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

Dijana Kosmajac **Mozhgan Saeidi** **Stacey Taylor**
Faculty of Computer Science
Dalhousie University, Halifax, Canada
{dijana.kosmajac,mozhgan.saeidi,stacey.taylor}@dal.ca

Abstract

Title Detection and Table of Contents Generation are important components in detecting document structure. In particular, these two elements serve to provide the *skeleton* of the document, providing users with an understanding of organization, as well as the relevance of information, and where to find information within the document. Here, we show that using *tesseract* with Levenstein distance, a feature set inspired by Alk *et al.*, we were able to correctly classify the title to an F1 measure 0.73 and 0.87, and the table-of-contents to a harmonic mean of 0.36 and 0.39, in English and French respectively. Our methodology works with both PDF and scanned documents, giving it a wide range of applicability within the document engineering and storage domains.

1 Introduction

In recent years, there has been increasing interest in applying Natural Language Processing (NLP) techniques to financial documents. One of the main publishing formats for these documents is portable document format (PDF), which are usually available on stock exchange websites, company websites, or linked to as part of regulatory filings. Generally, financial documents are professionally created (which reduces spelling and grammatical noise), are relatively commonly structured (from a reader's point of view), and provide very similar end-user information. There are, however, variations in the organization and naming conventions, which present challenges in extracting information from documents. The main focus of the FinTOC task (Bentabet et al., 2020) is to analyze the the structure of financial documents to detect elements such as headings, sub-headings, and titles. The task is split into two parts: title detection and table-of-contents generation, for both English and French documents.

Our team participated in both tasks, but found more success in the table-of-contents generation (TOC), placing first in the standings. Using the Python implementation of *tesseract*, we identified the text regions, and then used Levenstein distance to match the provided labels to the detected regions. We then selected features, inspired by the work of Akl *et al.* to be used in classification for both the title detection and the table of contents generation (Abi Akl et al., 2019). For the classification, we used Logistic Regression (LR), Random Forest (RF), and Support Vector Machines (SVM).

The rest of the paper is organized as follows: we first begin with Background and Related Works, then outline our Method, Dataset, and finally present our Results, for each task, starting with Title Detection.

2 Background

Title detection and table-of-contents generation are highly related tasks, as they both depend on being able to successfully first detect document structure. A better computational understanding of the structure of annual financial reports will provide opportunities for a much more fine-grained analysis of the narratives, temporal comparison, and provide a basis of comparability between companies and sectors.

The United States (U.S.) is often seen as a benchmark for financial reporting layout as the U.S. Securities and Exchange Commission (SEC) mandates a particular structure for regulatory filing. Another

Proceedings of the 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation, pages 169–173
Barcelona, Spain (Online), December 12, 2020.

aspect of the U.S. standard is that the SEC publishes its regulatory filings in HTML format on its online repository, *Edgar*; companies will then simultaneously make the same information available in PDF format, most often on its corporate website. (El-Haj et al., 2019) This rigorous structure is not found in other countries, making multi-jurisdictional information extraction a significant challenge. To address this, el-Haj *et al.* used 10,000 annual reports during the period of 2003-2014 in Spanish and Portuguese from the London Stock Exchange, and created a tool called FRIE-FRSE (Corporate Financial Information Environment - Final Report Structure Extractor) to detect the annual report structure (El-Haj et al., 2019).

A variety of approaches have been taken in previous work to detect titles. Gopinath *et al.* (Gopinath et al., 2018) used two main approaches for extraction of title and prose: a domain-independent approach (DI) and a domain-dependent approach (DD). The data is mutually exclusively grouped into two subgroups using *k-means* clustering. An overlap score is then generated for potential titles, and those which exceed the 75% overlap threshold are added as candidate titles. The approaches only differ in the feature selection in that a neural network is used to classify examples previously labelled during DI. Hercig and Kral used the *Brainy Maximum Entropy* classifier to detect titles, using 7 key features: character n-grams, binary features, first and last orto-characters, font size and type, as well as text length (Hercig, 2019). At the 2019 FinToc workshop, Daniel also proposed using various features such as visual characteristics, punctuation density and character n-grams, achieving an F1-measure score of 94.88% (Giguet and Lejeune, 2019). The FinDSE team used a supervised learning approach that used linguistic, semantic, and morphological features to classify a block of text as either *title* or *non-title*, resulting in an F1-measure of 97.01% (Abreu et al., 2019).

A table-of-contents is a document's *roadmap*; it shows the hierarchy and organization of a document, helps users quickly access relevant information according to their query, and provides additional information such as the length of the document. Yet, while there is a commonly accepted idea of what information the table-of-contents should contain, there is no actual imposed or required standard that corporate entities must use. In fact, it can be seen that some companies provide a sparse table-of-contents with only the major sections included, whereas others provide much more fine-grained detail. Therefore, being able to independently generate a table-of-contents can significantly enrich the search of electronic documents.

