Daniel@FinTOC '2 Shared Task: Title Detection and Structure Extraction

Emmanuel Giguet[1]
Normandie Univ, UNICAEN,
ENSICAEN, CNRS, GREYC
14000 Caen, France

Gaël Lejeune[2]
STIH, EA 4509
Sorbonne University
75006 Paris, France

Jean-Baptiste Tanguy[2]
OBVIL/STIH, EA 4509
Sorbonne University
75006 Paris, France

[1]`firstname.lastname@unicaen.fr` [2]`firstname.lastname@sorbonne-universite.fr`

Abstract

We present our contributions for the FinTOC'2 Shared Tasks: Table of Content (ToC) extraction in English and French documents. For ToC Extraction, we propose to combine information from multiple sources: ToC itself, wording of the document, and lexical domain knowledge. For title detection, we compare surface features to character-based features on various training configurations. We show that title detection results are very sensitive to the training dataset used.

1 Introduction

The Fintoc'2 Financial Document Structure Extraction competition (Bentabet et al., 2020) proposed to evaluate two tasks : *Title Detection* and *ToC Structure Extraction*. Structure Extraction is an important issue for Natural Language Processing and Document Analysis. Rich logical structures can be exploited for document classification and clustering (Doucet and Lehtonen, 2007; Ait Elhadj et al., 2012). In the Document Analysis field, ToC generation aims to retrieve or extract a ToC from documents where the logical structure is not explicitly marked. ToC makes it easier to access information, in particular in Digital Humanities where documents can be long and structured in parts, chapters, appendices. Title Detection plays an important role for extracting the structure by helping to get candidates to populate the ToC. Furthermore, the position of sentences with respect to titles is used to improve the results in some NLP tasks: text classification (Lejeune et al., 2013), Terminology Acquisition (Daille et al., 2016) or Keyphrase Extraction (Florescu and Caragea, 2017). The paper is organized as follows. Section 2 gives a quick background for both subtasks. Section 3 describes our contribution to *ToC Extraction* and Section 4 our contribution to *Title Detection*. We give some words of conclusion in Section 5.

2 Background

Usually, the logical structure of natural language data is not explicitly encoded within a PDF document, it is the case for most financial prospectuses. The organization of the information has to be inferred from the layout and the style of text blocks. Positional and contrastive features allows the recovery of the underlying structure. Recovering the global structure of a document is an important process to achieve for information extraction. In this regard, document structure analysis certainly precedes sentence analysis. By the way, neither one of them can be seen as preprocessing stage. Both are fully part of a natural language processing system. These processes relate to *skimming* and *scanning* reading techniques. While skimming allows a reader to get a first glance of a document, scanning is the process of searching for a specific piece of information. Different parts of the document may be spotted by the reader and sought for specific information using a zoom-in/zoom-out strategy (Andrew et al., 2019). Concerning global structure, important information is found in the titles and subtitles, making the detection of titles important for improving web indexation (Changuel et al., 2009) or downstream NLP tasks (Huttunen et al., 2011; Tkaczyk et al., 2018). We can see two main strategies for ToC extraction: detecting the ToC pages and relying on the book content. The ICDAR Book Structure Extraction competitions results (Doucet et al., 2013) showed that hybrid systems are promising which is consistent with more recent results from (Nguyen et al., 2017) who combined different systems to get better results.

Proceedings of the 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation, pages 174–180
Barcelona, Spain (Online), December 12, 2020.

3 Contribution to the ToC Extraction Shared Task

3.1 From Table of Content Extraction to Document Structure Extraction

In previous INEX Book Structure Extraction Competitions, we used to consider the whole wording of the document (Giguet and Lucas, 2010a; Giguet and Lucas, 2010b; Giguet et al., 2009). This is a minority approach, it is more common to rely on the recognition and the parsing of the ToC since most books contain one that is usually quite easy to locate. Taking into account the whole wording of the document presents several advantages. First, it allows to consistently handle documents with and without ToC. Second, it permits to extract titles that are not included in the ToC, such as lower-level titles or preliminary titles. Third, it avoids having to process erroneous ToCs. Indeed, the ToC of a document may not be accurately synchronized if the authors forgot to update it. It may also contain entries that are not titles, for instance a paragraph incorrectly labelled as a title, or wrong page numbers. These cases often occur when documents are published without the supervision of an editorial board.

