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

UNIVERSITÄTS BIBLIOTHEK

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

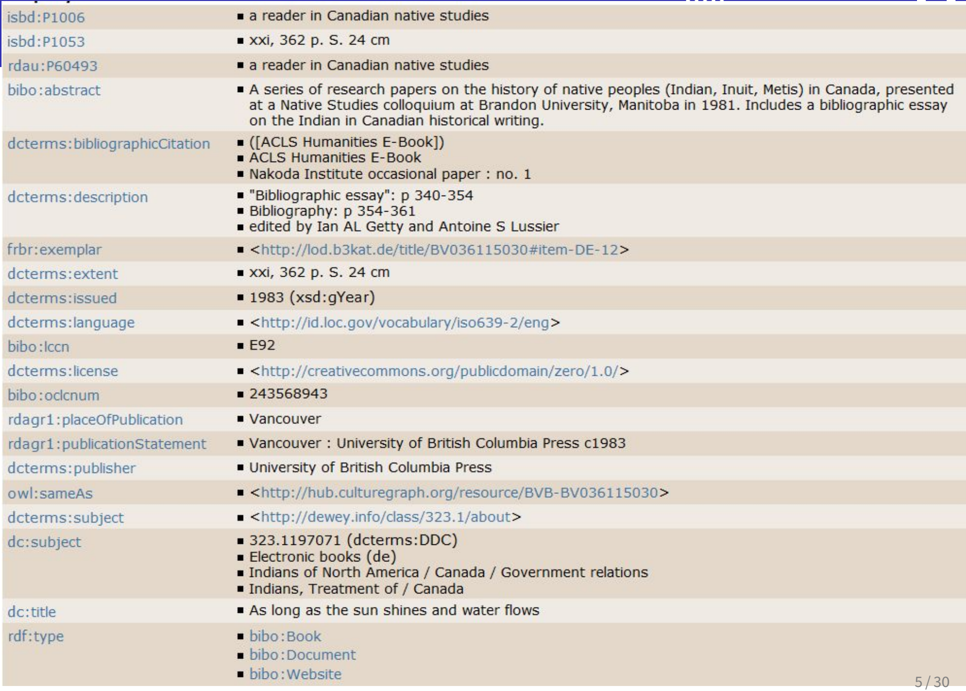
Introduction

B3KAT

- Library Union Catalogue of Bavaria, Berlin and Brandenburg
- Linked Open Data Representation `http://lod.b3kat.de`
- 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`
- Widely adopted in the anglophone world
- Cross-referenced to GND and Rameau (MACS-Project)



isbd:P1006	a reader in Canadian native studies
isbd:P1053	xxi, 362 p. S. 24 cm
rdau:P60493	a reader in Canadian native studies
bibo:abstract	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.
dcterms:bibliographicCitation	([ACLS Humanities E-Book]) ACLS Humanities E-Book Nakoda Institute occasional paper : no. 1
dcterms:description	"Bibliographic essay": p 340-354 Bibliography: p 354-361 edited by Ian AL Getty and Antoine S Lussier
frbr:exemplar	<http://lod.b3kat.de/title/BV036115030#item-DE-12>
dcterms:extent	xxi, 362 p. S. 24 cm
dcterms:issued	1983 (xsd:gYear)
dcterms:language	<http://id.loc.gov/vocabulary/iso639-2/eng>
bibo:lccn	E92
dcterms:license	<http://creativecommons.org/publicdomain/zero/1.0/>
bibo:oclcnum	243568943
rdagr1:placeOfPublication	Vancouver
rdagr1:publicationStatement	Vancouver : University of British Columbia Press c1983
dcterms:publisher	University of British Columbia Press
owl:sameAs	<http://hub.culturegraph.org/resource/BVB-BV036115030>
dcterms:subject	<http://dewey.info/class/323.1/about>
dc:subject	323.1197071 (dcterms:DDC) Electronic books (de) Indians of North America / Canada / Government relations Indians, Treatment of / Canada
dc:title	As long as the sun shines and water flows
rdf:type	bibo:Book bibo:Document bibo:Website

