FinTOC2021 - Document Structure Understanding

Christopher Bourez
christopher.bourez@gmail.com

Abstract

Most machine learning applications to documents have been focused on analysing the semantic components of the texts without their formatting. Nevertheless, formatting contains information in a number of ways. First, it encodes the semantic structure of the documents and extracting the Table of Contents (TOC) is a way to summarize, categorize and search into documents. Second, machine learning models, such as for instance segmentation in computer vision, have always been imprecise at boundaries and been the subject of many years of research (conditional random fields, models in U, etc). For texts, understanding the formatting gives precise boundaries to text sections. This document discusses the different aspects of structure understanding through style formatting, and describes the method that gives state-of-art results at FINTOC 2021 challenge.

1 Introduction

Document structure understanding has specifities that need to be taken into consideration for the tasks of title detection and TOC extraction.

1.1 Emphasis and background

Annotating headers is an ambiguous task. Humans will disagree whether the underlined word "Abstract" in the first paragraph of every publication is a header or not, in particular when it is inline. Any emphasis in formatting (bold font, italics, all caps, ...) is a marker of the structure and of a higher level node in the structure tree. But despite being a higher-level node, it is ambiguous to determine if it should be included into the final TOC (Table of Contents).

Emphasis is relative and requires a contrast in style with the locally most common style, that we could call "background style", used to format the "paragraphs", the stream of characters that expresses the content with semantics and no highlight. A font size of 12 points might be an attribute of the paragraph style in one document, but of header style in another document. Frequency and contrast are at the heart of the header inference. Headers detach themselves from paragraphs using a different style, with more emphasis, most of the time.

Note also the difference between the structure tree and the TOC. The structure tree is like the HTML/Word DOM, with nodes both for headers and paragraphs. Paragraphs are found at any depth in the structure tree, for example between H1 and H2 titles, or after a H2 title. A paragraph between H1 and H2 titles in the text is a child of the H1 node as the H2 node is, and in consequence, at the same level in the structure tree as the H2 node. Although a paragraph between H1 and H2 titles will have a different level in the structure tree as a paragraph after a H2 title, in most cases both paragraphs will have the same style. Contrary to headers in the TOC, paragraphs in the structure tree do not necessarily have a different style for different depths. The TOC only includes headers and different depth in the TOC is usually linked to different styles.

1.2 Lonely headers and enumerations of headers

Headers of same level in the structure tree are usually printed in a same style. Sometimes, they are numbered with the same numbering pattern ("Article 1", "1.", "I"...), with continuous or constant numbers. Let's call them "enumerations", even when their numbering is implicit.

Among all emphasis, some headers are unique in their style. This is usually the case of the main title of the document, with an emphasis superior to all other headers. It can also be the "Abstract" header in the introduction, for example. Once the clustering into enumerations is made, it might be good to check if lonely headers are not the result of an error in formatting or numbering, which usually happens

and re-attach them to previous enumerations.

If confirmed, a lonely header is a source of uncertainty. Since it appears alone, one can only compare its emphasis with other headers' emphasis to decide of its depth in the TOC. To the contrary, enumerations are more robust, due to the fact that multiple headers in the same styling or numbering format give confirmation clues about the position of the header in the structure tree, with their "wavelength", the average number of blocks (paragraph or header) inbetween headers of the enumeration. For example, the paragraphs most of the time follow each other, while headers are usually separated by paragraphs and do not follow. So, an enumeration of blocks in the same style will be more likely to be paragraphs if they follow each other (continuous paragraph numbers), while enumeration of headers at higher level (lower depth) will be separated by more blocks.

Contrary to lonely headers, enumeration of headers is a robust concept, where multiple headers confirm the fact they are structuring the document with a particular style and belong to the same depth in the TOC or are paragraphs. It is complementary to the inference by "most of the time"/frequency described in previous section.

Reasoning with the concept of enumerations or clusters of headers is also a more interesting concept for evaluation, than indexing the depth of the header in the TOC by an absolute integer. First, any error in previous levels in the TOC (adding one more header or missing one header on the path) will trigger a different depth for all children headers. In particular, the first page of each document contains numerous "lonely headers", sub-titles, making it arbitrary to decide at which of these sub-headers we should attach the main enumeration of headers structuring the whole document. Any wrong subtitle on the path before will consider as false an enumeration of multiple child headers that structure correctly and unambiguously the document because their absolute depth relies on the previous subtitles' depth and will have changed. Absolute depth is ambiguous and not as important as recognizing the main unambiguous enumerations that structure the document.

