CILAB@FinTOC-2021 Shared Task: Title Detection and Table of Content Extraction for Financial Document

Hyuntae Kim, Soyoung Park, Seongeun Yang, Yuchul Jung*

Department of Computer Engineering,
Kumoh National Institute of Technology (KIT), Gumi, Korea
{ hyuntaekim09, haluna8836, banilla03090, enthusia77}@gmail.com

Abstract

This paper describes Cilab's system for extracting the title and table of contents (TOC) of FinToc-2021 Shared Task's financial documents. We use FinTOC2021-trainset, FinTOC2020-testset provided by the organizer, and 444 documents obtained from four different financial companies (or organizations). In terms of the training algorithm, we selected a random forest classifier for title detection and considered font style, bold, font size, and text to determine the level of the title. Among English and French tracks, we participated only in English tracks and ranked fourth (F1-score, 51.4%) among the six teams that participated in title detection. In TOC extraction, we achieved third place (Harmonic mean, 26.3.) among the six teams.

1 Introduction

Many recent studies (Pappagari, R. et al., 2019; Martha O. Perez-Arriaga, 2016; Adhikari et al., 2019) have attempted to grasp the structure of PDF documents. However, most of these studies are on thesis-type or receipt-type documents with formal structures. Few studies identify the structure of financial documents, detect titles, and extract tables of contents (TOC).

The goal of FinTOC-2021 is to extract specific texts that act as the title, section, and title of the document in the financial document and generate a table of contents of the extracted titles.

Towards a robust TOC extraction from financial documents, we start by finding a good feature set required for distinguishing title text from non-title texts. Besides, to find & train a best-performing title classifier, a total of five models (e.g., Random Forest, SVM (Jiu-Zhen Liang, 2004), Naïve Bayes, Logistic Regression, and BERT based fine-tuning) were tested. Moreover, to overcome the lack of training data, 400+ financial documents were newly collected and used for building the title classifier. Finally, based on an in-depth investigation of the given training data, a set of heuristics was designed for post-processing, allocating appropriate depth levels based on the given title texts.

This paper shortly summarizes recent approaches of title detection and TOC extraction in Section 2, analyzes the data sets provided and additionally collected data in Section 3, talks about the structure and internal experiments of the system that detects title and extracts TOC in Section 4, describes official results in Section 5, and present our future work in Section 6.

2 Related Work

In FinTOC-2020 shared tasks on structure extraction from financial documents, various approaches on feature selection, algorithm, and procedures had been attempted.

Title Detection and TOC extraction tasks are highly co-related, and influential features can be obtained basically from preprocessing, like removing the header or footer or erasing the table. After preprocessing, the document's text is classified into title or non-title by applying various machine learning techniques.

As a combination approach, Daniel@FinToC'2 shared task (Emmanuel et al., 2020) combined TOC itself, document wording, and lexical domain knowledge.

Although there can be numerous machine learning approaches for title detection and TOC extraction, interestingly, the random forest algorithm showed the best performance (Kosmajac, Detal, 2020). As a different solution, others adopted maximum entropy classifier (Hercig et al., 2020) and multilingual BERT (Hase et al., 2020)

3 Datasets

FinTOC-2021 provided with two datasets: FinTOC2021-trainset and FinTOC2020-testset. However, due to the lack of training data, we analyzed which financial companies' documents are frequently occurring among various financial documents. Finally, we chose four different financial companies, such as, BI SICAV, UBS, DNB ASA, and SEB. To obtain documents similar to those included FinTOC2021-trainset and FinTOC2020-testset, we selected more than 400 PDF documents by crawling the four companies' websites. The newly constructed dataset was named FHFO (**F**our **H**undred documents from **F**our different **O**rganizations). In our experiments, a total of three datasets were used. Table 1 shows the data statistics for Title/Non-Title labels and the PDF documents included for the three datasets.

Dataset Name	Num. of Titles	Num. of Non-Titles	Num. of PDF Docs.
FinTOC2020 –testset	6.5k	154k	22
FinTOC2021 -trainset	6.5k	144k	42
FHFO Data	52.2k	3.6M	444

Table 1: Data statistics for title detection and table of contents extraction

In the case of FinTOC2021-trainset used 42 of the 47 documents as educational materials, excluding files that could not be opened and those that could not be processed with PDF-miner[1].

