Towards Extracting Entity Relationship Diagrams from Unstructured Text using Natural Language Processing

Authors: Prof. Dr. Gregor Grambow[1], Vaihunthan Vyramuthu (Presenter)[1]

[1]*Department of Electronics and Computer Science, University of Aalen (Germany)*
Contact email: vaihunthan@gmail.com

Hochschule Aalen

IARIA

Vaihunthan Vyramuthu

Academic career:

10/2022 – 11/2024: University of Aalen

Master: Machine Learning & Data Analytics (M.Sc.)

10/2019 – 09/2022: Cooperative State University Baden-Württemberg Stuttgart

Bachelor: Mechatronics (B.Eng.)

Specializations: Vehicle systems engineering and electromobility

09/2010 – 07/2019: Theodor-Heuss-Gymnasium, Aalen

Abitur: Award of the German Mathematical Society

German Physical Society Abitur Prize

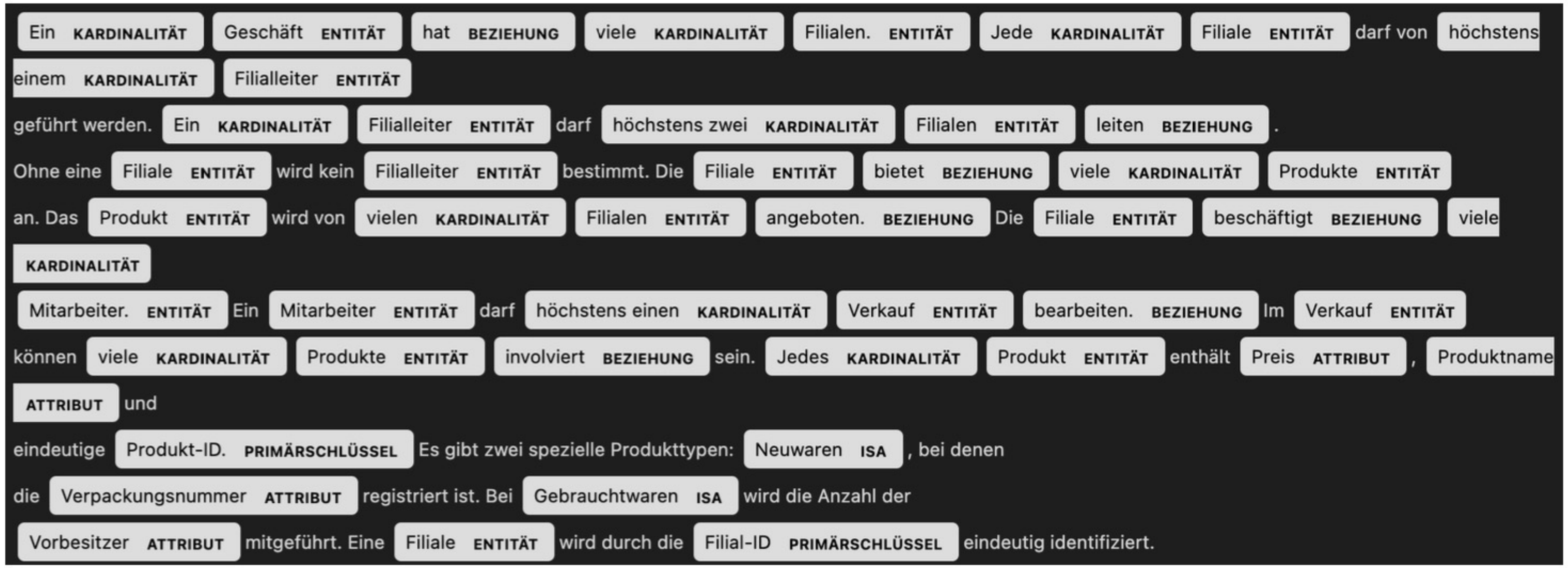
Vaihunthan Vyramuthu

Vaihunthan Vyramuthu received the master’s degree in Machine Learning & Data Analytics from the University of Aalen, Germany in 2024. He is currently working at Carl Zeiss SMT GmbH in Oberkochen, Germany as a Software Developer for photolithography system components.

Idea

How did I come up with the idea?

