AMEX AI-Labs: An Investigative Study on Extractive Summarization of Financial Documents

Piyush Arora and Priya Radhakrishnan
American Express AI Labs
Bangalore, India
{piyush.arora1,priya.radhakrishnan}@aexp.com

Abstract

We describe the work carried out by AMEX AI-LABS on an extractive summarization benchmark task focused on Financial Narratives Summarization (FNS). This task focuses on summarizing annual financial reports which pose two main challenges as compared to typical news document summarization tasks : i) annual reports are lengthier (average length about 80 pages) as compared to typical news documents, and ii) annual reports are more loosely structured e.g. comprising of tables, charts, textual data and images, which makes it challenging to effectively summarize. To address this summarization task we investigate a range of *unsupervised*, *supervised* and *ensemble* based techniques. We find that ensemble based techniques perform relatively better as compared to using only the unsupervised and supervised based techniques. Our ensemble based model achieved the highest rank of 9 out of 31 systems submitted for the benchmark task based on Rouge-L evaluation metric.

1 Introduction

The publicly available financial information has been rising rapidly with periodic updates, earnings reports etc. from companies and institutions. These financial reports not only increase in volume but also in terms of the diversity, comprising different structure and format, depending on the location of the respective institutes and companies (El-Haj et al., 2014). The rise in quantity and diversity of financial information, needs to be adequately processed, analyzed and summarized to make it easy to disseminate and consume by the end-users (El-Haj et al., 2019).

The financial narrative summarization task focuses on summarizing publicly available annual financial reports produced by UK firms listed on the London Stock Exchange (El-Haj et al., 2020). These reports are quite lengthy (average length about 80 pages), with some reports spanning more than 250 pages, which makes this task quite challenging. These reports broadly consist of two main sections :i) "*narrative sections*" also known as "*front-end sections*": part of the report which contain textual information and reviews by the firm's management and board of directors and ii) "*back-end sections*": sections containing financial statements in terms of tables and numbers. This task focuses on finding effective narratives sections and then performing the summarization over these narrative sections to generate a summary of about 1000 words. For more details about the task kindly refer (El-Haj et al., 2020).

We address this task by exploring the approaches which have shown to perform well for document summarization. We investigate page-rank based techniques to find effective summaries in an unsupervised fashion, which have shown to perform well in past (Mihalcea and Tarau, 2004; Erkan and Radev, 2004; Zheng and Lapata, 2019). Given the recent advancement in exploring deep learning techniques to extract effective summaries we explore a Bi-directional LSTM (Graves and Schmidhuber, 2005) based approach for extracting top sentence for a given document and generate summaries using the top-k ranked sentences. We combine complementary information captured using unsupervised and Bi-LSTM based supervised models along with Lead-k sentences approach (Lin and Hovy, 2002), to generate effective summaries (more details provided in Section 3).

Proceedings of the 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation, pages 137–142
Barcelona, Spain (Online), December 12, 2020.

The paper is organized as follows: Section 2 describes the dataset details and the evaluation mechanism used, Section 3 discusses the approaches explored in this work, Section 4 presents the result of our experiments, we describe our main findings and conclude in Section 5.

2 Dataset & Evaluation mechanism

The dataset provided by the task organizers consist of a training and a validation set consisting of 3000 and 363 financial documents respectively, as shown in Table 1. For each document we are provided with gold summaries which varies between two to seven for the training and validation set. The test set consists of 500 documents, we have to generate a document summary consisting of **1000** words for each of the documents in a test set. At most three different submissions were allowed for the task. Table 1 presents more detail such as number of sentences, average length etc., across training, validation and test set.

Dataset	**Documents**	**Gold Summaries**	**Average length of Documents**	**Average length of Summaries**
Training Set	3000	9873	1394 sentences	39 sentences
Validation Set	363	1250	1970 sentences	43 sentences
Test Set	500	-	2055 sentences	-

Table 1: Dataset statistics

The automatically generated summaries are evaluated against the gold summaries using the evaluation metric ROUGE (Lin, 2004), commonly used for summarization tasks. Rouge-1, Rouge-2, Rouge-L, Rouge-SU4 are the four variants of rouge which are used for evaluating the submitted systems on the test set. Rouge-L indicates longest common subsequence statistics, Rouge-1 indicates overlap of unigram, Rouge-2 indicates overlap of bigram and Rouge-SU indicates overlap of skip-bigram plus unigram statistics between the generated and the gold summary.

3 Methodology

We describe three types of approaches that we explored for addressing the financial narrative summarization task.

Unsupervised approach: Within the unsupervised approach we explored three methods namely Lead-k (Lin and Hovy, 2002), TextRank (Mihalcea and Tarau, 2004), and LexRank (Erkan and Radev, 2004) for performing extractive summarization.

