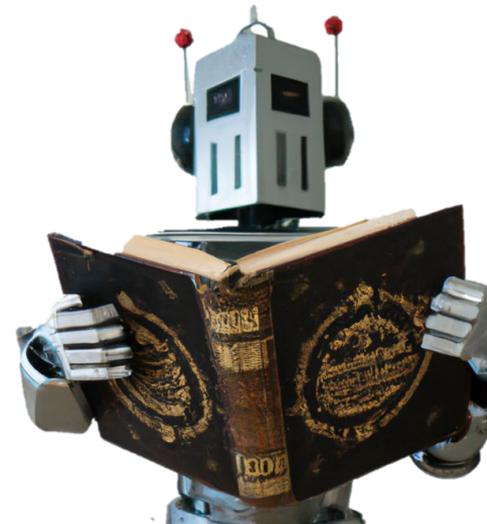


NER, aber prompto!

Named Entity Recognition mit Large
Language Models für historische Texte

Torsten Hiltmann, Martin Dröge, Nicole Dresselhaus, Sophie Eckenstaler, Till Grallert,
Jascha Schmitz, Philipp Schneider, Melanie Althage (Humboldt-Universität zu Berlin,
NFDI4Memory)



Intro



- ChatGPT (November 2022)
- Veränderungen durch große Sprachmodelle (LLM) naheliegend
- offene Frage: Was und wie genau?



“Der Elefant im Raum” (erwähnt auf der Digital History Tagung 2023, Bild: erstellt mit Dall-E 2)

Gliederung

1. Named Entity Recognition (NER) und die Geschichtswissenschaften
2. Large Language Models als Game Changer
3. Der Baedeker-Reiseführer als Evaluationskorpus
4. Experimente
5. Ergebnisse und Einordnung

1

Named Entity Recognition (NER) und die Geschichtswissenschaften



Named Entity Recognition (NER)

Named Entities sind Wörter oder Wortfolgen in Texten, die auf bestimmte einzigartige benenn- bzw. bezeichnenbare Konzepte verweisen, die sowohl konkret wie abstrakt sein können.

Abhängig von Domäne und Anwendungsfall bzw. Modellierung:

- **universell:** Person, Ort, Organisation
- **Naturwissenschaften:** chemische Stoffe, Proteine, Tierarten, astronomische Objekte, ...
- **Medizin:** Krankheiten, Medikamente, Gene, ...
- **Wirtschaft:** Produkte, Firmennamen, Währungsangaben, ...
- **Geschichtswissenschaft:** Amtsbezeichnungen, Werktitel, Ereignisse, abstrakte Konzepte, Datumsangaben, ...

Named Entity Recognition (NER)

Named Entities sind Wörter oder Wortfolgen in Texten, die auf bestimmte einzigartige benenn- bzw. bezeichnenbare Konzepte verweisen, die sowohl konkret wie abstrakt sein können.

Named Entity Recognition ist ein NLP-Verfahren, um in unstrukturierten Texten genau diese Verweise auf Named Entities zu detektieren und in zuvor definierte Kategorien zu klassifizieren.

NB: Disambiguierung oder Identifikation sind nicht Teil der NER!

Karl Baedeker (* 3. November 1801 in Essen ; † 4. Oktober 1859 in Koblenz ,
zeitgenössische Schreibweise: Karl Bædeker) war ein deutscher Verleger und als Autor
Begründer der Baedeker-Reiseführer.

unstrukturierte Textdaten

Karl Baedeker (* 3. November 1801 in Essen ; † 4. Oktober 1859 in Koblenz ,
zeitgenössische Schreibweise: Karl Bædeker) war ein deutscher Verleger und als Autor
Begründer der Baedeker-Reiseführer.

1. Detektion

Karl Baedeker PER (* 3. November 1801 in Essen LOC ; † 4. Oktober 1859 in Koblenz LOC ,
zeitgenössische Schreibweise: Karl Bædeker PER) war ein deutscher Verleger und als Autor
Begründer der Baedeker-Reiseführer.

2. Klassifikation

Token classification task:

t_1, t_2 : PER; t_9, t_9 : LOC; t_{16}, t_{16} : LOC; t_{22}, t_{23} : PER.

Herausforderungen in den Geschichtswissenschaften

Viele Domänen arbeiten meist mit umfangreichen und gleichförmigen Korpora auf aktueller Sprachebene: Wissenschaftliche Publikationen, Patientenakten, Geschäftsberichte, Patente,

Die NE haben dabei meist das Ziel, eindeutig zu sein.

Sprache

- unterschiedliche Sprachen
- unterschiedliche Sprachstufen
- lange Zeit fehlende einheitliche Schreibweisen
- unterschiedliche editorische Aufbereitung der Texte (diplomatisch, normalisiert)

Textstrukturen und verwendetes Vokabular:

- unterschiedliche Textgenres (Chroniken, Urkunden, Akten, Briefe, Presse, ...)
- sich verändernde kulturelle Gepflogenheiten (z.B. Anreden, Verwendung von Titeln)
- unterschiedliche Domänen (Recht, Wirtschaft, Frömmigkeit, Militär, Esskultur, ...)
- sich verändernde Wortbedeutungen
- Eindeutige Identifikation und Klassifikation der NE setzen meist eine erhöhte Expertise für Zeit und Kontext voraus.

warum das wichtig ist ? →

Bisherige Technologien (vereinfacht)

	regelbasierte Ansätze	Maschinelles Lernen (alt und neu)
<i>Vorgehen</i>	Erstellung von Sets spezifischer Regeln, die auf die Grammatik, den Kontext und die Position von Wörtern im Text basieren, um Entitäten zu erkennen.	Merkmalbasierte Ansätze unter Verwendung statistischer Verfahren; oder end-to-end Deep-Learning Ansätzen, um aus Trainingsdaten Muster zu lernen. Ggf. Finetuning bestehender Modelle
<i>Methoden</i>	Verwendung von Wörterbüchern, regulären Ausdrücken und linguistischen Mustern.	alt: HMM, CRF neuer: CNN, RNN, LSTM; Transformer (BERT)
<i>Voraussetzungen</i>	Hohe Expertise für die jeweiligen Quellen und deren Sprache und sprachlichen Muster, um effektive Regeln zu entwickeln.	Umfangreiche Trainingsdaten von hoher Qualität (Expertise für Erstellung) , Rechenleistung, ggf. geeignete vortrainierte Modelle
<i>Aufwand</i>	Sehr hoch	Sehr hoch bis hoch
<i>Übertragbarkeit (Transferabilität)</i>	Sehr gering , da die Regeln genau auf spezifische Muster oder Vokabular zugeschnitten sind	gering bis gut , da gelernte Muster von Schreibweisen*, Textstrukturen und Vokabular der Trainingsdaten abhängen. Neuere Modelle aber durch Transfer Learning mit weiteren Trainingsdaten anpassbar

Fehlende Transferabilität

besonders häufig in den Geschichtswissenschaften genutzte Workframes für NER:

Trainingsdaten

The logo for spaCy, featuring the word "spaCy" in a blue, lowercase, sans-serif font. The "a" and "y" are slightly larger and more prominent.

- [TIGER Corpus](#) - Artikel aus der Frankfurter Rundschau vor 2003, ohne Sport und Regionales
- [WikiNER](#) - teils automatisch erstellter Korpus auf Grundlage von Wikipedia (2013)

The logo for flair, featuring the word "flair" in a black, lowercase, sans-serif font. The "a" is orange and has a small flame icon above it.

- [Conll2003](#) Korpus, auf Grundlage der Frankfurter Rundschau, händisch annotiert

Auszug aus *Chronik aus Kaiser Sigismund's Zeit*

erstellt mit SpaCy de_core_news_sm (Demonstrator)

Item in demselben jar da weihet man dy groẓen glocken zu fant Sebolt an unfers herren leichnames abent und dy heiffet man **Benedicta** **LOC**, darnach hieng man fy auf.

Mittelhochdeutsches Original

Im selben Jahr weiht man die große Glocke zu **Sankt Sebolt** **LOC** am Abend des Fronleichnamfestes und nennt sie Benedicta, danach hängt man sie auf.

im modernen Deutsch

In **the same year** **DATE**, the large bells at Saint Sebolt were consecrated on the evening of **Corpus Christi** **PRODUCT** and were named **Benedicta** **ORG**; afterwards, they were hung up.

Übersetzt in Englisch mit Modell für Englisch

Beispiel für kulturelle Praktiken (Titel) und Vokabular

erstellt mit spaCy de_core_news_sm (Demonstrator)

Die erste flüchtige Begegnung zwischen dem jungen Geheimrat Schlettstett und dem Akademiezöglinge Schultheiß hatte 1779 in **Stuttgart** **LOC** stattgefunden.

Bezeichnung Personen mit Titel und Nachnamen

Die erste flüchtige Begegnung zwischen dem jungen Geheimrat **Goethe** **PER** und dem Akademiezöglinge **Schiller** **PER** hatte 1779 in **Stuttgart** **LOC** stattgefunden.

Namen ausgetauscht mit bekannten Namen (entspricht dem original-Text)

Peter Pane **PER** hat seine Burger auf Twitter gepostet.

Peter Pane **PER** hat seine Burger auf **Microsoft** **ORG** gepostet.

Peter Pane gegründet 2012, Twitter 2006

Bisherige Strategien

- a. ungeprüft anwenden und hoffen
- b. umfangreiche Erstellung von Trainingsdaten zum Training bzw. Nachtraining eines für den Anwendungsfall spezifischen Modells
- c. Anpassung der eigenen Texte, z.B. durch weitreichende Normalisierung, um doch noch mit bestehenden Modellen arbeiten zu können.

