Automated Metadata Extraction, Semantic Analysis and Anonymization of German Court Rulings

Ingo Glaser, 24.06.2021, SEBIS Day

Lehrstuhl für Software Engineering betrieblicher Informationssysteme (sebis)
Fakultät für Informatik
Technische Universität München
wwwmatthes.in.tum.de

Überblick

Introduction

Quick Overview of Selected Results

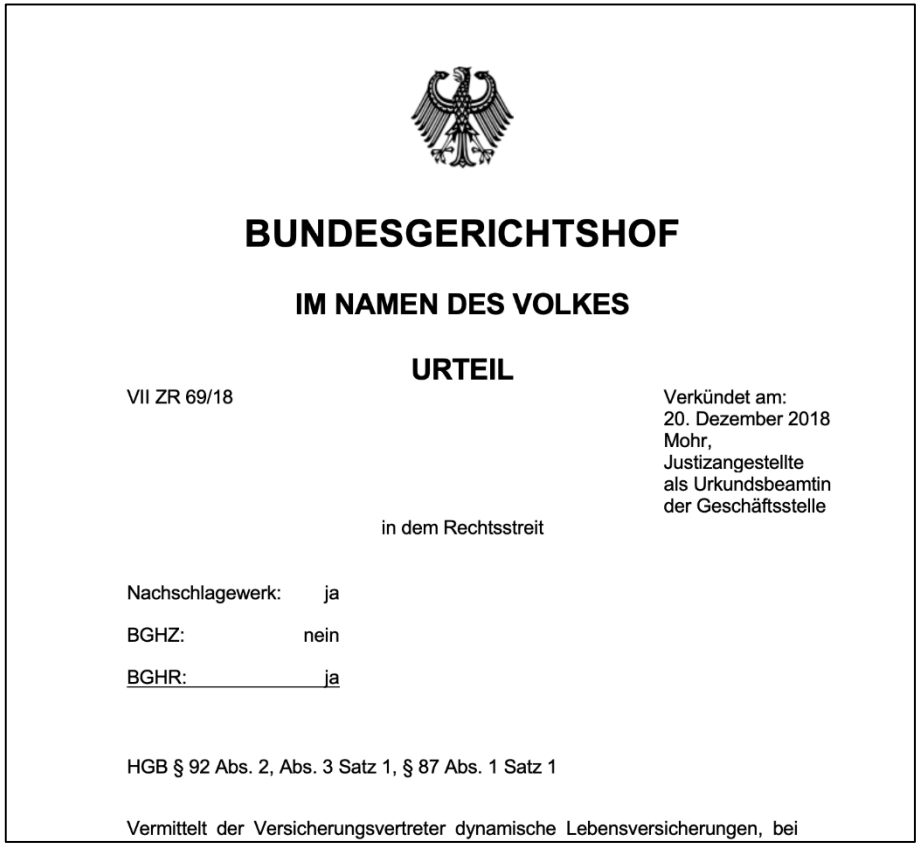
Anonymization

Discussion and Outlook

Introduction

Motivation

- Availability of court rulings in Germany is limited
- The process of digitizing court rulings is complex
  - Segmentation
  - Metadata
  - Semantic information
  - Anonymization
- Yet it is mostly done manual
- Modern natural language processing enables automating these steps
- Helpful for legal publishers but also to establish a single open platform



BUNDESGERICHTSHOF

IM NAMEN DES VOLKES

URTEIL

VII ZR 69/18

Verkündet am:
20. Dezember 2018
Mohr,
Justizangestellte
als Urkundsbeamtin
der Geschäftsstelle

