Financial Document Structure Extraction

Contents

1 Abstract

This paper presents some of the best approaches or state of the art techniques for structure extraction from financial documents. The problem at hand is of great significance as these financial documents play a crucial part in running and managing business firms worldwide. Usually, such documents do not come with a predefined table of contents and well-structured sections, which can help users better navigate documents. Along with the variability in structures, different firms around the globe use different layout strategies, which makes this task even more relevant these days.The results and findings of the best solutions in a comparative study manner is presented along with the overview of the adopted methodologies in this problem's context. The objective of the task is to extract a table of contents from financial documents by detecting the titles of the documents and then organizing them hierarchically into a Table of Contents(TOC).

2 Introduction

Financial narrative disclosures portrays a greater part of a company's overall financial communications with the investors. The document presents in depth financial data and uses advanced financial reporting tools to build a narrative around financial performance which helps the executives and anyone invested in finance to understand how things are going and where they're headed. The stakeholders of companies have always looked into the past to anticipated the future.These documents aim to go beyond the spreadsheets and explanations and to make the numbers more engaging and inclusive such that any reader can comprehend. These generally contains activities, financial situation, investment plans, and operational information which provides a robust mechanism to support the creation of a more commercially attractive and differentiated picture of the business providing investors better understanding and improve stakeholder's relationships. This underlying process is necessary to enhance Board effectiveness and improve governance. The reports are usually created on an annual basis in machine readable formats often only with minimal structure information.These financial narratives are generated worldwide with different structures and templates. Unlike the USA which follows a generalised template for all of its financial disclosures as proposed by EDGAR SEC or AMF, other countries follow inconsistent layout for managing their financial narratives.Which means some documents are known to be with or without TOC which makes it extremely difficult to navigate. In addition to this, the text format,graphics and layout depends mostly on the publisher,thus making TOC an integral part of the document to enhance readability.

Analysing and understanding a document structure is the first step towards carrying out tasks such as search, summarization, entity extraction and title detection etc. Unlike humans, machine don't have the ability to have a glance on the document to comprehend the structure as they lack intuitive ability and

instincts which present challenges in extracting information from documents.

Literature Survey

2.1 AMEX-AI-LABS: Investigating Transfer Learning for Title Detection in Table of Contents Generation

In this paper[1],Premi,et al. proposed a transfer learning approach for Title Detection. Transfer learning depends upon the fact that the character features and geometric layout of non-titles and titles can be learnt from a large corpus separately. Then the learning can be transferred to a domain-specific dataset on which they train a Deep Neural Network on the document's text along with a pre-trained model for geometric layout and character features. They managed to achieve an F-Score of 83.25 on the test set.
The authors have observed that in the final model the pre-trained weights captured the generic structure of documents, giving a boost to accuracy. This transfer learning approach can be improved further by using better architectures and features which are domain-independent.This procedure achieved a 10 percent increase in F Score. And the combination of geometric and character-based features complemented each other to attain higher accuracy compared to either of them separately.

2.2 UWB@FinTOC-2020 Shared Task: Financial Document Title Detection

In this paper[2], the researchers use Apache PDFBox library for text extraction and all additional information like font size and font type from the financial documents. Their system relies on the provided training data only without usage of any external resources.They carried out ablation experiments to demonstrate which features are more advantageous . The numbers represent the performance change when the given feature is removed, lower number meaning better feature. We used approximately 30 percent of the fixed training dataset3 for testing and the rest for training the features. As the previous experiment's inaccuracy made them repeat it again using leave-one-out cross-validation.It was observed that first and last orto-characters and font type unigrams are the most helpful features, while position features and font size features were the least helpful. The whole system is based on Maximum Entropy classifier and other features like font size and font type. This achieves F1 score 81 percent in the French track and 77 percent in the English track.

2.3 Daniel@FinTOC '20 Shared Task: Title Detection and Structure Extraction

Giguet et al.[3] proposed a different and exciting approach to handle this problem by extracting the textual content of the financial documents by eliminating

other forms of information present in the documents like tabular ones. The primary focus is on the text regions of documents to detect their titles and hierarchically arrange them to represent them as Table of Contents. This table of contents acts as an index of significant document regions for easy navigation of documents, especially financial ones. Their approach for finding titles was centered around using font and stylistic features of lines in the extracted text regions. In other terms, they mapped these stylistic features to the character-based features as available in the provided dataset. They used ToC Detection and Extraction modules to extract potential titles from tables and Numbered List detection module and Text Saliency Detection module, to get the same from continuous textual data. Mostly they used character n-grams and surface features for title detection. For the ToC generation, they made use of document wording and knowledge of linguistic analysis. This aid in finding the suitable table of contents from documents that are free from errors and further helped them distinguish between less notable titles and preliminary titles that are usually not included while making ToC of documents. For classification, they used RandomForest with 50 percent estimators. They achieved an F1 score of 0.22 and 0.28 simultaneously for French and English datasets, respectively, for the ToC task. They managed to acquire 0.62 as the harmonic mean between the Inex F1 score and Inex level accuracy for the Title Detection subtask.

2.4 DNLP@FinTOC'20: Table of Contents Detection in Financial Documents

Kosmajac et al.[4] in their work handled normal pdfs along with scanned pdfs for generating Table of contents and Title detection subtasks. For scanned pdfs, they incorporated the use of a powerful, open-source optical character recognition tool, tesseract. This OCR tool helped them to find relevant regions in financial documents containing textual data. They used a feature set defined by Akl et al.[6] for this purpose. By using tesseract in conjunction with Levenstein distance allowed them to achieve a harmonic mean of 0.36 and 0.39 in the Table of Contents generation task for English and French languages, respectively. They tested their feature set on three different classifier algorithms: Random Forest, Linear Regression, and SVM, out of which they achieved the best results for the Random Forest approach. They detected titles by performing a mapping between the labeled titles available to them with the text regions obtained after OCR scanning. They incorporated Levenstein distance into their proposed approach up to a threshold of 3 for maximum distance, which is particularly obtained by matching standard gold titles with detected titles. The most influential features found out to be the text distance from the nearest upper neighboring text, the width of the bounding box for text regions, and the estimate of character boldness.They acquired an F1 score of 0.73 and 0.87 for English and French languages, respectively, for the title detection task.

