Proceedings of the 4th Financial Narrative Processing Workshop @ LREC 2022, pages 53–58
Marseille, 24 June 2022

Multilingual Text Summarization on Financial Documents

Negar Foroutan*, Angelika Romanou*, Stéphane Massonnet, Rémi Lebret, Karl Aberer
École Polytechnique Fédérale de Lausanne (EPFL)
firstname.lastname@epfl.ch

Abstract

This paper proposes a multilingual Automated Text Summarization (ATS) method targeting the Financial Narrative Summarization Task (FNS-2022). We developed two systems; the first uses a pre-trained abstractive summarization model that was fine-tuned on the downstream objective, the second approaches the problem as an extractive approach in which a similarity search is performed on the trained span representations. Both models aim to identify the beginning of the continuous narrative section of the document. The language models were fine-tuned on a financial document collection of three languages (English, Spanish, and Greek) and aim to identify the beginning of the summary narrative part of the document. The proposed systems achieve high performance in the given task, with the sequence-to-sequence variant ranked 1st on ROUGE-2 F1 score on the test set for each of the three languages.

Keywords: multilingual text summarization, language models, natural language processing

1. Introduction

Machine Learning and Natural Language Processing have seen a tremendous increase in applications in the financial sector, mainly due to the need for automated approaches addressing financial tasks on both qualitative and quantitative data. Financial narrative summarization is a task that has drawn the attention of academia over the past couple of years with works regarding financial reports summarization (Suarez et al., 2020; Abdaljalil and Bouamor, 2021; Orzhenovskii, 2021) or financial news summarization (Passali et al., 2021). This is mainly because these computer-aided techniques could have an actual impact by saving considerable human manual annotation time and effort.

In this paper, we present our system regarding the Financial Narrative Summarization Shared Task [1] which aims to summarize the annual financial reports from international firms in three different languages: English, Greek, and Spanish. The input datasets are comprised of annual reports along with a set of human-curated summaries for each report, made by different annotators. Based on data statistics that will be presented in detail in Section 3, the summaries are created based on both extractive and abstractive approaches. This, along with the multilingual nature of the provided data, poses an additional level of difficulty on this specific task and paves the way for more sophisticated and holistic approaches to tackle these challenges.

We propose two distinct approaches to tackle the problem of Financial Narrative Summarization. Based on these, we implemented four systems that were tested and submitted to the shared task. All of the submitted systems leverage the fact that in the provided use case, the sentences that comprise the summary are usually extracted from the initial document in consecutive order. Therefore, we formulate the problem to identify the beginning of the summary in the document's corpus, following one abstractive and one extractive approach. In summary:

- **Sequence-to-sequence approach**: We use a pre-trained abstractive summarization model that is fine-tuned on the downstream task and aims to generate the start of the summaries.
- **Template-based approach**: We learn span representations from the financial documents in an unsupervised manner, and we apply similarity search on them to find suitable candidates for the summary start by building an index of summary templates.

The rest of the document is structured in the following manner: Section 2 presents the related work around text summarization and multilingual text representations. Section 3 describes the dataset used in this work. Section 4 presents the system created to deal this task. Section 5 presents the experiments and the results for each implemented model. Finally, Section 6 summarises the results and prompts for a discussion about future work and further applications.

2. Related Work

This section reviews the recent developments in Automated Text Summarization (ATS) and multilingual sentence embeddings and highlights the connection between our approach and the related literature.

2.1. Text summarization

There are two types of Automatic Text Summarization: Abstractive Summarization and Extractive Summarization. Automatic Text Summarization via the Extractive method constructs a summary by selecting the most pertinent sentences from the text and concatenating them. State-of-the-art extractive summarization methods use transformer based approaches modifying the

[*] Both authors contributed equally to this work.

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

BERT model (Liu and Lapata, 2019), proposing hierarchical encoder architectures (Liu and Lapata, 2019), as well as using summary-level representations (Zhong et al., 2020), leveraging the semantics of the entire summary. Automatic Text Summarization via the abstractive method consists of forming a summary inspired by human cognitive processes, understanding the text and writing a condensed version of it with minimal semantic loss. Important works around abstractive summarization involve the use of the encoder-decoder architecture for generating summaries in an auto-regressive manner (Liu and Lapata, 2019) and text generation (Lewis et al., 2019; Zhang et al., 2020). ATS has been applied in various use cases and domains with interesting academic work around news summarization (Sethi et al., 2017), biomedical document summarization (Azadani et al., 2018), legal document (Anand and Wagh, 2019) and scientific paper summarization (Alampalli Ramu et al., 2019). In this work, we use both extractive and abstractive summarization inspired by the literature, and we apply a custom filtering preprocessing procedure to the input data.

