NLP + Offeneregister.de

From: Dominik Pichler

Exploring and assessing different approaches for NLP-based NER extraction in the official german Company register extract from Offeneregister.de

Date: October 8, 2024

Dataset:

The following dataset is publicly available and has been used as the foundation of this project: https://offeneregister.de/

Data-Tables:

Available tables:

- company
- name
- officer
- registrations

I retrieved the included tables and columns via:

```sql
SELECT m.name as tableName,
       p.name as columnName,
       p.type as columnType FROM sqlite_master m
LEFT OUTER JOIN
       pragma_table_info(m.name) p
       ON m.name <> p.name
       WHERE m.type IN ('table', 'view')
       AND m.name NOT LIKE 'sqlite_%'
       ORDER BY tableName, columnName;
```

As we are interested in NER/NLP I filtered this list for non-structural, TEXT/CLOB columns only via:

```sql
SELECT * FROM (SELECT
    m.name as tableName,
    p.name as columnName,
    p.type as columnType
FROM
    sqlite_master m
LEFT OUTER JOIN
    pragma_table_info(m.name) p
ON
    m.name <> p.name
WHERE
    m.type IN ('table', 'view')
AND
    m.name NOT LIKE 'sqlite_%'
ORDER BY
    tableName,
    columnName) WHERE columnType IN ('TEXT','CLOB');
```

leading to **53** potentially interesting and relevant columns. After manually inspecting those columns, the following few seemed relevant for NLP/ER Tasks for me

Company Table:

```sql
SELECT _registerNummerSuffix
        ,company_number
        ,current_status
        ,federal_state
        ,former_registrar
        ,jurisdiction_code
        ,name
        ,native_company_number
        ,register_art
        ,register_flag_
        ,register_flag_Note:
        ,register_flag_Status information
        ,register_nummer
        ,registered_address
        ,registered_office
        ,registrar
        retrieved_at
FROM company;
```

Out of which the following are relevant for NLP/ER Tasks: - **Name**

Officer Table

```sql
SELECT city
        ,company_id
        ,dismissed
        ,end_date
        ,firstname
        ,flag
        ,lastname
        ,maidenname
        ,name
        ,position
        ,reference_no
        ,start_date
        ,title
        type
FROM officer;
```

Out of which the following are relevant for NLP/ER Tasks:

- **flag**

Registrations Table

```sql
SELECT  alternate_company_number
        ,alternate_entity_type
        ,alternate_jurisdiction_code
        ,company_id
        ,confidence
        ,data_type
        ,previous_company_number
        ,previous_entity_type
        ,previous_jurisdiction_code
        ,previous_registration_end_date
        ,publication_date
        ,retrieved_at
        ,sample_date
        ,source_url
        ,start_date
        ,start_date_type
        ,subsequent_company_number
        ,subsequent_entity_type
        ,subsequent_jurisdiction_code
        ,subsequent_registration_start_date
FROM registrations;
```

Out of which the **non** seem relevant for NLP/ER Tasks.

Summary:

This leads two potentially interesting columns:

- `company.name` that contains the full company name
- `officer.flag` that contains controlling rules of individuals.

NER Investigations:

Company.name Column:

After consideration, following information might be relevant to extract:

Company Type and Status

Identify the type of company based on suffixes like "GmbH," "e.K.," or "Union," which indicate the legal structure (e.g., GmbH for a limited liability company in Germany) In my estimation, this can extracted by an rule-based approach.

Geographical Information

Extract potential geographical indicators from the company name, such as "Algeria" in "Shell Algeria Zerafa GmbH," which might suggest a regional focus or origin.

Industry or Sector

Analyze keywords within the names that might indicate the industry, such as “Reederei” (shipping) or “Entertainment.”

Branding or Product Focus

Identify specific branding elements or product focus from names like “Lime Juice Entertainment,” which could hint at the company’s market segment.

Owner or Founder Names

Extract personal names if present, such as “Markus Blum” in “Markus Blum Montagearbeiten e.K.,” which might indicate the founder or owner.

