Karlsruher Institut für Technologie

Bachelorarbeit: Worteinbettungen für die Anforderungsdomäne

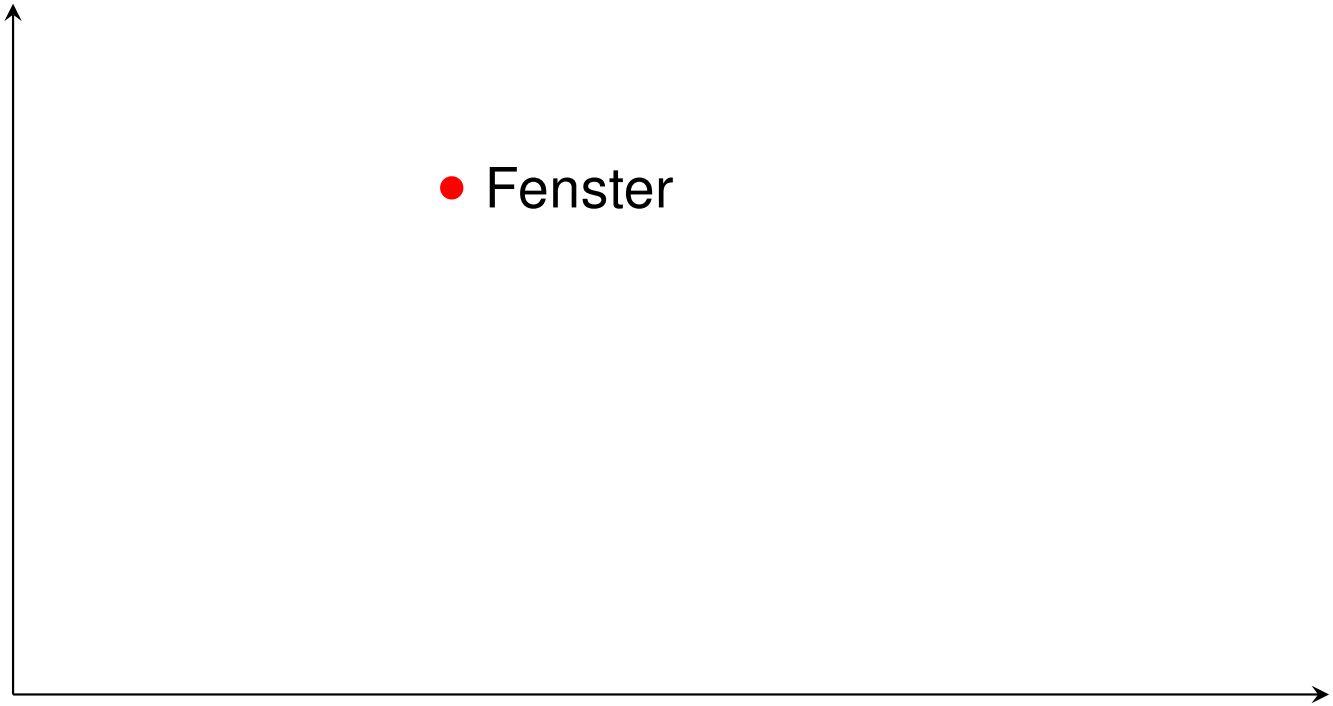
Tobias Telge, betreut von Tobias Hey

IPD Tichy, Fakultät für Informatik

Motivation

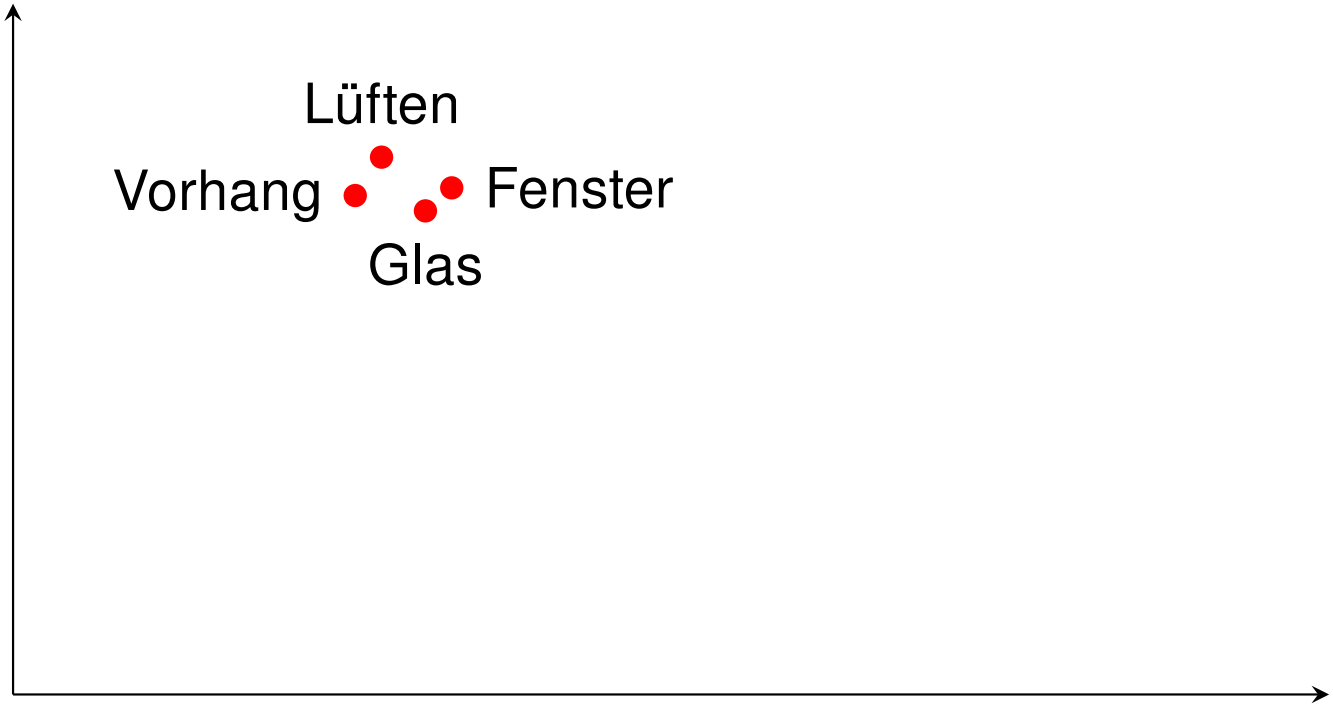
- Besonderheiten in der Anforderungsdomäne
  - Fachwörter
  - Andere Bedeutungen
- Beispiel: Das System muss in einem **Fenster** von 50ms das Ergebnis der **SQL**-Abfrage zurückgeben.
- Worteinbettungen bilden Wörter auf Wortvektoren ab
- Verwendung in der Verarbeitung natürlicher Sprache
- Vortrainiert auf generischen Texten

