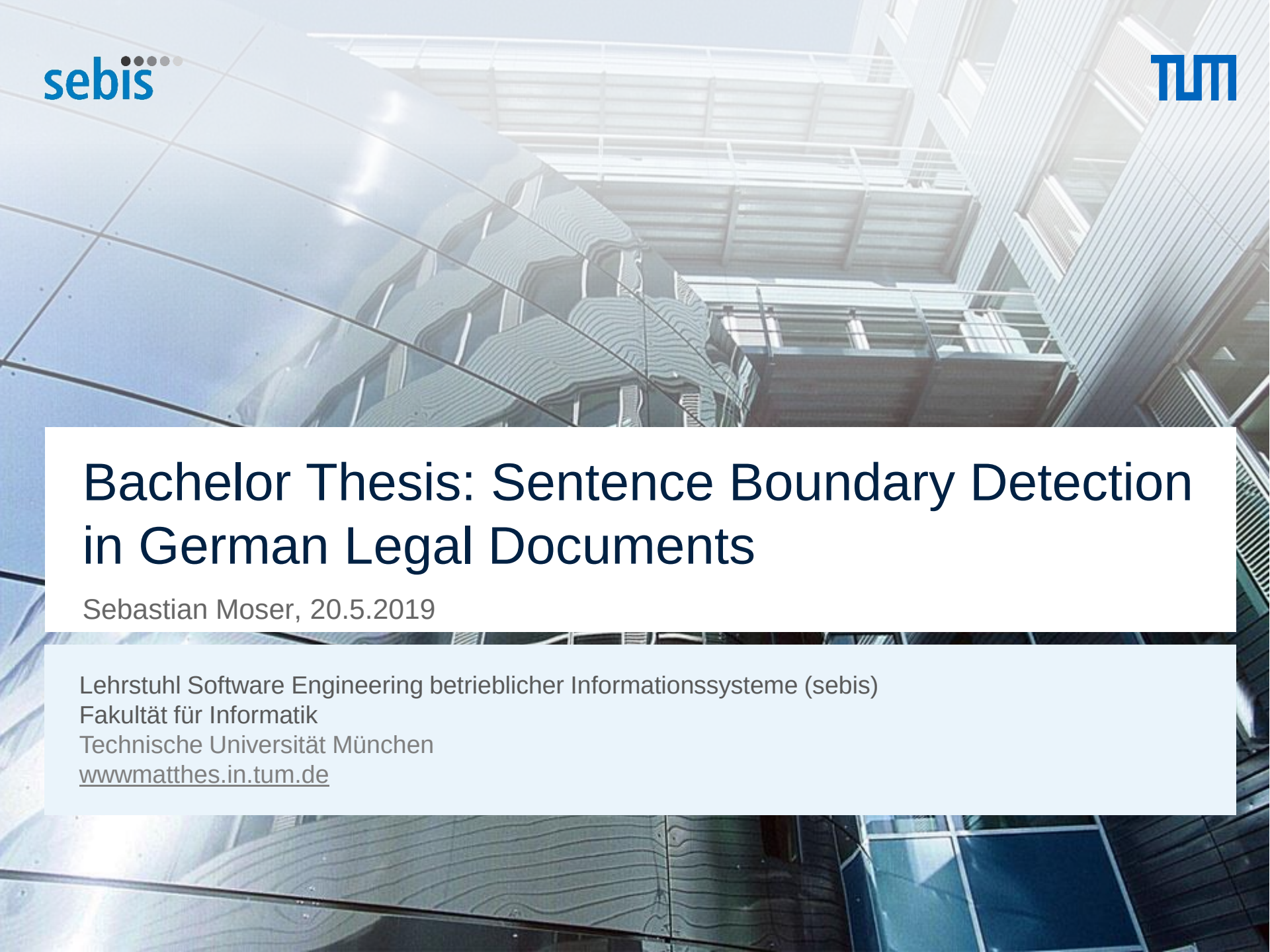


Bachelor Thesis: Sentence Boundary Detection in German Legal Documents

Sebastian Moser, 20.5.2019

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

Key Facts

- **Chair:** Software Engineering for Business Information Systems
- **Title:** Sentence Boundary Detection in German Legal Documents
- **Author:** Sebastian Moser (sebastian.moser@in.tum.de)
- **Advisor:** M.Sc. Ingo Glaser (ingo.glaser@tum.de)
- **Supervisor:** Prof. Dr. Florian Matthes (matthes@in.tum.de)
- **Dates**: 15.4.2019 – 15.8.2019