2.5 Taxy.io@FinTOC-2020: Multilingual Document Structure Extraction using Transfer Learning

Haase and Kirchhoff[5] proposed the use of clustering, an unsupervised learning technique for the Title detection and ToC generation tasks. Their approach was based on finding text blocks in financial documents using the DBSCAN clustering algorithm twice. They first extracted features by performing clustering on different characters in document pages, followed by detecting text blocks with the second run of DBSCAN. They incorporated a multilingual BERT model for extracting textual features, which they used in conjunction with other found features to classify the extracted blocks of text. These text blocks act as candidates for titles for the title detection subtask. They designed their single multilingual Bert model to efficiently handle both tasks. Using transfer learning along with text and document layout features in conjunction with a self-constructed regression framework helped them to achieve harmonic mean as 0.24 and 0.32 in title detection and F1 score of 0.55 and 0.69 in ToC generation task for English and French languages, respectively.

2.6 FinTOC-2019 Shared Task: Finding Title in Text Blocks

The first model Hanna et al.[6] evaluated was SVM classifier, which was trained on a union of features from the existing features present in the original csv file in addition to other features which have been extracted from pre-processing of xml files. These features encapsulate the for and layout of the text that plays a crucial part in classification.They also computed width,height,font,number of dots,count of capital letters,character count,word count,average word length etc along with the provided features. It was observed that the variants and results were very close with or without having considered CNN outputs as input features. This encouraged the authors to develop to deep learning techniques centered on raw text entries and focusing on regularisation methods. This was done so that the high variance value observed in SVM result can be changed.

The next model was a BiLSTM–Attention model depending on word embedding. This model had 2 dense layers that's comprises of 64 neurons and established with batch size of 256 and 100 epochs respectively. The primary motive of this model was evaluation of the possible semantic composition of sentences and checking how applicable they are for the task.

3 Results

The title detection (TD) ranking is based on F1-score, while the Table-Of-Content (TOC) ranking is based on the harmonic mean between Inex F1 score and Inex level accuracy.

3.1 Task 1: TOC extraction from French documents

Team	Title Detection
UWB	0.81
Taxy.io	0.69
Daniel 1	0.66
DNLP	0.64
Daniel 2	0.64
Daniel 3	0.64
Baseline	0.57

Table 1.1

Team	TOC Generation
DNLP	0.37
Taxy.io	0.32
Baseline	0.32
Daniel 1	0.22
Daniel 2	0.22
Daniel 3	0.20

Table 1.2

3.2 Task 2: TOC extraction from English documents

Team	Title Detection
AMEX 1	0.79
UWB	0.77
Daniel 1	0.69
Daniel 3	0.63
Daniel 2	0.62
DNLP	0.59
Baseline	0.19

Table 2.1

Team	TOC Generation
DNLP	0.34
Daniel 3	0.28
Daniel 2	0.28
Daniel 1	0.26
Taxy.io	0.24
AMEX 1	0.23
AMEX 2	0.23
Baseline	0.18

Table 2.2

4 Conclusion

In this shared task,the main aim is to detect the title and extract the Table of Contents of financial prospectuses.The financial documents which present the state of the companies and modalities of their investment might follow different structure or template depending on their country of origin.Without the table of contents it becomes difficult to comprehend for the end user.In this report we have done a detailed comprehensive assessment and summarization of research papers that have detected the titles and generated complete table of contents.There is a huge scope of improvement that can be done in this field. For instance, some of the deep learning methods can be applied to both title detection and table of contents generation. In addition to this embedding information pointers in documents can be used which can help better identify titles, sections, headers, and sub headers.Taking the help of computer vision for title detection we can use it to guide better generation of tables of contents.

5 References

[1]. Premi, Himanshu. "AMEX-AI-LABS: Investigating Transfer Learning for Title Detection in Table of Contents Generation." . In Proceedings of the 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation (pp. 153–157). COLING, 2020.
[2]. Hercig, Tom and Kr´al, Pavel."UWB@FinTOC-2020 Shared Task: Financial Document Title Detection." . In Proceedings of the 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation (pp. 158–162). COLING, 2020.
[3]. Giguet, Emmanuel and Lejeune, Ga¨el and Tanguy, Jean-Baptiste. "Daniel @ FinTOC'2 Shared Task: Title Detection and Structure Extraction" . In Proceedings of the 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation (pp. 174–180). COLING, 2020.
[4]. Kosmajac, Dijana and Taylor, Stacey and Saeidi, Mozhgan. "DNLP @ FinTOC'20: Table of Contents Detection in Financial Documents" . In Proceedings of the 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation (pp. 169–173). COLING, 2020.
[5]. Frederic Haase and Steffen Kirchhoff. "Taxy.io@FinTOC-2020: Multilingual document structure extraction using transfer learning." In Proceedings of the 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation (pp.163–168).COLING, 2020.
[6]. Akl, Dominique. "FinTOC-2019 Shared Task: Finding Title in Text Blocks."In Proceedings of the Second Financial Narrative Processing Workshop (FNP 2019) (pp. 58–62). Linköping University Electronic Press, 2019.