Named Entity Disambiguation and Linking on Historic Newspaper OCR with BERT

Staatsbibliothek zu Berlin – Preußischer Kulturbesitz

CLEF2020 - HIPE

23 September 2020

{kai.labusch,clemens.neudecker}@sbb.spk-berlin.de

CLEF 2020 Thessaloniki

Staatsbibliothek zu Berlin Preußischer Kulturbesitz

Qurator Curation Technologies

Overview

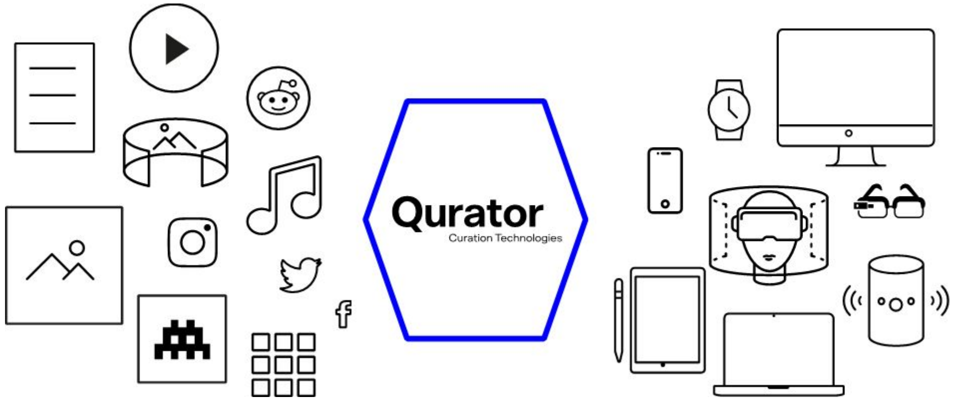
- Introduction
- Named Entity Recognition
- Named Entity Linking /Disambiguation:
  - Construction of knowledge base (KNB)
  - Entity candidate lookup
  - Candidate evaluation
  - Ranking of candidates
- CLEF HIPE NERC-Coarse results
- CLEF HIPE NEL-LIT results
- Conclusion

Introduction

Qurator

- *‘Flexible AI methods for the adaptive analysis and creative generation of digital content across multiple domain contexts’*

- Funded by the German Ministry for Education and Research (BMBF)
- Running time: 36 months (01/11/2019 - 31/10/2021)
- €15M Total Funding
- 10 Project Partners from Berlin
- Website: https://qurator.ai/



Qurator
Curation Technologies

Partners and Topics

DFKI: AI-platform for Curation Technologies

Wikimedia DE: Curation of Wikidata

3pc: Interactive Storytelling

Condat: TV-/Media-publications

Berlin State Library:

Digitized Cultural Heritage

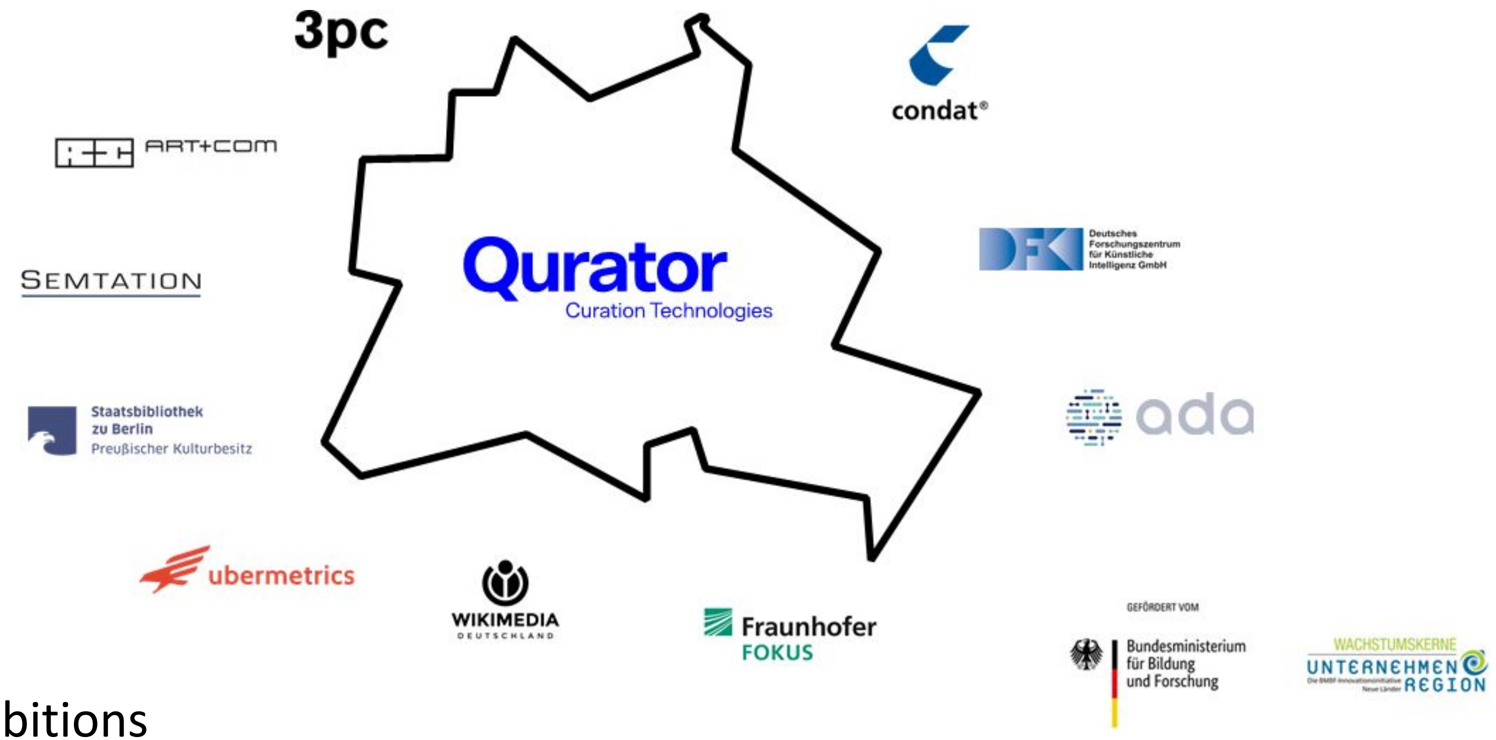
Ada: Biomedical Knowledge

FOKUS: Corporate Smart Insights (CSI)

ART+COM: Curation of Multimedia-Exhibitions

Ubermetrics: Media Intelligence and Risk-Monitoring

Semtation: Intelligent Process-Modelling



2nd Qurator Conference

- QURATOR2021 – Conference on Digital Curation Technologies
- 10-11 February 2021, Berlin, Germany
- Topics
  - Management of Digitally Curated and Semantically Expressive Information and Knowledge
  - AI-based / Semantic Large Scale and Complex Information and Content Analysis
  - Applications, Evaluations, and Experiences of applying Digital Curation Technologies, Standards, and Tools
- Call for Papers (submit before 23 November 2020): https://qurator.ai/conference-qurator-2021/call-for-papers/
- Proceedings are published in the CEUR Workshop Series

Qurator at the Berlin State Library

- Open Source development of technologies and applications:
  - https://github.com/qurator-spk
- Free access to data and models:
  - https://zenodo.org/communities/stabi
  - https://lab.sbb.berlin
- More about Qurator at Berlin State Library:
  - https://qurator.ai/partner/staatsbibliothek-zu-berlin/
  - https://qurator.ai/innovationlab/staatsbibliothek-zu-berlin/
  - Blog series „Artificial Intelligence“
    https://blog.sbb.berlin/tag/wissenschaftsjahr-2019/