Overview

Introduction
- Sentence Boundary Detection
- Natural Language Processing Pipeline

Motivation
- German Legal Documents
- Example Use Case

German Legal Document Corpus

Tasks

Research Questions, Approach, Timeline

Sentence Boundary Detection

- Automatic detection of the start/end of sentence
- Perfect results expected (based on WSJ, Brown Corpus)

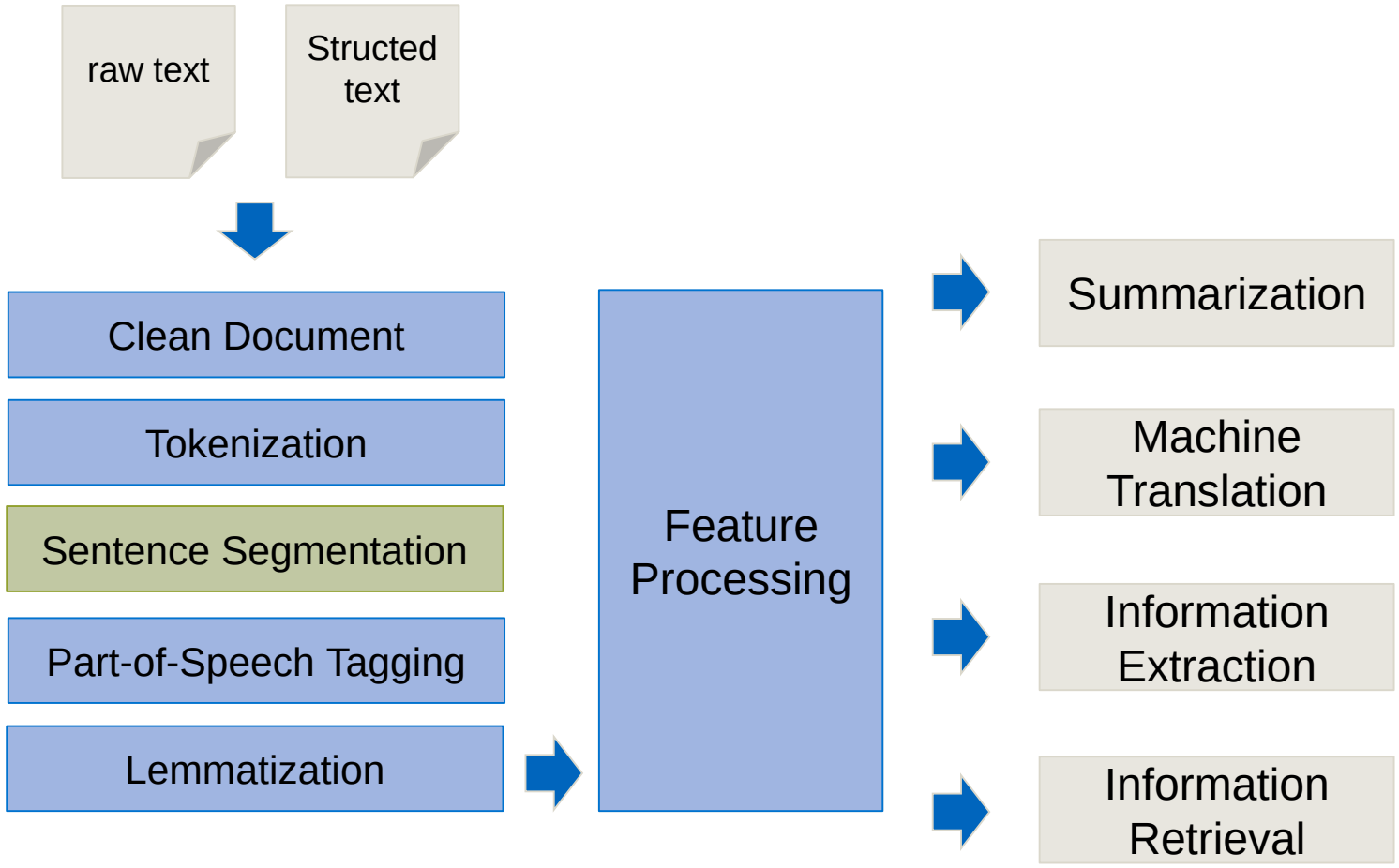
	BVA			CC			IP			SCOTUS			Overall		
	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
CoreNLP	.77	.84	.81	.80	.76	.78	.77	.81	.79	.77	.76	.76	.78	.78	.78
punkt	.68	.84	.75	.72	.79	.75	.69	.80	.74	.69	.80	.74	.70	.80	.75
openNLP	.77	.81	.79	.79	.75	.77	.80	.80	.80	.77	.78	.78	.78	.78	.78

