Automatic Assignment of LCSH
Combining Classification and Extraction

Christian Wartena  Michael Franke-Maier
Hochschule Hannover, Abteilung Information und Kommunikation
Freie Universität Berlin, Universitätsbibliothek

September 27, 2017
Overview

1. Introduction

2. Data
   - Records
   - LCSH

3. Classification
   - Baseline
   - Logistic Regression
   - Examples
   - Hybrid Method

4. Discussion
### Introduction

**LIS**

**LIS ‘15**  First ideas came up at the conference dinner

- Intuition: classification not promising for assignment of LCSH

**TPDL ‘16**  The problem is really hard. Matching labels in text is not good enough.

**Today**  Let’s try classification nevertheless
Introduction

**B3KAT**
- Library Union Catalogue of Bavaria, Berlin and Brandenburg
- Linked Open Data Representation [http://lod.b3kat.de](http://lod.b3kat.de)
- ca. 26 Mio bibliographic entities
- Freie Universität Berlin is shared cataloging partner library

**Library of congress Subject Headings (LCSH)**
- Used since 1898 for cataloging materials at the LoC
- Linked open data Representation
  [http://id.loc.gov/authorities/subjects.html](http://id.loc.gov/authorities/subjects.html)
- Widely adopted in the anglophone world
- Cross-referenced to GND and Rameau (MACS-Project)
A series of research papers on the history of native peoples (Indian, Inuit, Metis) in Canada, presented at a Native Studies colloquium at Brandon University, Manitoba in 1981. Includes a bibliographic essay on the Indian in Canadian historical writing.
## Data – Records

### Source
- SPARQL-Endpoint of the B3-Catalogue (B3Kat)
- Different from the one used in the 2016 TPDL Paper

### Criteria
- Abstract of at least 200 characters
- English, according to metadata **and** language detection
- LCSH in metadata
- Query results might differ due to time-outs

### Sets
<table>
<thead>
<tr>
<th>Set</th>
<th>Count</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Train 1</td>
<td>11,544</td>
<td>(Training classifiers)</td>
</tr>
<tr>
<td>Train 2</td>
<td>500</td>
<td>(Finding thresholds)</td>
</tr>
<tr>
<td>Test</td>
<td>500</td>
<td></td>
</tr>
</tbody>
</table>
We use informative words from title, subtitle and abstract as features.

- Words occurring in at least 10 documents ($df(w) \geq 10$)
- Words occurring in at most 2000 documents ($df(w) < 2000$)

6660 words
LCSH in Text

Labels of LCSH
- LCSH have IDs and labels
- Labels could occur in text
- LCSH are not classes!
  - Training a classifier for *each* LCSH is impossible
  - For many LCSH we don’t have training data at all.

Types of Labels
- Preferred labels
- Labels including scope notes (*Boundaries (Estates)*)
- Labels of precombined headings containing "- -"
- Inverted labels (*Steed, John (Fictitious character)*)
### Many highly specific headings

<table>
<thead>
<tr>
<th>LCSH</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>sh00000172</td>
<td>Halle 13 (Expo, International Exhibitions Bureau, 2000, Hannover, Germany)</td>
</tr>
<tr>
<td>sh2005002460</td>
<td>Brown versus Board of Education of Topeka</td>
</tr>
<tr>
<td>sh85120114</td>
<td>Septets (Piano, flute, zither, percussion)</td>
</tr>
</tbody>
</table>

### Ambiguity

- Many LCSH with identical variants!
- Especially, after removing scope notes from the labels
### Selection of LCSH

<table>
<thead>
<tr>
<th>Count</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>414,355</td>
<td>Subject headings</td>
</tr>
<tr>
<td>162,569</td>
<td>Pre-combined headings (like <em>Voyages and travel - Mythology</em>)</td>
</tr>
<tr>
<td>497,427</td>
<td>Terms after removing pre-combined labels with dashes, subdivisions, inverted labels and Children’s headings</td>
</tr>
<tr>
<td>572,697</td>
<td>Terms after adding (non-ambiguous) singular forms</td>
</tr>
<tr>
<td>15,661</td>
<td>Ambiguous terms after removing scope notes (like in <em>Taxis (Biology)</em> and <em>Taxis (Vehicles)</em>)</td>
</tr>
</tbody>
</table>
Figure: Number of labels for each degree of ambiguity.
**Subject Headings in our Data**

Distribution of LCSH in Training Set 1

**Figure:** Number of headings with given frequency in training data
Subject Headings in our Data

Distribution of LCSH in Training Set 1

- 10,944 headings (2.6% of all available headings)
  - 28,818 assignments
- 451 headings used over 10 times
  - 9,458 assignments

Most frequent headings:

<table>
<thead>
<tr>
<th>LCSH</th>
<th>Label</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>sh85056605</td>
<td>Great Britain</td>
<td>239</td>
</tr>
<tr>
<td>sh85147430</td>
<td>Women and literature</td>
<td>100</td>
</tr>
<tr>
<td>sh85045631</td>
<td>Europe</td>
<td>98</td>
</tr>
<tr>
<td>sh85043777</td>
<td>English literature</td>
<td>90</td>
</tr>
<tr>
<td>sh85009808</td>
<td>Authors, American</td>
<td>90</td>
</tr>
</tbody>
</table>
Baseline

Extraction from Text (TPDL ‘16)

- Assign LCSH if one of its labels is found in the text.