Named Entity Recognition

- BERT-based NER tagger
- Pre-trained for “Masked-LM” and “Next Sentence Prediction” on historical text material of SBB digitized collections
- Off-the-shelf system not trained on CLEF-HIPE NER training data
- Does not support PROD and TIME entities
- German model: Trained on historical and contemporary German NER ground-truth
- French, English model: Trained on combined German, French, Dutch, and English NER ground-truth

Details ⟶ Kai Labusch, Clemens Neudecker and David Zellhöfer:
BERT for Named Entity Recognition in Contemporary and Historic German

Der	O
Fuball-	B-ORG
und	I-ORG
Leichtathletik	I-ORG
Verband	I-ORG
Westfalen	I-ORG
erteilte	O
Klaus	B-PER
Kämper	I-PER
,	O
Trainer	O
des	O
A-Ligisten	O
SG	B-ORG
Viktoria	I-ORG
Dortmund	I-ORG
,	O
für	O
sein	O
Auftreten	O
beim	O
Gastspiel	O
bei	O
der	O
TuS	B-ORG
Neuasseln	I-ORG
vom	O
5	O
Oktober	O
einen	O

NER Training – German Model

- **CoNLL** 2003 corpus (approx 200000 Tokens)
- **GermEval** Konvens 2014 corpus (approx 450000 Tokens)

- Historical Newspapers (Europeana Newspapers):
  - Newspapers 1926 (Landesbibliothek Dr. Friedrich Teßmann, approx 70.000 Tokens, **LFT**)
  - Newspaper 1710 - 1873 (Österreichische Nationalbibliothek, approx 30.000 Tokens, **ONB**)
  - Newspapers 1872 - 1930 (Staatsbibliothek zu Berlin, approx 50.000 Tokens, **SBB**)

Cross-validation results [1]

5-fold cross validation on	pre-train	BERT multi-lingual-cased precision	recall	F_1	(Riedl and Padó, 2018) F_1	(Schweter and Baiter, 2019) F_1
SBB	DC-SBB + GermEval + CoNLL	81.1 $\pm$1.2	87.8 $\pm$1.4	**84.3 $\pm$1.1**	-	-
	DC-SBB + CoNLL	81.0 $\pm$2.1	87.6 $\pm$1.8	84.2 $\pm$1.9	-	-
	DC-SBB + GermEval	80.6 $\pm$1.8	87.4 $\pm$1.3	83.8 $\pm$1.2	-	-
	CoNLL	81.0 $\pm$1.9	86.6 $\pm$2.2	83.7 $\pm$1.5	-	-
	GermEval	79.7 $\pm$1.8	87.2 $\pm$0.8	83.3 $\pm$1.1	-	-
	GermEval + CoNLL	79.9 $\pm$2.1	86.4 $\pm$1.7	83.0 $\pm$1.9	-	-
	DC-SBB	79.1 $\pm$2.6	86.7 $\pm$0.7	82.7 $\pm$1.3	-	-
	none	79.1 $\pm$3.6	85.0 $\pm$1.1	81.9 $\pm$2.2	-	-
ONB	Newspaper (1703-1875)	-	-	-	-	**85.31**
	DC-SBB+GermEval + CoNLL	81.5 $\pm$1.8	87.8 $\pm$1.4	84.6 $\pm$1.5	-	-
	DC-SBB + GermEval	81.6 $\pm$2.5	87.5 $\pm$1.6	84.5 $\pm$1.8	-	-
	DC-SBB + CoNLL	81.7 $\pm$2.8	87.5 $\pm$1.9	84.5 $\pm$2.3	-	-
	DC-SBB	81.8 $\pm$2.3	87.1 $\pm$2.1	84.3 $\pm$2.0	-	-
	GermEval	80.8 $\pm$2.1	85.4 $\pm$1.2	83.0 $\pm$1.4	78.56	-
	GermEval + CoNLL	80.0 $\pm$1.5	84.7 $\pm$1.6	82.3 $\pm$1.5	-	-
	CoNLL	79.1 $\pm$2.5	84.5 $\pm$2.1	81.7 $\pm$2.2	76.17	-
	none	78.0 $\pm$2.4	84.1 $\pm$1.9	80.9 $\pm$2.0	73.31	-
LFT	Newspaper (1888-1945)	-	-	-	-	**77.51**
	DC-SBB + CoNLL	70.0 $\pm$2.6	81.0 $\pm$0.7	75.1 $\pm$1.5	-	-
	DC-SBB + GermEval	69.9 $\pm$3.0	81.1 $\pm$1.0	75.1 $\pm$1.8	-	-
	DC-SBB	70.0 $\pm$3.5	80.8 $\pm$1.4	75.0 $\pm$2.1	-	-
	DC-SBB + GermEval + CoNLL	69.8 $\pm$3.0	80.8 $\pm$0.9	74.9 $\pm$2.0	-	-
	GermEval	68.9 $\pm$2.7	79.3 $\pm$1.4	73.7 $\pm$1.9	74.33	-
	GermEval + CoNLL	69.1 $\pm$2.6	78.8 $\pm$1.3	73.6 $\pm$1.5	-	-
	none	68.8 $\pm$3.4	79.2 $\pm$1.5	73.6 $\pm$2.2	69.62	-
	CoNLL	68.4 $\pm$3.1	79.1 $\pm$1.3	73.3 $\pm$2.1	72.9	-

