Master‘s Thesis Kickoff – Semantic Analysis and Structuring of German Legal Documents using Named Entity Recognition and Disambiguation

Ingo Glaser, 10.04.2017

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

Outline

1. Motivation
2. Administrative Setup
3. Problem Statement
4. Research Approach
5. Solution
6. Research Questions and Challenges
7. Thesis Timeline

Motivation

- Legal technology is rising [BCG]
  - Digitalisation of legal documents [Saravanan]
  - Increasing number of startups
  - New and changing business models [Deloitte]

- Unstructured and semi-structured data [Svyatkovskiy]
  - Modelling and structuring of legal documents
  - Understanding the content of documents
  - Creating added value

- Capability of systems and algorithms [Waltl]
  - Computational power increases continuously
  - Technologies such as Apache Spark or Hadoop allowing even more powerful clusters
  - Natural Language Processing
  - Machine Learning and Data Mining

Outline

1. Motivation
2. Administrative Setup
3. Problem Statement
4. Research Approach
5. Solution
6. Research Questions and Challenges
7. Thesis Timeline

Administrative Setup

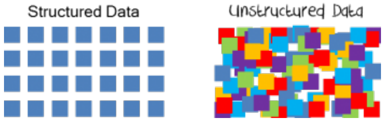
- **Chair:** Software Engineering for Business Information Systems
- **Title:** Semantic Analysis and Structuring of German Legal Documents using Named Entity Recognition and Disambiguation
- **Author:** Ingo Glaser (ingo.glaser@tum.de)
- **Supervisor:** Prof. Dr. Florian Matthes (matthes@in.tum.de)
- **Advisor:** M.Sc. Bernhard Waltl (b.waltl@tum.de)
- **Start:** 15th of March, 2017
- **End:** 15th of September, 2017

Outline

1. Motivation
2. Administrative Setup
3. Problem Statement
4. Research Approach
5. Solution
6. Research Questions and Challenges
7. Thesis Timeline

Problem Statement

- Legal documents as unstructured and semi-structured data [Hashmi]
  - Often plain text
  - Lack of consistency
  - Not suitable to be processed by systems



- Content and meaning of documents is unknown [Waltl]
  - Purpose of a document
  - Included entities
  - Norms

- Many tasks need to be performed manually
  - Missing added value of IT

Outline

1. Motivation
2. Administrative Setup
3. Problem Statement
4. Research Approach
5. Solution
6. Research Questions and Challenges
7. Thesis Timeline

Research Approach

1. Research

- Literature Research
- Finding state of the art solutions

2. Machine Learning to support information extraction

- What is Named Entity Recognition?
- Which value does NER add to contract analysis?
- How can Keyword Extraction help to recognize semantic meaning?
- How to dissolve ambiguities by using Word-sense disambiguation?
- How to utilize Semantic Role Labeling in order to classify the entities into semantic functions?

3. Interviews

- Understand needs of lawyers
- Identify stakeholders and understand their intention

4. Implementation

- Implement a prototypical usecase within Lexia

5. Evaluation

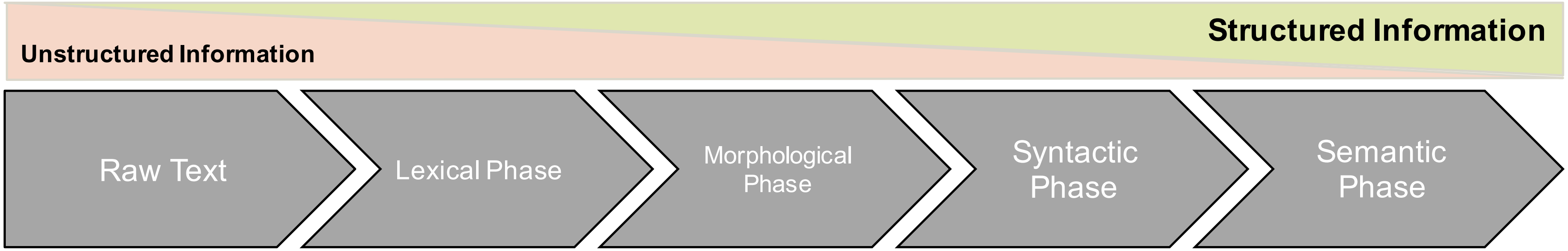
- Evaluate the use case and different scenarios

Outline

1. Motivation
2. Administrative Setup
3. Problem Statement
4. Research Approach
5. Solution
6. Research Questions and Challenges
7. Thesis Timeline

Solution

Text Processing Chain for Information Retrieval [Singh]



- Sequence of plain characters
- Can belong to various domains
- Multiple types of documents

- Text Segmentation
- Tokenization
- POS-Tagging
- Stop-Words

- Stemming
- Lemmatization
- Coreference Resolution

- Dependency Parsing
- Constiuency Parsing
- Predicate Annotations
- Syntactic Structures

- Named Entity Recognition
- Keyword Extraction
- Word Sense Disambiguation
- Semantic Role Labeling
- Latent Semantic Analysis
- Measures
- Sentiments
- (Text-Segmentation)

Comments

Appropriate segmentation in phases?