[8]

Die Rechtsbeschwerde ist nach § 78 Satz 1 und Satz 2 ArbGG iVm. § 574 Abs. 1 Satz 1 Nr. 2 ZPO statthaft *(vgl. zB BAG 21. Juni 2006 - 3 AZB 65/05 - Rn. 8; 25. August 2004 - 1 AZB 41/03 - zu B I und B II 1 der Gründe mwN; GMP/Müller-Glöge 9. Aufl. § 78 Rn. 3)*. Sie ist auch im Übrigen zulässig, jedoch nicht begründet.

[1]

Natural Language Processing Pipeline



Overview

Introduction

- Sentence Boundary Detection
- Natural Language Processing Pipeline

Motivation

- Example
- Sentences in the Legal Domain

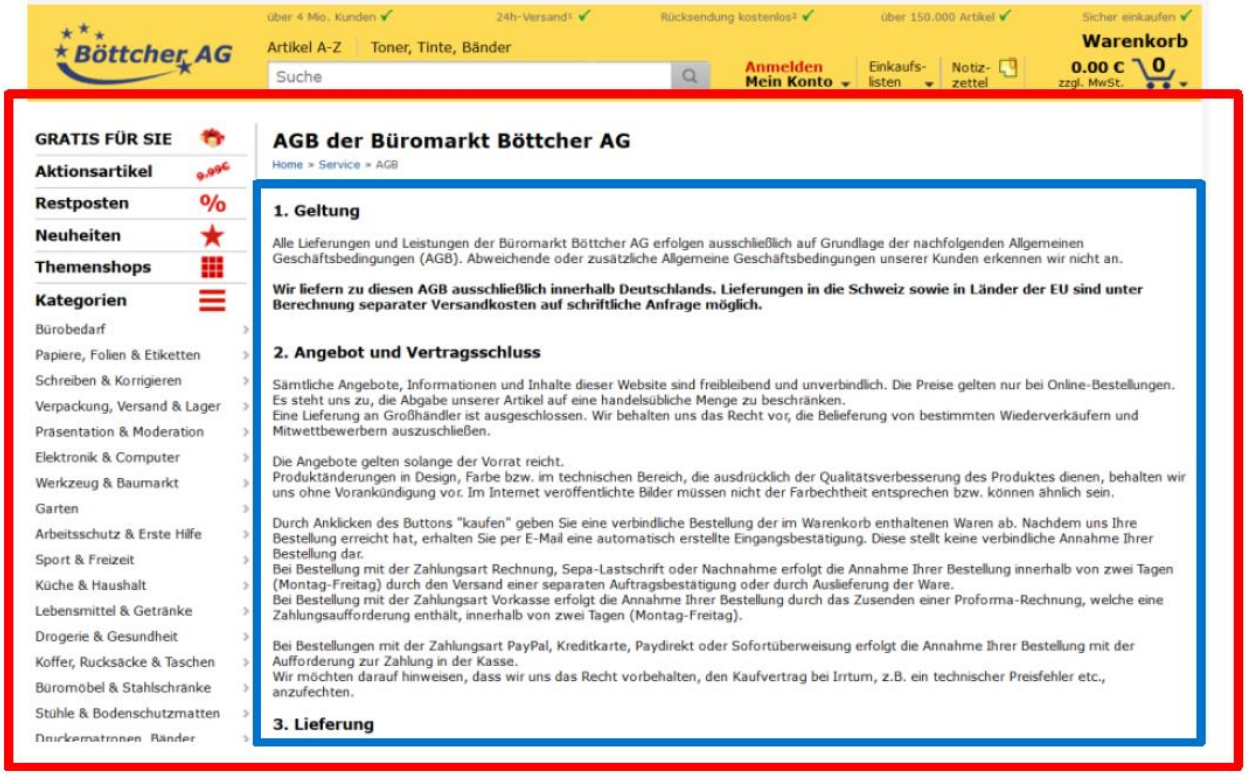
German Legal Document Corpus

Tasks

Research Questions, Approach, Timeline

Example: Analysis of General Terms and Conditions

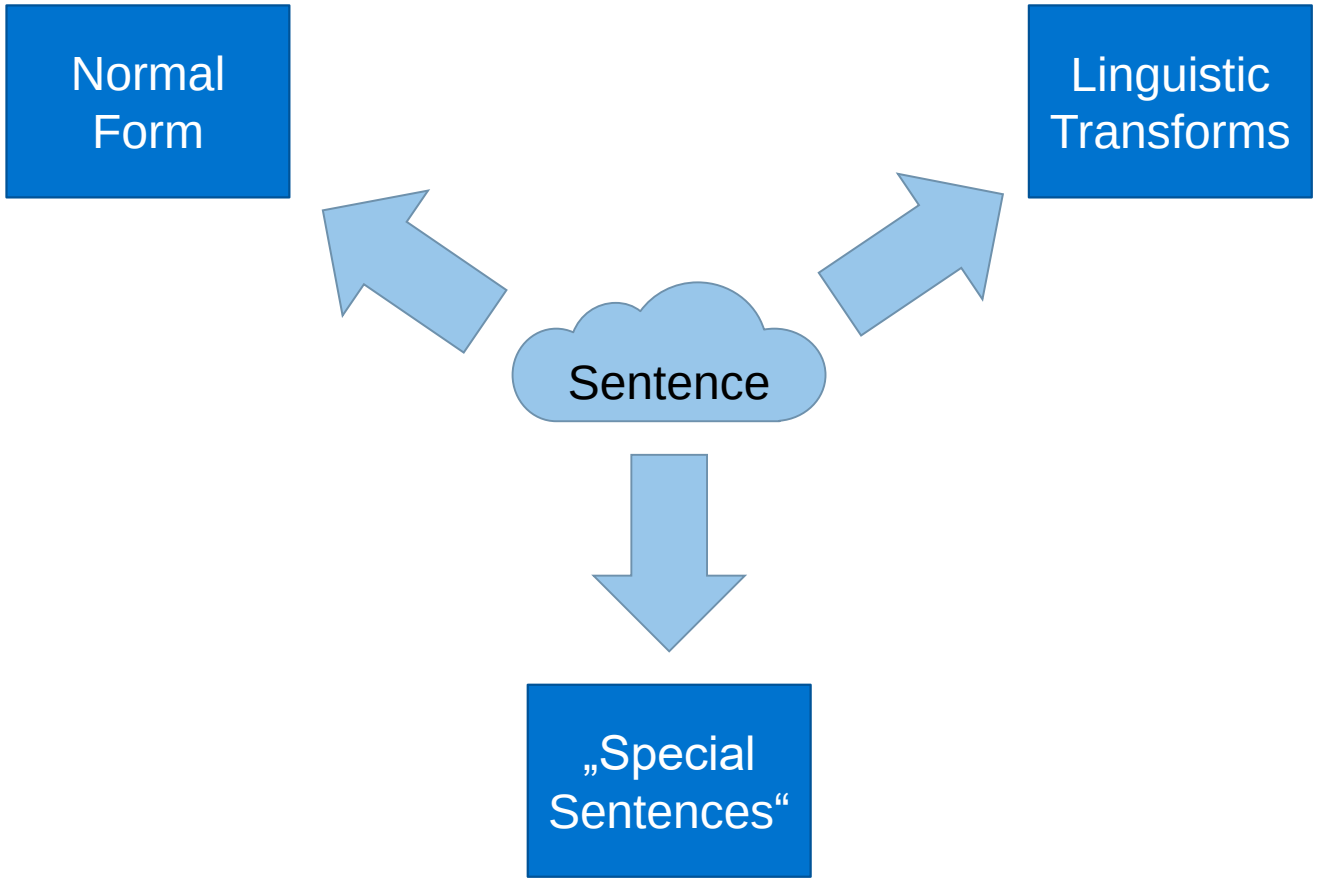
- Master's Thesis by David Koller
- Manual cleaning, segmenting and extracting General Terms and Conditions
- Automatic Segmentation system helpful



[12]

Sentences in the Legal Domain

- Sentence: “aus mehreren Wörtern bestehende, in sich geschlossene, eine Aussage, Frage oder Aufforderung enthaltende sprachliche Einheit” [12]



[9]

Sentences: “Special Sentences”

Headline

§ 286 Verzug des Schuldners

Sentence End (List)

(2) Der Mahnung bedarf es nicht, wenn

1. für die Leistung eine Zeit nach dem Kalender bestimmt ist,
2. der Leistung ein Ereignis vorauszugehen hat und eine angemessene Zeit für die Leistung in der Weise bestimmt ist, dass sie sich von dem Ereignis an nach dem Kalender berechnen lässt,
3. der Schuldner die Leistung ernsthaft und endgültig verweigert,
4. aus besonderen Gründen unter Abwägung der beiderseitigen Interessen der sofortige Eintritt des Verzugs gerechtfertigt ist.