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

Train 1	11 544	(Training classifiers)
Train 2	500	(Finding thresholds)
Test	500	

Abstracts

Words

- 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)*)

LCSH

Many highly specific headings

sh00000172	Halle 13 (Expo, International Exhibitions Bureau, 2000, Hannover, Germany)
sh2005002460	Brown versus Board of Education of Topeka
sh85120114	Septets (Piano, flute, zither, percussion)

Ambiguity

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

LCSH

Selection of LCSH

414 355	subject headings
162 569	pre-combined headings (like *Voyages and travel - Mythology*)
497 427	terms after removing pre-combined labels with dashes, subdivisions, inverted labels and Children's headings
572 697	terms after adding (non-ambiguous) singular forms
15 661	ambiguous terms after removing scope notes (like in *Taxis (Biology)* and *Taxis (Vehicles)*)

LCSH

Ambiguity

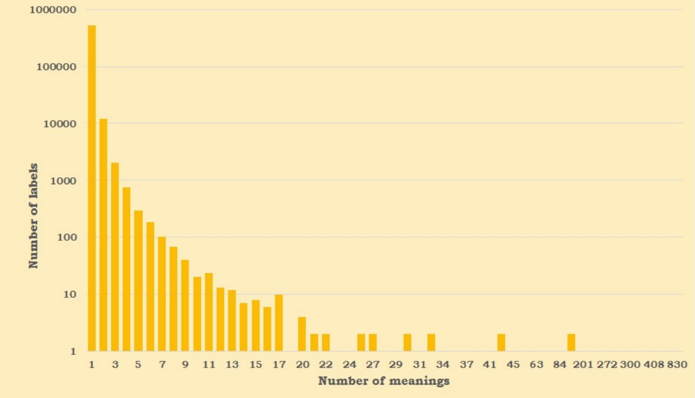


Figure: Number of labels for each degree of ambiguity.

Subject Headings in our Data

Distribution of LCSH in Training Set 1

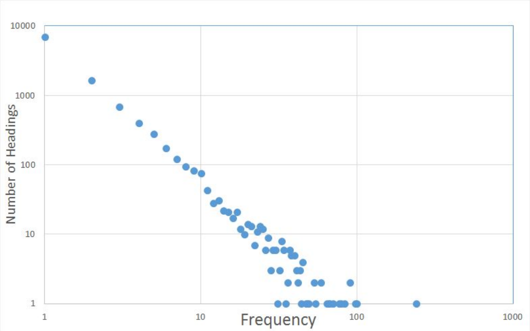


Figure: Number of headings with given frequency in training data

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:

LCSH	Label	Frequency
sh85056605	Great Britain	239
sh85147430	Women and literature	100
sh85045631	Europe	98
sh85043777	English literature	90
sh85009808	Authors, American	90

Baseline

Extraction from Text (TPDL '16)

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

	Precision	Recall	F1
Extraction	0, 071	0, 30	0, 10

Upper bounds

Best possible result for the classifier

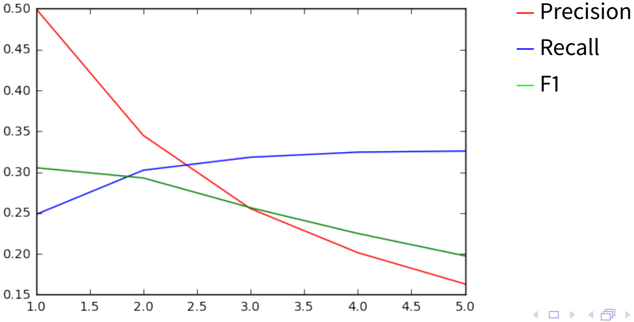
- Best possible precision is 1
- Recall: the classifier can only choose from 451 subject headings

