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

Frederic Haase
Taxy.io GmbH
Aachen, Germany
haase@taxy.io

Steffen Kirchhoff
Taxy.io GmbH
Aachen, Germany
kirchhoff@taxy.io

Abstract

In this paper we describe our system submitted to the FinTOC-2020 shared task on financial document structure extraction. We propose a two-step approach to identify titles in financial documents and to extract their table of contents (TOC). First, we identify text blocks as candidates for titles using unsupervised learning based on character-level information of each document. Then, we apply supervised learning on a self-constructed regression task to predict the depth of each text block in the document structure hierarchy using transfer learning combined with document features and layout features. It is noteworthy that our single multilingual model performs well on both tasks and on different languages, which indicates the usefulness of transfer learning for title detection and TOC generation. Moreover, our approach is independent of the presence of actual TOC pages in the documents. It is also one of the few submissions to the FinTOC-2020 shared task addressing both subtasks in both languages, English and French, with one single model.

1 Introduction

While large amounts of documents are created and published in machine-readable file formats such as PDF, only a small fraction come with structural information, e.g. on their table of contents (TOC). This is especially true in the financial domain where a table of contents, e.g. for annual reports or shareholder reports, would be particularly helpful. Even though it is often regulated which content these documents must include (Juge et al., 2019), an automated analysis of their structure remains difficult for several reasons. One is that financial documents usually come with a complicated layout consisting of text blocks, tables and figures of various kinds. Also, such documents typically come with a highly nested hierarchy of subsections. Additionally, most of these documents are published as PDF, which is the defacto standard for the creation of electronic documents, even though this file format comes with several inherent drawbacks (Hu and Liu, 2014). PDF files contain little structural information about the contents, such as words, lines or paragraphs, which would be helpful for automatic structure analysis. Moreover, there exist various different ways to generate a PDF, which further complicates the automatic analysis of the structure of such documents, especially in multicolumn settings.

Despite the importance and diverse use cases of automatic structure analysis of documents, there is only a small research stream focusing on this topic. The FinTOC-2020 shared task about Financial Document Structure Extraction (Bentabet et al., 2020) promotes research in this field by proposing two language tracks for English and French together with a benchmark data set of financial documents. The goal of the shared task is to design systems that solve the following two subtasks:

- Title detection: Given a document, extract text blocks and identify which ones are titles.
- Table of contents (TOC) generation: Given a document, identify titles and their nesting depths.

In this paper we describe our solution to the shared task, a multilingual approach based on transfer learning that jointly solves both problems title detection and TOC generation for both languages.

Proceedings of the 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation, pages 163–168
Barcelona, Spain (Online), December 12, 2020.

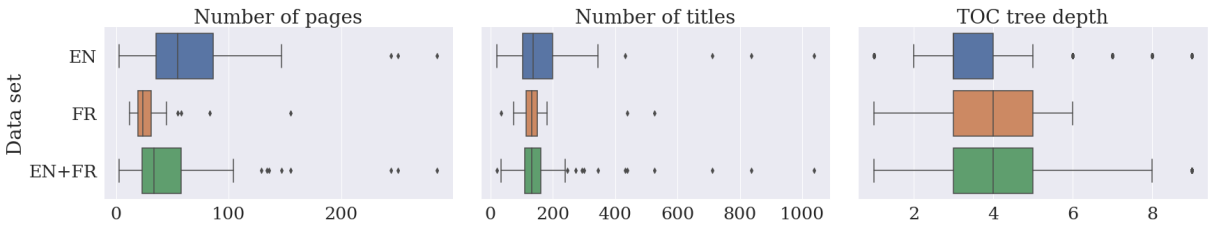

Figure 1: The financial documents of the FinTOC-2020 shared task vary greatly with respect to their number of pages, number of titles and the depth of their TOC tree. On the entire set across both languages, the number of pages ranges from 3 to 285, the number of titles ranges from 20 to 1039 and the TOC tree depth from 1 to 9.

The paper is organized as follows. We first review previous work in Section 2 and describe the data that was provided for the FinTOC-2020 shared task in Section 3. We then present our system in Section 4 before we report and discuss our results in Section 5. The paper ends with a conclusion in Section 6.

2 Previous Work

The first approaches to TOC generation are based on physical layout analysis, e.g. (Conway, 1993). Here, physical entities of a document, such as pages, paragraphs, and figures are extracted and mapped onto a hierarchy of logical entities such as titles, authors and sections from which the TOC is then derived. These approaches are often based on heuristic rules and grammars (Mao et al., 2003). As these predefined rules are mostly domain-specific, these approaches fail to generalize to a diverse set of documents from different domains (Najah-Imane et al., 2019).

