Label Propagation for Tax Law Thesaurus Extension

Markus Müller, 09.11.2018, Master's Thesis Final Presentation

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

Advisors
Prof. Dr. Stephan Günnemann (Group for Data Mining and Analytics)
Jörg Landthaler, Elena Scepankova

Outline

Motivation

- Problem: Thesauri in the Legal Context
- Base Technology: Word Embeddings
- Opportunity: Label Propagation on Graphs

Research Approach

- Research Questions
- Research Methods
- Thesaurus Extension Tool

Evaluation Results

- Quantitative Evaluation
- Qualitative Evaluation
- Baseline Comparison

Conclusion & Future Work

Problem: Thesauri in the Legal Context

Legal Content Providers

Provide their users with access to **relevant** legal documents

Leading Providers in Germany

Wolters Kluwer

DATEV

HAUFE.

ottoschmidt

C·H·BECK

Thesauri enhance Information Retrieval via Synonym Sets

Search Query Expansion

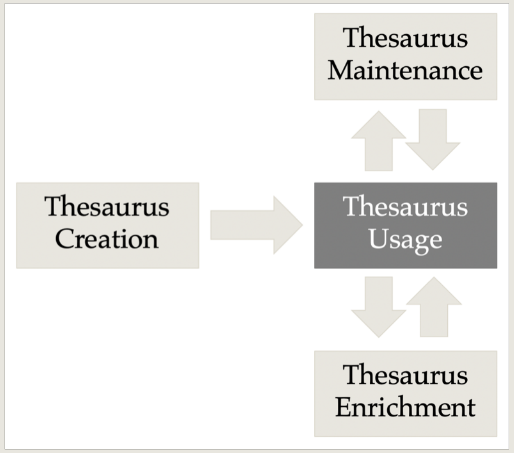
Abwrackprämie

Also showing results for "***Umweltprämie***"

[…] *Abwrackprämie*, the colloquial term for *Umweltprämie* […]

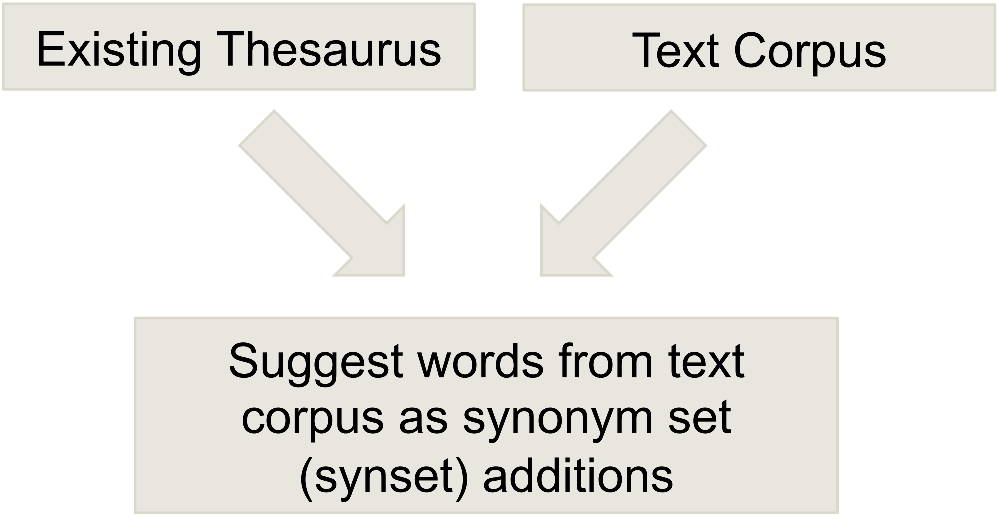
Creating and Maintaining Thesauri is hard

Mostly manual work, multiple domain-specific thesauri



Wolters Kluwer 2016 [1]

Focus: Thesaurus Extension as a Solution Approach



Existing Thesaurus

Text Corpus

Suggest words from text corpus as synonym set (synset) additions

Subject to research at this chair:
Landthaler et al. (2017) extended synsets starting from individual synset words

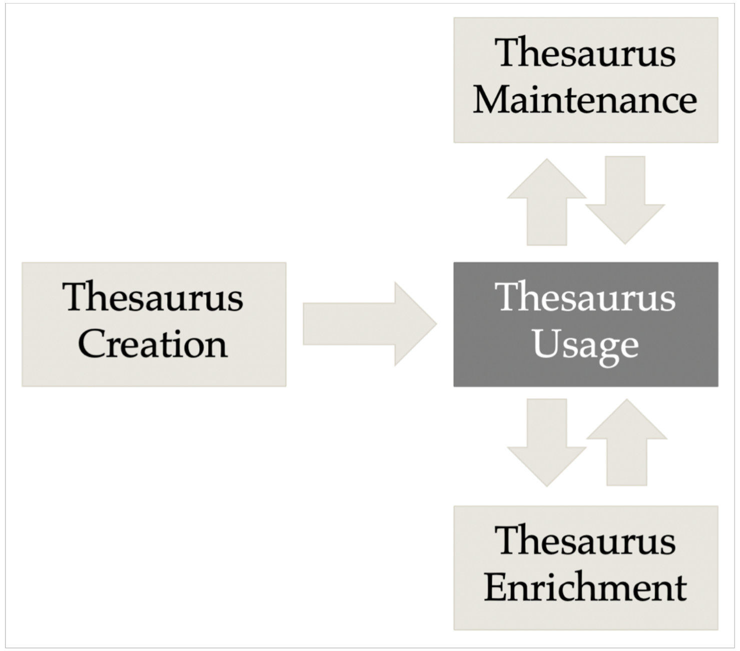
Potential Use-Cases for Thesaurus Extension

semi-automation, quality assurance

Suggest synset adjustments

semi-automation

Suggest additions to manually created synsets

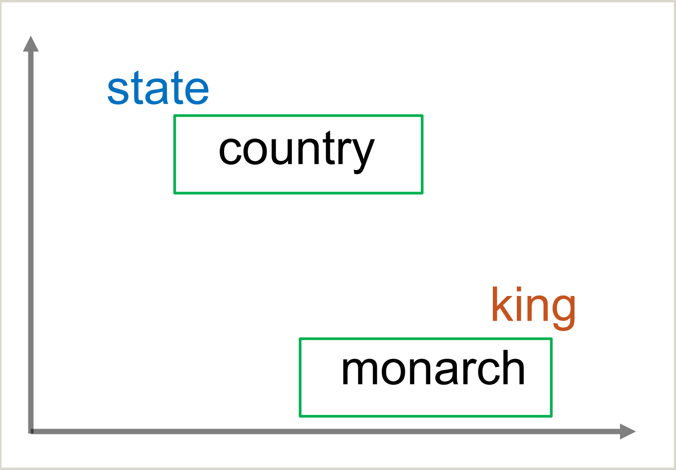


thesauri linkage

Identify relations between synsets of different thesauri

Problem with Vanilla Word Embeddings for Thesaurus Extension

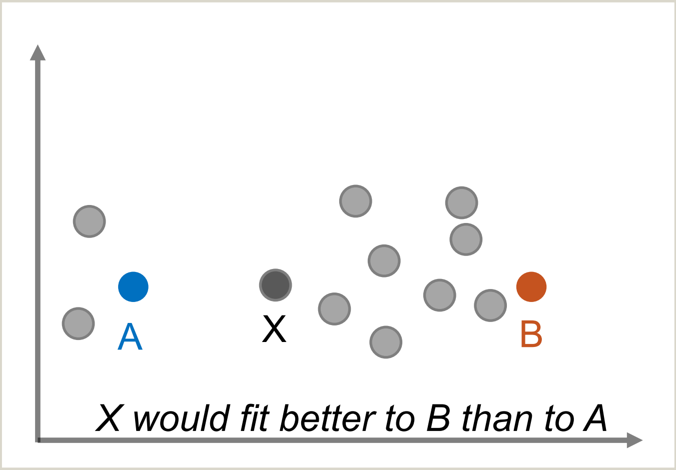
Word Embedding Technologies map similar words to similar vectors



