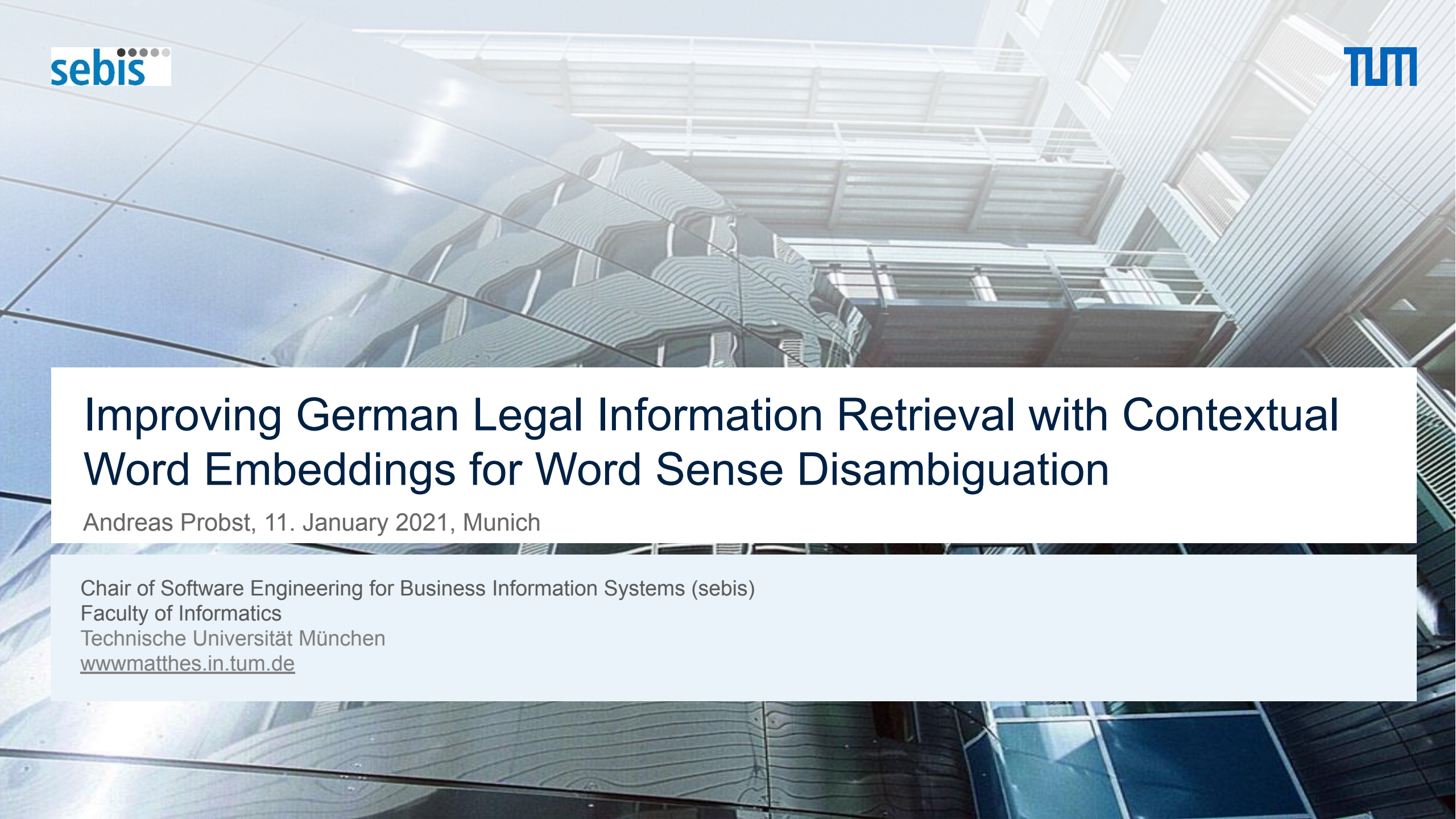


Improving German Legal Information Retrieval with Contextual Word Embeddings for Word Sense Disambiguation

