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

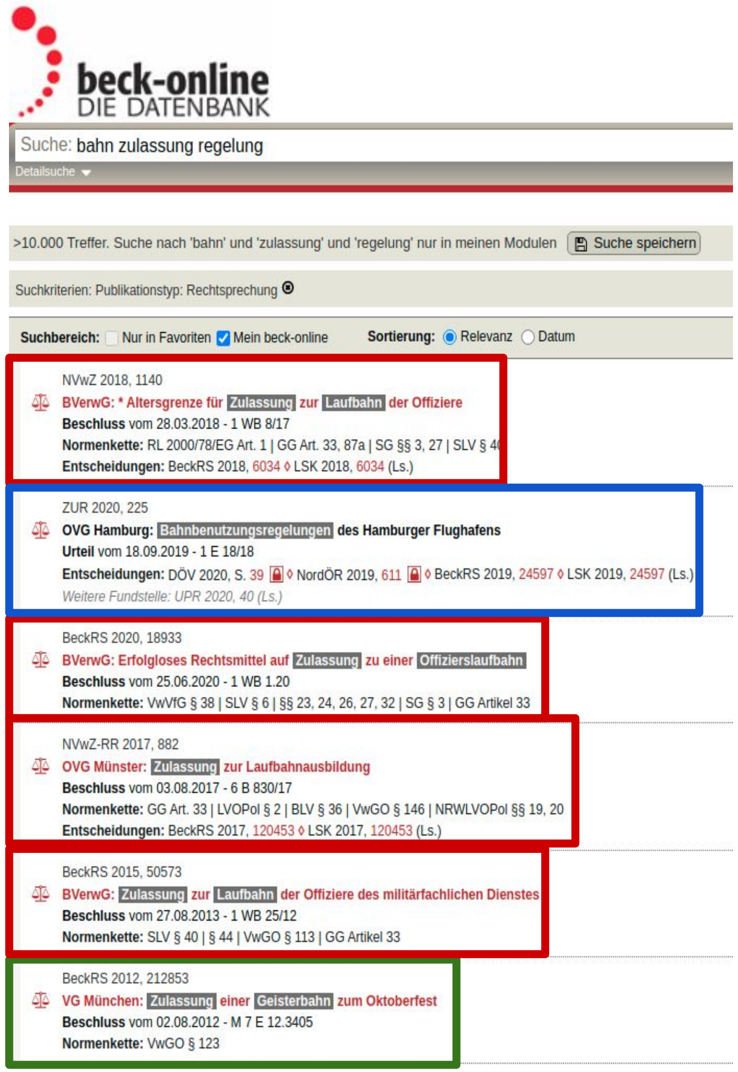
Andreas Probst, 26. July 2021, Munich

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

Outline

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

Use Case: Filter Search Results by Senses

Legal Expert

Query contains the word

„Bahn“

Selection of Intended Sense

Bahn Zulassung Regelung

Berufslaufbahn (Sense 1)

Laufbahn (Sense 2)

Schienenfahrzeug (Sense 3)

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

Motivation

Base Hypothesis

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

Motivation

Summary of Interview Based Pre-study Results with 6 Lawyers

- German legal databases retrieve too many irrelevant search results.
- Legal terms tend to have only one defined sense within one law domain.
- Lots of ambiguity in natural language terms in legal texts (**our use case**).

Outline

Challenges[1] : Word Senses in Legal Language and Natural Language

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

Challenges(2) : Available Resources for the German Language

Legal corpus by publisher Dr. Otto Schmidt

- 56,606 Court Rulings
- XML format
- from different German courts and court types

Dumps of German Wikipedia

- word senses from disambiguation pages
- 303,643 disambiguation pages

For more information on Wikipedia's disambiguation pages visit https://de.wikipedia.org/wiki/Wikipedia:Begriffsklärung (accessed May 20, 2021)

Challenges(2) : Available Resources for the German Language

Legal corpus by publisher Dr. Otto Schmidt

- 56,606 Court Rulings
- XML format
- from different German courts and court types

Dumps of German Wikipedia

- word senses from disambiguation pages
- 303,643 disambiguation page

Hardly any data available to train a German WSD Classifier.
No sense annotated legal data available at all for the German Language.

For more information on Wikipedia's disambiguation pages visit https://de.wikipedia.org/wiki/Wikipedia:Begriffsklärung (accessed May 20, 2021)

Outline

Research Questions

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

RQ1: What algorithms already exist to automatically classify word senses (knowledge-based, unsupervised, and supervised)?

RQ2: What possibilities are there to compensate for the lack of German sense annotated (legal) data?

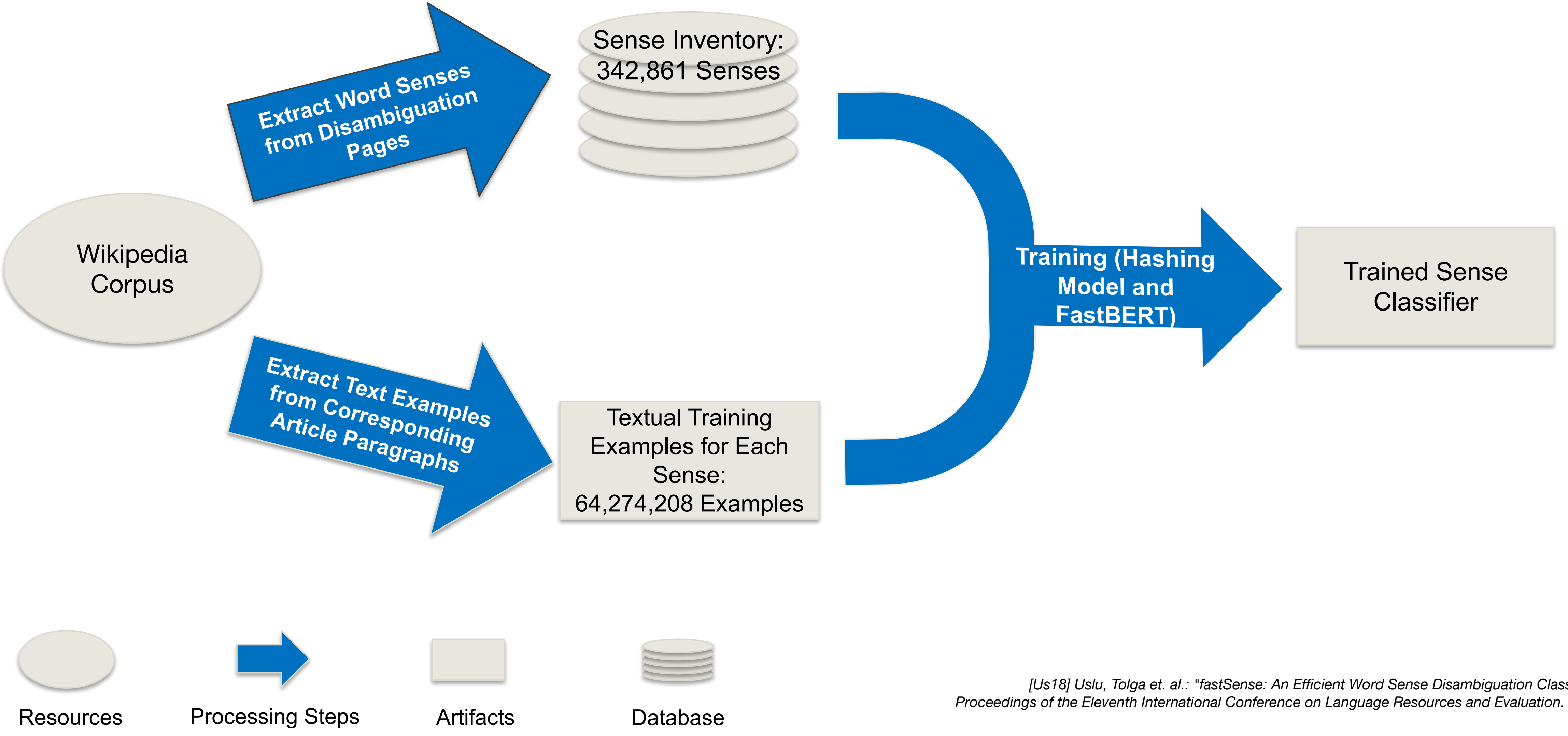
RQ3: Quantitative evaluation: How do WSD algorithms perform on German (legal) text?

RQ4: Qualitative evaluation: How do legal experts judge the usefulness of our word sense filter?

Outline

Technical Approach

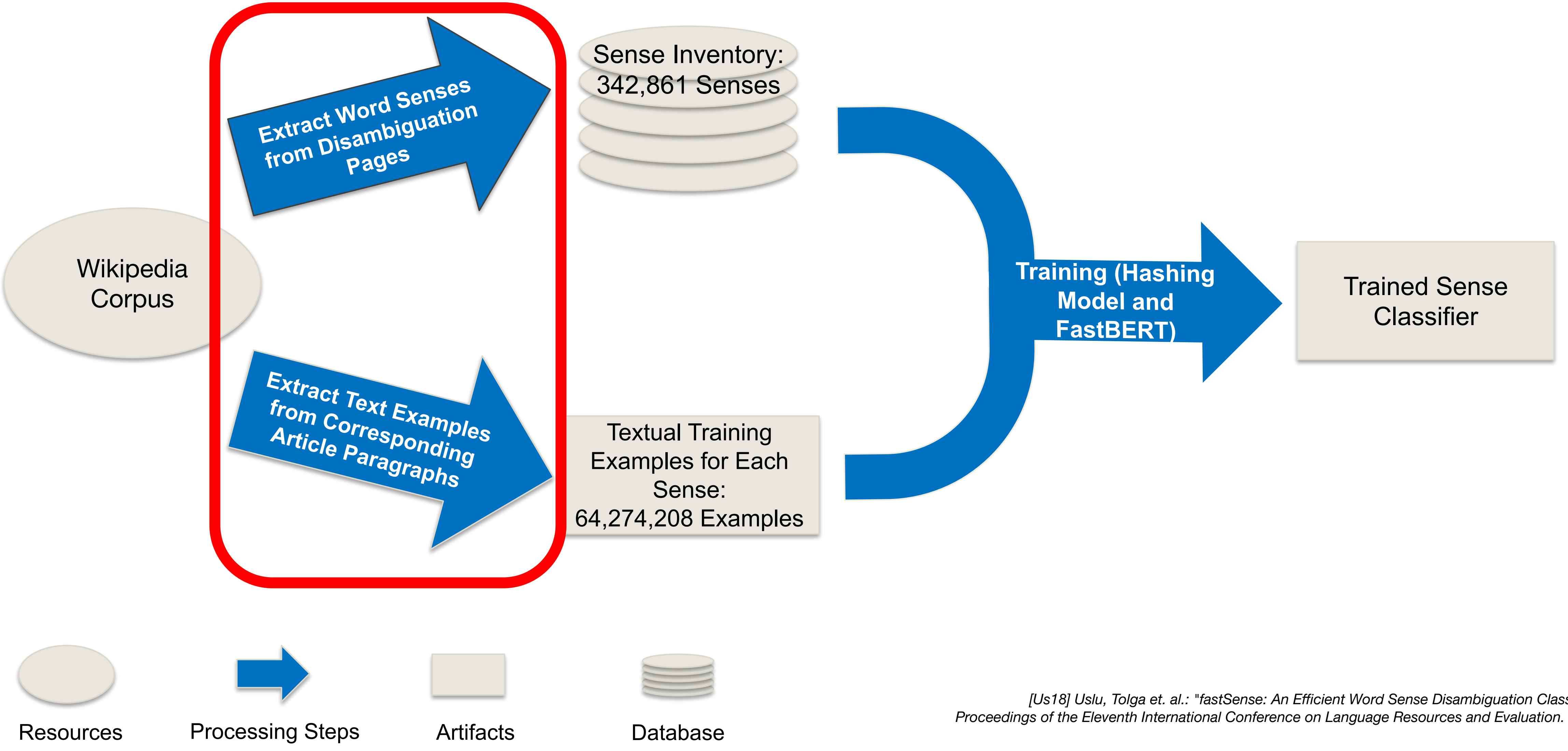
Training Data Generation and Model Training Through Neural-Net Based Algorithms [Us18]



[Us18] Uslu, Tolga et. al.: "fastSense: An Efficient Word Sense Disambiguation Classifier." Proceedings of the Eleventh International Conference on Language Resources and Evaluation. 2018.

