

Classification and Linking of Named Entities

Specific Classifier vs. Large Language
Model

Pia Schwarz (IDS Mannheim), Florian Barth (SUB Göttingen)

Named Entity Recognition

Domain Adapted Classifier: Academic NER

- Academic Named Entity Recognizer: academic persons, organisations, and research areas
- Pretrained model: [bert-base-german-cased](#)
- Fine-tuning on sentences from newspaper extracted from [DeReKo](#) (Deutsches Referenzkorpus)

Example

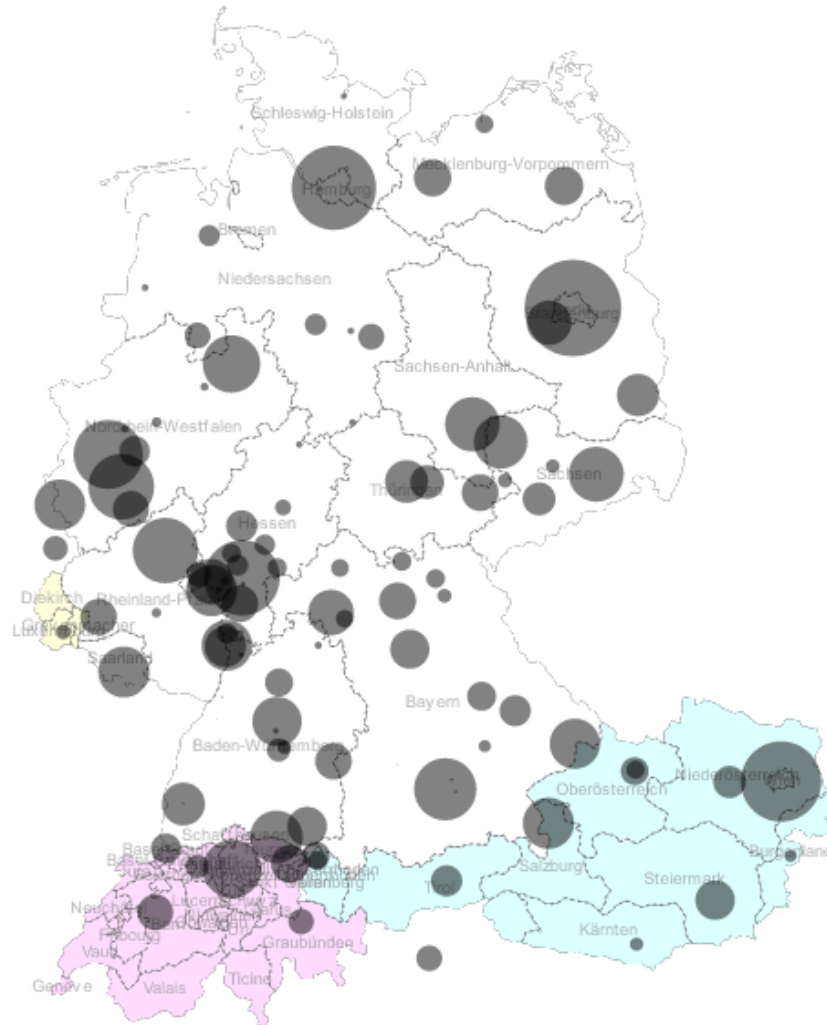
...

[Mathias Jucker]_{PER-RES'}
Professor der [Neurobiologie]_{AREA-RES'}
vom [Hertie-Institut]_{ORG-RES}
der [Universität Tübingen]_{ORG-RES}

...

(...Mathias Jucker, professor of neurobiology from the Hertie Institute of the University of Tübingen...)

Geographische Herkunft der DeReKo-Zeitungsquellen



Domain Adapted Classifier: Academic NER

- Data set of 4,928 manually annotated sentences
- Train/Dev/Test split of 70/20/10
- Using [spaCy transformer library](#)
- Training took less than 1 hour on single Tesla P4 GPU

Training data (3,449 sentences) entity type distribution

Entity Type	Count
PER-RES	2,942
ORG-RES	1,624
AREA-RES	450

LLMs

- Motivation: let LLM tag additional sentences from DeReKo newspaper texts
- Prompt experiments with two LLMs:
 - OpenOrca-Platypus-2 13B
 - Llama-2 13B Chat

LLM Prompting: Few-shot

"SYSTEM: Finde Entitäten wie akademische Personen, akademische Institutionen und akademische Fachrichtungen.