⇒ **Nearest Neighbors:** Extend synset with words close to synset words

Blue & Red: Words from different existing synsets
Green: Extension suggestion

But then: Overall structure is not taken into account



X would fit better to B than to A

A & B: Labeled with different synsets
Rest: Unlabeled

⇒ **Opportunity:** Semi-Supervised Learning

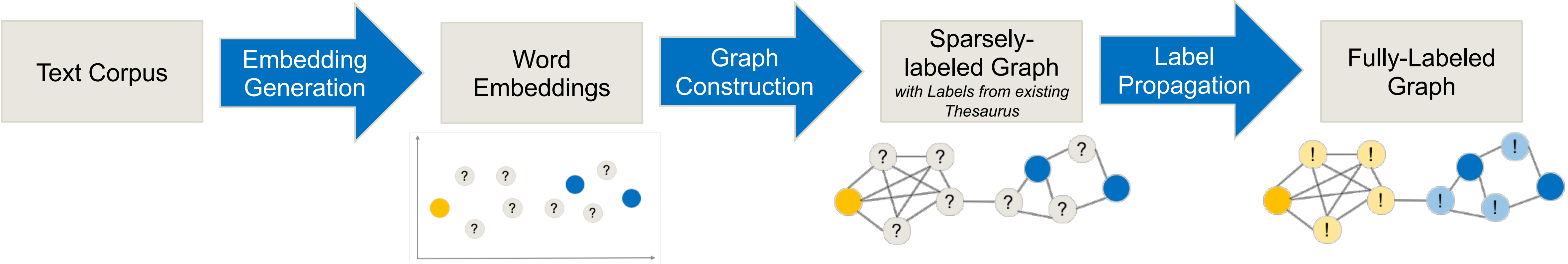
Research Idea: Label Propagation for Thesaurus Extension

Label Propagation is used by Google in Combination with Word Embeddings for knowledge graph extension, e.g. for **Emotion Association** and **Smart Replies**

https://ai.googleblog.com/2016/10/graph-powered-machine-learning-at-google.html & Ravi and Diao (2015)

RQ1: Can we **apply Label Propagation to Word Embeddings** to find new Synonyms?

Intuition



Text Corpus

Embedding Generation

Word Embeddings

Graph Construction

Sparsely-labeled Graph

with Labels from existing Thesaurus

Label Propagation

Fully-Labeled Graph

Research Questions

How can we get **semantic & context information into a graph** for LP? (RQ2)

Can we **model the thesaurus extension problem** on the LP technology? (RQ3)

What LP **algorithms work best**? (RQ4)

Is LP a **suitable technology** for thesaurus extension in the legal domain? (RQ1)

How much **automation** for **thesaurus creation** is achievable with LP? (RQ5)

Research Approach

Can we **model the thesaurus extension problem** on the LP technology? (RQ3)

Build a Thesaurus Extension Tool for trying out many approaches

How can we get **semantic & context information into a graph** for LP? (RQ2)

What LP **algorithms work best**? (RQ4)

Quantitative Evaluation
Automatic Parameter Studies

Is LP a **suitable technology** for thesaurus extension in the legal domain? (RQ1)

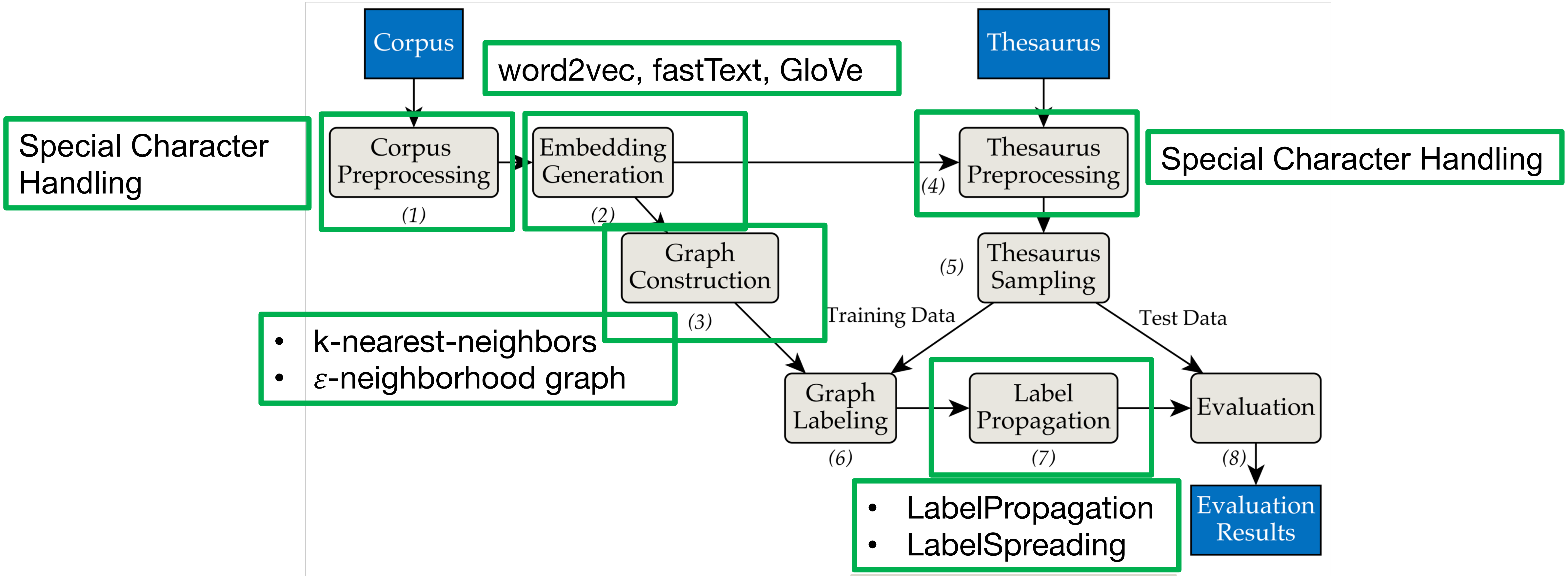
How much **automation for thesaurus creation** is achievable with LP? (RQ5)

Qualitative Evaluation
Manual Studies

Comparison with Vanilla Word Embeddings Approach

Thesaurus Extension Tool: Architecture

Extendable & Open Source on sebischair @ GitHub



Special Character Handling

word2vec, fastText, GloVe

Special Character Handling

- k-nearest-neighbors
- ε-neighborhood graph

- LabelPropagation
- LabelSpreading

Many variants incl. parameters implemented for later evaluation

via Group for Data Mining and Analytics

Pipes & Filters Architecture, Buschmann et al. (1996)