In FinTOC'1 (Giguet and Lejeune, 2019) our strategy relied on the detection of the ToC combined to a simple fallback strategy when no ToC is found. Our expectations was to have a good precision and a low recall due to missing or incomplete ToCs. ToCs belong to the category of index lists like list of figures or list of tables. Index lists together form a network of links starting from the periphery and pointing to the inner content. These links facilitate direct access to information and enable alternative reading strategies. They provide an "at-a-glance" snapshot of the complexity of the structure. While Document Structure Extraction do not consist in ToC extraction, it would be unfortunate to get rid of the information contained in the ToCs. Therefore, ToC recognition and parsing is integrated to our extraction method. Linking ToC entries to headings of the main text stream is the first step of our integration process.

3.2 Title Extraction from the Whole Content

As documents do not all contain ToCs, alternative ways have to be found to capture the hierarchy of headings. The main stream of content is the most natural source of information. However, it needs to be accurately and reliably detected. The task is not straightforward since the content is fragmented into pieces of texts. In order to retrieve the main text stream from the document content, page layouts have to be inferred in order to exclude the headers and the footers which break the linearity of the main text stream. Floating objects such as figures, tables, graphics and framed texts have to be excluded as well.

With financial documents, we solely focus on table detection and removal: they are the most frequent floating objects. Table removal reduces the search space and prevents considering table content when searching title candidates, thereby reducing the number of false positives. The table detection module parses the PDF vectorial shapes that are extracted by the `pdf2xml` command (Déjean, 2007). Text background and framed content are first inferred. The algorithm then builds table grids from adjacent framed content interpreted as possible table cells. Once the main text stream is extracted, titles are located with the help of two complementary strategies: Numbered List Detection and Salient Text Detection.

The *Numbered List Detection* strategy detects coherent series of numbered lines that may correspond to numbered titles. We exploit various features of text lines: the numbering style type (i.e., decimal, lower/upper-latin, lower/upper-roman), the numbering pattern (e.g., prefixes such as Chapter, Section, hierarchical numbering system such as A.2, 3.1.b). The *Text Saliency Detection strategy* is a contrastive approach to title detection. Titles are salient objects that stands out from the surrounding background. The background corresponds to text blocks (i.e., paragraphs, list items) that share common stylistic properties (i.e., font properties, line spaces, background, alignment). Title candidates are searched among the salient remaining text lines. A text line is salient if its stylistic properties generates enough contrasts with the surrounding text background. Salient texts sharing identical stylistic properties are clustered in order to build sets of titles acting at the same level of the hierarchy.

3.3 Taking advantages of Prospectus Document Model and Specific Document Models

The structure of financial prospectuses is driven by strong expectations from potential buyers and authorities. Thus, a model tends to emerge among the producers of financial information: across organizations, prospectuses tend to share common features in terms of macro structure. Certain sections and subsections

	Xerox measures				Inex08 measures					Error count		
	P	R	F1	Title	P	R	F1	Title	Level	Pb Title	Pb Level	Err
French	89.6	53.6	64.4	62.0	40.7	25.7	30.5	45.4	9.3	1765	3061	461
English	89.8	63.9	70.3	68.8	50.0	35.8	39.7	54.5	29.9	2713	4256	974

Table 1: Results obtained on the train dataset

Team	Inex F1
DNLP	0.37
taxy.io	0.32
Baseline	0.32
Daniel 2	0.22

(a) Results for the test dataset (French)

Team	Inex F1
DNLP	0.34
Daniel 2	0.28
taxy.io	0.24
Amex 1	0.23
Baseline	0.18

(b) Results for the test dataset (English)

Table 2: Official results (Inex F1) for the ToC extraction task

are expected and are present with an expected naming. Moreover, prospectuses issued by an organization tend to share the same structure over products and over years. Moreover they also often share exact or similar document or page layout models. It is interesting to take benefit of these features, whether they are related to the genre or related to a specific organization. In this regard, titles from the training set are stored as lexical entries and reused as an external knowledge source for title detection.