Andreas Probst, 11. January 2021, Munich

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

Outline

Motivation

Thesis Goal

Research Approach

Available Data

Technical Approach

Current Challenges

Research Questions

Timeline

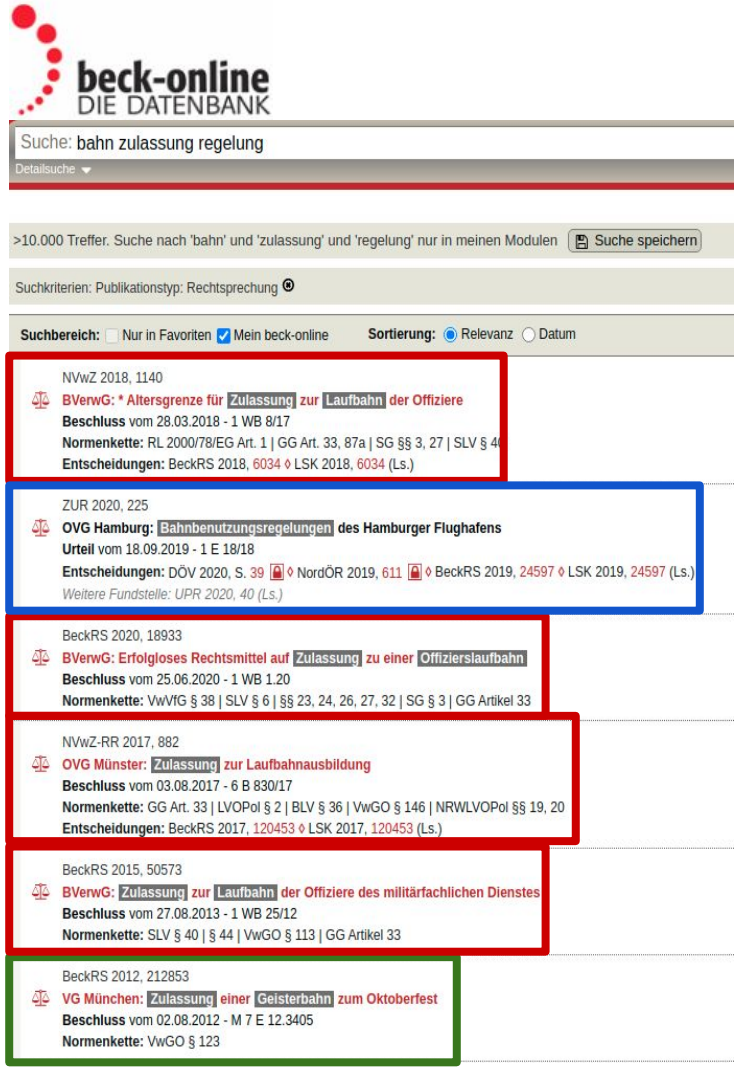
Motivation

Base Hypothesis

Word sense disambiguation (WSD) can be applied on German court rulings to improve legal information retrieval.

Motivation

Motivational Use Case: Search Results of Current Legal Databases Ignore Word Senses



- A legal expert is looking for information regarding the commissioning regulation of rail vehicles and inputs the query “bahn zulassung regelung”.
- He gets results for the word “Bahn” in the sense of career (Karriere), runway (Laufbahn) and rail vehicle (Schienen-fahrzeug).
- WSD could help to omit a lot of undesired search results by removing results related to unwanted word senses.

Motivation

Already Conducted Expert Interviews Indicate That Experts Want to Filter Based on Word Senses

Interviews and Methodology

- 5 lawyers so far
- semi-structured expert interview (ca. 45 minutes over Zoom)
- diverserse group (age, gender, experience, IT-affinity, legal field)
- 2 parts: legal research in general + word senses in law

lead to

Insights

- too many irrelevant search results when using legal databases, even when working with other filters
- hints about possible word senses prefered at query input
- legal terms tend to have only one defined sense
- a lot of ambiguity in natural language terms -> **our use case (!)**

Motivation

In Contrast to Legal Terms, Natural Language Terms Often Tend to Have Different Senses

Legal Terms Within One Legal Domain

⚠ No ambiguous terms

Legal Terms Across Legal Domains

Berufungsgrund

Berufen werden (Erbrecht) — **Berufung einlegen (ZPO)**

⚠ No AI needed

Natural Language Terms Across Legal Domains

Raum

Geographische Region — **Zimmer**

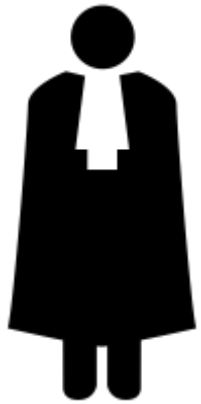
Natural Language Terms Within One Legal Domain