Strict boundaries between phases may not be feasible!

Semantic Analysis outside of Information Retrieval?

Quantity of methodologies?

Selection of methodologies?

Extraction and Annotation of Informations

Named Entity Recognition

- Monetary values
- Dates

Rule-based Approaches
(e.g. Apache Ruta, RegEx)

- References

Rule-based Approaches
(e.g. Apache Ruta, RegEx)

- Named Entities
  - Persons
  - Organisations
  - Locations

Rule-based Approaches
(e.g. Apache Ruta, RegEx)

Knowledge Bases
(e.g. DBPedia, OpenCalais)

- Keywords

Statistical Approaches &
Graph-based Approaches

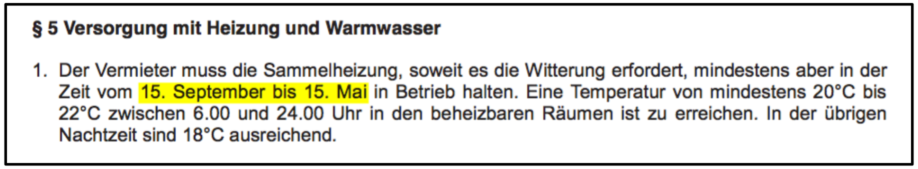
Monetary Values, Dates and References

Monetary Values

- Absolute: 1.234 Euro;
- Relative: „50 % der Miete“;

Dates

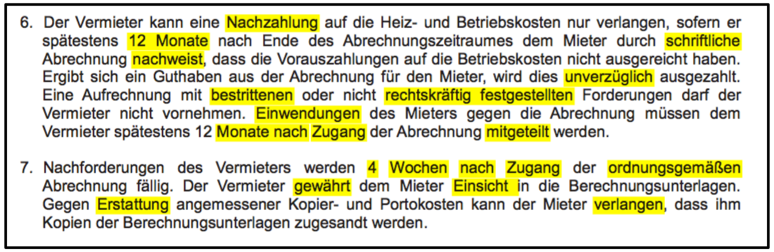
- Absolute:
  - „15. September bis 15. Mai“



§ 5 Versorgung mit Heizung und Warmwasser

1. Der Vermieter muss die Sammelheizung, soweit es die Witterung erfordert, mindestens aber in der Zeit vom 15. September bis 15. Mai in Betrieb halten. Eine Temperatur von mindestens 20°C bis 22°C zwischen 6.00 und 24.00 Uhr in den beheizbaren Räumen ist zu erreichen. In der übrigen Nachtzeit sind 18°C ausreichend.

- Relativ:
  - „12 Monate nach Ende des Abrechnungszeitraums“
  - „4 Wochen nach XX“
  - „3 Monate vor Beginn der Bauarbeiten“

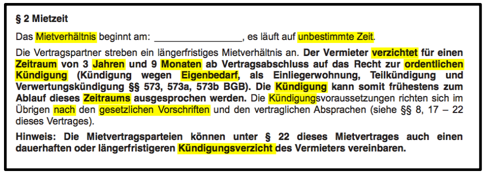


6. Der Vermieter kann eine Nachzahlung auf die Heiz- und Betriebskosten nur verlangen, sofern er spätestens 12 Monate nach Ende des Abrechnungszeitraumes dem Mieter durch schriftliche Abrechnung nachweist, dass die Vorauszahlungen auf die Betriebskosten nicht ausgereicht haben. Ergibt sich ein Guthaben aus der Abrechnung für den Mieter, wird dies unverzüglich ausgezahlt. Eine Aufrechnung mit bestrittenen oder nicht rechtskräftig festgestellten Forderungen darf der Vermieter nicht vornehmen. Einwendungen des Mieters gegen die Abrechnung müssen dem Vermieter spätestens 12 Monate nach Zugang der Abrechnung mitgeteilt werden.