Other approaches explicitly parse the title hierarchy from embedded TOC pages (Dresevic et al., 2009; Nguyen et al., 2017). While these approaches generalize across different domains, they fail on documents without TOC pages or whenever the TOC page does not reflect the entire document structure, e.g. in cases of deeply nested subtitles not mentioned on the TOC page (Giguet and Lejeune, 2019).

Yet another set of approaches leverage learning-based methods to predict the TOC, e.g. based on both layout and text features of the document. In (Najah-Imane et al., 2019), for example, titles are first detected using a convolutional neural network (CNN) (Kim, 2014); then the corresponding depth of each title is predicted using a combination of a bidirectional long short term memory (BiLSTM) network and a conditional random field (CRF) model, as suggested by (Huang et al., 2015) for sequence tagging.

The approaches from the FinTOC-2019 challenge (Juge et al., 2019) mostly fall into the latter two categories: explicit TOC parsing and learning-based methods.

3 Data

The training data of the FinTOC-2020 shared task consists of 52 English and 47 French financial PDF documents. Every document page comes along with a set of annotations, which include both the text and the depth for each title. The box plots in Figure 1 show the heterogeneity of the financial documents from this data set. In the median, a financial document from this data set has 34 pages, 132 titles and a TOC tree depth of 4. However, the number of pages ranges from 3 to 285, so some documents are very small while others are rather large. Also, the number of titles varies from 20 to 1039 across the different documents. While some TOC trees are flat with a depth of only 1, some are very nested with a depth of 9. This heterogeneity of the financial documents in the data set underlines their complex layout and hence the difficulty of the task to detect titles and extract the TOC.

4 System

Our system works directly on the entire content of a document and is trained for both English and French documents jointly. We apply a two-step approach, where we first detect candidate text blocks and predict their TOC depth in a subsequent step as a regression task. The system is described in Figure 2.

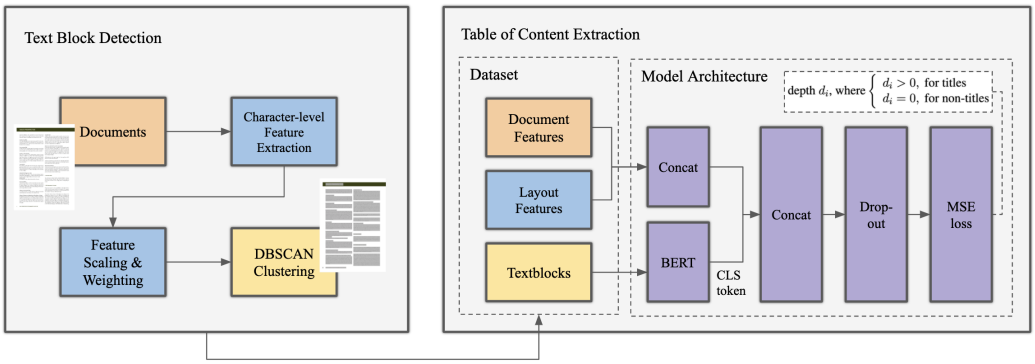

Figure 2: Text blocks are detected (left side) by extracting, scaling and weighting character-level features on which a DBSCAN clustering is applied. For TOC generation (right side), document features, layout features and text blocks are combined to train a neural network to predict the depth of each text block.

4.1 Text Block Detection

Previous work on text block detection applies spatial layout processing in combination with a defined set of rules, e.g. (Ramakrishnan et al., 2012). Other works leverage learning-based approaches, e.g. the work of (Klampfl et al., 2014), where text blocks are detected using the unsupervised learning approach of hierarchical agglomerative clustering (HAC). Their clustering mechanism joins characters into groups that represent words, and groups of words into contiguous blocks of text.

Similarly, we apply DBSCAN clustering (Ester et al., 1996) and combine layout features (character coordinates) as well as font features (font size, font color and font type). We find DBSCAN clustering to outperform HAC on this given data set. For each character on a page of a document, we extract the bounding box coordinates and label-encode font color, font size and font type. After extraction, these features are scaled using min-max normalization. Then, we apply the DBSCAN clustering with Minkowski distance of order 1 on the character-level data set per page.

To find appropriate hyperparameters for the DBSCAN clustering and appropriate scaling weights for our layout features and font features, we use Bayesian optimization as suggested in (Snoek et al., 2012). As objective we maximize the amount of titles that can be mapped correctly to the training data annotations. To map candidate titles to ground truth titles we use the same customized Levenshtein distance from the evaluation metric of the FinTOC-2020 shared task (Bentabet et al., 2020). While we optimize the Eps hyperparameter of DBSCAN, we fix the $MinPts$ parameter to 1. This hyperparameter optimization process has revealed the following: The vertical position, font type, font color and font size are important features when applying clustering on character-level. After optimization, our approach was able to extract text blocks that could be matched to 95.7% of the titles from the training data annotations.

