Hierarchy Identification for Automatically Generating Table-of-Contents

Judging documents by their content!

Joint work from:

Nicolai Erbs  UKP Lab
Iryna Gurevych UKP Lab and German Institute for Educational Research
Torsten Zesch University of Duisburg-Essen
In 1964 the eminent physicist Richard Feynman served on the State of California's Curriculum Commission and saw how the Commission chose math textbooks for use in California's public schools. He discovered that some members of the commission ‘read’ and selected a book which was empty.

Don't judge a book by its cover, judge it by its table-of-contents!
Why do we need a Table-of-Contents?

Search sorting algorithms

Sorting algorithm - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Sorting_algorithm
In computer science, a sorting algorithm is an algorithm that puts elements of a list in a certain order. The most-used orders are numerical order and ...

Quicksort - Bubble sort - Merge sort - Shellsort

Sorting Algorithm Animations
www.sorting-algorithms.com/ Animation, code, analysis, and discussion of 8 sorting algorithms on 4 initial conditions.

Quick Sort - Insertion Sort - Quick Sort (3 Way Partition) - Shell Sort

Sorting Algorithms
cs.smith.edu/~thiebaut/java/sort/demo.html...

1 Classification
  1.1 Stability

2 Comparison of algorithms

3 Summaries of popular sorting algorithms
  3.1 Bubble sort
  3.2 Selection sort
  3.3 Insertion sort
  3.4 Shell sort
  3.5 Comb sort
  3.6 Merge sort
  3.7 Heapsort
  3.8 Quicksort
  3.9 Counting sort
  3.10 Bucket sort
  3.11 Radix sort
  3.12 Distribution sort
  3.13 Timsort

4 Memory usage patterns and index sorting

5 Inefficient/humorous sorts

6 See also

7 References

8 External links
TOC generation as a process

Segments
- Line breaks  
- Markup  

Almost always

Titles
- Headings  
- Bold font

Sometimes

Hierarchy
- Section numbering  
- Linked TOC

Basically never
Hierarchy identification as classification

Segment 1

Same level n=0

Segment 2
Hierarchy identification as classification

Segment 1  Lower level n=1  Segment 2
Hierarchy identification as classification

Segment 1  \rightarrow \text{Higher level } n=-1 \rightarrow \text{Segment 2}
Hierarchy identification as classification

Segment 1

Higher level 
n=-2

Segment 2
Construct the full Table-of-Contents

1. Introduction
2. Related Work
3. Experimental Setup
   3.1 Datasets
4. Experiments and Results
5. Segment Title Generation
   5.1 Experimental Setup
   5.2 Experiments and Results
6. Conclusions and Future Work

n=0
n=1
n=-1
DKPro TC
Features

• N-grams
• Length
• Entities
• Noun chunks
• Keyphrases
• Frequency
N-gram features

On a lunar mission, the Command Module Pilot (CMP) was assigned the role of navigator...

The Earth-based mission control teams for Apollo 8 consisted of astronauts assigned...

Top n-grams
- lunar
- mission
- apollo

lunar_0: true
mission_0: true
apollo_0: false

lunar_1: false
mission_1: true
apollo_1: true

• Good for cue phrases such as „earlier“ or „above“
• Good for flags for standard segments such as „references“ or „Introduction“.
## Length features

On a lunar mission, the Command Module Pilot (CMP) was assigned the role of navigator…

<table>
<thead>
<tr>
<th>Characters</th>
<th>Tokens</th>
<th>Sentences</th>
</tr>
</thead>
<tbody>
<tr>
<td>88</td>
<td>15</td>
<td>1</td>
</tr>
</tbody>
</table>

Δ Characters: 2
Δ Tokens: 3
Δ Sentences: 0

The Earth-based mission control teams for Apollo 8 consisted of astronauts assigned…

<table>
<thead>
<tr>
<th>Characters</th>
<th>Tokens</th>
<th>Sentences</th>
</tr>
</thead>
<tbody>
<tr>
<td>86</td>
<td>12</td>
<td>1</td>
</tr>
</tbody>
</table>

• Higher level segments may be shorter because they provide a summary
Entity feature

On a lunar mission, the Command Module Pilot (CMP) was assigned the role of navigator...

