DEUTSCHE NATIONAL BIBLIOTHEK

Maximilian Kähler, Christa Schöning-Walter

Einsatz von KI und DH in Bibliotheken

Erschließungsmaschine EMa und KI-Projekt der DNB

DEUTSCHE NATIONAL BIBLIOTHEK

Inhaltsverzeichnis

DEUTSCHE NATIONAL BIBLIOTHEK

Hintergrund

- DNB sammelt jährlich mehr als 2 Mio. Publikationen
  - davon sind ca. 1 ½ Mio. Netzpublikationen (E-Books, E-Journals etc.)
- Instrumente der inhaltlichen Erschließung:
  DDC-Sachgruppen, DDC-Notationen und GND-Schlagwörter

Seit etwa 10 Jahren sind maschinelle Verfahren im Einsatz,
um inhaltsbeschreibende Metadaten zu generieren … und damit
das Finden im Katalog zu unterstützen.

DEUTSCHE NATIONAL BIBLIOTHEK

Maschinelle Erschließung in der DNB

Klassifizierung

DDC-Sachgruppen und
DDC-Kurznotationen

Assoziative Verfahren

Beschlagwortung

GND
als normiertes Vokabular

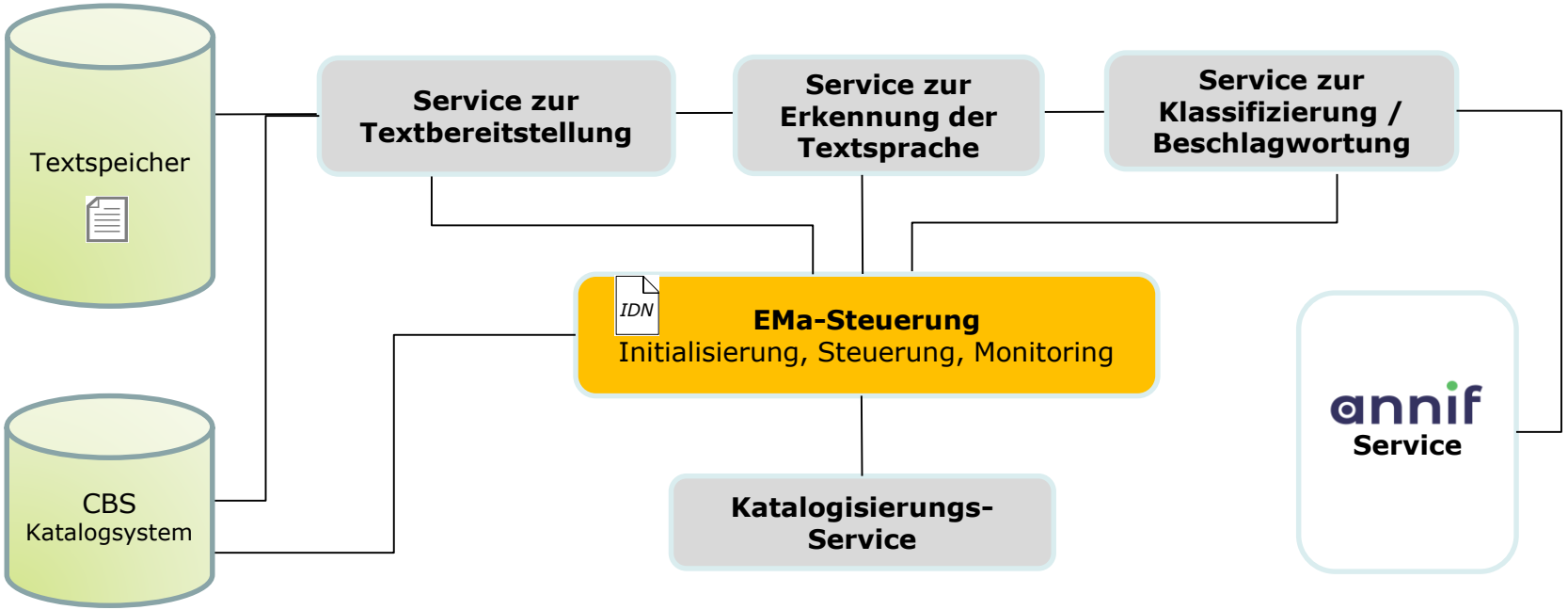
Lexikalische und assoziative Verfahren

Netzpublikationen und ausgewählte Printpublikationen

Deutsch und Englisch

DEUTSCHE NATIONAL BIBLIOTHEK

Modulare Erschließungsmaschine EMa



DEUTSCHE NATIONAL BIBLIOTHEK

Projekt Automatisches Erschließungssystem

https://www.dnb.de/ki-projekt

- Förderung im Rahmen der nationalen KI-Strategie
  - Beauftragte der Bundesregierung für Kultur und Medien (BKM)
- Laufzeit: 3 1/2 Jahre (Oktober 2021 – März 2025)
- Ausstattung
  - Personal (ca. 4 VZÄ)
  - Server mit 2 * 24 Prozessorkernen, 1 ½ TB RAM, 3.8 TB SSD-Festplatte
  - Mittel für Forschungskooperationen