in dem Rechtsstreit

Nachschlagewerk: ja

BGHZ: nein

BGHR: ja

HGB § 92 Abs. 2, Abs. 3 Satz 1, § 87 Abs. 1 Satz 1

Vermittelt der Versicherungsvertreter dynamische Lebensversicherungen, bei

Introduction

Motivation

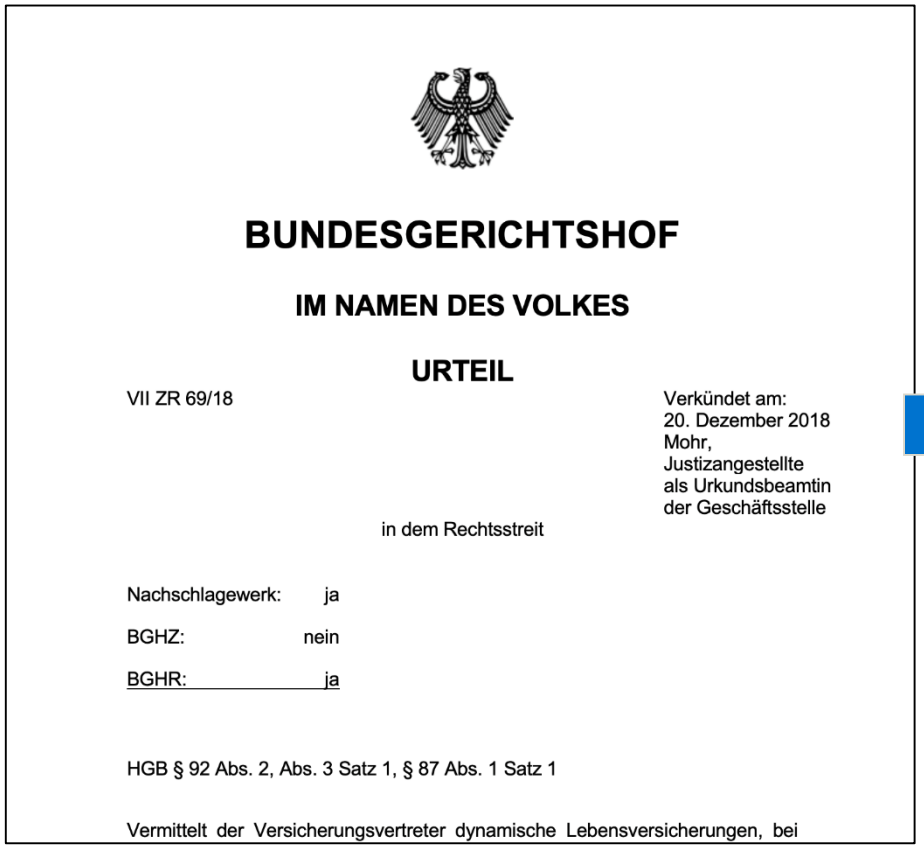
Problem:
- No automated processing of court rulings
- Limited search functionalities due to no semantic information

Solution:
- Automated transformation into machine-readable format
- Metadata extraction including segmentation
- Semantic analysis
- Anonymization

Goal: Creation of algorithms which are able to extract metadata, analysis verdicts with respect to certain semantic information, and anonymize legal court rulings.

Introduction

Problem Statement



BUNDESGERICHTSHOF

IM NAMEN DES VOLKES

URTEIL

VII ZR 69/18

Verkündet am:
20. Dezember 2018
Mohr,
Justizangestellte
als Urkundsbeamtin
der Geschäftsstelle

in dem Rechtsstreit

Nachschlagewerk: ja

BGHZ: nein

BGHR: ja

HGB § 92 Abs. 2, Abs. 3 Satz 1, § 87 Abs. 1 Satz 1

Vermittelt der Versicherungsvertreter dynamische Lebensversicherungen, bei

NLP

```
ORMENKETTE><VERWEIS-GS NORM="84" PUBKUERZEL="VVG">VVG § 84 Abs. 1</VERWEIS-GS>; AVB Kraftfahrzeugversicherung (hier A.2.18 AKB
ORM>AVB Kraftfahrzeugversicherung (hier A.2.18 AKB 2010)</NORM>
ORM>VVG § 84 Abs. 1</NORM>
EITSATZ>
P>Ein Mitarbeiter einer Partei ist kein Sachverständiger im Rahmen des Sachverständigenverfahrens nach A.2.18 AKB.</P>
LEITSATZ>
ERICHT>BGH</GERICHT>
NTSCHEIDUNGSTYP>Urt.</ENTSCHEIDUNGSTYP>
ATUM>
JAHR>2014</JAHR>
MONAT>12</MONAT>
TAG>10</TAG>
DATUM>
KTENZEICHEN>IV ZR 281/14</AKTENZEICHEN>
ORINSTANZ>
GERICHT>LG Frankfurt/O.</GERICHT>
ENTSCHEIDUNGSTYP>Urt.</ENTSCHEIDUNGSTYP>
DATUM>
<JAHR>2013</JAHR>
ONAT>12</MONAT>
>17</TAG>
TUM>
KTENZEICHEN>16 S 131/13</AKTENZEICHEN>
VORINSTANZ>
ORINSTANZ>
GERICHT>AG Strausberg</GERICHT>
ENTSCHEIDUNGSTYP>Urt.</ENTSCHEIDUNGSTYP>
DATUM>
<JAHR>2013</JAHR>
<MONAT>6</MONAT>
<TAG>13</TAG>
/DATUM>
AKTENZEICHEN>9 C 385/12</AKTENZEICHEN>
VORINSTANZ>
NTSCHEIDUNGSFORMEL>
P>Der IV. Zivilsenat des BGH hat durch die VorsRi Mayen, die Ri Wendt, Felsch, Lehmann und die Ri Dr. Brockmöller auf die mündl
P>für Recht erkannt:</P>
P>Auf die Revision des Klägers wird das Urteil der 6. Zivilkammer des LG Frankfurt/O. vom 17.12.2013 aufgehoben und die Berufun
G Strausberg vom 13.6.2013 zurückgewiesen.</P>
P>Der Beklagte trägt die Kosten der Rechtsmittelverfahren.</P>
P>Von Rechts wegen</P>
ENTSCHEIDUNGSFORMEL>
ATBESTAND>
TITEL>Tatbestand:</TITEL>
RZ>1</RZ>
P> Der Kläger begehrt vom Beklagten Ersatz eines Unfallschadens.</P>
RZ>2</RZ>
```