The Earth-based mission control teams for Apollo 8 consisted of astronauts assigned...

Command Module Pilot  
Shared NE: false  
earth  
apollo 8

• Two segments are related (on a lower or equal level) if they share entities
On a lunar mission, the Command Module Pilot (CMP) was assigned the role of navigator...

The Earth-based mission control teams for Apollo 8 consisted of astronauts assigned...

- Chunks in lower level segments are longer because they are more detailed
- “[bubble] sort algorithm“
On a lunar mission, the Command Module Pilot (CMP) was assigned the role of navigator…

The Earth-based mission control teams for Apollo 8 consisted of astronauts assigned…

- Lower-level segments mention noun chunks from a higher level
Keyphrase features

On a lunar mission, the Command Module Pilot (CMP) was assigned the role of navigator...

The Earth-based mission control teams for Apollo 8 consisted of astronauts assigned...

mission lunar navigator

Shared keyphrases
1: false
2: true
3: true

Apollo mission astronauts

• Get topically related segments
Frequency features

On a lunar mission, the Command Module Pilot (CMP) was assigned the role of navigator...

The Earth-based mission control teams for Apollo 8 consisted of astronauts assigned...

<table>
<thead>
<tr>
<th>Keyphrase</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>mission</td>
<td>25,903,394</td>
</tr>
<tr>
<td>lunar</td>
<td>1,365,796</td>
</tr>
<tr>
<td>navigator</td>
<td>845,732</td>
</tr>
<tr>
<td>Apollo</td>
<td>275,687</td>
</tr>
<tr>
<td>mission</td>
<td>25,903,394</td>
</tr>
<tr>
<td>astronauts</td>
<td>777,434</td>
</tr>
</tbody>
</table>

ΔØ Keyphrases freq.
1:  25,627,707
2:  545,055
3:  386,136

• Lower-level segments have less-frequent keyphrases
• They are more specific
### Datasets

<table>
<thead>
<tr>
<th>Name</th>
<th># documents</th>
<th># segments</th>
<th>tokens/segment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cormen</td>
<td>1</td>
<td>607</td>
<td>733</td>
</tr>
<tr>
<td>Gutenberg</td>
<td>18</td>
<td>1,312</td>
<td>1,927</td>
</tr>
<tr>
<td>Wikipedia</td>
<td>277</td>
<td>3,680</td>
<td>399</td>
</tr>
</tbody>
</table>

**Cormen**
- Single book
- Many segments
- Domain-specific

**Gutenberg**
- Project
- Long segments
- Broad styles

**Wikipedia**
- Featured articles
- Very fine-grained
- Broad topics
## Datasets

<table>
<thead>
<tr>
<th>Name</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cormen</td>
<td>.00</td>
<td>.02</td>
<td>.08</td>
<td>.41</td>
<td>.48</td>
</tr>
<tr>
<td>Gutenberg</td>
<td>.07</td>
<td>.48</td>
<td>.41</td>
<td>.04</td>
<td>.00</td>
</tr>
<tr>
<td>Wikipedia</td>
<td>.01</td>
<td>.35</td>
<td>.49</td>
<td>.12</td>
<td>.03</td>
</tr>
</tbody>
</table>

**Cormen**
- Deep levels
- Only one level 1

**Gutenberg**
- Balanced
- Different authors

**Wikipedia**
- Mostly level 3
- Short articles
### Datasets

<table>
<thead>
<tr>
<th>Name</th>
<th>n=2</th>
<th>n=1</th>
<th>n=0</th>
<th>n=-1</th>
<th>n=-2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cormen</td>
<td>.00</td>
<td>.20</td>
<td>.60</td>
<td>.16</td>
<td>.03</td>
</tr>
<tr>
<td>Gutenberg</td>
<td>.00</td>
<td>.15</td>
<td>.71</td>
<td>.13</td>
<td>.01</td>
</tr>
<tr>
<td>Wikipedia</td>
<td>.00</td>
<td>.10</td>
<td>.80</td>
<td>.09</td>
<td>.01</td>
</tr>
</tbody>
</table>

