

Customizable Anonymization of German Legal Court Rulings using Domain-specific Named Entity Recognition

Master Thesis - Final

Tom Schamberger, Final 27th of September 2021

Chair of Software Engineering for Business Information Systems (sebis)
Faculty of Informatics
Technische Universität München
www.matthes.in.tum.de

Motivation

Problem Statement

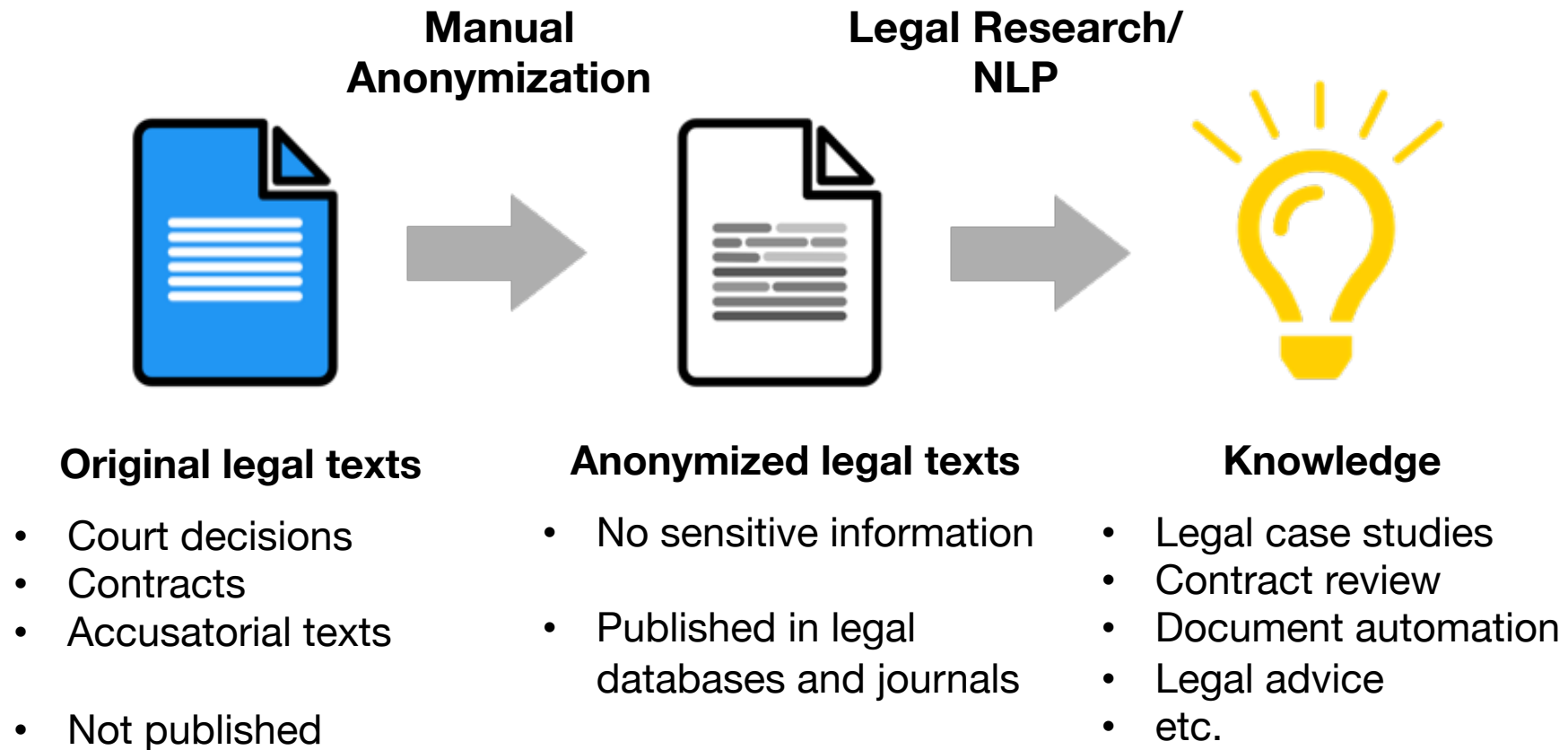
Research Questions

Approach

Results

Demo

Conclusion



- Challenges for courts
 - Mainly manual and time-consuming process
 - Prevents the publication of large amounts of court decisions
- Challenges for consumers of anonymized decisions
 - Few legal data publicly available
 - Heterogeneous anonymization conventions within and between courts
 - Anonymized entities marked using ambiguous patterns hardly detectable by machines (and even humans)

