Extractive Financial Narrative Summarisation based on DPPs

Lei Li, Yafei Jiang, Yinan Liu
Beijing University of Posts and Telecommunications (BUPT)
No.10 Xitucheng Road, Haidian District, Beijing, P.R.China
leili,jiangyafei,lyinan@bupt.edu.cn

Abstract

We participate in the FNS-Summarisation 2020 shared task to be held at FNP 2020 workshop at COLING 2020. Based on Determinantal Point Processes (DPPs), we build an extractive automatic financial summarisation system for the specific task. In this system, we first analyze the long report data to select the important narrative parts and generate an intermediate document. Next, we build the kernel Matrix L for the intermediate document, which represents the quality of its sentences. On the basis of L, we then can use the DPPs sampling algorithm to choose those sentences with high quality and diversity as the final summary sentences.

1 Introduction

With the development and progress of the economy, the popularity of the financial sector is increasing day by day. There are quite a few new companies gradually emerging and going public, but investors often find it difficult to deal with the long-form annual reports of various companies, because their content may be tedious and redundant, and it is difficult to filter out effective key information by human resources, so an automatic digest system is needed to help investors effectively screen company information.

In this paper, we try to solve the first task contained in FNS (Financial Narrative Summarisation) 2020, the dataset we use is the annual reports in the financial field provided by the organizer. As there is much redundant information in the reports, we plan to select the most useful parts first as the intermediate documents to be summarized. For the shared task of summarisation, we implement a Quality-Diversity model (QD) based on Determinantal Point Processes (DPPs) to represent the intermediate document, and merge three kinds of features to rank the sentences.

2 Related Work

Li L et al. (2018) focused on exploring the sampling process. They used WMD sentence similarity to construct new kernel matrix used in Determinantal Point Processes (DPPs). Ma S et al. (2018) divided all sentences into three categories (motivations, methods, and conclusions), and then extracted sentences from each cluster based on rules and severe features to form a summary. Debnath D et al. (2018) built a summary generation system using the OpenNMT tool. Zhong M et al. (2019) analyzed the relationship between the quality of extractive automatic summarization and the model network, and the influence of network architecture, knowledge transfer and learning mode on the effect of the summary system is verified through a series of experiments. Liu Y and Lapata M (2019) studied the influence of pre-training language models in automatic summarization tasks, and emphasized the importance of document encoding. In the extractive summarization task, they used multiple layers of BERT as the encoder, and obtained good results on three datasets.

3 Data

For the FNS 2020 Shared task we use 3863 UK annual reports for firms listed on The London Stock Exchange (LSE) covering the period between 2002 and 2017. UK annual reports are lengthy documents

Proceedings of the 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation, pages 100–104
Barcelona, Spain (Online), December 12, 2020.

with around 80 pages on average, some annual reports could span over more than 250 pages. These UK annual reports are divided into training, testing and validation sets. The training and validation sets include the full text of each annual report along with the gold-standard summaries. On average there are at least 3 gold-standard summaries for each annual report with some reports containing up to 7 gold-standard summaries. For testing dataset, the task participants were given access only to the full texts. Table 1 shows the dataset details.

Data Type	Training	Validation	Testing
Report full text	3000	363	500
Gold summaries	9873	1250	-

Table 1: FNS 2020 Shared Task Dataset

4 System

In our system, we present an original Quality-Diversity model for extractive automatic summarisation based on DPPs sampling algorithm (Kulesza and Taskar, 2012). In this model, a single input document can be regarded as a set of sentences, and the process of extracting abstracts can be regarded as sampling a subset of important sentences from that set. To choose the high-quality and diversity summaries with DPPs, we need to construct the kernel matrix L to represent the document. The main steps of summary generation include pre-processing, feature selection and sentence sampling. Pre-processing, feature selection and sentence sampling are as followed.

4.1 Pre-processing

In this task the summary requires extraction from different key sections found in the annual reports. Those sections are usually referred to as “narrative sections” or “front-end” sections and they usually contain textual information and reviews by the firm’s management and board of directors. Sections containing financial statements in terms of tables and numbers are usually referred to as “back-end” sections and are not supposed to be part of the narrative summaries. Therefore, in the data pre-processing part, there are mainly two steps: the first step is to detect and extract the narrative sections of the annual reports; the second step is to organize the format of the dataset, and clean the narrative sections extracted in the previous step. The final processed data represents the important narrative parts and will be used for subsequent summary generation/extraction.More implementation details of pre-processing are given below.