Überblick

Introduction

Quick Overview of Selected Results

Anonymization

Discussion and Outlook

Quick Overview of Selected Results

Segmentation

- Tailored sentence boundary detection for German legal domain: 97% (F1 score)

- Metadata extraction (F1 score)
  - Type: 100%
  - Id: 95%
  - Release date: 98%
  - Court: 96%
  - Norm chain: 93%
  - Previous instances: 98%

- Segmentation of the rulings components (F1 score)
  - Only ZPO: 89%
  - ZPO and SPO: 77%

➢ **Feasible for use in practice**
➢ However, **manual changes** must be possible
➢ Further research required

Detection of the Area of Law

- Machine Learning (F1 scores)
  - Employment Law, Tenancy Law, Commercial Law, and Others: **87%**
  - Employment Law, Tenancy Law, and Commercial Law: **96%**
  - 16 areas: **88%**

- Rule-based (F1 scores):
  - Employment Law, Tenancy Law, Commercial Law, and Others : **67%**
  - Employment Law, Tenancy Law, and Commercial Law : **80%**

➢ **Feasible for use in practice**
➢ In use at legal publisher
➢ Architecture can be used to easily train different areas

Quick Overview of Selected Results

Extraction of Norm Chains

- Rule-based (F1 score):
  - On law level: 68%
  - On norm level: 59%
  - On paragraph level: 52%

- Machine Learning (F1 score):
  - On norm level: 69%

➢ Not suitable to automatically detected relevant norms as of now

➢ Possibility to utilize in a **recommendation system**

➢ **Further research required**

Tenor Analysis

1. Main Decision: 95% F1 score
   - Success (Sustaining of the action, setting aside, …)
   - Non-success
   - Partial success
   - Submission to ECJ
   - No decision in these proceedings
2. Cost decision: 88% F1 score
   - Plaintiff (applicant) pays the costs
   - Defendant (or respondent) pays the costs
   - Cost sharing
   - Cancellation of costs
   - No decision on costs
3. Enforceability: 85% F1 score
   - Provisionally enforceable
   - Enforceable against security
   - Enforceable with possibility of avoidance
   - No decision

➢ **Feasible for use in practice**

➢ Training required on various domains in order to further improve results

Überblick

Introduction

Quick Overview of Selected Results

Anonymization

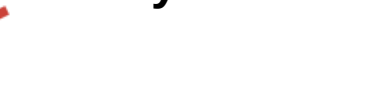
Discussion and Outlook

Anonymization

Overview

Example excerpt of a verdict:

Anonymization



6. Den Angeboten auf der Webseite ... liegt das Konzept der Produktdetailseite zugrunde. Dabei wird für jedes über die ...-Plattform angebotene Produkt jeweils nur eine Produkt- detailseite angezeigt; jedes Produkt enthält eine spezifische ...-Produktidentifikations- nummer (...) zugewiesen.