§ 81 Stiftungsgeschäft

(1) Das Stiftungsgeschäft unter Lebenden bedarf der schriftlichen Form. […] Durch das Stiftungsgeschäft muss die Stiftung eine Satzung erhalten mit Regelungen über

1. den Namen der Stiftung,
2. den Sitz der Stiftung,
3. den Zweck der Stiftung,
4. das Vermögen der Stiftung,
5. die Bildung des Vorstands der Stiftung.

Enumeration

[2]

Sentences: Normal Form/ Linguistic Transform

Headline

§ 89 Haftung für Organe; Insolvenz

(1) Die Vorschrift des § 31 findet auf den Fiskus sowie auf die Körperschaften, Stiftungen und Anstalten des öffentlichen Rechts entsprechende Anwendung.
(2) Das Gleiche gilt, soweit bei Körperschaften, Stiftungen und Anstalten des öffentlichen Rechts das Insolvenzverfahren zulässig ist, von der Vorschrift des § 42 Abs. 2.

Normal Sentence

- More diverse sentence structure, longer sentences
- Citations, lists, combination of punctuation and alpha-numeric characters

  → Most assumptions for SBD systems do not hold for legal texts

  → Complicated decision on sentence segmentation [9] [10]

[2]

Overview

Introduction

- Sentence Boundary Detection
- Natural Language Processing Pipeline