- Task in an examination: create a DB schema from plain text



Ein KARDINALITÄT Geschäft ENTITÄT hat BEZIEHUNG viele KARDINALITÄT Filialen. ENTITÄT Jede KARDINALITÄT Filiale ENTITÄT darf von höchstens
einem KARDINALITÄT Filialleiter ENTITÄT
geführt werden. Ein KARDINALITÄT Filialleiter ENTITÄT darf höchstens zwei KARDINALITÄT Filialen ENTITÄT leiten BEZIEHUNG .
Ohne eine Filiale ENTITÄT wird kein Filialleiter ENTITÄT bestimmt. Die Filiale ENTITÄT bietet BEZIEHUNG viele KARDINALITÄT Produkte ENTITÄT
an. Das Produkt ENTITÄT wird von vielen KARDINALITÄT Filialen ENTITÄT angeboten. BEZIEHUNG Die Filiale ENTITÄT beschäftigt BEZIEHUNG viele
KARDINALITÄT
Mitarbeiter. ENTITÄT Ein Mitarbeiter ENTITÄT darf höchstens einen KARDINALITÄT Verkauf ENTITÄT bearbeiten. BEZIEHUNG Im Verkauf ENTITÄT
können viele KARDINALITÄT Produkte ENTITÄT involviert BEZIEHUNG sein. Jedes KARDINALITÄT Produkt ENTITÄT enthält Preis ATTRIBUT , Produktname
ATTRIBUT und
eindeutige Produkt-ID. PRIMÄRSCHLÜSSEL Es gibt zwei spezielle Produkttypen: Neuwaren ISA , bei denen
die Verpackungsnummer ATTRIBUT registriert ist. Bei Gebrauchtwaren ISA wird die Anzahl der
Vorbesitzer ATTRIBUT mitgeführt. Eine Filiale ENTITÄT wird durch die Filial-ID PRIMÄRSCHLÜSSEL eindeutig identifiziert.

Aims and contributions of our paper

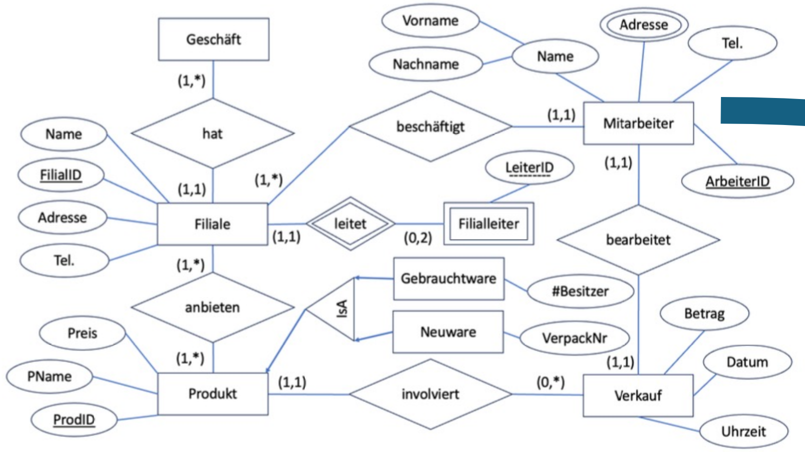
Aims:

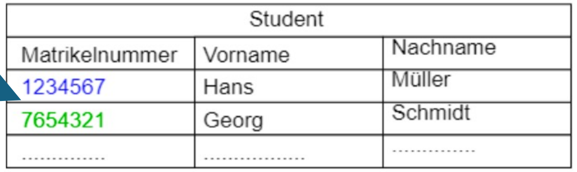
1. providing a hybrid system for capturing ER model components from German texts using NLP
2. time-consuming interpretation of textual database scenarios can be automated.

Contributions:

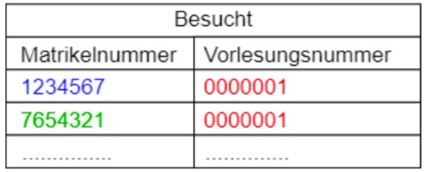
1. we implemented rule-based and model-based approaches, whereas the main extraction is performed by the rule-based variant so that the entities, attributes, relationships and cardinalities can be strategically identified.
2. this hybrid approach combines the strengths of both methods: The rule-based method enables accurate extraction, while the model-based method helps to validate and improve the results, resulting in a more robust and adaptable solution to different text scenarios.

Overview





Student		
Matrikelnummer	Vorname	Nachname
1234567	Hans	Müller
7654321	Georg	Schmidt
....	....	....

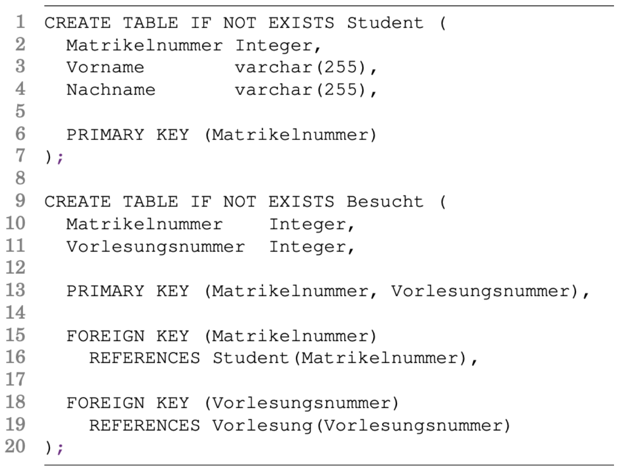


Besucht	
Matrikelnummer	Vorlesungsnummer
1234567	0000001
7654321	0000001
....	....