2

Large Language Models als Gamechanger

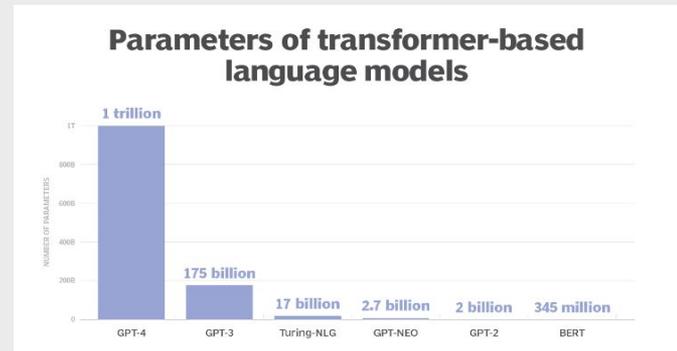


LLM als Gamechanger

- LLM basieren in Teilen auf der Transformerarchitektur, die schon in BERT erfolgreich verwendet wurde (nur statt encoder, hier decoder)
- BERT etc. musste jedoch für spezifische Aufgaben nachtrainiert werden
- **Skalierung** der Large Language Models bringt grundlegende Neuerungen
- **Neu - Kontext:** LLM haben durch Skalierung eigentlich schon alles trainiert

Dataset	Quantity (tokens)	Weight in training mix	Epochs elapsed when training for 300B tokens
Common Crawl (filtered)	410 billion	60%	0.44
WebText2	19 billion	22%	2.9
Books1	12 billion	8%	1.9
Books2	55 billion	8%	0.43
Wikipedia	3 billion	3%	3.4

Trainingsdaten für GPT3 (Brown, Ryder, Mann et al. (2020))



[Diagramm: TechTarget](#)

Kontext

Basierend auf der Repräsentation von Informationen und Aussagen in den Trainingsdaten

Bitte erläutere die Entitäten, die in folgendem Auszug einer Quelle aus der Mitte des 15. Jahrhunderts genannt werden:

Jtem anno domini 1206 jar da hub sich an der barfue[3er orden.



Der Auszug, den du angegeben hast, bezieht sich auf das Jahr 1206 und die Gründung des Barfüßerordens. Hier sind die spezifischen Entitäten erläutert:

1. **Anno Domini 1206:** Diese lateinische Phrase bedeutet „im Jahre des Herrn 1206“ und gibt das Jahr an, auf das sich der Text bezieht.
2. **Barfüßerorden:** Der Barfüßerorden, besser bekannt als der Orden der Franziskaner, wurde von Franz von Assisi gegründet. Der Name „Barfüßer“ kommt von der Praxis der Mönche dieses Ordens, barfuß oder nur mit Sandalen bekleidet zu gehen, als Zeichen ihrer Demut und Armut. Der Franziskanerorden ist bekannt für seine Betonung auf Armut, Einfachheit und dem direkten Dienst an den Armen, und wurde 1209 offiziell von Papst Innozenz III. anerkannt.

Der Textabschnitt deutet darauf hin, dass die Gründung oder der Beginn des Barfüßerordens im Jahr 1206 in Betracht gezogen wird, obwohl historisch gesehen der Orden formell erst 1209 anerkannt wurde. Möglicherweise bezieht sich der Text auf die frühen Aktivitäten von Franz von Assisi und seinen Anhängern vor der offiziellen Anerkennung ihres Ordens.

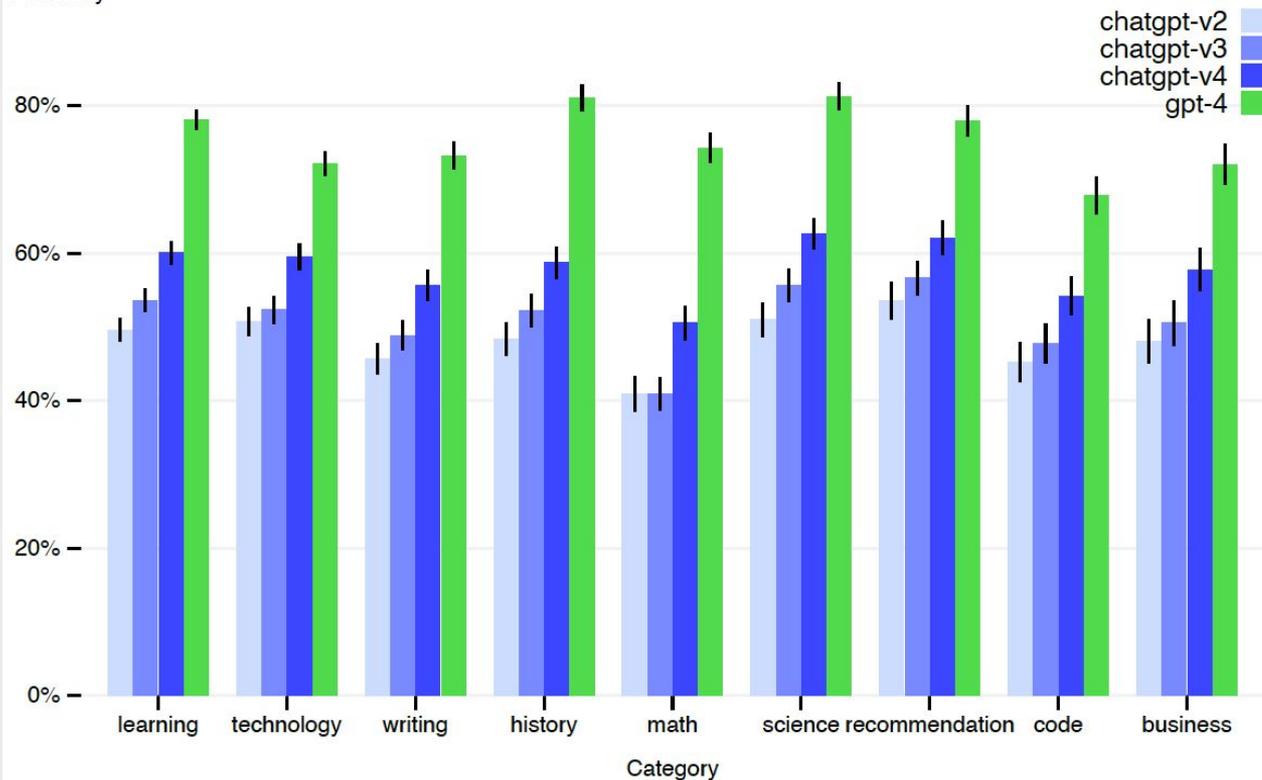
ChatGPT-4, 24.06.2024

Kontext

Basierend auf der Repräsentation von Informationen und Aussagen in den Trainingsdaten

Internal factual eval by category

Accuracy



OpenAI (2023), GPT-4 Technical Report, 10.

LLM als Gamechanger

- LLM basieren in Teilen auf der Transformerarchitektur, die schon in BERT erfolgreich verwendet wurde (nur statt encoder, hier decoder)
- BERT etc. musste jedoch für spezifische Aufgaben nachtrainiert werden
- **Skalierung** der Large Language Models bringt grundlegende Neuerungen
- **Neu - Kontext:** LLM haben durch Skalierung eigentlich schon alles trainiert
- **Neu - Fähigkeiten:** In Context Learning als emerging ability

Language Models are Few-Shot Learners

Tom B. Brown*	Benjamin Mann*	Nick Ryder*	Melanie Subbiah*	
Jared Kaplan ¹	Prafulla Dhariwal	Arvind Neelakantan	Pranav Shyam	Girish Sastry
Amanda Askell	Sandhini Agarwal	Ariel Herbert-Voss	Gretchen Krueger	Tom Henighan
Rewon Child	Aditya Ramesh	Daniel M. Ziegler	Jeffrey Wu	Clemens Winter
Christopher Hesse	Mark Chen	Eric Sigler	Mateusz Litwin	Scott Gray
Benjamin Chess	Jack Clark	Christopher Berner		
Sam McCandlish	Alec Radford	Ilya Sutskever	Dario Amodei	

OpenAI

Abstract

Recent work has demonstrated substantial gains on many NLP tasks and benchmarks by pre-training on a large corpus of text followed by fine-tuning on a specific task. While typically task-agnostic in architecture, this method still requires task-specific fine-tuning datasets of thousands or tens of thousands of examples. By contrast, humans can generally perform a new language task from only a few examples or from simple instructions – something which current NLP systems still largely struggle to do. Here we show that scaling up language models greatly improves task-agnostic,

Parameters of transformer-based language models



[Diagramm: TechTarget](#)

Brown, Mann, Ryder et al. (2020), Language Models are few-shot learners

Emergent Abilities

Von der Sprachmodellierung zum **Lösen von Aufgaben**

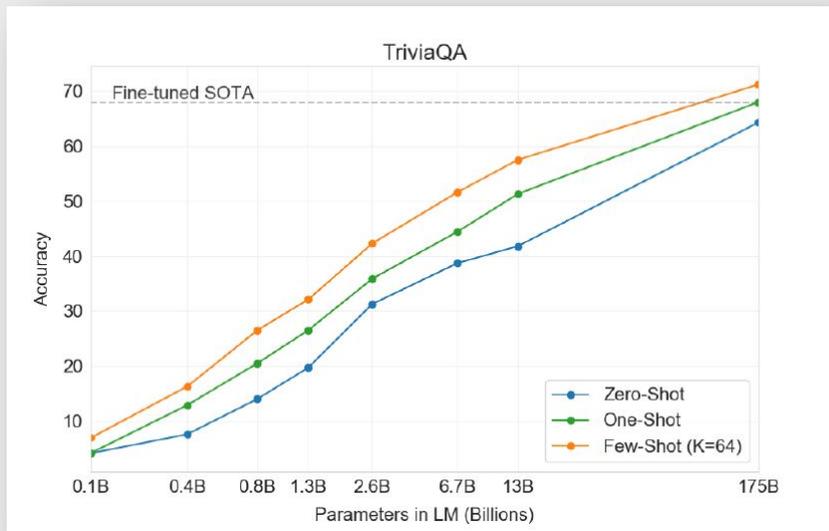


Diagramm: Brown, Mann, Ryder et al. (2020), 14

Neue Fähigkeiten (*emergent abilities*) ohne dafür trainiert worden zu sein, **einfach durch Fortsetzen eines gegebenen Textes (Prompt) und Generierung des jeweils nächsten Wortes:**

- **In-context learning**
- **Instruction following**
- **Step-by-step Reasoning**

In-context Learning

In-Context Learning nutzt das während des Pre-Trainings erworbene breite semantische Vorwissen großer Sprachmodelle, um neue Aufgaben allein durch einige wenige im Prompt gegebene Beispiele oder dort genannte Schlüsselwörter zu lösen, ohne dass ein zusätzliches Training der Modellparameter erforderlich ist.