3.4 Results

Table 1 exhibits the results obtained on the train dataset. We performed better than our first contribution (Giguet and Lejeune, 2019) which mainly relied on the ToC detection and extraction. The current version still demonstrates good precision and considerably improves recall. The choice to favour precision is much more sensitive with the Xerox metrics than with the Inex08 metrics. Our results on the test set are given in Table 2. One can see that, contrary to other systems, it performed better on the English data than on the French data. This in accordance with the results obtained on the train set.

4 Contribution to the Title Detection Shared Task

4.1 Datasets

The training and testing sets of the shared task are composed of segments labelled *Title* or *Not Title*. We also used the Fintoc-2019 (https://wp.lancs.ac.uk/cfie/shared-task/) corpus for improving English title detection and the DEFT-2011 (https://deft.limsi.fr/2011/) for French. In order to observe the impact of adding training data, we kept the split in training and testing sets. Both the FinTOC-2019 and the DEFT-2011 corpora are made up of segments labelled as *Title* or *Not Title*. A segment in the FinTOC-2019 corpus (as in the Fintoc-2020 corpus) refers to a physical component that is, in practice, a line. In the DEFT-2011 corpus, a segment refers to a logical component : a Title, a section, a subsection, etc. We must be clear that DEFT-2011's documents are scientific papers. We experiment several training sets combinations in order to assess the impact of the language and the text genre. To test the models, we selected randomly 20% of each dataset to use it as development data (see details in Appendices).

4.2 Methods: surface features and character n-grams

Baselines Three group of features are used to get six baselines (Table 3): **basic** (five booleans: IsBold, IsItalic, Is AllCaps, BeginsWithCaps and BeginsWithNumber and the page number, the **length** (in characters) and **stylo** (frequency of punctuation signs, numbers and capitalized letters).

***n*-gram method** It consists in vectorizing the segments by counting the frequency of character n-grams. We explore different values for n_{min} and n_{max} minimum and maximum size of the n-grams.

For both methods we use a Random Forest with 50 estimators since it outperformed other classifiers.

B1	basic
B2	basic + length
B3	stylo
B4	stylo + basic
B5	stylo + length
B6	basic + stylo + length

Table 3: Baselines features description.

4.3 Results

	(i)				(ii)				(iii)				(iv)		
	P	R	F		P	R	F		P	R	F		P	R	F
B1	.628	.634	.631	B1	.627	.631	.629	B1	.802	.641	.689	B1	.819	.642	.694
B2	.586	.727	.614	B2	.586	.726	.615	B2	.744	.789	.764	B2	.814	.763	.786
B3	.585	.656	.607	B3	.585	.655	.607	B3	.720	.612	.646	B3	.809	.570	.608
B4	.581	.734	.608	B4	.588	.751	.617	B4	.797	.792	.795	B4	.849	.777	.808
B5	.592	.787	.624	B5	.593	.790	.625	B5	.768	.728	.746	B5	.840	.675	.729
B6	**.639**	**.801**	**.684**	B6	**.643**	**.808**	**.688**	B6	**.821**	**.855**	**.837**	B6	**.881**	**.822**	**.849**

Table 4: Baseline results on the F-2020-en-dev dataset, learned from: (i) F-2019-en-TRAIN, (ii) F-2019-en-TRAIN+TEST, (iii) F-2019-en-TRAIN+TEST + F-2020-en-TRAIN and (iv) F-2020-en-TRAIN

	(i)				(ii)				(iii)				(iv)		
	P	R	F-m		P	R	F-m		P	R	F-m		P	R	F-m
B1	.460	.500	.479	B1	.460	.500	.479	B1	.812	.724	.759	B1	.811	.724	.758
B2	.579	.658	.592	B2	.590	.679	.606	B2	.813	.772	.791	B2	.828	.769	.794
B3	.572	.681	.572	B3	.578	.699	.581	B3	.821	.650	.699	B3	.846	.633	.683
B4	.579	.672	.590	B4	.581	.678	.593	B4	.857	.822	.838	B4	.881	.821	.848
B5	.584	**.699**	.594	B5	.590	**.714**	.602	B5	.781	.744	.761	B5	.839	.710	.756
B6	**.589**	.682	**.606**	B6	**.591**	.687	**.608**	B6	**.873**	**.844**	**.858**	B6	**.888**	**.845**	**.865**