Motivation

- Examples
- Sentences in the Legal Domain

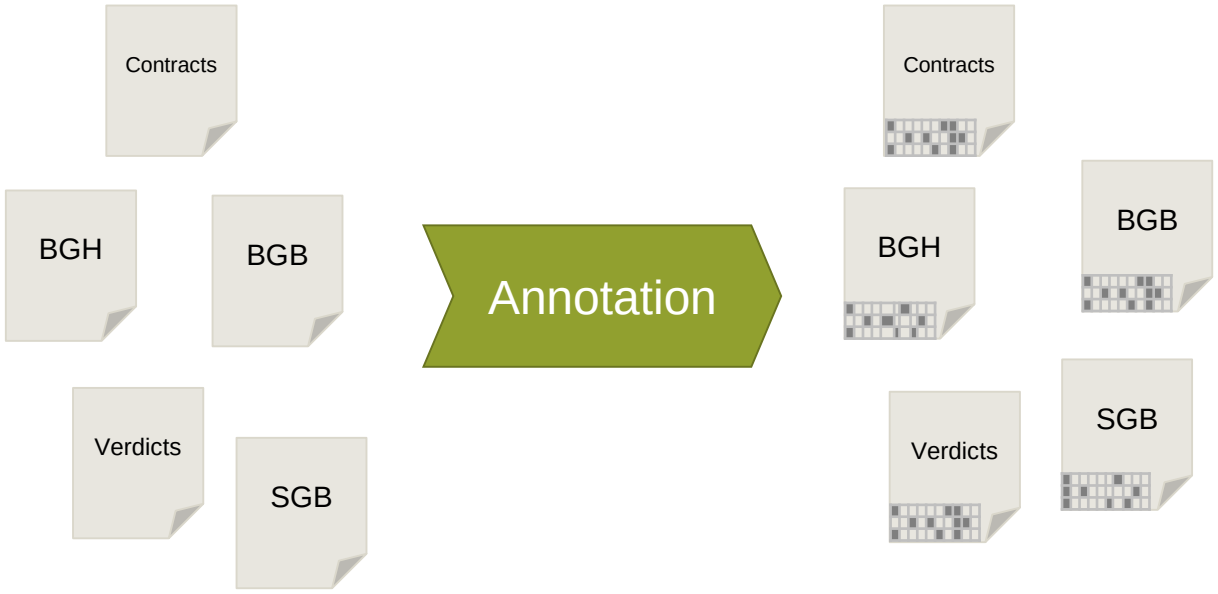
German Legal Document Corpus

Tasks

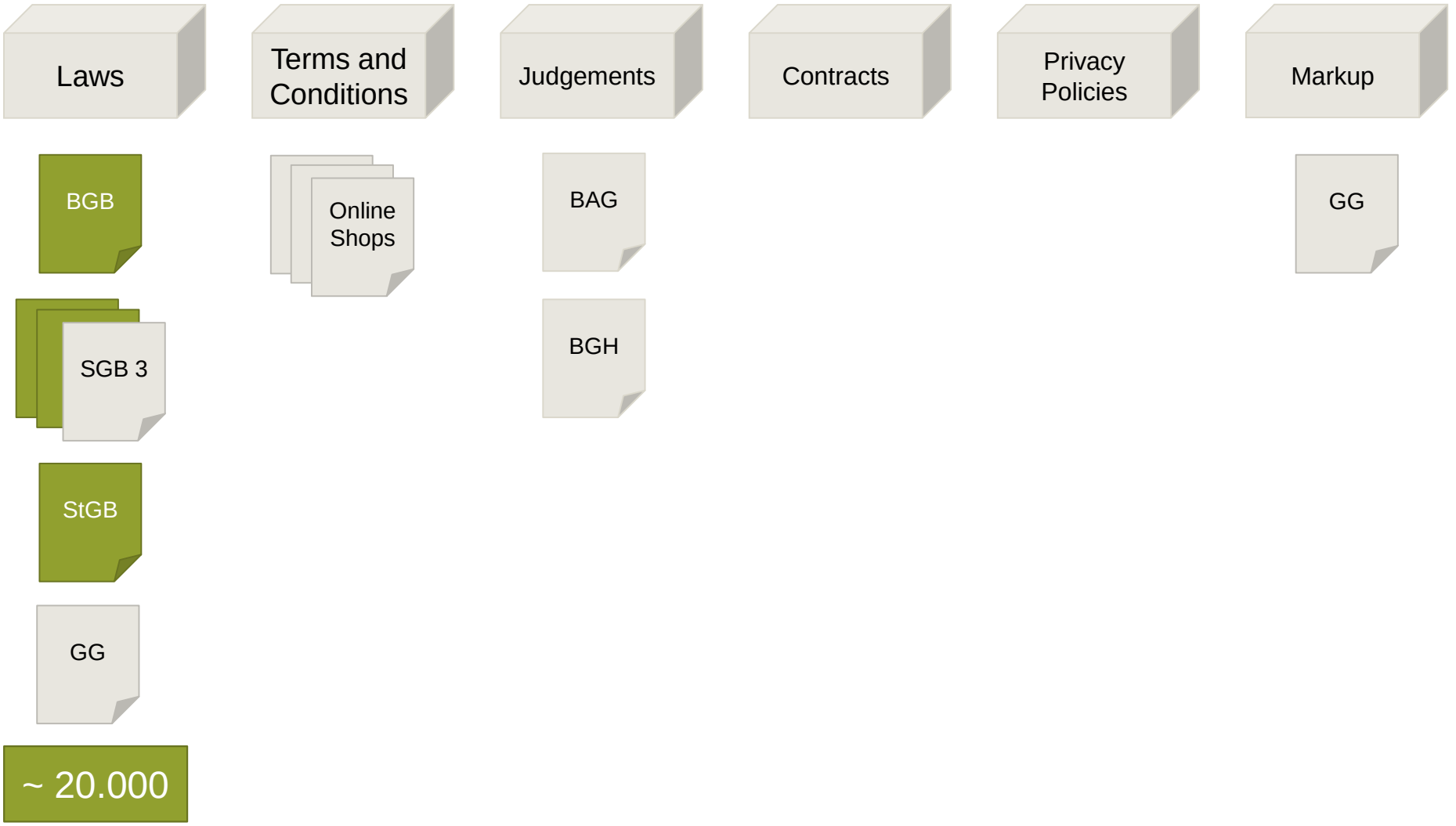
Research Questions, Approach, Timeline

German Legal Documents

- Currently no dataset for SBD in German Legal Documents
- Dataset needed for training, evaluation and validation of methods



German Legal Documents



Markup

- Markup ~ Additional structural annotations
- Evaluation of usefulness of additional markup information

```
<dl class="RspDL">
   <dt>
      <a name="rd_3">3</a>
   </dt>
   <dd>
      <p>2. Der Antragstellerin waren seitens des rbb zwei Sendeplätze zur Ausstrahlung eines 90-sekündigen Fernseh-
   </dd>
</dl>
<dl class="RspDL">
   <dt>
      <a name="rd_4">4</a>
   </dt>
   <dd>
      <p>3. Das Verwaltungsgericht wies den auf Verpflichtung des rbb zur Ausstrahlung gerichteten Eilantrag der An
   </dd>
</dl>
<dl class="RspDL">
   <dt>
      <a name="rd_5">5</a>
   </dt>
   <dd>
      <p>4. Mit ihrem Antrag auf Erlass einer einstweiligen Anordnung rügt die Antragstellerin die Verletzung ihrer
   </dd>
</dl>
```

[13]