4.2 Table of Contents Extraction

To estimate the depth of the text blocks extracted in the previous step, we construct a data set that can be used for supervised learning. For each text block t_i that can be mapped to a ground truth title, we set its label to the ground truth depth d_i. We assume that text blocks that cannot be mapped to a ground truth title to be in fact not a title. For such text blocks we define their TOC depth to be $d_i = 0$. That is, we treat depth 0 as a placeholder for everything that is not a title. This notation allows to train one model that can be applied to both the title detection subtask and the TOC generation subtask. Now, we have an annotated set of text blocks together with their depth in the TOC tree. We aim to train a model that we can use to predict the TOC tree depth of a given text block. For every text block we compute the following features: First, we represent a text block by the features used for the DBSCAN clustering as described in section 4.1. Then, we add layout features to describe the majority font color, font size and font type of all characters in the text block. To also capture information that puts the text block in relation to the document, we further compute the following features: `is_most_frequent_font`, `is_most_frequent_color` and `is_most_frequent_size`. Additionally, we add document features to the data set, comprising min, max and mean font size of the entire document.

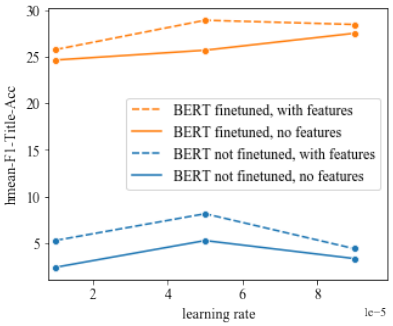

Figure 3: In our experiments we observed a positive impact of both fine-tuning the pretrained layers of the BERT model (orange vs blue lines) and of additionally incorporating document and layout features (dashed vs solid lines).

Title detection task				TOC generation task			
FR		**EN**		**FR**		**EN**	
Team	F1	Team	F1	Team	hmean-F1-Title-Acc	Team	hmean-F1-Title-Acc
UWB	0.81	Amex 1	0.79	DNLP	0.37	DNLP	0.34
Taxy.io	0.69	Amex 2	0.79	Taxy.io	0.32	Daniel 3	0.28
Daniel 1	0.66	UWB	0.77	Daniel 1	0.22	Daniel 2	0.28
DNLP	0.64	Daniel 1	0.69	Daniel 2	0.22	Daniel 1	0.26
Daniel 2	0.64	Daniel 3	0.63	Daniel 3	0.20	Taxy.io	0.24
Daniel 3	0.64	Daniel 2	0.62			Amex 1	0.23
		DNLP	0.59			Amex 2	0.23
		Taxy.io	0.55				

Table 1: In comparison to the other approaches submitted to the shared task, our approach ranks 2nd on the title detection task and 2nd on the TOC generation task for French, while we rank 8th and 5th on the respective tasks for English.

Using this data set, we train a model that combines the text blocks with the document features and the layout features to predict the depth d_i. To encode the text block t_i, we use the pre-trained multilingual cased BERT-Base model[1] (Devlin et al., 2019). The classification token (`CLS`) output of BERT, which serves as aggregate sequence representation, is concatenated with the document and layout features and put through a dropout layer. The training objective of the model is to optimize the mean squared error (MSE) loss to predict the depth of each text block according to our data set as a regression task.

5 Results and Discussion

In our experiments, we optimized our model for different hyperparameters using a 20% dev set and the evaluation metric from the FinTOC-2020 shared task (Bentabet et al., 2020). More specifically, we evaluated 1) the impact of fine-tuning the pretrained layers of the BERT model, 2) the impact of additionally incorporating document and layout features, and 3) the impact of different learning rates. Figure 3 shows a positive impact of both fine-tuning the pretrained layers of the BERT model (orange vs blue lines) and of additionally incorporating document and layout features (dashed vs solid lines). The inverted U-shape of the curve indicates that 5e-5 is a good choice for the learning rate. Across all experiments we trained our models for 2 epochs using a fixed 0.25 dropout ratio in the dropout layer. We trained the model 3 times with different seeds and averaged the outcomes for the final depth prediction, which we rounded to get an integer depth. We discarded text blocks with depth 0 to only submit titles.