Table 5: Baseline results on the F-2020-fr-dev dataset, learned from: (i) D-2011-TRAIN, (ii) D-2011-TRAIN+TEST, (iii) D-2011-TRAIN+TEST + F-2020-fr-TRAIN and (iv) F-2020-fr-TRAIN

From the results obtained with the baselines (Table 4 for English and Table 5 for French) we observe that B6 gives the best results in all cases. This result is in accordance with the observations we made in previous edition (Giguet and Lejeune, 2019). Regarding the training datasets, we can observe that the Fintoc-2020 train set gives the best results. We observe the same pattern for the character n-gram method (heatmaps are given in appendixes), the F score does not achieve 80% without this data. We also observed that using bilingual datasets does not improve results (see appendixes).

5 Conclusion

In this article we proposed approaches for two shared tasks of FinTOC 2020. Regarding the ToC Extraction Shared Task, we propose a hybrid approach. It consists in combining the output of multiple modules dedicated or related to title detection. Title candidates are extracted from the table of contents thanks to a ToC Detection and Extraction module. Candidates are also extracted from the main text stream with the help of two complementary modules: a Numbered List detection module and a Text Saliency Detection module. In order to enrich the approach, titles from the training set are used to detect domain-specific titles. The title candidates are merged to generate a complete Table of Contents. For the Title Detection task, we proposed to use two types of features: surface features and character n-grams. We showed that stylometric features (frequency of punctuation, numbers and capitalized letters) combined with visual characteristics (bold, italic…) achieve better results than the character n-gram approaches.

References

Ali Ait Elhadj, Mohand Boughanem, Mohamed Mezghiche, and Fatiha Souam. 2012. Using structural similarity for clustering XML documents. *Knowledge and Information Systems*, 32(1):109–139, juillet.

Judith Jeyafreeda Andrew, Stéphane Ferrari, Fabrice Maurel, Gaël Dias, and Emmanuel Giguet. 2019. Model-driven Web Page Segmentation for Non Visual Access. In *16th International Conference of the Pacific Association for Computational Linguistics (PACLING 2019)*, Hanoï City, Vietnam, October.

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.

Sahar Changuel, Nicolas Labroche, and Bernadette Bouchon-Meunier. 2009. A general learning method for automatic title extraction from html pages. In Petra Perner, editor, *Machine Learning and Data Mining in Pattern Recognition*, pages 704–718, Berlin, Heidelberg. Springer Berlin Heidelberg.

Béatrice Daille, Evelyne Jacquey, Gaël Lejeune, Luis Felipe Melo, and Yannick Toussaint. 2016. Ambiguity Diagnosis for Terms in Digital Humanities. In *Language Resources and Evaluation Conference*, Portorož, Slovenia, May.

Hervé Déjean, 2007. *pdf2xml open source software*. Last access on July 31, 2019.

Antoine Doucet and Miro Lehtonen. 2007. Unsupervised classification of text-centric xml document collections. In *Comparative Evaluation of XML Information Retrieval Systems, Fifth International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2006*, volume 4518 of *Lecture Notes in Computer Science*, pages 497–509. Springer.

Antoine Doucet, Gabriella Kazai, Sebastian Colutto, and Günter Mühlberger. 2013. Overview of the ICDAR 2013 Competition on Book Structure Extraction. In *Proceedings of the Twelfth International Conference on Document Analysis and Recognition (ICDAR'2013)*, pages 1438–1443, Washington DC, USA, August.

Corina Florescu and Cornelia Caragea. 2017. PositionRank: An unsupervised approach to keyphrase extraction from scholarly documents. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1105–1115, Vancouver, Canada, July. Association for Computational Linguistics.

