Automated Intrinsic Text Classification for Component Content Management Applications in Technical Communication

Jan Oevermann
jan.oevermann@hs-karlsruhe.de

Wolfgang Ziegler
wolfgang.ziegler@hs-karlsruhe.de

Motivation

• Semantic access to information through classification
• Demand for automation in industry use cases

• Adapt existing ML methods for Component Content Management
• Little research on Technical Communication topics
Technical Communication

• Writing documentation (and more)
• Complex information management
• Legal obligations and international standards

• Component Content Management
  • Modularized content for reuse and translation
  • XML-based information models
  • Metadata and classification models
  • Single Source Publishing
Methodology

1. Characterize relevant properties of CCM
2. Derive implications for classification
3. Verify with real-world data sets (Vector space classification)
Classification Models

• PI classification model (Ziegler 2011)
• Organized in taxonomies
• Focus on intrinsic information classification
Use Cases

- Content delivery portals
- Automated publishing
- Dynamic linking

<table>
<thead>
<tr>
<th></th>
<th>Series</th>
<th>Model</th>
<th>Project</th>
</tr>
</thead>
<tbody>
<tr>
<td>Safety advice</td>
<td>C-123</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Product description</td>
<td>C-321</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Operation Main Engine</td>
<td>C-159</td>
<td></td>
<td>C-158</td>
</tr>
<tr>
<td>Maintenance</td>
<td></td>
<td></td>
<td>C-123</td>
</tr>
</tbody>
</table>
Characteristics

• Standardized patterns
• Specific terminology
• Size of content
• Training and validation data
• Quality requirements
Data sets

<table>
<thead>
<tr>
<th>Set</th>
<th>Sector</th>
<th>Units</th>
<th>Words/Unit</th>
<th>Classes</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Construction equipment</td>
<td>570</td>
<td>173</td>
<td>11</td>
</tr>
<tr>
<td>B</td>
<td>Medical lab equipment</td>
<td>278</td>
<td>41</td>
<td>10</td>
</tr>
<tr>
<td>C</td>
<td>Security printing presses</td>
<td>3947</td>
<td>97</td>
<td>22</td>
</tr>
</tbody>
</table>

- XML-based content components
- Manually classified
- German language
Implications

• Semantic quantifiers
  \[ t_{fq} = t_f \cdot q \quad \text{for } q > 0 \]

• Confidence scoring
  \[ p = \frac{s_1 - s_2}{s_1 - s_n} \]

Instead of softmax or standard deviation
Feature selection

- Smaller total number of features
  - Standardization of wording and patterns
  - Size of content components
- Single words and patterns important
  - Combination of words and patterns
Token weighting

- **Tf-Idf**
  - Good for documents

- **Tf-Idf-Cf** (Liu/Yang 2012)
  - In-class characteristics

- **Tf-Icf-Cf**:

  \[ w_{ij} = \log(1 + t_{fi}) \times \log\left(1 + \frac{|C|}{t_{fi}}\right) \times \frac{t_{fij}}{C_j} \]
Applications

• Data migration
• Key figures (QA)

• Authoring assistance
• Content delivery portals (API, Import hook)
Results & Observations

• CCM has different requirements than document classification
• Technical content is well suited for automated classification
• Set of adjustments for content components to improve results
• Working prototype: REST API for classification of content components
Related work & Outlook

• Soto et al. (2015): Similarity-Based Support for Text Reuse in Technical Writing
• Oevermann (2016): Reconstructing Semantic Structures in Technical Documentation with Vector Space Classification
• Apply results to unstructured technical content
• Use more advanced machine learning or deep learning technologies
Contact

Jan Oevermann
jan.oevermann@hs-karlsruhe.de
www.janoevermann.de