Motivation

Problem Statement

Research Questions

Approach

Results

Demo

Conclusion

Distinction between Sensitive and Insensitive Entities

- Anonymization needs to distinguish between sensitive and insensitive entities
- Entities of same type can be sensitive or insensitive, e.g., judge names and plaintiff names

In der Streitsache Philip Kraft, Erdinger Str. 7 8000 München, (Kläger) gegen Zentralfinanzamt München vertreten durch den Vorsteher, StNr.: 624/17345, (Beklagter) wegen Rückforderung von Steuerüberzahlungen hat der I. Senat des Finanzgerichts München unter Mitwirkung des Vorsitzenden Richters am Finanzgericht Heiner und der Richter am Finanzgericht Hütte und Schönberg sowie der ehrenamtlichen Richter Bär und Meyer ohne mündliche Verhandlung am 21. Dezember 1988 für Recht erkannt: ...

sensitive

insensitive

Customization of the Anonymization Procedure

- Anonymization guidelines strongly differ between and within courts
- Many anonymization rules depend on the specific case
- **Common customization requirements include:**
 - Anonymization of judge names
 - Revelation of birth years/locations
 - Revelation of specific circumstantial locations
 - Rounding of specific values (such as money values)

Motivation

Problem Statement

Research Questions

Approach

Results

Demo

Conclusion



Which named entity taxonomy is suitable for the anonymization of court decisions?



What customization degree can be achieved using the proposed anonymization procedure?



What anonymization performance can be reached using the proposed model architectures?

Motivation

Problem Statement

Research Questions

Approach

Results

Demo

Conclusion

General NE types

Domain-specific Examples

Person

- Natural party (e.g. plaintiff, applicant, defendant)
- Judge
- Other persons (e.g. lawyer, expert)

Time/Date

- Birth date of natural party
- Decision date
- Other dates

Organization

- Private organization
- Public organization

Location

- Birth location / Origin of natural party
- Location of concern (of e.g. regulation, application)
- Other locations

Miscellaneous

- Brand
- Cash value
- Other

Function class

Legal

- Specifying details about the judicial process
- E.g., courts, judge names, legal references

Identifying

- Identifying specific information about a person, organization or location
- E.g., personal names, private organizations, addresses