Emmanuel Giguet and Gaël Lejeune. 2019. Daniel@fintoc-2019 shared task: Toc extraction and title detection. In *The Second Financial Narrative Processing Workshop (FNP 2019)*, pages 63–68, Turku, Finland, September.

Emmanuel Giguet and Nadine Lucas. 2010a. The book structure extraction competition with the resurgence software at caen university. In Shlomo Geva, Jaap Kamps, and Andrew Trotman, editors, *Focused Retrieval and Evaluation*, pages 170–178, Berlin, Heidelberg. Springer Berlin Heidelberg.

Emmanuel Giguet and Nadine Lucas. 2010b. The book structure extraction competition with the resurgence software for part and chapter detection at caen university. In *Comparative Evaluation of Focused Retrieval - 9th Workshop of the Initiative for the Evaluation of XML Retrieval (INEX 2010), Revised Selected Papers*, pages 128–139.

Emmanuel Giguet, Alexandre Baudrillart, and Nadine Lucas. 2009. Resurgence for the book structure extraction competition. In Shlomo Geva, Jaap Kamps, and Andrew Trotman, editors, *INEX 2009 Workshop Pre-Proceedings*, pages 136–142.

Silja Huttunen, Arto Vihavainen, Peter von Etter, and Roman Yangarber. 2011. Relevance prediction in information extraction using discourse and lexical features. In *Proceedings of the 18th Nordic Conference of Computational Linguistics (NODALIDA 2011)*, pages 114–121.

Gaël Lejeune, Romain Brixtel, Charlotte Lecluze, Antoine Doucet, and Nadine Lucas. 2013. Added-value of automatic multilingual text analysis for epidemic surveillance. In *Artificial Intelligence in Medicine (AIME)*, pages 284–294.

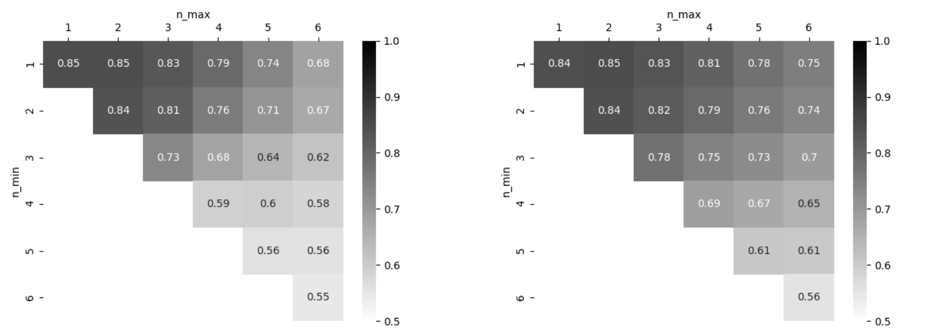
Thi-Tuyet-Hai Nguyen, Antoine Doucet, and Mickael Coustaty. 2017. Enhancing table of contents extraction by system aggregation. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 01, pages 242–247, Nov.

Dominika Tkaczyk, Andrew Collins, and Joeran Beel. 2018. Who did what?: Identifying author contributions in biomedical publications using naïve bayes. In *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries*, JCDL '18, pages 387–388, New York, NY, USA. ACM.