Four hundred forty-four financial PDF documents of FHFO data were used to train models that separated titles and non-titles using pseudo-labeling.

In addition, the distribution of the level of the table for contents in the two prepared datasets is shown in Table 2.

Depth Level	2020 Test	2021 Train
Level 1	193	45(0.5%)
Level 2	1149	99
Level 3	1405	975
Level 4	1450	2291
Level 5	430	2652
Level 6	241	980
Level 7	27	645
Level 8	2	211
Level 9	0	62
Level 10	0	0
Total	4897	7960

Table 2: Depth level distribution

4 Our Proposed System

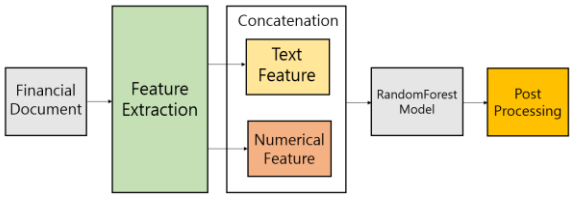

Figure 2: System Architecture

The core components of our proposed financial document structure extraction system consist of three sequential steps, such as feature extraction, random forest model (L. Breiman, 2001) for title detection, and post-processing for resolving depth levels as in Figure 1.

4.1 Feature Extraction

The feature extraction relies on PDF-Miner to extract features from financial documents. The extracted feature is as follows.

- Text: Extract the text of the document by row.

[1] https://github.com/pdfminer/pdfminer.six

- Font Size: Extract the font size of the extracted text.
- Font Style: Extract the font style of the text.
- Bold: Extract 0 or 1 whether the extracted text is bold.
- Text Coordinate: Extract the bounding box area (left upper, right lower) of the extracted text.

Text features and numerical features of text, font size, bold, and text coordinate are used to train the model for title detection, and the features of the font style are added to the features above.

4.2 Random Forest Model for Title Detection

To find the most appropriate dataset combination for building our title detection model, we divided the datasets in Section 3 into D1, D2, D3, and D4. Table 3 lists the dataset combinations used in our experiments.

	Dataset Combinations
D1	2021-trainset
D2	2021-trainset+2020-testset
D3	2021-trainset+FHFO data
D4	2021-trainset+2020-testset+FHFO data

Table 3: Dataset description

To decide the best classification algorithm for title classification, we tried a total of five models (e.g., Random Forest, SVM (Jiu-Zhen Liang, 2004), Naïve Bayes, Logistic Regression, and BERT based fine-tuning) among various machine learning algorithms as in Table 4. Experiments were conducted on the D1 dataset with the five models, and the Random Forest model showed the best performance.

Model	Precision	Recall	F1
SVM	0.84	0.84	0.84
Naïve Bayes	0.82	0.82	0.82
Logistic Regression	0.83	0.83	0.83
BERT based Fine-tuning	0.79	0.70	0.67
Random Forest	**0.87**	**0.86**	**0.86**

Table 4: Title classification performances of five different algorithms for the D1 dataset

Table 5 summarizes the results of applying the random forest algorithm from D1 to D4 datasets. We achieved an F1 score of 96% with the D4 dataset. The second-best performance was gotten with the D2 dataset.

Dataset	Precision	Recall	F1
D1	0.87	0.86	0.86
D2	0.94	0.91	0.92
D3	0.89	0.87	0.88
D4	**0.97**	**0.95**	**0.96**

Table 5: Title classification performances for D1~D4 datasets with random forest algorithm

4.3 Post-Processing for Resolving the Depth Levels

After that, post-processing procedures were carried out to determine the depth level for each title detected. To this ends, FinTOC2021-trainset and FinTOC2020-testset were verified semi-automatically in order to establish our internal criteria for allocating depth levels.

Through an in-depth analysis, we delineated the following heuristics for resolving depth levels for each title.

1) From depth 1 to 2, they mainly deal with the contents of titles and sub-headings and the large font size.

2) From depth 3 to 5, they mostly contain figures which represent the levels of titles.

3) Depths of 6~7 have a smaller font size and are rarely less than depth 8. Thus, our heuristics are target to allocate between depths 1 ~ 7 entirely.