Circumstantial

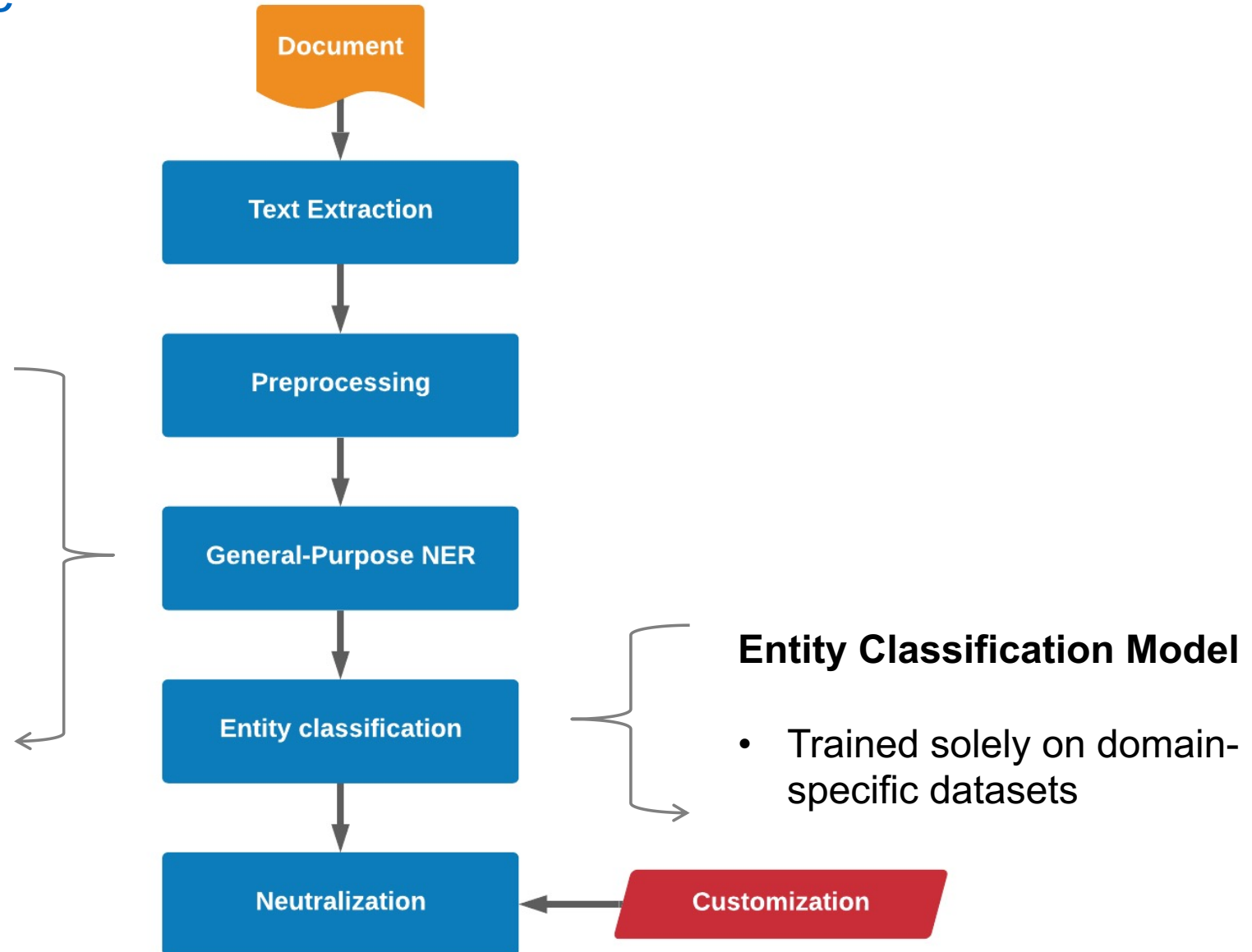
- Describing circumstances of the case
- E.g., correspondence, travel dates/locations

	Person	Organization	Time/Date	Location	Misc.
Legal	Judge	Court	Judgement date	Judgement location	Docket number ("Aktenzeichen")
Identifying	Plaintiff, defendant, witness	Private organization	Birth date, founding date	Birth location, place of business	ID number, website
Circumstantial	-	Authority, public department	Correspondence, event time/date	Event location	Money amount,

