

Creating gold standards and supervising outcomes – the future role of library staff in supporting information services based on Machine Learning

Dr. Timo Borst
Information systems and publishing technologies
ZBW Leibniz Information Center for Economics
Kiel/Hamburg, Germany



Content

- Overview: Artificial Intelligence and Machine Learning in academic libraries
- Exemplary applications with respect to training / test data management, evaluation and information ethics
- Wrap-up

(not so much about transformers, CNN, GNN, Python libraries, RoBERTa & Co...)

Overview: AI/ML/DL in digital libraries

- **Text classification** and machine translation
- **Metadata generation by means of named entity recognition**
- Query understanding and reformulation
- **Information retrieval and (re-)ranking of search results**
- Image classification and object detection, as conducted e.g. in Optical Character Recognition (OCR)
- **Information assistants**

Metadata generation by means of named entity recognition

Task / problem statement

- Automatic recognition of funders of scientific research acknowledged in scientific papers
- Creating a gold standard in terms of training / test set, optimal solution for funder recognition

Approach

- Extracting „green OA“ papers from an OA repository including metadata and fulltext
- Reconciling with funder information from CrossRef
- Checking / correcting funder information, adding acknowledgement phrase
- Providing / maintaining metadata by means of an Excel sheet

(Initial) curation effort including labeling: *approx. 2 weeks per 1 FTE*

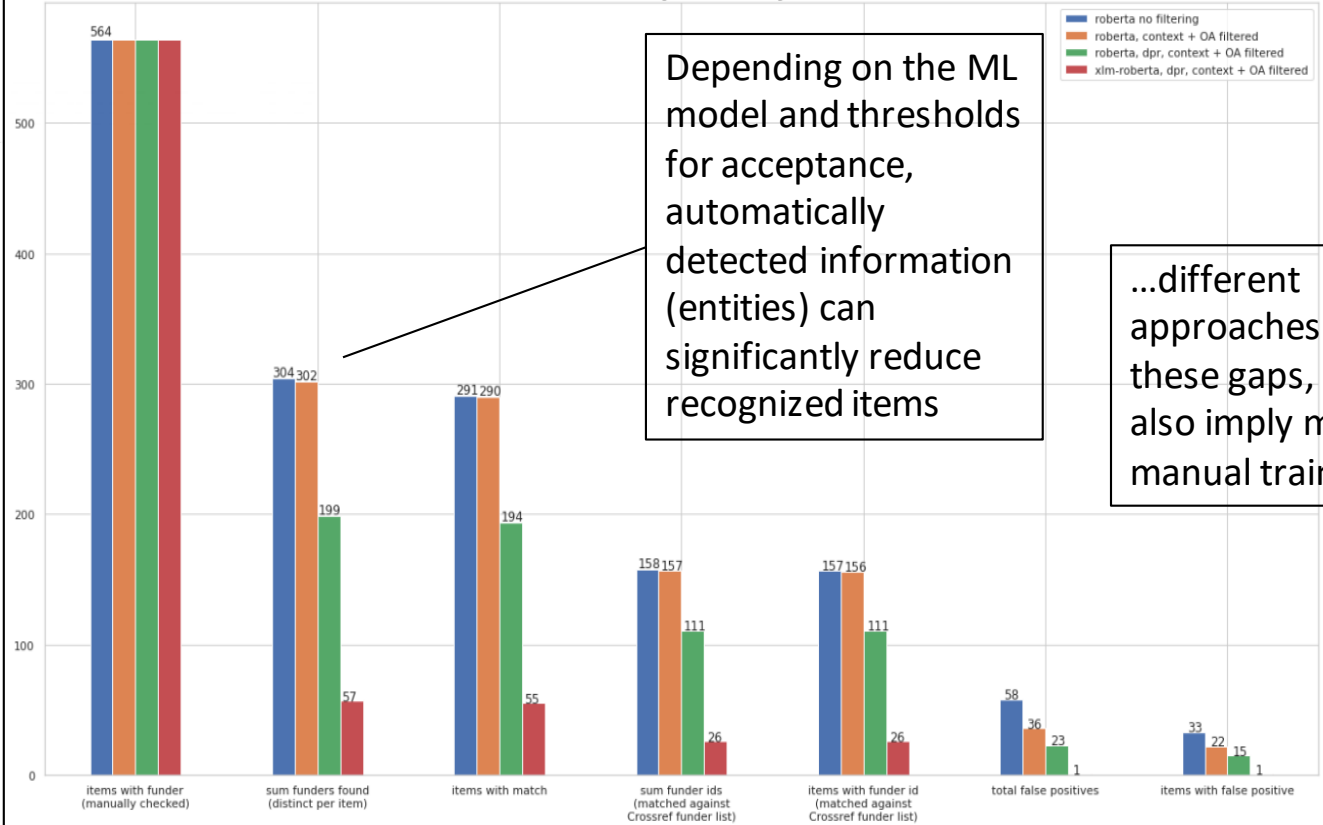
Metadata generation by means of named entity recognition

