Proceedings of the 4th Financial Narrative Processing Workshop @ LREC 2022, pages 73–78
Marseille, 24 June 2022

Transformer-based Models for Long Document Summarisation in Financial Domain

Urvashi Khanna, Samira Ghodratnama, Diego Mollá, Amin Beheshti
Macquarie University
Sydney, New South Wales, Australia
{urvashi.khanna, samira.ghodratnama, diego.molla-aliod, amin.beheshti}@mq.edu.au

Abstract

Summarisation of long financial documents is a challenging task due to the lack of large-scale datasets and the need for domain knowledge experts to create human-written summaries. Traditional summarisation approaches that generate a summary based on the content cannot produce summaries comparable to human-written ones and thus are rarely used in practice. In this work, we use the Longformer-Encoder-Decoder (LED) model to handle long financial reports. We describe our experiments and participating systems in the financial narrative summarisation shared task. Multi-stage fine-tuning helps the model generalise better on niche domains and avoids the problem of catastrophic forgetting. We further investigate the effect of the staged fine-tuning approach on the FNS dataset. Our systems achieved promising results in terms of ROUGE scores on the validation dataset.

Keywords: Document summarisation, Financial documents, Longformer, LED, Sequential fine-tuning

1. Introduction

Large amounts of unstructured data generated electronically in different organisations makes decision-making and gaining insights challenging, especially in the financial domain. Financial reports are critical to a company's financial performance and provide a snapshot of its financial situation. Financial statements not only help executives and investors understand the company's financial position, assets, and liabilities, but also provide a sense of financial transparency. Investors and stakeholders use these reports to make informed investment decisions, and to either vote in favour of or against corporate actions. Annual reports of various organisations from around the world typically include income statements, cash flow, statements from the chief executive officer, highlights, reviews of operating, investing, and financing activities, auditor's reports, risk disclosures, press releases, and so on (El-Haj et al., 2020b).

Annual reports in the financial sector are typically over 180 pages long (Leidner, 2020). This overload of textual data that investors and stakeholders must read is a time-consuming and exhausting process. Furthermore, in order to maximise profits, it is critical to make financial decisions in the shortest amount of time possible. As a result, automatic summarisation makes use of technology to simplify the process of concisely summarising long financial documents.

Despite recent advancements in automatic summarisation approaches, summarising long financial documents remains difficult due to the lack of large-scale datasets. Furthermore, the requirement for domain knowledge experts to create human-written summaries complicates the situation. As a result, traditional summarisation approaches that generate a summary based on the content cannot produce summaries comparable to human-written summaries and are thus rarely used in practice.

The use of unsupervised pretraining for natural language tasks is being driven by the availability of huge amounts of raw text on the web, as well as ever-increasing computational processing capacity. Fine-tuning a Pre-trained Language Model (PLM) on the target dataset is the norm these days. These PLMs are already pretrained on a massive amount of data and achieve state-of-the-art results on most of the Natural Language Understanding (NLU) tasks (Devlin et al., 2019; Lan et al., 2019; Liu et al., 2019). Longformer-Encoder-Decoder (LED) (Beltagy et al., 2020), a variant of longformer, scales efficiently on sequence-to-sequence tasks for long input sequences of up to 16k tokens. LED has performed exceedingly well on long-summarisation datasets like arXiv and PubMed (Cohan et al., 2018).

Language models are usually pretrained on general language like news articles and Wikipedia data, and then adapted to domain-specific downstream tasks. However, domain-specific tasks face the issue of scarcity of good quality manually labelled data. Thus, an intermediate stage of fine-tuning on a larger related dataset before fine-tuning on the target dataset has been a widely used approach in different domains like financial, biomedical, and scientific articles (Lee et al., 2019; Yoon et al., 2019; Phang et al., 2020; Khanna and Mollá, 2021). This addition of an intermediate stage helps the model generalise better on niche domains and also avoids the problem of catastrophic forgetting. In this paper, we describe the experimental setup, and approach of our participating systems at the Financial

Narrative Summarisation (FNS) shared task[1]. Both our systems use LED as pretrained language model considering the size of the financial documents. We formulate the task as one of extractive summarisation and also investigate the effect of multi-stage fine-tuning via our submissions at the FNS shared task. All of our systems outperformed the current publicly available validation results of other state-of-the-art systems.