Overview

Introduction

- Sentence Boundary Detection
- Natural Language Processing Pipeline

Motivation

- Examples
- Sentences in the Legal Domain
- German Legal Documents

German Legal Document Corpus

Tasks

Research Questions, Approach, Timeline

SBD for legal domain

- Implementation of Sentence Boundary Detection system for legal domain

- Possible methods:
  - Advanced Pre-processing + existing Solution
  - Tailored Rule-Based model
  - Conditional Random Fields [4] [5]
  - Hidden Markov model [6]
  - Long Short Term Memory networks
  - Maximum entropy model [7]

Existing Solutions

Rule-Based	Supervised Machine Learning	Unsupervised
CoreNLP	OpenNLP	NLTK
RASP	splitta	
LingPipe		

Overview

Introduction

- Sentence Boundary Detection
- Natural Language Processing Pipeline

Motivation

- Legal Tech
- German Legal Documents
- Example Use Case

German Legal Document Corpus

Tasks

Research Questions, Approach, Timeline

Research Questions

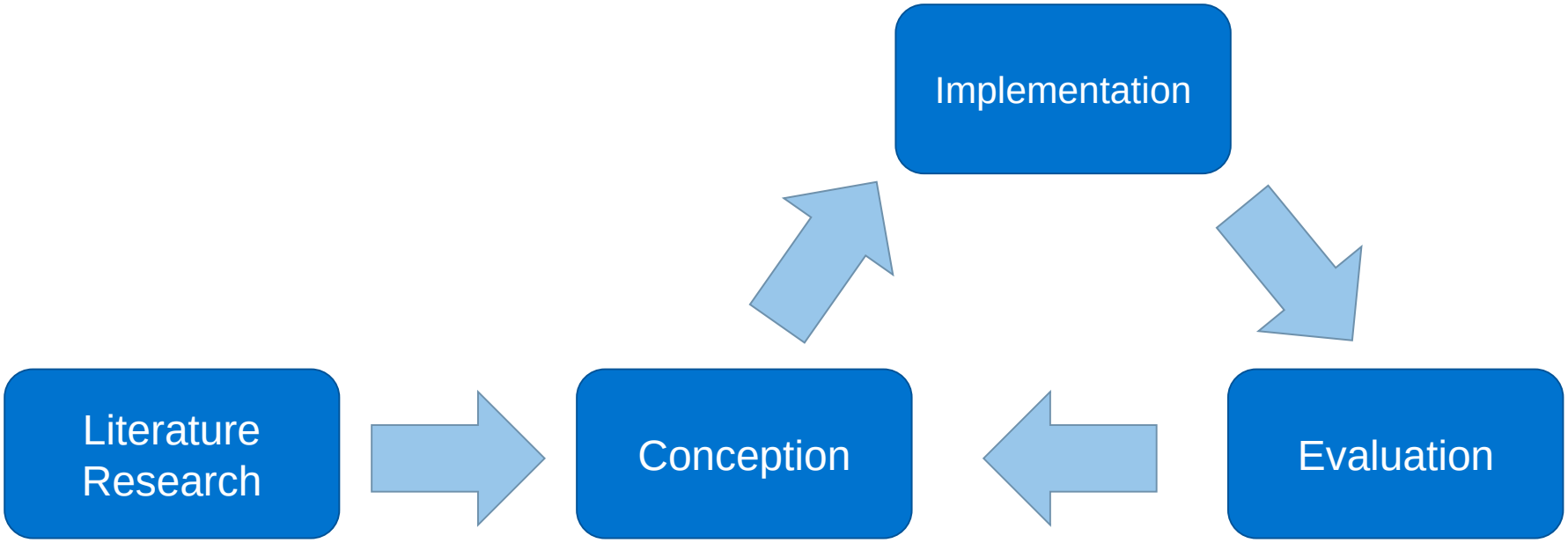
- What are sentences in the legal domain?

- Which methods are state of the art solutions on other domains?
- Best methods for SBD on German legal documents?