	A	B	C	D	E	F	G	H
1	resource_id	Handle	DOI	Funder Identifier	Grant No.	Funder-Info It. CrossRef	PDF-URL	Funder-Phrase It. PDF
245	200660	10419/195003	doi:10.1186/s40174-016-0057-2			'Swiss National Science Foundation (SNSF)'	https://www.econstor.eu/obitstream/10419/195003/1/890257795.pdf	Sarah Voitchovsky is grateful to the Swiss National financial support.
246	182056	10419/176457	doi:10.1186/s40854-017-0067-8			'Southwestern University of Finance and Economics'	https://www.econstor.eu/obitstream/10419/176457/1/10.1186_s40854-017-0067-8.pdf	Southwestern University of Finance and Econo
247	205763	10419/200101	doi:10.1016/j.esr.2017.12.002	'10.13039/100010663'	'StG 2012-313553'	'H2020 European Research Council'	https://www.econstor.eu/obitstream/10419/200101/1/10.1016_j.esr.2017.12.002.pdf	Stefan Pfenninger acknowledges support from grant StG 2012-313553. This paper was made p Open Energy Modelling Initiative.
248	173247	10419/167955	doi:10.3390/g6040458	'10.13039/501100002428'	'J3475'	'Austrian Science Fund'	https://www.econstor.eu/obitstream/10419/167955/1/840197632.pdf	Support from the Jomm Rempington Foundation i Hilbe acknowledges generous funding by the Sc Science Fund (FWF) J3475.
249	173253	10419/167960	doi:10.3390/g6040574	'10.13039/501100002428', '10.13039/501100000646', '10.13039/501100001809'	'P27018-G11', '26330387', '61503062 and 61203374'	'Austrian Science Fund', 'Japan Society for the Promotion of Science', 'National Natural Science Foundation of China'	https://www.econstor.eu/obitstream/10419/167960/1/840206844.pdf	T.S. was supported by the Austrian Science Fun acknowledges support by Grants-in-aid for Scie Society for the Promotion of Science 26330387 Natural Science Foundation of China (Grants N Fundamental Research Funds of the Central Un
250	172888	10419/167645	doi:10.1007/s10964-017-0720-6	'10.13039/501100003986', '10.13039/501100000269'	'ES/J019658/1'	'Jacobs Foundation', 'Economic and Social Research Council'	https://www.econstor.eu/obitstream/10419/167645/1/10.1007_s10964-017-0720-6.pdf	Terry Ng-Knight is supported by the post-doctor to Adulthood, funded by the Jacobs Foundation Wissenschaftszentrum Berlin (WZB) and Grant I British Eco- nomic and Social Research Council and Life-chances in the Knowledge Economies
251	181195	10419/175595	doi:10.1186/s13561-016-0100-z			'Deutsche Forschungsgemeinschaft and Ruprecht-Karls-Universität Heidelberg'	https://www.econstor.eu/obitstream/10419/175595/1/87764022X.pdf	The Article Processing Charges for this paper w Forschungsgemeinschaft and Ruprecht-Karls-Ui
		10419/195558	doi:10.1007/s40092-017-00004-4	'10.13039/50110000044'	'MHR-02-73-200-429'	'Ministry of Human Resource Development'	https://www.econstor.eu/obitstream/10419/195558/1/87764022X.pdf	The author makesh Kumar, Virendra, would like ti

