Prospectus Summarization Using BertSum

Zhonghao Guo, Lin Zhu, Yueying Yuan, Chenyang Xu, Yuguo Zheng, Mengning Li
The University of Hong Kong

Abstract

We introduce a NLP model called **BertSum** to summarize IPO prospectus. The prospectus provides information on business model, competition, risks and opportunities, and financial situations. However, it is usually a very long legal document. Therefore, in order to help generate a big picture of the company, we use BertSum to further condense summary part. It helps you get a quick overview of the company and help you to decide if you want to look at any detailed section that you are interested in. This would be helpful when you are not familiar with the company or the industry that it operates. [1]

1 Introduction

In order to get the input that is readable in the BertSum model, we need first to go through data collection and preprocessing procedure. The input required for BertSum is in txt format, but the prospectus is available through the Edgar system on the SEC website.

1.1 Data collection

We use three web scraping packages to get all the information from the website. Edgar is a package that specializes in data collection from the Edgar system. It requires two inputs: the company name and the CIK ID. It will give you the link of the index page. We use the regular expression and get the link of the prospectus. The next step is to delete tables in this webpage. We observe that there are two types of formats when it comes to the tables in the prospectus. After removing all the tables, we get a webpage that is purely text now.

1.2 Data Preprocess

After we get the prospectus from web-scraping, we need to do is to find the prospectus summary. Because BertSum model deal with the whole prospectus might take a lot of time. So we just put the prospectus summary into the model. We use regular expression to pick up the content between two titles called PROSPECTUS SUMMARY and RISK FACTORS, which are top 2 titles of the whole prospectus. And the first element of this summary list is prospectus summary. We delete all blank lines and *Table of Contents* and we use regular expression to find all of the page numbers and replace them to space. The subtitles will affect the performance of BertSum model definitely, so we also delete them. Finally, if some paragraphs are longer than 512 tokens, we will separate it into two paragraphs. Also, we add [CLS] [SEP] tokens to every end of sentences.

1. You can download code, data and model from our github repository
https://github.com/guozhonghao1994/Prospectus_Summarization_Using_BertSum

2 Related Work

Let's talk a little bit more about the model we used. We tried 2 other summarization models, *TextTeaser* and *TextRank*, to see how they work and why does BertSum, as a new model, perform better than the classic ones.

2.1 TextRank

TextRank is a graph-based ranking model for text processing. Graph-based ranking algorithm was invented in the end of last century, and the idea of TextRank is from Google's PageRank (1998), replacing the website pages with sentences in the text. In this model, the text is first tokenized with dropping stop words to get the keywords. The next step is evaluation, this model would generate a matrix to record the co-occurrence relationship of the keywords, and then generate a keyword graph and calculate the weight of different keywords. After that, the model would use the keyword weight to find the sentences with the highest ranks and use them as the summarization of the text. This model is widely used because of its convenience. It does not require using multiple articles to do the pre-training before finally employing it to do the extraction. And this can also be a limit of its reliability.

2.2 TextTeaser

TextTeaser was an automatic summarization service for long articles online when it was created by a developer named Jolo Balbin. Its core mechanism is also sentence ranking. Its ranking comes from 4 aspects: length, setting an ideal length and giving ranking by it; position, giving different rankings according to the position of the sentence in the text; relationship with the title, to see if the sentence contains information in the title; and number of keywords, picking out keywords of the text by occurrence frequency, and give the sentence ranking by number of keywords contained. Sentences with top rankings are chosen as the summary. Machine learning is involved in the model, as all the text processed by this model would be recorded and saved in the server to enhance the model itself in future processing, adjusting the length, position ranking.

2.3 BertSum

We all know that text summarization can be divided into two categories, extractive and abstractive. For both TextTeaser and TextRank they are only able to do the extractive summarization using existing sentences in the text, but BertSum is able to deal with abstractive summarization. The second is both these two models use sentence as the unit of summarization, it can be rigid when dealing with text in some specific area. For BertSum its sample unit can be adjusted according to different summarization needs. The third is that they are not able to deal with polysemy problems. When a word with multiple different meanings occurs, the old models can't process them correctly. Finally, BertSum has a lot of training series for different text types, for example news, literature, etc. Users are able to switch suitable training sets for their needs.

3 Experiments

In this section, we present how to use BertSum to generate prospectus summary.

3.1 Environment Configuration

There are two packages that should be installed.

- PyTorch: PyTorch is an open source deep learning library based on the Torch library. It is developed by Facebook.
- pytorch_transformers: This is a library of

pre-trained models for NLP. The library currently contains PyTorch implementations, pre-trained model weights, usage scripts and conversion utilities for several models.

3.2 CNN/DailyMail Dataset

It contains news articles and associated highlights. The data is divided into 3 parts, which are training, validation, and testing. We should have prepared the data using the following 4 steps:

1. First, Download raw CNN/DM data from https://cs.nyu.edu/~kcho/DMQA/
2. Second, split and tokenize Sentence using Stanford CoreNLP
3. Third, transfer the format to simpler json files
4. Last, transfer format to PyTorch Files

Fortunately, there is a preprocessed version textual data that can be downloaded directly.