Technical Approach

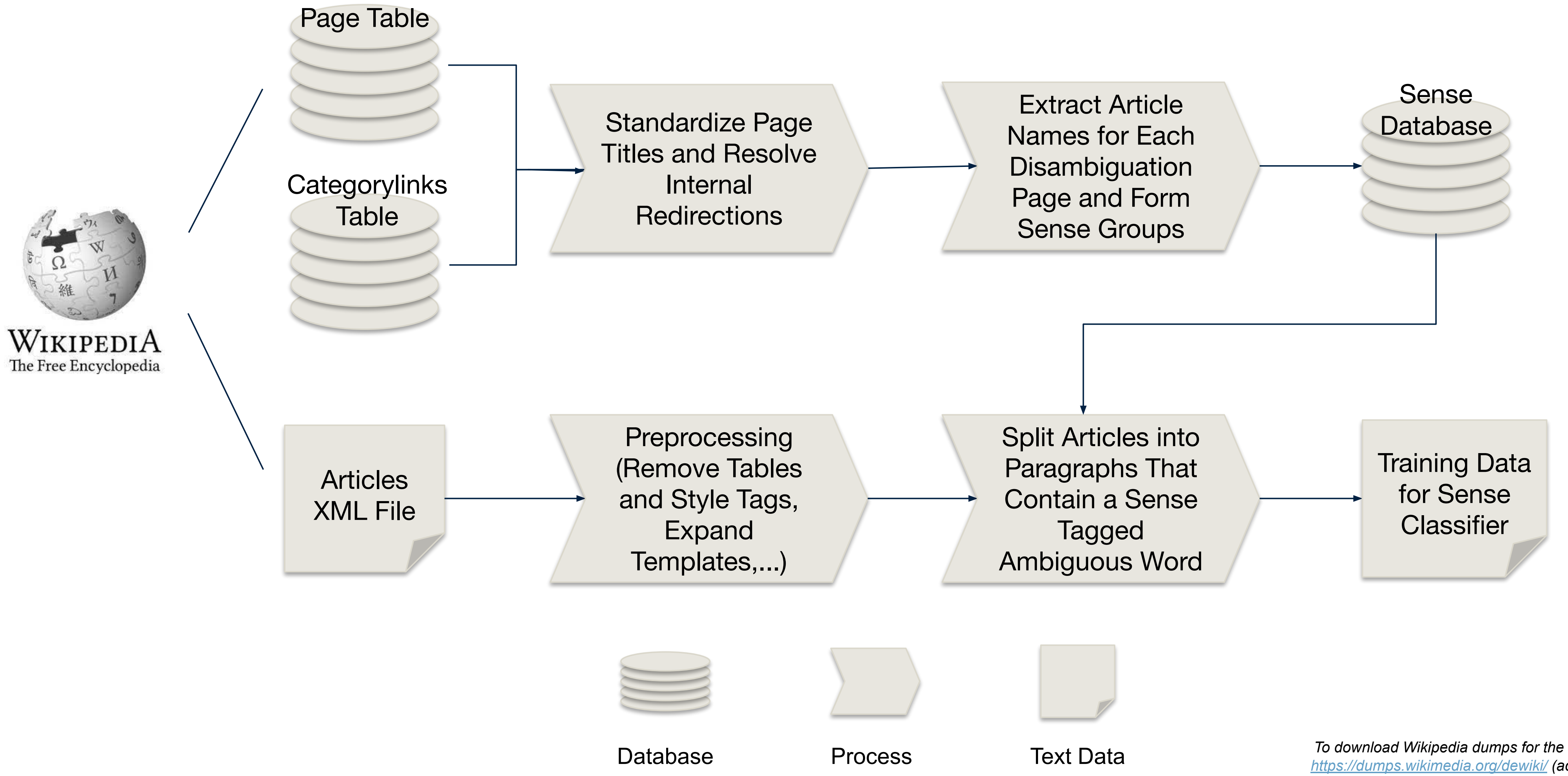
Training Data Generation and Model Training Through Neural-Net Based Algorithms [Us18]



[Us18] Uslu, Tolga et. al.: "fastSense: An Efficient Word Sense Disambiguation Classifier." Proceedings of the Eleventh International Conference on Language Resources and Evaluation. 2018.

Technical Approach

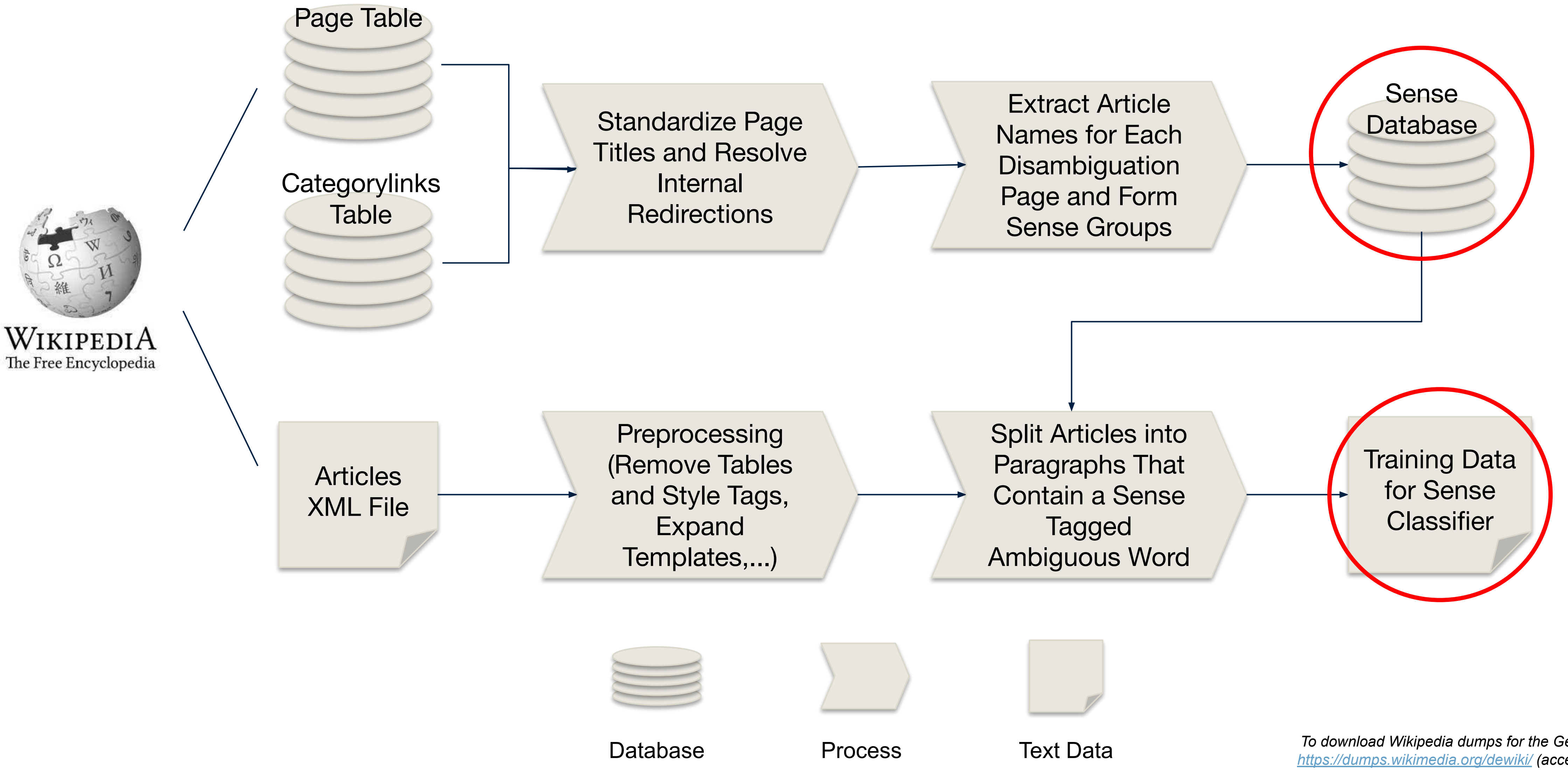
Extracting Training Data for WSD Classifier from Wikipedia Dumps



To download Wikipedia dumps for the German language visti https://dumps.wikimedia.org/dewiki/ (accessed July 24, 2021)

Technical Approach

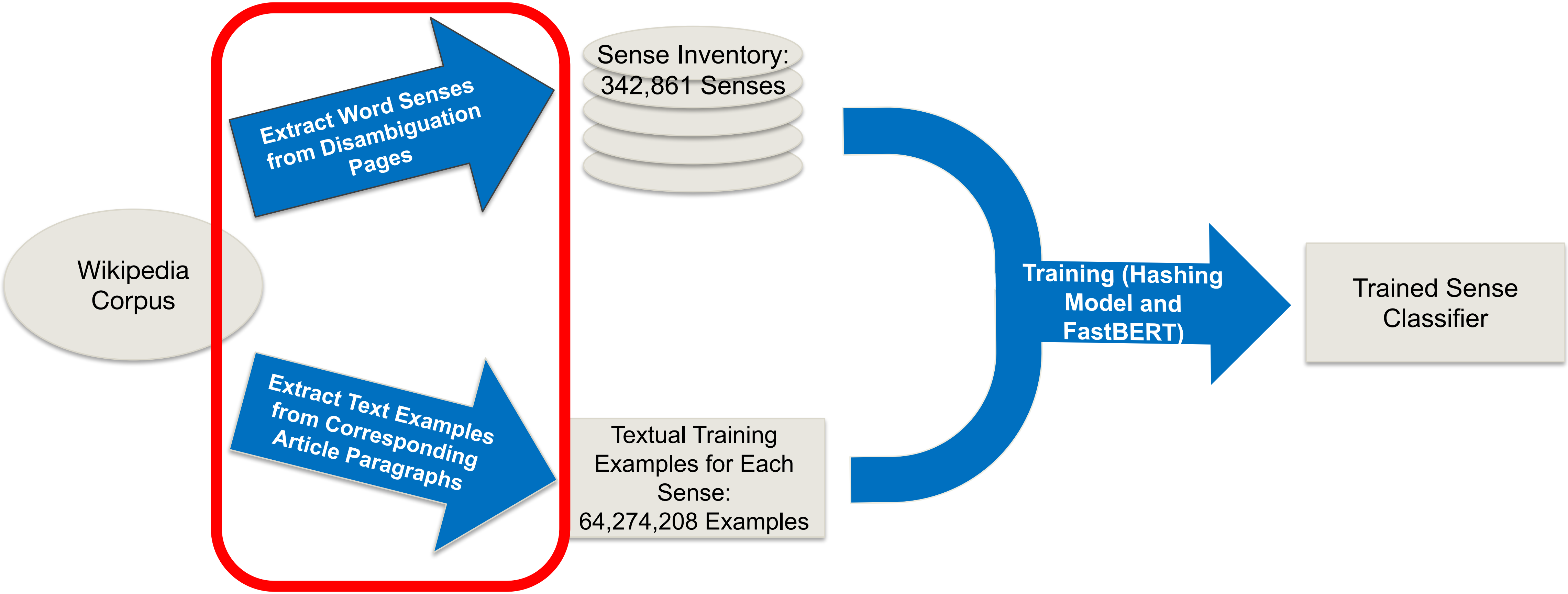
Extracting Training Data for WSD Classifier from Wikipedia Dumps