Gib die Entitäten im Wortlaut wieder. Generiere keinen weiteren Text darüber hinaus.

Beispiele:

Text: Gleichzeitig studierte Prof. Roland Girtler an der Rheinischen Friedrich-Wilhelms-Universität Bonn Politikwissenschaften, Öffentliches Recht und Philosophie mit Abschluss MA, gab zwei Fachbücher heraus und machte in der FDP Karriere.

Entitäten: PER: Prof. Roland Girtler; ORG: Friedrich- Wilhelms-Universität Bonn; AREA: Politikwissenschaften | Öffentliches Recht | Philosophie

Text: Bei den Studenten am Erziehungswissenschaftlichen Seminar der Heidelberger Universität und der Humboldt Universität Berlin (HU) regt sich Unmut: Als "unhaltbare und unzumutbare Zustände", dass seit nunmehr sieben Semestern der Lehrstuhl für Sozialpädagogik vakant ist.

Entitäten: PER: -; ORG: Heidelberger Universität | Humboldt Universität Berlin (HU); AREA: Sozialpädagogik

Text: "Die Schädigung im Gehirn folgt dabei dem Dominanzprinzip", sagt der Neurobiologe Mathias Jucker, PhD vom "Hertie-Institut" der Universität Tübingen (vgl. Grafik S. 98).

Entitäten: PER: Mathias Jucker, PhD; ORG: "Hertie-Institut" | Universität Tübingen; AREA: -

USER: Text: " + `target_sentence` + " **ASSISTANT:** "

(`target_sentence` = Satz mit potenziell zu klassifizierenden Entitäten)

Comparison

Model evaluation on test data split of 489 sentences[1]:

	Llama-2 13B Chat	OpenOrca- Platypus-2 13B	AcademicNER
P	88.53	92.48	99.05
R	81.76	62.35	97.16
F1	85.01	74.49	98.10

[1]: compare Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003, pages 142–147.

Learnings (I)

- Prompt length limitation to 512 tokens -> no result if target_sentence is too long
- Zero-shot vs. one-shot vs. few-shot prompt -> acceptable results only with few-shot prompt, i.e. enough examples (tradeoff with limitation of prompt length)
- LLM response with modification of tagged entities, e.g.
Input prompt: "Heidelberger Uni"
Response: "Universität Heidelberg"

Learnings (II)

- Output is not formatted as requested
(i.e. PER: entity1 | entity2 | entity3; ORG: -; AREA:
entity4)
-> complicates post-processing of output
- LLM invents more output than requested:
additional sentences
- Computation time for inference with LLMs rather
slow without GPU:
232 prompts per hour with Intel Core i7-1255U
CPU, 16GB RAM

Conclusion

- It takes time and effort to formulate an efficient prompt...
- ... which also responds with a output format easy to post-process
- Computation power is very useful to speed up inference
- the LLMs in this experiment cannot keep up with classifier trained on specific task (AcademicNER)
- But: LLMs can be very useful to automate data annotation (especially if output is reviewed)
- Assumption/to be tested: Other LLMs or bigger versions (70B) might generate better output

Entity Linking (EL)

- Evaluate linking to Wikidata
- Evaluate on *German* data
- Compare various models with existing data set (annotated with Wikidata IDs)
- Stages for EL:
mention detection -> candidate selection ->
disambiguation

Wikidata



Item [Discussion](#) [Read](#) [View history](#)

Rome (Q220)

capital and largest city of Italy
The Eternal City | Roma | Rome, Italy | City of Seven Hills

[▼ In more languages](#)
[Configure](#)

Language	Label	Description	Also known as
English	Rome	capital and largest city of Italy	The Eternal City Roma Rome, Italy City of Seven Hills
German	Rom	Hauptstadt und bevölkerungsreichste Stadt Italiens	Die Ewige Stadt
French	Rome	capitale de l'Italie	La ville éternelle La ville aux sept collines Roma l'Urbs