3.3 Load Pre-trained Model

Training process is both time and money consuming. For the extraction model, the author trained for 50,000 steps on 3 GPUs (GTX 1080 Ti). Model checkpoints were saved and evaluated on the validation set every 1,000 steps. For the abstractive model, all models were trained for 200,000 steps on 4 GPUs (GTX 1080 Ti). Model checkpoints were saved and evaluated on the validation set every 2,500 steps. Again, we already have 2 fully-trained CNN/DM Extractive & Abstractive models. So we can just download and use them.

3.4 Generate Summary

We put the prospectus text file into a folder and run the python command to generate the summary. Here are some parameters that need to be specified:

```
-task {ext;abs}
-mode {train;validate;test;test_text}
-test_from MODEL_PATH
-text_src RAW_TEXT_PATH
-result_path OUTPUT_PATH
-log_file TERMINAL_OUTPUT
-visible_gpus {-1;0;1…}
```

4 Results

In this section, we will show you good results and unsatisfactory results on prospectus.

4.1 Good Results

1. BertSum is able to capture the most useful and important information of a company in the prospectus. First let's look at an example.


We are the largest online and mobile commerce company in the world in terms of gross merchandise volume in 2013, according to the IDC GMV Report. We operate our ecosystem as a platform for third parties, and we do not engage in direct sales, compete with our merchants or hold inventory. As a platform, we provide the fundamental technology infrastructure and marketing reach to help businesses leverage the power of the Internet to establish an online presence and conduct commerce with consumers and businesses. We have been a leader in developing online marketplace standards in China. Given the scale we have been able to achieve,

Figure1: a long paragraph in Alibaba's prospectus where it introduces its business and ecosystem in the market

Summarized text: *Our business is the largest online and mobile commerce company in the world in terms of gross merchandise volume in 2013. We operate our ecosystem as a platform for third parties, and we do not engage in direct sales. As a platform, we provide the fundamental technology infrastructure and marketing reach to help businesses leverage the power of the internet to establish an online presence.* This summarization tells us the major business of Alibaba. It also points out Alibaba's leading competitive position in its industry around the world based on the

gross merchandise volume. Therefore, this summarization gives us a quick idea about what Alibaba does and its current position in the industry.

2. BertSum is able to remove the redundant attributes without changing the main idea of the original text. Here is another example.

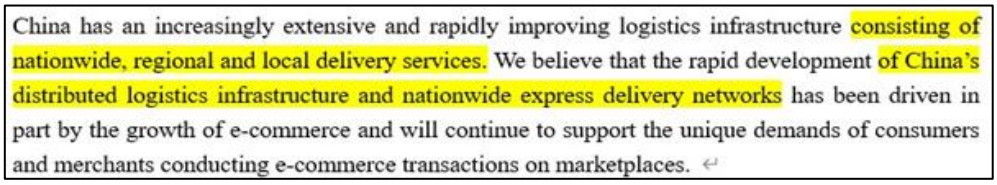
China has an increasingly extensive and rapidly improving logistics infrastructure consisting of nationwide, regional and local delivery services. We believe that the rapid development of China's distributed logistics infrastructure and nationwide express delivery networks has been driven in part by the growth of e-commerce and will continue to support the unique demands of consumers and merchants conducting e-commerce transactions on marketplaces.

Figure2: a paragraph where it talks about the development of China's logistics infrastructure

Summarized text: *China has an increasingly extensive and rapidly improving logistics infrastructure. The rapid development has been driven in part by the growth of e-commerce.* In this part, the attributes highlighted in yellow have been removed, but the main idea of the original text doesn't change. Specifically, for the second sentence, BertSum recognizes that the "rapid development" here mainly refers to the logistics infrastructure by looking at its attributive clause afterwards, and since the term "logistics infrastructure" has already appeared in the first sentence.

3. BertSum is able to rephrase some sentences based on its understanding of the original text.

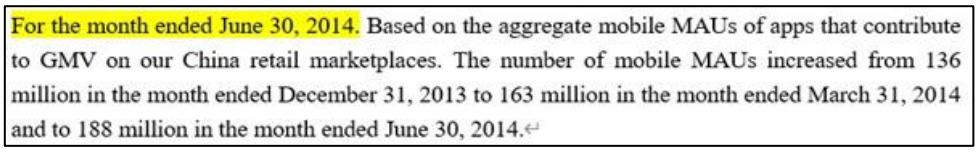
For the month ended June 30, 2014. Based on the aggregate mobile MAUs of apps that contribute to GMV on our China retail marketplaces. The number of mobile MAUs increased from 136 million in the month ended December 31, 2013 to 163 million in the month ended March 31, 2014 and to 188 million in the month ended June 30, 2014.