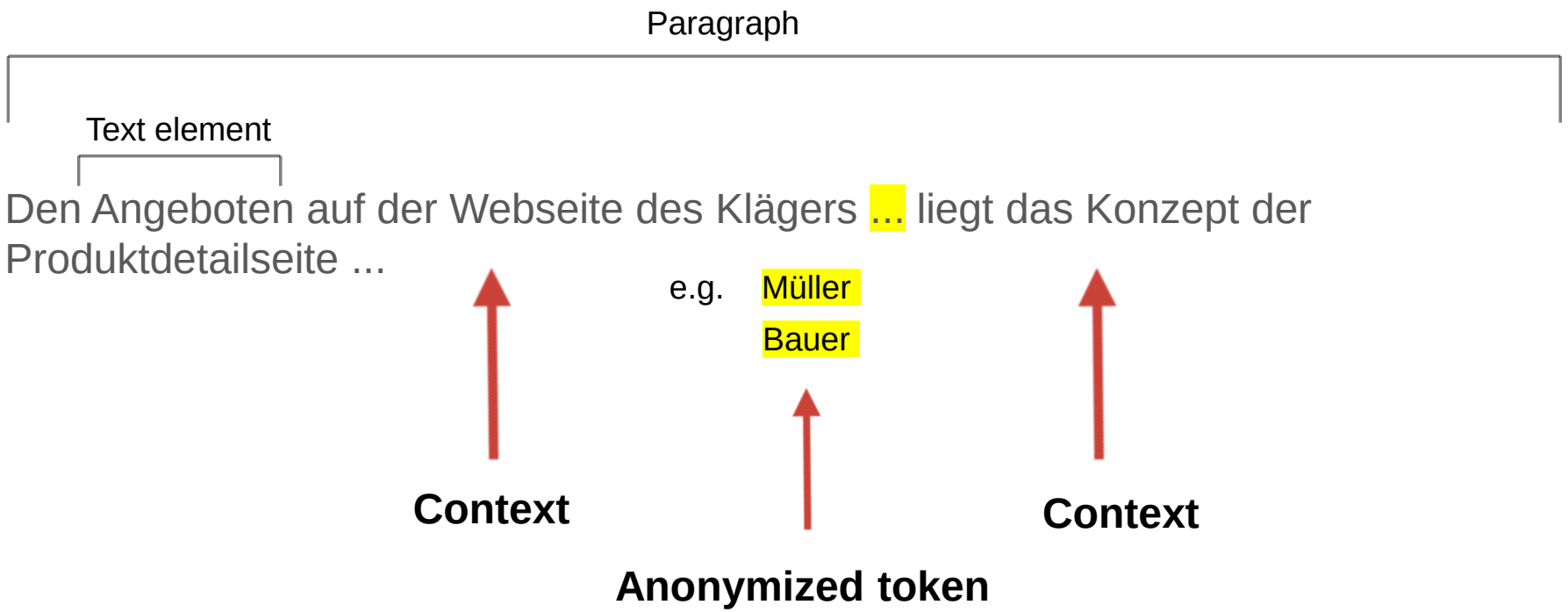
- **Tedious manual** process
  - Leads to infrequent publication of legal texts
  - Ensures few public legal text records
  - As a consequence, only very few public legal text records

- Solution: Using model-based NLP for **automatic sensitivity classification in legal texts**

- Problem: No non-anonymized datasets available
  - Approach: Training **only with anonymized data**