DEUTSCHE NATIONAL BIBLIOTHEK

Unser Ziel

- Qualität der maschinellen Beschlagwortung mit der GND durch passende Methoden messbar verbessern
  (F-Score = 0,4 oder besser)
- Technologie- und Wissenstransfer in die bibliothekarische Praxis

Projekt ist unser Labor für die Erforschung –

EMa das modulare (Steckkasten-)System für die Produktion

DEUTSCHE NATIONAL BIBLIOTHEK

Was wollen wir tun?

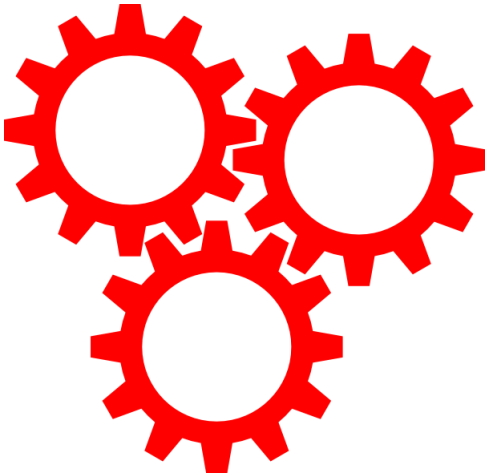
- vielversprechende neue Methoden/Algorithmen finden, systematisch untersuchen und ggf. adaptieren
  - die Qualität der GND-Verknüpfungen verbessern
  - (später) auch semantische Konzepte ohne GND-Repräsentation finden
  - Evaluation mit deutschsprachigen wissenschaftlichen Netzpublikationen
- daraus Werkzeuge entwickeln und bereitstellen
  - Services für die Einbindung in Erschließungssysteme mit modularer Architektur (EMa und andere)

DEUTSCHE NATIONAL BIBLIOTHEK

Forschungsschwerpunkte des Projektes

***Methoden für die* Extraktion und Aufbereitung der Texte und bibliografischen Daten**

***Methoden für die* Extraktion und Aufbereitung des Vokabulars**
(1,35 Mio. potenzielle Konzepte der GND)



Methoden für die Textanalyse und thematische Einordnung

(semantische Verknüpfung der Texte mit den Konzepten der GND)