Input/Output

Filter

Quantitative Evaluation: Set-up

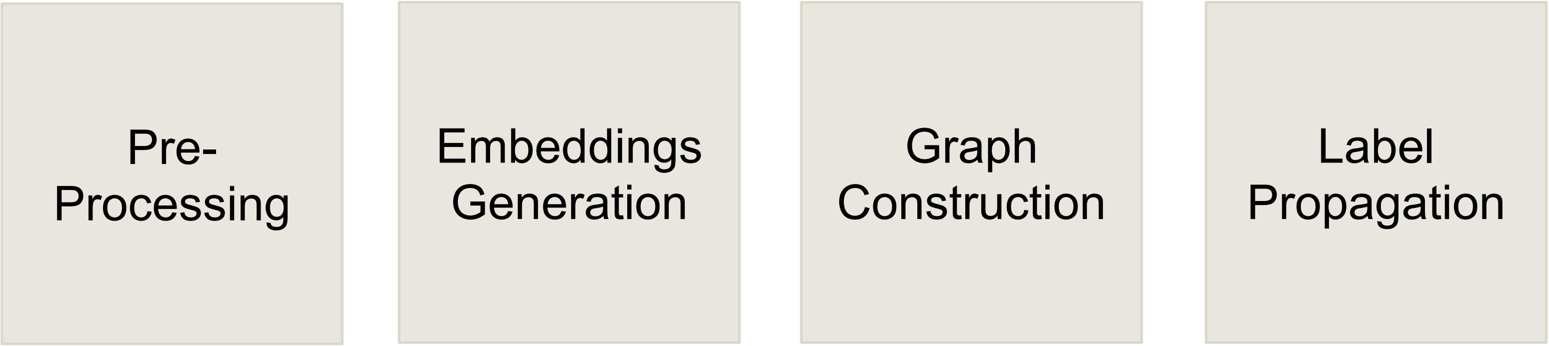
Tax Law Data Set by DATEV (in German)

- *text corpus:* 132,581 legal documents
- *handcrafted existing thesaurus: 12,288 synsets*

Evaluation Thesaurus (Subset):
2,552 thesaurus synsets

- **Training Set:** 3,277 words
- **Test Set:** 2,887 words

Hyper-Parameter Studies on these Phases



Pre-Processing

Embeddings Generation

Graph Construction

Label Propagation

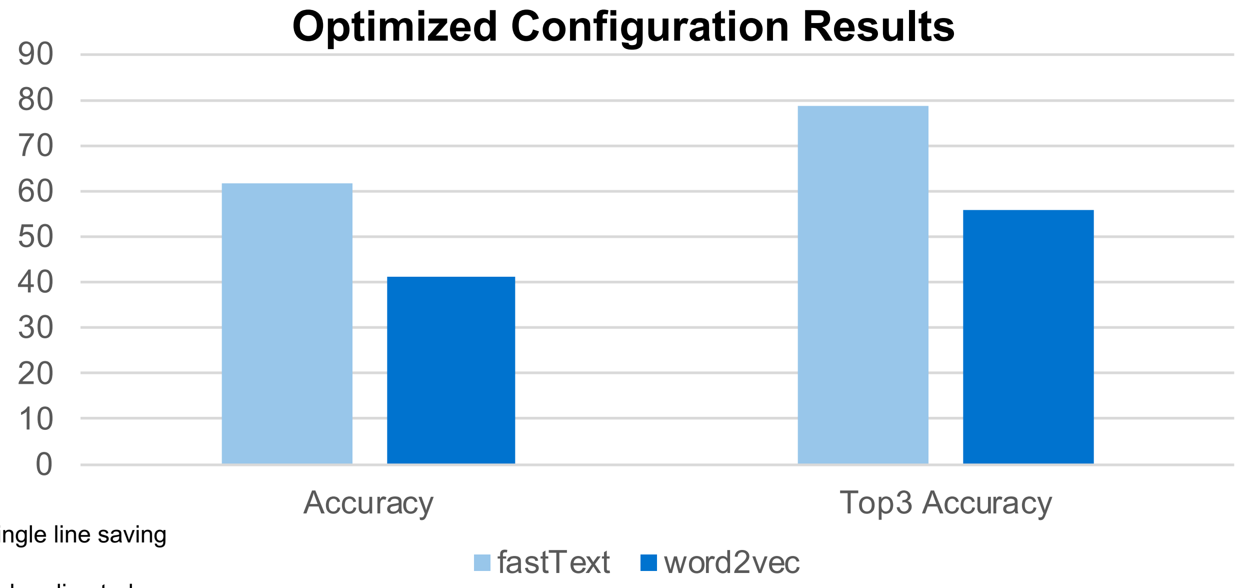
Goal: Find hyper-parameter configuration with highest accuracy
⇒ as input for Qualitative Evaluation

Challenge: Lots of possible configurations (> 1,000 runs)

Quantitative Evaluation: Lessons Learned & Final Result

Greatest performance impact: Word Embeddings Choice

High performance through hyper-parameter optimization



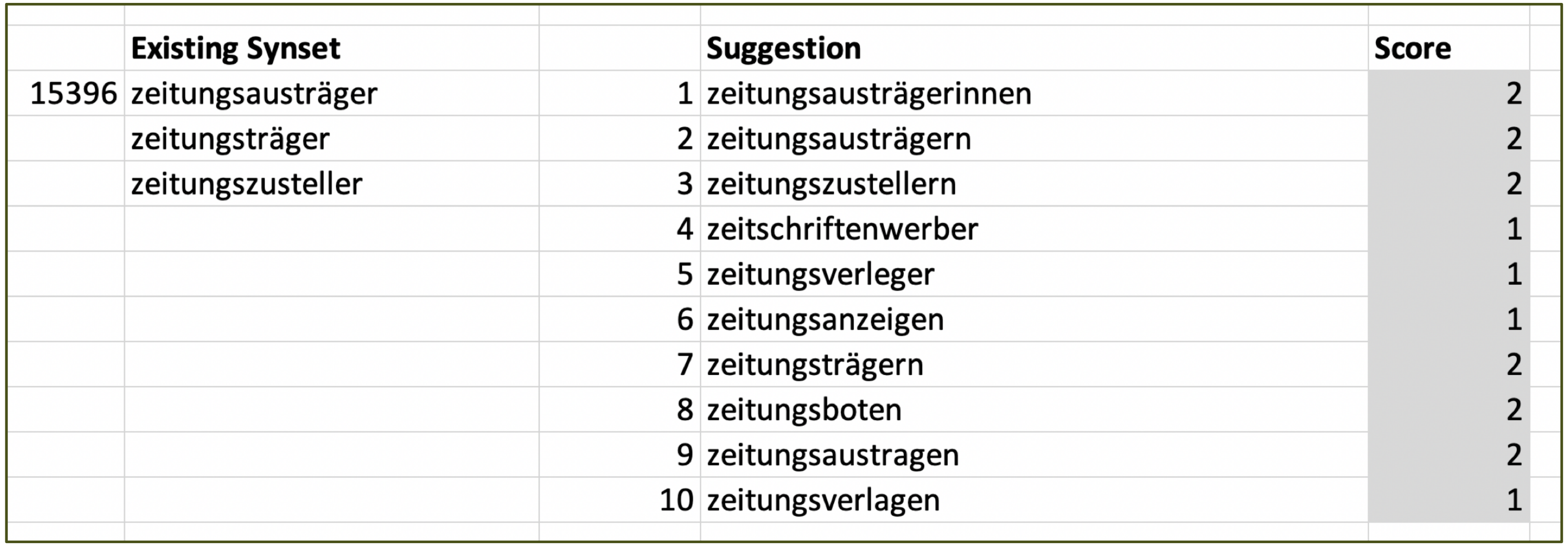
But: Also good suggestions outside of the existing thesaurus?