7. Nachforderungen des Vermieters werden 4 Wochen nach Zugang der ordnungsgemäßen Abrechnung fällig. Der Vermieter gewährt dem Mieter Einsicht in die Berechnungsunterlagen. Gegen Erstattung angemessener Kopier- und Portokosten kann der Mieter verlangen, dass ihm Kopien der Berechnungsunterlagen zugesandt werden.

References

- „Teilkündigung und Verwertungskündigung §§ 573, 573a, 573b BGB"



§ 2 Mietzeit

Das Mietverhältnis beginnt am: ________________, es läuft auf unbestimmte Zeit.

Die Vertragspartner streben ein längerfristiges Mietverhältnis an. **Der Vermieter verzichtet für einen Zeitraum von 3 Jahren und 9 Monaten ab Vertragsabschluss auf das Recht zur ordentlichen Kündigung (Kündigung wegen Eigenbedarf, als Einliegerwohnung, Teilkündigung und Verwertungskündigung §§ 573, 573a, 573b BGB). Die Kündigung kann somit frühestens zum Ablauf dieses Zeitraums ausgesprochen werden.** Die Kündigungsvoraussetzungen richten sich im Übrigen nach den gesetzlichen Vorschriften und den vertraglichen Absprachen (siehe §§ 8, 17 – 22 dieses Vertrages).

Hinweis: Die Mietvertragsparteien können unter § 22 dieses Mietvertrages auch einen dauerhaften oder längerfristigeren Kündigungsverzicht des Vermieters vereinbaren.

Named Entities

Persons

- „Herrn Martin Rollinger“
- „Prof. Dr. Florian Matthes“

Organisations / Institutions

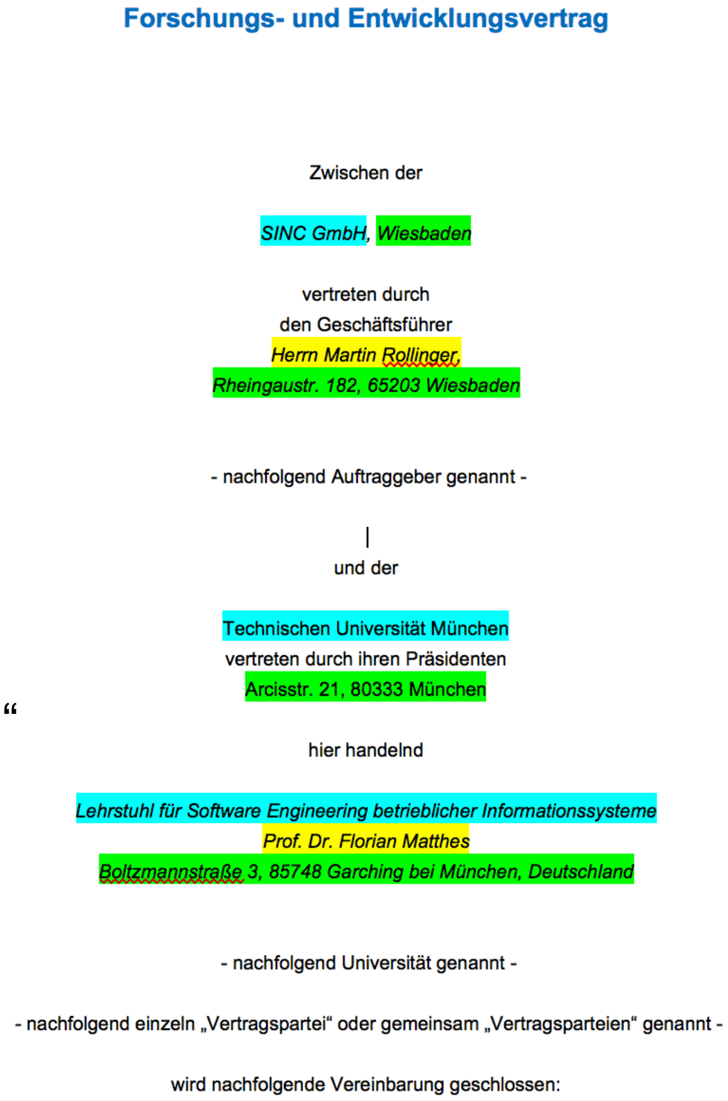
- „Technische Universität München“
- „SINC GmbH“

Locations / Geographic Information

- „Rheingaustr. 182, 65203 Wiesbaden“
- „Boltzmannstraße 3
  85748 Garching bei München, Deutschland“

Roles

- “Principal“ / „Contractor“



Forschungs- und Entwicklungsvertrag

Zwischen der

SINC GmbH, *Wiesbaden*

vertreten durch
den Geschäftsführer
Herrn Martin Rollinger,
Rheingaustr. 182, 65203 Wiesbaden

