TECHNISCHE UNIVERSITÄT WIEN

NLP – Group 23

Named Entity Recognition for German Official Documents

Tim Greß (12412672)
Viktor Hoffmann (12433741)
Lars Böhmer (12436447)
Jan Tölken (12432831)

16. Januar 2026

Content

Dataset

Name

FinCorpus-DE10k

Description

German financial/legal corpus containing 10,000 PDF documents with extracted text

Source

'anhaltai/fincorpus-de-10k' on HuggingFace

Labelling Guidelines

BIO Tagging Scheme

- Begin / Inside / Outside

Entities

- Organizations (ORG)
- Monetary (MON)
- Legal References (LEG)

Organizations – ORG

Include	Example
Full company names	Landesbank Baden-Württemberg
Company abbreviations	LBBW
Legal forms	KR Labs GmbH
Financial institutions	EZB
Banks	Sparkasse

Exclude	Example
Generic organizational terms	Eine Sparkasse
All other names	Max Mustermann

Organization Example

Die Landesbank Baden-Württemberg ist eine deutsche Bank, die mit der Deutschen Bank AG und der EZB zusammenarbeitet und jährlich 2.000.000.000 Euro umsetzt.

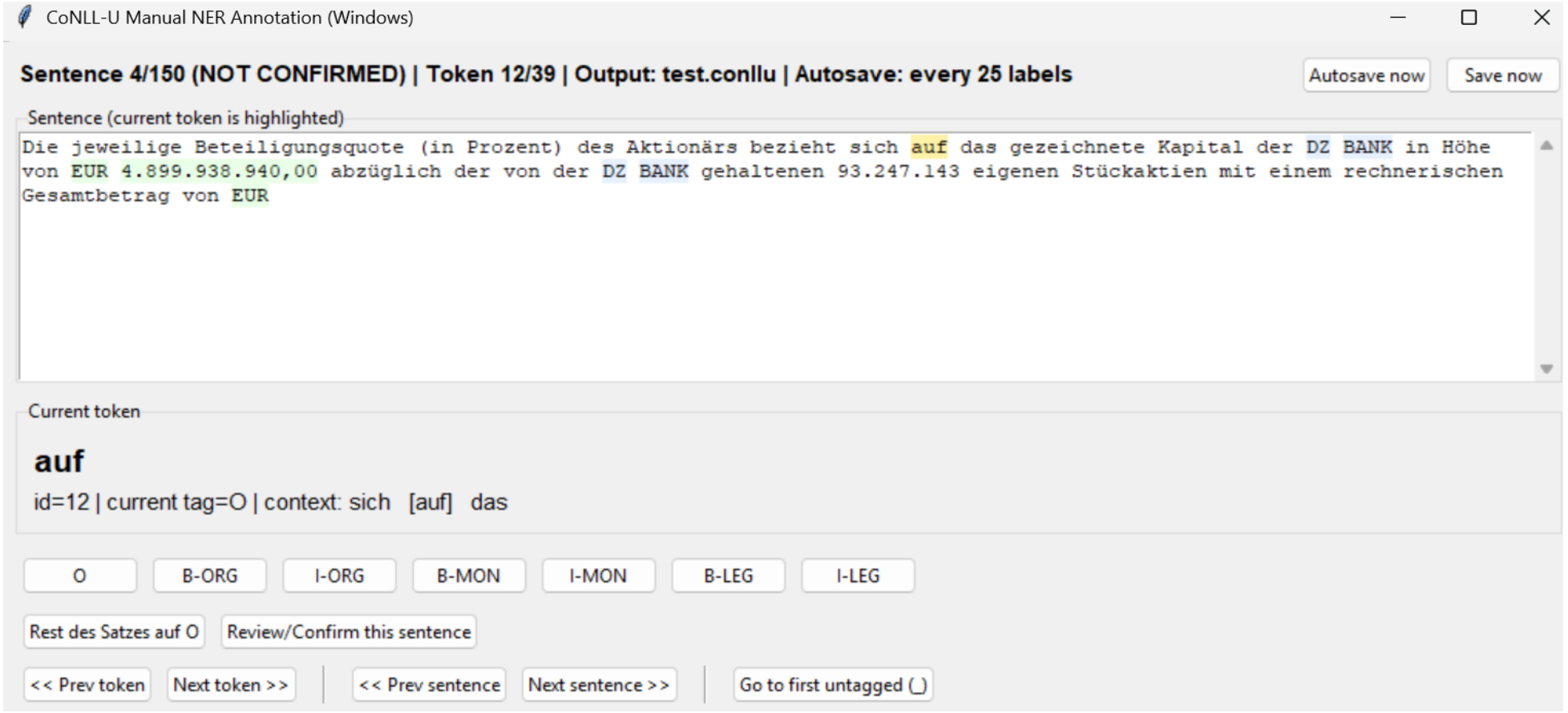
B-ORG I-ORG B-MON I-MON O

TECHNISCHE UNIVERSITÄT WIEN

Labelling Pitfalls

- Even manually it is often hard to distinguish between abbreviations for organizations and other names
- What to include as monetary values?
  - *Return of 20%*
- Very time consuming – labelled 170 sentences

Labelling Tool with GUI



TECHNISCHE UNIVERSITÄT WIEN

Baselines

EnhancedRuleNER

- Hand-crafted regex patterns per entity type