Ein Geschäft hat viele Filialen. Jede Filiale darf von höchstens einem Filialleiter geführt werden. Ein Filialleiter darf höchstens zwei Filialen leiten.

Ohne eine Filiale wird kein Filialleiter bestimmt. Die Filiale bietet viele Produkte an. Das Produkt wird von vielen Filialen angeboten. Die Filiale beschäftigt viele Mitarbeiter. Ein Mitarbeiter darf höchstens einen Verkauf bearbeiten. Im Verkauf können viele Produkte involviert sein. Jedes Produkt enthält Preis, Produktname und eindeutige Produkt ID. Es gibt zwei spezielle Produkttypen: Neuwaren, bei denen die Verpackungsnummer registriert ist. Bei Gebrauchtwaren wird die Anzahl der Vorbesitzer mitgeführt. Eine Filiale wird durch die Filial ID eindeutig identifiziert. Die Filiale hat einen Namen, eine Adresse und eine Telefonnummer. Verkauf enthält Datum, Uhrzeit und Betrag.

Jeder Mitarbeiter hat einen Namen bestehend aus Vor- und Nachname, eine oder mehrere Adressen und eine Telefonnummer. Die Mitarbeiter und der Filialleiter werden durch eine ID identifiziert.



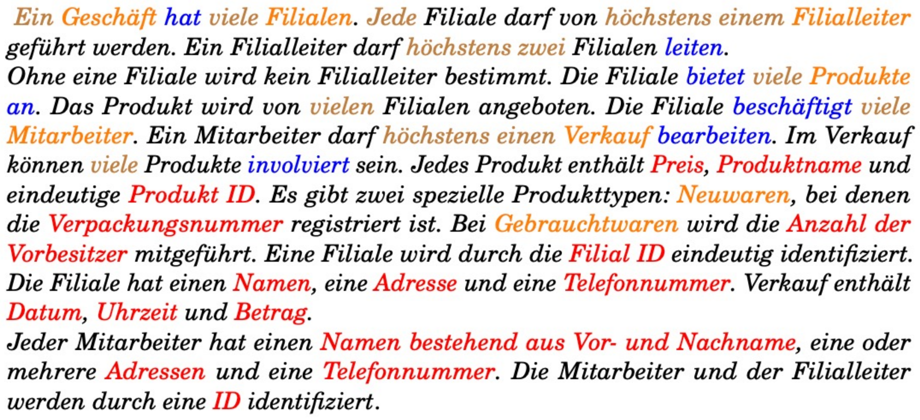
```
CREATE TABLE IF NOT EXISTS Student (
  Matrikelnummer Integer,
  Vorname        varchar(255),
  Nachname       varchar(255),

  PRIMARY KEY  (Matrikelnummer)
);

CREATE TABLE IF NOT EXISTS Besucht (
  Matrikelnummer    Integer,
  Vorlesungsnummer  Integer,

  PRIMARY KEY (Matrikelnummer, Vorlesungsnummer),

  FOREIGN KEY (Matrikelnummer)
    REFERENCES Student(Matrikelnummer),

  FOREIGN KEY (Vorlesungsnummer)
    REFERENCES Vorlesung(Vorlesungsnummer)
);
```

Digression: Why we need ER stage?

1. better understanding of the conceptual level
2. easier to identify redundand tables/relations

→ Goal: to have less tables as possible

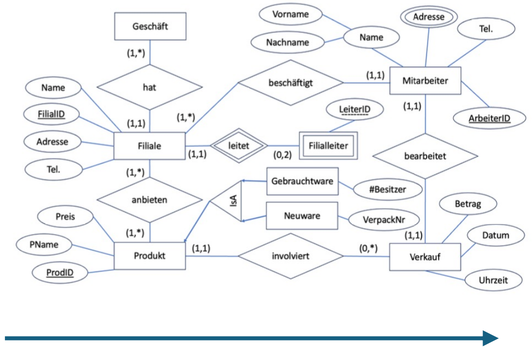
→ less tables means fewer time-consuming JOIN operations necessary

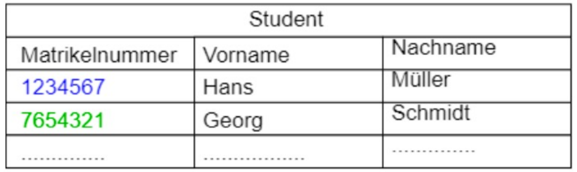


Ein Geschäft hat viele Filialen. Jede Filiale darf von höchstens einem Filialleiter geführt werden. Ein Filialleiter darf höchstens zwei Filialen leiten.