Figure3: an example on date recognition

Summarized text: *The number of mobile MAUs increased from 136 million in the month ended December 31, 2013 to 163 million in the month ended March 31, 2014. The number is now 188 million.* In this summarization, BertSum recognizes that the current date in the original text is June 30, 2014, so instead of saying "the number increased to 188 million in the month ended June 30, 2014", BertSum rephrases it to "the number is now 188 million", which shows that it is able to avoid too many repetitive expressions and add more varieties to the summarization.

4.2 Unsatisfactory Results

We should say that we experienced a lot of problems for the results.

1. Failure to identify the coordinate clause/sentences.

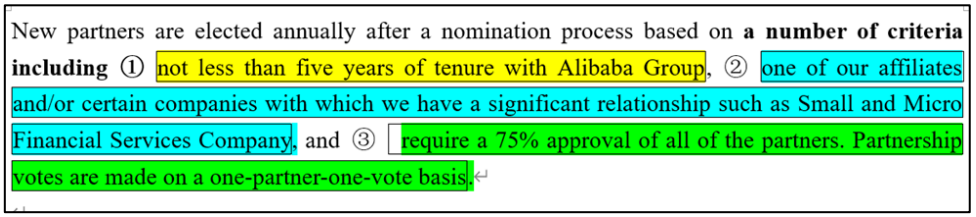
New partners are elected annually after a nomination process based on **a number of criteria including** ① not less than five years of tenure with Alibaba Group, ② one of our affiliates and/or certain companies with which we have a significant relationship such as Small and Micro Financial Services Company, and ③ require a 75% approval of all of the partners. Partnership votes are made on a one-partner-one-vote basis.

Figure4

Summarized Text: *New partners are elected annually after a selection process. The vote is based on a number of criteria including not less than five years of tenure with alibaba group.* In this case, we noticed that even though these three colored clauses are

all explanations for the previous clause, the required criteria for the system can only identify the first clause and eliminate others.

We believe that the possible reason might be the model was trained by news and usually the most common format for news is to use outlines to briefly summarize the main idea of the whole passage.

2. The model will replace standard terminology with synonyms.

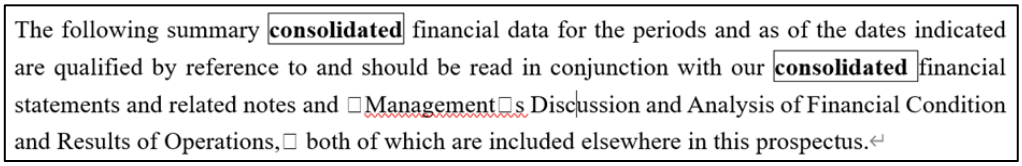
The following summary **consolidated** financial data for the periods and as of the dates indicated are qualified by reference to and should be read in conjunction with our **consolidated** financial statements and related notes and □Management□s Discussion and Analysis of Financial Condition and Results of Operations,□ both of which are included elsewhere in this prospectus.↵

Figure5

Summarized Text: *The preliminary data should be read in conjunction with our gathered financial statements and related notes.* In this case, we found that our model used "preliminary" and "gathered" to replace the word "consolidated". However, the standard terminology in accounting principles is "consolidated" financial data/statement. Possible reason: The word "consolidated financial data/statement" has appeared so many times in the previous chapters, so the model does not know the standard terminology and wrongly regards the repeated terms as duplicate errors.

3. Delete all bullet points.

We find that our model cannot understand some commonly used format in the prospectus. It will delete all the bullet points that explain the previous leading sentence. These two examples showed the same problem.

4. Keep the Specific Numbers.

We noticed that the algorithm tends to keep the information that contains specific numbers or dates in the summary. However, sometimes in the prospectus, a series of numbers are only listed to illustrate one viewpoint. So keeping the numbers and eliminate the statements is not suitable.

5. Summarized a sentence from its original training set.

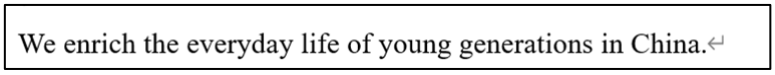
We enrich the everyday life of young generations in China.↵

Figure6

Summarized Text: *We enrich the every day life of young generations china, say cnn's Barbara starr. We also enrich the lives of young generations china.* The reason for this error remains unknown because we cannot understand why our model used a sentence quoted by Barbara of CNN from its original training set.

5 Conclusion

We found that Transformer captured more connections between each terms/sentences. We proved that BertSum can be applied in text summarization for financial filings. However, more training is still needed:

1. Terminology (legal terms, financial terminology, etc.)
2. More features of a sentence which usually appeared in the document. We need the model to understand the connections between each term.
3. Specific formats for fillings (i.e. bullet points)
4. The formal/frozen style for the documents

We also noticed ROUGE evaluation is not suitable for our results because ROUGE measures degree of similarity regardless of the content. Besides, our sample size is too small so the result will not be so convictive.