T5-LONG-EXTRACT at FNS-2021 Shared Task

Mikhail Orzhenovskii
orzhan057@gmail.com

Abstract

This paper describes T5-LONG-EXTRACT system submitted to the Financial Narrative Summarization Shared Task (FNS-2021). We developed a task-specific extractive summarization method based on pre-trained language model T5 and implemented a filtering procedure for the source dataset. The language model was fine-tuned on long sequences (4096 tokens) to identify the beginning of a continuous narrative part of the document. Our system ranks 1st on Rouge-1, Rouge-2, Rouge-SU4, and Rouge-L official metrics.

1 Introduction

Since financial data is constantly growing, automatic summarization is becoming increasingly important. The goal of the Financial Narrative Summarization Shared Task (El-Haj et al., 2021) is to summarize the annual financial reports from UK firms listed on The London Stock Exchange. It is a challenging task since the provided documents are long and the structures seem to vary.

We analyzed the provided reports and gold summaries and identified the non-overlapping matching word sequences between the reports and the corresponding gold summaries. For 99.4% summaries, there was only one matching block containing the entire summary, so the summary is included in the report as a whole subsequence. Further, we refer to those summaries as continuous.

In the vast majority of the reports, there is at least one continuous gold summary. Because of that, we concentrated on extracting one continuous part of the report as a summary. To perform this task, we only need to find two positions in the text: the beginning of the narrative part and its end. Most continuous summaries begin within the first 4000 tokens of the text.

The task could then be framed as an extractive summarization task, where we used the first 64 tokens of the selected gold summary as the target sequence and the first 4096 tokens of the report as the source sequence. After locating the beginning of the narrative part in the report, we select 60 sentences (but no more than 1000 words) as the system's summary.

2 Dataset

The dataset (Table 1) includes annual reports produced by UK firms listed on The London Stock Exchange. The texts can be 80 pages long which makes it challenging to analyze them manually. In the training and validation set, there were from 3 to 7 gold summaries for each report. We used the training set to fine-tune the model and select the number of training epochs, and used the validation set to choose the best performing model configuration.

We explored continuous gold summaries, which are included in the corresponding report as a whole subsequence. 3761 gold summaries (38.3%) are longer than 1000 words (maximum length required by the shared task). The average length is 1086 words. Only 5% of the summaries are longer than 3214 words. The average number of sentences in the gold summaries is 36, and 80% of the summaries have 57 or less sentences.

The position of the continuous gold summaries is on average 1695 words from the beginning. Half of those summaries have a position less than 1003, so most of such summaries are located at the beginning of the report. For 85.1% of the continuous summaries the position in the report is less than 2775 words and less than 4000 T5 tokens.

3 System

We selected only one gold summary for each annual report which had at least one continuous gold summary located not more than 4096 tokens from the beginning. Because the target metric was calculated using the average ROUGE score between the system's summary and all gold summaries, we

Dataset	Reports	Summaries
Train	3000	9873
Valid	363	1250
Test	500	N/A

Table 1: Dataset size

picked the summary that had the most intersections with the other summaries. The score for summary Sc_i was calculated based on intersections of words between summary S_i and other summaries S_j:

$$Sc_i = \frac{\sum_j |x \in S_i \cap S_j|}{|x \in S_i|}$$

This schema improved the results of the system, compared to selecting just the first summary or using all summaries.

We picked only summaries that were at least 100 words long. Out of 2940 resulting training samples, 2740 were inputted into the model, and the remaining 200 samples were used to determine the number of training epochs.

T5 (Raffel et al., 2020) is a sequence-to-sequence language model, pre-trained on multiple tasks, including abstractive text summarization. It is configured for 512 input tokens; however, the model is based on relative position embeddings, which allows to scale it to longer input sequences. Because of the complexity $O(n^2)$ of the Transformer's self-attention mechanism such scaling significantly increases memory consumption. As a countermeasure, we truncated the output sequence length to 64 tokens.

We fine-tuned the T5 model on the filtered dataset. During the inference, the model outputs 64 tokens. We locate the closest match of these tokens in the report's full text (in most cases, there is an exact match). The summary is created from 60 sentences starting from this location (the number of the sentences was selected based on gold summary statistics). After that the system's summary is truncated to 1000 words. To keep the output readable, we preserve white space while splitting the text into words.

4 Experiments

For fine-tuning we used HuggingFace's Transformers (Wolf et al., 2020) with the following parameters (Table 2).

We explored the effect of language model size on the system's performance. In an experiment, t5-base achieved ROUGE-2 F1 0.3808 on the validation set, and t5-small scored 0.3846 in the same configuration. Due to the increased source length, the models require more GPU memory to train even with batch size 1 (16 GB for t5-small and 40 GB for t5-base), so we were not able to experiment with the larger models.

It is critical to select a source length of 4096 in this task, as reducing it results in significantly poorer performance. The best ROUGE-2 F1 score of the t5-base model with input size 2048 was 0.3621. Other scores were also lower than those of the original model.

5 Results

The official metrics are shown in Table 3. On the validation set the system achieved Rouge-2 F1 0.38, with precision 0.34 and recall 0.47. We examined the summaries, produced by the system, on the validation set and compare them to the gold summaries. Including extra content in the summary lowers the precision. The system's summary is 1.72 times longer than the corresponding gold summary. Using a more accurate method of identifying the end of the summary (instead of the heuristics) can further improve the system's precision and F1 score.

The position of the narrative part identified by the system exactly matched one of the gold summaries in 24.5% of the validation samples. These samples typically begin with the words "chairman's", "highlights" and "strategic". For 47.3% samples, the system's summary was not more than 50 words away from one of the gold summaries. The accuracy of locating the summary is negatively correlated with the position: the farther the summary is placed in the text, the more difficult it is identify.

6 Conclusion and Future Work

In this paper, we describe T5-LONG-EXTRACT system for the Financial Narrative Summarization Shared Task. The proposed method achieved the highest score in all metrics.

Parameter	**Value**
Base model	t5-small
Maximum source length	4096
Maximum target length	64
Batch size	1
Warm-up ratio	0.1
Learning rate	5e-05
Training epochs	12

Table 2: Training parameters

System	**ROUGE-1 F1**	**ROUGE-2 F1**	**ROUGE-L F1**	**ROUGE-SU4 F1**
orzhan (T5-LONG-EXTRACT)	0.54	0.38	0.52	0.43
MUSE (TOPLINE)	0.5	0.28	0.45	0.32
POLY-baseline	0.37	0.12	0.26	0.18
LEXRANK (baseline)	0.26	0.12	0.22	0.14
TEXTRANK (baseline)	0.17	0.07	0.21	0.08

Table 3: Official scores of the system and the baselines. ROUGE Evaluation on the test set

T5-LONG-EXTRACT's source code and the weights of the fine-tuned model are available[1] and can be run using the provided instructions.

In the future, the language model can be trained in a multi-task setting to locate the end of the narrative section as well as the beginning. To reduce the memory requirements, sparse attention models (Longformer (Beltagy et al., 2020) or BigBird (Zaheer et al., 2020)) can be used as the encoder of the sequence-to-sequence model.

References

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv:2004.05150*.

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.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Huggingface's transformers: State-of-the-art natural language processing.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. 2020. Big bird: Transformers for longer sequences. *Advances in Neural Information Processing Systems*, 33.

[1] https://github.com/orzhan/t5-long-extract