Motivation



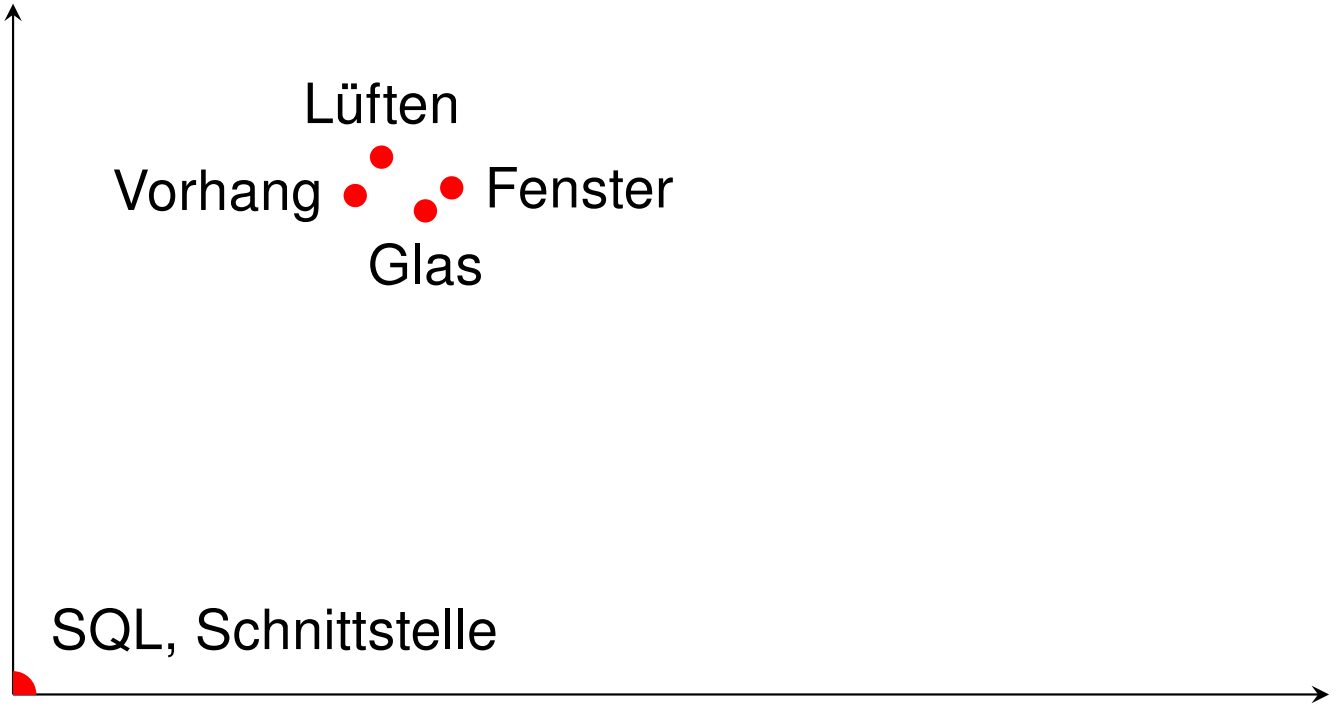
- Beispiel: Das System muss in einem **Fenster** von 50ms das Ergebnis der **SQL**-Abfrage zurückgeben.

Motivation



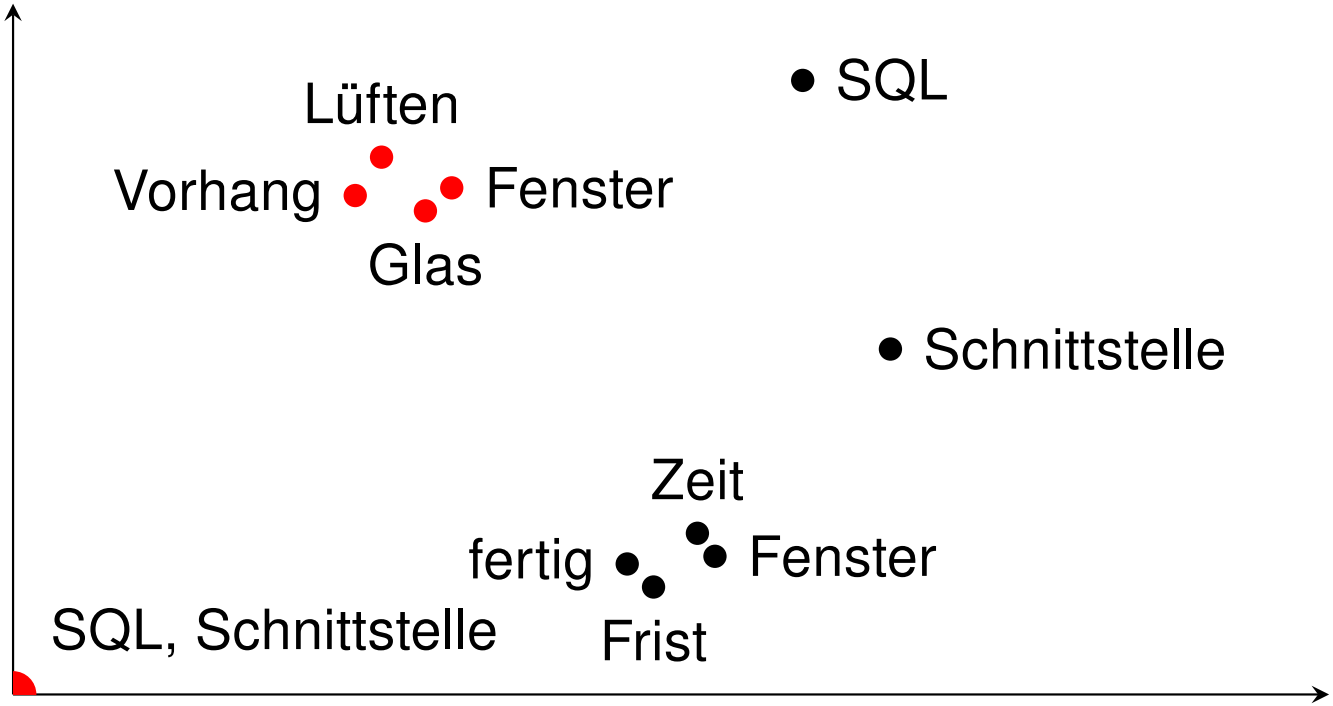
- Beispiel: Das System muss in einem **Fenster** von 50ms das Ergebnis der **SQL**-Abfrage zurückgeben.