To download Wikipedia dumps for the German language visti https://dumps.wikimedia.org/dewiki/ (accessed July 24, 2021)

Technical Approach

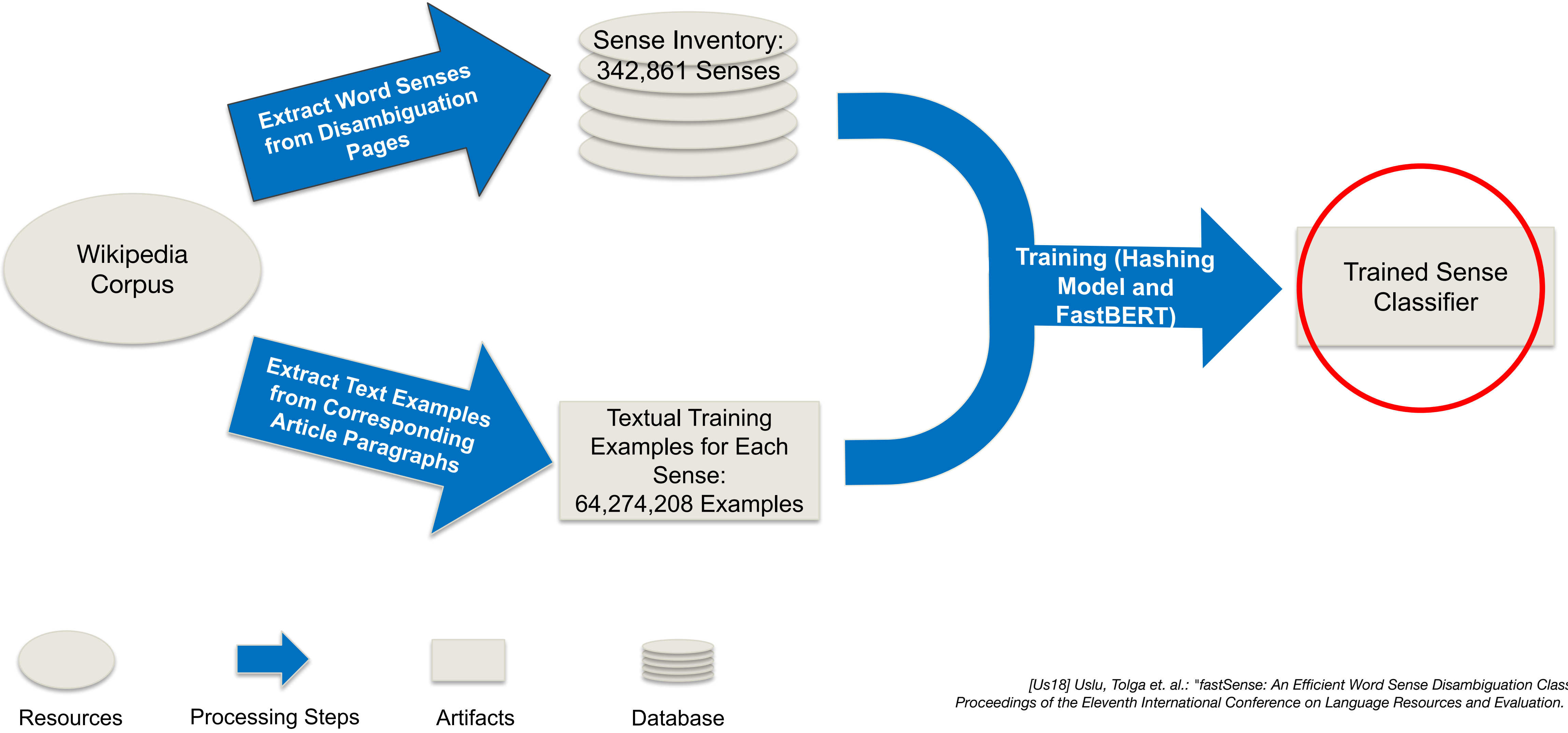
Training Data Generation and Model Training Through Neural-Net Based Algorithms [Us18]



[Us18] Uslu, Tolga et. al.: "fastSense: An Efficient Word Sense Disambiguation Classifier." Proceedings of the Eleventh International Conference on Language Resources and Evaluation. 2018.

Technical Approach

Training Data Generation and Model Training Through Neural-Net Based Algorithms [Us18]

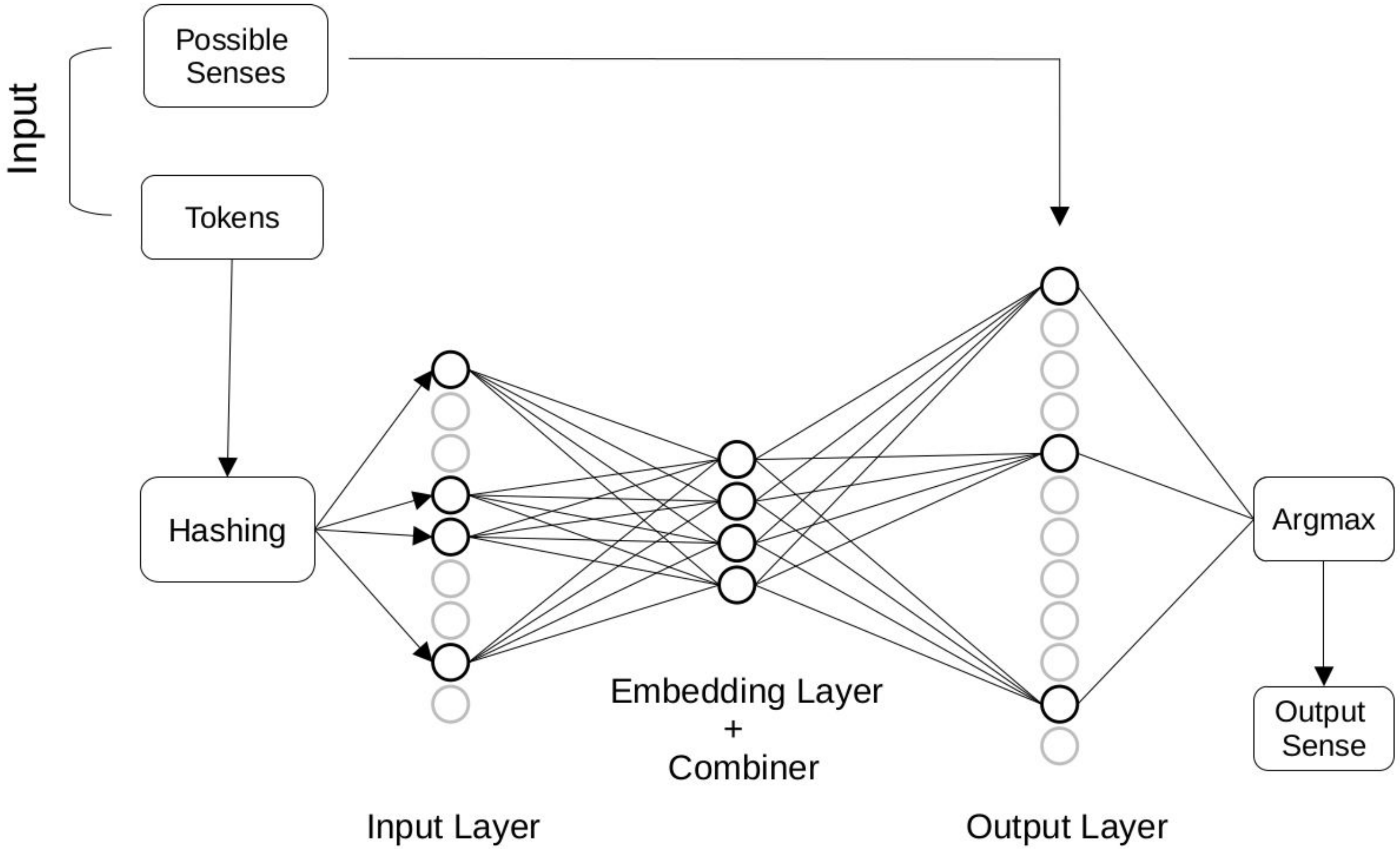


[Us18] Uslu, Tolga et. al.: "fastSense: An Efficient Word Sense Disambiguation Classifier." Proceedings of the Eleventh International Conference on Language Resources and Evaluation. 2018.

Outline

Model Architecture (1st of two)