The rest of the paper is organized as follows: in Section 2, we provide an overview of the related work and literature. Section 3 reviews the FNS dataset in detail, pre-processing and post-processing techniques, and the evaluation metrics used in this work. Section 4 discusses the methodology behind the proposed systems. In Section 5, we present evaluation results before concluding the paper with remarks for future directions in Section 6.

2. Related Work

Summarisation of documents can either be extractive or abstractive. Extractive summarisation selects a subset of sentences from the text to create a summary; on the other hand, abstractive summarisation reorganises the text's language and, if necessary, adds new words or phrases to the summary. In past FNS workshops, both extractive (Gokhan et al., 2021; Orzhenovskii, 2021) and abstractive (Singh, 2020) approaches were applied.

Unsupervised approaches have been used previously for the extractive summarisation of documents. TextRank (Mihalcea and Tarau, 2004) and LexRank (Erkan and Radev, 2004) are graph-based ranking models used for text processing. MUSE (MUltilingual Sentence Extractor), which is a language-independent approach for summarising extractive documents, uses linear optimisation of various sentence ranking measures using a generic algorithm (Litvak et al., 2010).

Participants in past years of the FNS workshop series used a variety of machine learning techniques for automatic summarisation of financial documents. Baldeon Suarez et al. (2020) used a combination of machine learning and statistical methods to calculate the importance of sentences based on features such as keywords, position, similarity, and topics. Litvak et al. (2020) combined topic modelling and discourse structure based on heuristic assumptions to create a new method for hierarchical summarisation of reports. Krimberg et al. (2021) used the Term Frequency-Inverse Document Frequency (TF-IDF) weighing method to identify the top 1000 most important words in a document and extract them as the summary.

Litvak et al. (2010) used the MUSE tool to filter large financial summaries, then combined different techniques like BERT and node embeddings, a similarity graph, and finally a neural LSTM model to train for sentence classification (Litvak and Vanetik, 2021). The participants also explored a combination of knowledge graph and deep learning approaches (Arora and Radhakrishnan, 2020; Vhatkar et al., 2020).

Language models like BERT (Devlin et al., 2019) and T5 (Raffel et al., 2020) are also used by participants in the FNS shared task (La Quatra and Cagliero, 2020; Orzhenovskii, 2021). Zmandar et al. (2021b) proposed a method that uses a combination of pointer network (Vinyals et al., 2015) and T5. They first use pointer network to extract the important sentences from the documents and then paraphrase the extracted sentences using a T5 model. To bridge the two models, they also use policy-based reinforcement learning. Sentence-BERT based clustering has also been effectively used by Gokhan et al. (2021).

3. FNS Data

The FNS 2022 shared task is organised annually to illustrate the challenges and potential of using automatic text summarisation for financial text documents in Spanish, English and Greek languages. These financial text documents can be anything ranging from financial company disclosures, budgeting, company's future prospects, etc. The FNS dataset contains the text extracted from United Kingdom (UK), Spanish and Greek companies' financial reports that are published annually in PDF format (El-Haj et al., 2020a).

Participants are asked to provide concise single summaries extracted from important sections from the financial annual reports of UK companies. The system generated summaries should reflect on the analysis and appraisal of the businesses' financial pattern over the last year, as supplied by annual reports. The FNS golden reference summaries are not written by human experts; instead, the experts who have created the financial reports inform which sections in the annual reports are considered a summary of the entire annual report, and those sections are used as gold standard summaries.

A typical financial report includes both numerical and narrative sections. Numerical sections refer to tables about tax returns, budgeting, expenditure and financial statements. The narrative sections comprise annual or quarterly highlights of the company, their future outlook, statements from the board of directors and management, etc. In this shared task, the participants are required to extract information from key narrative sections and produce a concise summary for each annual report such that the length of the summary should not exceed 1000 words (Zmandar et al., 2021a).

Dataset	Reports	Summaries
Training	3000	9873
Validation	363	1250
Testing	500	N/A

Table 1: Statistics of FNS 2022 dataset for Training, Validation and Test.