- nachfolgend Auftraggeber genannt -

und der

Technischen Universität München
vertreten durch ihren Präsidenten
Arcisstr. 21, 80333 München

hier handelnd

Lehrstuhl für Software Engineering betrieblicher Informationssysteme
Prof. Dr. Florian Matthes
Boltzmannstraße 3, 85748 Garching bei München, Deutschland

- nachfolgend Universität genannt -

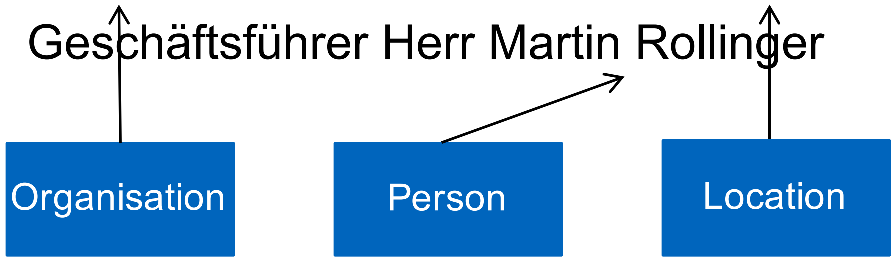
- nachfolgend einzeln „Vertragspartei“ oder gemeinsam „Vertragsparteien“ genannt -

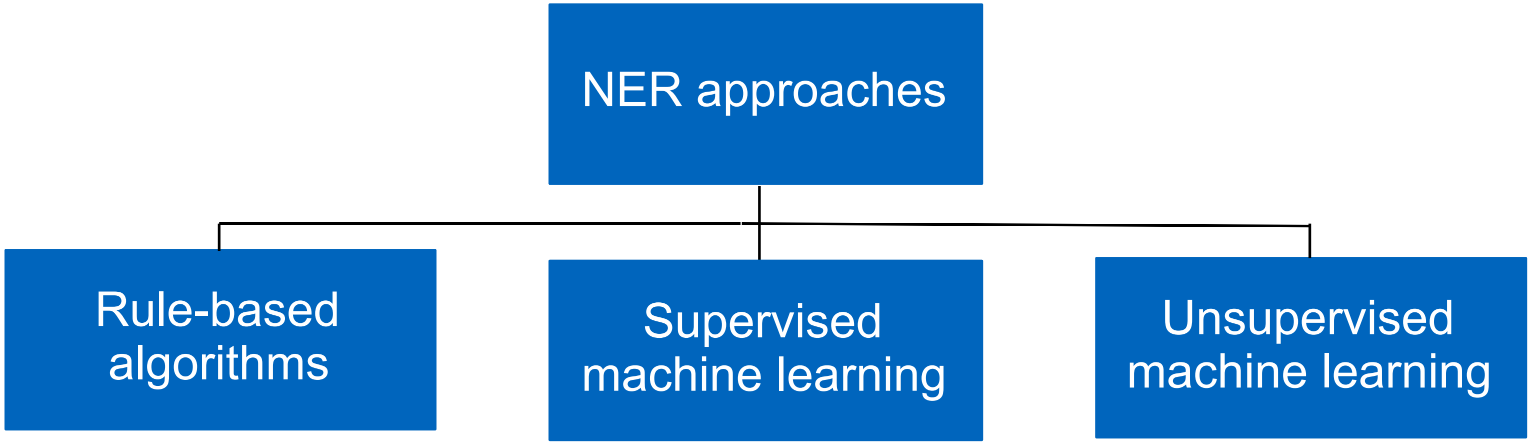
wird nachfolgende Vereinbarung geschlossen:

Named Entity Recognition

Named Entity Recognition is the detection and classification task of proper names in continuous text [Benikova]:

- A named entity is a phrase that contains the names of persons, organizations, locations, etc.
- “Die SINC Gmbh mit Sitz in Wiesbaden wird vertretendurch Ihren Geschäftsführer Herr Martin Rollinger