Bahn

Zug — **Schienen**

Motivation

Outlook Use Case: Filtering Based on Senses



Legal Expert

Query contains the word

„Bahn“

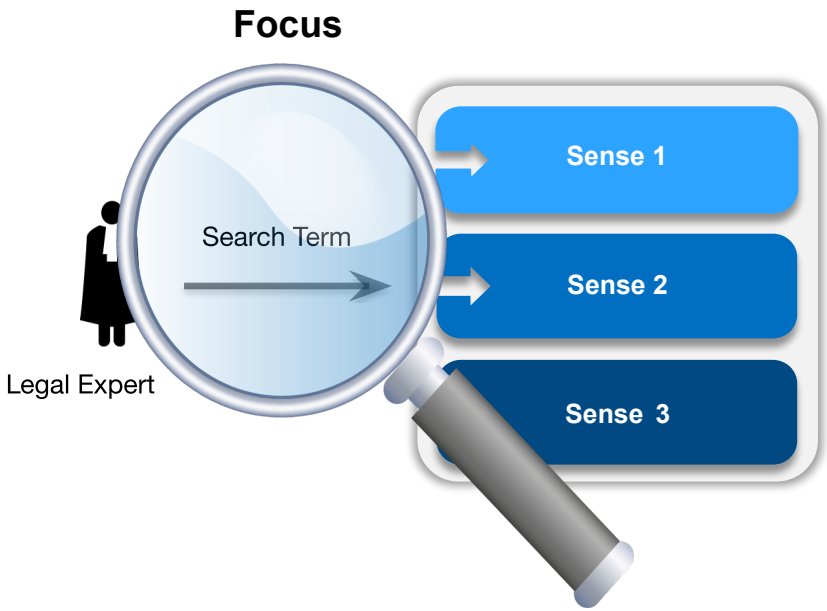
Possible Senses

- Berufslaufbahn, Karriere, Werdegang
- Laufbahn, Startbahn, Landebahn
- Zug, Eisenbahn

Offer the user a filtering option for the search term “Bahn” based on classified senses.