- **n=2**
  - Very unlikely
  - Skipping one level

- **n=-2**
  - Very unlikely
  - Going up two levels
## Experimental results

<table>
<thead>
<tr>
<th>Name</th>
<th>Cormen</th>
<th>Wikipedia</th>
<th>Gutenberg</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline (always equal)</td>
<td>.60</td>
<td>.71</td>
<td>.80</td>
</tr>
<tr>
<td>(1) N-gram features</td>
<td>.86</td>
<td>.64</td>
<td>.86</td>
</tr>
<tr>
<td>(2) Length features</td>
<td>.62</td>
<td>.76</td>
<td>.80</td>
</tr>
<tr>
<td>(3) Entity features</td>
<td>.60</td>
<td>.71</td>
<td>.80</td>
</tr>
<tr>
<td>(4) Noun chunk features</td>
<td>.83</td>
<td>.86</td>
<td>.91</td>
</tr>
<tr>
<td>(5) Keyphrase features</td>
<td>.60</td>
<td>.71</td>
<td>.80</td>
</tr>
<tr>
<td>(6) Frequency features</td>
<td>.60</td>
<td>.71</td>
<td>.80</td>
</tr>
<tr>
<td>All features</td>
<td>.86</td>
<td>.77</td>
<td>.86</td>
</tr>
<tr>
<td>All features w/o (1)</td>
<td>.83</td>
<td>.86</td>
<td>.91</td>
</tr>
<tr>
<td>All features w/o (3) &amp; (5)</td>
<td>.87</td>
<td>.77</td>
<td>.86</td>
</tr>
</tbody>
</table>
Confusion matrices

<table>
<thead>
<tr>
<th>Actual</th>
<th>Predicted</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>-4</td>
</tr>
<tr>
<td>1</td>
<td>-567</td>
</tr>
<tr>
<td>0</td>
<td>-2,585</td>
</tr>
<tr>
<td>-1</td>
<td>-478</td>
</tr>
<tr>
<td>-2</td>
<td>-24</td>
</tr>
</tbody>
</table>

All features w/o n-gram features on Wikipedia dataset.

<table>
<thead>
<tr>
<th>Actual</th>
<th>Predicted</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>-4</td>
</tr>
<tr>
<td>1</td>
<td>-539</td>
</tr>
<tr>
<td>0</td>
<td>-14</td>
</tr>
<tr>
<td>-1</td>
<td>-2</td>
</tr>
<tr>
<td>-2</td>
<td>-12</td>
</tr>
</tbody>
</table>

All features on Wikipedia dataset.

Only 1 and 0 predicted

Predicting more levels
Table-of-contents: An example

1 Crew
1.1 Backup crew
1.2 Mission control
1.3 Mission insignia
2 Planning
3 Saturn V
4 Mission
4.1 Parameter summary
4.2 Launch and trans-lunar injection
4.3 Lunar trajectory
4.4 Lunar sphere of influence
4.5 Lunar orbit
4.5.1 Earthrise
4.6 Unplanned manual re-alignment
4.7 Cruise back to Earth and re-entry
5 Historical importance
6 Spacecraft location
7 In film

Actual table-of-contents

Predicted table-of-contents
Conclusions

• Identify hierarchy relation between two segments
  → Results are sufficient for stable table-of-contents generation

• Two new datasets
  → Natural data for real tasks!

• Open source experiments

www.ukp.tu-darmstadt.de/data/table-of-contents-generation
Future work

• Create user interface

• Perform user acceptance tests

• Global TOC generation
Thanks to all my colleagues at UKP!

Btw: The empty book was selected as one of the school books by the commission.