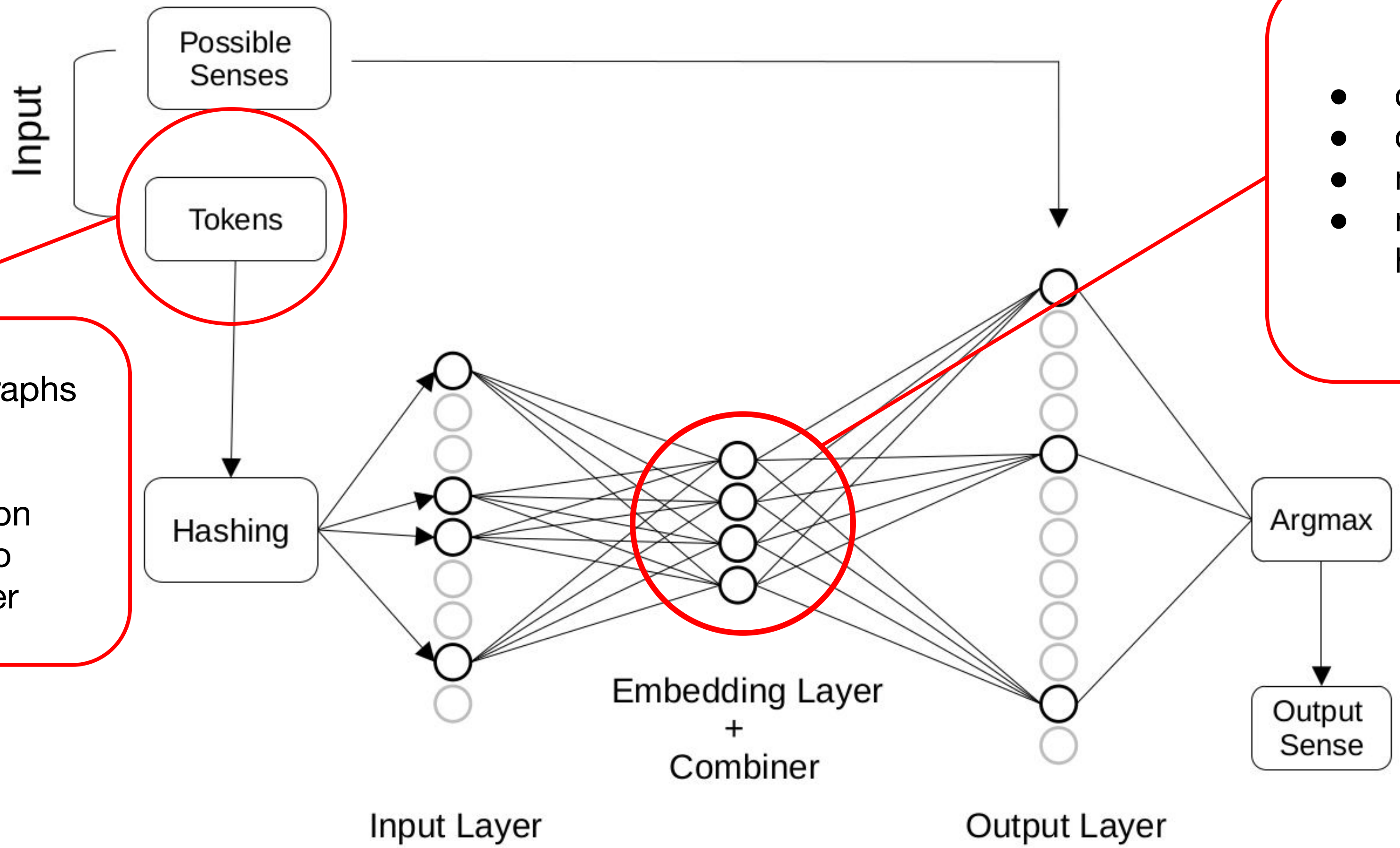
Hashing Model [Sc21]: Word Embeddings Through Hashing



[Sc21] Schulz, C.: "Automatische Disambiguation Für Das Englische". 2021

Model Architecture (1st of two)

Parameters for Quantitative Evaluation of Hashing Model [Sc21]



- using whole paragraphs or sentences
- lemmatize
- add POS information
- different n-grams to preserve word order

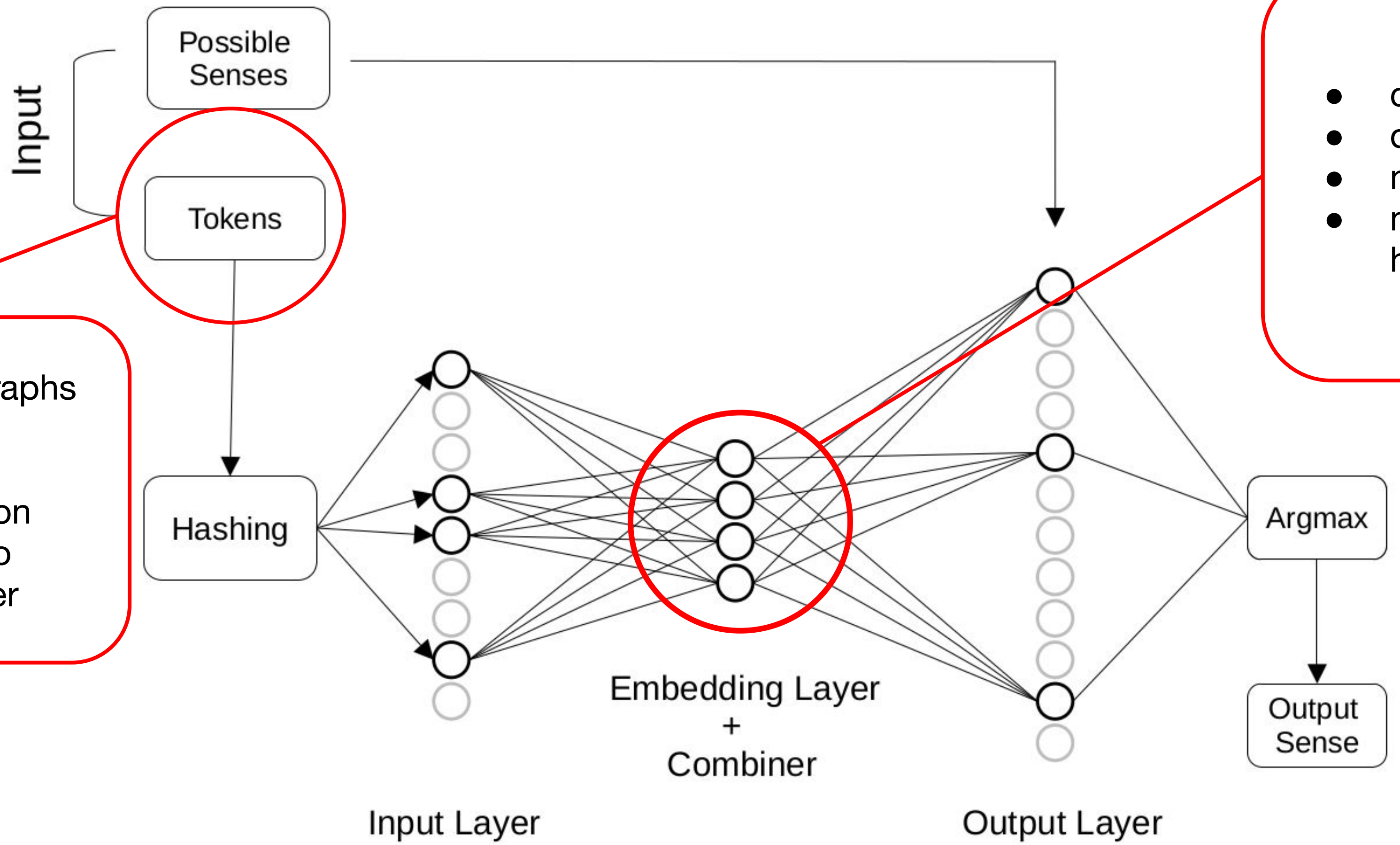
- combination formula
- dropout
- number of hidden layers
- number of nodes in the hidden layers

[Sc21] Schulz, C.: "Automatische Disambiguation Für Das Englische". 2021

Model Architecture (1st of two)

Parameters for Quantitative Evaluation of Hashing Model [Sc21]

- using whole paragraphs or sentences
- lemmatize
- add POS information
- different n-grams to preserve word order



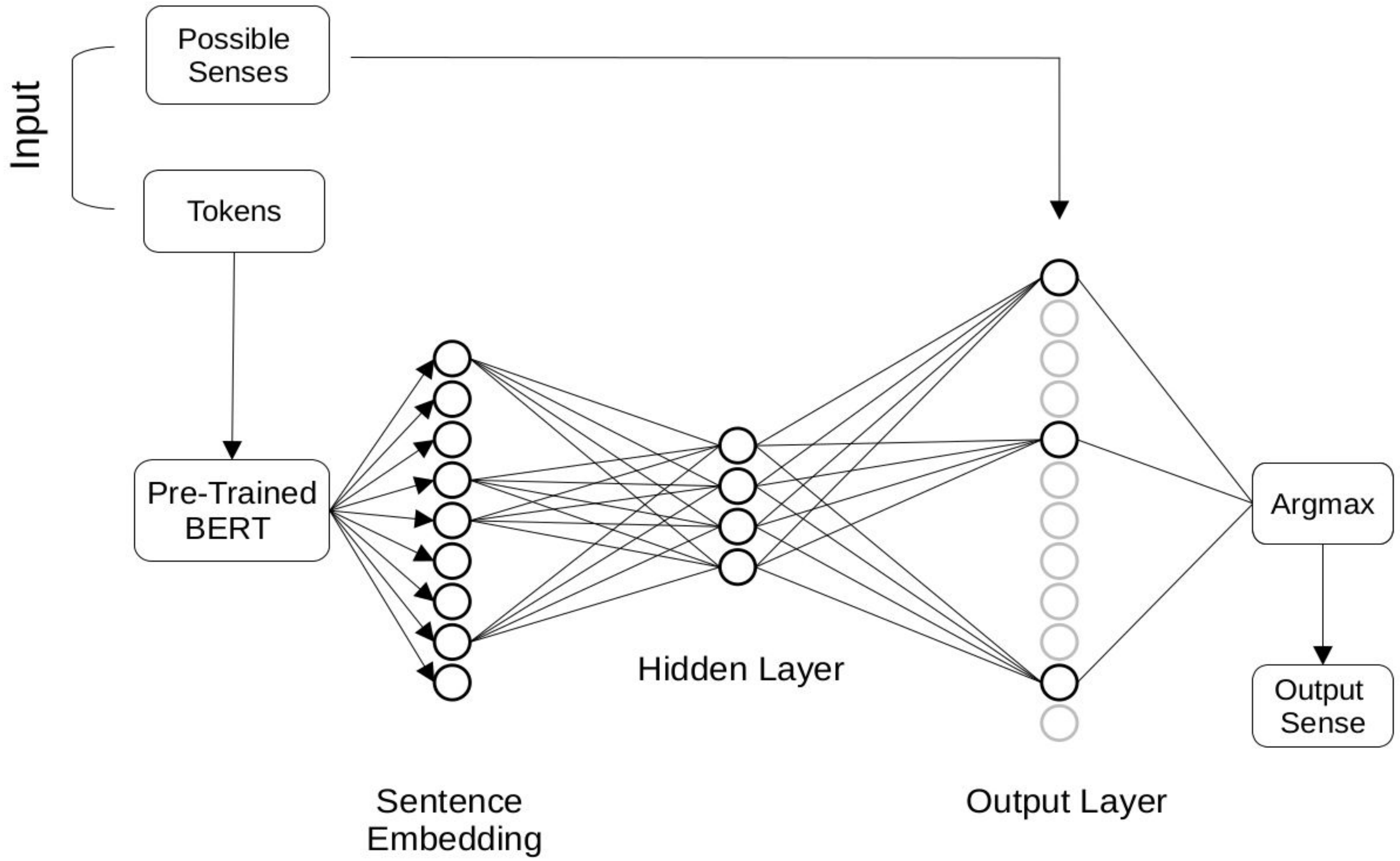
- combination formula
- dropout
- number of hidden layers
- number of nodes in the hidden layers

