Proceedings of the 4th Financial Narrative Processing Workshop @ LREC 2022, pages 112–115
Marseille, 24 June 2022

Multilingual Financial Documentation Summarization by Team_Tredence for FNS2022

Manish Pant, Ankush Chopra
Tredence Analytics
Bengaluru (India)
{manish.pant, ankush.chopra}@tredence.com

Abstract

This paper describes multi-lingual long document summarization systems submitted to the Financial Narrative Summarization Shared Task (FNS 2022[1]) by Team-Tredence. We developed task-specific summarization methods for 3 languages – English, Spanish and Greek. The solution is divided into two parts, where a RoBERTa model was finetuned to identify/extract summarizing segments from English documents and T5 based models were used for summarizing Spanish and Greek documents. A purely extractive approach was applied to summarize English documents using data-specific heuristics. An mT5 model was fine-tuned to identify potential narrative sections for Greek and Spanish, followed by finetuning mT5 and T5(Spanish version) for abstractive summarization task. This system also features a novel approach for generating summarization training dataset using long document segmentation and the semantic similarity across segments. We also introduce an N-gram variability score to select sub-segments for generating more diverse and informative summaries from long documents.

Keywords : Long Document Summarization, Abstractive Summarization, Extractive Summarization

1. Introduction

Huge corpus of financial documents is published around the world in various languages. These documents hold enormous information that can be very useful for the finance analysts and market stakeholders if it could be streamlined, structured, or summarized into a concise piece of text. Automating this task using NLP techniques can substantially reduce the gap between supply of unstructured text data and the availability of consumable piece of text information.

The objective of Financial Narrative Summarization (FNS 2022) (Zmandar et al., 2022) was to implement a system for automating text summarization of financial text written in **English, Spanish** and **Greek**. The task dataset was extracted from annual reports of the firms listed on UK, Spanish and Greek stock exchanges, published in the pdf format. The details of work submitted by various teams is collated by (Mahmoud et al., 2022).

The expected outcome was to provide structured single summaries, based on real-world, publicly available financial annual reports by extracting information from different key sections and generate summaries that reflects the analysis and assessment of the financial trend of the business over the past year, as provided by annual reports. The summary length should not exceed **1000 words.**

Gold summaries for English language reports were found to be extractive in nature with around 99.9% summaries as continuous word subsequences of reports. There were one or more gold summaries provided for each report. This task was framed to be purely extractive, where we classified smaller segments of the reports as summary segments and heuristically selected top-n segments as system generated summary.

Gold summaries for Greek and Spanish language reports were identified to be abstractive in nature. We implemented a text classifier to mark line/segment of reports as narrative sections. The classified segments were clustered into semantically related segments of reports. These cluster of report segments were summarized as system generated summaries using transformers-based models (Vaswani et al., 2017).

Next, we'll describe the dataset provided by the organizers followed by the systems we developed. We'll then briefly talk about the experiments, results and highlight our learning in the conclusion section.

2. Dataset

The dataset includes annual reports produced by UK, Spanish and Greek firms listed on the Stock Exchange for each of those markets.

The texts can be up to 80 pages long which makes it challenging to analyze them manually. The English summaries were extractive in nature and were created by taking multiple contiguous sentences from the original reports. Spanish and Greek summaries were abstractive in nature and were coming from the Chairman's letter or equivalent section.

Data Type	Train	Val	Test	Total
Report text	3000	363	500	3863
Gold summaries	9873	1250	1673	12796

Table 1: FNS 2022 Shared Task Dataset - English

Data Type	Train	Val	Test	Total
Report text	162	50	50	262
Gold summaries	324	100	100	524

Table 2: FNS 2022 Shared Task Dataset - Spanish

[1] FNS 2022 – FNP 2022 (lancs.ac.uk)

Data Type	Train	Val	Test	Total
Report text	162	50	50	262
Gold summaries	324	100	100	524

Table 3: FNS 2022 Shared Task Dataset – Greek

We used training set of each language to fine-tune the model and used the validation set to determine the best performing model configurations.

In English training set, we had 3000 annual reports and 9873 gold summaries. On an average 3 golden summary available for each report. The average golden summary is 1084 words long and average annual report length is 46167 words. Table 1 has the details of the English dataset.

In Spanish training set (Table 2), we had 162 annual reports and 324 gold summaries, such that there are exactly 2 golden summaries for each annual report. The average golden summary is 878 words long and average annual report length is 39980 words.

In Greek training set (Table 3), we had 162 annual reports and 324 gold summaries, such that there are exactly 2 golden summaries for each annual report. The average annual report length is 28360 words and average golden summary is 7353 words long while the median length is 1514 words. It was noted that Greek summary length had a very skewed distribution due to outliers.

3. Systems

The final submission was composed of 3 systems. These systems were combination of 2 English and Greek solutions each and 1 Spanish Solution that we developed.