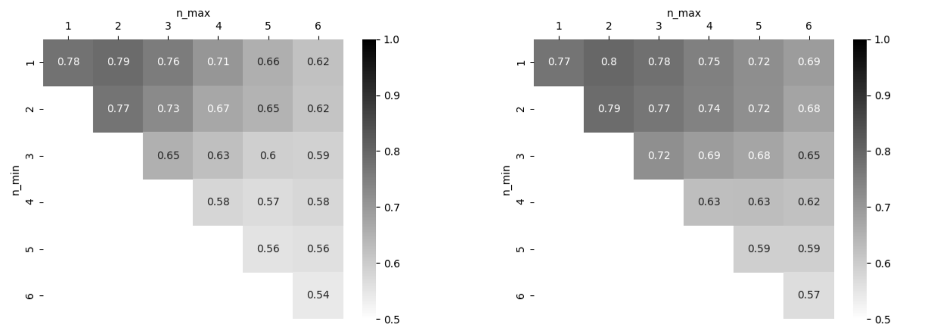
Appendices

	Lang.	Number of segments	Nb *Title*	Nb *Not Titles*
F-2019-en-TRAIN	English	75,625	10,271	65,354
F-2019-en-TRAIN+TEST	English	90,441	11,159	79,282
F-2020-en-TRAIN	English	148,940	5,463	143,477
F-2019-en-TRAIN+TEST + F-2020-en-TRAIN	English	239,381	16,622	222,759
D-2011-TRAIN	French	15,771	1,666	14,105
D-2011-TRAIN+TEST	French	22,531	2,415	20,116
F-2020-fr-TRAIN	French	54,483	4,233	50,250
D-2011-TRAIN+TEST + F-2020-fr-TRAIN	French	77,014	6,648	70,366
F-2020-en-DEV	English	37,234	1,322	35,912
F-2020-fr-DEV	French	13,620	1,076	12,544

Table 6: Size and composition of each training configuration used (F: FinTOC, D: DEFT)

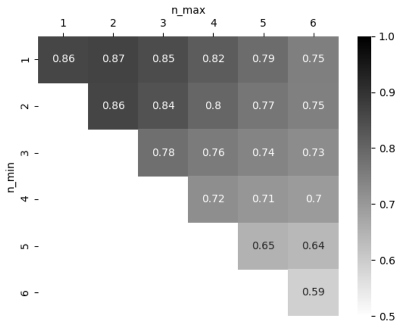


(a) F-2019-en-train+test + F-2020-en-train (abs.) (b) F-2019-en-train+test + F-2020-en-train (rel.)

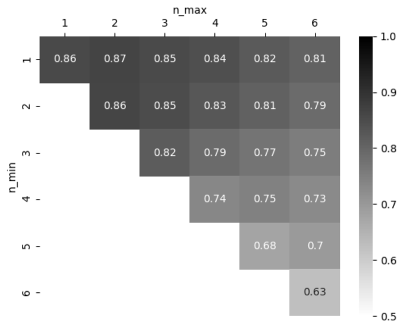


(c) F-2020-en-train (abs.) (d) F-2020-en-train. (rel.)

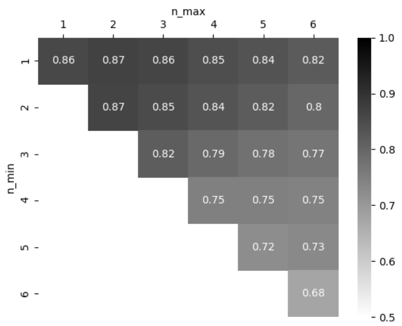
Figure 1: English n-grams models results with various training sets and absolute or relative counts



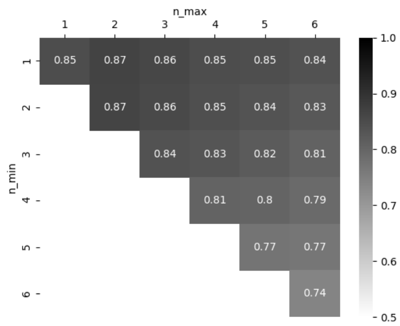
(a) D-2011-train+test + F-2020-fr-train (abs.)



(b) D-2011-train+test + F-2020-fr-train (rel.)



(c) F-2020-fr-train (abs.)



(d) F-2020-fr-train (rel.)

Figure 2: French n-grams models results with various training sets and absolute or relative counts

FinTOC-2020-en-dev

	P	R	F-m
B1	.774	.657	.698
B2	.736	.775	.754
B3	.723	.597	.633
B4	.788	.789	.789
B5	.759	.719	.737
B6	**.809**	**.839**	**.824**

FinTOC-2020-fr-dev

	P	R	F-m
B1	.801	.626	.670
B2	.770	.734	.750
B3	.806	.618	.662
B4	**.841**	.789	.812
B5	.800	.728	.758
B6	.838	**.829**	**.833**

Table 7: Results on F-2020-fr-dev of the baseline methods, learned from the bilingual training set (F-2019-en-train+test + F-2020-en-train + D-2011-train+test + F-2020-fr-train).