[Main page](#)
[Community portal](#)
[Project chat](#)
[Create a new Item](#)
[Recent changes](#)
[Random Item](#)
[Query Service](#)
[Nearby](#)
[Help](#)
[Donate](#)

[Lexicographical data](#)
[Create a new Lexeme](#)
[Recent changes](#)
[Random Lexeme](#)

[Tools](#)

Model: Spacyfishing

- spaCy wrapper for [entity-fishing](#)[2]
- entity recognition and disambiguation against **Wikidata** for several languages
- Tanti Kristanti from Inria Paris used off the shelf version of entity-fishing in the CLEF HIPE 2020 competition shared task, ranking first at the Entity Linking task for English and second best for French, in F1-score

[2]: <https://grobid.s3.amazonaws.com/presentations/29-10-2017.pdf>

Model: DBpedia Spotlight

- links to DBpedia knowledge graph (<http://de.dbpedia.org>) [3]
- DBpedia entities provide link to Wikidata (can easily be retrieved with SPARQL via relation `owl:sameAs`)

[3]:Mendes, P. N., Jakob, M., García-Silva, A., & Bizer, C. (2011). DBpedia Spotlight: Shedding Light on the Web of Documents. In Proceedings of the 7th international conference on semantic systems (pp. 1-8).



👁 Browse using ▾

📄 Formats ▾

rdfs:label

- [Rom \(de\)](#)

owl:sameAs

- [freebase:Rom](#)
- <http://www.viaf.org/viaf/145348938>
- [dbpedia-cs:Rom](#)
- [dbpedia-el:Rom](#)
- [dbpedia-es:Rom](#)
- [dbpedia-eu:Rom](#)
- [dbpedia-fr:Rom](#)
- [dbpedia-id:Rom](#)
- [dbpedia-it:Rom](#)
- [dbpedia-ja:Rom](#)
- [dbpedia-ko:Rom](#)
- [dbpedia-nl:Rom](#)
- [dbpedia-pl:Rom](#)

Model: OpenTapioca

- Links to Wikidata
- Trained on Wikidata dump 2018-02-24

Delpeuch, A. (2019). Opentapioca: Lightweight entity linking for wikidata. arXiv preprint arXiv:1904.09131.

Comparison: Test Data Set HIPE 2020

- HIPE = Identifying Historical People, Places and other Entities

Evaluation campaign on **named entity processing on historical newspapers** in French, German and English[4]

[4]: <https://impresso.github.io/CLEF-HIPE-2020/>

Comparison: Test Data Set HIPE 2020

HIPE-data-v1.4-test-de[5]

- 49 German texts from Swiss and Luxembourg newspapers from mid-18th to mid-19th century
- OCRred text
- Annotated with sentence boundaries, and Q-items from Wikidata dump of 2019.11.13[6]

[5]: <https://zenodo.org/records/6046853> [6]: <https://github.com/hipe-eval/HIPE-2022-data/blob/main/documentation/README-hipe2020.md>

Comparison: Test Data Set HIPE 2020

TOKEN	NE-COARSE-LIT	NE-COARSE-METO	NE-FINE-LIT	NE-FINE-METO	NE-FINE-COMP	NE-NESTED	<u>NEL-LIT</u>	NEL-METO	MISC
# language = de									
# newspaper = NZZ									
# date = 1798-01-20									
# document_id = NZZ-1798-01-20-a-p0002									
Neuigkeiten	0	0	0	0	0	0	-	-	NoSpaceAfter
.	0	0	0	0	0	0	-	-	<u>PySBDSegment</u>
(0	0	0	0	0	0	-	-	NoSpaceAfter
Mißhandlung	0	0	0	0	0	0	-	-	-
der	0	0	0	0	0	0	-	-	EndOfLine
Franken	0	0	0	0	0	0	-	-	-
in	0	0	0	0	0	0	-	-	-
Rom	B-loc	0	B-loc.adm.town	0	0	0	<u>Q220</u>	-	NoSpaceAfter LED0.00
.	0	0	0	0	0	0	-	-	NoSpaceAfter <u>PySBDSegment</u>
]	0	0	0	0	0	0	-	-	-
Buonaparte	B-pers	0	B-pers.ind	0	0	0	<u>Q517</u>	-	NoSpaceAfter LED0.00
'	I-pers	0	I-pers.ind	0	0	0	<u>Q517</u>	-	NoSpaceAfter LED0.00
s	I-pers	0	I-pers.ind	0	0	0	<u>Q517</u>	-	LED0.00
Bruder	0	0	0	0	0	0	-	-	-
hat	0	0	0	0	0	0	-	-	EndOfLine
die	0	0	0	0	0	0	-	-	-