First, we analyze the dataset and try to catch some underlying laws. Comparing the annual reports and gold summaries, we find that each gold summary contains some sentences appearing in the corresponding annual report, and each corresponds to a continuous piece of content. Upon further analysis, we can discover that the gold summary basically corresponds to four sections in the annual report, namely Highlights, At a glance, Chairman’s statement and Chief Executive’s review. Although not every annual report contains these four sections and the names of the sections may be different, the contents of these sections are similar and are all narrative sections. Therefore, we extract all these sections included in each annual report and use them as a new dataset together with the gold summaries.

Next, we perform further data cleaning and formatting on the new dataset obtained in the previous step. The dataset provided by the organizer is directly converted from pdf files to txt files, thus the data format is confusing and has many useless content such as headers and footers. The reason why they are considered useless is that headers and footers information of each page is basically the same, such as ANNUAL REPORT AND ACCOUNTS XXX, where XXX represents the year. In order to store sentences by lines, we divide the text according to punctuation marks such as periods, exclamation points, etc. In order to delete non-narrative texts such as tables and numbers, we first use regular expressions to filter financial numbers, and then use the open source NLP toolkit to perform dependency syntactic analysis on the data. If a sentence does not contain the subject, predicate and object, We consider the sentence to be unqualified and delete it. Finally, we use the processed data to generate intermediate documents for the subsequent summarisation instead of the original long annual reports.

4.2 Feature Selection

To represent the document, we build a kernel matrix L from statistical feature method. We use Sentence Length (SL), Sentence Position (SP), Sentence Coverage (SC) as features according to the work of (Li L et al., 2017) to construct matrix L. Sentence Length can be calculated by

$$sl_i = \exp\left(-1 * \frac{(len_i - \mu)^2}{\sigma^2}\right) \tag{1}$$

where sl_i is the sentence length feature score, μ and σ are the mean and standard deviation of sentence length respectively. Sentence Position can be calculated by

$$sp_i = 1 - \frac{s_i}{|D|} \tag{2}$$

where s_i represents the position of the $i - th$ sentence in the text, $|D|$ represents the total number of sentences in the text, and sp_i is the score of the sentence position feature. Sentence Coverage can be calculated by

$$SC_i = \frac{\Sigma_{i=1}^{|S|} \frac{num_S(word_i)}{n}}{|S|} \tag{3}$$

where sc_i represents the score of sentence coverage, $word_i$ represents the $i - th$ word in sentence S, $num_s\ (word_i)$ represents the total number of sentences covered by $word_i$, and $|S|$ represents the total number of words of in the sentence, n represents the total number of sentences in the input sequence. Elements of L can be calculated by

$$L_{ij} = q_i Sim_{ij} q_j \tag{4}$$

where q_i is the quality of a single sentence, Sim_{ij} represents the similarity between sentences. q_i can be calculated by the features of SL, SP and SC. We use the Jaccard similarity between sentences to measure the degree of sentence diversity, so Sim_{ij} can be computed as

$$Sim_{ij} = \frac{|\ \{word|word \in sen_i\ and\ word \in sen_j\}\ |}{|\ \{word|word \in sen_i\ or\ word \in sen_j\}\ |} \tag{5}$$

It can be seen that the matrix L has the function of measuring the quality and diversity of sentences, which is quite important in further sampling process. For the shared task, we select one feature or the sum of multiple features to get the sentence quality q and arrange these methods into three runs.

4.3 Sentence Sampling

In our approach, we use discrete DPPs to select sentences, by constructing matrix L, we can apply the DPPs sampling algorithm to extract summaries. To perform DPPs sampling, we first get the eigenvalue λ_n and the eigenvector v_n of the matrix L, then project all the sentence vectors into a new low-dimensional feature space. In this space, the magnitude of the vector describes the importance and the cosine similarity between two vectors describes the similarity between sentences. When selecting sentences, the number of sentences in the summary is obtained by probability model.

Next, we select a sentence with high quality x_i, then remove a list of feature vectors v_i in the feature space that contributes to its vector modulus length, and perform series of orthogonalization on feature space.

Finally we select new elements according to the new sub-feature space re-projection until the end, which can ensure that the element selected again are both high-quality and low-similar to the previous element, thus choose those high-quality and diversity sentences as summary sentences. the details of DPPs can be referred to the work of Kulesza and Taskar(2012).