Anonymization Procedure

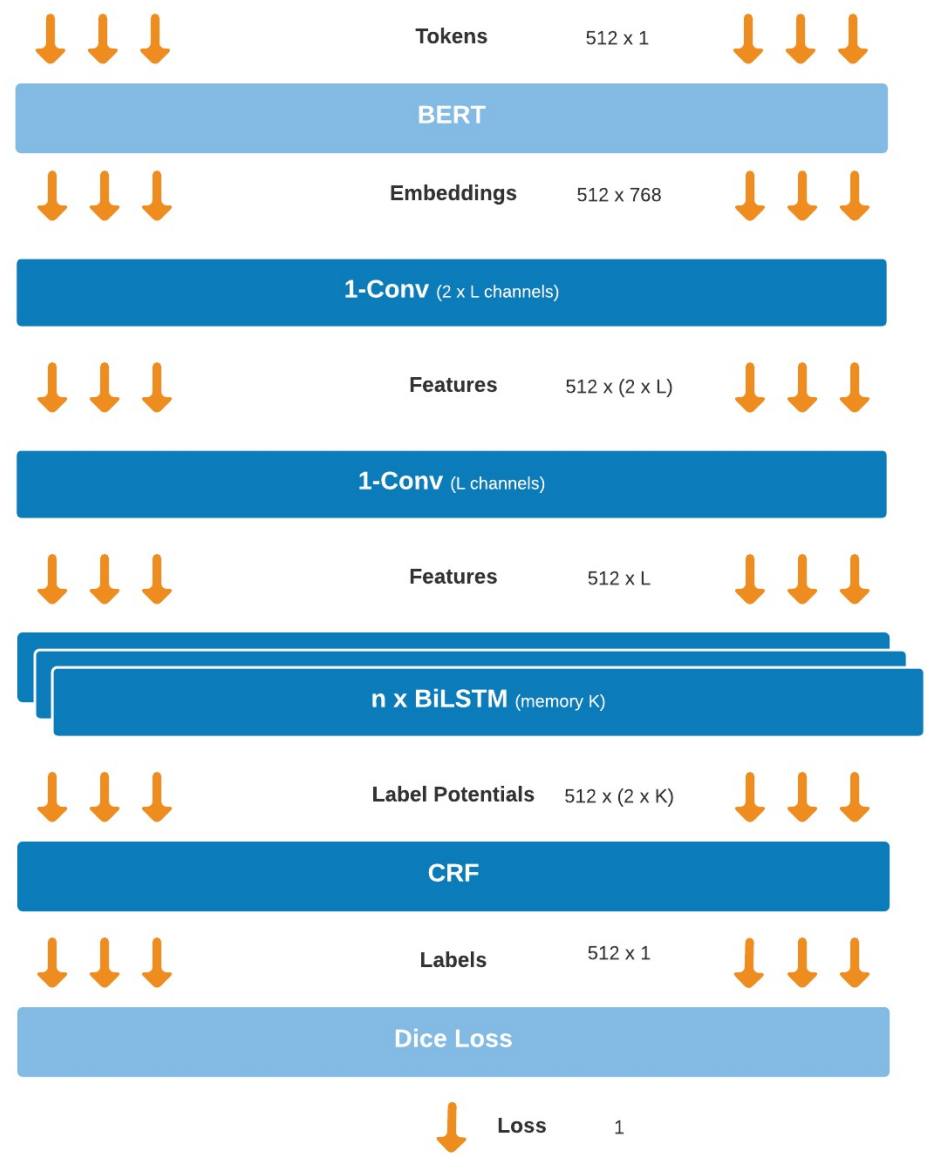
General-purpose NER Model

- Pre-trained on large amounts of general-purpose datasets
- Fine-tuned on domain-specific datasets



- **Original court decisions** from the “Finanzgericht München” (engl. financial court in Munich)
 - Training set: 217 Paragraphs
 - Test set: 71 Paragraphs
- **Rewritten court decisions** of the “Amtsgericht München” and the “Landgericht München” from the “Bayern.Recht” platform
 - Training set: 245 Paragraphs
 - Test set: 134 Paragraphs
- **General-purpose NER dataset** from “GermEval 2014” shared task (engl. financial court in Munich)
 - Training set: 31,000 Paragraphs

Model Architecture



Customization of the Neutralization Procedure

- For each domain-specific entity type, the degree of neutralization can be chosen
- Neutralization degree options depend on the general entity type
 - **Person**: Hide, Initials, Initials + Title, Reveal
 - **Organization**: Hide, Initials, Initials + Legal Form, Reveal
 - **Date/Time**: Hide, Only Year, Only Date, Reveal
 - **Location**: Hide, Only State, Only Region, Reveal
 - **Misc.:** Hide, First Characters, Reveal

Motivation

Problem Statement

Research Questions

Approach

Results

Demo

Conclusion

Results: General-Purpose NER Evaluation

	Acc	Prec	Rec	F1
PER	0.995	0.876	0.747	0.806
ORG	0.993	0.873	0.741	0.801
TIME	0.998	0.957	0.998	0.977
LOC	0.998	0.981	0.868	0.921
MISC	0.989	0.918	0.935	0.926
Average	0.995	0.921	0.858	0.886

Results: Entity Classification Evaluation

	Acc	Prec	Rec	F1
LEG	0.996	0.946	0.989	0.967
ID	0.996	0.909	0.977	0.942
CIRC	0.994	0.978	0.896	0.935
Average	0.995	0.944	0.954	0.948

Motivation

Problem Statement

Research Questions

Approach

Results

Demo

Conclusion

Demo

Motivation

Problem Statement

Research Questions

Approach

Results

Demo

Conclusion

- Separation of domain-specific NER into general-purpose NER and entity classification
 - Overcomes issues related to limited data
 - Enables anonymization flexibility
 - Yields good performance using the proposed model architecture
- General-purpose NER types and customization options may be extended in future work



Tom Schamberger

Technische Universität München
Faculty of Informatics
Chair of Software Engineering for Business
Information Systems

Boltzmannstraße 3
85748 Garching bei München

Tel +49.89.289. 17132
Fax +49.89.289.17136

matthes@in.tum.de
wwwmatthes.in.tum.de