Motivation



- Beispiel: Das System muss in einem **Fenster** von 50ms das Ergebnis der **SQL**-Abfrage zurückgeben.

Motivation



- Beispiel: Das System muss in einem **Fenster** von 50ms das Ergebnis der **SQL**-Abfrage zurückgeben.

Zielsetzung

- Ziel: Bildung von Worteinbettungen für die Anforderungsdomäne
- Aufbau eines Textkorpus aus Anforderungsdokumenten
  - Ergebnis: Textkorpus
- Worteinbettungsmodelle untersuchen, eins auswählen und trainieren
  - Ergebnis: Domänenspezifische Worteinbettungen
- Geeignete Kombination der domänenspezifischen Worteinbettungen mit generischen
  - Ergebnis: Domänenadaptierte Worteinbettungen

Verwandte Arbeiten

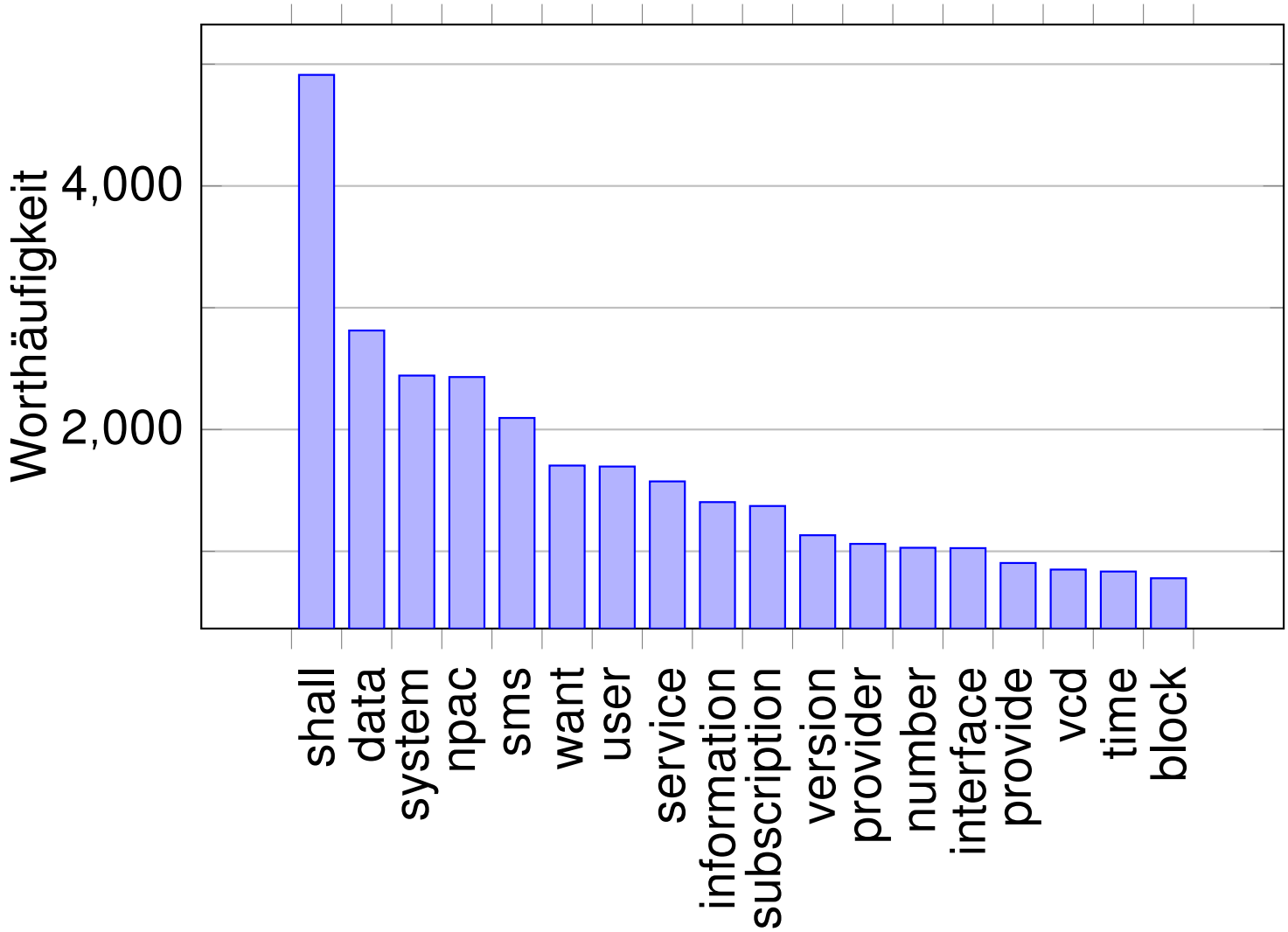
- Aufbau eines Textkorpus
  - Korpus PURE aus Anforderungsdokumenten [FSG17]
- Analyse von Worteinbettungen
  - Intrinsische und extrinsische Evaluationsmethoden [Bak18]
- Domänenspezifische Worteinbettungen
  - Softwareentwicklungsdomäne (Stack Overflow-Beiträge) [ECS18]
  - Patentdomäne [RK19]
- Kombination von Worteinbettungen
  - KCCA und anschließende Durchschnittsbildung [KSLS18]
  - Normalisierung und anschließende Durchschnittsbildung [CB18]