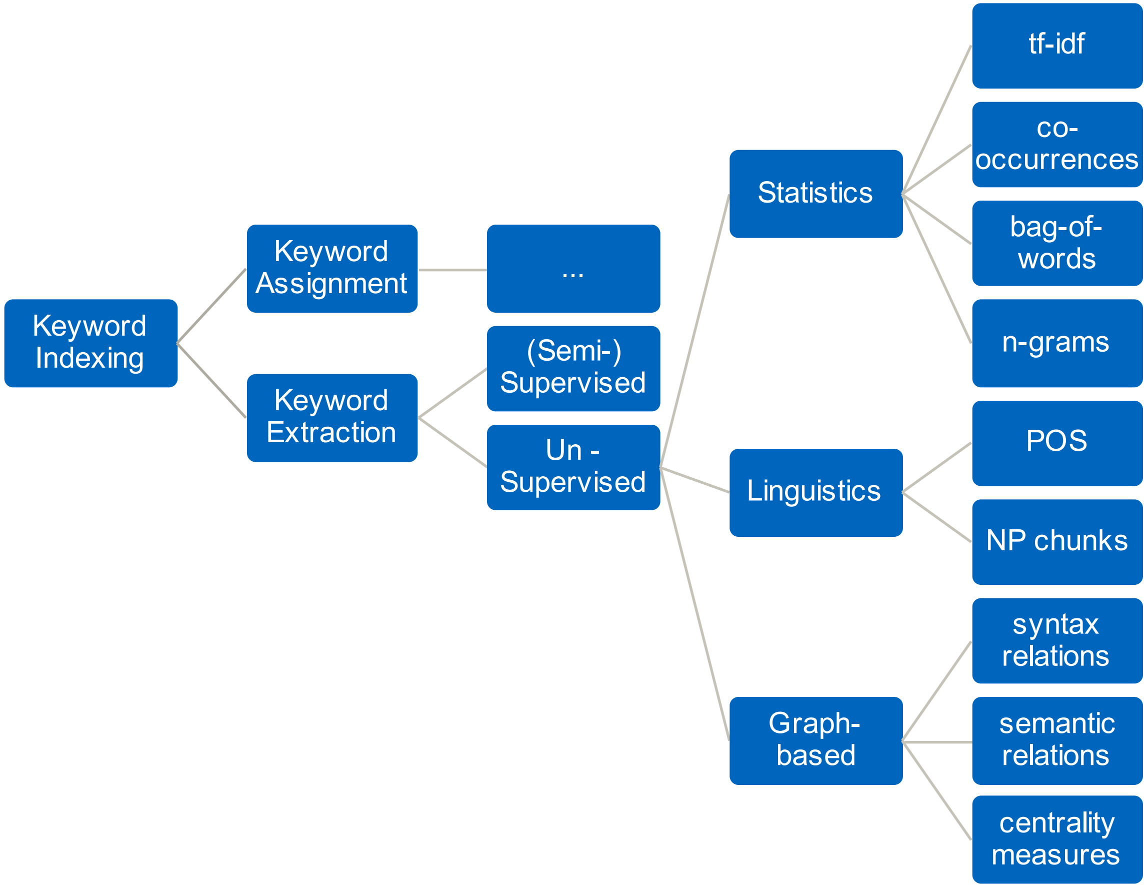




Keyword Extraction

Level

- Document
- Article
- Sentence



Disambiguation (I)

- **Resolution of the role of a named entity**
  - e.g. Employment agreement
    - Technische Universität München ☾ Institution ☾ Employer
    - Bernhard Waltl ☾ Person ☾ Employee

NE Recognition

NE Disambiguation

Text	Named Entity	Role
Technische Universität München	Institution	Employer
Bernhard Waltl	Person	Employee
3 Monate	Time duration	Cancelation period
30. April 2018	Date	End of contract
...	...	...

Disambiguation(II)

- Domain model is required
  - Types with attributes (perhaps relations)
  - Context (Sentence, Clause, Document)

1. Assignment of NE by means of rules
   - Apache Ruta
   - RegEx
   - “Bootstrap“ of ML-Approaches (labelled data set)

2. Assignment of NE through heuristic approaches
   - Active Machine Learning
   - Apache Spark
     - Naive Bayes

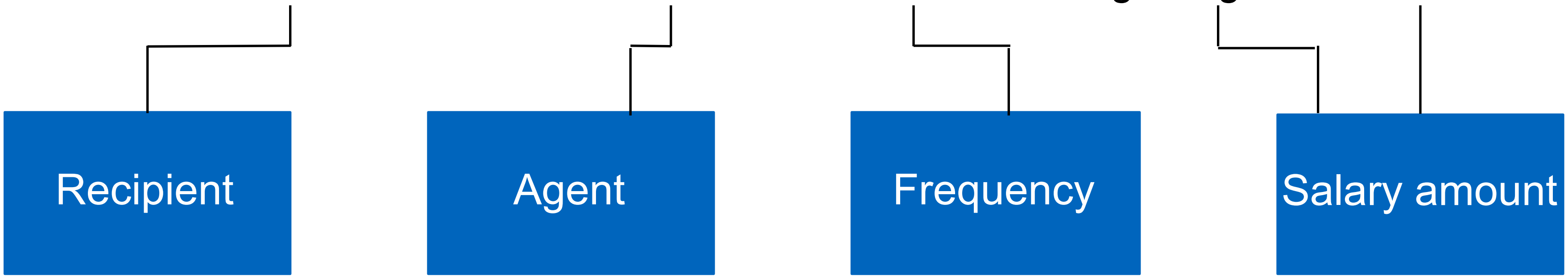
Working	
Employer	*Institution*
Employee	*Person*
Cancelation period	*Time duration*
End of contract	*Date*
Prohibition of competition	*Boolean*
....	....