Lead-k: In general, news documents tend to contain the most informative content at the top and are further detailed. In this method the top k sentences from a document, based on the order of their occurrence are extracted and combined to form a summary. This method has proved to be quite robust and have been used as a good baseline for the task of extractive summarization (Lin and Hovy, 2002; Lewis et al., 2019; Zheng and Lapata, 2019).

TextRank and *LexRank*: The initial work on PageRank (Page et al., 1999) for crawling web pages, motivated the idea of TextRank and LexRank, which are graph based approaches for performing extractive summarization of documents. In graph based approaches all the sentences in a document are represented as the vertices of a graph where the edges are vertex (sentence) to vertex (sentence) similarity, and hence weighted. The weights of the edges are calculated using the textual similarity between the sentences. Top k salient sentences are combined to form a summary. For a given document, a connected graph of the sentences in constructed and then the salient sentences from the graph are extracted as shown in Equation 1 and Equation 2. S_i and S_j are two sentences comprising of words $w_i, w_{i+1}.., w_n$ in Equation 1. In Equation 2, $WS(V_i)$ represent a weighted score for a Vertex i, $In(V_i)$ and $Out(V_i)$ indicates in-degree and out-degree scores respectively, and d is a damping factor having a value between 0 and 1.

TextRank approach looks at the absolute number of words two sentences have in common, which are then normalized by the sentences length. Whereas, LexRank calculates cosine similarity of the word vectors for both the sentences as shown in Equation 3. Each sentence is represented as a N dimensional

vector using a bag-of-words based model, where N represents the number of unique words in the document. In Equation 3, x and y represents the N dimensional sentence vector for Sentences S_i and S_j respectively, and tf indicates the term frequency and idf indicates the inverse document frequency for a word w.

$$Similarity(S_i, S_j) = \frac{|\{w_k | w_k \in S_i \& w_k \in S_j\}|}{log(|S_i|) + log(|S_j|)} \tag{1}$$

$$WS(V_i) = (1-d) + d * \sum_{V_j \in In(V_i)} \frac{w_{ji}}{\sum_{V_k \in Out(V_j)} w_{jk}} WS(V_j) \tag{2}$$

$$Idf - modified - cosine(x, y) = \frac{\sum_{w \in x,y} tf_{w,x} tf_{w,y} (idf_w)^2}{\sqrt{\sum_{x_i \in x} (tf_{x_i,x} idf_{x_i})^2} \times \sqrt{\sum_{y_i \in y} (tf_{y_i,y} idf_{y_i})^2}} \tag{3}$$

We used the sumy tool[1] for performing LexRank based summarization. We used the gensim (Řehřek and Sojka, 2011) implementation of TextRank algorithm from (Barrios et al., 2015). This revised version of TextRank performs better as compared to the original version proposed in (Mihalcea and Tarau, 2004). For more details on these models, we advise readers to refer (Barrios et al., 2015; Mihalcea and Tarau, 2004; Erkan and Radev, 2004).

Supervised approach: Extractive summarization focuses on finding good representative sentences from a document to represent document summaries. Thus the summarization task can be approached as a binary level sentence classification task where sentences which form a part of document summaries are positive samples and rest being categorized as negative samples. Thus given a new document we have to classify sentences as positive and negative instances. Top k positive sentences in their chronological occurrence are considered as document summaries. We experimented with a Bi-directional LSTM (Bi-LSTM) based approach for sentence extraction for generating document summaries by combining top-k ranked sentences. We used the training dataset which has about 350k positive samples and we randomly sample a similar number of negative samples from the training set. Similarly a validation dataset was created which has about 100k sentences. Both the training and validation set have equal numbers of positive and negative samples.

We use the PyTorch library (Paszke et al., 2017) for Bi-LSTM based method, with following parameters: number of hidden units set to 48, loss function used is binary cross entropy loss, we used fasttext textual embeddings (Joulin et al., 2016), dimensionality = 300 to represent the textual input. We perform dropout to avoid overfitting of the model.

Ensemble approach: We find quite promising results using Lead-k, TextRank and LexRank based unsupervised approaches and a Bi-LSTM based supervised approach on the training and validation set (described later in Section 4). As described above that the Lead-k sentences based summaries has been used as a strong baseline for the summarization task so we performed an ensemble approach where we combine the output of i) TextRank, ii) LexRank, iii) Lead-k and iv) Bi-LSTM approaches. For a given document we sort and rank the sentences using these four approaches in descending order, we take the top 20 sentences from each of these approaches and combine the output using an ensemble technique. We chose 20 sentences empirically, as the length of summary is constrained to 1000 words only, so all the sentences occurring lower in the order do not impact the Rouge scores. This ensemble model is a linear combination of the output of multiple approaches, where the sentences that have occurred in the output of most approaches are ranked higher. In case of the same frequency of the sentences, preferences are given to the following models' output: Bi-LSTM $>$ TextRank $>$ LexRank $>$ Lead-k, determined empirically based on the training set. A detailed worked out example of the ensemble approach is shown in Figure 1.