3.1 English Solutions

Only 0.01% of records were such where given summaries were not contiguous subsets of reports. We discarded these records from data. We divided each report into smaller text segments of 250 words. We experimented with segment of various lengths and empirically decided 250 as optimal cutoff.

We then compared these generated segments with the given summary text. Comparison was done at unigram token level. Any report segment with an overlap of more than 75-word tokens with summary text was considered to have potential towards summary creation and marked as positive. Segments with no overlap were marked as negative. Segments that had overlapping words between 0 to 74 were kept away from the modelling.

3.1.1 Summary Identification/Extraction Module

Above stated approach was used to generate train and validation dataset. We fine-tuned base version of the RoBERTa (Liu et al., 2017) model for classifying the report segment to be candidate summary segment or not. The best model achieved F1-score of 0.76 on the validation set.

During the inference, the report is first broken into segments of 250 words each except last segment. Each of these segments are scored using finetuned RoBERTa model.

Since the organizers have put a limit of max 1000-words for the system generated summaries, we select 4 candidate summary segments to make the final complete summary. We came up with 2 methods for final 4 segment selection.

In first solution, we select 4 continuous segments sequence such that the mean confidence score of prediction is maximized. This was done to mimic the process that was used for summary preparation by organizers.

In solution 2, we introduced the bi-gram variability score associated with each segment. We used this to reduce repetition of information across different segments for final summary. Bi-gram variability score for summary segment “S_i” was calculated based on count of bigrams in given candidate summary segment “C_i” and all other candidate summary segments “C_k” in given report:

$$S_i = \frac{Ci}{\sum_{k=1}^{y} Ck}$$

All the segments with score of more than 0.75 from RoBERTa model are considered as candidate segments. Top 4 candidates based on bi-gram variability score are selected as final summary of the report from all the candidate segments.

3.2 Spanish and Greek Solutions

The solutions for Spanish and Greek report summarization have 2 main submodules, summary identification and abstractive summarization. Both Greek and Spanish solutions are almost identical, with only difference being the base-model used for finetuning abstractive summarization task.

We divided each report into smaller segments delimited by the new line characters. We dropped lines with less than 5 words. Similarly, each summary was segmented into multiple lines and filtered. Embeddings for each segmented line of report and summary was generated using the sentence transformer (Reimers & Gurevych, 2019) framework. We used multilingual-*mpnet-base-v2 model* (Song et al., 2020) *within this framework*. Using these embeddings, we calculated the cosine similarity of each report line against each summary line. The report lines with similarity score above 0.65 against any summary line, were marked as positive for candidate summary classification model dataset. All the remaining lines from reports were marked as negative.

3.2.1 Summary Identification/Extraction Module

Above stated approach was used to generate train and validation datasets. We finetuned a multilingual T5 model to classify between the positive and negative candidate report segments. The classifier achieved an f1 score of 0.29 on Spanish validation set and 0.65 on Greek validation set. We trained a single multilingual model for Spanish and Greek combined to classify report lines for being candidate input to summary extraction.

3.2.2 Abstractive Summarization Module

We scored the candidate segments using the previous module to generate the training data for this module. Since T5 (Raffel et al., 2020) is seq-to-seq model (Sutskever et al., 2014) we took all the lines where label 1 was generated output as candidates. We selected only such candidate report lines that had a cosine similarity score higher than 0.65 with any of the summary lines. Again, we used the same sentence transformer model for embedding generation. We generated the dataset for abstractive summarization model using each summary line as target sequence and top-4 similar candidate report lines as the input sequence. This approach was applied for generating both training and validation dataset.

We finetuned Google's mt5-small model for Greek and a Spanish-t5 model from flax community in Huggingface[2].

During inference, the report is broken into lines and scored using first submodule (classifier). Sentence embeddings are generated for all lines/segments from report which were classified as candidate input for summarization model.

We needed to group the candidate input lines into clusters so that a sizable text can be provided as input to the abstractive summarization model. We implemented two methods for Greek and one for Spanish after experimenting with few ideas. For Greek Solution-1 (first method), the classified summary segments of test set were grouped into 5 clusters. These 5 clusters were input to the abstractive summarization model and the output was the system generated summary segments.

Implementation of Spanish and 2nd Greek method (Solution-2) were same. We clustered the candidate report lines into 16 clusters and calculated centroid for each of these clusters. These 16 cluster centroids were used to select top-5 similar report lines to each cluster centroid. This resulted in 16 clusters of 5 similar lines each. These 16 clusters were input to the abstractive summarization model and the output was the system generated summary segments.

4. Experiments

4.1 English Solutions

We experimented with different overlap word lengths and segment word lengths for English summary identification model training dataset. RoBERTa-base model (Liu et al., 2017) variant was able to generalize well with overlap length in the range of 60-90 words and segment length of 250-350 words. We found the most optimal overlap length of 75 words and segment length of 250 words length. Also dropping the boundary case data points with overlap between 0-75 words improved the f1 score to 0.76. It was also critical to use the bigram variability score in final segment selection, which helped in ensuring the selected segments with least repetition of information. Final model was trained for 5 epochs, with learning rate of 1e-6 and AdamW (Loshchilov & Hutter, 2019) optimizer. We chose a batch size of 32 for both train and validation sets.