General Entity Recognition:

ER has been tested/performed on the two columns mentioned above. More can be found in the `playground_ER-NLP.ipynb`,`company_name_spaCy_ER_NLP.ipynb` and `company_name_HuggingFace.ipynb` notebook.

Approaches	Result	Comment
1. Raw german SpaCy using `de_core_news_lg`	good first results	Should be tested on larger scales
2. Translate names to english + `en_core_news_lg`	TBD	Left out for now
3. `elenanereiss/bert-german-ler`	did not help much	
4. `google-bert/bert-base-german-cased`	did not help much	
5. `dslim/bert-base-NER`	problematic	Probably needs translation first
5. `spaCy-llm`	TBD	

1. (N)ER using default spaCy models:

Above all, a save fallback solution seems to be spaCy and it’s available NER of the following entity types. For example the following entities are available in the general english sca vocabulary:

```
"CARDINAL",
        "DATE",
        "EVENT",
        "FAC",
        "GPE",
        "LANGUAGE",
        "LAW",
        "LOC",
        "MONEY",
        "NORP",
        "ORDINAL",
        "ORG",
        "PERCENT",
        "PERSON",
        "PRODUCT",
        "PROPN",
        "QUANTITY",
        "TIME",
        "WORK_OF_ART",
```

```
"LOC",
"MISC",
"ORG",
"PER",
```

Some examples:

```
PERSON:      People, including fictional.
NORP:        Nationalities or religious or political groups.
FAC:         Buildings, airports, highways, bridges, etc.
ORG:         Companies, agencies, institutions, etc.
GPE:         Countries, cities, states.
LOC:         Non-GPE locations, mountain ranges, bodies of water.
PRODUCT:     Objects, vehicles, foods, etc. (Not services.)
EVENT:       Named hurricanes, battles, wars, sports events, etc.
WORK_OF_ART: Titles of books, songs, etc.
LAW:         Named documents made into laws.
LANGUAGE:    Any named language.
DATE:        Absolute or relative dates or periods.
TIME:        Times smaller than a day.
PERCENT:     Percentage, including "%".
MONEY:       Monetary values, including unit.
QUANTITY:    Measurements, as of weight or distance.
ORDINAL:     "first", "second", etc.
CARDINAL:    Numerals that do not fall under another type.
PROPN:       proper noun, e.g. Mary, John, London, NATO, HBO
```

In order to handle foreign (non-german/english) company names, the langdetect model comes in handy. For translations to german/english, the Googletrans package has shown strong performance.

German

For the german language, the following models offer NER recoginition: - de_core_news_sm - de_core_news_md - de_core_news_lg

For those models include the following common entity labels:

```
PERSON (PER): People, including fictional characters.
NORP: Nationalities or religious or political groups.
FAC: Buildings, airports, highways, bridges, etc.
ORG: Companies, agencies, institutions, etc.
GPE: Countries, cities, states.
LOC: Non-GPE locations, such as mountain ranges and bodies of water.
PRODUCT: Objects, vehicles, foods, etc. (Not services).
EVENT: Named hurricanes, battles, wars, sports events, etc.
WORK_OF_ART: Titles of books, songs, etc.
LAW: Named documents made into laws.
LANGUAGE: Any named language.
```

This approach has then been tested with the `de_core_news_lg` model on a small scale and seemed to work well.

Company Name	Token	Entity
olly UG (haftungsbeschränkt)	olly	ORG
olly UG (haftungsbeschränkt)	UG	ORG
olly UG (haftungsbeschränkt)	(	No entity
olly UG (haftungsbeschränkt)	haftungsbeschränkt	No entity