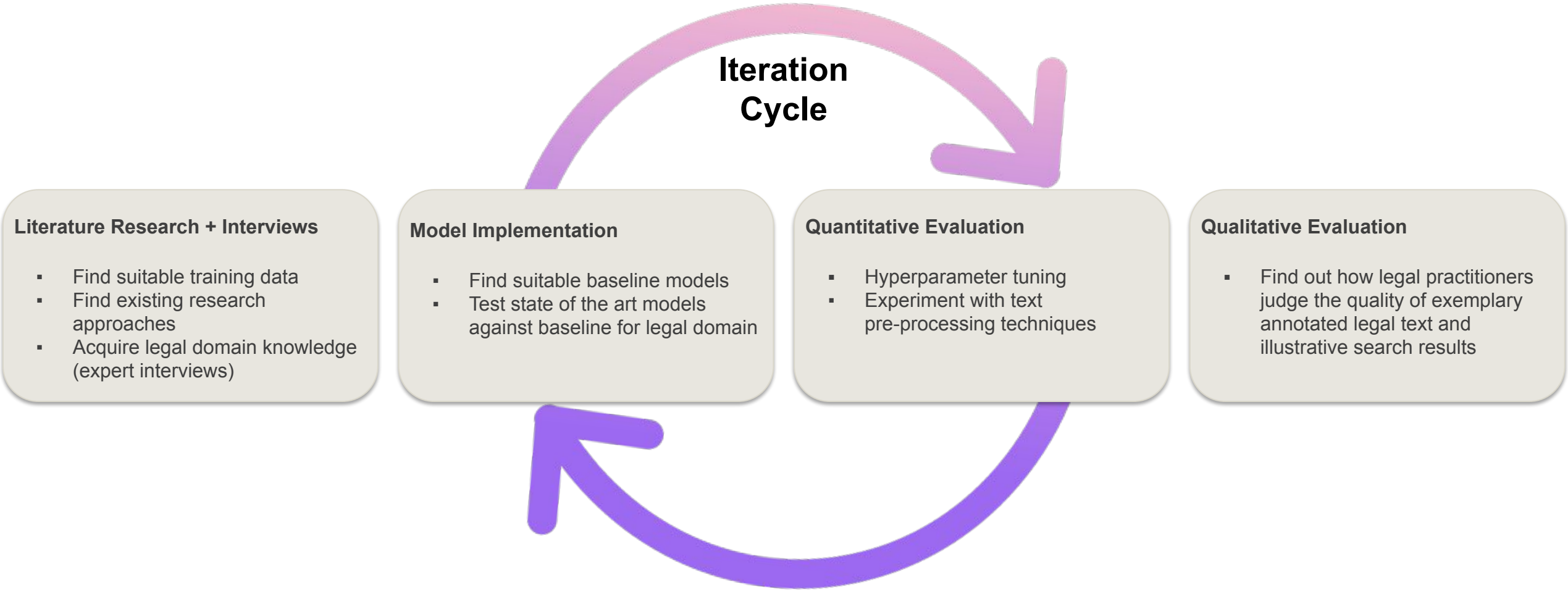
Thesis Goal

Required Intermediate Step for Outlook Use Case: Sense Classification



Automatically classify senses of words in unannotated court rulings.