WSD on Wikipedia Texts achieve up to 84% “macro” accuracy

[Sc21] Schulz, C.: “Automatische Disambiguation Für Das Englische”. 2021

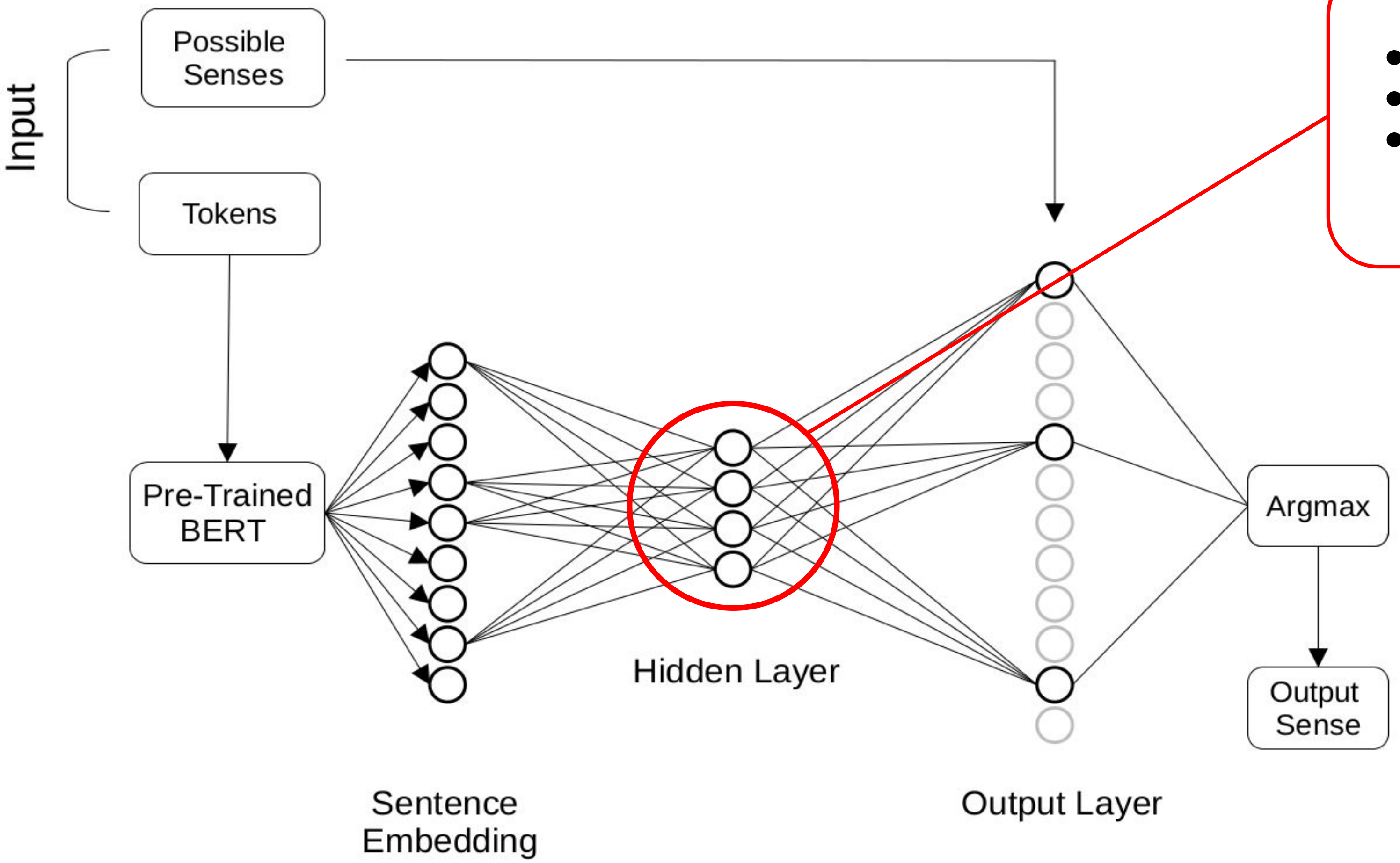
Model Architecture (2nd of two)

FastBERT Model: Using Contextual BERT Embeddings for WSD



Model Architecture (2nd of two)

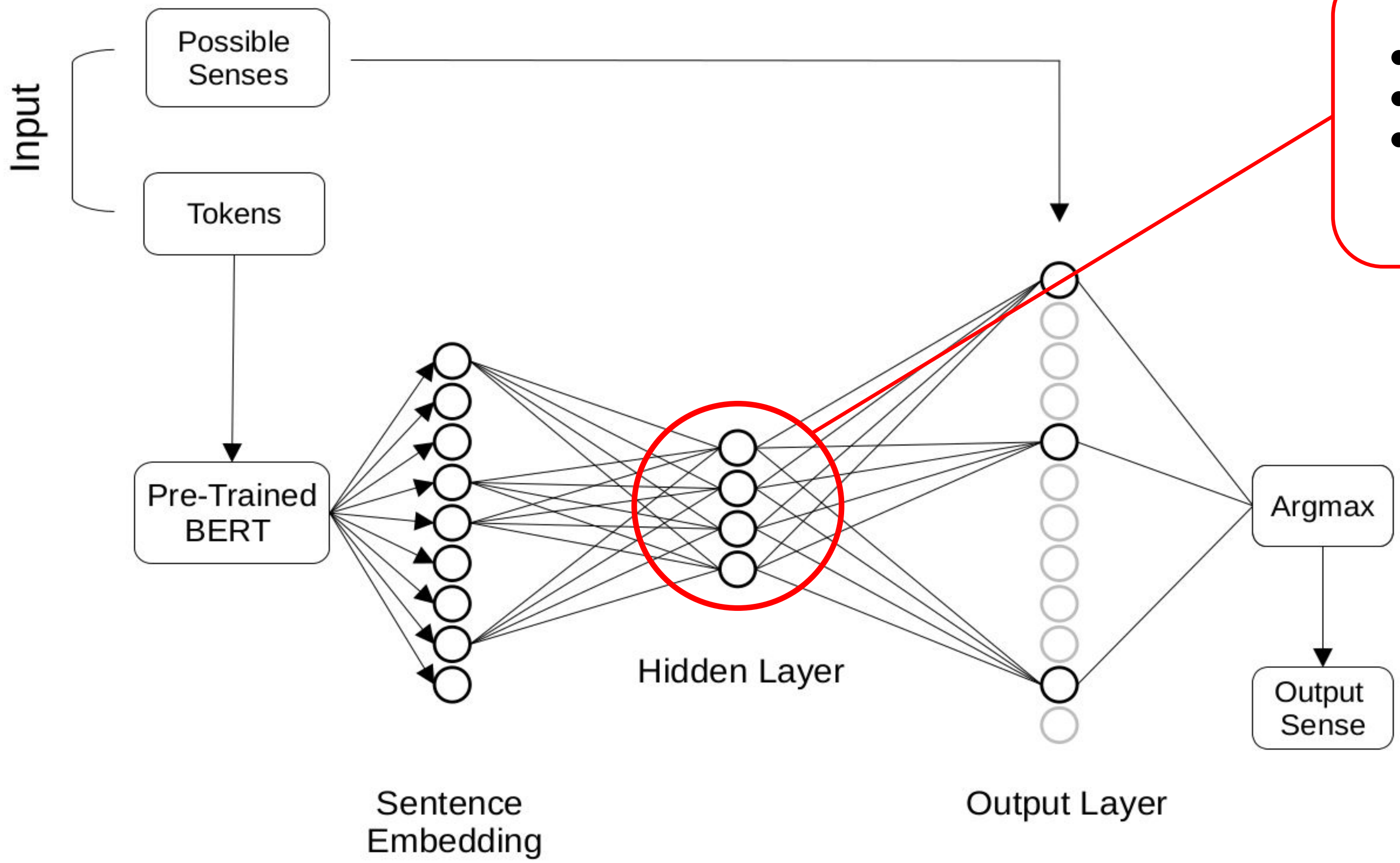
Parameters for Quantitative Evaluation of FastBERT Model



- dropout
- number of hidden layers
- number of nodes in the hidden layers

Model Architecture (2nd of two)

Parameters for Quantitative Evaluation of FastBERT Model

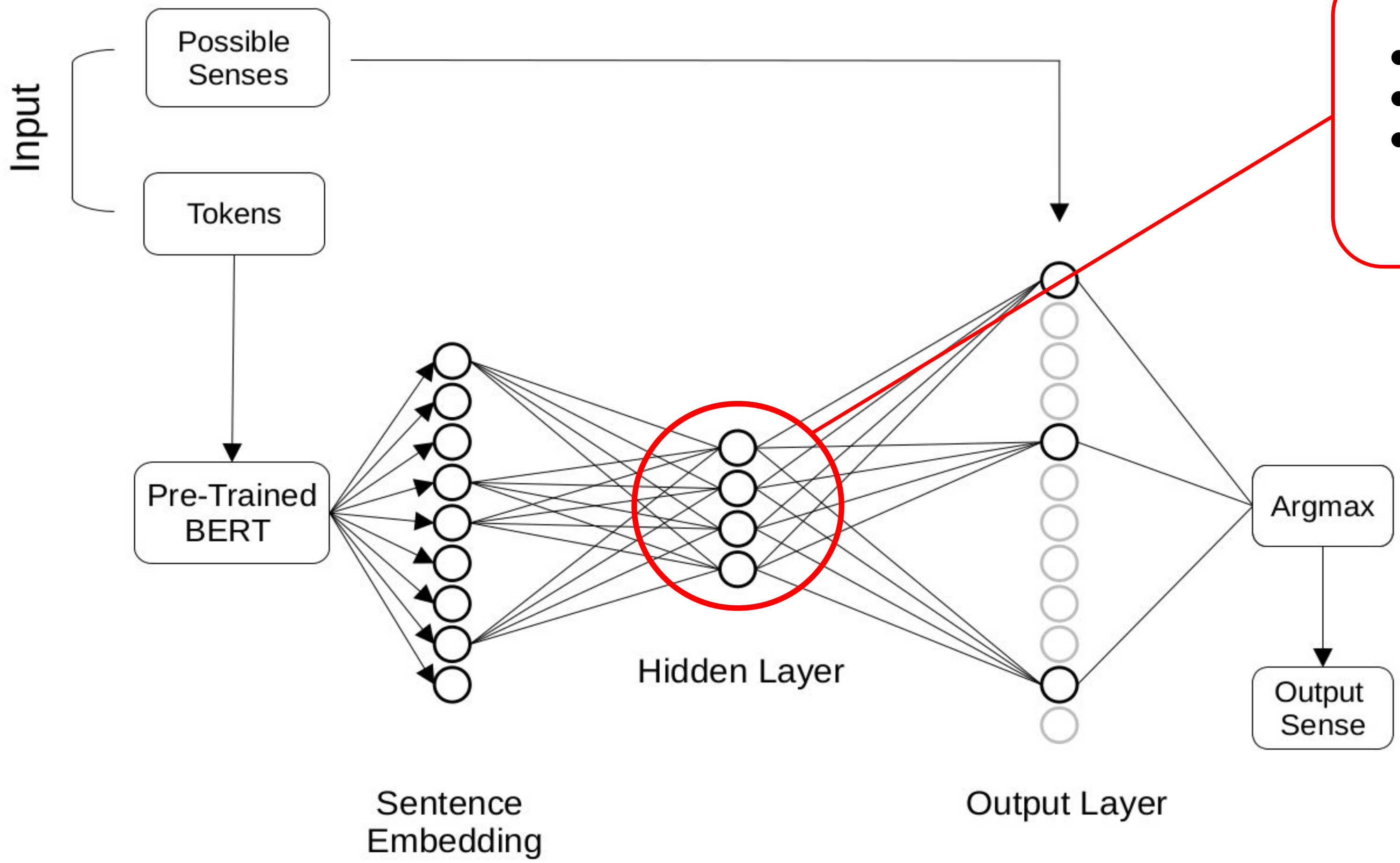


- dropout
- number of hidden layers
- number of nodes in the hidden layers

WSD on Wikipedia texts achieve up to 88% “micro” accuracy

Model Architecture (2nd of two)