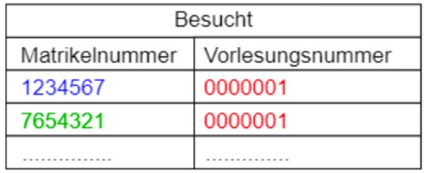
Ohne eine Filiale wird kein Filialleiter bestimmt. Die Filiale bietet viele Produkte an. Das Produkt wird von vielen Filialen angeboten. Die Filiale beschäftigt viele Mitarbeiter. Ein Mitarbeiter darf höchstens einen Verkauf bearbeiten. Im Verkauf können viele Produkte involviert sein. Jedes Produkt enthält Preis, Produktname und eindeutige Produkt ID. Es gibt zwei spezielle Produkttypen: Neuwaren, bei denen die Verpackungsnummer registriert ist. Bei Gebrauchtwaren wird die Anzahl der Vorbesitzer mitgeführt. Eine Filiale wird durch die Filial ID eindeutig identifiziert. Die Filiale hat einen Namen, eine Adresse und eine Telefonnummer. Verkauf enthält Datum, Uhrzeit und Betrag.

Jeder Mitarbeiter hat einen Namen bestehend aus Vor- und Nachname, eine oder mehrere Adressen und eine Telefonnummer. Die Mitarbeiter und der Filialleiter werden durch eine ID identifiziert.



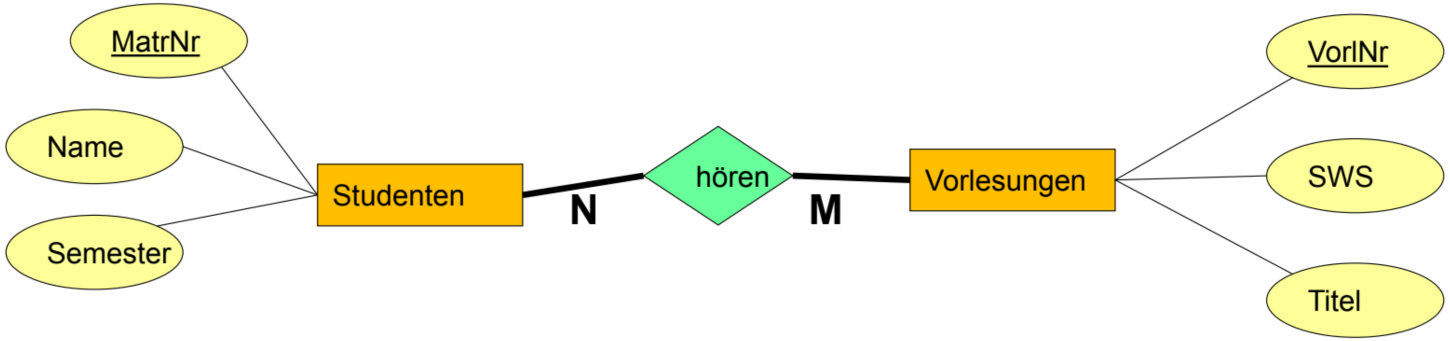


Student		
Matrikelnummer	Vorname	Nachname
1234567	Hans	Müller
7654321	Georg	Schmidt
....	....	....

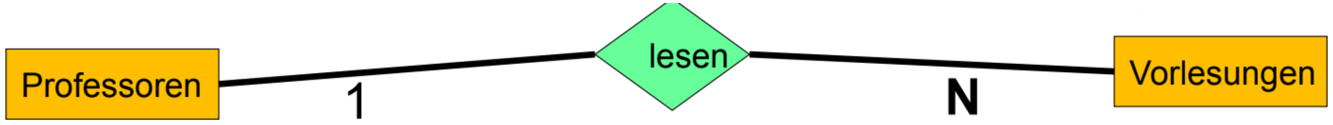


Besucht	
Matrikelnummer	Vorlesungsnummer
1234567	0000001
7654321	0000001
....	....

Example



min. 3 Tables



min. 2 Tables

→ cardinalities in ER diagrams helps to build compact relational models

Publication masters the following challenges

- general complexity of the German language
- different spellings
- word-based problems
- semantics and contextual understanding