Perplexity, Focus: academic, 25.06.2024

```
1 Translate English to French: ← task description
2 cheese => ..... ← prompt
```

zero-shot

```
1 Translate English to French: ← task description
2 sea otter => loutre de mer ← example
3 cheese => ..... ← prompt
```

one-shot

```
1 Translate English to French: ← task description
2 sea otter => loutre de mer ← examples
3 peppermint => menthe poivrée ←
4 plush girafe => girafe peluche ←
5 cheese => ..... ← prompt
```

few-shot

Beispiele: Brown, Mann, Ryder et al. (2020), 7

State of the art

However, despite their impressive performance in numerous NLP tasks, LLMs have shown some limitations when applied to NER. This is primarily because NER is a sequence labeling task, while LLMs are originally designed for text generation.

Keraghel, Morbieu, Nadif (2024), 11

	NewsEye			hipe-2020			ajmc		
	P	R	F1	P	R	F1	P	R	F1
	strict								
<i>Stacked NERC</i>	75.0	70.6	72.7	-	-	-	-	-	-
<i>Temporal NERC</i>	-	-	-	76.5	76.5	76.5	84.8	83.9	84.4
ChatGPT	70.9	72.3	71.6	32.5	50.0	39.4	21.8	26.1	23.8
	fuzzy								
<i>Stacked NERC</i>	85.4	80.5	82.9	-	-	-	-	-	-
<i>Temporal NERC</i>	-	-	-	86.7	86.7	86.7	90.2	89.2	89.7
ChatGPT	77.8	79.4	78.6	49.0	75.4	59.4	25.5	30.6	27.8

González-Gallardo, Boros, Doucet et al. (2023), [3] (Hervorhebungen TH)

“We followed a straightforward zero-shot approach to retrieve named entities from ChatGPT via the official web interface between January 11th and February 7th, 2023.”

González-Gallardo, Boros, Doucet et al. (2023), [2]

Frameworks	Algorithms	Macro-averaged F1-score									
		CoNLL-2003	OntoNotes	WNUT2017	FIN	BioNLP2004	NCBI Disease	BC5CDR	MITRestaurant	Few-NERD	MultiCoNER
Apache OpenNLP Stanford CoreNLP Flair	Maximum Entropy	80.00	67.83	-	<u>63.24</u>	-	-	-	-	-	-
	CRF	85.18	63.87	8.34	55.25	73.26	86.10	85.22	70.57	45.13	19.39
	LSTM-CRF	<u>90.35</u>	80.10	38.07	74.23	<u>71.64</u>	<u>86.21</u>	90.27	78.33	<u>60.03</u>	56.27
spaCy	CNN-small	81.26	69.30	9.01	55.12	65.92	77.92	80.83	75.62	40.55	35.63
	CNN-large	85.64	69.60	9.78	54.71	66.17	79.15	79.66	76.39	40.01	35.82
	roberta-base	89.92	<u>81.04</u>	<u>41.84</u>	63.18	66.56	87.05	87.08	79.09	59.15	55.21
Hugging Face	xlm-roberta-large	91.46	81.57	43.92	48.68	71.43	85.25	<u>87.41</u>	80.12	61.59	<u>58.15</u>
	distilbert-base-cased	88.12	77.63	25.45	43.74	69.63	84.42	84.03	77.67	58.62	55.17
	bert-base-uncased	88.89	76.99	32.12	46.84	70.50	85.64	85.78	<u>79.18</u>	58.16	59.96
	bert-base-cased	90.09	79.55	33.32	39.53	69.46	85.27	85.14	78.48	59.48	56.64
OpenAI	GPT-4	62.74	33.61	18.82	36.70	41.32	57.46	55.67	41.38	44.96	33.61

Keraghel, Morbieu, Nadif (2024), 20 (Hervorhebung TH)

State of the art

However, despite their impressive performance in numerous NLP tasks, LLMs have shown some limitations when applied to NER. This is primarily because NER is a sequence labeling task, while LLMs are originally designed for text generation.

Keraghel, Morbieu, Nadif (2024), 11

Strategien:

- verschiedene Wege, das Modell zu einer sequentiellen Ausgabe zu zwingen (zusätzliche Zeichen, codebasierte LLM)
- Verbesserung der Selektion der Beispiele (z.B. durch Erhöhung Ähnlichkeit)
- Aber: immer an Technik und Benchmarks ausgerichtet

Beispiele Prompts

NewsEye	hipe-
What are the locations (LOC), persons (PER), organizations (ORG) and human productions (HumanProd) present in the following historical text? <i>{SENTENCE}</i> Respond, for each word, using IOB or BIO format separated by tab. If a word has no entity, add O.	What are the locations (pers), organizations and time periods (time) present in the following historical text? Respond, for each word, using IOB or BIO format separated by tab. If a word has no entity, add O.

We followed a straightforward zero-shot approach to retrieve named entities from C González-Gallardo, Boros, Doucet et al. (2023), [2]

I You

Prompt: Please identify the named entities in the sentence provided, classifying them as either "CHM" for chemical entities or "DIS" for disease entities. "CHM" refers to chemical substances, including drugs, compounds, biomolecules, and more, like "Aspirin" or "Carbon Dioxide". "DIS" pertains to medical conditions or diseases, such as "Diabetes" or "Glioblastoma". Present your findings in JSON format, with keys "CHM" and "DIS", listing the identified entities under the appropriate category.

Sentence for analysis: "In this model of chronic renal failure."

ChatGPT

Based on the provided sentence, here are the identified named entities classified as "CHM" and "DIS" in JSON format:

```
json Copy code
{
  "CHM": [],
  "DIS": ["chronic renal failure"]
}
```

In this sentence, "chronic renal failure" is a disease entity (DIS). There are no chemical entities (CHM) mentioned in the sentence.

Figure 8: An example of a prompt for detecting named entities in the NCBI Disease dataset.

Keraghel, Morbieu, Nadif (2024), 20

3. LLM als Gamechanger: unser Ansatz

Aufgabe neu definieren

- Aufgabe nicht als NLP-Aufgabe definieren (sprachliche Dimension), sondern als geschichtswissenschaftliche Aufgabe (inhaltliche Dimension)
- i.e. konkreter Einzelfall statt generische Benchmarks
- Orientierung am menschlichen Annotationsprozesses, i.e. Einbezug Kontext und externer Informationen

Prompt Engineering

- In-Context Learning nutzen, in Bezug auf
 - das genaue Befolgen von Anweisungen (Prozess genau beschreiben, Output beschreiben, Chain-of-Thought)
 - in Modell enthaltenes Wissen nutzbar machen (Context mitgeben)
- Nutzung weiterer Prompting-Techniken (Rolle, Reihenfolge, Anreize setzen, few-shot learning, etc.)

Umsetzung

- Statt generischer Prompts spezifische Anpassung an jeden Anwendungsfall (vor Hintergrund Domänenwissen)
- **wichtig: semantisch explizit und präzise prompten, um notwendige Regionen im Netz zu aktivieren**

3

Der Baedeker-Reiseführer als Evaluationskorpus



Ziele des Projekts

- NER mit LLM als Methode in der Geschichtswissenschaft reflektiert und methodenkritisch anwenden
- Evaluation der Methode, um die Aussagefähigkeit der Ergebnisse bestimmen zu können
- Performanz von NER für spezifisch geschichtswissenschaftliche Fragestellungen und Quellenkorpora evaluieren

Beispiele für bestehende Evaluation-Datensets

- CoNLL 2003 (Tjong Kim Sang/De Meulder 2003)
- NoSta-D NE - GermEval 2014 (Bernikova et. al. 2014)
- HIPE 2022 (Ehrmann et. al. 2022)

**Introduction to the CoNLL-2003 Shared Task:
Language-Independent Named Entity Recognition**

Erik F. Tjong Kim Sang and Fien De Meulder
CNTS - Language Technology Group
University of Antwerp

NoSta-D Named Entity Annotation for German: Guidelines and Dataset

Darina Benikova¹, Chris Biemann¹, Marc Reznicek²

(1) FG Language Technology, Comp. Sci. Dept., TU Darmstadt, Germany
(2) Facultad de Filología, Universidad Complutense de Madrid, Spain

HIPE – Identifying Historical People, Places and other Entities

Shared Task on Named Entity Recognition and Linking in
Multilingual Historical Documents

**Extended Overview of HIPE-2022: Named Entity
Recognition and Linking in Multilingual Historical
Documents**

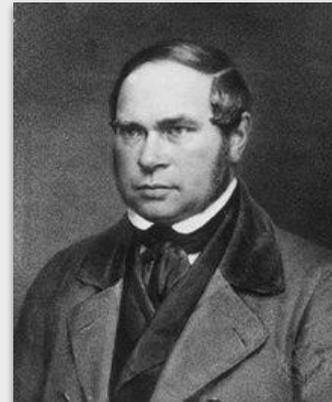
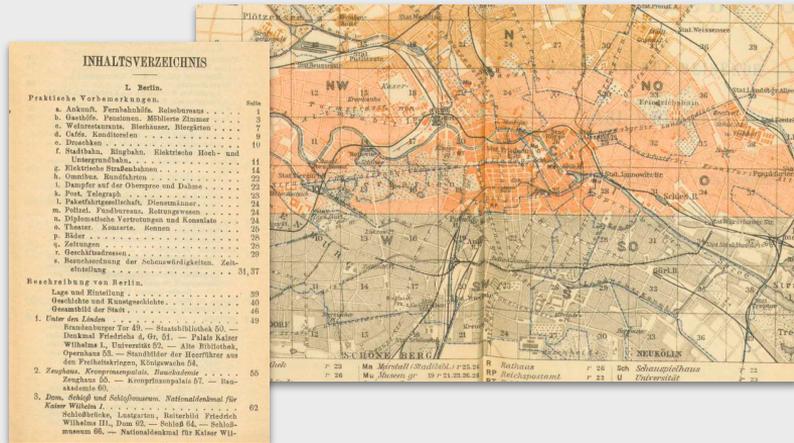
Maud Ehrmann¹, Matteo Romanello², Sven Najem-Meyer¹, Antoine Doucet³ and
Simon Clematide⁴

Warum ein eigenes Evaluationsset?