5 Results

In the shared task, we submitted a total of three versions of our system. CIST-BUPT-RUN3 uses SC as the feature that represents document, CIST-BUPT-RUN2 uses SL, while CIST-BUPT-RUN1 merges all three features: SC, SL and SP. We tried three versions on the pre-processed training set and table 2 shows the performance from our experiments.

System	ROUGE-1/F	ROUGE-2/F	ROUGE-L/F
CIST-BUPT-RUN1	0.305	0.073	0.166
CIST-BUPT-RUN2	0.293	0.067	0.164
CIST-BUPT-RUN3	**0.313**	**0.079**	**0.167**

Table 2: Results on Official Validation Set

The result shows that the systems using different features have similar performance, while CIST-BUPT-RUN3 performs the best and the performance of others are slightly lower.

System	R-L/R	R-L/P	R-L/F	R-1/R	R-1/P	R-1/F
CIST-BUPT-RUN1	0.294	**0.361**	0.317	0.405	0.423	0.401
CIST-BUPT-RUN2	0.311	0.352	0.324	0.418	0.440	0.416
CIST-BUPT-RUN3	**0.324**	0.348	**0.329**	**0.434**	**0.449**	**0.428**
	R-2/R	**R-2/P**	**R-2/F**	**R-SU4/R**	**R-SU4/P**	**R-SU4/F**
CIST-BUPT-RUN1	0.258	0.206	0.220	0.315	0.190	0.228
CIST-BUPT-RUN2	**0.272**	0.224	0.237	0.330	0.204	0.243
CIST-BUPT-RUN3	0.228	**0.233**	**0.248**	**0.346**	**0.209**	**0.251**

Table 3: Evaluation Results on Testing Set

As is shown in table 3, CIST-BUPT-RUN3 still performs the best. We find that the ROUGE score on testing set is higher than that on training set, the reason may be that because for each document in the official training set, there are multiple golden abstracts corresponding to them, we choose to merge these summaries for evaluation, so the ROUGE score is a bit lower. But even so, our experimental results can reflect the pros and cons of the three models. The model of DPPs with SC as feature for the kernel matrix shows the best performance among the three models in our system. We can also infer that SC catches more relations between sentences and thus is superior to the features of sentence's length and position.

6 Conclusion and Future Work

Actually, there are multiple ways to represent the document, one typical way is to train a Sentence2Vec model using neural network, which may improve the performance of the model. But because of the COVID-19, the servers available are limited, we don't have enough computing resources to train the model. In the future, we'll try other feasible methods of document representation, seeing if they can help improve the quality of summaries.

Acknowledgements

This work was supported by Beijing Municipal Commission of Science and Technology [grant number Z181100001018035]; National Social Science Foundation of China [grant number 16ZDA055]; National Natural Science Foundation of China [grant numbers 91546121, 71231002]; Engineering Research Center of Information Networks, Ministry of Education; Beijing BUPT Information Networks Industry Institute Company Limited; the project of Beijing Institute of Science and Technology Information.

References

Li L, Chi J, Chen M, et al. 2018. CIST@ CLSciSumm-18: Methods for Computational Linguistics Scientific Citation Linkage, Facet Classification and Summarization[C]//BIRNDL@ SIGIR. 2018: 84-95.

Ma S, Zhang H, Xu J, et al. 2018. NJUST@ CLSciSumm-18[C]//BIRNDL@ SIGIR. 2018: 114-129.

Debnath D, Achom A, Pakray P. 2018. NLP-NITMZ@ CLScisumm-18 In: Proceedings of the 3nd Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries[C]//BIRNDL@ SIGIR. 2018: 164- 171.

Zhong M, Liu P, Wang D et al. 2019. Searching for Effective Neural Extractive Summarization: What Works and What's Next. arXiv preprint arXiv:1907.03491.

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

Alex Kulesza and Ben Taskar 2012. Determinantal Point Processes for Machine Learning, Foundations and Trends in Machine Learning: Vol. 5: No. 23, pp 123- 286. http://dx.doi.org/10.1561/2200000044.

Li L, Zhang Y, Chi J et al. 2017. UIDS: A Multilingual Document Summarization Frame- work Based on Summary Diversity and Hierarchical Topics [M] // Li L, Zhang Y, Chi J et al. Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data. Springer, 2017: 2017: 343-354.