FinDSE@FinTOC-2019 Shared Task

Carla Abreu[1,2], **Henrique Lopes Cardoso**[1,2], **Eugénio Oliveira**[1,2]
Faculdade de Engenharia da Universidade do Porto[1], LIACC[2]
Rua Dr. Roberto Frias, s/n, Porto, Portugal
(ei08165, hlc, eco)@fe.up.pt

Abstract

We present the approach developed at the Faculty of Engineering of the University of Porto to participate in FinTOC-2019 Financial Document Structure Extraction – Detection of titles sub-task. Several financial documents are produced in machine-readable format. Due to the poor structure of these documents, it is an arduous task to retrieve the desired information from them. The aim of this sub-task is to detect titles in this kind of documents. We propose a supervised learning approach making use of linguistic, semantic and morphological features to classify a text block as title or non title. The proposed methodology got a F1 score of 97.01%.

1 Introduction

Several financial documents are produced, every day, for different financial applications. Some of these documents are mandatory by law, however they are not created following the same standard and sometimes have a poor structure, making it difficult to retrieve the desired information. These documents are usually published in machine-readable format (such as Portable Document Format (PDF) files) but unfortunately, they remain untagged – they have no tags for identifying layout items such as paragraphs, columns, or tables. Document structuring has clear benefits to users, enabling them to gain direct access to the relevant part of the document (which can be lengthy), improving also search performance.

Financial Prospectuses are financial documents where investment funds are described, and have a non-standard content format. These documents need to be consulted by distinct persons and fast retrievals of data are desired.

A lot of effort has already been put to label the structure of documents. Some known projects are the Million Book project (Linke, 2003), the Open Content Alliance (OCA) (Suber, 2005), or the digitisation of Google (Coyle, 2006) (Doucet et al., 2011). Projects that have aim at automatically recognizing document structure take, as input, a document in PDF format, or its content obtained via Optical Character Recognition (OCR).

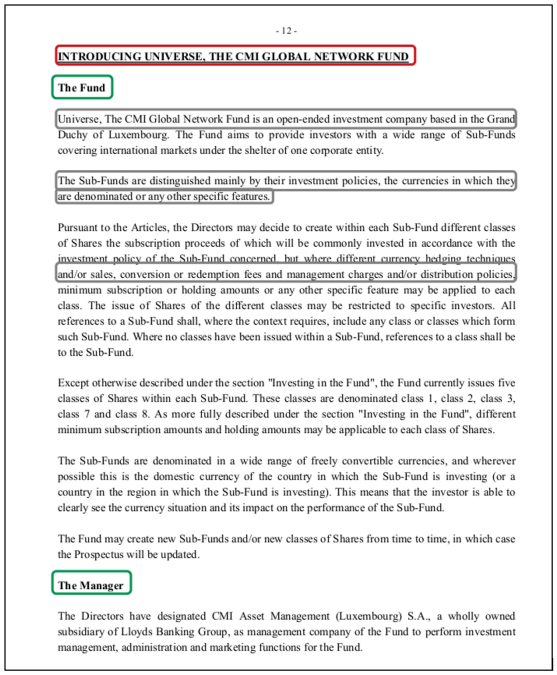
Document structure extraction is a well studied problem in document analysis, and has been applied in distinct types of documents and in different domains. Works on this matter go from scientific articles (Klampfl et al., 2014) (Bast and Korzen, 2017) to books (Linke, 2003).

Rangoni et al. (Rangoni et al., 2012) make use of three types of features: geometrical (width, height, X position, among others), morphological (the font and other characteristics, such as italics, bold, and so on) and semantic (language, is numeric, and so on). Bitew (Bitew, 2018) also includes three distinct categories: textual features (similar to semantic), markup features (similar to morphological) and linguistic (related with Part of Speech). As described, some authors groups features in categories; however, some studies use only one category, including Kim et al. (Kim et al., 2017), who make use of morphological elements only for logical structured extraction.

The methodologies used to address this problem include rule-based and machine learning approaches (Klampfl and Kern, 2013) (He, 2017).

In this paper we present a supervised approach to automatically classify a text block as title or non title (a binary classification problem), making use of linguistic, semantic and morphological features. In Section 2, we describe the FinTOC Sub-Task on title detection, and in Section 3 we analyze the provided data. In Section 4 we present our approach, followed by the experimental setup in Section 5. Results are discussed in Section 6. In Section 7 we conclude.

Proceedings of the Second Financial Narrative Processing Workshop (FNP 2019), pages 69–73
Turku, Finland, 30 September,



- 12 -

INTRODUCING UNIVERSE, THE CMI GLOBAL NETWORK FUND

The Fund

Universe, The CMI Global Network Fund is an open-ended investment company based in the Grand Duchy of Luxembourg. The Fund aims to provide investors with a wide range of Sub-Funds covering international markets under the shelter of one corporate entity.