Semantic Role Labeling

Semantic role labeling is a task consisting of the detection of semantic arguments associated with the predicate or verb of a sentence and their classification into their specific roles [Palmer]:

- This helps to understand the meaning of sentence
- That knowledge can be used to obtain the meaning of whole documents
- The recognized entities or tags from previous phases is linked to semantic functions

- *Herr Waltl erhält von der TUM eine monatliche Vergütung von 4000 Euro*



Recipient

Agent

Frequency

Salary amount

Solution

- Output in Lexia
  - Overview of contract with all relevant information
  - Search features
    - Full-text search
    - Narrowing done the results by specific criteria based on the learned semantic
  - Additional use cases
- Reusability as a crucial requirement

Outline

1. Motivation
2. Administrative Setup
3. Problem Statement
4. Research Approach
5. Solution
6. Research Questions and Challenges
7. Thesis Timeline

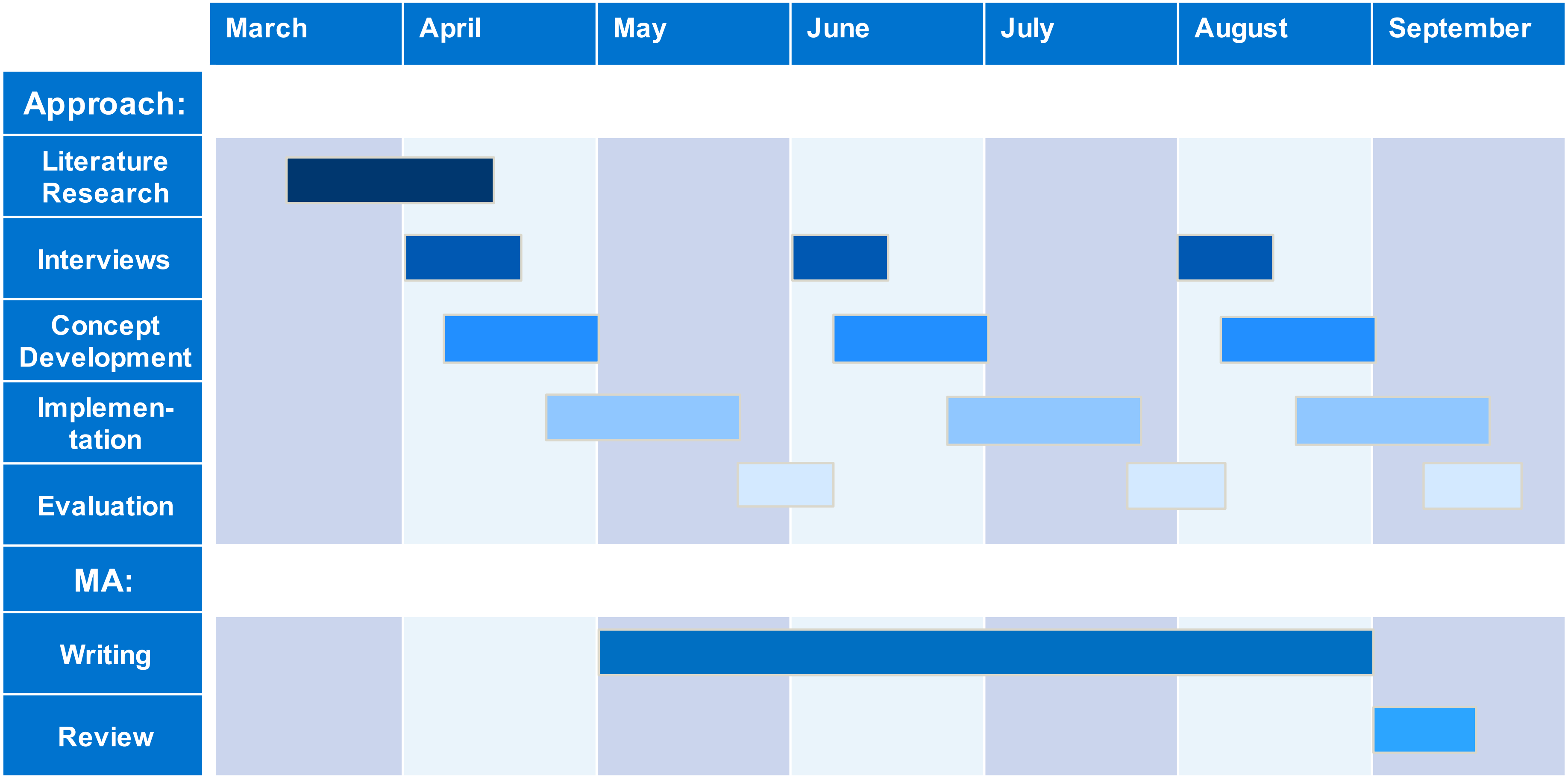
Research Questions

- Which **information** does a stakeholder want to extract from contracts?
- What are the **functional** and **non-functional requirements (evaluation criteria)** of a software for the analysis of legal contracts?
- How does a **prototypical implementation** enabling the semantic analysis of contracts look like?
- Which **NLP technologies** can be used, to extract the semantic meaning of a contract? How to combine these technologies into a **Apache UIMA pipeline**?
- How can such a system be **integrated** into the workflow of potential stakeholders?

Outline

1. Motivation
2. Administrative Setup
3. Problem Statement
4. Research Approach
5. Solution
6. Research Questions and Challenges
7. Thesis Timeline

Thesis Timeline



March April May June July August September

Approach:

Literature Research

Interviews

Concept Development

Implementation

Evaluation

MA:

Writing

Review

Thank you for your attention!

- Suggestions?
- Questions?
- Remarks?

References (I)

[Aizawa]
A. Aizawa, "An information-theoretic perspective of tf–idf measures," Information Processing & Management, vol. 39, no. 1, pp. 45-65, 2003.

[Bauer]
L. Bauer, Introducing linguistic morphology, 2nd ed. Washington, D.C.: Georgetown University Press, 2003, pp. x, 366 p.

[Benikova]
D. Benikova, S. Muhie, Y. Prabhakaran, and S. C. Biemann, "C.: GermaNER: Free Open German Named Entity Recognition Tool," in In: Proc. GSCL-2015, 2015: Citeseer.

[BCB]
The Boston Consulting Group: How Legal Technology Will Change The Business of Law, 2016, Bucerius Law School

[Beeferman]
D. Beeferman, A. Berger, and J. Lafferty, "Statistical Models for Text Segmentation," Machine Learning, journal article vol. 34, no. 1, pp. 177-210, 1999.

[Choi]
F. Y. Y. Choi, "Advances in domain independent linear text segmentation," presented at the Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference, Seattle, Washington, 2000.

[Deloitte]
Deloitte: Digitisation of Documents and Legal Archiving, 2014

[Habash]
N. Habash, O. Rambow, and R. Roth, "MADA+ TOKAN: A toolkit for Arabic tokenization, diacritization, morphological disambiguation, POS tagging, stemming and lemmatization," in Proceedings of the 2nd international conference on Arabic language resources and tools (MEDAR), Cairo, Egypt, 2009, vol. 41, p. 62.

[Hashmi]
Mustafa Hashmi: A Methodology for Extracting Legal Norms from Regulatory Documents, 2015, IEEE 19th International Enterprise Distributed Object Computing Workshop

References (II)

[Hakimov]
S. Hakimov, S. A. Oto & E. Dogdu: Named Entity Recognition and Disambiguation using Linked Data and Graph-based Centrality Scoring, 2012, SIGMOD

[Lavrac]
N. Lavrac, D. Mladenic, and T. Erjavec, "Ripple Down Rule learning for utomated word lemmatisation," AI Communications, vol. 21, no. 1, pp. 15-26, 2008.