Tian and Peng augmented their training and testing data provided in the task by adding additional fields to the data (Tian and Peng, 2019). From the augmented data, they retrieved unique tokens for training the word embedding. The vectors were then in attention based long-short term memory (LSTM) and bidirectional LSTM models, using 10-fold cross-validation. Their method performed very well, and they ranked 1st in the 2019 competition. Najah-Imane et al. proposed a TOC-generation pipeline based on binary classification and sequence label modelling (Bentabet et al., 2019). Classification is done to separate titles from non-titles, and sequence labelling provides a hierarchical structure to the detected titles. Tuyet et al. (2017) introduced an aggregation-based method to enhance ToC extraction using system submissions from the International Conference on Document Analysis and Recognition (ICDAR) *Book Structure Extraction* competitions (2009, 2011, and 2013). Their results show that using both methods together outperforms existing approaches using both the title-based and link-based evaluation measures on a dataset of more than 2,000 books. By efficiently combining the results of existing systems in an unsupervised way, they consistently beat the state-of-the-art in book structure extraction, with performance improvements that are statistically significant. Akl et al. (2019) used DNN-based approach which it scored a weighted F1 of 97.16% on the test data. They evaluated a SVM classifier. It was trained on a combination of features, compiled from the existing features presented within the original csv file as well as additional features extracted from pre-processing work on the xml files. Their second model is a BiLSTM–Attention model relying on word embedding, to make use of the attention mechanism. The third model is a CNN classifier. The purpose of this method is to evaluate the combinatorics of characters at word level (as predictors) and how relevant they are for the task.

3 Methodology

In the following sections we describe the methodology used to build the prediction models. The processing pipeline consists of two general steps: feature extraction and classification.

3.1 Feature Extraction

First, we used *tesseract* (Kay, 2007), an open-source OCR tool to extract the text regions. For the training set, the matching between the provided labelled titles and the extracted text regions is done using Levenstein distance (with the threshold of maximum allowed distance of 3). The choice for Levenstein distance threshold was done empirically by observing the optimal match between gold standard titles and detected ones (by optimal match we consider the highest number of matched pairs). Although the dataset provided by the organizers contains only searchable PDFs, we opted for using OCR scanning instead of PDF-specific mining libraries to make our approach applicable to the documents that are scanned. Inspired by Akl et al. (2019) we used a similar feature set. The features considered are enlisted below:

- **top distance:** integer indicating the distance of the text with respect to the previous text block of the document page;

- **bottom distance:** integer indicating the distance of the text with respect to the next text block of the document page;

- **indent:** integer indicating the placement of the text with respect to the left of the document page;

- **boldness:** float indicating approximation if the text is bold. This is done by counting the average value of all pixels in the bounding box;

- **height:** integer indicating the vertical space occupied by the text;

- **width:** integer indicating the number of characters in the bounding box;

- **PoS tags:** part of speech tags appearing in the text;

- **is title case:** flag true/false indicating if the text is in title case;

- **is upper case:** flag true/false indicating if the text is in upper case;

- **is lower case:** flag true/false indicating if the text is in lower case;

- **is numbered:** flag true/false indicating if the text starts with numbers;

- **is numbered with letters:** flag true/false indicating if the text starts with enumeration letters (roman numerals or a, b, c,...)

- **enumeration level:** integer indicating the level of numbering (for example 1. is 1, 1.1. is 2, 1.1.1. is 3 etc.). text

3.2 Classifiers

For both tasks we experimented with the same features and three classifiers: Logistic Regression (LR), Random Forests (RF) and linear kernel SVM. The library used is *scikit learn* in Python. For all classifiers hyperparameters used are the ones provided as default by the library.

Language	**Classifier**	**Title detection**	**TOC**
English	SVM	0.67 (±0.06)	0.24 (±0.05)
	LR	0.70 (±0.06)	0.25 (±0.06)
	RF	**0.73 (±0.04)**	**0.36 (±0.06)**
French	SVM	0.76 (±0.05)	0.28 (±0.04)
	LR	0.77 (±0.04)	0.28 (±0.05)
	RF	**0.87 (±0.07)**	**0.39 (±0.05)**

Table 1: Development results.

3.2.1 Results

The results are presented in Tables 1 and 2, for the development and the official test phase, respectively. The development tests are conducted using 10-fold cross validation for each classifier. The results for title detection are reported using F1 measure. The results for TOC generation are reported using *harmonic mean (Inex F1, Inex Lvl Acc)*. The evaluation script is provided by the organizers.