2.2. Multilingual Sentence Representations

Language models and transfer learning have become one of the cornerstones of natural language processing in recent years, especially in the context of machine translation and multilingual text representations. While some approaches were built for a single language or several languages separately, there is recent literature that demonstrates models trained on datasets that contain sentences from various languages, outperforming monolingual models in various multilingual benchmarks. Notable works propose methods to handle low-resource languages through zero-shot or few-shot cross-lingual transfer (Pfeiffer et al., 2020; Cao et al., 2020), as well as massive multilingual pretraining (Devlin et al., 2018; Xue et al., 2020) for both auto-encoder and auto-regressive models. Additionally, there has been a recent growing interest in using individual raw sentences for self-supervised constructive learning on top of pre-trained language models (Liu et al., 2021; Gao et al., 2021). In our case, we also use constructive learning in a multilingual setting to acquire multilingual representations of the summaries.

3. Dataset

The provided datasets included documents of financial reports along with a set of human-curated summaries. The corpora of the reports were extracted through Optical Character Recognition (OCR) from the original PDF documents. Each report in the English dataset had from three to seven respective gold summaries, whereas both the Greek and the Spanish reports had two respective gold summaries each. As presented in Table 1, the number of data samples for the English documents is far more than for the other two languages. This could pose a challenge when it comes to the fine-tuning of monolingual approaches. Motivated by this, instead of using a monolingual approach exclusively for each language, we also tested multilingual approaches on both the high resource language, English, and low resource languages, Greek and Spanish.

Language	**Split**	**Number of Documents**	
		Report	*Summary*
English	Train	3000	9873
	Validation	363	1250
	Test	500	1673
Greek	Train	162	324
	Validation	50	100
	Test	50	100
Spanish	Train	162	324
	Validation	50	100
	Test	50	100

Table 1: Datasets split sizes per language.

The lengths of the reports follow the same distribution in all languages with an average size of around 46500 tokens. However, the size of the gold summaries varies a lot among the three languages with the English and the Spanish datasets following the same distribution with a median size of around 775 tokens, and the Greek dataset around 1500 tokens.

Exploratory analysis was also made regarding the existence of the gold summary's sentences in the corresponding report as well as the position of the summary in the document. Based on these descriptive results, we found that for the English dataset, the summaries are extracted from the document in a continuous fashion. This signals that not only the summarization method was an extractive one, but also finding the start of the summary in the document and taking consecutive sentences someone can construct the gold summary with high accuracy performance. While this finding hints at an extractive way of formulating this text summarization problem, the rest of the datasets follow a different approach. For both the Greek and the Spanish cases, in more than 85% of the samples, the gold summary could not be found in the document which means that they are not a sequential subset of the reports and therefore an abstractive method might be more well suited. To tackle this heterogeneity of the datasets, we decided to apply different experiments that leverage the power of both sequence-to-sequence models as well as autoencoding ones and apply them to the task of identifying the start of the summary in the documents.

4. System

In this section, we present in detail the two approaches that the submitted systems are based on and explain the methodology behind them.

4.1. Sequence-to-sequence approach

In this abstractive summarization approach, we used a pre-trained sequence-to-sequence model, fine-tuned on the provided dataset on the task of start-of-summary prediction (Orzhenovskii, 2021). To achieve this, at the inference time, the model performs a similarity search between the output generated from the model and each sentence of the document. It then locates the span that is the closest match in terms of token similarity and selects it as the beginning of the summary. Having this selected start point in the document, it constructs the summary by taking the 1000 consecutive words. We submitted two flavours of this approach; one that utilizes the multilingual power of the used sequence-to-sequence model and keeps the input data in its original format, and one that translates the Greek and Spanish data into English, runs inference and then translates back the generated summaries to their original languages.