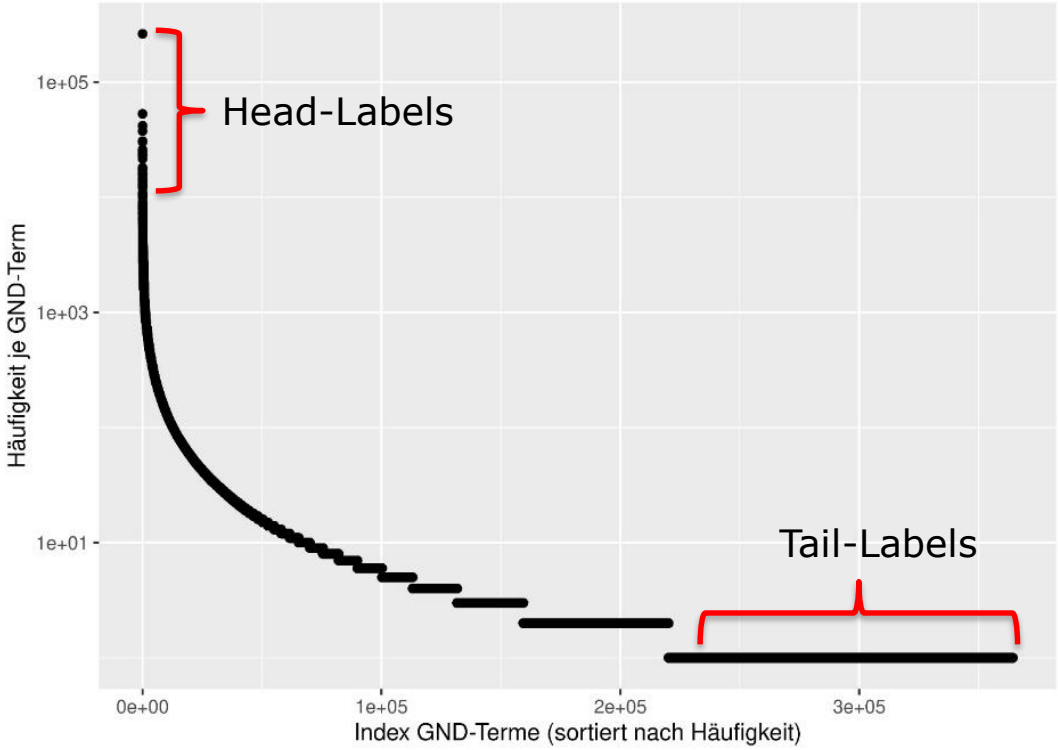
DEUTSCHE NATIONAL BIBLIOTHEK

Maschinelle Beschlagwortung als XMLC-Problem

- Maschinelle Beschlagwortung von Texten mit Konzepten aus der GND lässt sich als sogenanntes E**x**treme **M**ulti-**L**abel **C**lassification-Problem abstrahieren
  - Eingehende Text-Dokumente werden mit a-priori fest stehenden Labels (GND-Konzepte) verknüpft. Die Menge der zutreffenden Labels pro Dokument ist nicht beschränkt.
- Charakteristisch für XMLC-Probleme sind[1]:
  - Große „Label-Menge" ~ 10^5 – 10^6 Labels
  - Long-Tail-Charakteristik: Ein Großteil der möglichen Labels kommt in Trainingsdaten selten oder nie vor

[1]vgl. u.a. Jain etal. 2016 "Extreme Multi-label Loss Functions for Recommendation, Tagging, Ranking & Other Missing Label Applications"

Long-Tail Charakteristik für typische DNB Trainingsdaten



+ 1 Million Zero-Shot-Labels
(ohne Trainingsdaten)

DEUTSCHE NATIONAL BIBLIOTHEK

GND – Vokabular für die Beschlagwortung

- kooperativ gepflegter und genutzter Dienst zur Unterstützung der Erschließung und Vernetzung von Informationsressourcen

enthält mehr als 9 Mio. Konzepte für Entitäten:

Sachbegriffe, Personen, Körperschaften, Konferenzen, Geografika und Werke

davon sind ca. 1,35 Mio. potenzielle Konzepte für die Beschlagwortung

(nur) ca. 385.000 GND-Entitäten sind mit mindestens einer Publikation im Bestand der DNB verknüpft