Parameters for Quantitative Evaluation of FastBERT Model

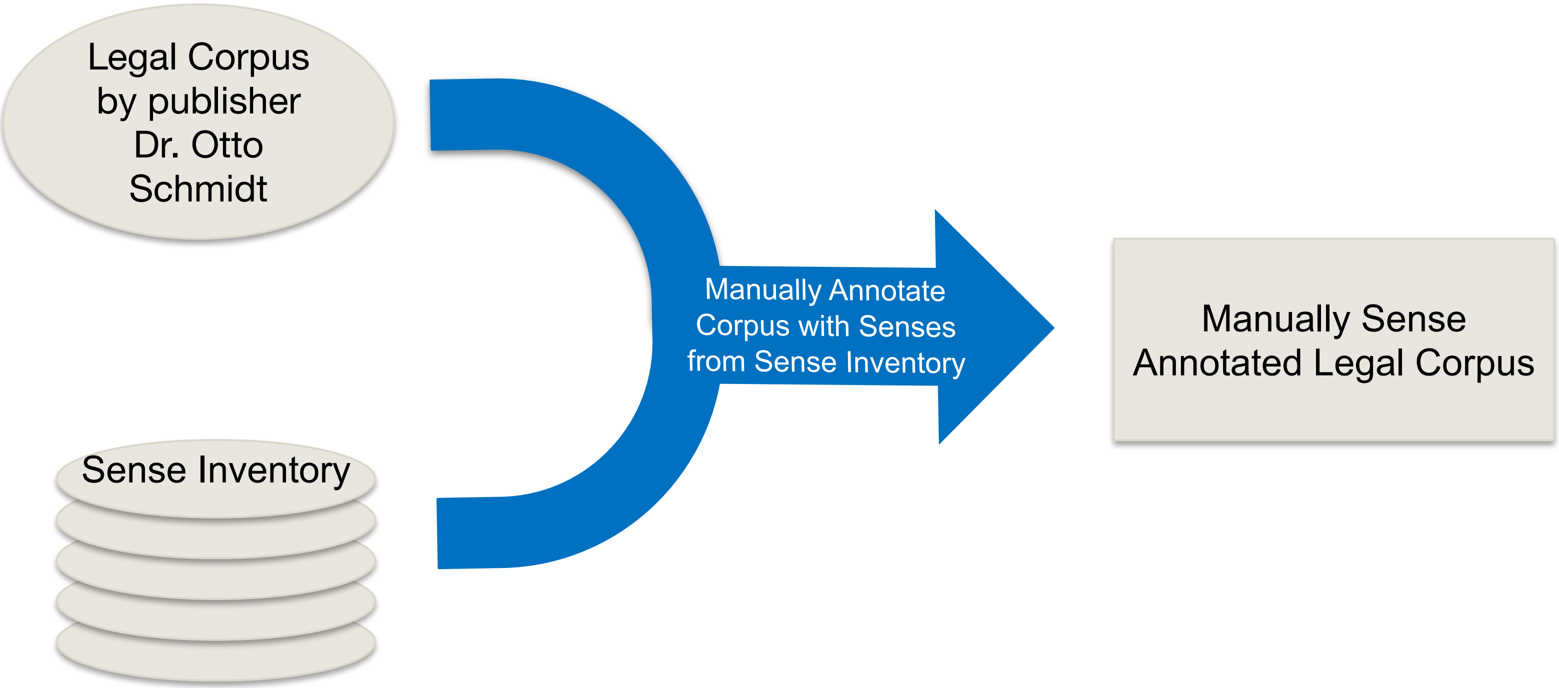


- dropout
- number of hidden layers
- number of nodes in the hidden layers

WSD on Wikipedia texts achieve up to 88% “micro” accuracy
WSD on manually sense tagged legal corpus achieve up to 74% “micro” accuracy

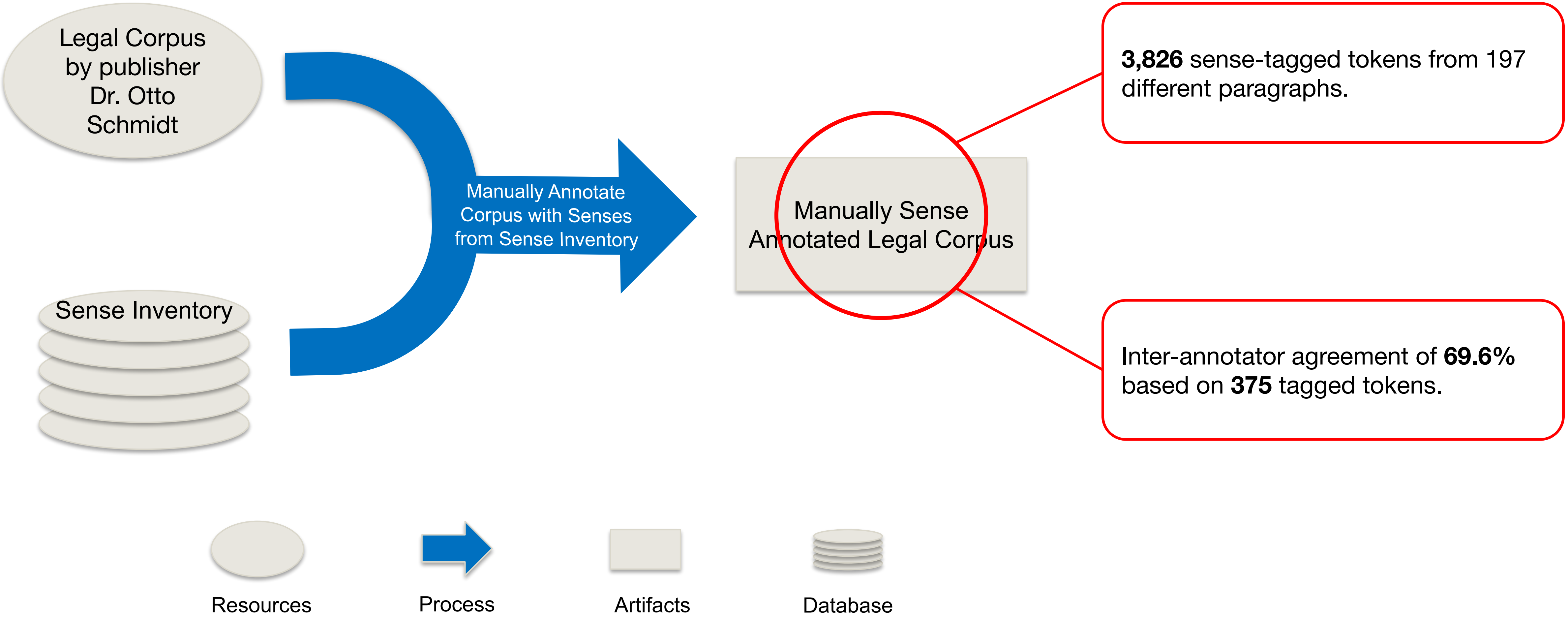
Sense Annotating Legal Corpus

Manual Sense Tagging of Legal Corpus with Senses from Sense Inventory



Sense Annotating Legal Corpus

Manual Sense Tagging of Legal Corpus with Senses from Sense Inventory



Outline

Qualitative Evaluation

Qualitative Evaluation with 11 Legal Practitioners

Interviews and Methodology

- **11** legal experts
- semi-structured expert interview (ca. 45 minutes over Zoom)
- diverserse group (age, gender, experience, IT-affinity, legal field)
- 2 parts: their **legal research behavior** in general and their **assessment** of the **quality** of the distinguishable **word senses**

lead to

Insights

- provided **word senses are not always clearly distinguishable** (e.g. “Methodik” and “Algorithmus” are two senses of the word “Verfahren”)
- **granularity** of senses should be adjustable
- more useful for lawyers who are **not yet fully familiar** with a legal field
- lawyers would like to filter already **during query input** into the search bar

Outline

Conclusion & Future Work

3 out of 4 Senses are Correctly Labelled: Good Automation Potential for Practical Deployment

RQs

- neural network based classifiers promising due to **high stability** and ability to **generalize well on unseen data**
- **sense inventory** can be created from **encyclopedic resources**
- **contextual embedding** based model achieves up to **74%** “micro” accuracy **on legal corpora**
- legal experts would use sense filter **if senses are clearly distinguishable and granularity can be adjusted**

Key Findings

- major limiting factor: **low inter-annotator agreement** due to **unclear senses**
- lawyers judge a sense filter **especially helpful** when they are **not yet fully familiar** with a legal field
- lawyers prefer to **filter** for word senses **already during query input** into the search engine
- despite the **good automation potential** of the WSD classifier for **practical deployment**, **further challenges** need to be tackled!

Future Work

- **remove proper names** from sense inventory
- **extend sense inventory** with senses from other lexical sources (e.g. GermaNet, German Wiktionary,...)
- make senses **clearly distinguishable**

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 +49.89.28917132
Fax +49.89.289.17136

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

Backup Slides

Evaluation Metrics

Quantitative Evaluation Metrics for WSD Applications

Measure	Used In This Thesis?	Comment
Coverage	✗	No information about undetected ambiguous words available.
Accuracy	✓	2 measures: *micro* and *macro* accuracy: $micro = \frac{\text{\# correctly classified senses}}{\text{\# total classified senses}}$ $macro = \frac{1}{n}\sum_{i=1}^{n} \frac{\text{\# correctly classified instances of sense } s_i}{\text{\# total classified instances of sense } s_i}$
Recall	✗	Cannot be calculated without coverage information.
F_1 Score	✗	Cannot be calculated without recall information.

Hashing Model: Combiner Hyperparameter

Square Root Combiner

$$avg_1 = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} v_i$$

Average Combiner

$$avg_2 = \frac{1}{n} \sum_{i=1}^{n} v_i$$

Motivation

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

Interviews and Methodology

- 6 lawyers
- 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

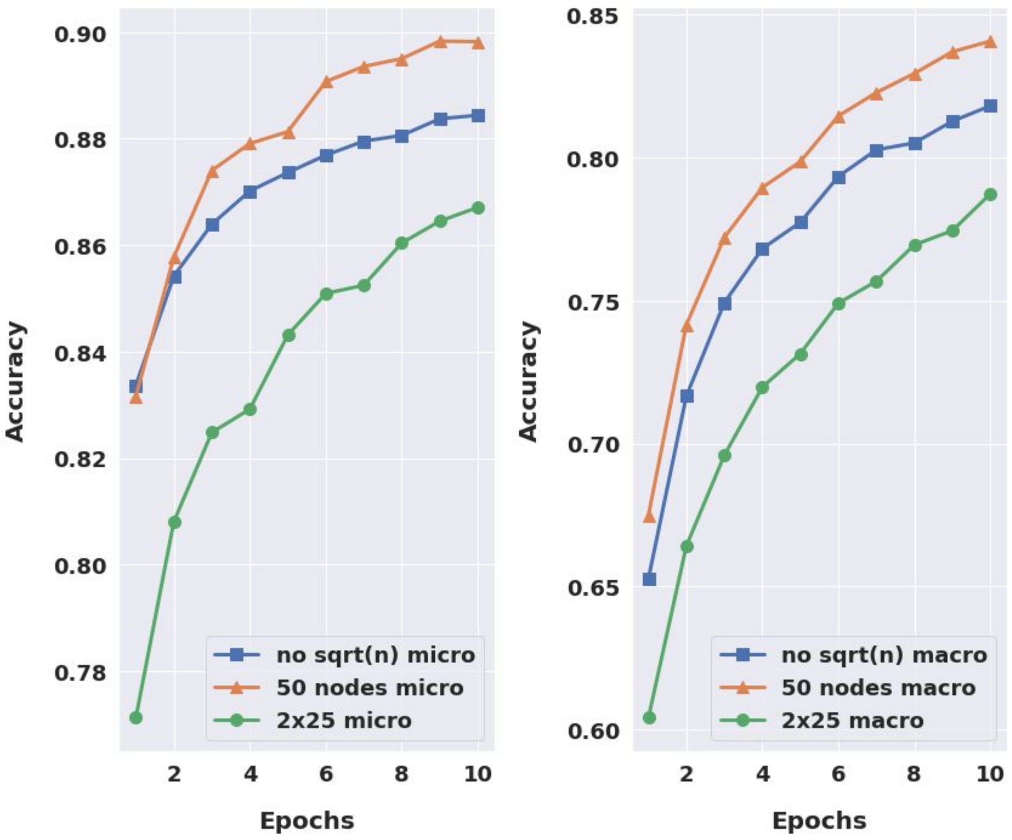
lead to

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 (!)**

Quantitative Evaluation

Hashing Model: Parameter Study on Wikipedia Corpus



Model accuracy as a function of the training epochs. The model was trained and evaluated on over **64M** samples containing over **340,000 senses**.