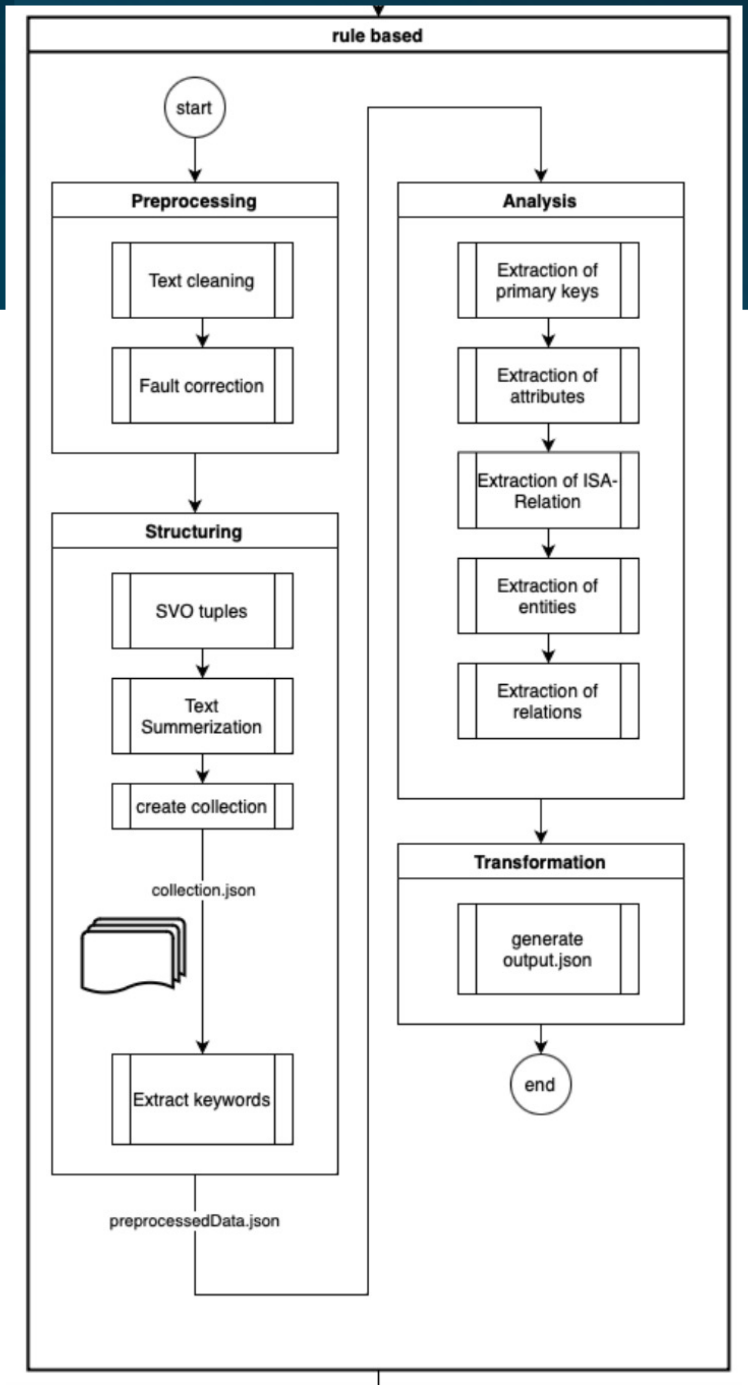
→identify only the relevant information without losing something

→Trade-off: accuracy vs flexibility

→avoiding the tradeoff with the **hybrid** approach

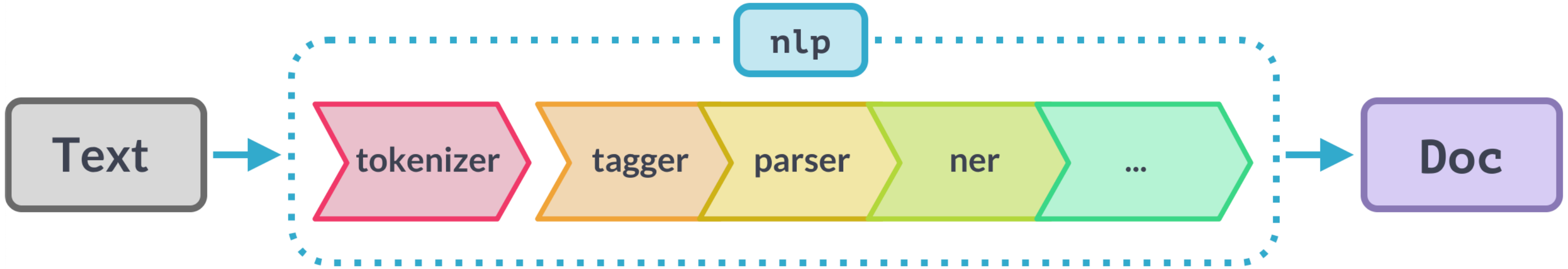
Architecture – rule based

1. made up of the sub-steps preprocessing, structuring, analysis and transformation.
2. during preprocessing, the text data is cleansed and corrected for spelling errors.
3. the structuring phase contains processes that break down the text into more meaningful parts and store them temporarily.
4. in the main block of the analysis, the ER components are extracted one after the other.