We used Stanford CoreNLP pipeline (Manning et al., 2014) to perform sentence splitting for each document before applying all three approaches described above. A post-processing step is performed after

[1] https://pypi.org/project/sumy/

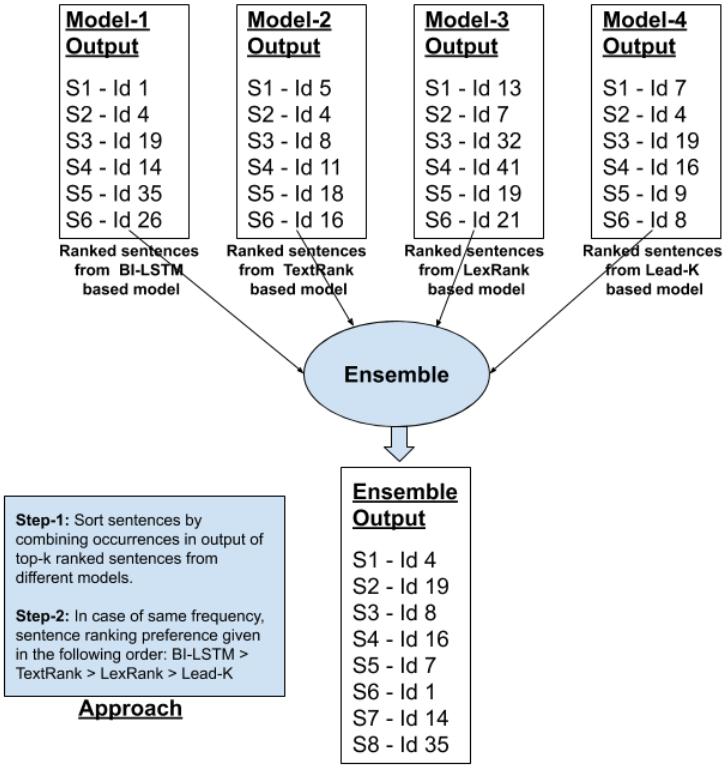

Figure 1: Ensemble Approach

generating document summaries from these three methods (Unsupervised, Supervised and Ensemble), where all the textual information beyond 1000 words is removed, as per the task requirement.

4 Results & Analysis

Results of different unsupervised approaches such as TextRank, Lead-k are shown in Table 2. We found that the TextRank based approach performs relatively better than LexRank and Lead-5 approach for both training as well as the validation set. We compare TextRank with Lead-5 as the average length of sentences for the output of TextRank is 4.6. We select TextRank as one of our system submissions on the test set which is referred to as $AMEX - TextRank$.

	ROUGE-1			**ROUGE-2**		
Training set	Recall	Precis	F-Meas	Recall	Precis	F-Meas
Lead-5	0.156	0.474	0.213	0.073	0.203	0.097
LexRank	0.310	0.269	0.259	0.098	0.088	0.083
TextRank	0.303	0.282	**0.277**	0.100	0.094	**0.092**
Validation set	Recall	Precis	F-Meas	Recall	Precis	F-Meas
Lead-5	0.148	0.516	0.211	0.065	0.219	0.092
LexRank	0.436	0.247	0.293	0.141	0.08	0.094
TextRank	0.308	0.330	**0.304**	0.099	0.107	**0.098**

Table 2: Results of unsupervised models on training and validation set, the best scores are in bold face

Our second system submission is a Bi-LSTM based binary classification model referred to as $AMEX - BILSTM$. Our final submission is an ensemble model referred to as $AMEX - Ensemble$. Table 3 presents the result of our three submissions on the test set. We find that the ensemble approach