The Sub-Funds are distinguished mainly by their investment policies, the currencies in which they are denominated or any other specific features.

Pursuant to the Articles, the Directors may decide to create within each Sub-Fund different classes of Shares the subscription proceeds of which will be commonly invested in accordance with the investment policy of the Sub-Fund concerned, but where different currency hedging techniques and/or sales, conversion or redemption fees and management charges and/or distribution policies, minimum subscription or holding amounts or any other specific feature may be applied to each class. The issue of Shares of the different classes may be restricted to specific investors. All references to a Sub-Fund shall, where the context requires, include any class or classes which form such Sub-Fund. Where no classes have been issued within a Sub-Fund, references to a class shall be to the Sub-Fund.

Except otherwise described under the section "Investing in the Fund", the Fund currently issues five classes of Shares within each Sub-Fund. These classes are denominated class 1, class 2, class 3, class 7 and class 8. As more fully described under the section "Investing in the Fund", different minimum subscription amounts and holding amounts may be applicable to each class of Shares.

The Sub-Funds are denominated in a wide range of freely convertible currencies, and wherever possible this is the domestic currency of the country in which the Sub-Fund is investing (or a country in the region in which the Sub-Fund is investing). This means that the investor is able to clearly see the currency situation and its impact on the performance of the Sub-Fund.

The Fund may create new Sub-Funds and/or new classes of Shares from time to time, in which case the Prospectus will be updated.

The Manager

The Directors have designated CMI Asset Management (Luxembourg) S.A., a wholly owned subsidiary of Lloyds Banking Group, as management company of the Fund to perform investment management, administration and marketing functions for the Fund.

Figure 1: Financial Prospectuses document layout

2 Sub-Task Description

The task addressed in this work concerns the detection of titles in financial prospectuses (Rémi Juge, 2019). Given a set of text blocks, the goal is to classify each given text block as a 'title' or 'non-title'. As shown in Figure 1[1], titles can have different layouts (marked with red and green boxes), and they have to be distinguished from regular text ('non-title' marked with grey boxes).

The evaluation metric used in the task is the F1 metric.

3 Dataset

FinTOC organizers provide an excel file with text blocks information. Each line represents one text block and each column their characteristics:

- *text blocks*: text block textual content;
- *begins_with_numbering*: 1 if the text block begins with a numbering such as 1., A/, b), III., etc. . . .; 0 otherwise;
- *is_bold*: 1 if the text block appears in bold in the PDF document; 0 otherwise;
- *is_italic*: 1 if the text block is in italic in the pdf document; 0 otherwise;

[1]FNP Workhop Series – Title detection subtask: http://wp.lancs.ac.uk/cfie/shared-task/

	Title	Non Title
Excel - Number of Rows	2,420	13,092
Average of NC / TB	25.03	152.82
Standard Deviation of NC / TB	14.87	300.87
Variance of NC / TB	221	90,525.25
Min of NC / TB	2	1
Max of NC / TB	143	7,715

Table 1: Training set statistics (TB = text block; NC = number of characters).

- *is_all_caps*: 1 if the text block is all composed of capital letters; 0 otherwise;
- *begins_with_cap*: 1 if the text block begins with a capital letter; 0 otherwise;
- *xmlfile*: the xmlfile from which the above features have been derived;
- *page_nb*: the page number in the PDF where the text bock appears;
- *label*: 1 if text block is a title, 0 otherwise.

The test set has the same format as the training set, but without information in the last column of the CSV file. This column is meant to be filled in by systems participating in the task.

The training set contains 44 distinct documents, not standardized. The CSV file used as training set contains 75625 annotated rows. More details about the training set are included on Table 1 and Table 2.

The test set is composed of 7 PDF files (whose length ranges from 35 to 134 pages, with an average of 64 pages). The CSV file is composed of 14816 non-annotated rows.

4 Proposed approach

4.1 Features

Text blocks are provided with some characteristics, such as: (Fe1) begins_with_numbering; (Fe2) is_bold; (Fe3) is_italic; (Fe4) is_all_caps; and (Fe5) begins_with_cap. These elements are described in Section 3.