Table 1 shows the final results on the test set in comparison to the other approaches that were submitted to the challenge. Our approach ranks 2nd on the title detection task and 2nd on the TOC generation task for French, while we rank 8th and 5th on the respective tasks for English. It is noteworthy that our single multilingual model performs well on both tasks and on both languages, which indicates the usefulness of transfer learning for both title detection and TOC generation. The advantage of jointly training a model for both languages is that we can leverage the small amount of training data, which is particularly useful for the language independent layout features. Moreover, our approach is independent of the presence of actual TOC pages in the documents. It is furthermore one of the few submissions addressing both subtasks in both languages with one single model. The constructed data set for supervised learning allowed us to train one model for both subtasks, which we find to be advantageous on the small amount of labeled training documents.

However, our approach also comes with limitations. Selecting fixed clustering hyperparameters assumes that all documents share a similar layout. While we optimized the text block detection hyperparameters for the given set of financial documents, this does not generalize well for other documents, which is mainly due to the limited size of the training set with only 99 documents. Moreover, our text block detection is based on DBSCAN, which requires label encoding where categorical features, e.g. font type, are artificially transformed into numerical ones. Using clustering algorithms such as (Ahmad and Khan, 2019), which handle numerical and categorical attributes directly, is promising future work.

[1] `https://github.com/google-research/bert/blob/master/multilingual.md`

6 Conclusion

We proposed a multilingual two-step approach for both title detection and TOC generation. First, we identified title candidates by clustering a document into text blocks using DBSCAN. Then, we trained a neural network to predict the depth of each text block in the document structure hierarchy. The architecture of the network combines a pre-trained multilingual BERT model with a carefully selected set of document and layout features. We have learned that fine-tuning BERT and adding document and layout features improves the TOC generation accuracy. The approach presented in this paper can be used for both English and French financial documents, even in cases where a TOC page is not present.

References

A. Ahmad and S. S. Khan. 2019. Survey of state-of-the-art mixed data clustering algorithms. *IEEE Access*, 7:31883–31902.

Najah-Imane Bentabet, Rémi Juge, Ismail El Maarouf, Virginie Mouilleron, Dialekti Valsamou-Stanislawski, and Mahmoud El-Haj. 2020. The Financial Document Structure Extraction Shared task (FinToc 2020). In *The 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation (FNP-FNS 2020*, Barcelona, Spain.

A. Conway. 1993. Page grammars and page parsing. a syntactic approach to document layout recognition. In *Proceedings of 2nd International Conference on Document Analysis and Recognition (ICDAR '93)*, pages 761–764.

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, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Bodin Dresevic, Aleksandar Uzelac, Bogdan Radakovic, and Nikola Todic. 2009. Book layout analysis: Toc structure extraction engine. volume 5631, pages 164–171, 09.

Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, volume 96, pages 226–231.

Emmanuel Giguet and Gaël Lejeune. 2019. Daniel@FinTOC-2019 shared task : TOC extraction and title detection. In *Proceedings of the Second Financial Narrative Processing Workshop (FNP 2019)*, pages 63–68, Turku, Finland, September. Linköping University Electronic Press.

Jianying Hu and Ying Liu, 2014. *Analysis of Documents Born Digital*, pages 775–804. Springer London, London.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF models for sequence tagging. *CoRR*, abs/1508.01991.

Remi Juge, Imane Bentabet, and Sira Ferradans. 2019. The FinTOC-2019 shared task: Financial document structure extraction. In *Proceedings of the Second Financial Narrative Processing Workshop (FNP 2019)*, pages 51–57, Turku, Finland, September. Linköping University Electronic Press.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. *CoRR*, abs/1408.5882.

Stefan Klampfl, Michael Granitzer, Kris Jack, and Roman Kern. 2014. Unsupervised document structure analysis of digital scientific articles. *International journal on digital libraries*, 14(3-4):83–99.

Song Mao, Azriel Rosenfeld, and Tapas Kanungo. 2003. Document structure analysis algorithms: a literature survey. In Tapas Kanungo, Elisa H. Barney Smith, Jianying Hu, and Paul B. Kantor, editors, *Document Recognition and Retrieval X*, volume 5010, pages 197 – 207. International Society for Optics and Photonics, SPIE.

Bentabet Najah-Imane, Juge Rémi, and Ferradans Sira. 2019. Table-of-contents generation on contemporary documents. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 100–107. IEEE.

Thi-Tuyet-Hai Nguyen, Antoine Doucet, and Mickaël Coustaty. 2017. Enhancing table of contents extraction by system aggregation. pages 242–247, 11.

Cartic Ramakrishnan, Abhishek Patnia, Eduard Hovy, and Gully APC Burns. 2012. Layout-aware text extraction from full-text pdf of scientific articles. *Source code for biology and medicine*, 7(1):7.

Jasper Snoek, Hugo Larochelle, and Ryan P. Adams. 2012. Practical bayesian optimization of machine learning algorithms. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'12, page 2951–2959, Red Hook, NY, USA. Curran Associates Inc.