Aufbau des Textkorpus

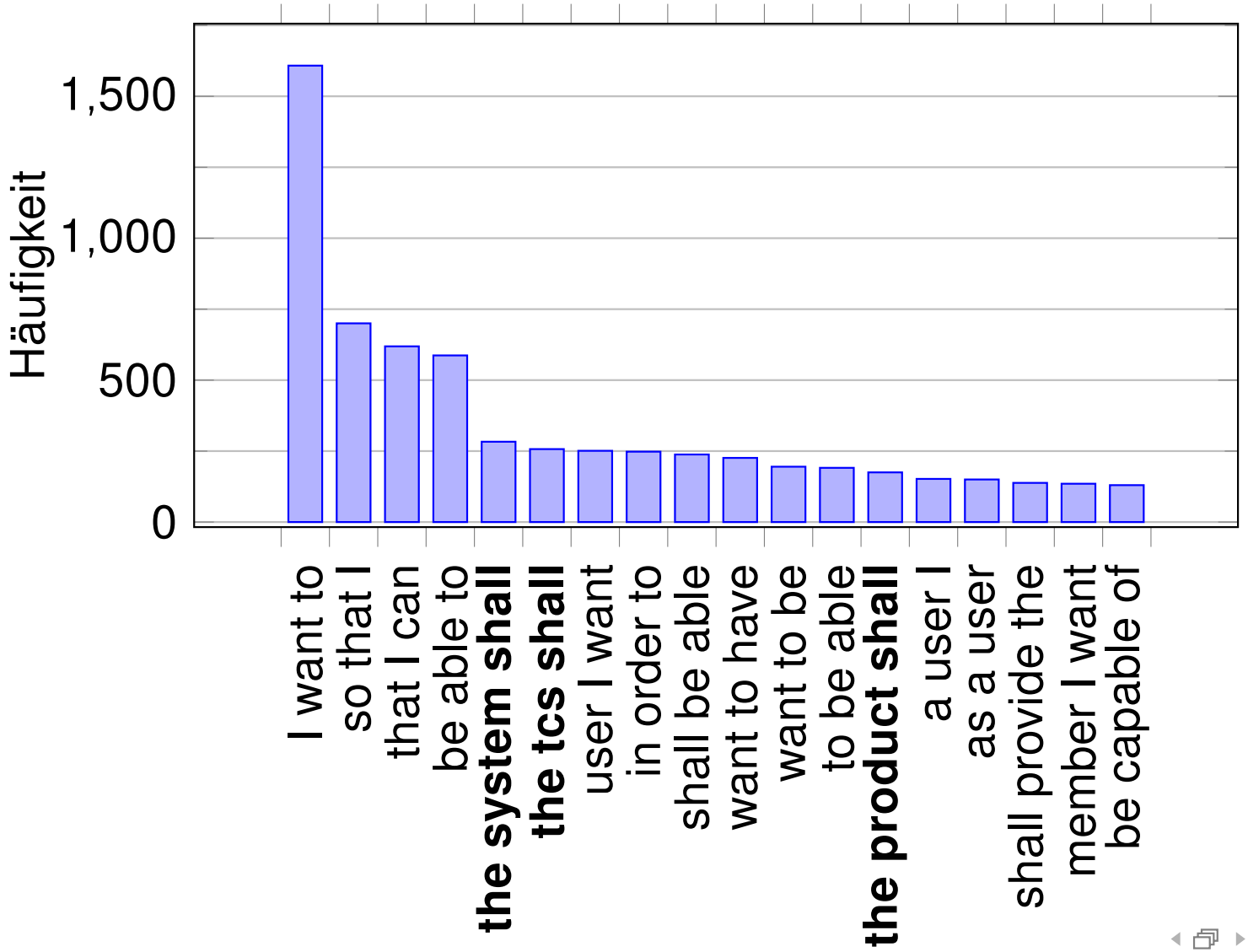
- Durch Google-Suche gefundene Datensätze
- `PURE`, `NFR`, `Risk`, `MODIS`, `CM1`, `EBT`, `GANNT`, `Ice Breaker`, `Dapliaz`
- Berücksichtigung von User Stories

Bestandteil	Anzahl
Projekte	65
Anforderungsbeschreibungen	21 458
User Stories	1680
Sätze	15 388
Wörter	346 256
Wörter ohne Stoppwörter	214 542
Vokabular	16 129

Häufigste Wörter ohne Stoppwörter



Häufigste Trigramme



Vorverarbeitung

- Kleinschreibung
- Entfernung von Rauschen
- Entfernung von Stoppwörtern
- Beispiel aus Korpus
  - The DPU-BOOT CSC shall include a DRAM BIT consisting of two write/read/compare tests.
  - dpu boot csc shall include dram bit consisting two write read compare tests

Worteinbettungsmodelle

- `word2vec`
  - Neuronales Netz trainiert an Pseudoaufgabe
  - `Continuous Bag-of-Words-Modell` und `Skip-Gram-Modell`
- `GloVe`
  - Kein Neuronales Netz
  - Berücksichtigung der globalen Kookurenz
- `ELMo`
  - Mehrschichtige LSTMs
  - Berücksichtigung der unterschiedlichen Bedeutungen eines Wortes in unterschiedlichen Kontexten
- `fastText`
  - Basiert auf `word2vec`
  - Berücksichtigung von Teilwörtern