Company Name	Token	Entity
olly UG (haftungsbeschränkt)	)	No entity
BLUECHILLED Verwaltungs GmbH	BLUECHILLED	ORG
BLUECHILLED Verwaltungs GmbH	Verwaltungs	No entity
BLUECHILLED Verwaltungs GmbH	GmbH	No entity
Mittelständische Beteiligungsgesellschaft Bremen mbH	Mittelständische	No entity
Mittelständische Beteiligungsgesellschaft Bremen mbH	Beteiligungsgesellschaft	No entity
Mittelständische Beteiligungsgesellschaft Bremen mbH	Bremen	LOC
Mittelständische Beteiligungsgesellschaft Bremen mbH	mbH	No entity
Albert Barufe GmbH	Albert	PER
Albert Barufe GmbH	Barufe	PER
Albert Barufe GmbH	GmbH	No entity
ITERGO Informationstechnologie GmbH	ITERGO	ORG
ITERGO Informationstechnologie GmbH	Informationstechnologie	No entity
ITERGO Informationstechnologie GmbH	GmbH	No entity
Rheinbahn AG	Rheinbahn	ORG
Rheinbahn AG	AG	ORG
Verwaltung IFÖ Zweite Immobilienfonds für Österreich GmbH	Verwaltung	No entity
Verwaltung IFÖ Zweite Immobilienfonds für Österreich GmbH	IFÖ	ORG
Verwaltung IFÖ Zweite Immobilienfonds für Österreich GmbH	Zweite	No entity
Verwaltung IFÖ Zweite Immobilienfonds für Österreich GmbH	Immobilienfonds	No entity
Verwaltung IFÖ Zweite Immobilienfonds für Österreich GmbH	für	No entity
Verwaltung IFÖ Zweite Immobilienfonds für Österreich GmbH	Österreich	LOC
Verwaltung IFÖ Zweite Immobilienfonds für Österreich GmbH	GmbH	No entity
AWS Personalmarketing GmbH	AWS	ORG
AWS Personalmarketing GmbH	Personalmarketing	ORG
AWS Personalmarketing GmbH	GmbH	ORG

2. Translate names to english + `en_core_news_lg`

TBD

3. NER using bert-german-ler

Fine-tuned BERT utilizing a german dataset with the following annotated tokens:

Fine-grained classes	#	%
1. PER Person	1,747	3.26
2. RR Judge	1,519	2.83
3. AN Lawyer	111	0.21
4. LD Country	1,429	2.66
5. ST City	705	1.31
6. STR Street	136	0.25
7. LDS Landscape	198	0.37
8. ORG Organization	1,166	2.17
9. UN Company	1,058	1.97
10. INN Institution	2,196	4.09
11. GRT Court	3,212	5.99
12. MRK Brand	283	0.53
13. GS Law	18,52	34.53
14. VO Ordinance	797	1.49
15. EUN European legal norm	1,499	2.79

Fine-grained classes	#	%
16. VS Regulation	607	1.13
17. VT Contract	2,863	5.34
18. RS Court decision	12,58	23.46
19. LIT Legal literature	3,006	5.60
Total	53,632	100