Configuration:
Pre-Processing: Keep letters & hyphens, muß⇒muss, single line saving
Embedding Generation: 400 dimensions, 40 iterations
Graph Construction: k-nearest neighbors, k=12, weighted undirected edges, no self-references allowed
Label Propagation: LabelSpreading, α=0.2, 15 iterations

Qualitative Evaluation: Set-up

Show synset suggestions to humans & get ratings

Pre-Study	Identify influence factors for good suggestions
Main Study (2x)	Rate suggestions of best configurations



	Existing Synset		Suggestion	Score
15396	zeitungsausträger	1	zeitungsausträgerinnen	2
	zeitungsträger	2	zeitungsausträgern	2
	zeitungszusteller	3	zeitungszustellern	2
		4	zeitschriftenwerber	1
		5	zeitungsverleger	1
		6	zeitungsanzeigen	1
		7	zeitungsträgern	2
		8	zeitungsboten	2
		9	zeitungsaustragen	2
		10	zeitungsverlagen	1

Scores
0: Not similar to predicted synset
1: Same semantic area
2: Should be added to synset

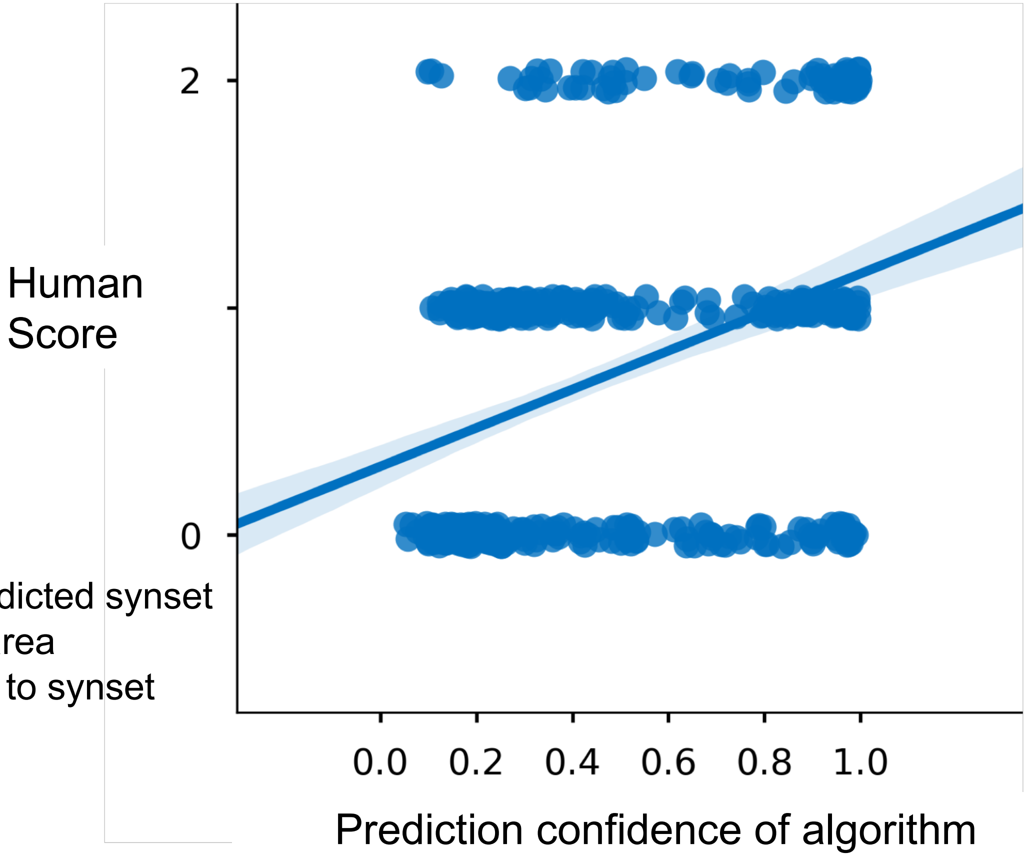
Rated 54 synsets per study, 10 suggestions per synset ⇒ **540 ratings/study**

- Originally planned with legal experts
- In the end, conducted by Jörg Landthaler & Markus Müller, supported by Text Corpus via ElasticSearch instance

Qualitative Evaluation: Pre-Study Lessons Learned

High confidence, high synset training number and low synset prediction number lead to better rating

E.g. correlation between prediction confidence and score

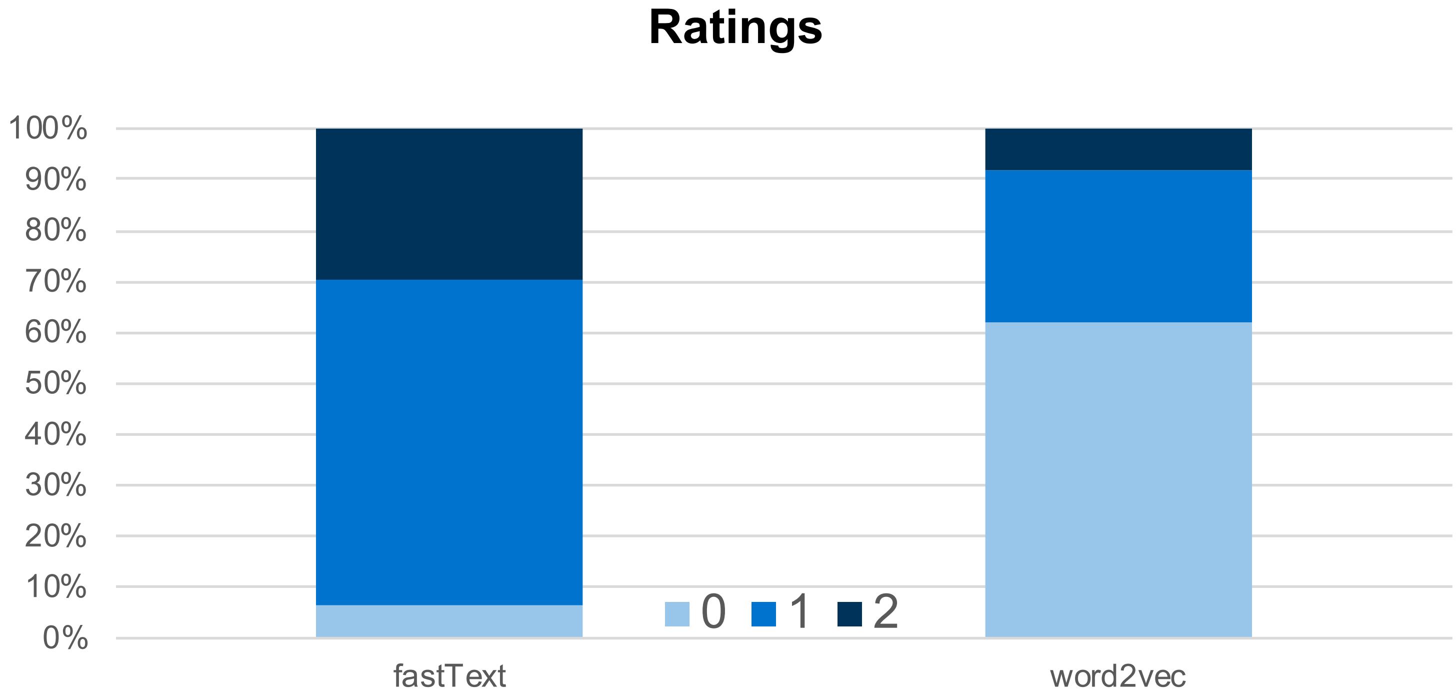


Scores
0: Not similar to predicted synset
1: Same semantic area
2: Should be added to synset

Qualitative Evaluation: Main Study Lessons Learned

fastText again considerably better than word2vec

But: Why does fastText perform better?