We have extracted additional features from the text block, as follows:

- (Fe6) Number of characters;
- (Fe6a) Number of characters distributed in categories (Table 2) ;
- (Fe7) First block character type: alphabetic upper/lower, numeric, other (space or punctuation);

$ID_{Category}$	Range	TS_{Title}	$TS_{NonTitle}$
0	0	0	0
1	1 - 3	2	334
2	4 - 9	80	48
3	10 - 16	695	203
4	17 - 21	553	160
5	22 - 30	400	87
6	31 - 40	366	73
7	41 - 50	164	48
8	51 - 70	129	108
9	71 - 100	23	157
10	101 - 150	8	178
11	151 - 200	0	181
12	201 - 400	0	523
13	401 - 600	0	172
14	601 - 1000	0	114
15	1001 - 1500	0	30
16	>1501	0	4

Table 2: Number of characters distributed in categories and in the training set (TS)

- (Fe8) Last block character type: alphabetic upper/lower, numeric, other (space or punctuation);
- (Fe9) Number of tokens;
- (Fe10) Number of sentences contained in the block text;
- (Fe11) Part Of Speech of the first token in the block text;
- (Fe12) Contains date;
- (Fe13) Title suggestion word - if the first token belongs to one of these words: appendix', 'annex', and others;
- (Fe14) Tense block - check if the text block is written in the past, present or future.

The enunciated features belong to three different types: morphological (Fe2, Fe3, Fe4, Fe5, Fe6, Fe6a, Fe7, Fe8), semantic (Fe1, Fe12), and linguistic (Fe11, Fe13, Fe14). Tense, part of speech, title suggestion words and contains date are language dependent features applied only to English language.

4.2 Classification Algorithms

Supervised learning techniques create a model that predicts the value of a target variable based on a set of input variables. One challenge is to select the most appropriate algorithm for the task of classifying as 'title' or 'non-title' a given text block. We have compared the following algorithms: Decision Tree (DT), Extra-tree classifier (EXT), and Gradient Boosting (GBC).

As shown in Table 3, different configurations were attempted for each algorithm. Implementations of these algorithms are provided by the Python library scikit-learn library[2].

Alg_{ID}	**Algorithm**	**Configuration**
DT_1	DT	random_state=0
DTC_1	DT	max_depth=None min_samples_split=2 random_state=0
EXT_1	EXT	n_estimators=10 max_depth=None min_samples_split=2 random_state=0
GBC_1	GBC	loss='exponential'
GBC_2	GBC	n_estimators=2000 learning_rate=0.75 max_depth=5

Table 3: List of algorithm configurations

	E_1	**E_2**	**E_3**	**E_4**	**E_5**
Fe1	x	x	x		x
Fe2	x	x	x	x	x
Fe3	x	x			x
Fe4	x	x	x		x
Fe5	x	x	x		x
Fe6		x	x	x	
Fe6a	x				
Fe7	x	x	x		
Fe8	x	x	x	x	
Fe9	x	x	x	x	
Fe10	x	x	x		
Fe11	x	x	x	x	
Fe12		x	x		
Fe13		x			
Fe14		x			

Table 4: List of features used in each experimental setup.

5 Experimental Setup

The set of features used in each experimental setup is shown in Table 4. Experiment 5 (E_5) is our baseline, as this setup includes all the features available in the dataset. We combine all the available features with all extracted by us in Experiment 2 (E_2). We create a model based on E_2 and select all the features with an importance above 0.03 to compose Experiment 3 (E_3) and above 0.07 to include in Experiment 4 (E_4).

Experiment 1 (E_1) was based in our analysis regarding text blocks number of characters categories distribution, such as presented in Table 2.

6 Experimental Evaluation

Several combinations of features (Table 4) and algorithms (Table 3) were applied to solve the title classification problem. The results obtained are shown in Table 5.

[2]sklearn: https://scikit-learn.org

Exp	*Alg_ID*	TN	FP	FN	TP	*F1_title*	*F1_non*
E_1	DT_1	18,882	756	705	2,345	76.25	96.28
E_1	DTC_1	18,882	756	705	2345	76,25	96,28
E_1	EXT_1	**18,932**	**706**	**684**	**2,366**	**77.30**	**96.46**
E_1	GBC_1	*14,954*	*4,684*	**629**	**2,421**	47.68	84.92
E_1	GBC_2	18,829	809	1,192	1,858	65.00	94.95
E_2	DT_1	18,851	787	747	2,303	75.02	96.09
E_2	DTC_1	18,851	787	747	2,303	75.02	96.09
E_2	EXT_1	18,891	747	741	2,309	75.63	96.21
E_2	GBC_1	18,856	782	737	2,313	75.28	96.13
E_2	GBC_2	18,816	822	1,214	1,836	64.33	94.87
E_3	DT_1	18,850	788	794	2,256	74.04	95.97
E_3	DTC_1	18,850	788	794	2,256	74.04	95.97
E_3	EXT_1	18,880	758	786	2,264	74.57	96.07
E_3	GBC_1	18,735	903	770	2,280	73.16	95.73
E_3	GBC_2	18,813	825	1,225	1,825	64.04	94.83
E_4	DT_1	18,801	837	848	2,202	72.33	95.71
E_4	DTC_1	18,801	837	848	2,202	72.33	95.71
E_4	EXT_1	18,810	828	847	2,203	72.46	95.74
E_4	GBC_1	18,798	840	877	2,173	71.68	95.63
E_4	GBC_2	18,739	899	1,208	1,842	63.62	94.68
E_5	DT_1	**19,280**	**358**	*2,328*	*722*	34.96	93.49
E_5	DTC_1	**19,280**	**358**	*2,328*	*722*	34.96	93.49
E_5	EXT_1	**19,280**	**358**	*2,328*	*722*	34.96	93.49
E_5	GBC_1	**19,280**	**358**	*2,328*	*722*	34.96	93.49
E_5	GBC_2	**19,280**	**358**	*2,329*	*721*	34.96	93.49