[1] http://wp.lancs.ac.uk/cfie/fns2022/

Table 1 shows the number of reports in the training, validation and test data provided by the FNS organisers. In the training and validation data provided, there are around 3 to 7 golden summaries for each report.
Recall-Oriented Understudy for Gisting Evaluation (ROUGE) (Lin, 2004) is the measure we utilise to evaluate our systems. Text summarisation tasks are frequently evaluated using ROUGE metrics. ROUGE is a collection of metrics that compare system-generated summaries to a set of ideal or reference summaries automatically. There are several distinct ROUGE measures depending on the amount of granularity of texts between the system and reference summaries. Among all of them, we use ROUGE-N, ROUGE-SU, and ROUGE-L as these metrics are used by FNS organisers. The ROUGE-N measure calculates the overlap between the system-generated summary to be assessed and the reference summaries in terms of unigram, bigram, trigram, and higher-order n-grams. ROUGE-L measures the longest matching sequence of words, while ROUGE-SU measures the co-occurrence statistics based on skip-bigram plus unigrams. The overlap of word pairs with a maximum of two gaps between them is measured by skip-bigram (Ganesan, 2015).

4. Systems Overview

In this section, we describe the approaches used in our two systems and the experimental setup that we explored when addressing the shared task of FNS 2022.

4.1. Longformer-Encoder-Decoder (LED)

BERT-style transformer models typically limit the sequence length to 512 tokens as they scale quadratically due to their self-attention mechanism (Devlin et al., 2019; Liu et al., 2019). To overcome this memory and computational constraint for long sequences, Beltagy et al. (2020) introduced Longformer, a transformer architecture that utilises a self-attention pattern which scales linearly with the sequence length, allowing it to process long documents. Longformer has made it easier to process long documents for natural language tasks like question answering, long document classification, and co-reference resolution.
The original Transformer architecture (Vaswani et al., 2017) uses an encoder-decoder pipeline for generative sequence-to-sequence tasks like translation and text summarisation. Encoder-Deccoder architectures like BART (Lewis et al., 2020) and T5 (Raffel et al., 2020) have achieved good results on sequence-to-sequence tasks but are not able to scale to longer sequences. Longformer-Encoder-Decoder (LED), a Longformer variant (Beltagy et al., 2020), has both encoder and decoder transformer stacks and utilises their efficient local+global attention pattern that can handle the long text sequence-to-sequence tasks efficiently (Sutskever et al., 2014).
We decided to use LED as the pretrained model for all our experiments considering the average report length in the FNS dataset is around 80 pages. We have mainly focused on only English language summarisation and formulated the task as an extractive summarisation task.
In the FNS training and validation datasets, each report has 3 to 7 golden reference summaries. We examined the reports and the golden summaries, and discovered that at least one golden summary was extracted from the report as a continuous sequence of text or section. In addition, the majority of the reference summaries were located at the beginning of the report. To train our systems, we applied the same approach as Orzhenovskii (2021) and chose the summary that had at least one continuous block of text in the report and also the most intersection with other summaries as our golden summary.
Our system takes the first 8192 tokens from the report as input and the first 1024 tokens from the selected golden summary as the target output. The system generates 1024 tokens as output predictions. The ROUGE F1 metrics was very low when we used the 1025 generated tokens as predicted summary because the summary length was less than 1000 words. As a result, we identify the sequence of text in the input report that matches this generated text and choose 1000 words as the output summary.

Hyper-parameters	Values
source length	8192
target length	1024
epochs	3,5
learning rate	5e-5
batch size	1
beam size	2,4

Table 2: Training hyper-parameters.

In our experiments, the pretrained language model was "led-large-16384," along with its tokenizer, all of which are freely available from the Huggingface Transformers Library (Wolf et al., 2020). The hyper-parameters for both fine-tuning steps were set to the default values used by the Longformer developers, unless stated otherwise. The systems were trained on the training dataset to fine-tune the hyper-parameters and later validated on the FNS validation dataset. The hyper-parameters used are listed in Table 2. Due to computational limitations, we were only able to experiment with a batch size of 1. Note that in Table 3, "macquarie1" and "macquarie2" are variations of longformers with different hyper-parameters.