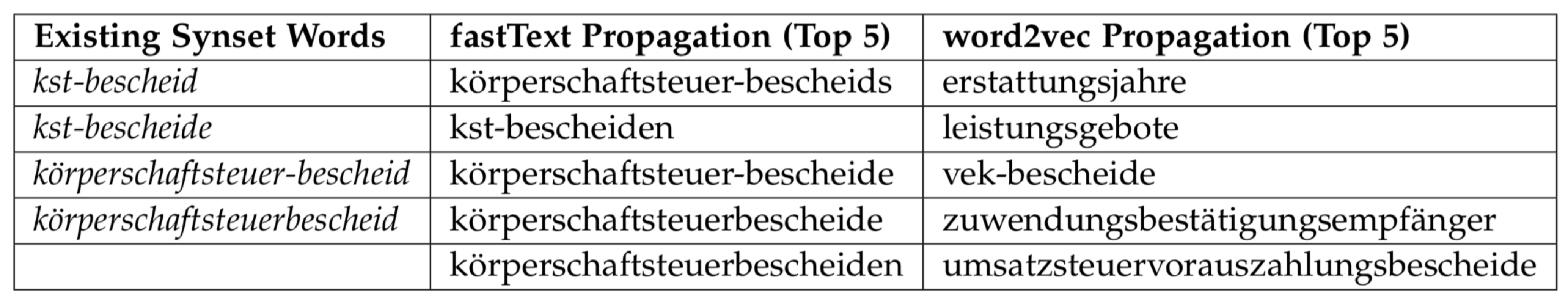
Scores
0: Not similar to predicted synset
1: Same semantic area
2: Should be added to synset

Qualitative Evaluation: Interpretation

fastText predominantely suggests **syntactically** similar words, word2vec suggests really different words (⇒ more interesting)

Our evaluations favored syntactically similar words

Example

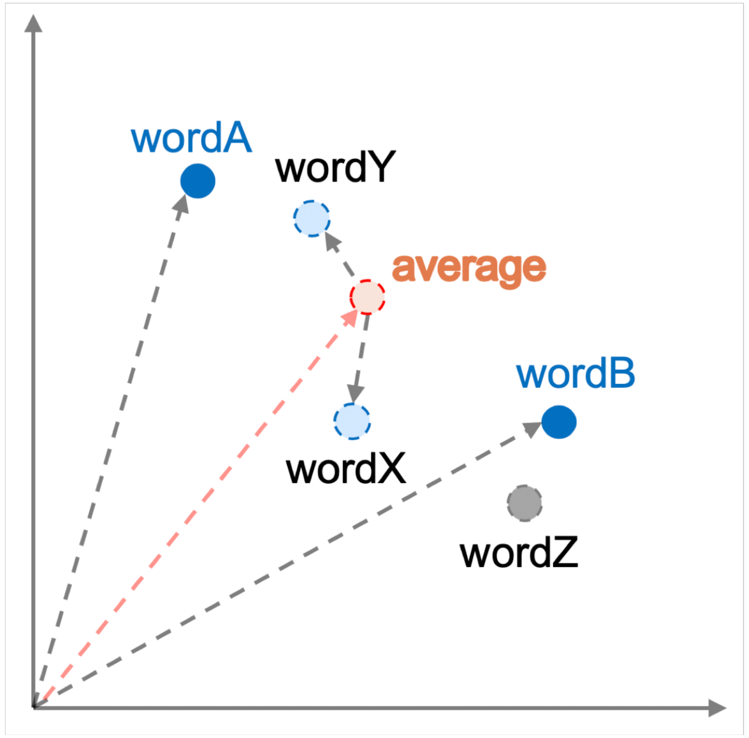


Existing Synset Words	**fastText Propagation (Top 5)**	**word2vec Propagation (Top 5)**
kst-bescheid	körperschaftsteuer-bescheids	erstattungsjahre
kst-bescheide	kst-bescheiden	leistungsgebote
körperschaftsteuer-bescheid	körperschaftsteuer-bescheide	vek-bescheide
körperschaftsteuerbescheid	körperschaftsteuerbescheide	zuwendungsbestätigungsempfänger
	körperschaftsteuerbescheiden	umsatzsteuervorauszahlungsbescheide

We compiled a list of common challenges around Thesaurus Extension

„Synset Vector“ Baseline: Approach

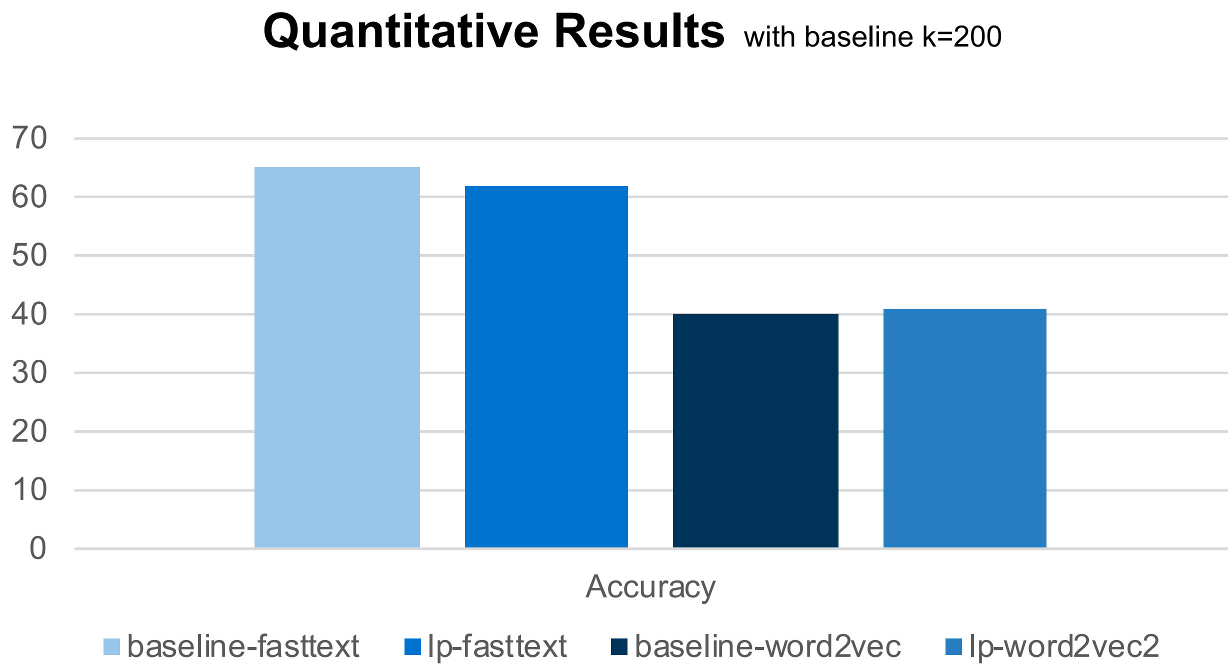
- Nearest neighbors approach, operates directly on word embeddings
- Self-designed, inspired by Rothe and Schütze (2016) [4]

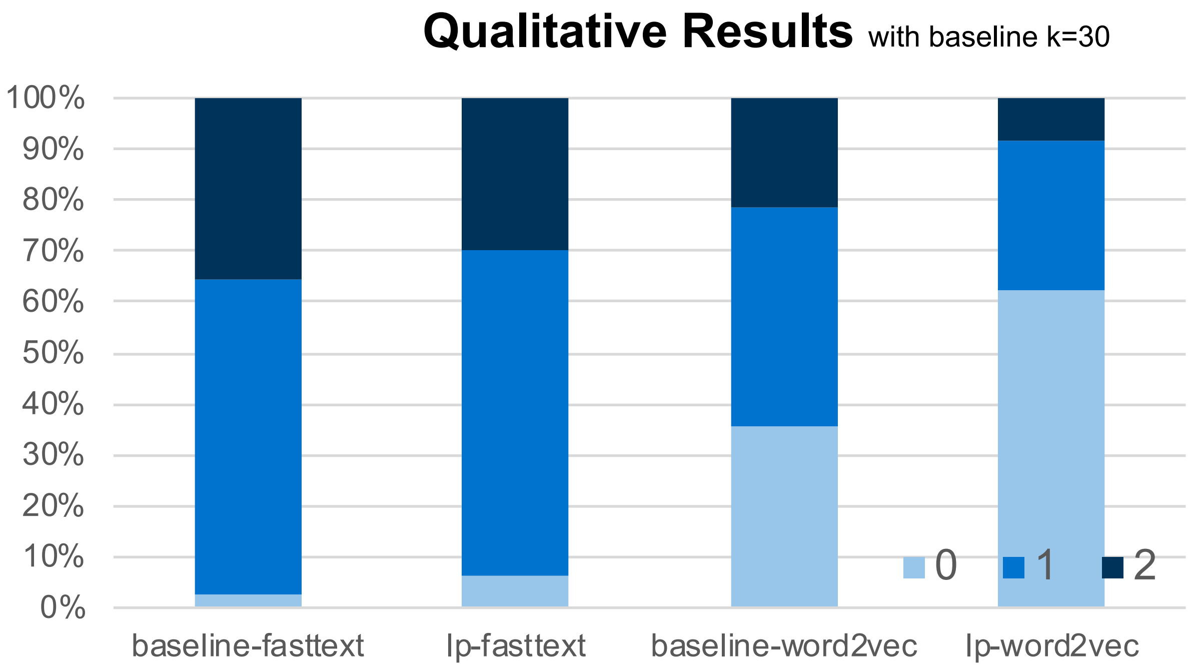


Intuition with k=2

„Synset Vector“ Baseline: Lessons Learned

Baseline performs equal or better than label propagation approach, while being less complex