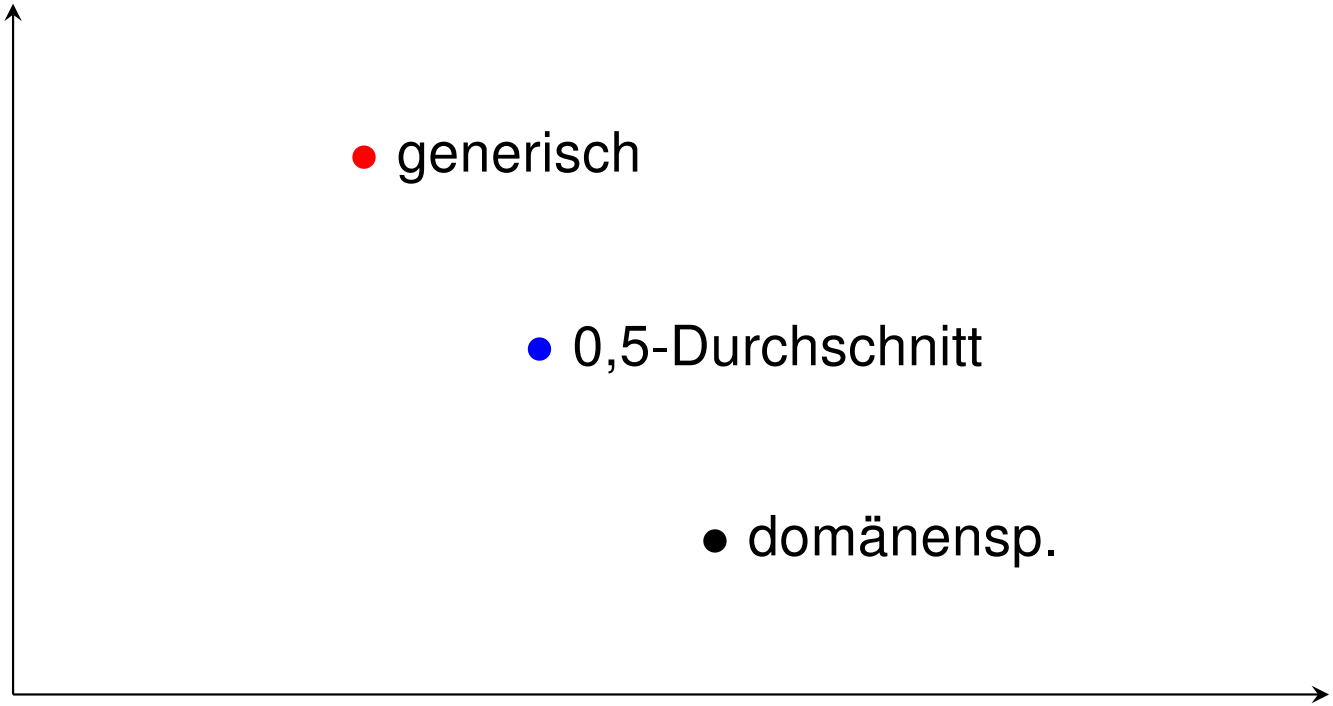
Worteinbettungsmodelle

- `word2vec`
  - Neuronales Netz trainiert an Pseudoaufgabe
  - `Continuous Bag-of-Words-Modell` und `Skip-Gram-Modell`
- `GloVe`
  - Kein Neuronales Netz
  - Berücksichtigung der globalen Kookurenz
- `ELMo`
  - Mehrschichtige LSTMs
  - Berücksichtigung der unterschiedlichen Bedeutungen eines Wortes in unterschiedlichen Kontexten
- `fastText`
  - Basiert auf `word2vec`
  - Berücksichtigung von Teilwörtern
- Auswahl: `fastText`
  - Bessere Darstellung seltener Wörter
  - Wortvektoren für nicht im Vokabular enthaltene Wörter

Kombination von Worteinbettungen

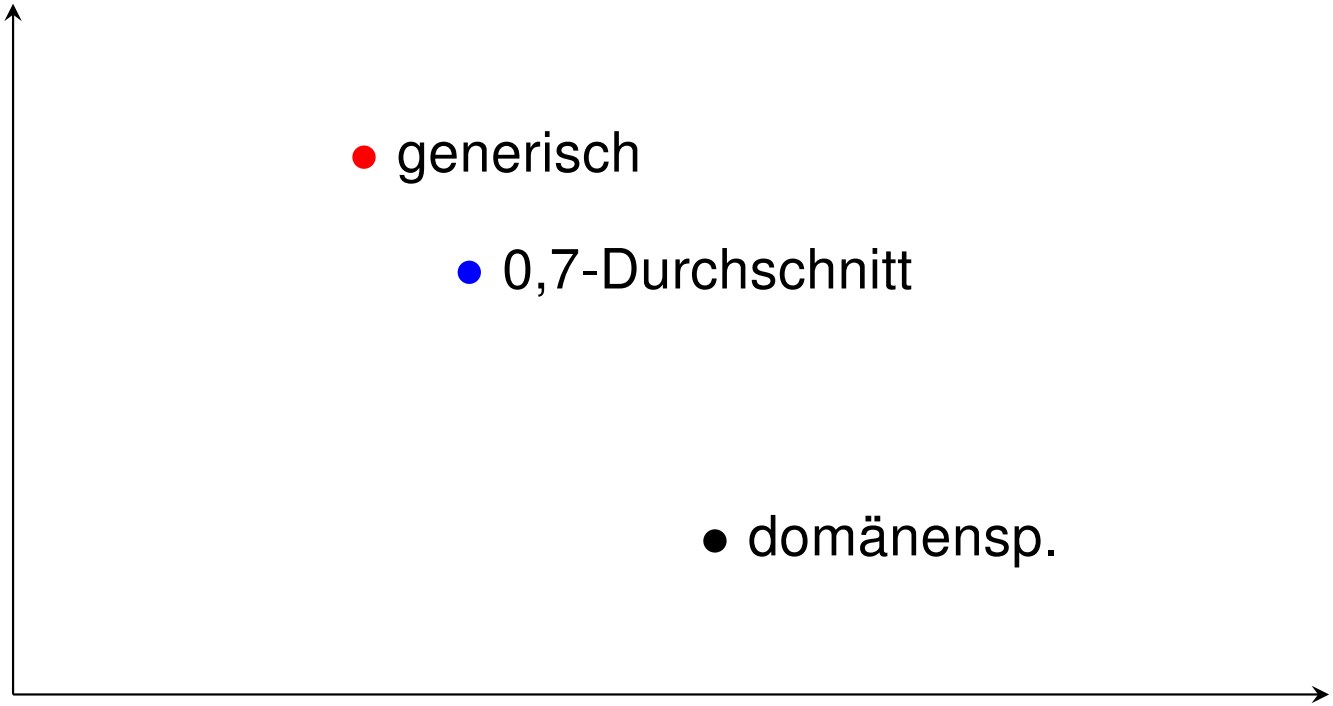
- Nutzung der Vorteile der domänenspezifischen und der generischen Worteinbettungen
  - Angepasstheit an Domäne
  - Umfang und größeres Wissen
- Kombination mit vortrainiertem generischen `fastText-Modell`
  - Auch Kombination von Teilwortvektoren
  - Trainiert auf `Common Crawl`-Korpus mit 630 Milliarden Wörtern
  - 2 Millionen Wortvektoren
- `PROJ`
  - Projektion durch KCCA
  - Bildung des gewichteten Durchschnittes der projizierten Vektoren
- `AVG`
  - L2-Normalisierung
  - Bildung des gewichteten Durchschnittes der normalisierten Vektoren
- Unterschiedliche Gewichtungen