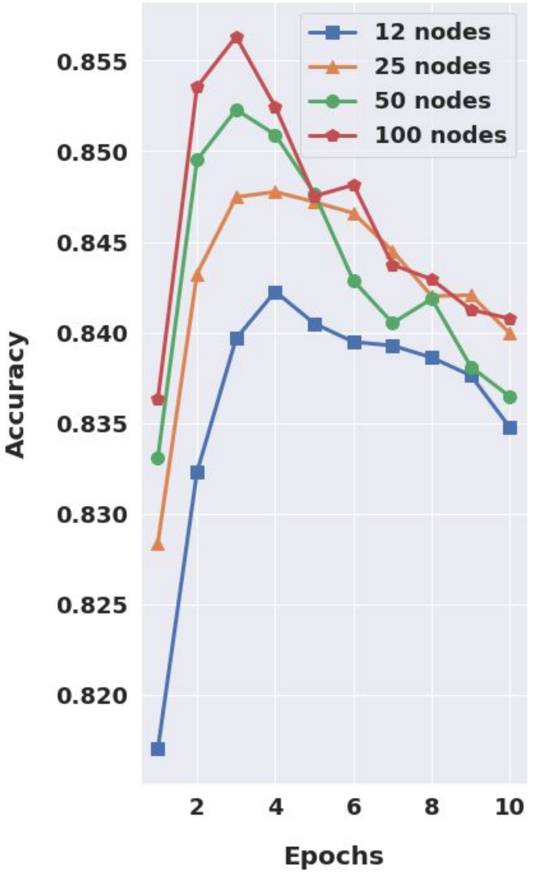
Lemmatized	NG	POS	n	S	L	N	B	DO	η_0	δ	Micro	Macro
✓	1				1	25	32	0	1	0.98	88.30%	81.61%
✓	2				1	25	32	0	1	0.98	87.73%	83.43%
✓	3				1	25	32	0	1	0.98	86.85%	83.63%
	1				1	25	32	0	1	0.98	88.06%	81.50%
✓	1			✓	1	25	32	0	1	0.98	88.30%	81.79%
✓	1				1	25	32	10%	1	0.98	87.55%	79.93%
✓	1				1	25	32	20%	1	0.98	86.93%	78.17%
✓	1				1	25	32	30%	1	0.98	86.34%	76.49%
✓	1	✓		✓	1	25	32	0	1	0.98	87.98%	81.03%
✓	1	✓			1	25	32	0	1	0.98	88.08%	81.31%
✓	1		✓		1	25	32	0	1	0.98	88.44%	81.81%
✓	1				1	50	32	0	1	0.98	**89.83%**	**84.08%**
✓	1				2	25x25	32	0	1	0.98	86.71%	78.73%

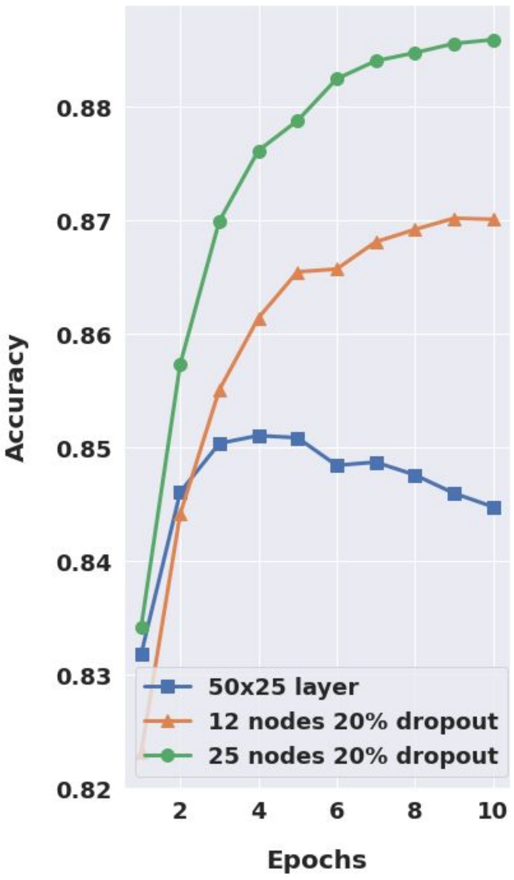
"**Micro**" is the maximum micro precision within 10 epochs.
"**Macro**" is the maximum macro precision within 10 epochs.
A **train-val-test** split of 75% / 15% / 15% was used.

Legend:
Lemmatized = word tokens are lemmatized, **NG** = n-gram size, **POS** = added part-of-speech information, **n** = average combiner was used, **S** = sentences instead of paragraphs used, **L** = number of hidden layers, **N** = number of nodes in the hidden layers, **B** = batch size, and **DO** = dropout rate, **η_0** = initial learning rate, **δ** = decay rate.

Quantitative Evaluation

FastBERT Model: Parameter Study on Wikipedia Corpus





Model accuracy as a function of the training epochs.
The model was trained and evaluated on over **314,700 samples** containing over **32,000 senses**.

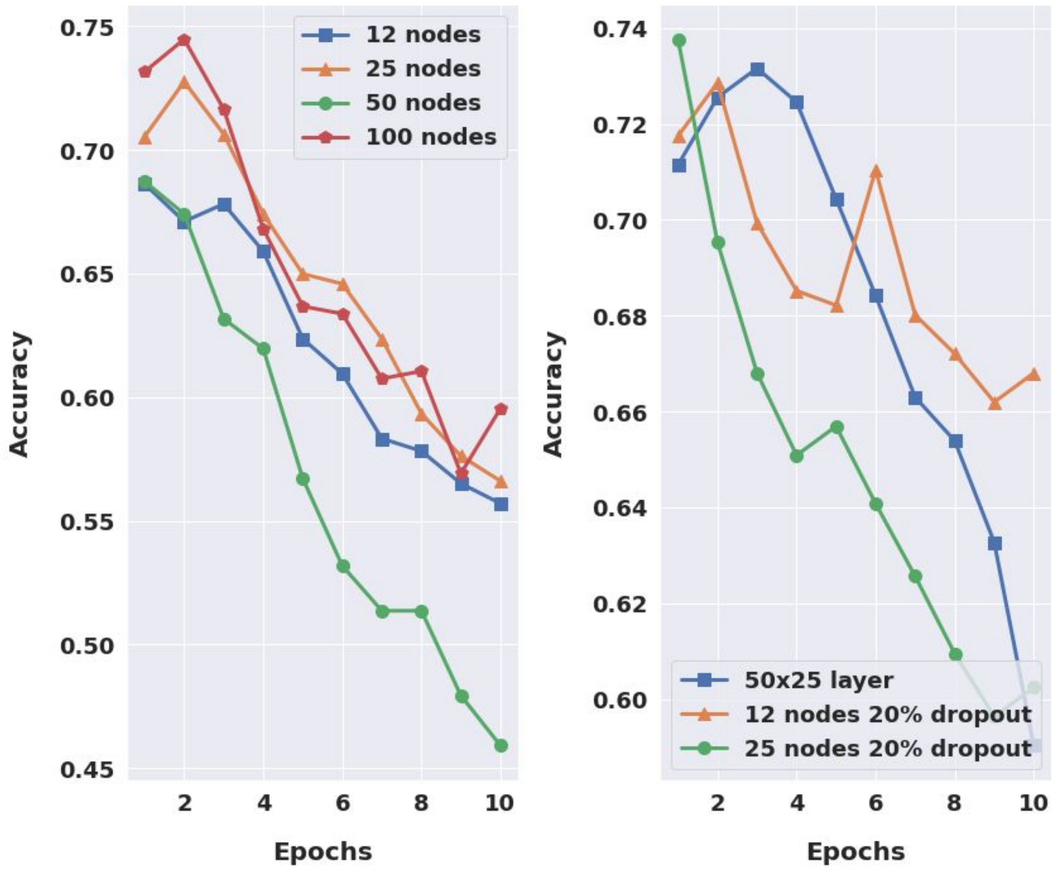
S	L	N	B	DO	Micro
✓	1	12	32	0	84.23%
✓	1	25	32	0	84.78%
✓	1	50	32	0	85.23%
✓	1	100	32	0	85.63%
✓	2	50x25	32	0	85.11%
✓	1	12	32	20%	87.02%
✓	1	25	32	20%	**88.59%**

“**Macro**" is the maximum macro precision within 10 epochs.
A **train-test** split of 75% / 25% was used.

Legend:
S = sentences instead of paragraphs used, **L** = number of hidden layers, **N** = number of nodes in the hidden layers, **B** = batch size, and **DO** = dropout rate.
"**Micro**" is the maximum micro precision within 10 epochs.

Quantitative Evaluation

FastBERT Model: Parameter Study on Manually Sense Tagged Legal Corpus



Model accuracy of the previously trained model when evaluated on **991** manually sense-tagged tokens from the legal corpus.