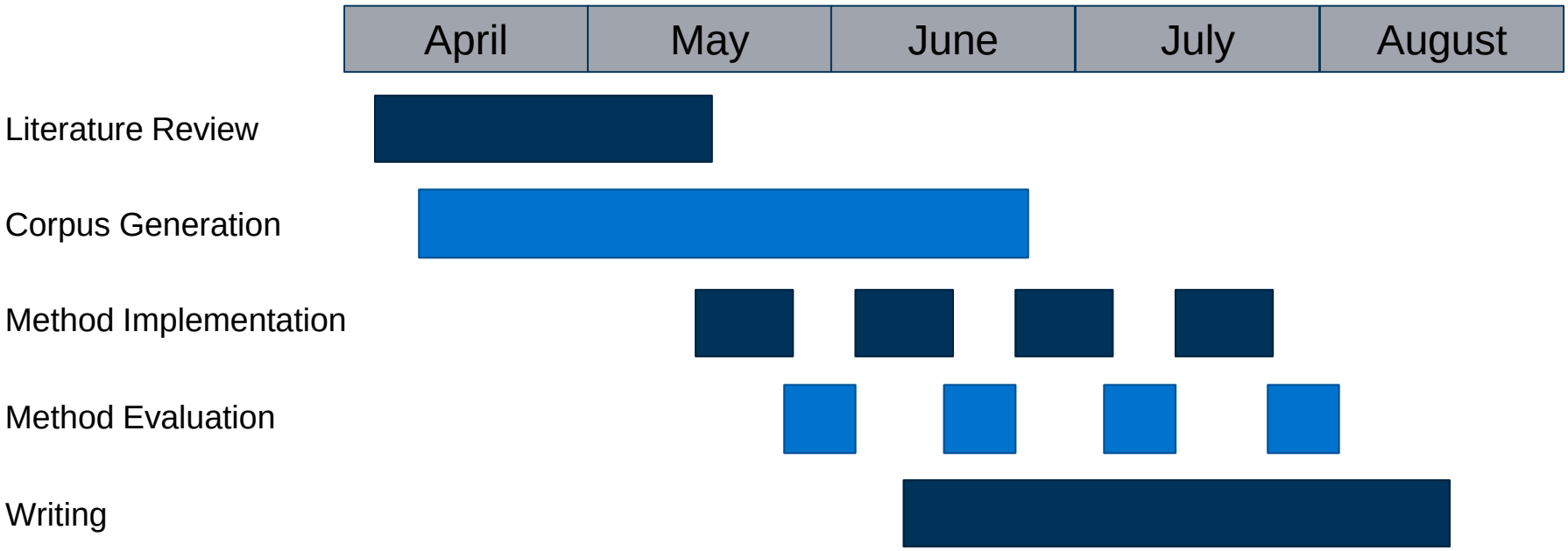
- Functional/ non-functional requirements of the SBD system?
- How good are existing approaches on Legal Documents?

- How should the document corpus be build?
- Do different legal document types need different solutions?

Research Approach



Timeline


April
May
June
July
August
Literature Review
Corpus Generation
Method Implementation
Method Evaluation
Writing

References

[1] Bundesarbeitsgericht, 10 AZB 44/18, 28.2.19

[2] Bürgerliches Gesetzbuch (Fassung vom 2. Januar 2002). https://www.gesetze-im-internet.de/bgb/

[3] Read J., Dridan R., Oepen S., Solberg L. (2012) *Sentence Boundary Detection: A Long Solved Problem?* Proceedings of COLING 2012: Posters, page 985-994, COLING 2012, Mumbai, December 2012

[4] Lafferty J., McCallum A., Pereira F. C.N. (2001) *Conditional Random Fields: Probabilistic Models for Segmentating and Labeling Sequence Data*

[5] Liu Y., Stolcke A., Shriberg E., Harper, M. (2005). *Using Conditional Random Fields for Sentence Boundary Detection in Speech*

[6] Jurish B., Würzner, K. (2013). *Word and Sentence Tokenization with Hidden Markov Models*. JLCL. 28. 61-83.

[7] Ratnaparkhi, A. (1998). Maximum Entropy Models For Natural Language Ambiguity Resolution. PhD thesis.

[8] Savelka, J., Walker, V. R., Grabmair, M., & Ashley, K. D. (2017). Sentence Boundary Detection in Adjudicatory Decisions in the United States. Revue TAL, 58(2), 21-45.

[9] Wyner A., Peters W., “On Rule Extraction from Regulations.”. JURIX, vol. 11, p. 113-122, 2011.

[10] Wikipedia. *European Union symbols*. https://en.wikipedia.org/wiki/Symbols_of_the_European_Union

[11] Duden: “Satz” unter https://www.duden.de/rechtschreibung/Satz (last access: 18.05.2019)

[12] Koller, D. (2019). *Interactive Analysis of a Corpus of General Terms and Conditions for Variability Modeling*. Master’s Thesis.

[13] BVerfG 1. Senat 2. Kammer, 1 BvQ 43/19, 15.5.19

Sebastian Moser

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

Boltzmannstraße 3
85748 Garching bei München

Tel +49.89.289.17132
Fax +49.89.289.17136

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