Durchschnittsbildung



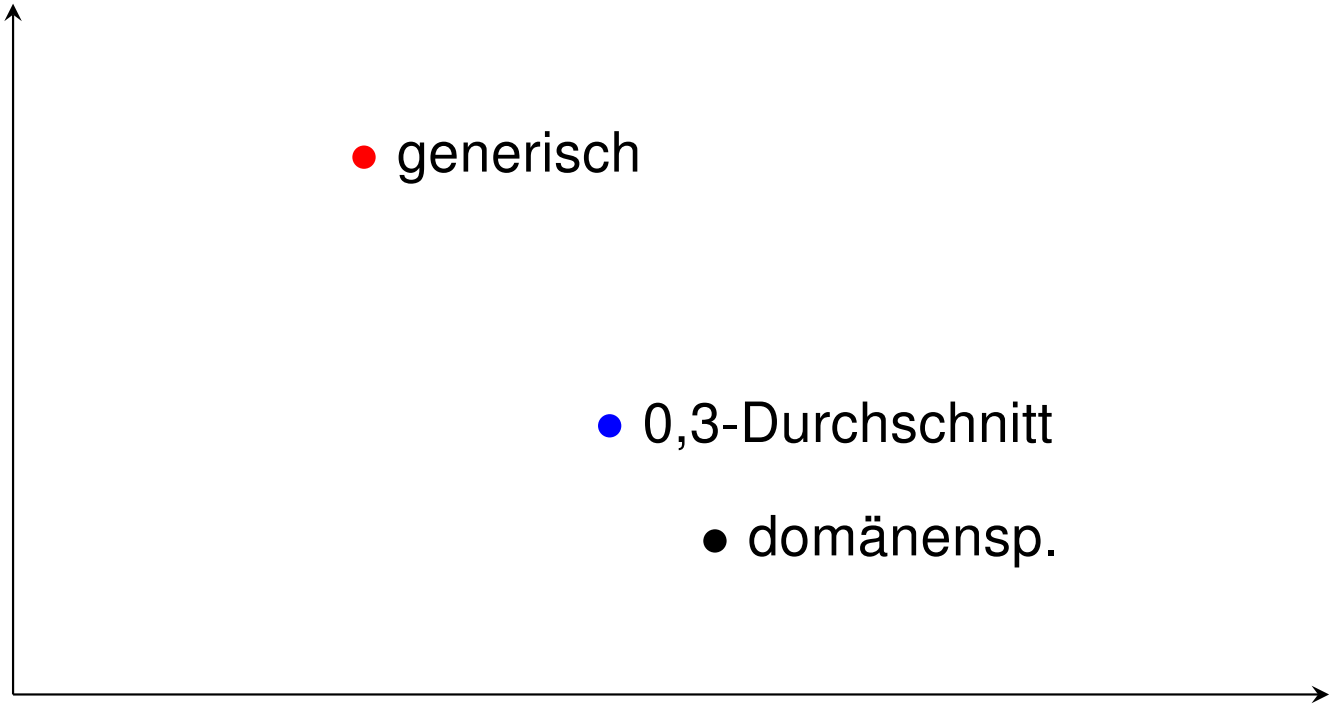
- Betrachtung von 3 Vektoren zu einem Wort