Scores
0: Not similar to predicted synset
1: Same semantic area
2: Should be added to synset

Conclusion

Label Propagation approach was not better than Baseline, but overall results were promising

fastText and word2vec predictions could be used in a **semi-automated way** for Thesaurus Extension

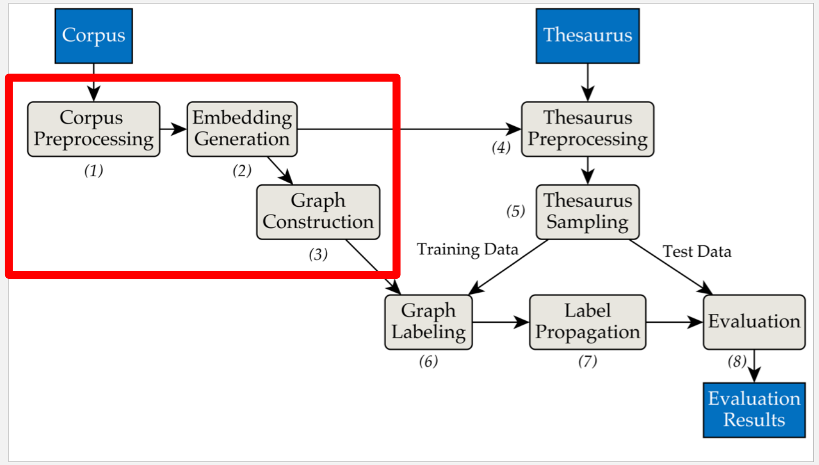
And: We contributed to the problem area

Conclusion: Contributions & Future Work

Contributions

- Created Open Source **„ThesaurusLabelPropagation“ tool**
  - Found implementation issues around label propagation in „scikit-learn“ (32.000 stars)
  - Significantly optimized performance for graph construction on word emebeddings
- Conducted **multiple hyper-parameter studies** (>1000 individual runs) & optimized configurations
- Rated configurations within **5 qualitative evaluations** (overall 2,500 suggst. manually rated)
  - Identification of influence factors for quality of suggestion results
  - Classification of typical thesaurus challenges
- Introduced & evaluated **new baseline** approach

Future Work with regards to Label Propagation



- Evaluation with a corpus in a different language and/or more training data?
- Evaluation within a different application area besides tax law?
- Augment word embeddings with other semantic knowledge, e.g. Wikidata, Wikipedia, Freebase

References

Buschmann, Frank, Regine Meunier, Hans Rohnert, Peter Sommerlad, and Michael Stal. 1996. "A System of Patterns: Pattern-Oriented Software Architecture."

Dirschl, Christian. 2016. "Thesaurus Generation and Usage at Wolters Kluwer Deutschland GmbH." *Jusletter IT 25. Februar 2016*, February.

Landthaler, Jörg, Bernhard Waltl, Dominik Huth, Daniel Braun, Christoph Stocker, Thomas Geiger, and Florian Matthes. 2017. "Extending Thesauri Using Word Embeddings and the Intersection Method." In *Proceedings of 2nd Workshop on Automated Semantic Analysis of Information in Legal Texts*. London, UK.

Ravi, Sujith, and Qiming Diao. 2015. "Large Scale Distributed Semi-Supervised Learning Using Streaming Approximation." *ArXiv:1512.01752 [Cs]*, December. http://arxiv.org/abs/1512.01752.

Rothe, Sascha, and Hinrich Schütze. 2015. "AutoExtend: Extending Word Embeddings to Embeddings for Synsets and Lexemes." In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 1:1793–1803.

Master’s Student Informatics

Markus Müller
www.muellermarkus.com

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

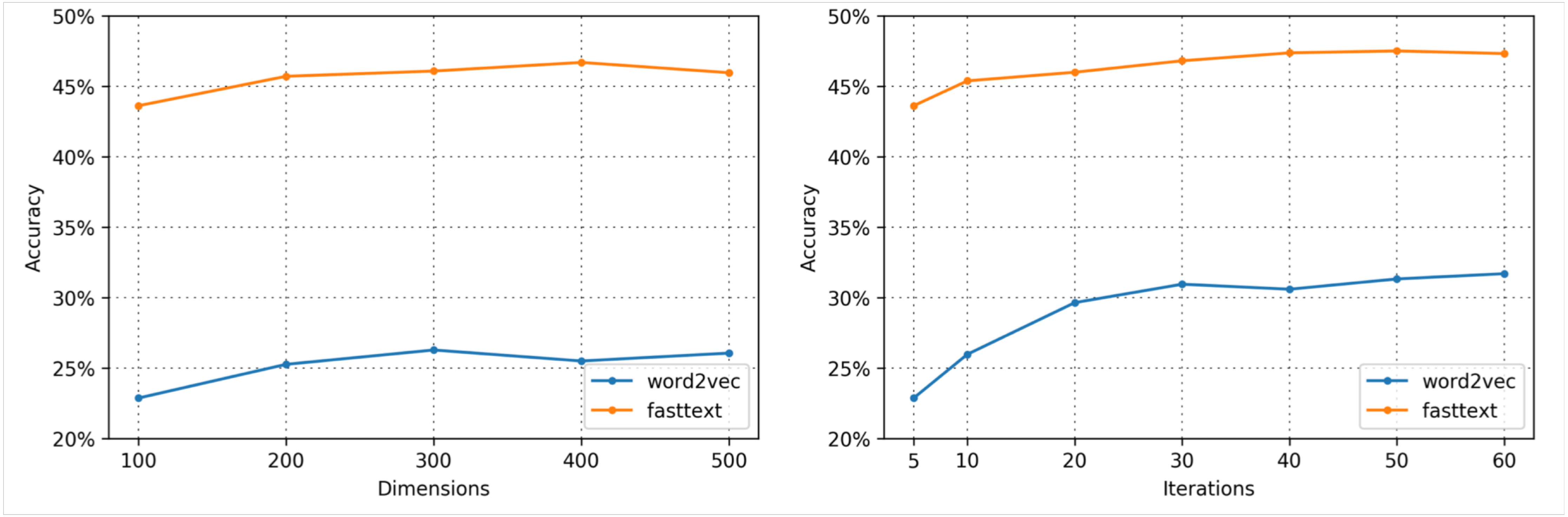
Boltzmannstraße 3
85748 Garching bei München