Description	R-L / R	R-L / P	R-L / F	Rank	R-1 / R	R-1 / P	R-1 / F	Rank
Top System	0.6050	0.3764	0.4556	1	0.6123	0.3934	0.4663	1
Top Baseline	0.4696	0.3704	0.4067	3	0.4827	0.4135	0.4329	10
AMEX-Ensemble	**0.4077**	**0.3652**	**0.3778**	**9**	**0.4417**	**0.4080**	**0.4125**	**16**
AMEX-BILSTM	0.4017	0.3601	0.3718	10	0.4361	0.4063	0.4088	17
AMEX-TextRank	0.2461	0.2448	0.2370	23	0.3533	0.2715	0.2950	26
Description	**R-2 / R**	**R-2 / P**	**R-2 / F**	**Rank**	**R-SU4 / R**	**R-SU4 / P**	**R-SU4 / F**	**Rank**
Top System	0.3655	0.2951	0.3060	1	0.3440	0.3324	0.3184	1
Top Baseline	0.3108	0.1978	0.2338	17	0.3748	0.2008	0.2530	11
AMEX-Ensemble	**0.2639**	**0.1919**	**0.2142**	**20**	**0.3285**	**0.1941**	**0.2352**	**20**
AMEX-BILSTM	0.2603	0.1897	0.2111	21	0.3247	0.1921	0.2323	21
AMEX-TextRank	0.1838	0.0967	0.1200	26	0.2499	0.1079	0.1438	28

Table 3: Results of our three submitted models, our best scoring system is in boldface. R-1, R-2, R-L, R-SU4 indicates rouge scores at unigram, bigram, longest common subsequence and skip-gram based metrics respectively, and P,R,F indicates precision, recall and f-scores respectively. Top baseline model is a genetic algorithm based approach for summarization (Litvak et al., 2010)

.

performs relatively better than the TextRank and Bi-LSTM based approach. Our ensemble based model achieved the highest rank of 9 out of 31 systems submitted for the benchmark task based on Rouge-L evaluation metric.

Overall ranking of the systems varies depending on the evaluation metric being considered. Based on Rouge-L and Rouge-1 metrics our system submissions achieved quite high precision but scored lower on recall values hence the F-score are relatively averaged. In our ensemble based approach, we combined the output of multiple approaches by a linear combination, in future we will like to explore learning effective weights while combining the output of multiple models, to generate effective summaries which can possibly address the problem of low recall, by including potential good candidates occurring lower in the order.

5 Conclusion & Future work

We present our work on an initial investigation of extractive summarization for annual financial reports. We explored alternative approaches using unsupervised, supervised and an ensemble based method. We find that Bi-LSTM based supervised approaches perform relatively better than using unsupervised based approaches such as TextRank. The ensemble based model performs best as compared to supervised and unsupervised models and obtained a rank of 9/31 on the test set using $Rouge - L$ evaluation metric.

We used a Bi-LSTM based model for extracting good representative sentences to be included in the document summary. However we lose the document level information while treating the problem of extracting good representative sentences as a binary class classification problem while training the model. In future we will like to explore recent models (Lewis et al., 2019; Zheng and Lapata, 2019) for performing extractive summarization. These models leverage BERT based distributed representation (Devlin et al., 2019) and train summarization models by optimizing the Rouge scores, to generate effective document summaries.

Acknowledgements

We would like to thank the task organizers for organizing this interesting benchmark task and the reviewers for their valuable feedback. We would like to thank our colleague Arpan Somani for helping in running initial experiments using the Bi-LSTM based model, and Salil Rajeev Joshi, Himanshu Sharad Bhatt & Shourya Roy for their guidance and suggestions.

References

Federico Barrios, Federico López, Luis Argerich, and Rosita Wachenchauzer. 2015. Variations of the similarity function of textrank for automated summarization. In *Argentine Symposium on Artificial Intelligence (ASAI 2015)-JAIIO 44 (Rosario, 2015)*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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.

Mahmoud El-Haj, Paul Rayson, Steve Young, and Martin Walker. 2014. Detecting document structure in a very large corpus of uk financial reports. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1335–1338.

Mahmoud El-Haj, Paul Rayson, Paulo Alves, Carlos Herrero-Zorita, and Steven Young. 2019. Multilingual financial narrative processing: Analysing annual reports in english, spanish and portuguese. *Multilingual Text Analysis: Challenges, Models, And Approaches*, page 441.

Mahmoud El-Haj, Ahmed AbuRa'ed, Marina Litvak, Nikiforos Pittaras, and George Giannakopoulos. 2020. The Financial Narrative Summarisation Shared Task (FNS 2020). In *The 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation (FNP-FNS 2020)*, Barcelona, Spain.

Günes Erkan and Dragomir R Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, 22:457–479.

Alex Graves and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional lstm networks. In *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.*, volume 4, pages 2047–2052. IEEE.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016. Fasttext. zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Chin-Yew Lin and Eduard Hovy. 2002. From single to multi-document summarization. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 457–464.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81.

Marina Litvak, Mark Last, and Menahem Friedman. 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.

Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. 2014. The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60.

Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 404–411.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in pytorch.

Radim Řehřek and Petr Sojka. 2011. Gensim—statistical semantics in python. *Retrieved from genism. org*.

Hao Zheng and Mirella Lapata. 2019. Sentence centrality revisited for unsupervised summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6236–6247.