Durchschnittsbildung



- Betrachtung von 3 Vektoren zu einem Wort

Durchschnittsbildung

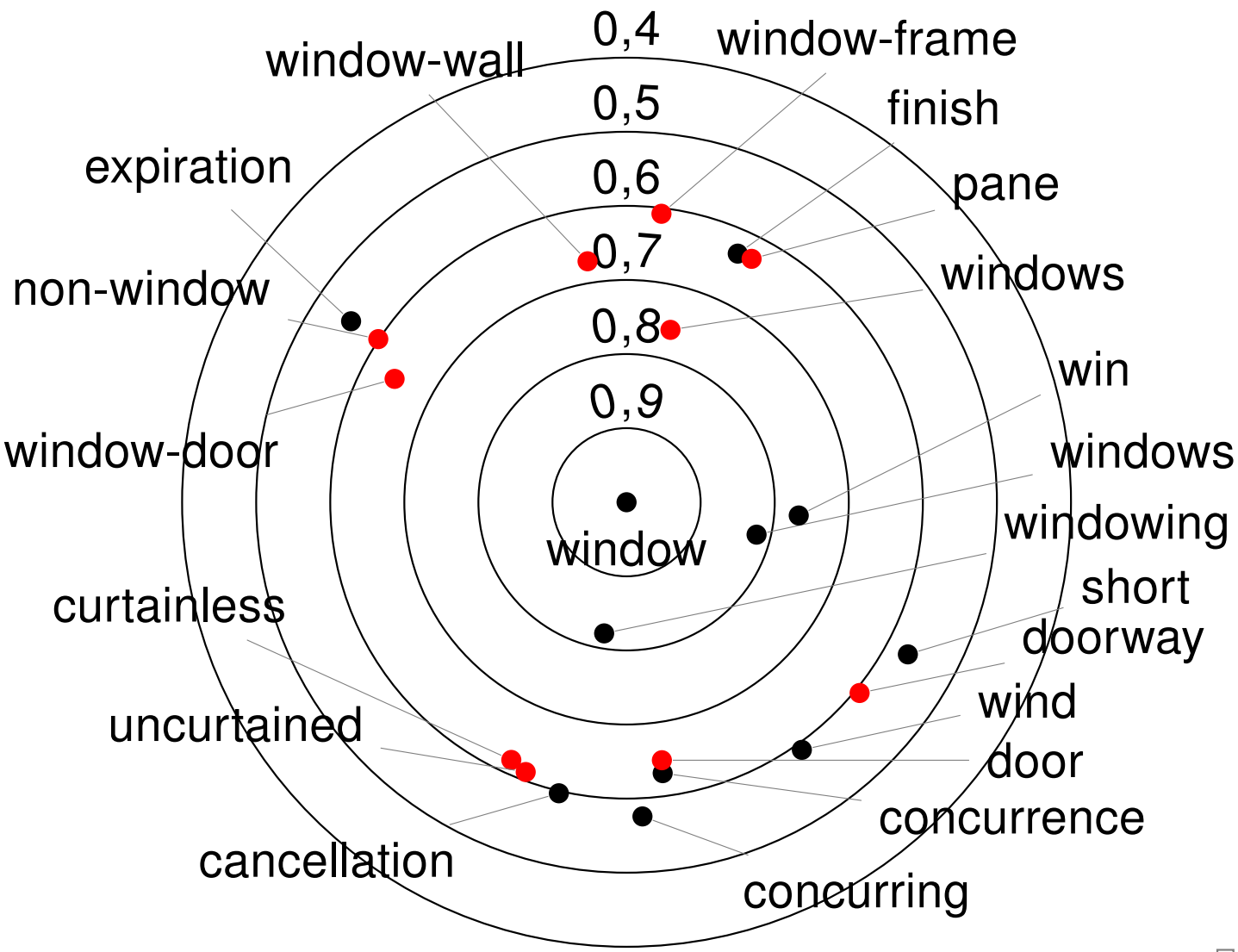


- Betrachtung von 3 Vektoren zu einem Wort

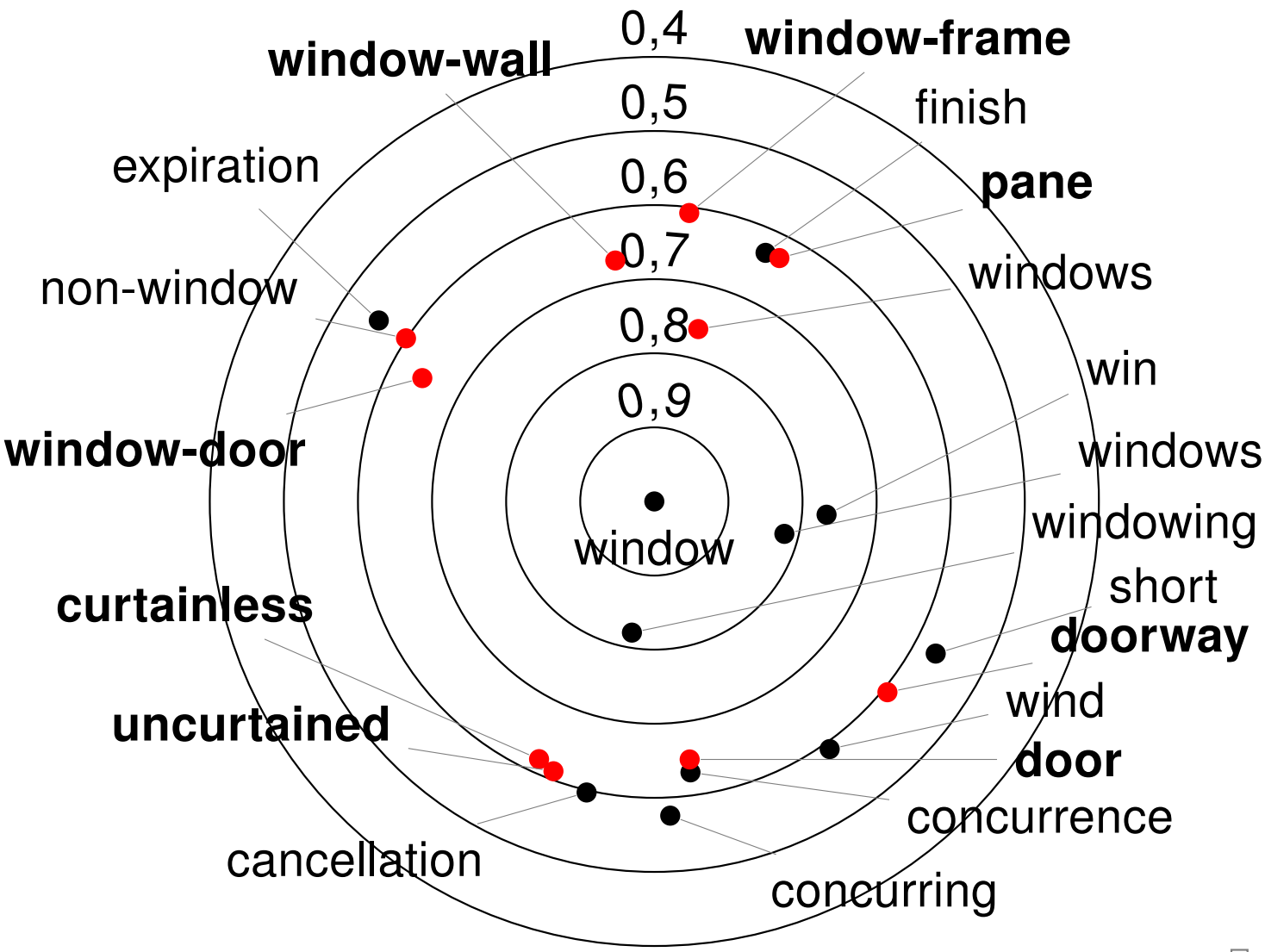
Evaluation

- Intrinsische Evaluation
  - Untersuchung von Wortähnlichkeiten
  - Vergleich domänenspezifisch und generisch
- Extrinsische Evaluation
  - Klassifizierungsaufgabe
  - Auswahl Kombinationsmethode und Gewichtung
  - Vergleich domänenspezifisch, domänenadaptiert und generisch