	Precision	Recall	F1
Extraction	0, 071	0, 30	0, 10
Oracle	**1, 0**	**0, 28**	**0, 34**

Logistic Regression

Training

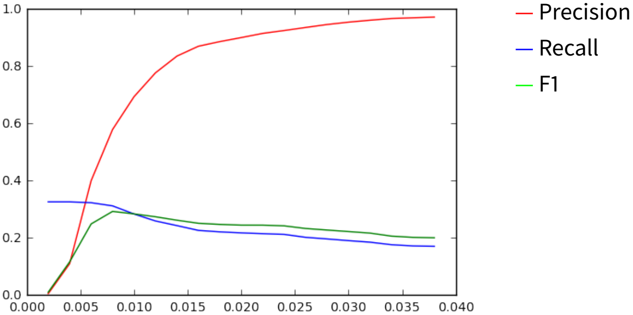
- We train 541 one- vs. all classifiers
- Low probabilities for each class
- We assign *n* most probable LCSH to each record
- Determine *n* using the second train set



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

	Precision	Recall	F1
Extraction	0, 071	0, 30	0, 10
Oracle	1, 0	0, 28	0, 34
Log. Regr. (n=1)	**0, 47**	**0, 22**	**0, 28**
Log. Regr. (n=1 OR p > 0,03)	**0, 46**	**0, 26**	**0, 31**

Logistic Regression

Combinations

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

Results

	Precision	Recall	F1
Extraction	$0,071$	$0,30$	$0,10$
Oracle	$1,0$	$0,28$	$0,34$
Log. Regr. (n=1)	$\mathbf{0,47}$	$\mathbf{0,22}$	$\mathbf{0,28}$
Log. Regr. (n=1 OR p > 0,03)	$\mathbf{0,46}$	$\mathbf{0,26}$	$\mathbf{0,31}$

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

Example Medicine : Method Extraction

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

Example Medicine : Method Classifier

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: –
- Missing: nuclear medicine ; radioisotopes ; diagnostic imaging
- Wrongly: social sciences ; religion and science ; psychiatry

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

Example Physics

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

- `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

	Precision	Recall	F1
Extraction	0, 071	0, 30	0, 10
Oracle	1, 0	0, 28	0, 34
Log. Regr. (n=1)	0, 47	0, 22	0, 28
Log. Regr. (n=1 OR p > 0,03)	0, 46	0, 26	0, 31
Combined	**0, 40**	**0, 37**	**0, 33**

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

	Precision	Recall	F1
Extraction	0, 071	0, 30	0, 10
Oracle	1, 0	0, 28	0, 34
Log. Regr. (n=1)	0, 47	0, 22	0, 28
Log. Regr. (n=1 OR p > 0,03)	0, 46	0, 26	0, 31
Combined	**0, 40**	**0, 37**	**0, 33**

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

Extraction Very general terms, that occur in many texts

Classification Completely wrong (specific) areas, with some related terms and lack of training data.

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

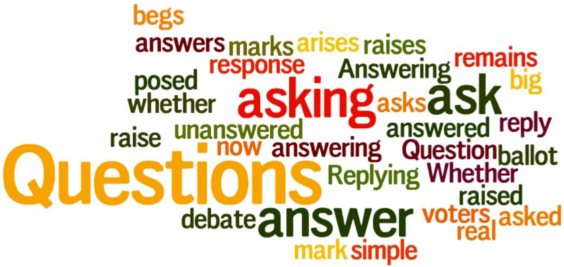
- Aga R.T., Wartena C., Franke-Maier M. (2016)
Automatic Recognition and Disambiguation of Library of Congress Subject Headings. In: Fuhr N., Kovács L., Risse T., Nejdl W. (eds) Research and Advanced Technology for Digital Libraries. TPDL 2016. Lecture Notes in Computer Science, vol 9819. Springer, Cham.
`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)

Contact

Thanks for your attention!



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