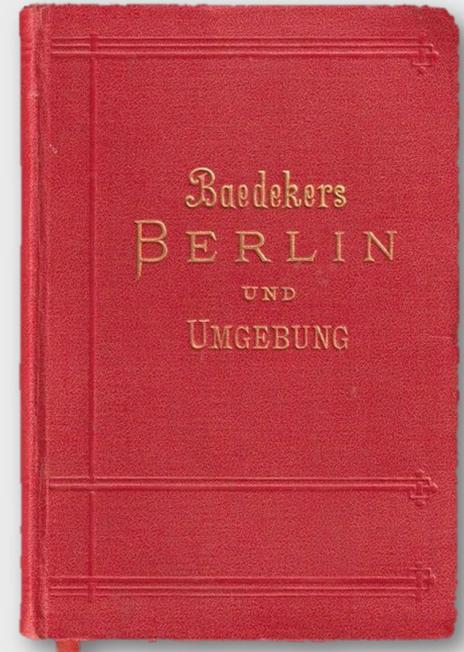
- nicht NLP-Benchmarks der Modelle verbessern
⇒ sondern in der Domäne der Geschichtswissenschaften das Verfahren NER mittels LLM für geschichtswissenschaftliche Fragestellungen evaluieren
- kein one-fits-all-Modell erstellen
⇒ sondern spezial und granular mit historischen Quellen arbeiten
ABER: ein one-fits-all Verfahren entwickeln, anpassbar auf spezifische Anwendungsfälle in der historischen Forschung
- LLM für NER funktioniert kontextbezogen – hier liegt eine Stärke
⇒ daher Evaluation losgelöst von Standard-Evaluationskorpora

Der Baedeker

- im 19. Jahrhundert: Synonym für Reiseführer
- revolutionierte die Reiseliteratur
- Übersichtlichkeit, Genauigkeit und Aktualität
- regelmäßige Überarbeitungen
- Wegbereiter des Massentourismus



Karl Baedeker (1801-1859)



Von uns verwendet:
Baedekers Berlin und
Umgebung, 1921.

Zur Komplexität des Textgenres

Fließtext

Die erste Beschreibung Berlins in Bædekers Reisehandbüchern war noch von *Karl Bædeker* (1801-59), dem Gründer der Firma, verfaßt und füllte in seinem 1842 erschienenen „Handbuch für Reisende durch Deutschland und den österreichischen Kaiserstaat“ 23 Seiten. Der stetig wachsende Stoff zwang allmählich zu Teilungen des Buches in mehrere Bände, aber auch in diesen konnte Berlin nur ein verhältnismäßig beschränkter Abschnitt gewidmet werden, der seit 1878 auch als Sonderdruck aus dem Bande „Norddeutschland“ erschien. 1883 wurde er, entsprechend der wachsenden Bedeutung der Reichshauptstadt, zu einem selbständigen Handbuch erweitert, dessen Neubearbeitungen bis 1914 schnell aufeinander folgten. Der unglückliche Ausgang des Weltkriegs hat dem kaiserlichen Berlin ein Ende bereitet und die Herausgeber vor neue Aufgaben gestellt. Zu unserer Freude erklärte sich Prof. Dr. *Bruno Graupe*, seit 1896 unser Mitarbeiter, auch diesmal wieder bereit, die Erneuerung des Buches zu übernehmen und führte sie auf Grund eigener Beobachtungen und an zuverlässigen Stellen eingezogener Erkundigungen mit gewohnter Sorgfalt durch. Besondere Schwierigkeiten boten dabei die fast unausgesetzten Neuordnungen der öffentlichen Sammlungen. Ihren Direktoren, die den Bearbeiter durch wertvolle Mitteilungen bereitwillig unterstützten, sei auch an dieser Stelle unser Dank ausgesprochen, insonderheit Herrn Generaldirektor Geh. Regierungsrat Prof. Dr. *Otto v. Falke*, dessen Entgegenkommen die Beschreibung des erst im Herbst 1921 zu eröffnenden Schloßmuseums ermöglichte.

Die Grundsätze der Herausgeber bei Aufzählung der Gasthöfe, Restaurants usw. sind bekannt. Bei der Masse derartiger Anstalten in einer Millionenstadt reicht aber alle Erfahrung und Erkundigung nicht aus. Besonders gerühmte Gaststätten sind durch ein Sternchen (*) oder ein empfohlenes Wort hervorgehoben. Mehr als ein Durchschnittsurteil wollte man darin nicht sehen. Unter den heutigen Verhältnissen hat auch der beste Wille der Wirte seine Grenzen; der zunehmenden Teuerung müssen ihre Preise folgen, und die im Buche angeführten Preisangaben werden sich z. T. inzwischen bereits wieder erhöht haben. Wo Bedienungsgeld berechnet oder im Preise bereits enthalten ist, bietet man kein Trinkgeld an.

Der große Plan von Berlin, am Ende des Buches, ist in drei verschiedenfarbige Streifen geteilt und läßt sich wie ein Taschenbuch handhaben (vgl. S. 225). Die äußere Abgrenzung des Verkehrsplanes (S. 11) ist dieselbe wie die des Dreistreifenplanes. Die Orientierung von einem zum anderen

Die erste Beschreibung Berlins in Bædekers Reisehandbüchern war noch von *Karl Bædeker* (1801-59), dem Gründer der Firma, verfaßt und füllte in seinem 1842 erschienenen „Handbuch für Reisende durch Deutschland und den österreichischen Kaiserstaat“ 23 Seiten. Der stetig wachsende Stoff zwang allmählich zu Teilungen des Buches in mehrere Bände, aber auch in diesen konnte Berlin nur ein verhältnismäßig beschränkter Abschnitt gewidmet werden, der seit 1878 auch als Sonderdruck aus dem Bande „Norddeutschland“ erschien. 1883 wurde er, entsprechend der wachsenden Bedeutung der Reichshauptstadt, zu einem selbständigen Handbuch erweitert, dessen Neubearbeitungen bis 1914 schnell aufeinander folgten. Der unglückliche Ausgang des Weltkriegs hat dem kaiserlichen Berlin ein Ende bereitet und die Herausgeber vor neue Aufgaben gestellt. Zu unserer Freude erklärte sich Prof. Dr. *Bruno Graupe*, seit 1896 unser Mitarbeiter, auch diesmal wieder bereit, die Erneuerung des Buches zu übernehmen und führte sie auf Grund eigener Beobachtungen und an zuverlässigen Stellen eingezogener Erkundigungen mit gewohnter Sorgfalt durch. Besondere Schwierigkeiten boten dabei die fast unausgesetzten Neuordnungen der öffentlichen Sammlungen. Ihren Direktoren,

Baedeker, Berlin, S. V.

Zur Komplexität des Textgenres

Dichte der Named Entities

11. *Theater des Westens* (Pl. G 10; S. 180), Kantstr. 12, für Operetten. 1650 Plätze.
12. *Komische Oper* (Pl. R 23; S. 143), Friedrichstr. 104/104a, an der Weidendammer Brücke. 1250 Plätze.
13. *Theater am Nollendorfsplatz* (Pl. G 14; S. 173), Motzstr. 80/82. 1106 Plätze.
14. *Metropoltheater* (Pl. R 22; S. 122), Behrenstr. 55/57, Ausstattungspossen.
15. *Trianon-Theater* (Pl. R 23), Georgenstr. 9. — 16. *Kleines Theater* (Pl. R 23), U. d. Linden 44. — 17. *Kleines Schauspielhaus*, im Theatersaal der Hochschule für Musik (Pl. R 10; S. 180), Fasanenstr. 1. — 18. *Tribüne* (Pl. R 7), Berliner Straße 37, ebenfalls ein kleines Saaltheater. — 19. *Neues Operetten-Theater* (Pl. R 23), Schiffbauerdamm 5. — 20. *Komödienhaus* (Pl. R 20; S. 159), Schiffbauerdamm 25; 1100 Plätze. — 21. *Schiller-Theater* (Pl. R 7; S. 180), Bismarckstraße 110. — 22. *Neues Volkstheater* (Pl. R 28), Köpenicker Str. 68. — 23. *Wallner-Theater* (Pl. R 28), Wallner-Theater-Str. 35. — 24. *Residenz-Theater* (Pl. R 29), Blumenstr. 9. — 25. *Lustspielhaus* (Pl. G 22), Friedrichstr. 236. — 26. *Luisentheater* (Pl. G 29), Reichenberger Str. 34. — 27. *Thalia-Theater* (Pl. R 25), Dresdener Str. 72/73, Possen. — 28. *Rose-Theater* (Pl. R 32), Große Frankfurter Str. 132. — 29. *Friedrich-Wilhelmstädtisches Theater* (Pl. R 21),

Baedeker, Berlin, S. 26.

Zur Komplexität des Textgenres

Abkürzungen

Abkürzungen.

Bhf. = Bahnhof.
Z. = Zimmer, einschl. Licht und
Bedienung.
F. = Frühstück.
G. = Gabelfrühstück.
M. = Mittagmahl (Table d'hôte).
A. = Abendessen.
P. = Pension mit Zimmer.
n., N. = nördlich, Norden.
ö., O. = östlich, Osten.
s., S. = südlich, Süden.