<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Extraction</td>
<td>0.071</td>
<td>0.30</td>
<td>0.10</td>
</tr>
</tbody>
</table>


Upper bounds

Best possible result for the classifier

- Best possible precision is 1
- Recall: the classifier can only choose from 451 subject headings

<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Extraction</td>
<td>0.071</td>
<td>0.30</td>
<td>0.10</td>
</tr>
<tr>
<td>Oracle</td>
<td>1.0</td>
<td>0.28</td>
<td>0.34</td>
</tr>
</tbody>
</table>
We train 541 one-vs-all classifiers
Low probabilities for each class
We assign \( n \) most probable LCSH to each record
Determine \( n \) using the second train set
Logistic Regression

Using a threshold

- Alternatively, we assign all subject headings, for which the probability is above some threshold.
- Determine threshold using the second train set.
Logistic Regression

Combinations

- The best subject heading is always used.
- More subject headings, when probability is above some threshold.

Results

<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Extraction</td>
<td>0.071</td>
<td>0.30</td>
<td>0.10</td>
</tr>
<tr>
<td>Oracle</td>
<td>1.0</td>
<td>0.28</td>
<td>0.34</td>
</tr>
<tr>
<td>Log. Regr. (n=1)</td>
<td>0.47</td>
<td>0.22</td>
<td>0.28</td>
</tr>
<tr>
<td>Log. Regr. (n=1 OR p &gt; 0.03)</td>
<td>0.46</td>
<td>0.26</td>
<td>0.31</td>
</tr>
</tbody>
</table>
Logistic Regression

Combinations

- The best subject heading is always used.
- More subject headings, when probability is above some threshold.

Results

<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Extraction</td>
<td>0.071</td>
<td>0.30</td>
<td>0.10</td>
</tr>
<tr>
<td>Oracle</td>
<td>1.0</td>
<td>0.28</td>
<td>0.34</td>
</tr>
<tr>
<td>Log. Regr. (n=1)</td>
<td>0.47</td>
<td>0.22</td>
<td>0.28</td>
</tr>
<tr>
<td>Log. Regr. (n=1 OR p &gt; 0.03)</td>
<td>0.46</td>
<td>0.26</td>
<td>0.31</td>
</tr>
</tbody>
</table>
Example Politics: Method Extraction

The Irish constitutional tradition

- http://lod.b3kat.de/page/title/BV009807040
- existing LCSH: ireland, constitutional history
- …1782 to the present day and treats the **constitutional history** of **Ireland**, north and south, as an integrated whole…
- Correctly: ireland; constitutional history
- Missing: –
- Wrongly: irish; political science; day; north and south
Example Politics: Method Classifier

The Irish constitutional tradition

- http://lod.b3kat.de/page/title/BV009807040
- existing LCSH: ireland, constitutional history
- …1782 to the present day and treats the **constitutional history** of Ireland, north and south, as an integrated whole…
- Correctly: ireland ; constitutional history
- Missing: –
- Wrongly: constitutional law
Dictionary and handbook of nuclear medicine and clinical imaging

- http://lod.b3kat.de/page/title/BV004061649
- existing LCSH: diagnostic imaging, nuclear medicine, radioisotopes

Dictionary that "bridges the gap between those highly sophisticated papers and ...volumes dealing with basic sciences generally." Intended for generalists and specialists. Brief definitions. Handbook contains basic and reference data.

- Correctly: nuclear medicine ; diagnostic imaging
- Missing: radioisotopes
- Wrongly: paper ; science
<table>
<thead>
<tr>
<th>Dictionary and handbook of nuclear medicine and clinical imaging</th>
</tr>
</thead>
<tbody>
<tr>
<td>- <a href="http://lod.b3kat.de/page/title/BV004061649">http://lod.b3kat.de/page/title/BV004061649</a></td>
</tr>
<tr>
<td>- existing LCSH: diagnostic imaging, nuclear medicine, radioisotopes</td>
</tr>
<tr>
<td>Dictionary that &quot;bridges the gap between those highly sophisticated papers and …volumes dealing with basic sciences generally.&quot; Intended for generalists and specialists. Brief definitions. Handbook contains basic and reference data.</td>
</tr>
<tr>
<td>Correctly: –</td>
</tr>
<tr>
<td>Missing: nuclear medicine ; radioisotopes ; diagnostic imaging</td>
</tr>
<tr>
<td>Wrongly: social sciences ; religion and science ; psychiatry</td>
</tr>
</tbody>
</table>
Example Physics: Method Extraction

Planck

- http://lod.b3kat.de/page/title/BV042620297
- existing LCSH: physicists, physics, national socialism and science
- Correctly: –
- Missing: physics; physicists; national socialism and science
- Wrongly: war; vision; process (law); law; radiation; universe; quantum theory; comprehension; matter; states; twentieth century; home; science; shorthand
Abstract: Planck

- **Planck’s Law**, an equation used by physicists ... **Max Planck** is credited with being the father of **quantum theory**, and his work laid the foundation for our modern understanding of **matter** and energetic processes. But Planck’s story is not well known, especially in the United States. ...What remains, ...,

- are handwritten letters in German **shorthand**, ...In Planck: Driven by **Vision**, Broken by **War**, Brandon R.