4.2. Template-based approach

We also propose an extractive summarization using an unsupervised summary generation method to find the best start candidate in a report to begin the summary with. The motivation behind this approach is the assumption that the start of the golden summaries can be clustered into different templates, and for each report, we want to start the generated summary with a block of tokens similar to the existing templates in our training dataset. We achieve this by mapping the span representations of the reports and the start block of the summaries in the same embedding space using BERT-like models. First, we compute the representations of the first 64 tokens of each summary, and we cluster them using the k-means algorithm. Next, we extract all 64-token blocks of each report with a 32-token window and compute their representations. For each report, we then find the block with the highest cosine similarity to all the clusters' centroids and consider it the beginning of the summary. Similarly, as with the sequence-to-sequence approach, having the selected span representation as the start of the summary, we take 1000 consecutive words (in addition to the start-of-summary span) and construct the predicted summary for the document. Once again, we submitted two variants of this method, having as input the original data format as well as their translations (for Greek and Spanish).

5. Experiments & Results

We first evaluate the proposed template-based approach with different models to get span representations. Then, we compare the best template-based model with the pre-trained abstractive summarization model fine-tuned on the given task.

Lang	Model	Rouge-1	Rouge-2	Rouge-L
EN	mBERT	**0.4059**	**0.2570**	**0.3875**
	mDeBERTa	0.4012	0.2544	0.3824
	MAD-X^{mBERT}	**0.4059**	0.2548	0.3872
	Mirror-mBERT	0.4016	0.2525	0.3827
GR	mBERT	**0.1337**	0.0363	0.1259
	mDeBERTa	0.1267	0.0348	0.1179
	MAD-X^{mBERT}	0.1335	**0.0373**	**0.1264**
	Mirror-mBERT	0.1320	0.0360	0.1247
ES	mBERT	**0.2354**	**0.0984**	**0.2068**
	mDeBERTa	0.2180	0.0838	0.1900
	MAD-X^{mBERT}	0.2317	0.0978	0.2030
	Mirror-mBERT	0.2300	0.0968	0.2017

Table 2: Rouge F1 scores on the validation sets between the constructed summaries and the golden summaries for each language dataset using different backbone models for the template-based approach. Scores in bold are the best model for each language.

5.1. Unsupervised summary generation using span representations

5.1.1. Zero-shot setting

For the unsupervised summary generation case, we first used mBERT (Devlin et al., 2018) and mDeBERTa (He et al., 2021) as baseline models for the span representations. These are transformer-based models, pretrained on a large corpus of multilingual data in a self-supervised fashion. Both of them are trained on the Mask Language Model objective, meaning that the model is asked to predict a masked token in a given text input. Consequently, they manage to learn bidirectional representations of the input sentences. We used these pre-trained models on all languages without any fine-tuning on the datasets. We used the average of the text's tokens output embeddings as span representations. For the k-means algorithm, we ran experiments with $k = 1, 3, 5, 10$ and reported the results with the highest validation Rouge F1 score, which were obtained with $k = 10$.

5.1.2. Fine-tuned setting

Additionally, we fined-tuned and tested two transformer-based models: Mirror-BERT (Liu et al., 2021) and MAD-X (Pfeiffer et al., 2020). Mirror-BERT leverages the contrastive learning technique and is trained on fully identical or slightly modified text span pairs as positive fine-tuning examples while maximizing their similarity during identity fine-tuning. In our experiments, we used the same implementation and training setup introduced by the authors of the Mirror-BERT paper[2] (Liu et al., 2021), using mBERT. MAD-X is an adapter-based framework that enables high parameter-efficient transfer to arbitrary tasks and languages by learning modular language and task representations. In our experiments, we used the

[2] https://github.com/cambridgeltl/mirror-bert

mBERT version of the MAD-X, and we fine-tuned a separate adapter for each language. We used the same training setup suggested by AdapterHub [3]. We used both of the models to get the span representations and then applied the method proposed in Section 4.

5.1.3. Experimental Results

Results for the unsupervised extractive summarization approach can be found in Table 2. There is no a significant difference in terms of Rouge scores between the different models to get the span representations. We can however notice that the scores for English are much higher than the two other languages. That could be explained by the limited number of samples provided for both Greek and Spanish. As noticed in Section 3, the distribution of the summaries in Greek is slightly different from the other two. Such a singularity could potentially explained why the performance in Greek are the lowest. As the performance obtained with mBERT are marginally better than with the other models, we decided to select that model for the shared task.