[Rajamaran]
A. U. Rajaraman, "JD (2011)." Data Mining," Mining of Massive Datasets, pp. 1-17

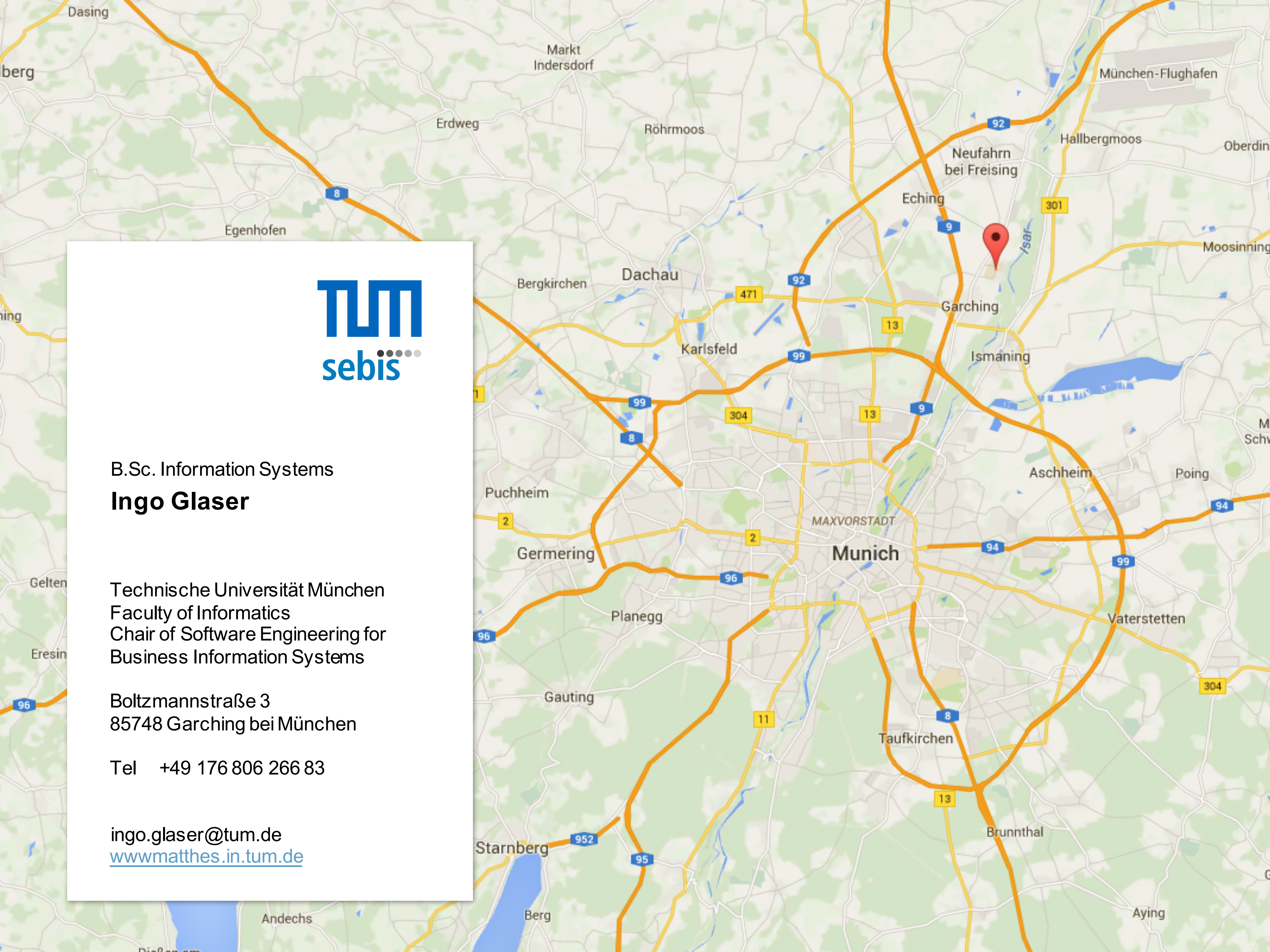
[Saravanan]
M. Saravanan, B. Ravindran & S. Raman: Improving Legal Information Retrival Using an Ontological Framework, 2009, Artificial Intelligence and Law

[Singh]
J. Singh and V. Gupta, "Text Stemming: Approaches, Applications, and Challenges," ACM Comput. Surv., vol. 49, no. 3, pp.

[Stevenson]
M. Stevenson and Y. Wilks, "Word sense disambiguation," The Oxford Handbook of Comp. Linguistics, pp. 249-265, 2003.

[Svyatkovskiy]
A. Svyatkovskiy, K. Imai, M. Kroeger, Y. Shiraito: Large-scale Text Processing Pipeline with Apache Spark, 2016, IEEE International Conference on Big Data

[Waltl]
B. Waltl et al: Automated Extraction of Semantic Information from German Legal Documents, 2017, Internationales Rechtsinformatik Symposium, Salzburg



B.Sc. Information Systems

Ingo Glaser

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

Boltzmannstraße 3
85748 Garching bei München

Tel +49 176 806 266 83

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