Anonymization

Solution

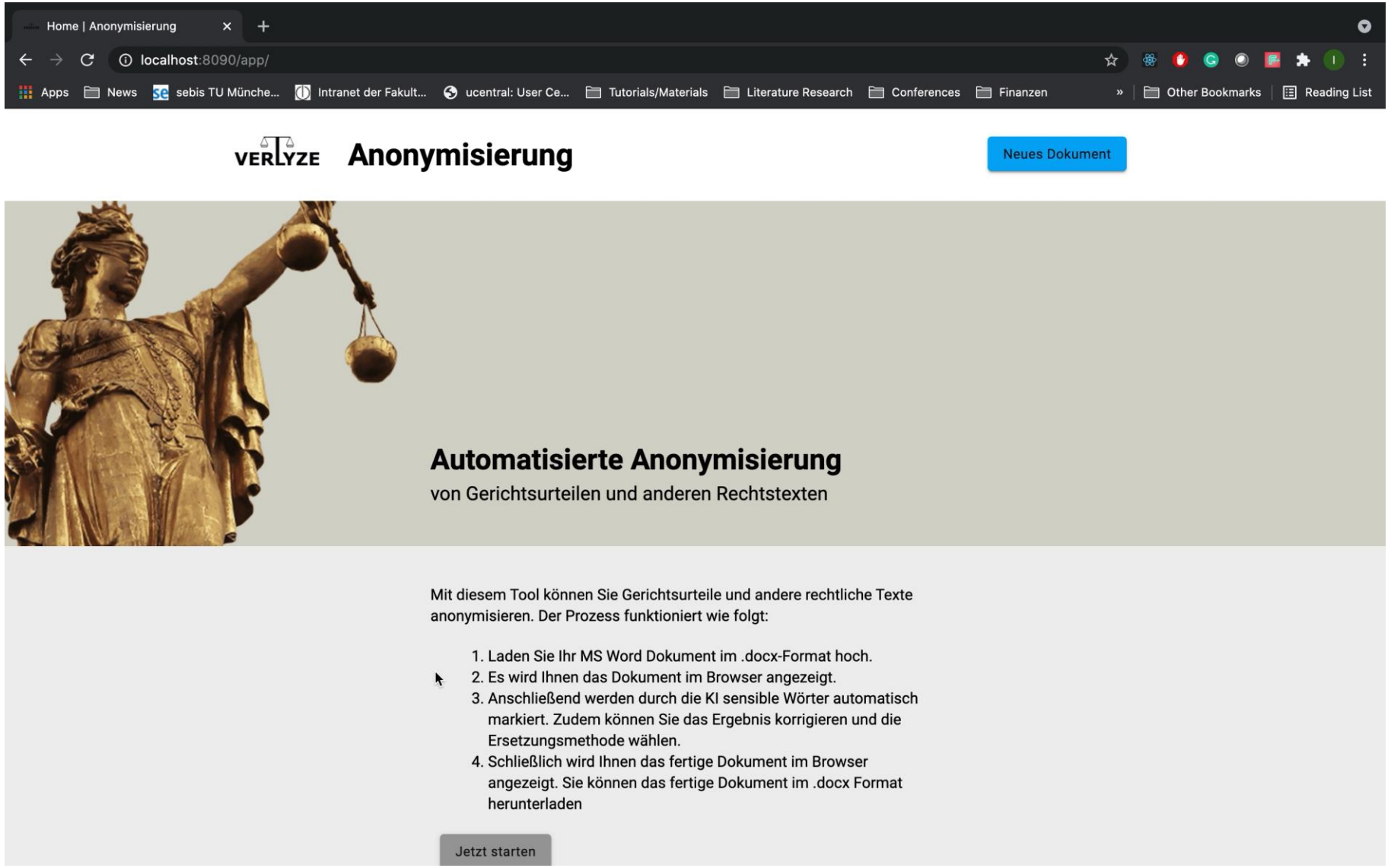


- Assumption: The **sensitivity** of text elements **depends only on the surrounding context**, not on the actual content

➢ The anonymization model is trained using the context of anonymized passages in anonymized legal texts

Anonymization

Demo of Prototype



Anonymization

Results

- Placeholder detection
  - Precision 95.9%
  - Recall: 98.0%

- Anonymization results
  - Munich district court data: 68.9% precision and 79.1% recall
  - Munich financial court data: 64.7% precision and 74.4% recall

- In most cases, the context-based model can distinguish sensitive text passages from insensitive text passages.

- However, these objects often appear as references (e.g. as "plaintiffs") to the actually sensitive text passages (e.g. names).

Anonymization

Results (ii)

- Example:

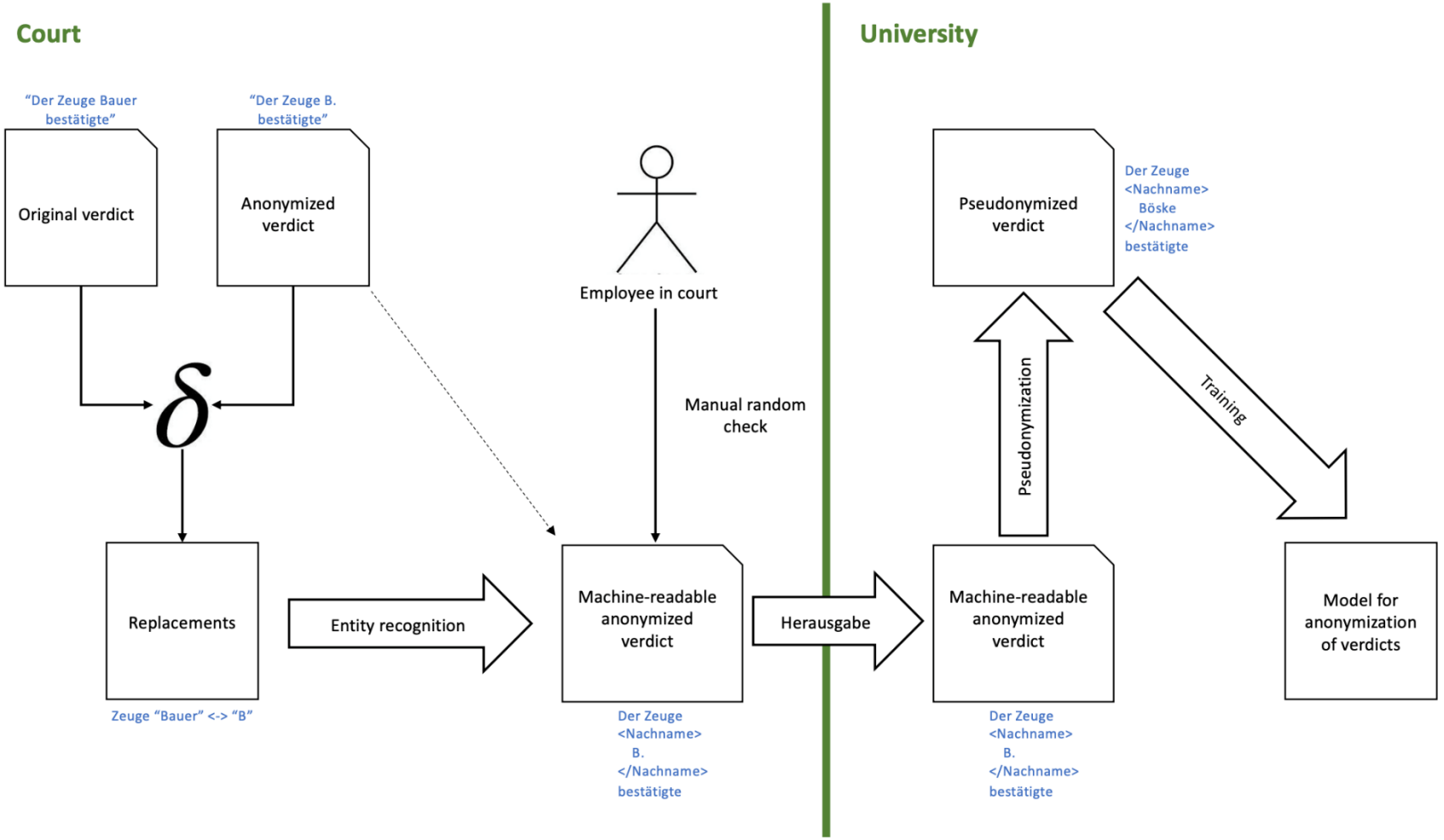
> *Unstreitig hat ein Mitarbeiter der Beklagten dem Kunden <<Fausner>> eine Krankenversicherung angeboten. Dass die angerufene Telefonnummer für die Firma des Kunden in einem Branchenverzeichnis eingetragen wäre, behauptet die Beklagte nicht, sie mutmaßt dies nur.*

- Problem: If only the context of text passages is used for recognition, sensitive text passages often cannot be distinguished from references to sensitive text passages.

➢ Possible solutions:
  - Detection of „named entities“ through general named entity recognition
  - Utilization of non-anonymized court rulings

Anonymization

Pseudonymization



Überblick

Introduction

Quick Overview of Selected Results

Anonymization

Discussion and Outlook

Discussion and Outlook

- Applied methods show promising results for
  - Metadata extraction
  - Segmentation
  - Tenor analysis
  - Area of law detection

- Anonymization challenging due to data scarcity/issues

- Yet an existing prototype exists

- Future work
  - Improvement of existing approaches
  - Looking for further helpful semantic information to improve legal information retrieval

MSc

Ingo Glaser

Wissenschaftlicher Mitarbeiter

Technische Universität München
Fakultät für Informatik
Lehrstuhl für Software Engineering
betrieblicher Informationssysteme

Boltzmannstraße 3
85748 Garching bei München

Tel +49.89.289.17138
Fax +49.89.289.17136

ingo.glaser@in.tum.de
wwwmatthes.in.tum.de