[1] Kai Labusch, Clemens Neudecker and David Zellhöfer. BERT for Named Entity Recognition in Contemporary and Historic German, KONVENS 2019

Named Entity Linking
and
Disambiguation

Construction of Knowledge Bases for German, French and English

- Recursive traversal of category structure of German Wikipedia for identification of entities:
  - PER: All pages of categories „Frau“ or „Mann“ or of one of the reachable sub-categories of „Frau“ and „Mann“.
  - LOC: All pages of category „Geographisches Objekt“ or one of its sub-categories. Exclude everything that is part of „Geographischer Begriff“ or one of its sub-categories.
  - ORG: All pages of category „Organisation“ or one of its sub-categories.
- French and English knowledge bases:
  - Map identified German Wikipedia entity pages to Wikidata-IDs.
  - Map Wikidata-IDs back to French and English Wikipedia pages.

German, French, and English KNB:

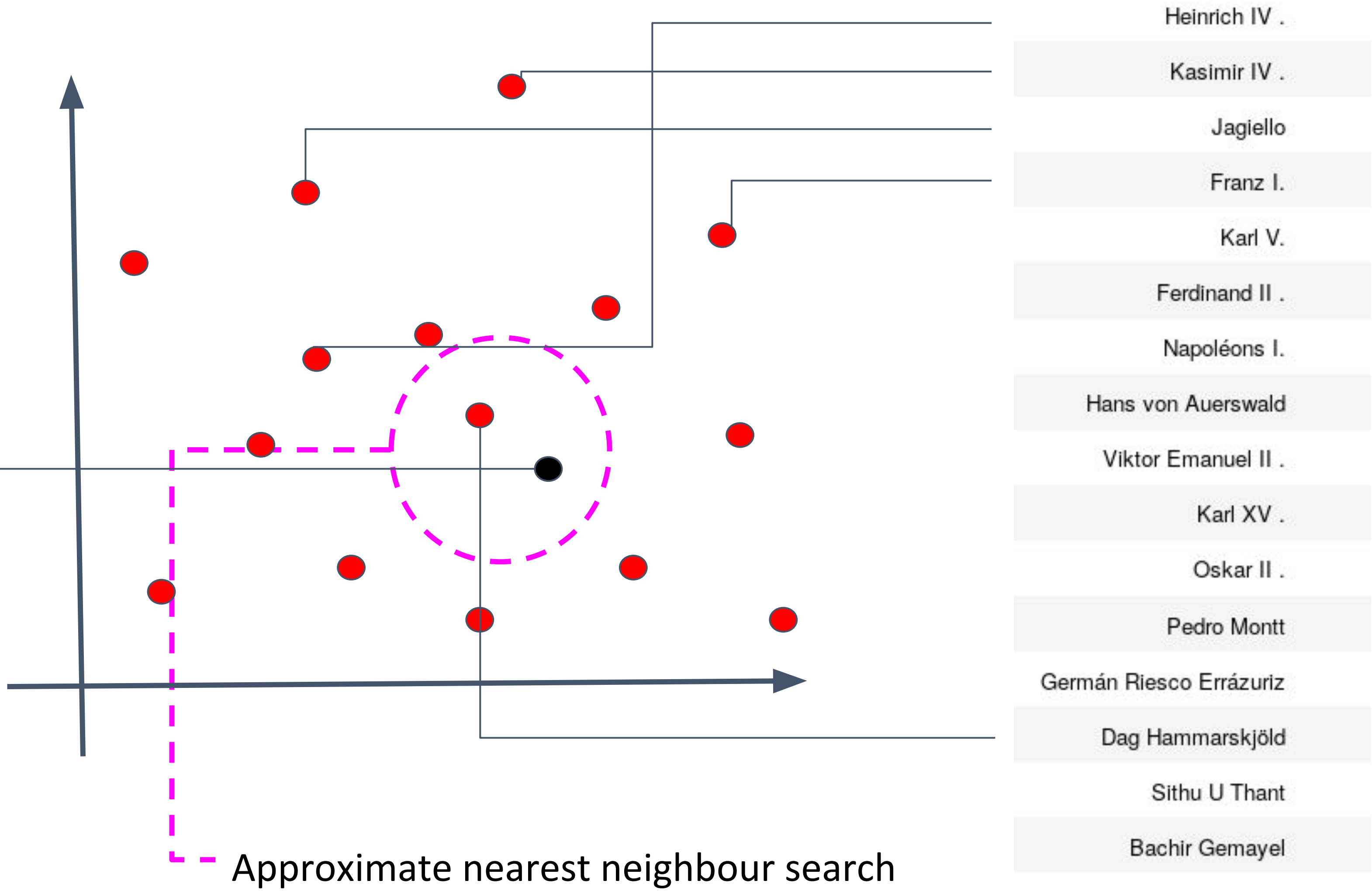
Lang	PER	LOC	ORG	coverage of test data
DE	671398	374048	136044	71%
FR	217383	155856	39305	68%
EN	324607	198570	58730	47%

- Size of French and English KNB significantly smaller than German
- Coverage of test data of German and French comparable, significantly worse for English

Entity Candidate Lookup

So wurden Erik Axel Karlfeldt 1931 und UN-Generalsekretär Hammarskjöld 1961 posthum geehrt.

- BERT embeddings stored in an approximate nearest neighbour index
- Lookup of up to 400 candidates with a distance less than 0.1
- 100 random projection search trees
- Angular distance measure



Entity Candidate Evaluation

- For each candidate consider up to 50 sentence pairs (A,B):
- Sentence A is part of text being subject to NEL.
- Sentence B is part of Wikipedia and contains explicit link to candidate.
- Purpose trained BERT-model determines probability of sentence (A,B) referring to the same item.
- Outcome is a set of matching probabilities per candidate.
- Final ranking of candidates on the basis of matching probabilities by ranking model.

Entity Candidate Ranking

- Outcome previous steps: Set of matching probabilities per candidate
- Compute statistical features of sets of matching probabilities:
    - Mean, median, min, max, standard deviation, various quantiles
    - Ranking statistics over all candidates
- Random forest model estimates overall matching probability per candidate
- Final output:
    - Sorted list of candidates that have matching probability > 0.2
    - NIL: not implemented. Either list of sorted candidates or “-” if there is not any candidate with matching probability above 0.2.