Research Approach



Research Data

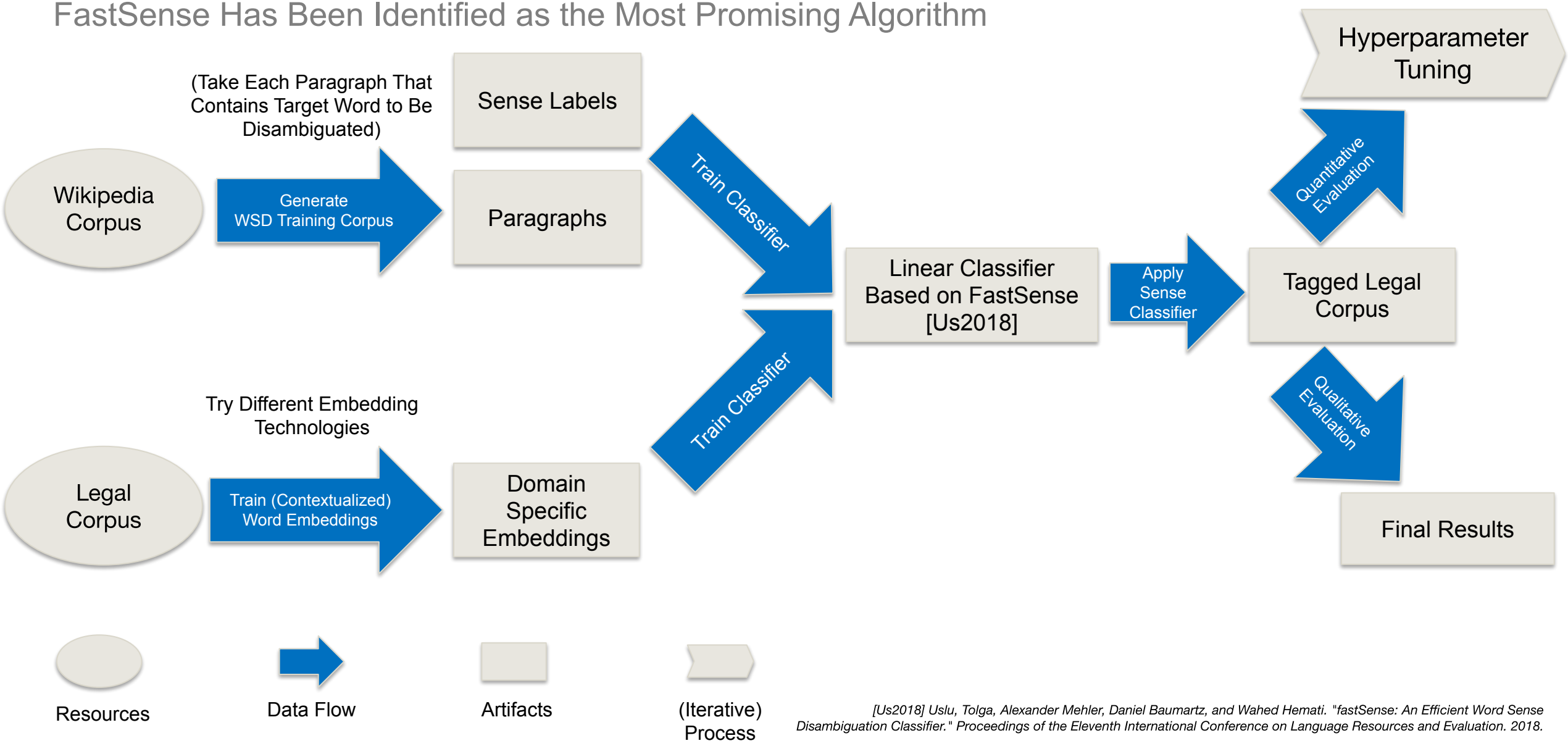
Sufficient Research Data Is Available

Name	Type	Sense Annotated	Domain Specific	For Training Word Embeddings?	Size
GermaNet	lexical graph network	✓	✗	✗	185.000+ lexical units
webCAGe [He2012]	text	✓	✗	✗	9000+ samples
wikipedia + wiktionary dump	text	indirectly	✗	✗	55M+ samples
Court Rulings (Otto Schmidt)	text	✗	✓	✓	70.000+ samples
Court Rulings (Haufe)	text	✗	✓	✓	10+ GB

[He2012] Henrich, Verena & Hinrichs, Erhard & Vodolazova, Tatiana. (2012). "WebCAGe: a web-harvested corpus annotated with GermaNet senses". 387-396.

Technical Approach

FastSense Has Been Identified as the Most Promising Algorithm



[Us2018] Uslu, Tolga, Alexander Mehler, Daniel Baumartz, and Wahed Hemati. "fastSense: An Efficient Word Sense Disambiguation Classifier." Proceedings of the Eleventh International Conference on Language Resources and Evaluation. 2018.

Current Challenges

1

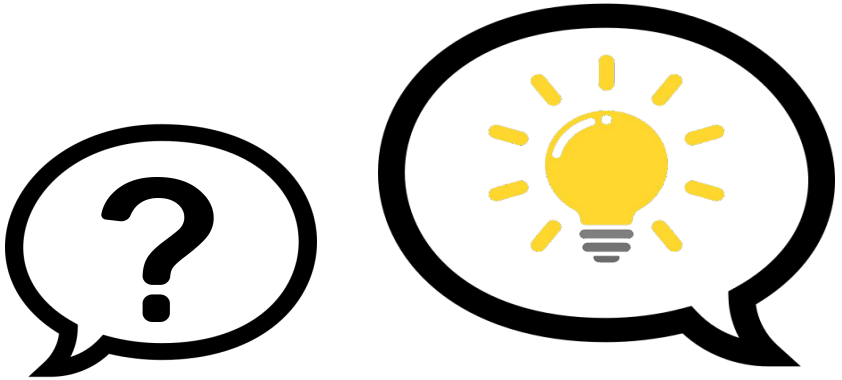
Published WSD research focuses on the English language.

2

Little sense annotated data freely available in German.