Funder phrase added manually (,autopsy')

Human inspection is needed to perceive outliers

Results for all merged answers score ge 12



Depending on the ML model and thresholds for acceptance, automatically detected information (entities) can significantly reduce recognized items


...different approaches to narrow these gaps, but may also imply more manual training data

Information retrieval and (re-)ranking of search results

Task / problem statement

- (Re-)Ranking of search results according to different criteria (text statistics, popularity, freshness,...)
- Creating a gold standard (or ‚baseline‘) in terms of a reference set for judging relevance factors and training set for ‚learning to rank‘

Approach

- Funded project „LibRank“ (by DFG German Research Foundation) 
- Creating a snapshot from a ‚living‘ search index
- Document pooling (based on text statistics)
- Human search tasks according to which search results are judged by different user groups
- Web application and interface for labeling and (re-)ranking of search hits
(,relevance assessment tool‘ (RAT) from HAW Hamburg University, Prof Dirk Lewandowski)

Information retrieval and (re-)ranking of search results

Relevance Assessment Tool

Fortschritt: 0% 100%

(0 von 23 Ergebnissen)

Suchanfrage:

Kostenrechnung und Kostenanalyse

Beschreibung:

Gesucht werden Lehrmaterialien zu Kostenrechnung und Kostenanalyse. Wie erfolgt die Durchführung und gibt es Fallstudien oder Rechenbeispiele?

Wie relevant ist das Dokument?

nicht relevant

relevant

Relevant?

ja nein

Nächste

Kostenrechnung und Kostenanalyse in der chemischen Industrie

von Günther Geissler ; Werner Müller ; Dieter Seidel ; Horst Weihs

Erscheinungsjahr: 1964

Weitere Verfasser/innen: Geißler, Günther; Müller, Werner; Seidel, Dieter; Weihs, Horst

Verlag:

Leipzig : VEB Dt. Verl. für Grundstoffind.

Beschreibung:

426 S
8

Sprache:

Deutsch

Schlagwörter:

Chemieindustriebetrieb | Betriebskostenrechnung | DDR

Publikationsform:

Buch / Working Paper

Anmerkungen:

Mit Literaturverz. (S. 420 - 426)

Verfügbarkeit:

in Bibliotheken finden

Exemplare in Ihrer Bibliothek

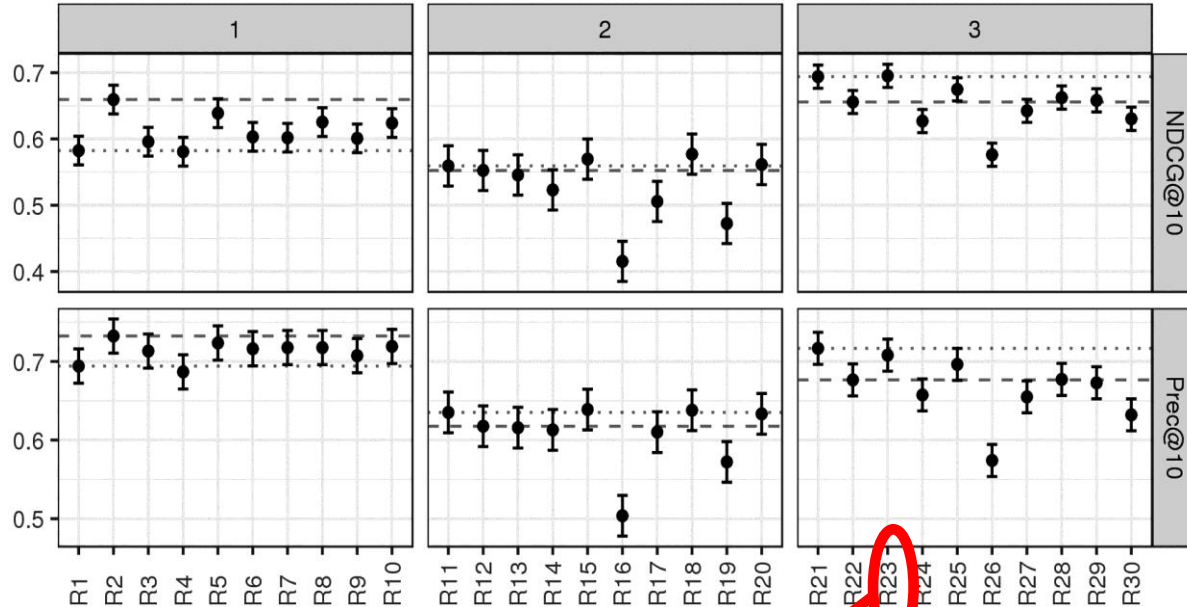
Standort: Ihre Bibliothek

Signatur: II 52127

Status: - Verfügebar Bestellen

- Specific application and GUI for labeling by humans (,RAT')
- Relevance judgements can be either binary (yes / no) or ordinal (,more relevant than non-relevant')
- Digital tools for human labeling must reflect the complexity of the task

Information retrieval and (re-)ranking of search results



- Learning-to-rank including all ranking factors (popularity, impact, freshness, availability,...)
- L2R only applied in third run on the basis of the first two runs, automatic weighing of relevance factor
- Requires relevance judges by raters both for training data and comparing different ranking models
- Labeling effort: *Several weeks for three test runs with different rater groups*

Information assistants

Task / problem statement

- Supporting / assisting both library staff and users by introducing an automatic Q&A system („conversational AI“)
- Creating a gold standard in terms of training / test set by labeling conversations

Approach

- Tracking and anonymizing real-world chats (keyboard input)
- Manually labeling the chats according to different categories (e.g., general information, user management, specific document delivery)
- Transforming labeled conversations into typical, rule-based dialogues

Information assistants

Seiten / ... / Labels for Chat Conversations 

Manual for labeling old conversations

Angelegt von Kazakova, Anastasia, zuletzt geändert am Mai 31, 2021

In the folder "\\EOWYN\zwbshare\110_employees\Kazakova_Anastasia\Chatbot\Chatlogs_for_Labeling" are the excel files with chat logs that should be labeled. The labels can be taken from Labels for Chat Conversations

Procedure

- Please start with the English logs
- Please put the labelled files into the folder "\\EOWYN\zwbshare\110_employees\Kazakova_Anastasia\Chatbot\Chatlogs_for_Labeling\gelabelt".
- Some groups, like **around library**, have multiple labels: **around library**, **library rules**, **fees**, **opening hours**. If you can define it exactly what the conversation is about, then take an appropriate one, e.g.: **Fees**. If this is not possible, then **around library**.
- If users have questions about only one topic, then labels belong to **columns C-E** and in the **row with Question ID**.
 - One topic → label unique → one label → column C
 - One topic → label not unique → one label per column C-E