Comparison: Results

All models use spaCy's German NER from language model [de_core_news_lg](#)

Model	F1	P	R	TP	FP	FN
Spacyfishing	0.28	0.351	0.233	234	432	772
DBpedia Spotlight	0.29	0.192	0.557	675	2837	536
OpenTapioca	0.421	0.407	0.437	497	724	641
HIPE Task 2022 team2[^7]	0.464	0.462	0.466	535	623	612

[7]: Best score on the Task End-to-end EL hipe2020 German relaxed @1 (literal sense) from Team "L3i" from La Rochelle University, France, <https://hipe-eval.github.io/HIPE-2022/results#hipe-2022-track-evaluation-results>)

Comparison: Results

- Evaluated using the "fuzzy regime" as in HIPE 2022: counted as TP if token boundaries do not match exactly but entity type (Q-item) is the same
- DBpedia Spotlight Q-item retrieved through DBpedia-Wikidata interlinking, in 33 cases a DBpedia entity was found but no corresponding Q-item was interlinked, e.g.:

in	0	0	0	0	0	0	-	-	-	-
der	0	0	0	0	0	0	-	-	-	-
Gegend	0	0	0	0	0	0	-	-	-	-
von	0	0	0	0	0	0	-	-	EndOfLine	_
Laffaux B-loc	0	B-loc	adm.town	0	0	0	Q842559	-	LED0.00	WIKIDATA-Q-ITEM-NOT-FOUND-FOR:http://de.dbpedia.org/resource/Laffaux
wurden	0	0	0	0	0	0	-	-	-	-
abgewiesen	0	0	0	0	0	0	-	-	NoSpaceAfter	_

Code and results:

https://github.com/piaschwarz/Entitylinking_Evaluation

LLMs and Entity Linking

ChatGPT & Wikidata



You

Bitte annotiere die Named Entities im folgenden Text und verlinke sie mit Wikidata:

Phileas Fogg vermuthete wohl bei seiner Abreise von London nicht, welch' großes Aufsehen sein Vorhaben erregen würde. Die Neuigkeit von der Wette verbreitete sich zuerst im Reformclub und erregte unter den Mitgliedern dieser ehrenwerthen Gesellschaft eine arge Aufregung, welche sich durch die Berichterstatter von da in die Journale verbreitete und durch diese das Publicum von London und im ganzen Vereinigten Königreiche durchdrang.

- Named Entities:
 - Phileas Fogg (Q424522)
 - London (Q84)
 - Reformclub (Q1806762)
 - Vereinigtes Königreich (Q1860)

- only very common entities are identified correct (London Q84)
- other Wikidata IDs are pure fantasy

ChatGPT & Wikidata



You

Drei der Wikidata-Links sind falsch. Woher hast du die Informationen?



ChatGPT

Entschuldigung für die Verwirrung. Mein vorheriges Verständnis basierte auf dem allgemeinen Wissen bis zu meinem letzten Aktualisierungsdatum im Januar 2022. Für genauere und aktuellere Informationen wäre es am besten, direkt auf der Wikidata-Website nachzusehen. Wenn es seit meiner letzten Aktualisierung Änderungen gab, können die Wikidata-IDs möglicherweise aktualisiert worden sein.

- Very vague knowledge of origin

ChatGPT & Wikidata – Varying types of responses



You

Bitte verlinke die Entitäten jetzt mit Wikidata IDs, wenn möglich.