1.3 Validation

Though a high number of headers, one might note the relatively low number of documents. Splitting the train dataset into 2 new splits, train and val and adding more features to my models gave higher and higher accuracies on the dataset, while the accuracy on the official validation dataset was not getting better. In fact, the model was probably learning the different styles in each document that were common to the two splits of the train dataset. So, the headers in the dataset are not randomly distributed headers, but headers belonging to few documents. And evaluation on the validation dataset gives an entirely new distribution to test the generalization capability of the features.

Increasing the number of documents in the dataset would require to ensure that they do not follow the same document template (same styles).

2 Method

2.1 Block extraction

ABBYY is used to convert PDF to paragraphs, in particular because of its good layout extraction technology that works well in multi-column layout as well as provides usually accurate reading order of blocks. Indeed PDF format contains only the position and styling of the characters, but do not enforce any grouping of characters into words, lines, paragraphs, and reading order of paragraphs.

ABBYY Finereader is also used for table detection to exclude paragraphs inside tables.

Paragraph attributes are:

- Page, page height, page width
- Number of lines
- Text
- Line spacing (not used, but could)
- Formatting attributes:
- Font name
- Font color
- Font weight
- Underline
- Italics
- Font size
- First line:
  - Text
  - Formatting attributes

While a paragraph might contain characters in different formatting, always the most common formatting attribute is used, meaning that first histograms are computed for each attribute and do not average the value but take the most common. Two paragraphs with most common font size 12 could have for example averaging values such as 11.3234 and 12.3545, making not much sense, losing the fact they belong to the background/paragraph style because of a formatting style change occuring in the middle of the paragraphs.

The first line attribute does not exactly exactly correspond to the first line, but to all starting characters in the first line that share the same formatting.

2.2 Predictions

A style is defined by the following features:

- Font name
- Font color
- Font weight boolean
- Underline boolean
- Italics boolean
- Font size float in pixel (converted from the point size value)
- A boolean indicating if string is all capitalized

A serialization method computes a hash of each style. Style is defined either at paragraph level or first line level.

Prediction takes as input the following features:

- Font size ratio with most common font size of the document
- Indentation of the paragraph's first line (difference between the first line's start abscissa and paragraph's abscissa), in pixels.
- The local frequency of the paragraph style, where local means on a window of 20 paragraphs after the current one.
- The local frequency of each style feature
- The global frequency of the paragraph style, where global means all paragraphs in the document
- Style feature equality between current paragraph p_n and paragraphs p_{n-2}, p_{n-1}, p_{n+1}, p_{n+2} as well as first line of paragraph p_n: a boolean for each style feature to indicate equality
- Position feature difference between current paragraph p_n and next and previous paragraphs (p_{n-1}, p_{n+1}) where position features are paragraph start abscissa, paragraph's first line start abscissa, indentation and paragraph width and are normalized to the page width
- Paragraph vertical (top and bottom) margins: distance to previous and next paragraph in the same column, if any

The local frequency features, for each style as well as for each style features, is an approximation of the concept of enumerations. The same remark is true for style feature comparison with surrounding paragraphs.

Other features, such as normalized font size, vertical margin, and global frequency of a style feature in the document are measures of the emphasis. That means that the model can consider boldness or italics as emphasis attribute only if their global frequency in the document is low. In a document completely written in italics, 'no-italics' is an emphasis attribute.

2.3 Title extraction

Once the block is predicted as header, spaces and bullet characters are stripped from the text of the block. If not void, the text of the first line is considered, otherwise the paragraph's text, and, if still void, the next paragraph's text. If the string contains a colon mark, only the text before the colon mark is kept. During extraction, we consider as "first line" not the full line, but first characters of the first line that share the same formatting.

2.4 Numbered enumeration averaging

It is a small feature that gives a gain of 1 or 2 points in accuracy. It is easy to parse the numbering at the start of each block, and build enumerations of headers with same numbering format. For blocks part of a numbering enumeration, the average prediction probability for blocks inside this enumeration instead of the prediction for the single block.