Spalten A	B	C	D	E
Spalten ID	Text	Legende allgemein	sonstige info	
1	Patron Question	Überrascht ist		
2	Librarian Incident	Bibliothek / in Bibliothek name: bibliotheksbesucher/ besucher		
3	Patron Incident	How can I download again		
4	Librarian Incident	Bibliothek / in der Bibliothek...		
5	Librarian Incident	hallo, wie kann ich einen halben ?		
6	Librarian Incident	hello, how can I help you ?		
7	Librarian Incident	would you mind by "repulse" ?		
8	Librarian Incident	sorry "age"		
9	Librarian Incident	you should be able to view the book via this link : //doi.org/10.1202/tu19875 an		
10	Patron Incident	Bibliothek / in ich nicht mehr verbunden.		
11	Librarian Incident	Das die Bibliothek / in bei der Chatting/beendet.		
12	Librarian Incident	auf's Thung: Bibliothek		
13	Librarian Incident	Absprechungen von Bibliothek / in 10789		
14				

- If users ask questions about the several different topics, please place the labels in the row with the question matching the topics

- Purpose: Generating training / test data, but also to know more about the distribution of conversational topics
- Labels are not pre-analyzed or recognized by the system, but manually annotated (currently no machine learning here, but planned)
- Labels to be defined by information supporting staff
- Multi-labeling to be applied because of ambiguities
- Specific labels for unusual requests (e.g., spam)
- Again, Excel-based – but thinking about integrating with chat backend ,LibAnswers‘ ...
- Effort: ~ 90 working hours for labeling 5k conversations

Information ethics

- Generally: important, but to be adopted
- Avoid wrong or inappropriate results caused by imprecision, false-positives/-negatives
- Decide on threshold for reliable entity recognition
- Biases to be anticipated and to be avoided (resulting from unbalanced training data and recognition of information entities), e.g.
 - funders: not (or wrongly) detected because of unusual funding phrase
 - relevance ranking: document pooling may exclude relevant papers from the very beginning
 - information assistant: ambiguous conversations should be handled appropriately, e.g. by multi-labeling or by splitting them up)

Wrap up

- Validated test / training data is essential for ML approaches
- It takes at least weeks, sometimes even much longer to systematically collect training / test data in terms of labeled collections
- Easier for ML projects that can refer to already labeled data (e.g., bibliographic databases)
- Often it means extra effort to structure and preprocess the data to be labeled

Wrap up (cont'd)

- Labeling might be integrated into standard environments and workflows, e.g. by extending tools for cataloguing
- Software environments and applications for crowdsourcing of manual training data are essential, whereas maintaining or curating should be ,fun' (gamification)
- Information ethics: essential, but to be adopted and contextualized
- In the light of human effort, first results from ML test runs can be very disappointing – but check your models and algorithms before investing more into human labeling and corresponding tools

References

- **Khakpour, A., Colomo-Palacios, R.:** *Convergence of Gamification and Machine Learning: A Systematic Literature Review*. Tech Know Learn (2020). <https://doi.org/10.1007/s10758-020-09456-4>
- **Xiaoli, Z., Jie, Z., X. Le, D., Thoma, G.:** *A semi-supervised learning method to classify grant support zone in web-based medical articles*. Proc. SPIE 7247, Document Recognition and Retrieval XVI, 72470W (2009). <https://doi.org/10.1117/12.806076>
- **Cordell, R.** (2020). *Machine learning and libraries: a report on the state of the field*. <https://labs.loc.gov/static/labs/work/reports/Cordell-LOC-ML-report.pdf>
- **Behnert, C., Plassmeier, K., Borst, T., Lewandowski, D.:** *Evaluierung von Rankingverfahren für bibliothekarische Informationssysteme*. *Information - Wissenschaft & Praxis*, vol. 70, no. 1, 2019, pp. 14-23. <https://doi.org/10.1515/iwp-2019-0004>
- **Ayu, I., Mckie, S., Narayan, B.:** *Enhancing the Academic Library Experience with Chatbots: An Exploration of Research and Implications for Practice*. *Journal of the Australian Library and Information Association*, 2019, 268-277. <https://doi.org/10.1080/24750158.2019.1611694>
- **Padilla, T.:** *Responsible Operations: Data Science, Machine Learning, and AI in Libraries*. OCLC Research Position Paper, 2019. <https://eric.ed.gov/?id=ED603715>