- Validation heuristics using the POS tags (e.g. span contains at least one NOUN or PROPN)

TokenNB

- Token-level Naïve Bayes classifier
- Uniform class priors to compensate heavy class imbalance (~93% O-tags)

CRF

- Linear-chain CRF for BIO tagging
- Rich token features: casing/shape, prefixes/suffixes,
- Domain features: currency symbols (€), paragraph signs (§), company suffixes (GmbH, AG)

SpaCy + Heuristics

- Hybrid labelling combining:
  - spaCy German NER model
  - Key-word based ORG spans
  - Monetary Detection
  - Legal reference detection

German BERT

- Fine-tunes bert-base-german-cased
- Aligns labels to WordPiece tokens
- Token-level + span overlap evaluation

RuleChef

Idea
Use an LLM to *learn* explicit extraction rules

Result
interpretable ruleset that can be applied like a classic rule-based NER system

Nach §15 Abs.1 ist die vollständige….

Train on gold examples → LLM rule synthesis → Evaluation on FP/FN → Iteratively refine rules → Freeze final ruleset

```
{
  "name": "Extract § + number + Abs.",
  "format": "regex",
  "pattern": "§\\s*\\d+\\s*Abs\\.?\\s*\\d+"
}
```

SOURCE: HTTPS://GITHUB.COM/KRLABSORG/RULECHEF - LAST ACCESS: 15.01.2026

LLM API

Input: "Die DZ BANK hat EUR 4.899.938 nach § 271 HGB ausgewiesen."

API Returns:

```
  "entities": [
    {"label": "ORG", "text": "DZ BANK"},
    {"label": "MON", "text": "EUR 4.899.938"},
    {"label": "LEG", "text": "§ 271 HGB"}
  ]
```