Run	Precision	Recall	F1
Training 1	0.825	0.828	0.822
Training 2	0.851	0.823	0.830

Table 1: Title detection English

Run	Precision	Recall	F1
Training 1	0.870	0.772	0.817
Training 2	0.873	0.771	0.818

Table 2: Title detection French

2.5 Training

The best working model is a XGBoost classifier of maximal depth 7. Two trainings are provided for each task: in each first training, use of the validation dataset (private 2020) to choose the datasets to train on. On the English task, the French dataset does not help, while on the French task, the English dataset does. In second training, use of all datasets, both english+french, train+validation, to provide the results, but without being able to run a validation since the validation dataset is used in training.

Results are ranked first on all tasks on FINTOC 2021. We see on these results that more data (training 2) provide better results on all tasks. During development of the features, best results are also achieved on FINTOC 2020, except for French title detection, probably due to other aspects such as block extraction, or text matching in the evaluation metric due to French accents (being French, I'm supposed to deal with that better).

3 Discussion

Despite a theory of enumerations not fully demonstrated, the local frequency of style features translates it as an approximation. The comparison to neighbor paragraphs enables to capture a local context as well, which is important to decide if it is a header.

The current metric to evaluate was developed for books, where there are only a few, usually one header per page. In the case of financial documents, there a multiple headers per page and a mis-alignment in early headers in the page will disqualify all following predicted headers, even if they were true positive. It should match headers per page, inside each page, and not consider the order to match. This will also help when the block extraction technology makes error in the block reading order.

The current metric is also comparing absolute depth described by an integer, while early titles in on the first page of the document are ambiguous and have an impact on the depth. Clustering metrics to compare the groundtruth list of enumerations and the predicted enumerations of headers (headers child of a header) will be good alternatives.

Last, it would be more comfortable if best machine learning models would be rule based and predict rules that one could understand. Though, feature importance is used to search for features, and manual rules to numbered enumerations.

References

Marco Aiello, Christof Monz, Leon Todoran, and Marcel Worring. 2002. Document understanding for a broad class of documents.

Najah-Imane Bentabet, Remi 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.

Ismail El Maarouf, Juyeon Kang, Abderrahim Aitazzi, Sandra Bellato, Mei Gan, and Mahmoud El-Haj. 2021. The Financial Document Structure Extraction Shared Task (FinToc 2021). In *The Third Financial Narrative Processing Workshop (FNP 2021)*, Lancaster, UK.

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)*, Proceedings of the Second Financial Narrative ProcessingWorkshop (FNP 2019), pages 63–68, Turku, Finland.

Yufan Guo, Roi Reichart, and Anna Korhonen. 2013. Improved information structure analysis of scientific documents through discourse and lexical constraints. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 928–937, Atlanta, Georgia. Association for Computational Linguistics.

Yufan Guo, Roi Reichart, and Anna Korhonen. 2015. Unsupervised declarative knowledge induction for constraint-based learning of information structure in scientific documents. *Transactions of the Association for Computational Linguistics*, 3:131–143.

Run	Inex08-P	Inex08-R	Inex08-F1	Inex08-Title acc	Inex08-Level acc	harm mean
Training 1	53.3	52	52.5	59	36.5	52.5
Training 2	55.4	52.6	53.6	60.3	30.6	53.6

Table 3: TOC extraction English

Run	Inex08-P	Inex08-R	Inex08-F1	Inex08-Title acc	Inex08-Level acc	harm mean
Training 1	60.8	54.3	57.3	63.5	38.7	57.3
Training 2	60.9	54.2	57.3	63.6	39	57.3

Table 4: TOC extraction French

Yutong Jiang, Ning Li, and Yingai Tian. 2019. Understanding the structure of streaming documents based on neural network. In *ICCPR '19: 8th International Conference on Computing and Pattern Recognition, Beijing, China, October 23-25, 2019*, pages 30–38. ACM.

Stefan Klampfl, Michael Granitzer, Kris Jack, and Roman Kern. 2014. Unsupervised document structure analysis of digital scientific articles. *Int. J. Digit. Libr.*, 14(3-4):83–99.