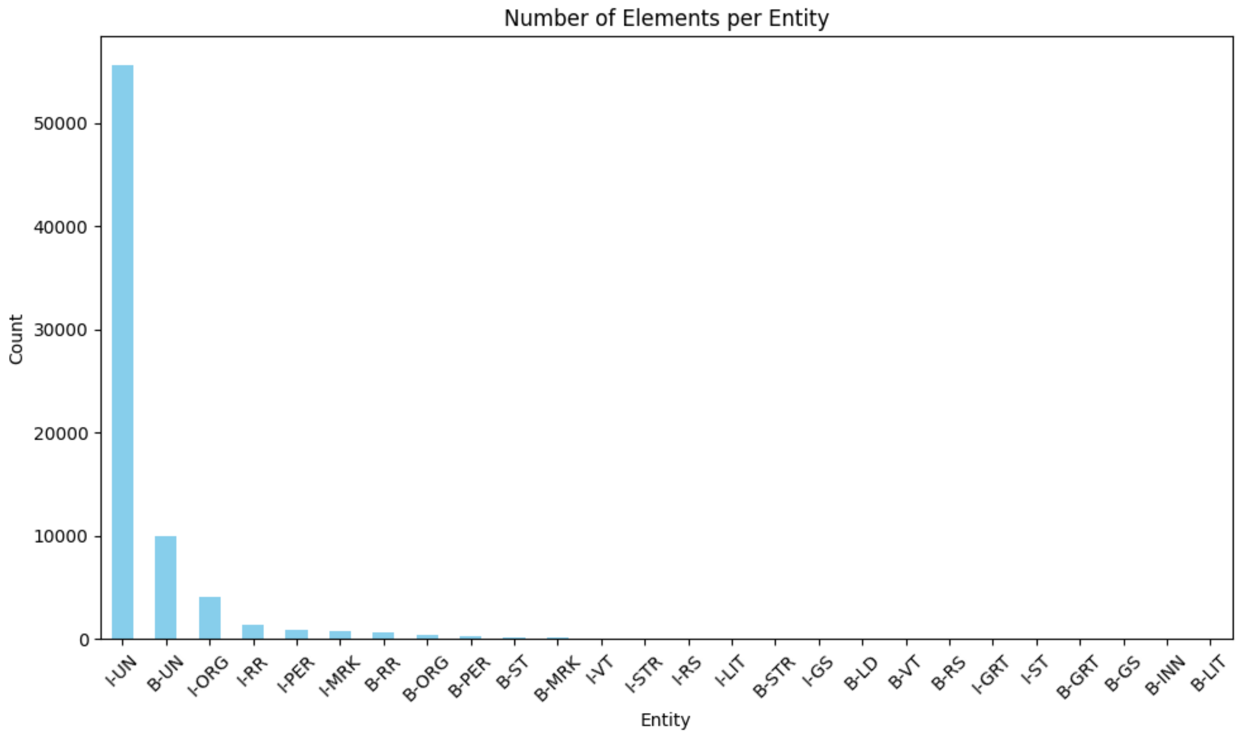

Figure 1: output.png

Benchmark Test with 100k Company Names: The model detected 880 Tokens as "Person". Examples look like:

entity	score	index	word	start	end
I-PER	0.52123576	2	:	2	3
I-PER	0.7519806	3	Lee	3	6
I-PER	0.5675053	3	##o	5	6
I-PER	0.52893364	12	##e	33	34
I-PER	0.70512134	13	Ha	35	37
I-PER	0.4497944	14	##gem	37	40
I-PER	0.48570016	15	##eier	40	44
I-PER	0.3244481	16	-	44	45
I-PER	0.3622264	17	Lem	45	48
I-PER	0.29476708	6	o	22	23
I-PER	0.33195055	8	H	25	26
I-PER	0.6844228	12	Wa	47	49

To me, the effectiveness of this model seems limited . . .

5. `bert-base-NER`

...

6. `spaCy-llm`

....

Specific ER

Company Type and Status

Due to the limited number of different available company types, a manually curated dict will probably be best.

Geographical Information

Raw spaCy (German)

Did work at least ok on all of the following models: - `de_core_news_sm` - `de_core_news_md` - `de_core_news_lg` - `de_dep_news_trf`

Snipped from the best results:

ID	Company Name	Category	Type
471	CAMTEC24 - Sicherheitstechnik e. K.	Sicherheitstechnik	LOC
546	Gebäude Technologie Center GmbH	Technologie	LOC
547	Gebäude Technologie Center GmbH	Center	LOC
25	Verwaltung IFÖ Zweite Immobilienfonds für Österreich GmbH	Österreich	LOC
472	CAMTEC24 - Sicherheitstechnik e. K.	e.	LOC
473	CAMTEC24 - Sicherheitstechnik e. K.	K.	LOC
316	"Verwaltungsgesellschaft MS""Barmbek" " mbH"	Barmbek	LOC
753	Abel & Dr. Schuhmann Rechtsanwaltsgesellschaft mbH Zweigniederlassung Hamburg	Hamburg	LOC
1055	Geis SDV GmbH Zweigniederlassung Hamburg	Hamburg	LOC
10	Mittelständische Beteiligungsgesellschaft Bremen mbH	Bremen	LOC
545	Gebäude Technologie Center GmbH	Gebäude	LOC
984	SPC Ardmona (Germany) GmbH	Germany	LOC
641	HCI Treuhand Holland XXII UG (haftungsbeschränkt)	Holland	LOC
469	CAMTEC24 - Sicherheitstechnik e. K.	CAMTEC24	LOC
432	Hartungstraße 12 Verwaltungsgesellschaft mbH	Hartungstraße	LOC

GeoSpaCy

Paper Githup Repo outdated - Core Streamlit application does not work anymore Centrally it looks to me like the also only use standard spaCy models with small regex extensions that dont help much So this approach is also not very helpful

Geoparsing

TBD

Gazetters

TBD

Vocabulary based Entity Recognition ?

