Daniel@FinTOC-2019 Shared Task : TOC Extraction and Title Detection

Emmanuel Giguet [1] Gaël Lejeune [2]

September 30th, 2019

[1] Normandie Univ, UNICAEN, ENSICAEN, CNRS, GREYC UMR 6072 – Caen, France

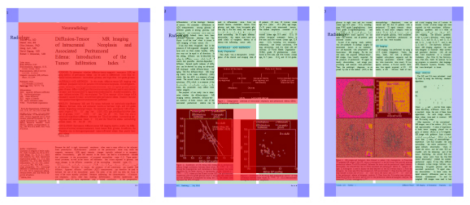
[2] Sorbonne University, STIH, EA 4509, – Paris, France

Outline

Introduction

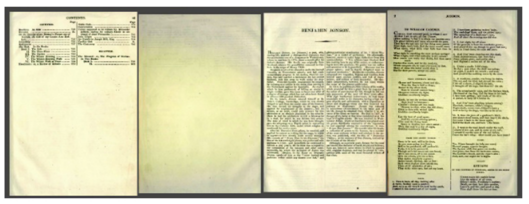
Former Related Work

1. RESURGENCE : Structure Extraction from Biomedical Articles
   - **Corpus**: 300 Biomedical Articles from Medline
   - **Documents**: 5 to 20 pages
   - **Hierarchy**: Section, subsection, subsubsection
   - **Tasks**: Document Layout Analysis, Document Structure Extraction, Information Extraction (Authors, Affiliations, Keywords, Figures, ...)



Former Related Work

2. Several Participations to INEX “Book Structure Extraction”
    - **Corpus**: 2,000 Books (Microsoft and Internet Archives)
    - **Document size**: Hundreds of pages
    - **Hierarchy**: Book, Part, Chapter, Psalm, Sonnet, Sermon, ...
    - **Task**: Document Structure Extraction from Whole Content



The FinTOC-2019 Corpus

- A few dozen of financial prospectuses
- Document size: About hundred pages
  → larger than scientific articles, smaller than books

Document characteristics

- May contain a table of contents, and parts → like books
- May contain small sections → like articles
- May contain large tables → more corpus specific

Document layout and formatting

- No Professional Editing Guidelines, no Controlled Stylesheet
- Manual Formatting instead of Styling Rules leads to inconsistencies
  - ⇒ between the ToC and the Document Structure
  - ⇒ between headings level and Formatting effects

TOC Detection

TOC Detection: Principles

- Our method is based on the ToC Detection and Analysis
  - ⇒ "Do not deny the obvious" principle :
  - ⇒ If there is a ToC, try to use it.
- And a Fallback when no ToC is found
  - ⇒ Major Headings are detected from Shallow Document Analysis
  - ⇒ We do not focus on the Whole Document Analysis, unlike our participations to Inex/ICDAR
- Our expectations: good precision and low recall
  - ⇒ Headings in the ToC are supposed to be good
  - ⇒ Some documents don't have a ToC
  - ⇒ Some headings may not be in the TOC
- The input: the raw PDF documents
  - ⇒ In order to control the whole processing chain

TOC Detection: Method Overview

1. Locating the ToC Pages
2. Building the ToC Entries
3. Inferring the Hierarchy
4. Computing PDF Page Numbers

TOC Detection: Method (I)

1. **Locating the ToC Pages** at the beginning
    - Search-space: the first third of the document
    - Invariant Pattern: A right-aligned increasing sequence of integers
    - Size: The ToC may spread on up to three contiguous pages
2. **Building the ToC Entries**: A sequential pattern
    - Toc Entry Parts: Level Number, Title*, Leader line, Page Number
    - Only the title is mandatory. It may spread over multiple lines.
    - Some title may have no Page Number $\rightarrow$ Contrast Detection based on Line Spacing and Character effects variations
3. Inferring the Hierarchy
4. Computing PDF Page Numbers

TOC Detection: Method (II)

1. Locating the ToC Pages
2. Building the ToC Entries: A sequential pattern
3. **Inferring the Hierarchy** from Contrastive Effects
    - Line spacing → Larger for major headings
    - Formatting character effects → bold, character set, font-size
    - Indentation → Positive for lower-level subheadings
    - Numbering Character Sets → Uppercase for major headings
    - Multi-level numbering structure → For lower-level subheadings
4. **Computing PDF Page Numbers**
    - Computing the shift between PDF and printed page numbers

TOC Detection: Results

	Run	F-measure
Daniel	1	42.72
IHSMarkit	1	39.41

Table 1: Results for the ToC Generation Task (test set)

