects.

Kathleen McKeown and Dragomir R Radev. 1995. Generating summaries of multiple news articles. In *Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 74–82.

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.

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

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, Barcelona, Spain (Online). COLING.

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.