Our internal post-processing criteria are generally based on the font style and font size. Therefore, bold or Italic was set as an essential criterion among the font styles, and the Font size was divided by the most significant or most minor Font size.

Since the depths of most titles were 3 to 5 levels, it was difficult to distinguish only by manual verification, and in this case, the font style was first screened out and through a regular expression to check whether there are specific rules in the text.

To derive a set of rules for the post-processing, we further subdivided and generalized the above

findings by applying them to FinTOC2021-trainset and FinTOC2020-testset.

5 Official Results

In this paper, the system of CILAB who participated in the FinTOC 2021 sharing work was explained. The proposed system ranked 4th in Title Detection and 3rd in TOC Extraction. Table 6 and Table 7 & 8 show the official evaluation results of title detection and TOC extraction, respectively.

Team	Precision	Recall	F1
Cilab_fintoc1	0.702	0.376	0.456
Cilab_fintoc2	0.708	0.422	0.514

Table 6: Title detection results of Cilab team

Team	Precision	Recall	F1
Cilab_fintoc1	26.6	14.4	17.6
Cilab_fintoc2	30.5	18.6	22.6

Table 7: TOC extraction results of Cilab team

Team	Title acc.	Level acc.	Harm. mean
Cilab_fintoc1	34.2	34.8	23.4
Cilab_fintoc2	38.9	31.4	26.3

Table 8: TOC Extraction results from "Inex08-result"

Among the two submitted runs, the 2nd run (i.e., Cilab_fintoc2) performed better. It ranked fourth in title detection and third in TOC extraction out of six teams, respectively.

The model used a random forest classifier and PDF-Miner to distinguish the title from various financial documents and extracted features such as text, font size, font style, and bold. Then, we went through assigning depth to the extracted feature based on the depth arbitrarily determined by us.

The difference between the two models lies in the difference in learning data. In the case of Cilab_fintoc1, the training was performed with FinTOC2021-trainset and FinTOC2020-testset. On the other hand, Cilab_fintoc2's training data includes our constructed FHFO data additionally. As a result, it was confirmed that more training data from similar publication organizations could produce better results.

However, there are many noises in detecting the only title in the financial documents, and the data of non-title is much more biased, indicating that the performance, when applied to actual data, is much lower than that of the training stage.

6 Future Work

As our future work, we are interested in building a more robust TOC extraction model using multi-modal machine learning techniques specialized in financial documents that better combine visual features and text features.

7 Reference

El Maarouf, Ismail and Kang, Juyeon and Aitazzi, Abderrahim and Bellato, Sandra and Gan, Mei and El-Haj, Mahmoud "The Financial Document Structure Extraction Shared Task (FinToc 2021) " (2021).

Kosmajac, D. et al. "DNLP@FinTOC'20: Tableof Contents Detection in FinancialDocuments." Fin ancial Narrative Processing Workshops (2020).

Hase, Frederic and Steffen Kirchhoff."Taxy.io@FinTOC-2020: Multilingual DocumentStructure Extraction using Transfer Learning." Financial Narrative Processing Workshops (2020).

Hercig, Tomás and P. Král. "UWB@FinTOC-2020 Shared Task: Financial Document Title Detection." Financial Narrative Processing Workshops (2020).

Emmanuel Giguet ,Gaël Lejeune ,Jean-Baptiste Tanguy" Daniel@FinTOC '2 Shared Task: Title Detection and Structure Extraction" Financial Narrative Processing Workshops (2020).

L. Breiman. Random forests. Machine Learning, 45:5–32, 2001. Text Extraction using Document Structure Features and Support Vector Machines

Jiu-Zhen Liang, "SVM multi-classifier and Web document classification," Proceedings of 2004 International Conference on Machine Learning and Cybernetics (IEEE Cat. No.04EX826), 2004, pp. 1347-1351 vol.3

Martha O. Perez-Arriaga, Trilce Estrada, and Soraya Abad-Mota "TAO: System for Table Detection and Extraction from PDF Documents"

Adhikari, Ashutosh & Ram, Achyudh & Tang, Raphael & Lin, Jimmy. (2019). DocBERT: BERT for Document Classification.

Pappagari, R. et al. “Hierarchical Transformers for Long Document Classification.” 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU) (2019): 838-844.