HIPE-NERC Results

NERC-Coarse
SBB system vs L3i (best system):

- Strict NER significantly worse than fuzzy NER
  - Difference more pronounced for SBB results
  - SBB system has not been trained on CLEF-HIPE data
  - Training data of SBB system is diverse
- German + French performance similar, English significantly worse
  - Overall OCR quality of French and German similar, English worse

Lang	Team	Evaluation	Label	P	R	F_1
DE	L3i	NE-COARSE-LIT-micro-fuzzy	ALL	0.870	0.886	0.878
DE	SBB	NE-COARSE-LIT-micro-fuzzy	ALL	0.730	0.708	0.719
DE	L3i	NE-COARSE-LIT-micro-strict	ALL	0.790	0.805	0.797
DE	SBB	NE-COARSE-LIT-micro-strict	ALL	0.499	0.484	0.491
FR	L3i	NE-COARSE-LIT-micro-fuzzy	ALL	0.912	0.931	0.921
FR	SBB	NE-COARSE-LIT-micro-fuzzy	ALL	0.765	0.689	0.725
FR	L3i	NE-COARSE-LIT-micro-strict	ALL	0.831	0.849	0.840
FR	SBB	NE-COARSE-LIT-micro-strict	ALL	0.530	0.477	0.502
EN	L3i	NE-COARSE-LIT-micro-fuzzy	ALL	0.794	0.817	0.806
EN	SBB	NE-COARSE-LIT-micro-fuzzy	ALL	0.642	0.572	0.605
EN	L3i	NE-COARSE-LIT-micro-strict	ALL	0.623	0.641	0.632
EN	SBB	NE-COARSE-LIT-micro-strict	ALL	0.347	0.310	0.327

HIPE NEL-LIT Results

NEL-LIT-micro-fuzzy-relaxed-@5:

- German + French: Competitive precision, poor recall
- German + French performance similar, English significantly worse
    - French, German coverage of test data similar, for English significantly worse

Lang	Team	Evaluation	Label	P	R	F_1
DE	L3i	NEL-LIT-micro-fuzzy-relaxed-@5	ALL	0.696	0.696	0.696
DE	SBB	NEL-LIT-micro-fuzzy-relaxed-@5	ALL	0.686	0.389	0.497
DE	aidalight-baseline	NEL-LIT-micro-fuzzy-relaxed-@5	ALL	0.440	0.435	0.437
FR	L3i	NEL-LIT-micro-fuzzy-relaxed-@5	ALL	0.746	0.743	0.744
FR	SBB	NEL-LIT-micro-fuzzy-relaxed-@5	ALL	0.716	0.393	0.507
FR	Inria-DeLFT	NEL-LIT-micro-fuzzy-relaxed-@5	ALL	0.604	0.670	0.635
FR	IRISA	NEL-LIT-micro-fuzzy-relaxed-@5	ALL	0.590	0.588	0.589
FR	aidalight-baseline	NEL-LIT-micro-fuzzy-relaxed-@5	ALL	0.516	0.508	0.512
EN	L3i	NEL-LIT-micro-fuzzy-relaxed-@5	ALL	0.744	0.744	0.744
EN	Inria-DeLFT	NEL-LIT-micro-fuzzy-relaxed-@5	ALL	0.633	0.685	0.658
EN	UvA.ILPS	NEL-LIT-micro-fuzzy-relaxed-@5	ALL	0.607	0.580	0.593
EN	aidalight-baseline	NEL-LIT-micro-fuzzy-relaxed-@5	ALL	0.506	0.506	0.506
EN	SBB	NEL-LIT-micro-fuzzy-relaxed-@5	ALL	0.390	0.135	0.200

Conclusion

- OCR performance is crucial → invest into text error cleanup
- Better NER for noisy historical text material is possible (L3i)
- NEL recall performance has biggest potential for easy improvement
    - Construction of KNB on basis of Wikidata
- NEL precision looks promising

Thank you for listening!
Questions?

Staatsbibliothek zu Berlin – Preußischer Kulturbesitz

CLEF2020 - HIPE

23 September 2020

{kai.labusch,clemens.neudecker}@sbb.spk-berlin.de

CLEF 2020 Thessaloniki

Staatsbibliothek
zu Berlin
Preußischer Kulturbesitz

Qurator
Curation Technologies