3

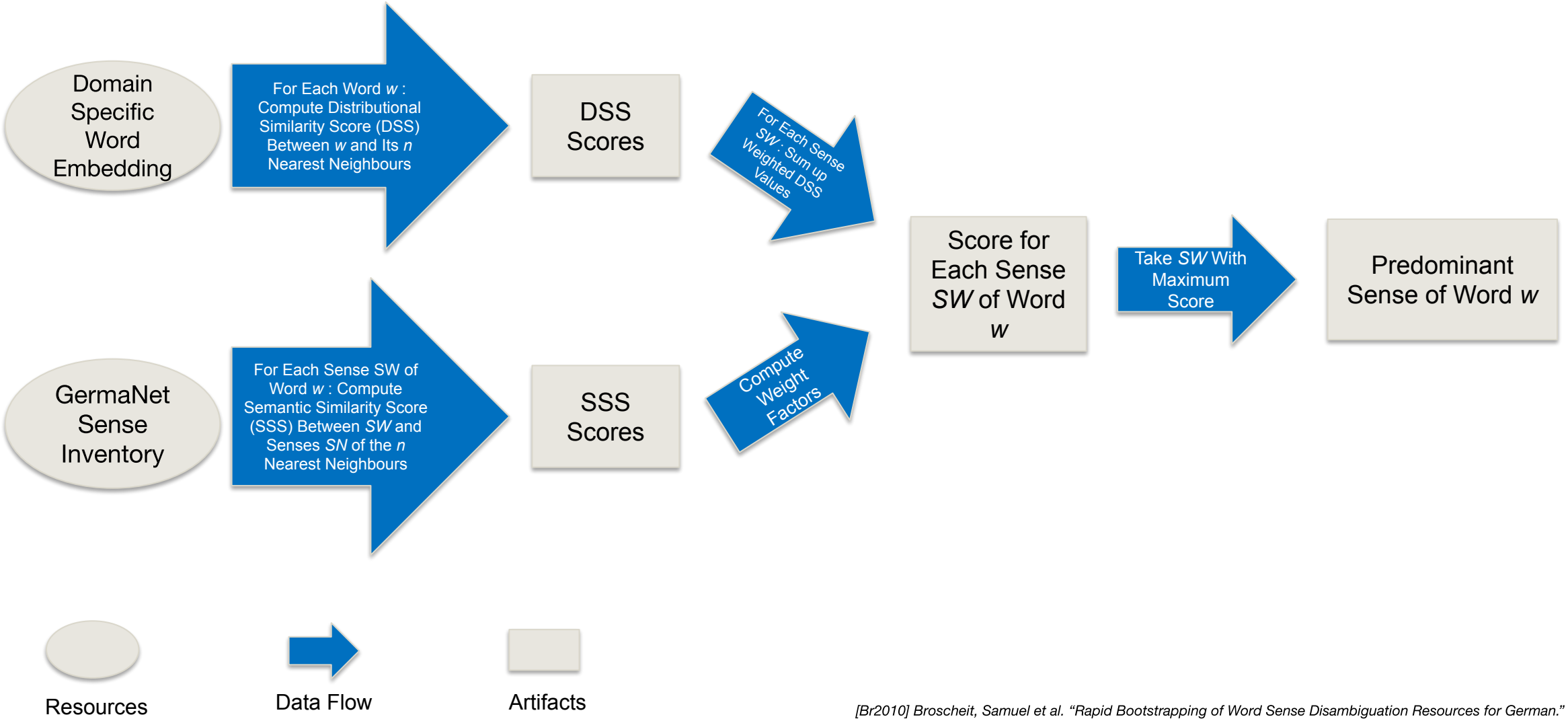
No (predominant) sense information for "most frequent sense" (MFS) baseline approach in German.



Open for suggestions in discussion afterwards!

Current Challenges

[Br2010] Can Be a Potential Solution to Missing “Most Frequent Sense” (MFS) Information

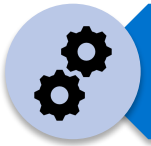


[Br2010] Broscheit, Samuel et al. “Rapid Bootstrapping of Word Sense Disambiguation Resources for German.” KONVENS (2010).

Research Questions

Word sense disambiguation (WSD) can be applied on German court rulings to improve legal information retrieval.

What algorithms exist to classify word senses (supervised, semi-supervised, unsupervised)?