St. = Stunde.
Min. = Minute.
m = Meter, km = Kilometer.
cm = Centimeter.
cbm = Kubikmeter.
ha = Hektar.
kg = Kilogramm.
So., Mo., Di., Mi., Do., Fr., Sa. =
Sonntag, Montag, Dienstag, Mitt-
woch, Donnerstag, Freitag, Sams-
tag (Sonnabend).
M = Mark, Pf. = Pfennig.
S. = Seite.

tes ist durch Stern (*) hervorgehoben.

1913-15 von O. Kaufmann errichtet, ein ernster, eindrucksvoller Bau, mit säulengeschmückter Muschelkalkfassade und Bildwerken von F. Metzner. Westl. von der Schönhauser Allee, in der Fehrbelliner Straße, die kathol. Herz-Jesu-Kirche (Pl. R 27), 1897/8 von Ch. Hehl in frühroman. Stil erbaut, von F. Stummel ausgemalt. Den Senefelderplatz (Station der Untergrundbahn s. S. 12) schmückt ein Marmordenkmal für den Erfinder des Steindrucks, A. Senefelder († 1834), von Pohle (1892). Der alte jüdische Friedhof (Pl. B 30) enthält reiche Grabdenkmäler.

Baedeker, Berlin, S. VI, 165.

Annotation-Guidelines

Modellierung: Selektion, Abstraktion

- iterative Anpassung und Optimierung der Guidelines im Prozess
- Kuratierung auf Grundlage der Guidelines

Named Entity Recognition: Annotation-Guidelines für die Reiseführer von Karl Baedeker.

Variante 1: einfache, grundlegende Auszeichnung

Digital History, Institut für Geschichtswissenschaft – Humboldt-Universität zu Berlin

1. Einführung

1.1 Motivation

Ziel der Guidelines ist, eine nachvollziehbare Grundlage der Annotation der Baedeker-Reiseführer zu bieten. Die nachvollziehbare Annotation setzt eine transparente Methodologie voraus. Damit eine möglichst vergleichbare Grundlage geschaffen ist, müssen die Annotationen genau den Guidelines folgen, die wiederum Qualitätskriterien erfüllen und somit eine Evaluation möglich machen.¹ Nach dem methodologischen Vorbild zur Erstellung von Guidelines von Fort et al. sollen demnach diese Richtlinien aufzeigen, was annotiert werden soll, anstelle von wie, denn „[t]very often, feasibility constraints overcome semantic criteria.“²

1.2. Annotation Tool

Die Annotation des Goldstandards erfolgt mit *INCEpTION*.

Auszug aus unseren Annotations-Guidelines

Unterstützung von Studierenden

bei den Annotation-Guidelines:

Wiebke Sczeponik, Koray Mendi, Paul Bayer, Anica Skibba

beim Annotationsprozess:

Paul Bayer, Anica Skibba, Anselm König, Isabell Trilling

DANKE!

Annotation mit INCEpTION

The screenshot displays the INCEpTION web interface for document annotation. The top navigation bar shows 'INCEpTION', 'Projects', and 'Dashboard'. The user is logged in as 'droege'. The current project is 'GPT meets NER - 2. Batch' and the document is '0115_proc.txt'. The interface is divided into three main sections:

- Units Sidebar:** A list of units from 21 to 59. Units 25, 26, 39, and 40 are highlighted in green, indicating they are the current focus.
- Main Text Area:** Displays the text of the document with annotations. The text is: "38 Jahrhunderts (der andere Flügel des Diptychons im Museum zu Antwerpen); 528 b.", "39 Hans Memling, Madonna; 537a.", and "40 Meister von Flemalle, männl.". Annotations are shown as colored boxes above the text: 'ORG' (orange) for 'Museum zu Antwerpen', 'PER' (purple) for 'Hans Memling', and 'LOC' (green) for 'Flemalle'.
- Control Bar:** A toolbar with various icons for navigation (back, forward, search), editing (undo, redo), and document management (refresh, lock, settings).

Screenshot vom Kuratierungsprozess in [INCEpTION](#)

Interannotator-Agreement

Annotationsvorgang:

- durchgeführt in mehreren Schritten
- erfolgt durch mehrere Annotierende
- basierend auf den Annotation-Guidelines

Interannotator-Agreement:

- Übereinstimmung der Annotierenden nach Krippendorff's Alpha: zwischen 0.77 und 0.84 (1 = komplette Übereinstimmung)

droege	
38	Jahrhunderts (der andere Flügel des Diptychons im <u>Museum zu Antwerpen</u>); 528 b. LOC
39	<u>Hans Memling</u> , <u>Madonna</u> ; 537a. PER PER
40	Meister von <u>Flemalle</u> , männl. LOC

koenig	
38	Jahrhunderts (der andere Flügel des Diptychons im <u>Museum zu Antwerpen</u>); 528 b. ORG
39	<u>Hans Memling</u> , <u>Madonna</u> ; 537a. PER
40	Meister von <u>Flemalle</u> , männl. PER

Beispiele für Unsicherheiten bei der Annotation

Rathaus und Börse - LOC oder ORG?

als LOC:

- Haltestellen Rathaus bzw. Börse als konkrete Ort im Raum

als ORG:

- Rathaus als Institution der politischen Entscheidungsfindung
- Börse als institutionalisierter Handelsplatz

⇒ hängt vom spezifischen Kontext ab
mitunter nicht entscheidbar



Börse 1932, BA Bild 102-13817



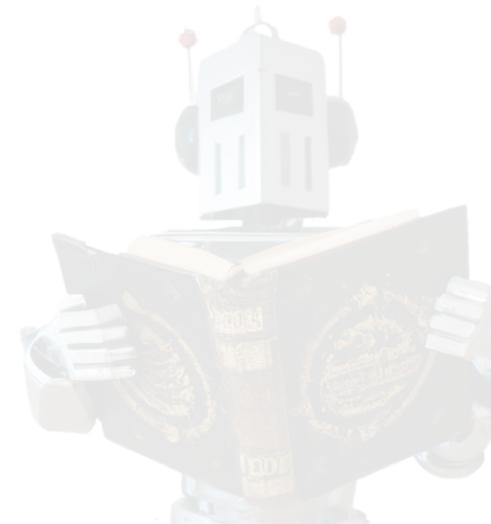
Rathaus 1955, BA Bild 183-28540-0003

Übersicht zur Ground Truth

```
{"text": "Die erste Beschreibung Berlins in Baedekers Reisehandbüchern war noch von Karl Baedeker (1801 - 59), dem Gründer der Firma, verfaßt und füllte in seinem 1842 erschienenen „Handbuch für Reisende durch Deutschland und den österreichischen Kaiserstaat“ 23 Seiten. Der stetig wachsende Stoff zwang allmählich zu Teilungen des Buches in mehrere Bände, aber auch in diesen konnte Berlin nur ein verhältnismäßig beschränkter Abschnitt gewidmet werden, der seit 1878 auch als Sonderdruck aus dem Bande „Norddeutschland“ erschien. 1883 wurde er, entsprechend der wachsenden Bedeutung der Reichshauptstadt, zu einem selbständigen Handbuch erweitert, dessen Neubearbeitungen bis 1914 schnell aufeinander folgten. Der unglückliche Ausgang des Weltkriegs hat dem kaiserlichen Berlin ein Ende bereitet und die Herausgeber vor neue Aufgaben gestellt. Zu unserer Freude erklärte sich Prof. Dr. Bruno Graupe, seit 1896 unser Mitarbeiter, auch diesmal wieder bereit, die Erneuerung des Buches zu übernehmen und führte sie auf Grund eigener Beobachtungen und an zuverlässigen Stellen eingezogener Erkundigungen mit gewohnter Sorgfalt durch. Besondere Schwierigkeiten boten dabei die fast unausgesetzten Neuordnungen der öffentlichen Sammlungen. Ihren Direktoren, die den Bearbeiter durch wertvolle Mitteilungen bereitwillig unterstützten, sei auch an dieser Stelle unser Dank ausgesprochen, insonderheit Herrn Generaldirektor Geh. Regierungsrat Prof. Dr. Otto v. Falke, dessen Entgegenkommen die Beschreibung des erst im Herbst 1921 zu eröffnenden Schloßmuseums ermöglichte. Die Grundsätze der Herausgeber bei Aufzählung der Gasthöfe, Restaurants usw. sind bekannt. Bei der Masse derartiger Anstalten in einer Millionenstadt reicht aber alle Erfahrung und Erkundigung nicht aus. Besonders gerühmte Gaststätten sind durch ein Sternchen (*) oder ein empfehlendes Wort hervorgehoben. Mehr als ein Durchschnittsurteil wolle man darin nicht sehen. Unter den heutigen Verhältnissen hat auch der beste Wille der Wirte seine Grenzen; der zunehmenden Teuerung müssen ihre Preise folgen, und die im Buche angeführten Preisangaben werden sich z. T. inzwischen bereits wieder erhöht haben. Wo Bedienungsgeld berechnet oder im Preise bereits enthalten ist, bietet man kein Trinkgeld an. Der große Plan von Berlin, am Ende des Buches, ist in drei verschiedenfarbige Streifen geteilt und läßt sich wie ein Taschenbuch handhaben (vgl. S. 225). Die äußere Abgrenzung des Verkehrsplanes (S. 11) ist dieselbe wie die des Dreistreifenplanes. Die Orientierung von einem zum anderen", "length": 2540, "ground_truth": {"tags": {"LOC": [{"start": 23, "end": 30}, {"start": 200, "end": 211}, {"start": 374, "end": 380}, {"start": 495, "end": 510}, {"start": 763, "end": 769}, {"start": 2273, "end": 2279}], "PER": [{"start": 34, "end": 43}, {"start": 74, "end": 87}, {"start": 878, "end": 890}, {"start": 1436, "end": 1449}], "ORG": [{"start": 220, "end": 248}, {"start": 1529, "end": 1542}]}, "prodigy": [{"label": "LOC", "start": 23, "end": 30}, {"label": "LOC", "start": 200, "end": 211}, {"label": "LOC", "start": 374, "end": 380}, {"label": "LOC", "start": 495, "end": 510}, {"label": "LOC", "start": 763, "end": 769}, {"label": "LOC", "start": 2273, "end": 2279}, {"label": "PER", "start": 34, "end": 43}, {"label": "PER", "start": 74, "end": 87}, {"label": "PER", "start": 878, "end": 890}, {"label": "PER", "start": 1436, "end": 1449}, {"label": "ORG", "start": 220, "end": 248}, {"label": "ORG", "start": 1529, "end": 1542}]}}
```