Implementation – rule based

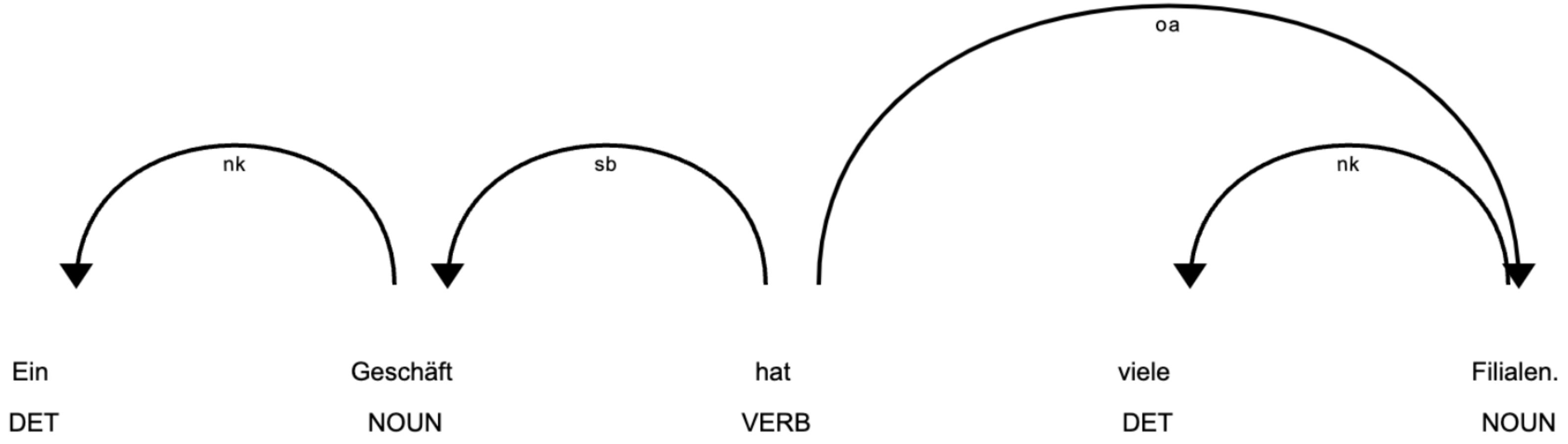
1. it is important not to define too many special rules, otherwise, the implementation will be too application-specific.
2. for the SpaCy model, it is sufficient to convert the text into a doc object so that the rule based analysis can be started.
3. different NLP techniques like levenshtein-similarity, normalized word frequency (TF), inverted document frequency (TF-IDF) is used.



Source: https://spacy.io/images/pipeline.svg

Implementation – rule based

1. to generate SVO tuples (for relations), the sentence must be analyzed using ***dependency parsing*** and ***POS-Tagging***.
2. certain rules can be used to analyze the grammatical structure of sentences.
3. similar to the tree structure, the successors or predecessors of a word are addressed with “children” or “parent”.



Implementation – rule based

Example 1: ISA-Inheritance

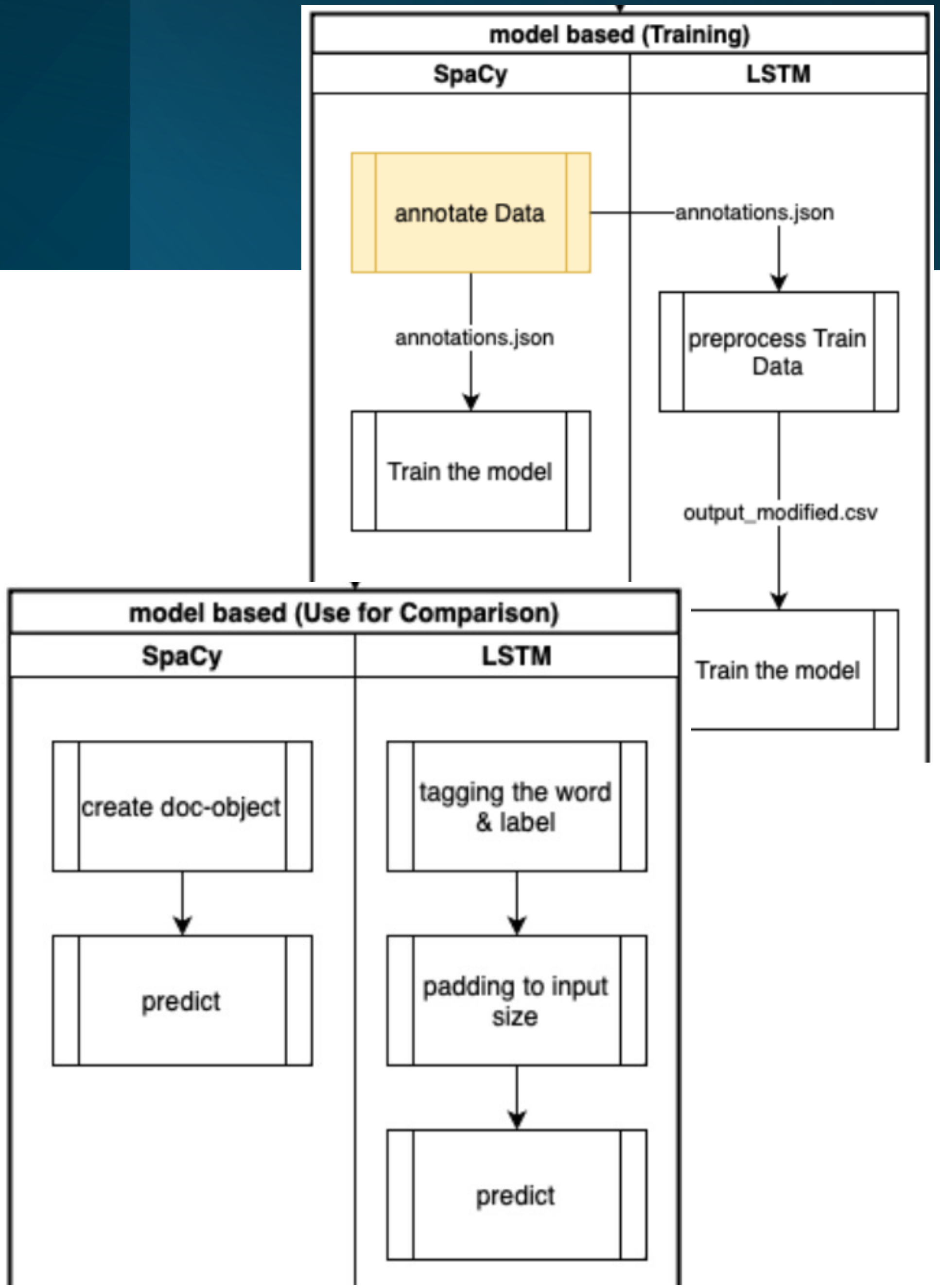
1. if a sentence contains the following verbs: [„einbeziehen“, „bestehen“, „umfassen“, „teilen“, „beinhalten“] , then the sentence describes a generalization.
2. if the sentence contains “Typ” as a word (noun), the first noun or entity is the generalization of the following nouns (entities).