ChatGPT

Leider kann ich keine direkten Verknüpfungen mit Wikidata-IDs erstellen, da ich nicht über eine Echtzeitverbindung zu Datenbanken verfüge und keine Internetrecherche durchführen kann. Sie können jedoch die Verknüpfungen manuell überprüfen und die entsprechenden Wikidata-IDs zu den genannten Personen, Orten und Titeln suchen.

Sie können die Wikidata-Suche verwenden: <https://www.wikidata.org/wiki/Special:Search>, um die relevanten IDs zu finden. Geben Sie den Namen der Entität ein, suchen Sie nach Übereinstimmungen und wählen Sie dann die entsprechende Seite aus, um die Wikidata-ID zu finden.

- Sometimes chat GPT knows better about its resources

Other LLMs & Wikidata

- tested models:
 - llama2
 - nous-hermes-llama2
 - em_german_mistral_v01.Q4_0
 - sauerkrautlm-7b-hero
- each model returns Wikidata IDs
- all of them are fantasy

How to improve knowledge of the model?

→ enrich LLM with structured graph data from Wikidata

Existing studies

Using Multiple RDF Knowledge Graphs for Enriching ChatGPT Responses (Michalis Mountantonakis and Yannis Tzitzikas)

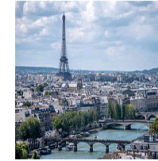
- Provides [Annotation and Enrichment Service](#)
- Utilises own entity linking system (LODsyndesisIE)
- Simply maps results LODsyndesisIE (i.a. OpenTapioca) to ChatGPT responses
- No ingest of data or knowledge to the LLM

Correct response from ChatGPT

GPT•LODS

Annotated ChatGPT Response with links to H

Paris



Place

[RDF Datasets](#)

[URIs](#)

[Facts](#)

Knowledge Graphs

User Prompt: **Wie** lautet die Wikidata-ID von **Paris**

ChatGPT Annotated Response:

Die Wikidata-ID von **Paris** lautet: **Q90**

Entities Information

Facts between Entities

Fact Checking (with relations)

HTML+RDFa

Ask Again

Link


False response from ChatGPT

(while textual references are linked correctly)

GPT•LODS

Annotated ChatGPT Response with links to Hundreds of Graphs

Dietrich Bonhoeffer



Person
[RDF Datasets](#)
[URIs](#)
[Facts](#)

User Prompt: **Wie** lautet die Wikidata-ID von **Dietrich Bonhoeffer**

ChatGPT Annotated Response:

Die Wikidata-ID von **Dietrich Bonhoeffer** lautet Q59815.

Entities Information Facts between Entities Fact Checking (with relations) HTML+RDFa Ask Again

Link

Prompts in ChatGPT

Correct (simple entity)



You

Wie lautet die Wikidata-ID von Paris?



ChatGPT

Die Wikidata-ID für Paris ist Q90.

False (demanding entity)



You

Wie lautet die Wikidata-ID von Dietrich Bonhoeffer?



ChatGPT

Die Wikidata-ID von Dietrich Bonhoeffer ist Q7633.

Experiment: Data ingest

Build a knowledge document based on Wikipedia and Wikidata

This article is about Dietrich Bonhoeffer.

+

The Wikidata identifier of the article is Q76326.

+

The Wikidata identifier for Dietrich Bonhoeffer is Q76326.

+

Dietrich Bonhoeffer (German: [ˈdiːtʁɪç ˈbɔnhøːfɐ] ; 4 February 1906 – 9 April 1945) was a German Lutheran pastor, theologian and anti-Nazi dissident who was a key founding member of the Confessing Church. His writings on Christianity's role in the secular world have become widely influential; his 1937 book ...

Embedding model with contextual knowledge

- Initiate embedding model (from llama2) in vector store
- Vectorise context document (chunk_size=500) and store in vector database
- Build chain of prompt and embedding model

Question

What is the Wikidata identifier of Dietrich Bonhoeffer?

Received answer

The Wikidata identifier of Dietrich Bonhoeffer is Q76326.



Next steps

- Fine-tune model with knowledge from Wikidata
- Ingest structured data in RDF format