4.2 Spanish and Greek Solutions

Fine-tuned mT5-small model (Xue et al., 2021) performed the best compared to few other models we tried for candidate classification for both Greek and Spanish. It did not fare well for Spanish, when we used it for abstractive summarization as well. It performed well for Greek in abstractive summarization application though. Using a Spanish language specific model proved to be better since it clearly outdid mT5 model when we compared validation set performance of both the models using Rogue-2 (Ganesan, 2006) scores.

It was also critical to use semantic similarity embeddings for artificially creating summarization training dataset and clustering the semantically related lines for input generation for summarization submodule.

We observed that certain clusters with fewer lines tend to perform relatively worse due to lack of enough context for summarization which led to the idea of clustering with repetition. For Greek Solution-2 and Spanish, we clustered based on top-n similar data points to a given cluster centroid which enabled the consistent length and context for summarization model input.

The mT5 model used for candidate classification of both Spanish and Greek was trained with input sequence length of 128. Model was trained for 4 epochs with batch size of 8, learning rate of 1e-4 and AdamW (Loshchilov & Hutter, 2019) optimizer.

Abstractive summarization for Spanish was done by a T5 models pretrained for Spanish corpora. We finetuned it for input and output sequence lengths of 700 and 180 respectively. Model was finetuned for 40 epochs with learning rate of 3e-4 and batch size of 1.

We used the mT5 model for Abstractive summarization for Greek. We finetuned it for input and output sequence lengths of 1024 and 256 respectively. Model was finetuned for 30 epochs with learning rate of 3e-5 and batch size of 1.

All the models were finetuned on NVIDIA RTX3090 system.

5. Results

Rouge-2 (Ganesan, 2006) F1 score was the official metric for evaluating system performance for each language. The final score was weighted 0.5, 0.25, and 0.25 for English, Spanish, and Greek respectively. We submitted 3 systems to the competition and achieved an overall team rank 4. Our best scoring system was composed of solution-2 of both English and Greek sole, Spanish submission that we made. The final weighted score of best performing system was 0.228. Below table has the results for all the solutions that we submitted.

[2] Hugging Face – The AI community building the future.

Language	Solution	Rogue-2 Recall	Rogue-2 Precision	Rogue-2 F1-Score
English	Solution1	0.305	0.363	0.317
	Solution2	0.346	0.323	0.322
Greek	Solution1	0.043	0.415	0.072
	Solution2	0.097	0.321	0.138
Spanish	Solution1	0.134	0.149	0.131

Table 4: Results

6. Conclusion

We built the final system by dividing the problem into two. This division was done after analyzing the nature of input and output data. English summaries were purely extractive in nature whereas Greek and Spanish were abstractive.

Using a more sophisticated approach for final segment selection in English system could marginally improve the scores. Instead of picking the top-n segments, any seq2seq model could be trained to predict the start and end of summary on a combined corpus of selected sections.

We could also experiment using larger mt5 models for Spanish and Greek summarization which requires higher GPU memory for fine-tuning. Also, few language-specific text generation models could be finetuned to compare the performance with existing multilingual model for Greek and Spanish individually.

7. Bibliographical References

Ganesan, K. (2006). ROUGE 2.0: Updated and Improved Measures for Evaluation of Summarization Tasks. Computational Linguistics, 1(1).

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach.

Loshchilov, I., & Hutter, F. (2018, September). Decoupled Weight Decay Regularization. In International Conference on Learning Representations.

Mahmoud El-Haj, et al., editors. (2022). Proceedings of the 4th Financial Narrative Processing Workshop, Marseille, France, 24 June. The 13th Language Resources and Evaluation Conference, LREC 2022.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... & 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.

Reimers, N., Gurevych, I., Reimers, N., Gurevych, I., Thakur, N., Reimers, N., ... & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (pp. 671-688). Association for Computational Linguistics.

Song, K., Tan, X., Qin, T., Lu, J., & Liu, T. Y. (2020). Mpnet: Masked and permuted pre-training for language understanding. Advances in Neural Information Processing Systems, 33, 16857-16867.

Sutskever, I., Vinyals, O., & 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., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.

Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., ... & Raffel, C. (2021, June). mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 483-498).

Zmandar, N., El-Haj, M., Rayson, P., Abura'Ed, A., Litvak, M., Giannakopoulos, G., Pittaras, N., Carbajo-Coronado, B., and Moreno-Sandoval, A. (2022). The financial narrative summarisation shared task (fns 2022). In Proceedings of the 4th Financial Narrative Processing Workshop, Marseille, France, 24June. The 13th Language Resources and Evaluation Conference, LREC 2022.