Example 2: Attributes

1. attributes are identified as soon as a sentence contains a list of more than two nouns.
2. the first noun that occurs in the sentence is the entity to which the remaining nouns or attributes belong.

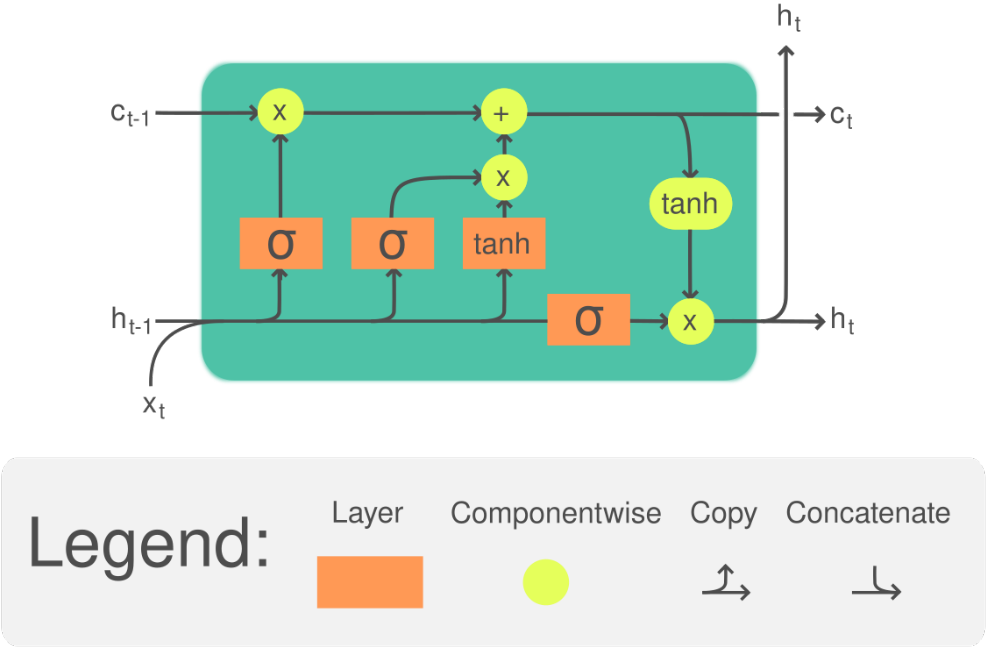
Architecture – model based

1. a different process path is carried out depending on the selected model type (SpaCy-Transformer or LSTM).
2. due to the individual input requirements of the models, the text data must be converted into the permitted form for both the training case and the use case.
3. the text for the LSTM model requires additional processing which decodes the words and labels into numerical values and scales the input word vector to a specified size (padding).