Xrx-measure Links				Title	
Doc	Prec	Rec	F1	Acc	book id
0	97.7	48.6	64.9	84.5	1252823262
1	87.2	51.9	65.1	96.5	1139920265
2	22.2	40.0	28.6	91.9	0881817786
3	90.5	12.3	21.7	85.7	1150262910
4	100	10.4	18.9	42.4	0992626050
5	83.3	2.9	5.6	59.7	0949250459
6	100	12.4	22.1	94.6	1151059737

Table 2: Results for the ToC Generation Task on the test set

Title Detection

Title Detection : Corpus

	IsTitle	IsNot
# Seg.	10,271	65 354
Ratio	(**13.6%**)	86.4%
avg. std.	29.8 ±23	203.4 ±446
min:max	2 : 242	1 : 6,607

(a) Stats Train Set, size in chars

	IsTitle	IsNot
# Seg.	888	13 928
Ratio	**6%**	94%
avg.(std.) std.	32.3 ±24	98.7 ±280
min:max	5 : 232	1 : 6,586

(b) Stats Test Set, size in chars

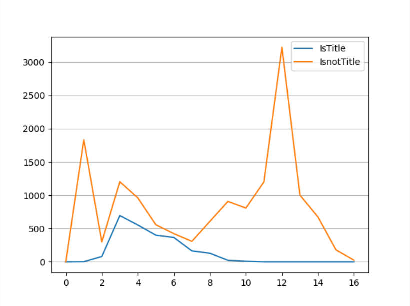


Figure 1: Number of instances with respect to their size in characters

Title Detection: method

Features (baseline)

- **basic features** BeginsWithNumbering, IsBold, IsItalic, IsAllCaps, BeginsWithCap, PageNumber
- **length** of the segment (in characters)
- **stylometry** Relative frequency of each punctuation sign, numbers and capitalized letters

Features (main system)

- Character n-grams with various sizes
- n_{min} and n_{max} in $[1:10]$
- (and $n_{min} \leq n \leq n_{max}$)

Title detection : Results (DT10 classifier)

Weighted F-1 measure

	Cross-valid	Test-set
B1 (basic features)	83.2	92.9
B2 (basic + length)	85.4	93.6
B3 (stylo)	85.4	93.2
B4 (stylo+basic)	90.4	94.2
B5 (stylo+length)	90.0	93.7
B6 (stylo+basic+length)	90.6	**95.1**
n-grams ($1 \leq n \leq 1$)	94.0	94.6
n-grams ($1 \leq n \leq 2$)	94.2	94.5
n-grams ($1 \leq n \leq 3$)	94.3	94.8
n-grams ($1 \leq n \leq 4$)	93.5	**95.0**
n-grams ($1 \leq n \leq 5$)	93.1	95.1

Title detection: Contributions

What we learned

- stylometric features worked well
- . . . and even better than character n-grams
- 1-grams were sufficient to build an efficient classifier ($> 94\%$).
- with $n_{min} > 1$ or $n_{max} > 5$ the results drop significantly

Title detection: Contributions

What we learned

- stylometric features worked well
- . . . and even better than character n-grams
- 1-grams were sufficient to build an efficient classifier ($> 94\%$).
- with $n_{min} > 1$ or $n_{max} > 5$ the results drop significantly
- Performs better on the test set (underfitting ?)
- 95% is not enough (roughly 65% on the **real task**)

Conclusion

Conclusion: Task 1

Interesting features we overlooked

- "Prefixes" : REGEX patt for first 3 chars of a line ($To9 \rightarrow Aa1$)
- "Suffixes" : REGEX patt for last 3 chars of a line (id.)
- Font Type
- Font Size

Title Detection

- **Pros**: simple method (characters and stylometry)
- **Cons**: ranked last, more feature engineering is needed
- **Future Directions**: Syntactic structure and/or LSTMs

Conclusion: Task 2

- **Pros**: Good precision, Simple and fast, Multilingual $\rightarrow$ no Lexicon
- **Cons**: Low recall (prospectuses without ToC) problem with headings not in the ToC
- **Future Directions**: Deeper Analysis of the Whole Document $\rightarrow$ not Straightforward to handle Manual Text Formatting and Unnumbered Headings
- **Open for collaborations**: Document Structure Extraction, Table Extraction, Information Extraction... from PDF documents

Comments, questions ?

Contacts

Emmanuel Giguet – Emmanuel.Giguet@unicaen.fr

Gaël Lejeune – gael.lejeune@sorbonne-universite.fr