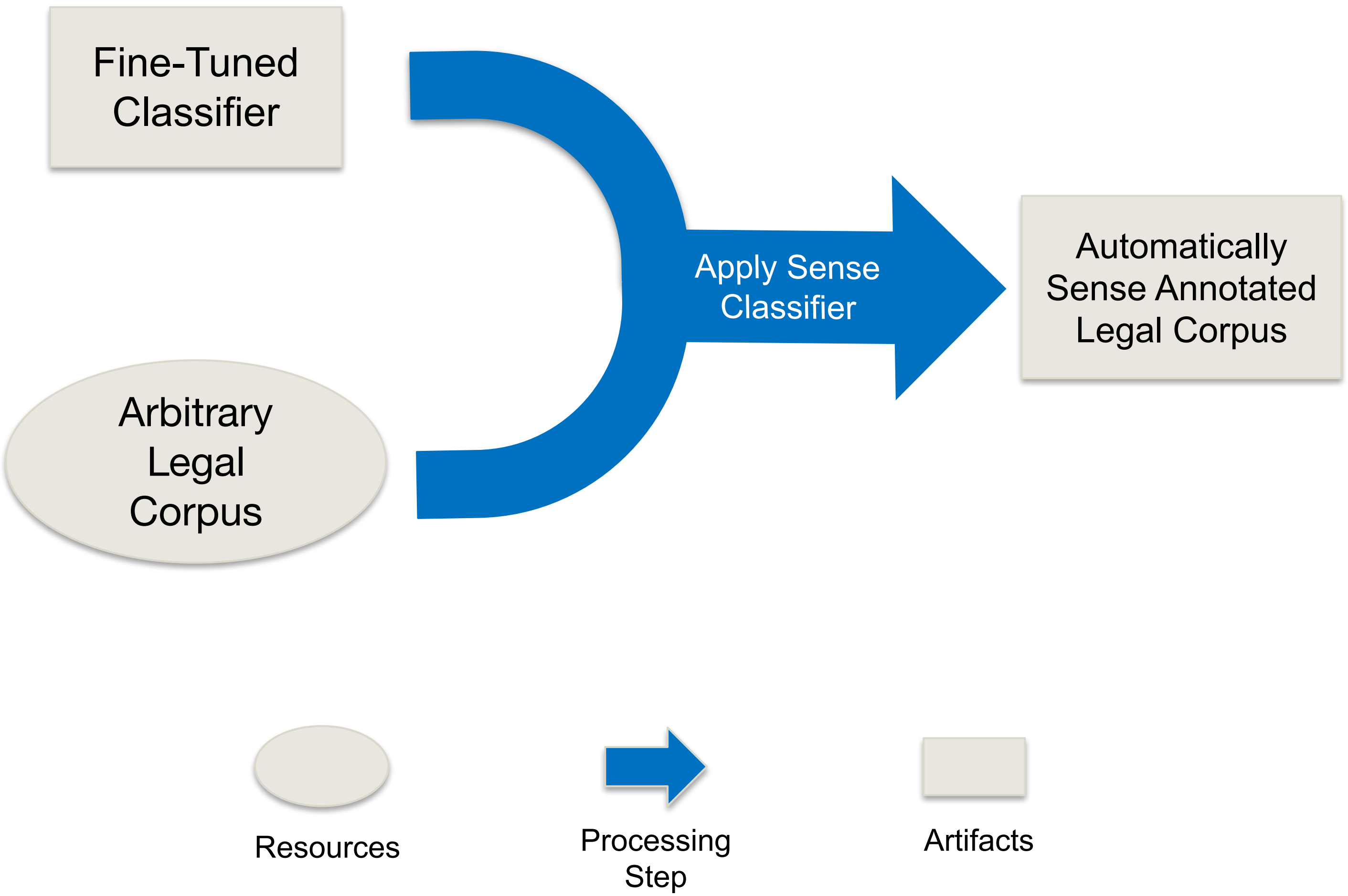
S	L	N	B	DO	Micro
✓	1	12	32	0	68.62%
✓	1	25	32	0	72.75%
✓	1	50	32	0	68.72%
✓	1	100	32	0	**74.47%**
✓	2	50x25	32	0	73.16%
✓	1	12	32	20%	72.86%
✓	1	25	32	20%	73.76%

"**Micro**" is the maximum micro precision within 10 epochs.

Legend:
S = sentences instead of paragraphs used, **L** = number of hidden layers, **N** = number of nodes in the hidden layers, **B** = batch size, and **DO** = dropout rate.

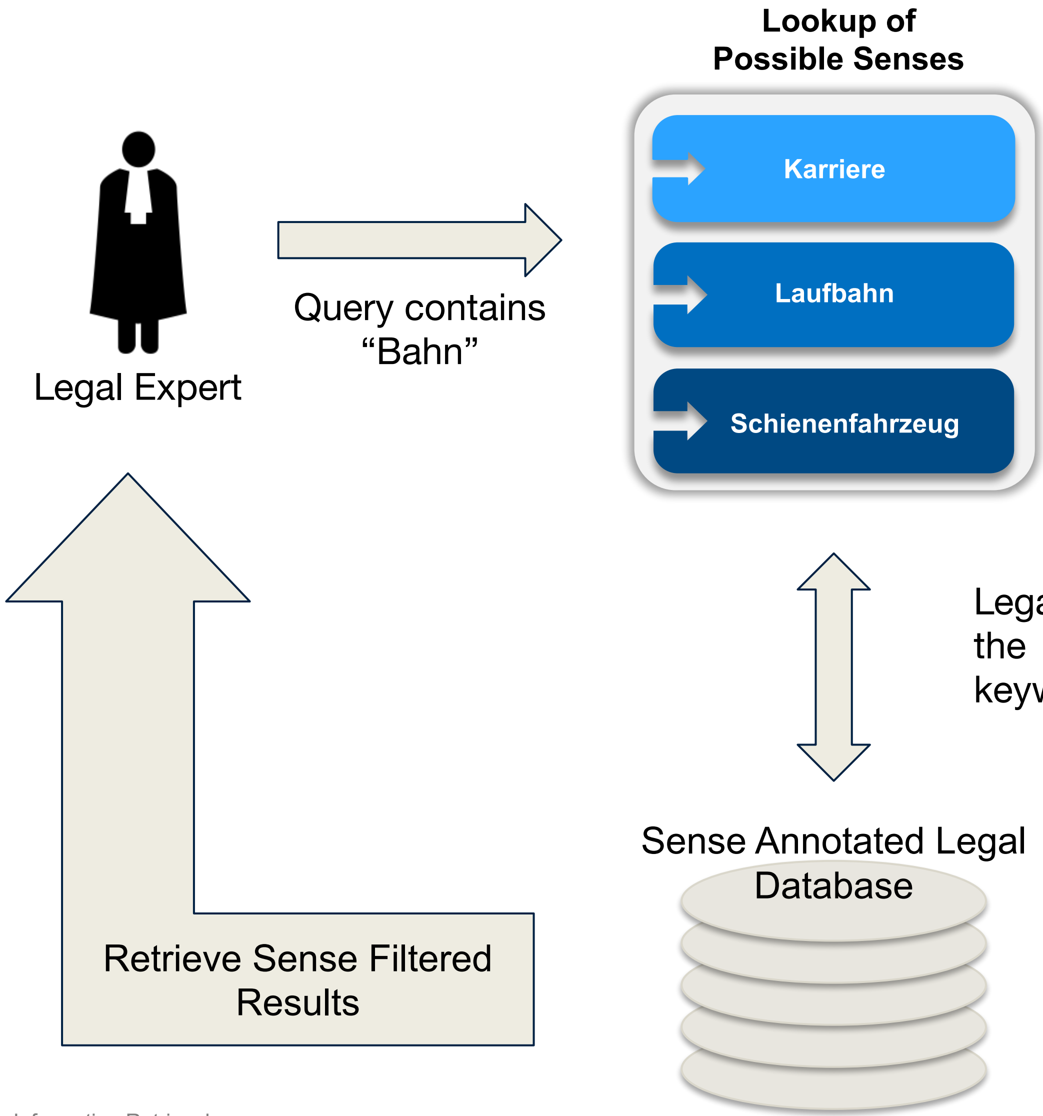
Application of Sense Classifier in Digital Legal Research

Use Fine-Tuned Classifier on Arbitrary Legal Corpus to Obtain a Sense-Tagged Legal Corpus



Application of Sense Classifier in Digital Legal Research

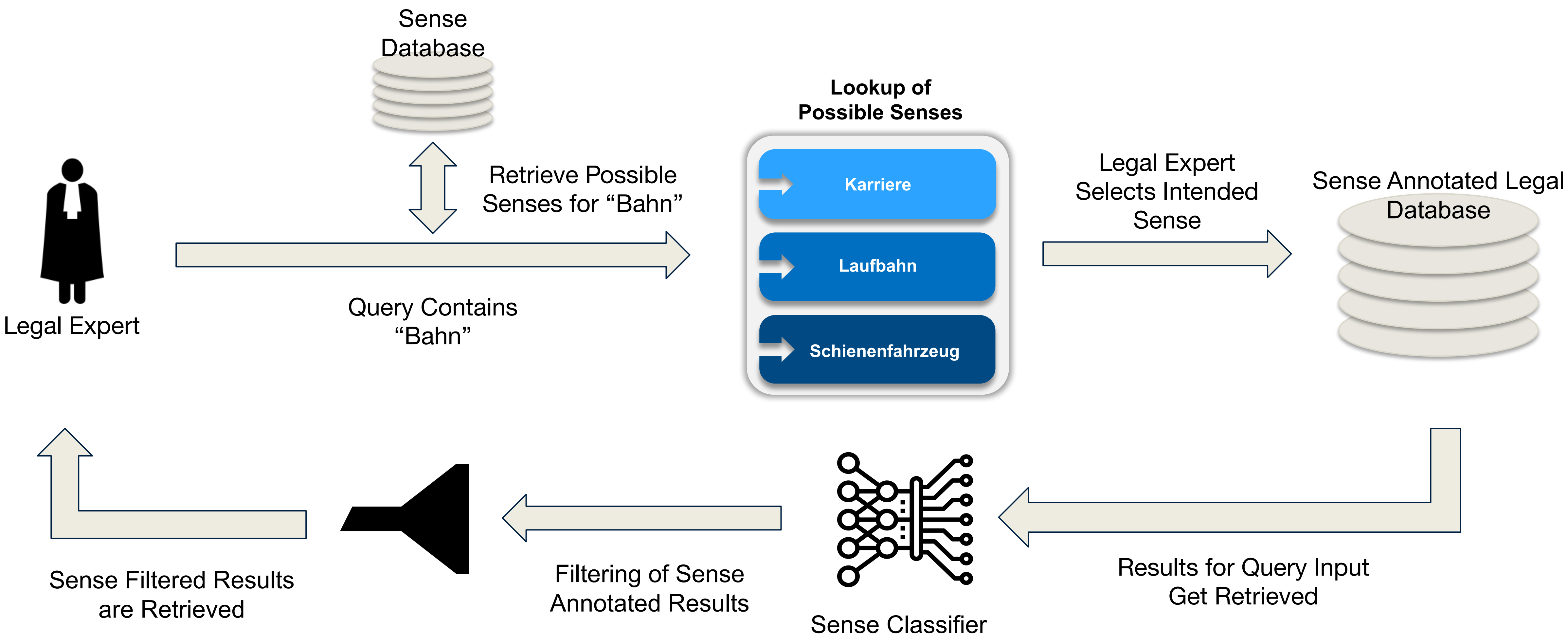
Possible Use Case 1: Legal Corpus Got Automatically Sense Annotated Before the Search



Legal Expert selects the intended sense from the possible senses. Only the results with the keyword(s) in the correct sense(s) are retrieved.

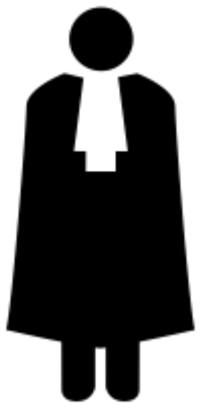
Application of Sense Classifier in Digital Legal Research

Possible Use Case 2: Search Results Get Sense Annotated on Demand



Motivation

Use Case: Filter Search Results by Senses



Legal Expert

Query contains the word

„Bahn“

Possible Senses

Sense 1:
Berufslaufbahn, Karriere, Werdegang

Sense 2:
Laufbahn, Startbahn, Landebahn

Sense 3:
Zug, Eisenbahn

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