4.2. Sequential Fine-tuning

In our second approach for the system "macquarie3", we follow a sequential fine-tuning approach by first fine-tuning on a large dataset and then on the target FNS dataset. This intermediate stage of fine-tuning is ideal in this case due to the small size of the FNS

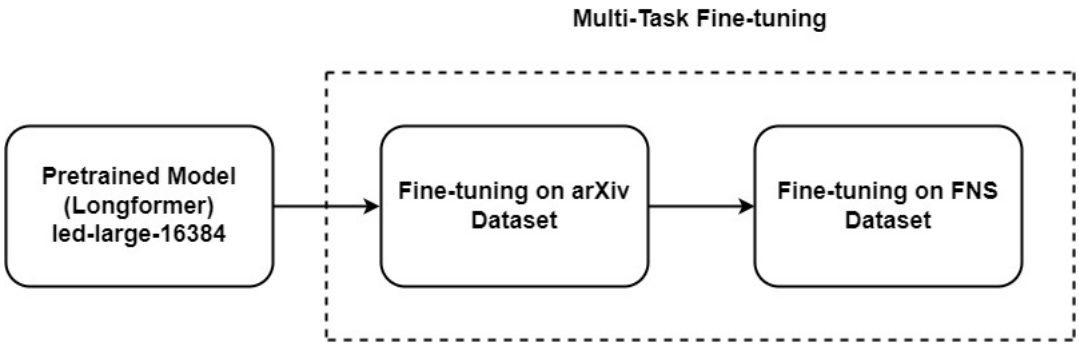

Figure 1: Diagram depicting our system's fine-tuning strategy.

System Name	R-1 / F	R-2 / F	R-L / F	R-SU4 / F
UoBNLP	0.480	0.250	0.400	0.290
TFIDF-SUM-3	0.433	0.209	0.374	0.250
MUSE	0.243	0.040	0.238	0.079
macquarie1	0.436	0.294	0.426	0.345
macquarie2	0.442	**0.302**	**0.434**	**0.353**
macquarie3	**0.443**	0.302	0.432	0.352

Table 3: Results of our systems on the FNS validation dataset. Our top scoring model is highlighted in bold. The rouge F-measure scores at unigram, bigram, longest common sub-sequence, and skip-gram based metrics are represented by the columns R-1/F, R-2/F, R-L/F, and R-SU4/L, respectively. UoBNLP (Gokhan et al., 2021), TFIDF-SUM-3 (Krimberg et al., 2021) are the validation results from past years and MUSE (Litvak et al., 2010) is the top baseline model. The highest score among our submissions is in bold.

dataset. We choose the arXiv summarisation dataset (Cohan et al., 2018), as there is no other large scale financial summarisation dataset that was readily available. We first fine-tune the "led-large-16384" model on the arXiv dataset and then on the target FNS dataset. We used the same hyper-parameters listed in Table 2. This approach is illustrated in Figure 1.

5. Results and Discussion

System Name	R-2 / F
Top Ranked System	0.374
macquarie1	0.303
macquarie2	0.301
macquarie3	0.302

Table 4: Results of our three submissions along with the top ranked system (LIPI) from the official FNS 2022 shared task results.

Table 3 contains the results of our validation experiments. Note that "macquarie1" and "macquarie2" are fine-tuned with the traditional approach and "macquarie3" is fine-tuned using the staged fine-tuning approach discussed in Section 4.2. "UoBNLP" is Sentence-BERT based system that applies clustering algorithm to generate dynamic summaries (Gokhan et al., 2021). "TFIDF-SUM-3" uses TF-IDF features to extract the important sentences to form summaries.

We used the ROUGE Java package[2] evaluation metrics as our main metrics for the evaluation of our models (Ganesan, 2015). FNS organisers also use ROUGE 2 as their main metric for ranking the teams' submissions on the leaderboard. ROUGE-2 F1 score on the test dataset is used for ranking the teams.

Based on the validation results, we observe that our systems performed better than the current state-of-the-art systems in all the ROUGE metrics except one (R-1/F). We also observed that there was no significant improvement in the performance of the system using the sequential fine-tuning approach. Table4 lists the results of our submissions in the FNS 2022 shared task. We observe that our results are similar to our validation results. However, other participants' systems performed better than ours in the FNS 2022 shared task [3].

On analysis of the predicted summaries, we found that LED is good at identifying the beginning part of the narrative section, however, the challenge still remains to identify the end span for a long length documents like financial reports. LED can handle up to 16K input tokens but not 16K decoder output tokens. The idea

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

[3]http://wp.lancs.ac.uk/cfie/fns2022/

behind LED was to be able to process very long inputs (articles to summarise) with the assumption that the decoder outputs did not have to be very long (summaries). This is also the reason for the model not showing any significant improvement when using the sequential fine-tuning approach.

