EasyChair Preprint

№ 8840

LIPI at Financial Narrative Summarisation 2022: Extractive Approaches for Summarizing Multi-Lingual Financial Reports

Sohom Ghosh and Sudip Kumar Naskar

September 17, 2022

LIPI at Financial Narrative Summarisation 2022: Extractive approaches for summarizing Multi-Lingual Financial Reports

Sohom Ghosh[1,2], Sudip Kumar Naskar[2]
[1]Fidelity Investments, [2]Jadavpur University
[1]Bengaluru,India [2]Kolkata,India
{sohom1ghosh, sudip.naskar}@gmail.com

Abstract

In this fast-moving world, it is difficult for investors to read lengthy annual reports of organizations. However, they need to make investment-related decisions relying on facts present in these documents. We participated in the third edition of the Financial Narrative Summarization (2020) shared task. It consisted of summarizing financial documents in three languages: English, Spanish, and Greek. In this paper, we describe the approaches followed by our team, LIPI. Among all the participants, we performed the best for the English sub-task in terms of ROUGE scores. Our source code is available at https://github.com/sohomghosh/Financial_Narrative_Summarization_2022.

Keywords: Financial Texts, Text summarization, Natural Language Processing

1. Introduction

The financial world is changing rapidly due to digitalization. To cope with the fast-moving financial markets, investors need to make important decisions in less time. Annual reports submitted by the organizations listed in various stock exchanges comprise detailed descriptions of their financial health. Investors rely on these reports to understand the market better. However, these reports are quite lengthy. Reading them from end to end is time-consuming. This brings in the need for an automated system that can extract import sections from these documents.

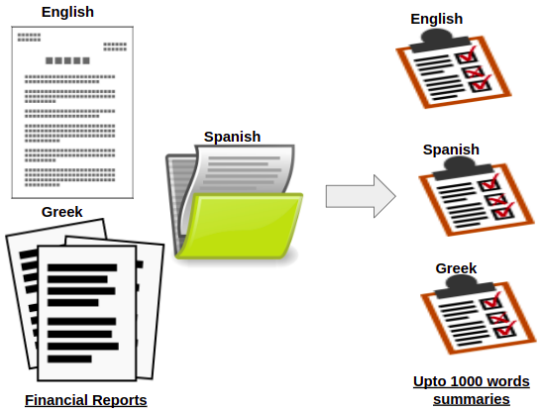

Figure 1: Extracting up to 1000 word summaries from Financial Reports

Text summarization is one of the trending areas of research in Natural Language Processing. There are primarily two kinds of text summarization approaches: Extractive and Abstractive. Financial Narrative Summarization (FNS) is the first of its kind shared task that deals with the summarization of narratives present in annual reports of organizations. The first two edition of FNS(El-Haj et al., 2020), (Zmandar et al., 2021) dealt with summarizing financial narratives in English. The third edition of FNS which is being held this year has been expanded to two more languages: Spanish and Greek. The length of the extracted summary has been restricted to 1000 words and the standard ROUGE metric is used for evaluation.

2. System Description

2.1. English

For this, we first compared the dataset of FNS 2022 with that of FNS 2021. Since the datasets were almost the same, we scored the best performing model of FNS 2021. This was developed by Orzhenovskii (Orzhenovskii, 2021) and has been open-sourced[1]. It uses Text-To-Text Transfer Transformer (T5) model (Raffel et al., 2020) for extracting relevant texts.

2.2. Spanish

We used the SummerTime Python library (Ni et al., 2021) to obtain summaries of the Spanish documents using mBART (Tang et al., 2020) and mt5 (Xue et al., 2021) models. Subsequently, we ranked each of the sentences using LexRank (Erkan and Radev, 2004) and phrases using TextRank (Mihalcea and Tarau, 2004). We considered the top 1000 phrases for further analysis. We re-ranked the sentences based on the sum of lex rank scores and the average of text rank scores of the constituent phrases. We append this to the summaries obtained using mBART (Tang et al., 2020) and mt5 (Xue et al., 2021) models and consider only the first 1000 tokens.

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

2.3. Greek

For the Greek documents, we first extract the sentences using Spacy [2]. We rank them using LexRank (Erkan and Radev, 2004) and consider the first 1000 tokens.

2.4. Results

We present the test set results provided by the organizers for our team in Table 1. We observe that for the English sub-task we perform the best. However, for the other two languages, we did not do as good.

3. Future Works

In future, we would like to develop a unified framework that would extract summaries from financial narratives irrespective of language. Exploring how extractive and abstractive summarization can be combined is another dimension for pursuing future research.

4. Bibliographical References

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

Erkan, G. and Radev, D. R. (2004). Lexrank: Graph-based lexical centrality as salience in text summarization. *J. Artif. Int. Res.*, 22(1):457–479, dec.

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, Barcelona, Spain, July. Association for Computational Linguistics.

Ni, A., Azerbayev, Z., Mutuma, M., Feng, T., Zhang, Y., Yu, T., Awadallah, A. H., and Radev, D. (2021). Summertime: Text summarization toolkit for non-experts. *arXiv preprint arXiv:2108.12738*.

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.

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(140):1–67.

Tang, Y., Tran, C., Li, X., Chen, P.-J., Goyal, N., Chaudhary, V., Gu, J., and Fan, A. (2020). Multilingual translation with extensible multilingual pretraining and finetuning.

Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., and Raffel, C. (2021). 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*, pages 483–498, Online, June. Association for Computational Linguistics.

Zmandar, N., El-Haj, M., Rayson, P., Abura'Ed, A., Litvak, M., Giannakopoulos, G., and Pittaras, N. (2021). 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.

[2] https://spacy.io/

Language	R-L / R	R-L / P	R-L / F	R-SU4 / R	R-SU4 / P	R-SU4 / F	Rank
English	0.559	0.449	0.487	0.515	0.369	0.417	1
Spanish	0.098	0.325	0.146	0.069	0.291	0.107	10
Greek	0.081	0.509	0.137	0.046	0.402	0.08	10

Language	R-1 / R	R-1 / P	R-1 / F	R-2 / R	R-2 / P	R-2 / F	Rank
English	0.587	0.451	0.496	0.472	0.326	0.374	1
Spanish	0.142	0.58	0.217	0.045	0.196	0.07	10
Greek	0.101	0.625	0.17	0.026	0.33	0.046	10

Table 1: Results. R-L, R-SU, R-1 and R-2 denotes ROUGE-L, ROUGE-SU, ROUGE-1 and ROUGE-2 scores respectively. R, P and F represents Recall, Precision and F1 scores.