In the official evaluation for the title detection task, our team was placed 7th for the English subset and 4th for the French subset respectively. In the TOC generation task we were placed 1st for both subsets (Table 2).

Language	**Title detection***	**Placement**	**TOC**	**Placement**
English	0.59	7	0.34	1
French	0.64	4	0.37	1

Table 2: Official results.

Because our best model in all cases was Random Forest, we used feature permutation importance technique (Breiman, 2001) to explore the most important features. For the title detection task, for both languages we found that the top three features are:

1. **top distance:** 0.0295 (±0.0003)
2. **boldness:** 0.0247 (±0.0002)
3. **width:** 0.0162 (±0.0003)

The general idea behind this technique is to see how much the performance of the classifier decreases when a feature of interest in not available (that is, contains a random noise instead of a true sample value). The most influential features in the title detection task are very intuitive: the distance of the text from the closest upper neighbouring text, estimation of the boldness and width of the bounding box.

The results show a significant difference in the performance between English and French subsets. The choice of feature set is the main source of it. Namely, English language subset had only 64 features in total, while French had 200. Note that the *PoS tag* features are language specific. The French language, besides general labels for part-of-speech tags, has gender, number, verb form that are absent in English language. Additionally, the PoS tagging tool itself did not have more granular labels for English. To validate this claim, we run title detection on French subset and removed gender, number, verb form and other additional information and kept only the PoS tag labels. 10-fold cross validation in this case was 0.74 (+/- 0.04), which is in the same range as the English, and we observe a significant drop in performance from the original experiment for French.

4 Conclusion

In this paper, we presented our work as part of the 2020 FinToc task for title detection and table-of-contents generation. We also demonstrated that using *tesseract* in conjunction with Levenstein distance and a set of key features returned a harmonic mean of 0.34 and 0.37 for English and French, respectively, earning us 1st place for table-of-contents generation. Our method also returned F1 measures of 0.59 and 0.64 for English and French, respectively. While previous researchers have used methods primarily focused on PDF documents, our approach used OCR scanning in order to include scanned documents in addition to PDFs. While document storage practices have evolved over time with PDF being the *standard*, this has not always been the case. Many older documents were scanned and are archived in that format. Employing an approach that considers this evolution, as we did, makes our method more widely applicable to electronic documents.

In the future, we would like to apply deep learning methods to both title detection and table-of-contents generation, as this approach is currently not widely researched. Another area of interest would be in embedding information pointers in documents to help better identify titles, sections, headers, and sub-headers. Finally, we would like to experiment using computer vision for title detection and use it to guide better generation of tables-of-contents.

References

Hanna Abi Akl, Anubhav Gupta, and Dominique Mariko. 2019. Fintoc-2019 shared task: Finding title in text blocks. In *Proceedings of the Second Financial Narrative Processing Workshop (FNP 2019)*, pages 58–62.

Carla Abreu, Henrique Cardoso, and Eugénio Oliveira. 2019. FinDSE@ FinTOC-2019 shared task. In *Proceedings of the Second Financial Narrative Processing Workshop (FNP 2019)*, pages 69–73.

Najah-Imane Bentabet, Rémi Juge, and Sira Ferradans. 2019. Table-of-contents generation on contemporary documents. *arXiv preprint arXiv:1911.08836.*

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.

Leo Breiman. 2001. Random forests. *Machine Learning*, 45(1):5–32, October.

Mahmoud El-Haj, Paul Rayson, Paulo Alves, Carlos Herrero-Zorita, and Steven Young. 2019. Multilingual financial narrative processing: Analysing annual reports in English, Spanish and Portuguese. *Multilingual Text Analysis: Challenges, Models, And Approaches*, page 441.

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.

Abhijith Athreya Mysore Gopinath, Shomir Wilson, and Norman Sadeh. 2018. Supervised and unsupervised methods for robust separation of section titles and prose text in web documents. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 850–855.

Pavel Král Tomas Hercig. 2019. UWB@ FinTOC-2019 shared task: Financial document title detection. In *Proceedings of the Second Financial Narrative Processing Workshop (FNP 2019)*, pages 74–78.

Anthony Kay. 2007. Tesseract: An open-source optical character recognition engine. *Linux J.*, 2007(159):2, July.

Thi-Tuyet-Hai Nguyen, Antoine Doucet, and Mickaël Coustaty. 2017. Enhancing table of contents extraction by system aggregation. In *14th IAPR International Conference on Document Analysis and Recognition, ICDAR 2017, Kyoto, Japan, November 9-15, 2017*, pages 242–247. IEEE.

Ke Tian and Zijun Peng. 2019. Finance document extraction using data augmentation and attention. In *Proceedings of the Second Financial Narrative Processing Workshop (FNP 2019), September 30, Turku Finland*, number 165, pages 1–4. Linköping University Electronic Press.