6. Conclusion and Future Work

Our participation in the FNS 2022 was primarily focused on English language summarisation. We submitted three LED-based systems and also investigated the effect of sequential fine-tuning with the FNS dataset as our use case. Our systems performed better than the current state-of-the-art systems on the validation dataset. However, from our experiments we also found that staged fine-tuning had no impact on the performance of the system.

In future work, to locate the end span of the summary, the input sequence can be truncated into smaller chunks and fed into the language models. Later, each extracted summary could be concatenated to get the final summary. To capture the inter-sentence relationships better, graph-based neural networks can also be explored.

7. Bibliographical References

Arora, P. and Radhakrishnan, P. (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, Barcelona, Spain (Online), December. COLING.

Baldeon Suarez, J., Martínez, P., and Martínez, J. L. (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), December. COLING.

Beltagy, I., Peters, M. E., and Cohan, A. (2020). Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Cohan, A., Dernoncourt, F., Kim, D. S., Bui, T., Kim, S., Chang, W., and Goharian, N. (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, June. Association for Computational Linguistics.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (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, June. Association for Computational Linguistics.

El-Haj, M., AbuRa'ed, A., Litvak, M., Pittaras, N., and Giannakopoulos, G. (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), December. COLING.

El-Haj, M., Litvak, M., Pittaras, N., Giannakopoulos, G., et al. (2020b). 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.

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

Ganesan, K. (2015). Rouge 2.0: Updated and improved measures for evaluation of summarization tasks.

Gokhan, T., Smith, P., and Lee, M. (2021). Extractive financial narrative summarisation using sentencebert based clustering. In *Proceedings of the 3rd Financial Narrative Processing Workshop*, pages 94–98.

Khanna, U. and Mollá, D. (2021). Transformer-based language models for factoid question answering at bioasq9b. In *CLEF 2021 Working Notes*, CEUR Workshop Proceedings, pages 247–257. CEUR.

Krimberg, S., Vanetik, N., and Litvak, M. (2021). Summarization of financial documents with TF-IDF weighting of multi-word terms. In *Proceedings of the 3rd Financial Narrative Processing Workshop*, pages 75–80, Lancaster, United Kingdom, 15-16 September. Association for Computational Linguistics.

La Quatra, M. and Cagliero, L. (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), December. COLING.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., and Soricut, R. (2019). Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., and Kang, J. (2019). Biobert: pre-trained biomedical language representation model for biomedical text mining. *arXiv preprint arXiv:1901.08746*.

Leidner, J. L. (2020). Summarization in the financial and regulatory domain. In *Trends and Applications of Text Summarization Techniques*, pages 187–215. IGI Global.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer,

L. (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.

Lin, C.-Y. (2004). ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July. Association for Computational Linguistics.

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

Litvak, M., Last, M., and Friedman, M. (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.

Litvak, M., Vanetik, N., and Puchinsky, Z. (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, Barcelona, Spain (Online), December. COLING.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

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

Orzhenovskii, M. (2021). 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. Association for Computational Linguistics.

Phang, J., Calixto, I., Htut, P. M., Pruksachatkun, Y., Liu, H., Vania, C., Kann, K., and Bowman, S. R. (2020). English intermediate-task training improves zero-shot cross-lingual transfer too. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 557–575, Suzhou, China, December. Association for Computational Linguistics.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:1–67.

Singh, A. (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), December. COLING.

Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Vhatkar, A., Bhattacharyya, P., and Arya, K. (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, Barcelona, Spain (Online), December. COLING.

Vinyals, O., Fortunato, M., and Jaitly, N. (2015). Pointer networks. *Advances in neural information processing systems*, 28.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. M. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October. Association for Computational Linguistics.

Yoon, W., Lee, J., Kim, D., Jeong, M., and Kang, J. (2019). Pre-trained language model for biomedical question answering. *arXiv preprint arXiv:1909.08229*.

Zmandar, N., El-Haj, M., Rayson, P., Abura'Ed, A., Litvak, M., Giannakopoulos, G., and Pittaras, N. (2021a). The financial narrative summarisation shared task FNS 2021. In *Proceedings of the 3rd Financial Narrative Processing Workshop*, pages 120–125, Lancaster, United Kingdom, 15-16 September. Association for Computational Linguistics.

Zmandar, N., Singh, A., El-Haj, M., and Rayson, P. (2021b). 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. Association for Computational Linguistics.