- Planck’s Law not a concept in LCSH, but in GND: http://d-nb.info/gnd/4174789-6

- Max Planck as person not a concept in LCSH, but in LC Name Authority File http://id.loc.gov/authorities/names/n80038130.html

- vision and war very unspecific words from subtitle
Example Physics: Method Classifier

Planck

- [link](http://lod.b3kat.de/page/title/BV042620297)
- existing LCSH: physicists, physics, national socialism and science
- Correctly predicted: –
- Missing: physics; physicists; national socialism and science
- Wrongly predicted: modernism (art); cognition; scientists
Hybrid Method

Combining classification and extraction

1. The best subject heading is always used.
2. More subject headings, when probability is above some threshold.
3. More subjects are added, when
   1. The subject is not in the set of 541 subjects
   2. A label of the subject occurs in the text

Results

<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Extraction</td>
<td>0.071</td>
<td>0.30</td>
<td>0.10</td>
</tr>
<tr>
<td>Oracle</td>
<td>1.0</td>
<td>0.28</td>
<td>0.34</td>
</tr>
<tr>
<td>Log. Regr. (n=1)</td>
<td>0.47</td>
<td>0.22</td>
<td>0.28</td>
</tr>
<tr>
<td>Log. Regr. (n=1 OR p &gt; 0.03)</td>
<td>0.46</td>
<td>0.26</td>
<td>0.31</td>
</tr>
<tr>
<td>Combined</td>
<td>0.40</td>
<td>0.37</td>
<td>0.33</td>
</tr>
</tbody>
</table>
Hybrid Method

Combining classification and extraction

1. The best subject heading is always used.
2. More subject headings, when probability is above some threshold.
3. More subjects are added, when
   - The subject is not in the set of 541 subjects
   - A label of the subject occurs in the text

Results

<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Extraction</td>
<td>0.071</td>
<td>0.30</td>
<td>0.10</td>
</tr>
<tr>
<td>Oracle</td>
<td>1.0</td>
<td>0.28</td>
<td>0.34</td>
</tr>
<tr>
<td>Log. Regr. (n=1)</td>
<td>0.47</td>
<td>0.22</td>
<td>0.28</td>
</tr>
<tr>
<td>Log. Regr. (n=1 OR p &gt; 0.03)</td>
<td>0.46</td>
<td>0.26</td>
<td>0.31</td>
</tr>
<tr>
<td><strong>Combined</strong></td>
<td><strong>0.40</strong></td>
<td><strong>0.37</strong></td>
<td><strong>0.33</strong></td>
</tr>
</tbody>
</table>
Error analysis

General impression

- Differences with gold standard are not real errors in most cases!
- Classification makes errors for specific disciplines that are underrepresented in the training data (e.g. medicine, physics)

Quality and errors

- Given the huge amount of possible LCSH: both methods have a high recall
- Low precision: many false positives

False positives

- Extraction: Very general terms, that occur in many texts
- Classification: Completely wrong (specific) areas, with some related terms and lack of training data.
Error analysis

General impression

- Differences with gold standard are not real errors in most cases!
- Classification makes errors for specific disciplines that are underrepresented in the training data (e.g. medicine, physics)

Quality and errors

- Given the huge amount of possible LCSH: both methods have a high recall
- Low precision: many false positives

False positives

<table>
<thead>
<tr>
<th>Extraction</th>
<th>Classification</th>
</tr>
</thead>
<tbody>
<tr>
<td>Very general terms, that occur in many texts</td>
<td>Completely wrong (specific) areas, with some related terms and lack of training data.</td>
</tr>
</tbody>
</table>
Conclusions

Lessons learned

- LCSH is not very well suited for automatic assignment (ambiguity, pre-combined labels, etc.)
- For the majority of headings there is not enough training data available.
- Classification works very well for a small subclass of headings.
  - We reach almost highest possible recall
- Extraction of rare and specific labels from text can improve classification results.
Discussion

Related work

  https://doi.org/10.1007/978-3-319-43997-6_40

Future work

- Integrate procedures into workflow at the FU Library
- Using hierarchical structure
- Using other features (e.g. DDC, GND, etc.; NKOS ‘17)
- Results for other vocabularies (e.g. GND or MESH)
Thanks for your attention!

- Michael Franke-Maier: franke@ub.fu-berlin.de
- Christian Wartena: christian.wartena@hs-hannover.de