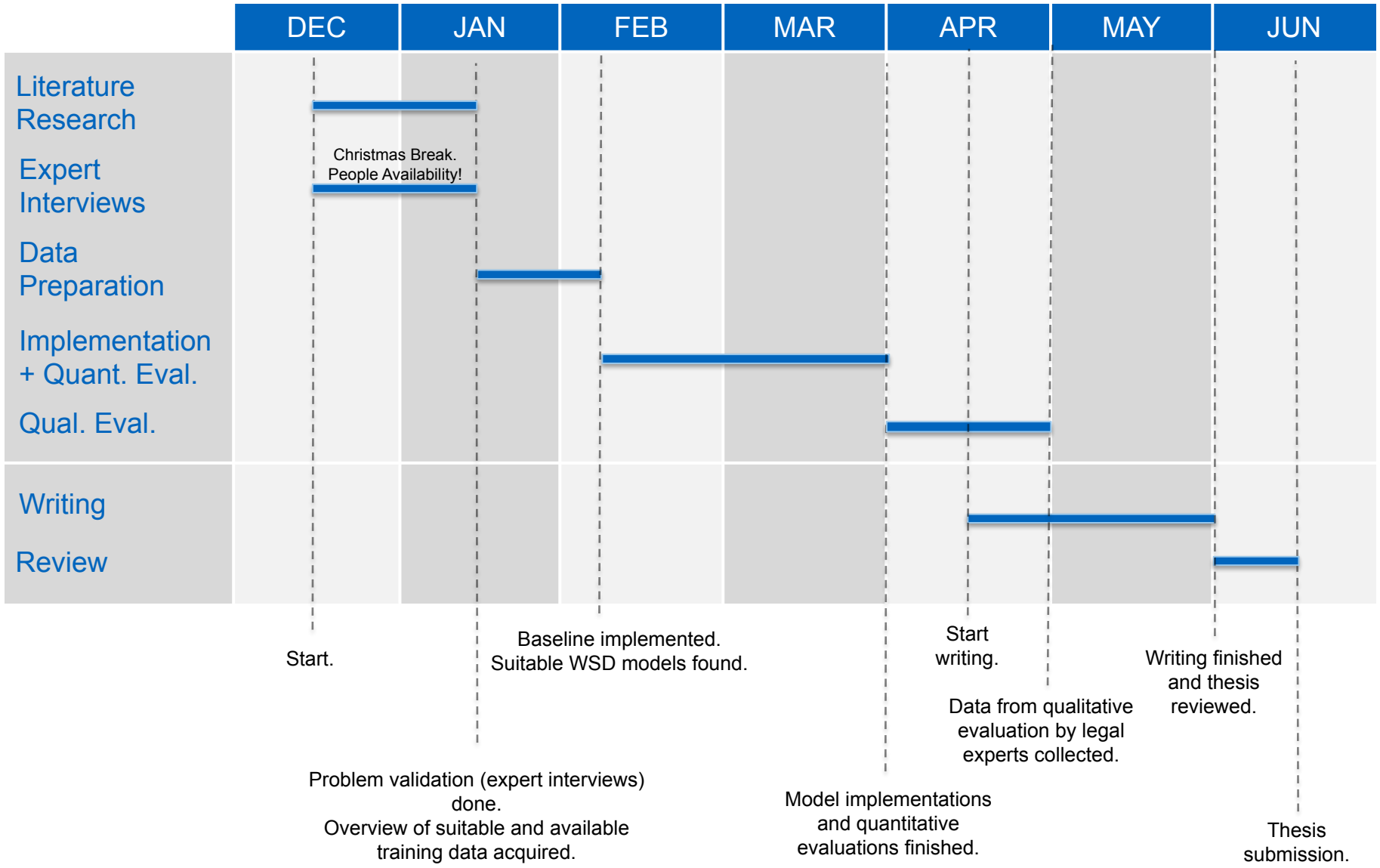


Are there suitable ways to compensate for the lack of german sense annotated data?

Quantitative evaluation: Which WSD algorithms perform best?

Qualitative evaluation: How do legal experts evaluate sense annotated search results?

Timeline (Rough)



Master's Student in Computational Science and Engineering

Andreas Probst

www.linkedin.com/in/andreas-probst

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

andreaspadprobst@gmail.com
wwwmatthes.in.tum.de

Stick to the good sebis traditions

- Provide action links at the bottom of the slide to guide the audience to our web pages or publications (see below). (Select the text, press CTRL-K)

- Use a file name according to our sebis conventions which helps us and our audience to find the file of your presentation on our web site with Google search:
  - YYMMDD Author Short Title
  - Include this string in the footer (Einfügen -> Kopf- und Fusszeile -> Fusszeile)
  - The unusual date format simplifies the search for the latest version of a slide in an alphabetical directory listing (Dropbox, Explorer, Tricia, Sky-Drive)

[Ha13g]Hauder, M., Roth, S., Matthes, F.: Current Tool support for Metrics in Enterprise Architecture Management

For more information visit BEAMS , EAM Pattern Catalog and EAM KPI Catalog (http://wwwmatthes.in.tum.de)

Use the sebis visual language (shapes, fonts, colors, sizes)