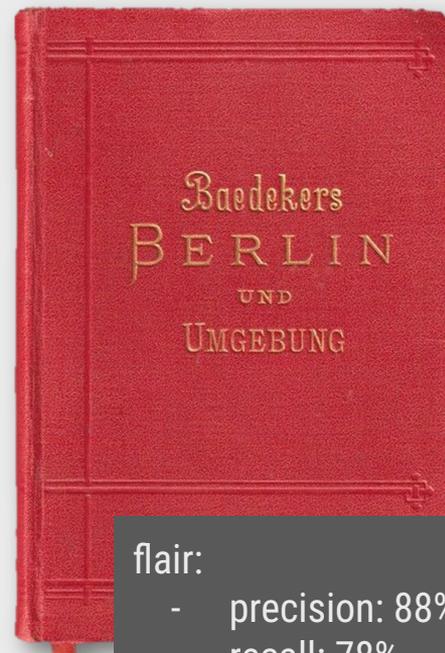
Auszug Ground Truth im prodigy-Format in JSON

4 Experimente



flair-Modell als Baseline (zum Vergleich)

- State of the Art NLP (92.31 F1-Score auf Conll-03)
- Einfach zu benutzen
- Bildet ein breites Spektrum ab und ist u.a. multilingual
 - hat mehr Varianz in Sprache gelernt
 - viele Abweichungen verhalten sich ähnlich zu Fremdsprachen
 - lateinischen Inschriften
 - alternativen Schreibweisen ("Regenspurg", ...)



flair:

- precision: 88%
- recall: 78%

Für die folgenden Tests verwendetes Modell

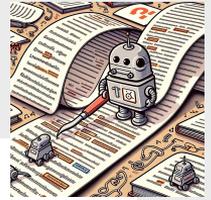
- via API von OpenAI nutzen wir das Modell: GPT-4o (2024-05-13)
- Settings:
 - Temperature: 0
 - penalties: 0
- ≠ ChatGPT

```
pred_data = []
for line in open(jsonfile.path, "r").readlines():
    data = json.loads(line)
    tagged = tagger.chat.completions.create(
        model="gpt-4o-2024-05-13",
        # model="gpt-3.5-turbo-0125",
        messages=conversation.get_prompt(data["text"]),
        max_tokens=None,
        seed=42,
        stream=True,
        temperature=0,
        top_p=1.0,
        frequency_penalty=0.0,
        presence_penalty=0.0
    )
```

Prompt Strategie

Beispiel-Persona:

“ Als erfahrener Sprachwissenschaftler mit Spezialisierung auf Named Entity Recognition (NER) für historische Texte und Experte insbesondere für die Geschichte Berlins besteht deine Aufgabe darin, einen gegebenen Text für die maschinelle Weiterverwertung zu annotieren. Es handelt sich hierbei um einen Reiseführer aus der Zeit der Weimarer Republik. Dieser enthält neben beschreibenden Texten auch eine Vielzahl von Aufzählungen von für Reisende interessanten Geschäften, Gaststätten, Transportmöglichkeiten und ähnliches, welche an einem bestimmten Ort oder entlang einer Strecke zu beobachten sind. [...]



Struktur

```
### SETTING ###
Als erfahrener Sprachwissenschaftler mit Spezialisierung auf Named Entity Recognition (NER) für historische Texte und Experte insbesondere für die Geschichte Berlins besteht deine Aufgabe darin, einen gegebenen Text für die maschinelle Weiterverwertung zu annotieren. Es handelt sich hierbei um einen Reiseführer aus der Zeit der Weimarer Republik. Dieser enthält neben beschreibenden Texten auch eine Vielzahl von Aufzählungen von für Reisende interessanten Geschäften, Gaststätten, Transportmöglichkeiten und ähnliches, welche an einem bestimmten Ort oder entlang einer Strecke zu beobachten sind.

### INSTRUCTION ###
Folge diesen Schritten
1. Wiederhole den gegebenen Text. Achte dabei sehr genau darauf, dass abgesehen von den Annotationen nichts hinzugefügt oder entfernt wird.
2. Die Aufgabe ist es BEREITESTIMTE Kategorien zuzuordnen. Benutze hierfür Tags mit den entsprechenden Regeln:
  - Wenn ein Bereich beginnt, markiere dies durch einen Kategorie-Tag. Zur Verfügung stehen hier ausschließlich die Kategorien '<<PER >>' für Personen, '<<LOC >>' für Orte und '<<ORG >>' für Organisationen.
  - Wenn ein Bereich endet, markiere das Ende durch den entsprechenden Tag, also '/PER>>', '/LOC>>' oder '/ORG>>'.
  - Achte darauf, dass jeder geöffnete Tag auch geschlossen werden muss.
  - Tags dürfen sich überschneiden oder geschachtelt sein - allerdings ist es unwahrscheinlich, dass dieses zu häufig passiert.

=BEACHTE: Auch griechische Heiden und Götter sind Named Entities und als Personen <<PER >> zu annotieren
=BEACHTE: Gaststätten und Restaurants sind als Organisationen <<ORG >> zu annotieren, auch wenn der jeweilige Bezeichner ohne Kontext als Person interpretierbar wäre
=BEACHTE: Auch sind Locations <<LOC >> innerhalb zitierter Entitäten zu annotieren. Das Buch „Pfade des Aves durch Gartien“ wäre dementsprechend „Pfade des <<PER Aves >>/PER>> durch <<LOC Gartien >>/LOC>>“

### EXAMPLE ###
Input:
(1853) und des Physiologen J. Müller († 1858). Eintr. s. S. 33. — Das Gebäude enthält drei selbständige, mit Museen verbundene Institute: das Geologisch-paläontologische Institut (Direktor: Geh. Bergrat Prof. Dr. Jos. Pompeckj /PER>>), das Mineralogisch-petrographische Institut (Direktor: Geh. Bergrat Prof. Dr. Theod. Liebisch) und das Zoologische Institut (Direktor: Geh. Reg.-Rat Prof. Dr. Carl Heider). — Die „Hauptsammlung“ in ersten und zweiten Stock, die lediglich wissenschaftlichen Zwecken dient, ist dem Publikum nicht zugänglich.

Output:
(1853) und des <<PER Physiologen J. Müller >> /PER>> († 1858). Eintr. s. S. 33. — Das Gebäude enthält drei selbständige, mit Museen verbundene Institute: das <<ORG Geologisch-paläontologische Institut >> /ORG>> (Direktor: Geh. Bergrat Prof. Dr. <<PER Jos. Pompeckj >> /PER>>), das <<ORG Mineralogisch-petrographische Institut >> /ORG>> (Direktor: Geh. Bergrat Prof. Dr. <<PER Theod. Liebisch >> /PER>>) und das <<ORG Zoologische Institut >> /ORG>> (Direktor: Geh. Reg.-Rat Prof. Dr. <<PER Carl Heider >> /PER>>). — Die „Hauptsammlung“ in ersten und zweiten Stock, die lediglich wissenschaftlichen Zwecken dient, ist dem Publikum nicht zugänglich.

### HINTS ###
Beachte, wie in Beispiel alle Schritte eingehalten wurden:
1. Wiederholung des exakten Textes.
2. Jeder Bereich hat einen Tag, der
  - beginnt und endet hat
  - markiert ist
Der Text muss ANSONSTEN unverändert reproduziert werden.
Keine Kürzere hängt davon ab, dass du keine Fehler machst. Daher bekommen sowohl du, als auch deine Oma jeweils 1000€ für jede erfolgreiche Annotation, die in weiteren Vorlaufschritten maschinenlesbar ist.
Nimm dir Zeit und gehe Schritt für Schritt vor. Hier der zu annotierende Text:
...
INPUT_TEXT
```

Persona-Modeling

Instruktionen

Beispiel(e)

Wiederholung Instruktionen

„Bestechung“ des Netzes

Chain-Of-Thought

Promptteil zur Sequenzausgabe

Kern-Anweisung:

""Folge diesen Schritten:

1. **Wiederhole den gegebenen Text exakt.** Achte dabei sehr genau darauf, dass abgesehen von den Annotationen nichts hinzugefügt oder entfernt wird.

2. Die Aufgabe ist es **BEREICHE bestimmten Kategorien zuzuordnen.** Benutze hierfür Tags mit den entsprechenden Regeln:

- Wenn ein Bereich beginnt, markiere dies durch einen Kategorie-Tag. Zur Verfügung stehen hier unter anderem die Kategorien '<<PER ' für Personen, '<<LOC ' für Orte und '<<ORG ' für Organisationen.

- Wenn ein Bereich endet, markiere das Ende durch den entsprechenden Tag, also ' /PER>>', ' /LOC>>' oder ' /ORG>>'.

- Achte darauf, dass jeder geöffnete Tag auch geschlossen werden muss.

- Tags dürfen sich überlappen oder geschachtelt sein - allerdings ist es unwahrscheinlich, dass dieses zu häufig passiert.