Tel +49.89.289.17132
Fax +49.89.289.17136

mail@muellermarkus.com
wwwmatthes.in.tum.de

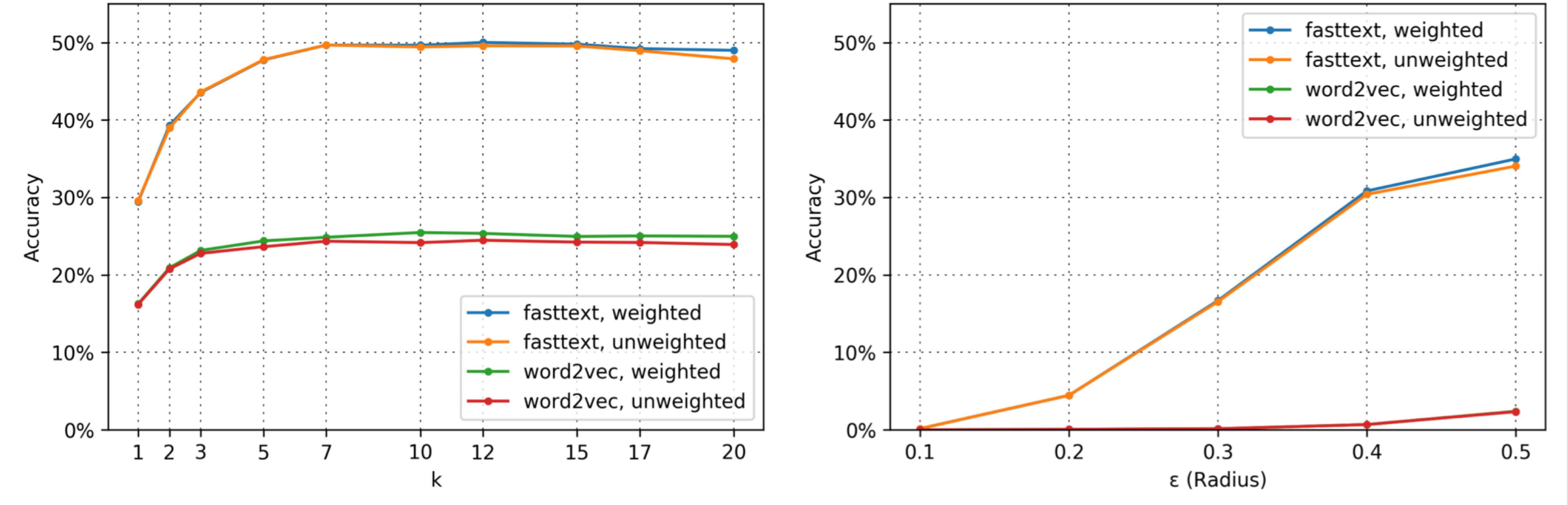
Backup

Hyper-Parameter Study on Word Embeddings



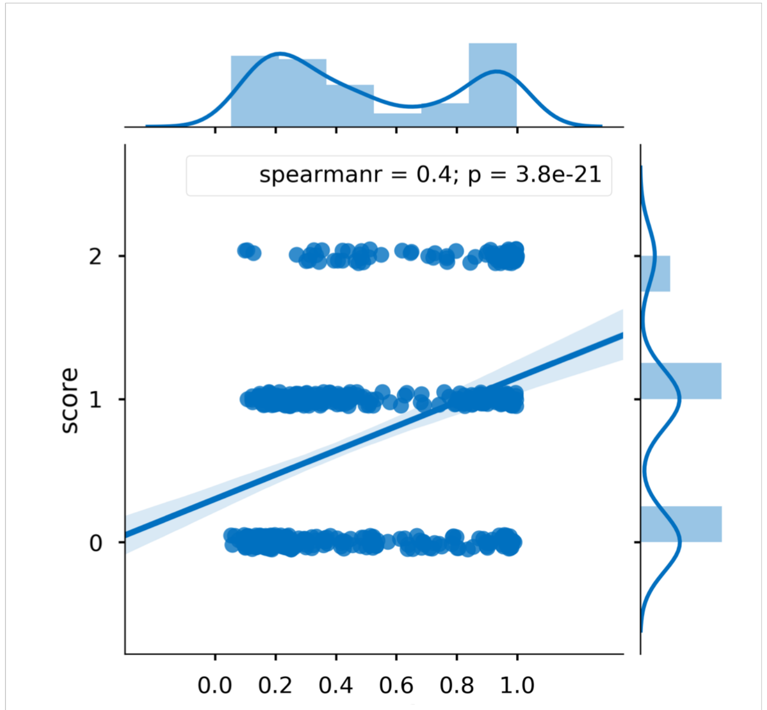
Backup

Hyper-Parameter Study on Graph Construction

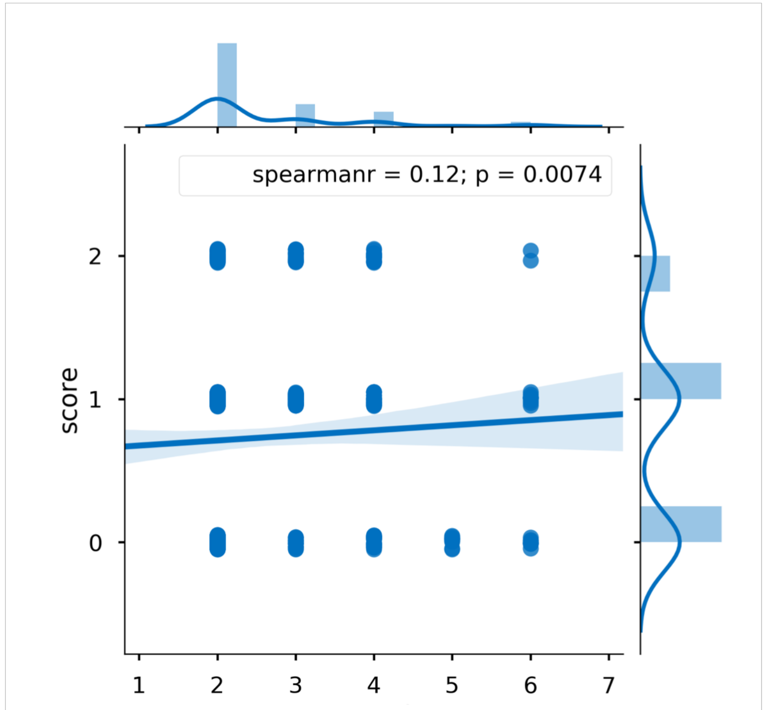


Backup

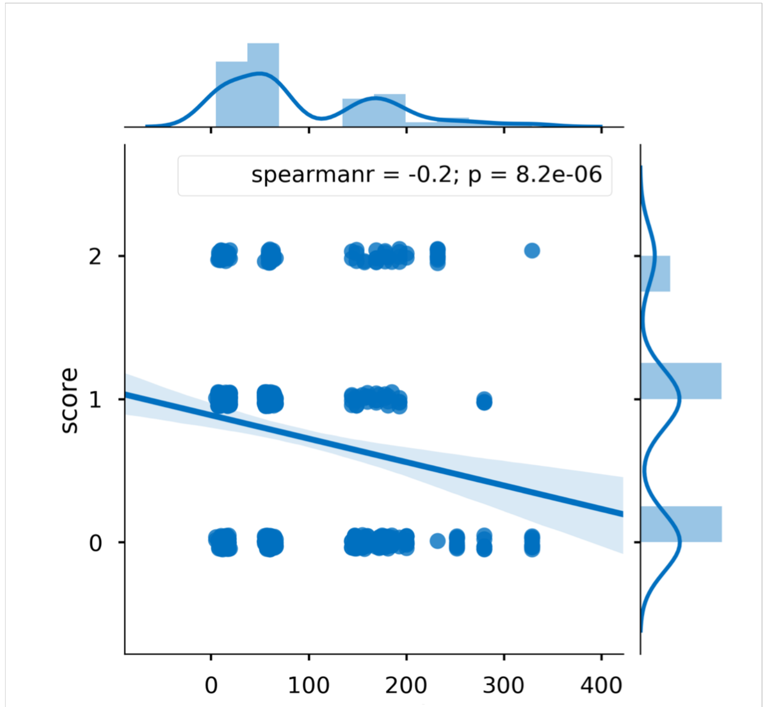
Qualitative Evaluation: Correlations



Prediction Confidence



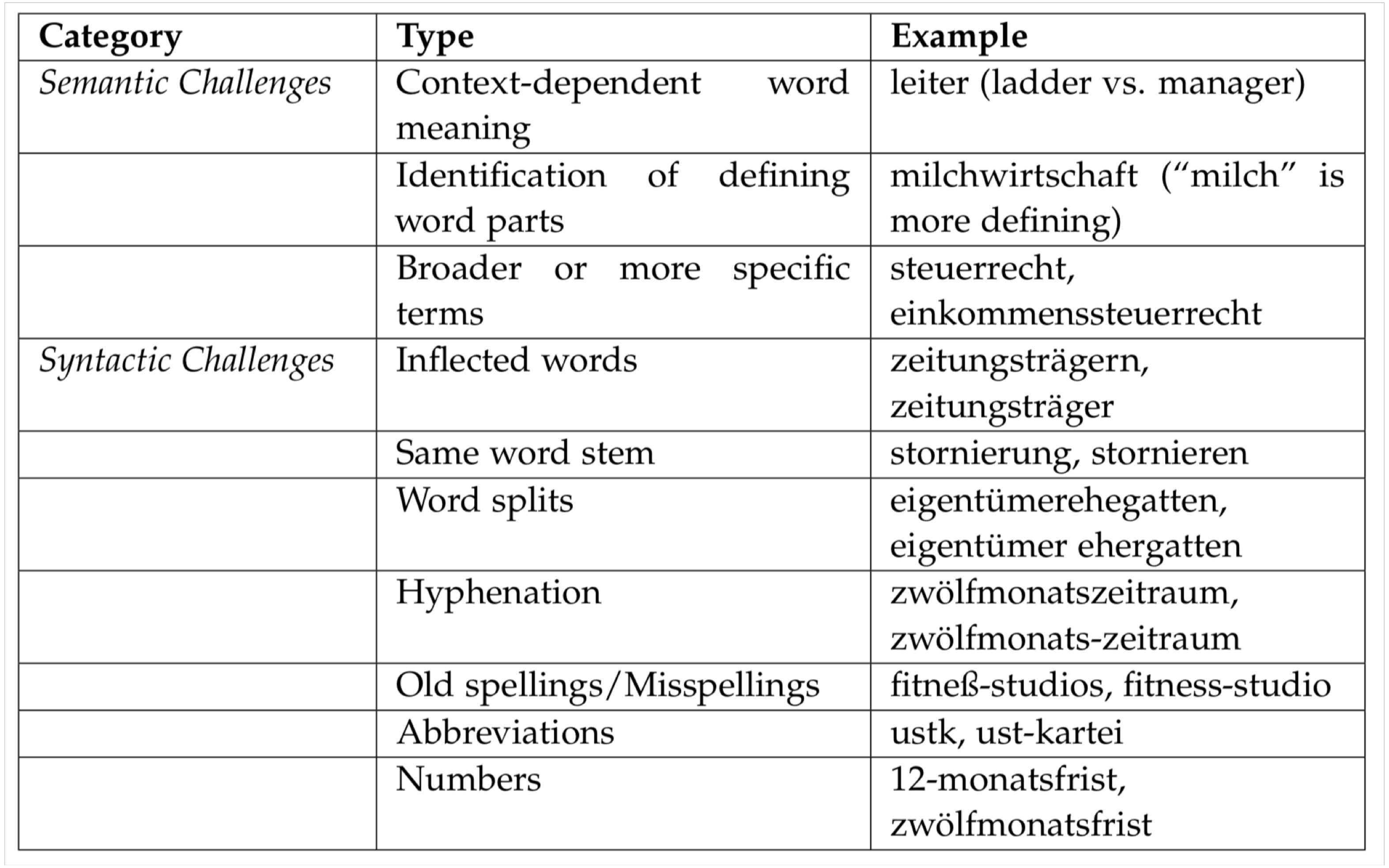
Number of Training Words in Synset



Number of Predictions for Synset

Backup

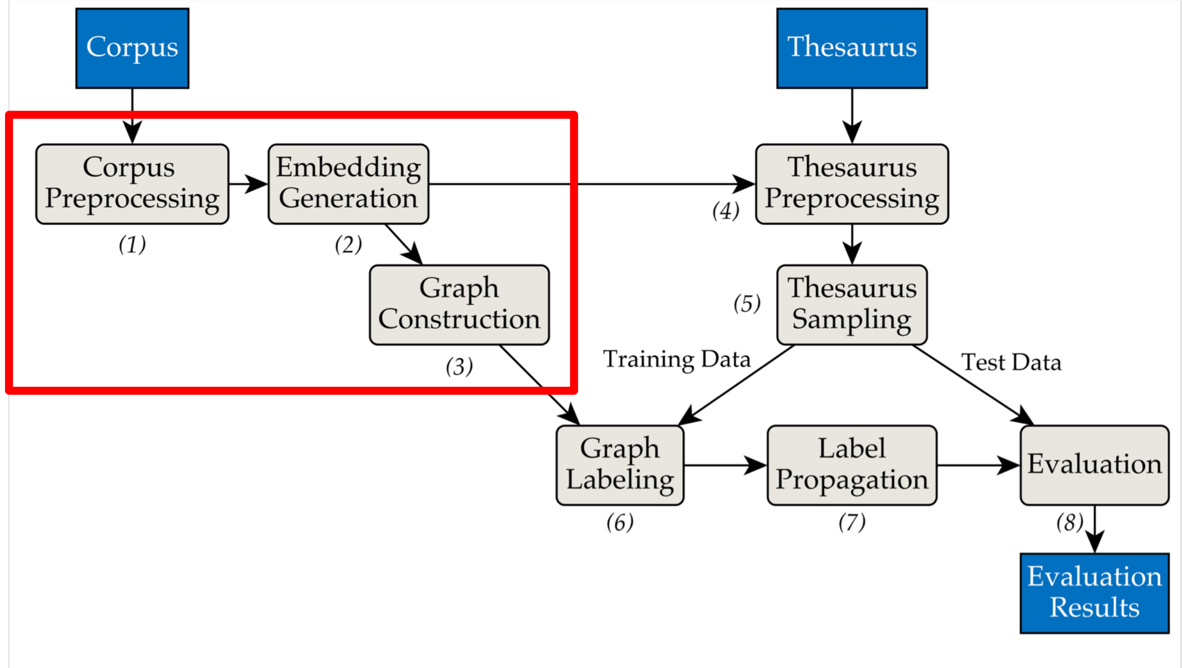
Challenges around Thesaurus Extension



Category	Type	Example
Semantic Challenges	Context-dependent word meaning	leiter (ladder vs. manager)
	Identification of defining word parts	milchwirtschaft ("milch" is more defining)
	Broader or more specific terms	steuerrecht, einkommenssteuerrecht
Syntactic Challenges	Inflected words	zeitungsträgern, zeitungsträger
	Same word stem	stornierung, stornieren
	Word splits	eigentümerehegatten, eigentümer ehergatten
	Hyphenation	zwölfmonatszeitraum, zwölfmonats-zeitraum
	Old spellings/Misspellings	fitneß-studios, fitness-studio
	Abbreviations	ustk, ust-kartei
	Numbers	12-monatsfrist, zwölfmonatsfrist

Backup

Possible Reasons and Future Work



Language & Training Data
Evaluation with a corpus in a different language and/or more training data?

Context of Tax Law
Evaluation within a different application area?

Graph Type
Augment word embeddings with other semantic knowledge, e.g. Wikidata, Wikipedia, Freebase [3]

Backup

Supervised, Semi-Supervised, Transductive

Supervised learning: Learn on labeled training instances, perform prediction on unknown test data.

Inductive semi-supervised learning: Learn on labeled training instances and unlabeled training instances, perform prediction on unknown test data.

Transductive semi-supervised learning: Learn on labeled training instances and unlabeled training instances, perform prediction on known test [=training] data.