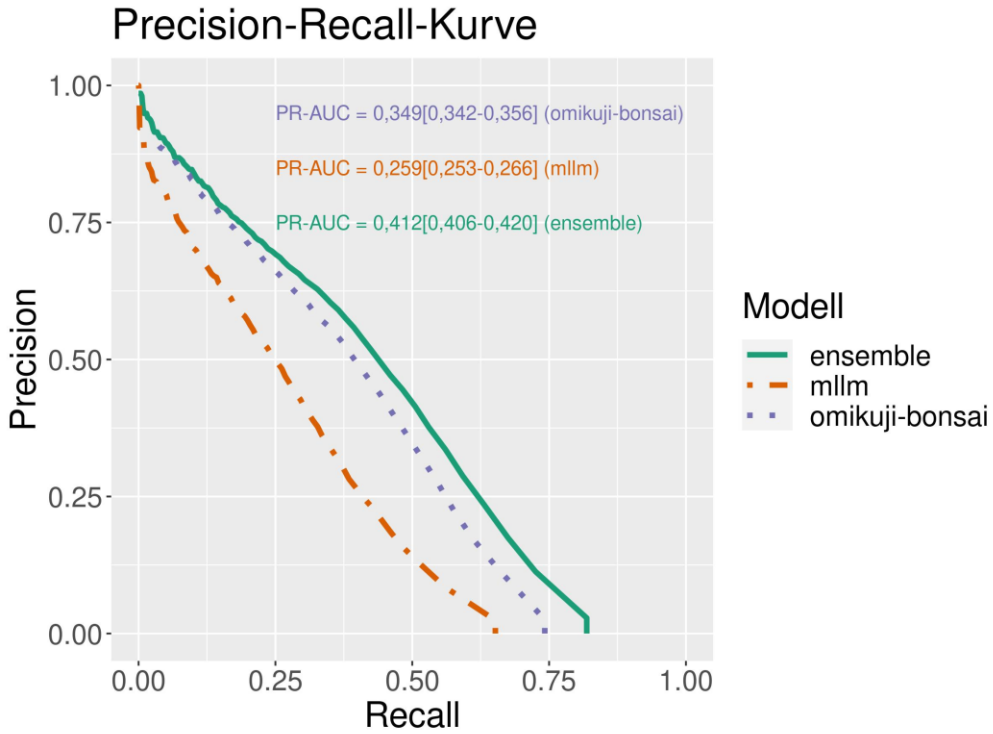
DEUTSCHE NATIONAL BIBLIOTHEK

Trainingsdaten

Wir gewinnen Trainingsdaten aus der intellektuellen Erschließung von Print-Publikationen mit GND-Konzepten:

- Volltexte: ca. 200.000 Volltexte (E-Books mit parallelen Print-Ausgaben)
- Inhaltsverzeichnisse: ca. 640.000 elektronisch lesbare Inhaltsverzeichnisse von Print-Publikationen
- Inhaltstexte: ca. 300.000 Inhaltstexte (z.B. Klappentext)
- Titel: ca. 1,5 Millionen reine Buch**titel**
- Durchschnittlich ca. 5 intellektuell vergebene GND-Konzepte pro Trainingstitel

Ergebnisse aktuell in der Produktion verwendeter Modelle



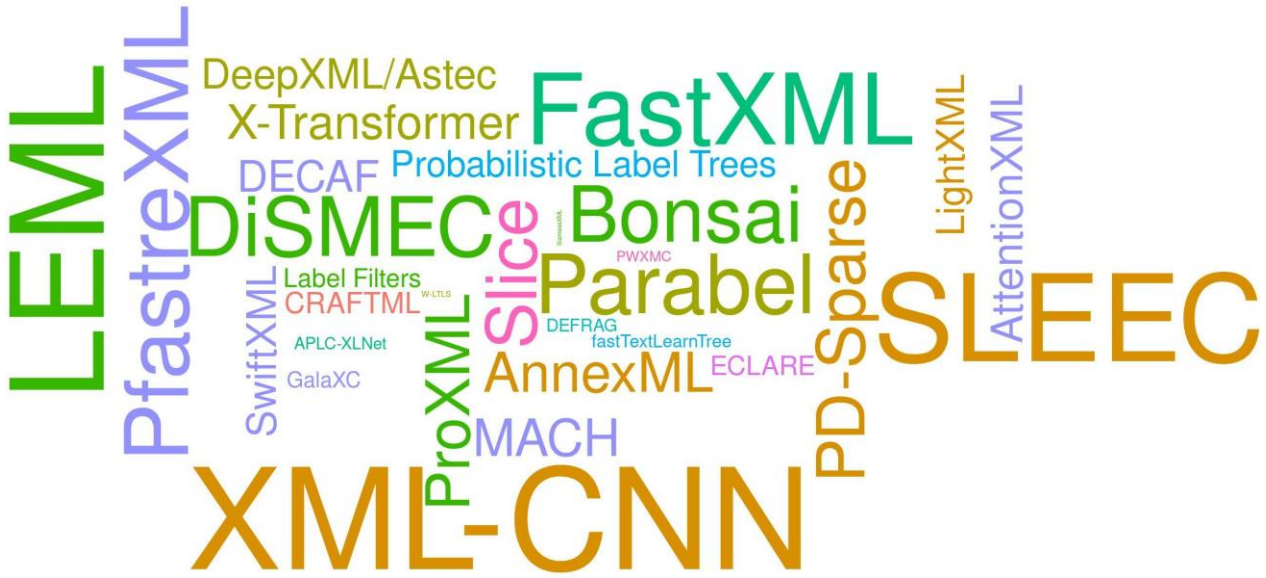
F1@5 Document Average on Test-Set

ensemble	0,376 (0,373-0,380)
mllm	0,275 (0,272-0,278)
omikuji-bonsai	0,349 (0,345-0,352)

mllm: Maui like lexical matching (vgl. annif.org)
Omikuji-bonsai: Khandagale etal. 2016 "Bonsai: diverse and shallow trees for extreme multi-label classification"

DEUTSCHE NATIONAL BIBLIOTHEK

Luxus und Herausforderung:
Die Vielfalt an XMLC-Verfahren ist groß!



Liste der Verfahren aus der Benchmark-Initiative: The Extreme Classification Repository (manikvarma.org), gewichtet nach ihren Google-Scholar-Zitationen pro Jahr, Stand 11/2021