""



<<PER Kaiser Wilhelm II. /PER>>, der das <<LOC Schloß /LOC>> seiner früheren Bestimmung, außer zu Festlichkeiten und zur Beherbergung fürstlicher Gäste auch als Wohnung für die königliche Familie zu dienen, wiedergab. Die fünf äußeren Portale erhielten schmiedeeiserne Gitter; an der Spreeseite wurde eine Terrasse mit Landungsplatz angelegt; in neuester Zeit endlich erfuhren der <<LOC Weiße Saal /LOC>> (S. 68) mit den benachbarten Räumen sowie die entsprechenden Teile der darunter liegenden Stockwerke, ferner der <<LOC Joachimsaal /LOC>> (S. 71) nach Plänen <<PER E. v. Ihne's /PER>> einen durchgreifenden Umbau. Das Kaiserpaar bewohnte das erste Stockwerk am <<LOC Schloßplatz /LOC>>. Vom Äußern des Schlosses ist die <<LOC Wasserfront /LOC>> S. 64 besprochen. Die Westfassade <<PER Eosanders /PER>> mit ihrem Hauptportal, einer riesenhaft gesteigerten Nachbildung des <<LOC Severusbogens in Rom /LOC>>, ist etwas trocken, aber seit Abbruch der Gebäude der <<LOC Schloßfreiheit /LOC>> für das <<LOC Nationaldenkmal /LOC>> (S. 71) von großartiger Wirkung. Die beiden Bronzereliefs von <<PER O. Lessing /PER>> (1897) beziehen sich auf <<PER Kurfürst Friedrich II. /PER>> als Erbauer der Burg und auf <<PER König Friedrich I. /PER>> als Bauherrn des Schlosses. — An den beiden von <<PER Schlüter /PER>> herrührenden Langfronten zeigt sich ein feiner Unterschied zwischen der streng monumentalen Südfassade am <<LOC Schloßplatz /LOC>> und der zierlich belebten Nordfassade am <<LOC Lustgarten /LOC>> (damals Gartenseite). Vor der Nordfassade erinnern sechs Bronzestandbilder an die niederländischen Beziehungen des <<PER Großen Kurfürsten /PER>>. Vor der „Schloßapotheke“ steht seit 1905 der <<PER Admiral Coligny /PER>>, vom <<PER Grafen Görtz-Schlitz /PER>>; <<PER Coligny /PER>>, 1572 in der <<LOC Bartholomäusnacht /LOC>> ermordet, war Urgroßvater der ersten Gemahlin des <<PER Großen Kurfürsten /PER>> <<PER Luise Henriette /PER>>. Fünf <<LOC STANDBILDER VON ORANIERN /LOC>> wurden 1907 auf der Brüstung der <<LOC Schloßterrasse /LOC>> aufgestellt: <<PER Wilhelm I. der Schweiger /PER>> (1533 - 84), von <<PER Schott /PER>>; <<PER Moritz /PER>> (1567 - 1625), von <<PER M. Wolff /PER>>;

Postprocessing des Outputs

1. Auffinden aller Bereiche innerhalb der Tags <<FOO /FOO>>
 - Man erhält so einen “zerstückelten” Text
2. Fuzzy suchen des Tag-Inhalts im originalen Text
 - Nicht gebunden an etwaiges Tokenizing
 - Ein LLM **wird** Abweichungen produzieren
 - Korrektur an Rechtschreibung
 - Tauschen von Zeichen (etwa {3 durch ß)
 - Hinzufügen/Entfernen von Leerzeichen
3. Ergebnis ist die Position der Annotation im Originaltext

Metriken

- Typische Metriken sind “zu simpel”
- Wir nutzen `nervaluate` zur Ermittlung
 - Ermöglicht ebenfalls Berechnung von Precision/Recall
 - “striker” Modus entspricht den üblichen Metriken

“relaxte” Definitionen

correct: Annotierte Textregion **überlappt** mit Ground-Truth und hat **richtigen** Typ

incorrect: Annotierte Textregion **überlappt** mit Ground-Truth und hat **falschen** Typ

missed: Textregion in der Ground-Truth wurde **nicht annotiert**

spurious: Textregion **wurde annotiert**, die nicht in der Ground-Truth vorkommt

Ergebnisse: Simpler und Vollständiger Prompt

Text	correct	incorrect	missed	spurious	precision	recall
Nur Instruktion (ours)	1132	77	281	43	0,9	0,76
Full Prompt (ours)	1357	53	80	99	0,9	0,91
flair	1163	93	234	69	0,88	0,78
Full Prompt (strict, ours)	1233	177	80	99	0,82	0,83
<i>flair (strict)</i>	921	335	234	69	0,74	0,62

Tabelle 1: Ergebnisse der Evaluation auf den Daten Evaluationskorpus auf Grundlage des Berlin-Reiseführer von 1921, unsere Lösung im Vergleich zu flair (large model)

Ablation study

Test	Recall	Loss	Remarks
only instruction	0,76	-16,63 %	Simple Instruktionen ohne Kontext & Anweisungen
1-shot	0,86	-5,86 %	Full Prompt mit nur 1 zufälligen Beispiel
0-shot	0,87	-4,83 %	Full Prompt ohne Beispiele
no context	0,87	-3,91 %	Full Prompt ohne Persona-Modeling
less context	0,88	-1,95 %	Full Prompt mit generischer Historiker-Persona
no bullying	0,89	-2,03 %	Keine “Belohnungen” für das LLM bzw. emotionale Manipulation
no repeat	0,89	-1,81 %	Keine zusammenfassende Wiederholung der Aufgabe in anderen Worten

Tabelle 2: Minderung der Performanz durch Weglassen einzelner Komponenten aus dem Prompt

→ Kontext und Anweisungen sind wesentlich für die Performance!

Fehlerexegese

Durch Kontextwissen z.T. besser als der Mensch?

T: **Borstells** Lesezirkel); **Speyer & Peters** (Universitätswissenschaften),
P: **Borstells** Lesezirkel); **Speyer & Peters** (Universitätswissenschaften),

wir: Borstell als PER; LLM: Borstells Lesezirkel als ORG

Kontextwissen aus
dem Modell
(nur teilweise
korrekt, aber
Grundlagen richtig)

Was ist Borstells Lesezirkel?



Borstells Lesezirkel ist ein Unternehmen, das sich auf die Vermietung und Verteilung von Zeitschriften spezialisiert hat. Kunden, wie beispielsweise Arztpraxen, Friseursalons, Hotels und Wartezimmer in verschiedenen Branchen, können Zeitschriften abonnieren, die regelmäßig und zu einem reduzierten Preis geliefert werden. Das Ziel ist es, Wartezeiten angenehmer zu gestalten, indem aktuelle Zeitschriften bereitgestellt werden. Borstells Lesezirkel bietet eine Vielzahl von Zeitschriften aus unterschiedlichen Kategorien wie Lifestyle, Mode, Sport, Wissenschaft und vielen mehr an.

Übertragbarkeit

Evaluationskorpus von Ole Meiners (IBI, HU Berlin):

- 4 Texte in frühneuzeitlichem Deutsch
- Verschiedene Gattungen (Briefe, Reiseberichte, Predigten)
- teils diplomatisch transkribiert, teils normalisiert, mit vielen zeittypischen Spezifika

Zu balte Eine superintetur Zu hoffen, so machte ich mir fast ein gewißen ihmm da von ab Zu halte. wann er etwann da noch mehr gudes beförder könde, ihmm daran Zu hindern, die andern sind aber noch frey, und haben kein ammt auch noch keine vocation, gott wird unß Zeigen wer es sein soll, wegen deßen von der schul pforde halte auch orlichen noch wichticher, doch ist ienes auch ein feiner mensch, unßer magister und bonin haben ihm recht könen lerne,

Die Reyse nach <<LOC Wien /LOC>>. 4. m
den 27. 7bris, ♀, biß <<LOC Chemnitz /LOC>>, 3. m die <<LOC Augustusburg /LOC>>, auf der l s liegend gesehen, biß <<LOC Schopa /LOC>>, 2. m facit 5.
den 29. 7bris ♂. durch <<LOC Postelberg /LOC>>, <<LOC Laun /LOC>> auf die r s <<LOC Chumtzaan dorff /LOC>>, <<LOC Teinnitz Flecken /LOC>>, 5 ½ m
den 7. 8bris, ♂. biß <<LOC Znaib /LOC>>, 3. m biß <<LOC Gundsdorf /LOC>>, den <<PER Herren Teufeln /PER>> gehörig, biß <<LOC Holabrun /LOC>>, 4. m facit 7.
den 10. 8bris, ♀, biß <<LOC Wien /LOC>>, 2. m Zum güldenem stier, oder Ochsen.
den 20. 8bris ♂ auf den Abend, bey dem von <<PER Schlabata /PER>> zu gaste gewesen.

Übertragbarkeit

- Identischer Prompt wie beim Baedeker
- Einzige “Eigenleistung”: Beschreibung von Quelle und Art
- **Keine** Beispiele/Annotation/..., d.h. Ergebnisse können noch deutlich verbessert werden

Als erfahrener **Sprachwissenschaftler** mit **Spezialisierung auf Named Entity Recognition (NER)** für historische Texte und Experte insbesondere für **frühneuzeitliches Deutsch** besteht deine Aufgabe darin, einen gegebenen Text für die maschinelle Weiterverwertung zu annotieren. Es handelt sich hierbei um einzelne Sätze, die zusammenhangslos transkribiert wurden. Dir wird in jeder Zeile ein Satz vorgegeben.

Beachte, dass auch einzelne Abschnitte des Textes in **Latein** stehen können. Führe die NER dort dann auf Latein aus.

Achte wirklich darauf, dass Du jeden Satz einzeln bearbeitest und zeichengenau wiedergibst. Nimm Dir Zeit und arbeite so genau wie möglich. Ich möchte die Ergebnisse später computationell weiterverarbeiten.