5.2. Unsupervised vs supervised learning

For the sequence-to-sequence modeling, we used the mT5 (Xue et al., 2020) model, which is a massively multilingual pre-trained text-to-text transformer model. mT5 is a multilingual extension of T5 (Raffel et al., 2020) that was pre-trained on a new Common Crawl-based dataset covering 101 (mC4). In our experiments, we used the same data preparation pipeline and training setup as (Orzhenovskii, 2021). The only difference is the maximum source length, which is set to 3900 due to limited GPU memory. In Table 3, we compare the performance of fine-tuned mT5 with the best unsupervised model, which is obtained with mBERT. The supervised approach significantly outperforms the unsupervised approach in English, but we can see that both approaches obtained similar performance in Greek and Spanish. Such a result is a good indicator that the unsupervised approach is a promising alternative to the computationally expensive sequence-to-sequence modeling approach when the number of training samples is quite limited. As the results with mT5 on English were however significantly better, we decided to submit this system to the shared task.

Further experiments were conducting following a slightly different approach by formulating the problem as a span classification task. In this case, we select a beginning span from the summaries and spans from documents that are not present in the respective summaries, and perform span classification on whether the span is a start-of-summary or not. For this classification task, we used a monolingual BERT-like model that is trained on the respective language of the dataset. Therefore, we used BERT (Devlin et al., 2018), Greek-BERT (Koutsikakis et al., 2020) and BETO (Canete et al., 2020) for the English, Greek and Spanish respectively. These approaches were not the final submission to the task since their performance was substantially inferior that the rest of the implemented systems.

Lang	Model	Rouge-1	Rouge-2	Rouge-L
EN	mBERT	0.4059	0.2570	0.3875
	mT5	**0.4402**	**0.3014**	**0.4236**
GR	mBERT	**0.1337**	0.0363	**0.1259**
	mT5	0.1336	**0.0367**	0.1258
ES	mBERT	**0.2354**	**0.0984**	**0.2068**
	mT5	0.2259	0.0921	0.1970

Table 3: Rouge F1 scores on the validation sets between the sequence-to-sequence approach (mT5) and the proposed unsupervised generation approach based on mBERT. Scores in bold are the best obtained for each language.

Lang	Model	Rouge-1	Rouge-2	Rouge-L
EN	mBERT	0.451	0.275	0.425
	mT5	**0.489**	**0.365**	**0.479**
GR	mBERT	0.315	0.130	0.238
	mT5	**0.346**	**0.141**	**0.267**
		English-based model with translation		
	mBERT	0.309	0.115	0.225
	mT5	0.309	0.115	0.224
ES	mBERT	0.454	0.138	0.168
	mT5	**0.466**	**0.157**	**0.238**
		English-based model with translation		
	mBERT	0.449	0.128	0.159
	mT5	0.443	0.131	0.167

Table 4: Rouge F1 scores on the test sets between the constructed summaries and the golden summaries for each language dataset. Scores in bold are the best obtained for each language.

5.3. Results on the Shared Task

On the test set from the shared task, the results reported in Table 4 show that the mT5 outperforms our proposed unsupervised approach in all the provided languages. Given the fact that mT5 is a supervised model trained on a massive multilingual dataset and later fine-tuned on the task's training dataset, its superiority was expected. However, our unsupervised approach shows a promising performance considering that the models in this approach do not have any understanding of the summarization task. Such a method can be a practical solution in a data-limited scenario. Additionally, as expected, using the translation of Greek and Spanish reports as the inputs is inferior to using the original form. This observation could be explained by the fact that translation from these languages to English and then from English back to them introduce a new error to the problem, especially since the extracted text from the pdf documents can be quite noisy. Also, as Greek and Spanish contributed to the pre-training phase of both

[3] https://github.com/Adapter-Hub/adapter-transformers

mBERT and mT5, it was expected for these models to perform better compared to the translated ones on the original data.