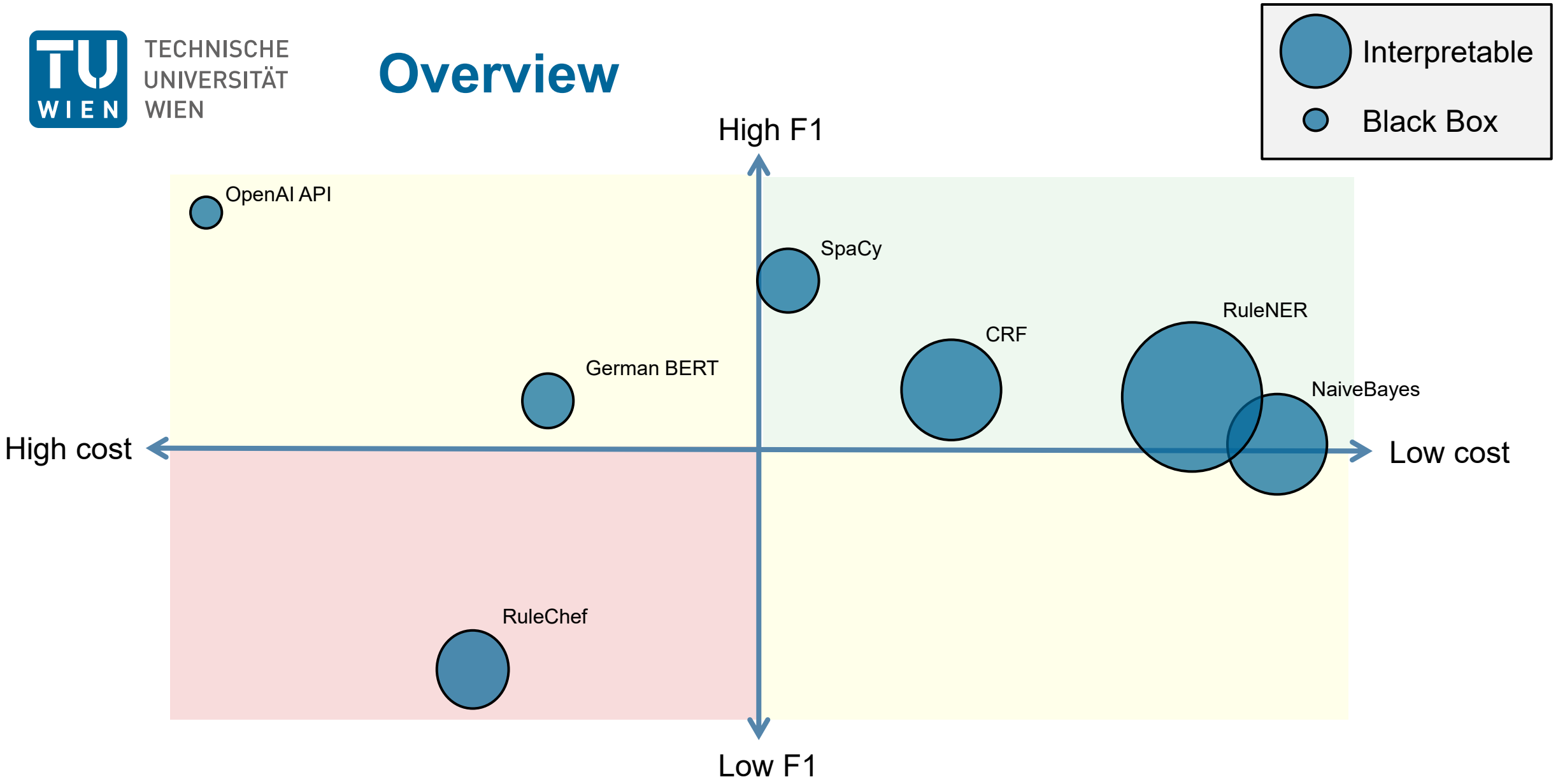
Output: O B-ORG I-ORG O B-MON I-MON O B-LEG I-LEG I-LEG O

TECHNISCHE UNIVERSITÄT WIEN

Performance

Method	LEG F1	MON F1	ORG F1	Macro F1		Run Time (s)
EnhancedRuleNER	0,59	0,67	0,20	0,61		0,01
TokenNB	0,07	0,90	0,20	0,51		0,01
CRF	0,67	0,97	0,06	0,67		0,76
SpaCy	0,86	0,82	0,44	0,77		0,24
German BERT (2 epochs)	0,43	0,69	0,62	0,68		99,52
RuleChef (3 loops)	0,07	0,10	0,01	0,22		370,75
OpenAI API (10 threads)	1,00	1,00	0,59	0,89		167,90

Overview


Interpretable
Black Box
High F1
Low F1
High cost
Low cost
OpenAI API
SpaCy
RuleNER
CRF
German BERT
NaiveBayes
RuleChef

RuleChef

+

- Interpretable results as code or regex
- Fast and deterministic inference as soon as rules are created
- Strong domain control, can adapt to certain keywords, formats etc.

-

- Refinement often overgeneralizes
- Formatting issues with malformed rules or invalid JSON
- Struggles with unstructured classes like ORG
- Very bad performance in our case
- Inconsistent across runs

Conclusion and Insights

- NER for German official documents is a trade-off problem, no „free lunch“
- Entity structure is very important - consistent underperformance for ORG
- Neural Models improve ORG detection but are computationally expensive
- RuleChef was unstable and underperformed in our setup due to implementation and robustness issues

Suggested Pipeline