Implementation – model based

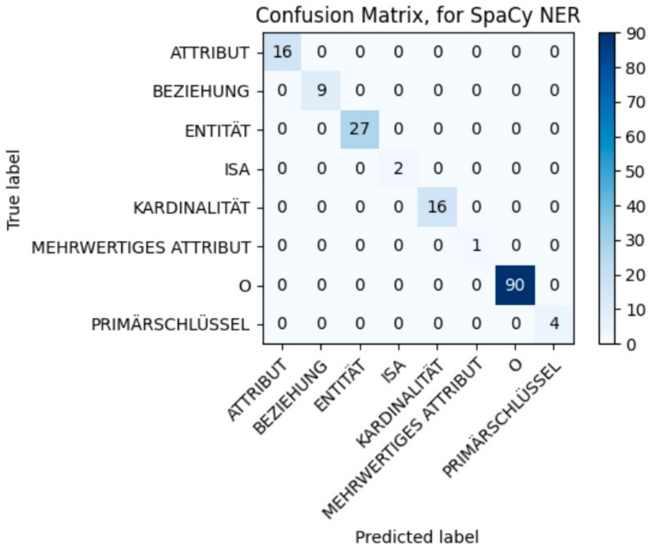
1. the selection of SpaCy as a model is based on its robust capabilities and user-friendly implementation.
2. SpaCy is recognized for its comprehensive documentation and strong support from an active developer community.
3. LSTM were chosen for their effectiveness in handling sequential data.
4. this makes them particularly suitable for tasks that require understanding the relationships within long text sequences of information.



Source: https://upload.wikimedia.org/wikipedia/commons/thumb/9/93/LSTM_Cell.svg/1200px-LSTM_Cell.svg.png

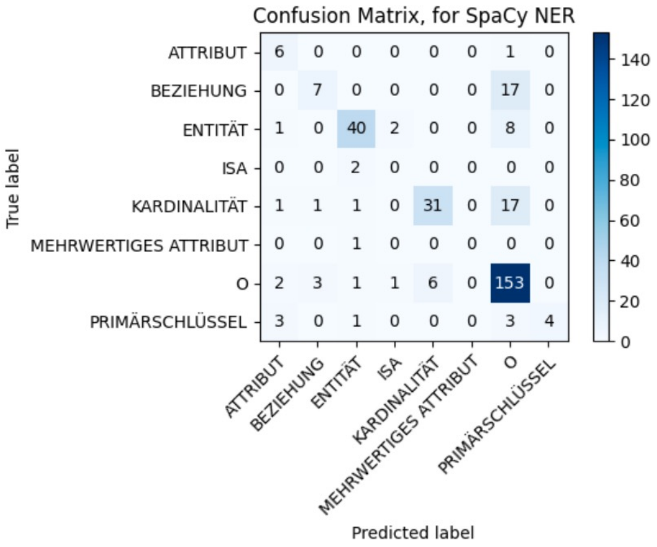
Performance – SpaCy model

1. the evaluation is based on several German texts that describe a specific DB scenario.
2. the left confusion matrix shows the reclassification case in which all previously labeled ER components are correctly classified during testing.
3. the right confusion matrix provides more information about the generalization capability of the SpaCy model, because the cross-validation also examines text data that was not part of the training process.



Confusion Matrix, for SpaCy NER

True label \ Predicted label	ATTRIBUT	BEZIEHUNG	ENTITÄT	ISA	KARDINALITÄT	MEHRWERTIGES ATTRIBUT	O	PRIMÄRSCHLÜSSEL
ATTRIBUT	16	0	0	0	0	0	0	0
BEZIEHUNG	0	9	0	0	0	0	0	0
ENTITÄT	0	0	27	0	0	0	0	0
ISA	0	0	0	2	0	0	0	0
KARDINALITÄT	0	0	0	0	16	0	0	0
MEHRWERTIGES ATTRIBUT	0	0	0	0	0	1	0	0
O	0	0	0	0	0	0	90	0
PRIMÄRSCHLÜSSEL	0	0	0	0	0	0	0	4



Confusion Matrix, for SpaCy NER

True label \ Predicted label	ATTRIBUT	BEZIEHUNG	ENTITÄT	ISA	KARDINALITÄT	MEHRWERTIGES ATTRIBUT	O	PRIMÄRSCHLÜSSEL
ATTRIBUT	6	0	0	0	0	0	1	0
BEZIEHUNG	0	7	0	0	0	0	17	0
ENTITÄT	1	0	40	2	0	0	8	0
ISA	0	0	2	0	0	0	0	0
KARDINALITÄT	1	1	1	0	31	0	17	0
MEHRWERTIGES ATTRIBUT	0	0	1	0	0	0	0	0
O	2	3	1	1	6	0	153	0
PRIMÄRSCHLÜSSEL	3	0	1	0	0	0	3	4

Limitations

The following aspects were not taken into account in the **rule-based** (but can be captured via model based) algorithm:

- attributes for relationships.
- special cardinalities (“two”, “three”, etc.).
- weak entities, relationships, attributes, etc.
- described ISA-Inheritance across several sentences.
- multi-valued or complex attributes.
- Coreference resolutions

Conclusion

1. two subsystems were developed which can extract ER components from unstructured texts.
2. special NLP methods are used for the rule-based extraction of entities, attributes and relationships.
3. due to the higher reliability of the rule-based results, these are used for the final ER model.
4. Transformers and LSTM models can be used to validate the rule-based results.
5. in the future, the results of the rule-based approach could be supplemented by an automated comparison of the results from the model-based approach.

Questions?

Thanks for your attention!