6. Conclusion & Future work

In this paper, we submitted an automated document summarization solution for multilingual financial reports. We proposed two approaches: one based on the multilingual sequence-to-sequence model mT5 and one using unsupervised summary generation by identifying the templates of the beginning of the summaries. Experiments have shown that overall, this task heavily relies on the way summarization is happened by the dataset curators and aims for dataset-dependent preprocessing mechanisms. The presented results also made apparent the trade-off between the monolingual and the multilingual approaches showing that in low resource datasets, it might be better to employ transfer learning from a pre-trained multilingual model that relies on fine-tuning.

A potential extension to our work is to formulate this setting as a multi-task problem and deploy a method that can be extended beyond the modeling of the beginning of the narrative section. Additionally, a challenge that remains to be tackled is to find a more efficient way to remove the OCR noise from the datasets at the preprocessing step. Moreover, an interesting application would be to use a hybrid model that can handle the extractive and abstractive fashion of the datasets. Lastly, it would be insightful to see experimental results on the Financial Narrative Summarization task with language model augmentation approaches that leverage both the entities and factuality of the input text.

7. Bibliographical References

Abdaljalil, S. and Bouamor, H. (2021). An exploration of automatic text summarization of financial reports. In *Proceedings of the Third Workshop on Financial Technology and Natural Language Processing*, pages 1–7.

Alampalli Ramu, N., Bandarupalli, M. S., Nekkanti, M. S. S., and Ramesh, G. (2019). Summarization of research publications using automatic extraction. In *International Conference on Intelligent Data Communication Technologies and Internet of Things*, pages 1–10. Springer.

Anand, D. and Wagh, R. (2019). Effective deep learning approaches for summarization of legal texts. *Journal of King Saud University-Computer and Information Sciences*.

Azadani, M. N., Ghadiri, N., and Davoodijam, E. (2018). Graph-based biomedical text summarization: An itemset mining and sentence clustering approach. *Journal of biomedical informatics*, 84:42–58.

Canete, J., Chaperon, G., Fuentes, R., Ho, J.-H., Kang, H., and Pérez, J. (2020). Spanish pre-trained bert model and evaluation data. *Pml4dc at iclr*, 2020:2020.

Cao, S., Kitaev, N., and Klein, D. (2020). Multilingual alignment of contextual word representations. *arXiv preprint arXiv:2002.03518*.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Gao, T., Yao, X., and Chen, D. (2021). Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821*.

He, P., Gao, J., and Chen, W. (2021). Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.

Koutsikakis, J., Chalkidis, I., Malakasiotis, P., and Androutsopoulos, I. (2020). Greek-bert: The greeks visiting sesame street. In *11th Hellenic Conference on Artificial Intelligence*, pages 110–117.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. (2019). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Liu, Y. and Lapata, M. (2019). Text summarization with pretrained encoders. *arXiv preprint arXiv:1908.08345*.

Liu, F., Vulić, I., Korhonen, A., and Collier, N. (2021). Fast, effective, and self-supervised: Transforming masked language models into universal lexical and sentence encoders. *arXiv preprint arXiv:2104.08027*.

Orzhenovskii, M. (2021). T5-long-extract at fns-2021 shared task. In *Proceedings of the 3rd Financial Narrative Processing Workshop*, pages 67–69.

Passali, T., Gidiotis, A., Chatzikyriakidis, E., and Tsoumakas, G. (2021). Towards human-centered summarization: A case study on financial news. In *Proceedings of the First Workshop on Bridging Human–Computer Interaction and Natural Language Processing*, pages 21–27.

Pfeiffer, J., Vulić, I., Gurevych, I., and Ruder, S. (2020). Mad-x: An adapter-based framework for multi-task cross-lingual transfer. *arXiv preprint arXiv:2005.00052*.

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.

Sethi, P., Sonawane, S., Khanwalker, S., and Keskar, R. (2017). Automatic text summarization of news articles. In *2017 International Conference on Big Data, IoT and Data Science (BID)*, pages 23–29. IEEE.

Suarez, J. B., 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.

Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., and Raffel, C. (2020). mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.

Zhang, J., Zhao, Y., Saleh, M., and Liu, P. (2020). Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pages 11328–11339. PMLR.

Zhong, M., Liu, P., Chen, Y., Wang, D., Qiu, X., and Huang, X. (2020). Extractive summarization as text matching. *arXiv preprint arXiv:2004.08795*.