Table 5: Results

E_5 is the experiment that has as feature set all the features available upfront with the dataset. This experiment got similar results using distinct supervised learning algorithms. The results obtained indicate that this set of features are not enough to classify block text titles, showing a high number of false negatives and a low number of true positives.

The DT_1 and DTC_1 algorithms have distinct configurations, however they presented the same results when exposed to the same feature set. The GBC_1 algorithm configuration was more sensible when exposed to a specific feature set – in E_1, this algorithm has shown the higher number of false positives obtained in our experiments. GBC_2 was the worst configuration algorithm used in this classification, having the lowest value of true positives.

The feature set used in E_1 includes all features provided by the competition organizers. Other features were added, some of them related to how the text appears in the text block (such as number of characters or sentences), and also language dependent features (such as the case of F11). Except for GBC_1, all other algorithm configurations reached their best result. EXT_1 got the best performance in the task of title classification.

FinTOC-2019 received two submissions for each participant, on which we achieved F1 score of 97.01% on E_1 with EXT_1 reaching the fifth position and the sixth position with F1 score of 96.84% on E_1 with DT_1.

7 Conclusion

It is difficult to retrieve the desired information from lengthy documents when the Table Of Content (TOC) is missing. TOC helps the reader to identify what is written in each section, enabling an oriented reading. The aim of this study is to classify each text block into title or non-title, a step towards identifying each section in a document.

In this work we propose a supervised learning strategy to classify text blocks. We also proposed an extension of the provided feature set based on recognizing new characteristics of text blocks (related with the text block composition and the use of linguistic resources). The dataset available in this competition was composed by five features. We experimented the use of these features but the results obtained point out that these are not enough to the envisaged classification task.

We recognize more features in text blocks, some of them related with the text composition and others related with linguistic resources. Not all of these features have shown to be essential for title classification.

Title detection got an high performance using Extra-Tree classifier with the following features: the five ones available on the dataset (*begings_with_numbering*, *is_bold*, *is_italic*, *is_all_caps*, *begin_with_cap*) and six more (number of characters, first sentence character, last sentence character, number of tokens, number of sentences, Part of speech of the first sentence element).

References

Hannah Bast and Claudius Korzen. 2017. A benchmark and evaluation for text extraction from pdf. In *Proceedings of the 17th ACM/IEEE Joint Conference on Digital Libraries*, pages 99–108. IEEE Press.

Semere Kiros Bitew. 2018. Logical structure extraction of electronic documents using contextual information. Master's thesis, University of Twente.

Karen Coyle. 2006. Mass digitization of books. *The Journal of Academic Librarianship*, 32(6):641–645.

Antoine Doucet, Gabriella Kazai, and Jean-Luc Meunier. 2011. Icdar 2011 book structure extraction competition. In *2011 International Conference on Document Analysis and Recognition*, pages 1501–1505. IEEE.

Yi He. 2017. Extracting document structure of a text with visual and textual cues. Master's thesis, University of Twente.

Tae-young Kim, Suntae Kim, Sangchul Choi, Jeong-Ah Kim, Jae-Young Choi, Jong-Won Ko, Jee-Huong Lee, and Youngwha Cho. 2017. A machine-learning based approach for extracting logical structure of a styled document. *TIIS*, 11(2):1043–1056.

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.

Stefan Klampfl and Roman Kern. 2013. An unsupervised machine learning approach to body text and table of contents extraction from digital scientific articles. In *International Conference on Theory and Practice of Digital Libraries*, pages 144–155. Springer.

Erika C Linke. 2003. Million book project. *Encyclopedia of Library and Information Science: Lib-Pub*, page 1889.

Yves Rangoni, Abdel Belaïd, and Szilárd Vajda. 2012. Labelling logical structures of document images using a dynamic perceptive neural network. *International Journal on Document Analysis and Recognition (IJDAR)*, 15(1):45–55.

Sira Ferradans Rémi Juge, Najah-Imane Bentabet. 2019. The fintoc-2019 shared task: Financial document structure extraction. In *The Second Workshop on Financial Narrative Processing of NoDalida 2019*.

Peter Suber. 2005. The open content alliance. *SPARC Open Access Newsletter*.