Bei dem Text handelt es sich um das „Diarium August des Jüngeren von Braunschweig-Wolfenbüttel“, ein von Herzog August II. eigenhändig verfasstes **Tagebuch** über einen Zeitraum von vierzig Jahren (1594-1635). Das Tagebuch wurde u. a. während der **zahlreichen Reisen des Verfassers durch Nordwestdeutschland** und das übrige Reich, aber auch durch das Elsass und Italien geführt und dokumentiert diese. Es diente seinem Autor mutmaßlich vorrangig als Gedächtnisstütze. Die Einträge des Tagebuchs sind stark schematisch angelegt, auf die Nennung des jeweiligen Datums (unter Angabe des Planetensymbols für den Wochentag) folgen zumeist sehr kurze Angaben zu **Reisestrecken und besuchten Orten**. Längere Passagen widmen sich **Sehenswürdigkeiten oder getroffenen Personen**. Zudem werden soziale Ereignisse wie **Turniere, Hochzeiten, Taufen oder Todesfälle** berichtet. Die Einträge sind auf deutsch verfasst, allerdings werden mitunter längere Abschriften von **lateinischen Inschriften** eingefügt.

Der Text enthält die sprachlichen oder dialektalen Eigenheiten seiner Zeit. Dazu gehört die Schreibung der Eigennamen, Getrennt- und Zusammenschreibung, Groß- und Kleinschreibung, Majuskeln, Konsonantenhäufungen, Orthographie und Interpunktion sowie die Verwendung zeittypischer Abkürzungen.

Übertragbarkeit

Ergebnisse zum Evaluationskorpus von Ole Meiners:

Text	Datierung	recall (flair large)	recall (ours)	precision (flair large)	precision (ours)
Sammlung von Reiseberichten „Newe unbekante landte“	1508	67%	86%	74%	68%
Tagebuch des Herzogs August II. von Braunschweig-Wolfenbüttel	1594-1635	81%	91%	83%	79%
Leichenpredigten aus dem Thüringer Raum	1673-1723	90%	94%	78 %	71%
Briefe der Gräfin Erdmuthe Beningna von Reuss-Ebersdorf	1711-1722	29%	60%	53%	56%

Tabelle 3: Ergebnisse der Evaluation auf den Daten des von Ole Meiners erstellten Evaluationskorpus zu fnz. Quellen

Kosten

OpenAI Pricing Calculator

Calculate how much it will cost to generate a certain number of words using OpenAI GPT-3.5 and GPT-4 APIs.

Select the OpenAI language model:

GPT-4o (Newest 🦄)

Number of output words to generate:

750

Number of prompt words (per 1000 words generated):

8000

Generating 750 words using GPT-4o (Newest 🦄) costs ~\$0.0683

1 Seite Baedeker
Full Prompt
32 Beispiele

Select the OpenAI language model:

GPT-4o (Newest 🦄)

Number of output words to generate:

750

Number of prompt words (per 1000 words generated):

1000

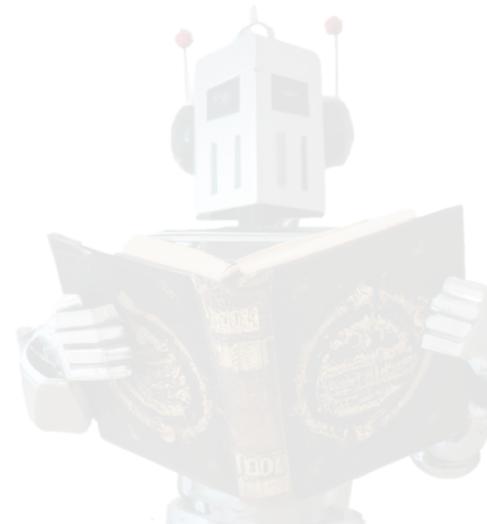
Generating 750 words using GPT-4o (Newest 🦄) costs ~\$0.0217

1 Seite Baedeker
Full Prompt
0 Shot

Kosten für alle hier vorgestellten Experimente, Tests, Auswertungen, etc. unter 250 USD

5

Ergebnisse und Einordnung



Ergebnisse

Ansatz: Von NLP-Sprachverarbeitungs-Task zur domänenspezifischen, kontextgebundenen Aufgabe unter Nutzung aktueller Prompting-Techniken

Unser Ergebnis:

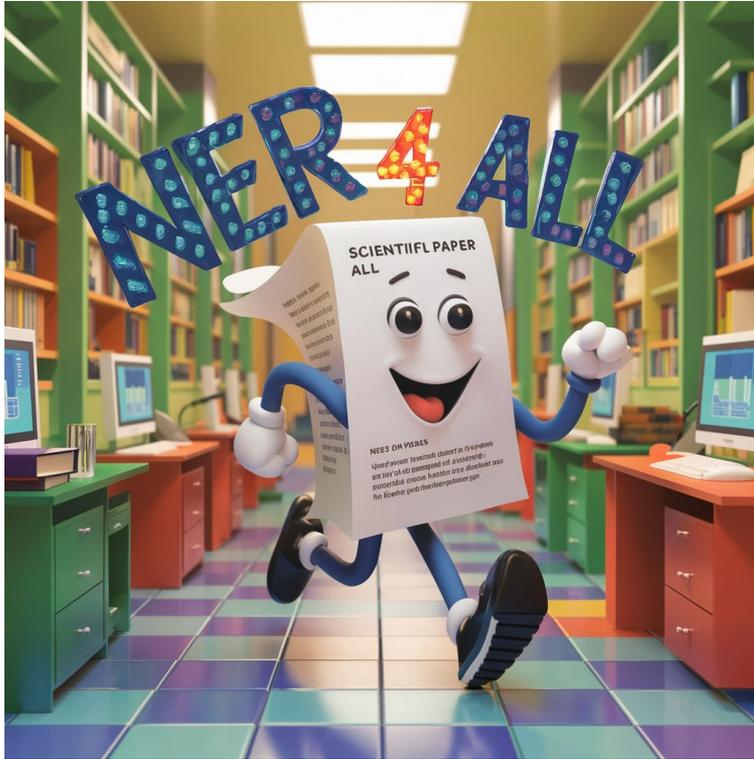
- unser Ansatz bringt **deutlich bessere Ergebnisse** (gerade im Recall) als Standard-Workframes (hier: flair)
- **problemlos übertragbar**, ohne Fine-tuning und ohne Erstellung eigener Trainingsdaten
- **anpassbar** an verschiedene Tagsets und Projektanforderungen
- iterativ durch **natürlichsprachliche Eingaben** optimierbar

Paradigmenwechsel durch LLM:
eröffnet Anwendung von NER für **heterogene, low-ressource Anwendungsfälle** in der Arbeit mit historischen Textmaterial ganz neue Möglichkeiten

Ergebnisse und Ausblick

Nächste Schritte:

- Übertragung auf andere, offene LLM (Llama, Command R+)
- weitere Prompting-Techniken: self-correction ; Redundanz und Vergleich
- Agnostischer Masterprompt zur eigenständigen Erstellung eines Prompts auf Grundlage der zu annotierenden Texte
- Erweiterung um Named Entity Resolution (Einbindung *function calling*)
- Test für weitere Information Extraction Kontexte, z.B. Naturereignisse, Konflikte etc.
- Intrinsisches Erstellen von Entity-Ontologien (z.B. "Autor" als Spezialfall von Person; "Werk"; ...) aus dem Korpus



Das Paper ist auf dem Weg!
Bild: ideogram.ai

Vielen Dank!

Bibliographie

- Baedeker, Karl (1921): Berlin und Umgebung: Handbuch für Reisende; mit 4 Karten, 9 Plänen und 17 Grundrissen, Leipzig.
- Brown, Tom B., Benjamin Mann, Nick Ryder et al. (2020): Language Models are Few-Shot Learners, in: arXiv:2005.14165v4 [cs.CL], DOI: <https://doi.org/10.48550/arXiv.2005.14165>.
- Ehrmann, M., Hamdi, A., Pontes, E.L. et al. (2023), Named Entity Recognition and Classification in Historical Documents: A Survey, in: ACM Computing Surveys 56, 27:1-27:47. DOI: <https://doi.org/10.1145/3604931>
- González-Gallardo, C.-E., Boros, E., Girdhar, N., et al. (2023): Yes but.. Can ChatGPT Identify Entities, in: [arXiv:2303.17322v1](https://arxiv.org/abs/2303.17322v1) [cs.DL]. DOI: <https://doi.org/10.48550/arXiv.2303.17322>.
- Keraghel, I., Morbieu, S., Nadif, M. (2024): A survey on recent advances in named entity recognition, in: arXiv:2401.10825v1 [cs.CL]. DOI: <https://doi.org/10.48550/arXiv.2401.10825>
- OpenAI (2023): GPT-4 Technical Report, in: arXiv:2303.08774v6[cs.CL]. DOI: <https://doi.org/10.48550/arXiv.2303.08774>.
-
- TIGER-Korpus. Internet: <https://www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/tiger/>
- Conll3-Korpus: <https://huggingface.co/datasets/eriktks/conll2003>.
- WikiNER: Nothman, J., Ringland, N, Radford, W. et al. (2013), Learning multilingual named entity recognition from Wikipedia, in: Artificial Intelligence 194, 151-175. DOI: <https://doi.org/10.1016/j.artint.2012.03.006>.
- Sang, E.F.T.K, F. 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*, 142–147, Internet: <https://aclanthology.org/W03-0419.pdf>.
- Benikova, D., C. Biemann, M. Kisselew, S. Padó (2014): GermEval 2014 Named Entity Recognition Shared Task: Companion Paper, Internet: http://www.ims.uni-stuttgart.de/institut/mitarbeiter/kisselmx/publikationen/GermEval_2014_NER_Shared_Task.pdf.
- Ehrmann, M., M. Romanello, M., A. Doucet, S. Clemenide (2022): Introducing the HIPE 2022 Shared Task: Named Entity Recognition and Linking in Multilingual Historical Documents, in: M. Hagen et al.: *Advances in Information Retrieval. ECIR 2022. Lecture Notes in Computer Science*, vol. 13186. Cham. DOI: https://doi.org/10.1007/978-3-030-99739-7_44.