For Company Type

Identify the type of company based on suffixes like "GmbH," "e.K.," or "Union," which indicate the legal structure (e.g., GmbH for a limited liability company in Germany) Rule-based approach / Fixed Vocab should do the trick here.

For Geographical Information

Same here, I think, if at all, a Rule-based / Fixed Vocab approach, utilizing the data from Deutschlandatlas will probably do the trick.

Industry or Sector

Same here, RB & fV utilizing lists of industries/sectors from: - Statista - Gewerbelisten -> could be utilized to get a comprehensive list of available *Gewerbearten*

Branding or Product Focus

Same here, RB & fV utilizing lists of registered trade marks(?) like: - DPMAregister (Limited to Germany) - EUIPO Database (EU wide) - ###### Owner or Founder Names Extract personal names if present, such as "Markus Blum" in "Markus Blum Montagearbeiten e.K.," which might indicate the founder or owner.

Other Resources:

- Survey on DL for NER (2019) Yielding a list of modern NER Tools:

NER System	URL	Description
StanfordCoreNLP	https://stanfordnlp.github.io/CoreNLP/	A comprehensive NLP toolkit in Java.
OSU Twitter NLP	https://github.com/aritter/twitter_nlp	Tools for NLP on Twitter data.
Illinois NLP	http://cogcomp.org/page/software/	NLP tools from the University of Illinois.
NeuroNER	http://neuroner.com/	Neural network-based NER system.
NERsuite	http://nersuite.nlplab.org/	A suite for named entity recognition.
Polyglot	https://polyglot.readthedocs.io	Multilingual NLP library.
Gimli	http://bioinformatics.ua.pt/gimli	Biomedical NER tool.
spaCy	https://spacy.io/api/entityrecognizer	Industrial-strength NLP library.
NLTK	https://www.nltk.org	A leading platform for building Python programs to work with human language data.
OpenNLP	https://opennlp.apache.org/	Machine learning based toolkit for processing natural language text.
LingPipe	http://alias-i.com/lingpipe-3.9.3/	Toolkit for text analytics and linguistic processing.
AllenNLP	https://demo.allennlp.org/	An open-source NLP research library built on PyTorch.

NER System	URL	Description
IBM Watson	https://natural-language-understanding-demo.ng.bluemix.net/	AI-powered natural language understanding service by IBM.
FG-NER	https://fgner.alt.ai/extractor/	Fine-grained named entity recognition tool.
Intellexer	http://demo.intellexer.com/	Semantic analysis and natural language processing tool.
Repustate	https://repustate.com/named-entity-recognitionapi-demo/	Sentiment analysis and text analytics API with NER capabilities.
AYLIEN	https://developer.aylien.com/text-api-demo	Text analysis API with NER features.
Dandelion API	https://dandelion.eu/semantic-text/entityextraction-demo/	Semantic text analysis API with entity extraction features.
displaCy	https://explosion.ai/demos/displacy-ent	Visualizer for spaCy's named entity recognition model.
ParallelDots	https://www.paralleldots.com/named-entityrecognition	AI-powered text analysis API with NER functionality.
TextRazor	https://www.textrazor.com/named_entity_recognition	Text analysis API with powerful NER capabilities.

- The paper A Dataset for German Legal Documents for NER might also be helpful
- Distilling Large Language Models into Tiny Models for NER
- GPT-NER: Named Entity Recognition via LLMs
- https://github.com/explosion/spacy-llm
- https://medium.com/@lokaregns/named-entity-recognition-with-hugging-face-transformers-a-beginners-guide-e1ac6085fb3c