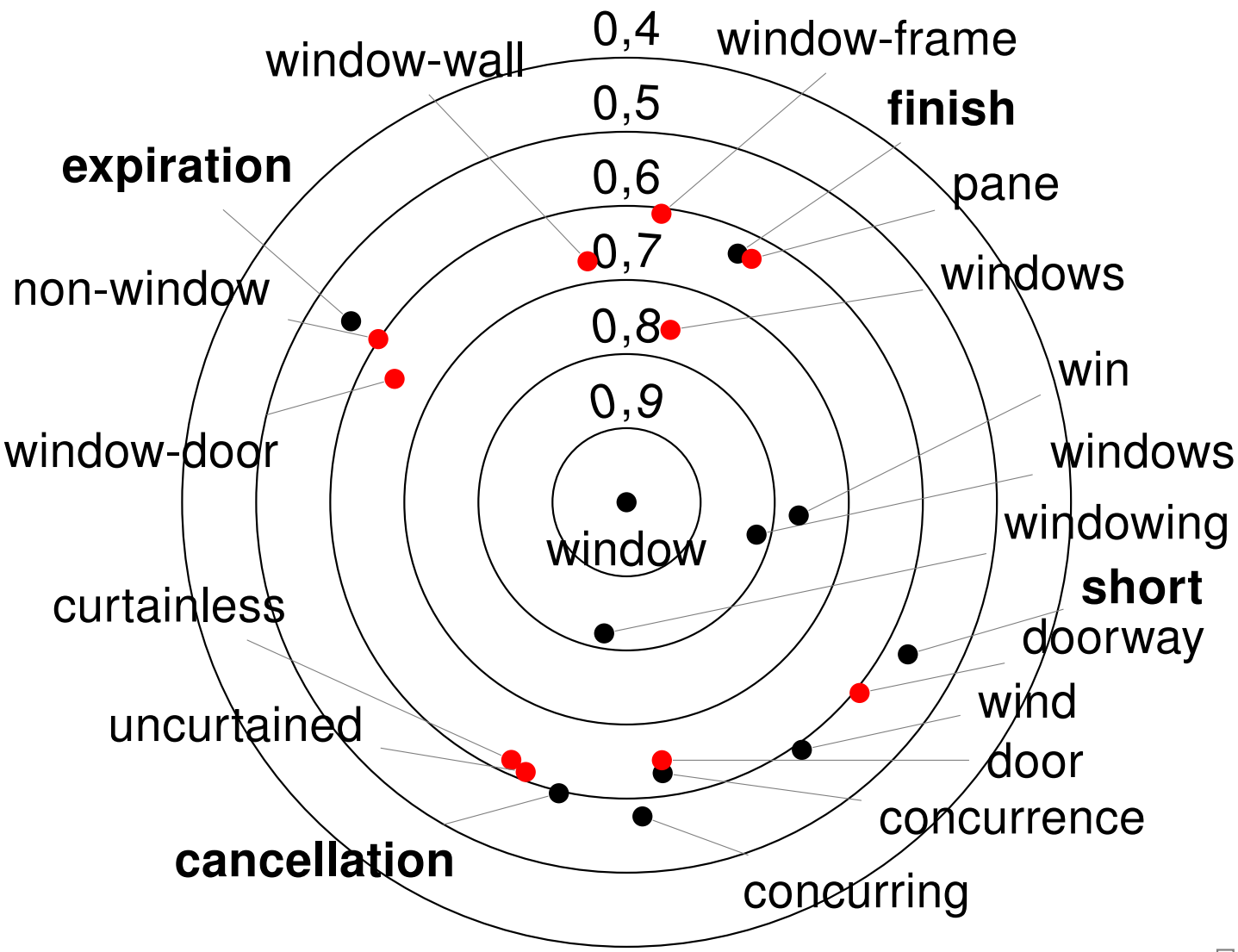
Intrinsische Evaluation



Intrinsische Evaluation



Intrinsische Evaluation



Extrinsische Evaluation

- Klassifizierung der semantischen Funktion von Anforderungen
- Klassen: *Aktion*, *Aggregation*, *Ereignis*, *Zustand*
- 700 annotierte Sätze aus 14 Projekten
- Klassifikatoren
  - Zufallswald
  - Stützvektormaschine
  - logistische Regression
  - LSTM
- Testverfahren
  - Training-Test-Teilung (80 zu 20)
  - Zehnfache Kreuzvalidierung
  - Projektspezifische Kreuzvalidierung

Generisch und Domänenspezifisch

Kl.	M.	T.-T.-Teilung gen.	T.-T.-Teilung sp.	Zehnf. Kreuzv. gen.	Zehnf. Kreuzv. sp.	Projektsp. Kreuzv. gen.	Projektsp. Kreuzv. sp.
ZW	G.	**0,68**	0,59	**0,61**	0,49	**0,62**	0,58
	F_1	**0,62**	0,54	**0,55**	0,45	**0,55**	0,51
SV	G.	**0,74**	0,66	**0,70**	0,57	**0,69**	0,56
	F_1	**0,73**	0,65	**0,70**	0,57	**0,69**	0,58
LR	G.	**0,77**	0,72	**0,71**	0,58	**0,70**	0,58
	F_1	**0,76**	0,72	**0,71**	0,59	**0,70**	0,60
LM	G.	0,72	**0,75**	**0,71**	0,67	**0,65**	0,62
	F_1	0,66	**0,74**	**0,67**	0,66	0,58	**0,62**

Ergebnisse verschiedener Methoden

		T.-T.-Teilung			Zehnf. Kreuzv.			Projektsp. Kreuzv.		
Kl.	M.	AVG	PROJ	gen.	AVG	PROJ	gen.	AVG	PROJ	gen.
ZW	G.	**0,68**	0,59	0,68	**0,61**	0,49	0,61	**0,66**	0,56	0,62
	F_1	**0,61**	0,53	0,62	**0,55**	0,45	0,55	**0,58**	0,50	0,55
SV	G.	**0,77**	0,68	0,74	**0,69**	0,59	0,70	**0,69**	0,60	0,69
	F_1	**0,77**	0,64	0,73	**0,69**	0,56	0,70	**0,69**	0,58	0,69
LR	G.	**0,81**	0,65	0,77	**0,74**	0,58	0,71	**0,71**	0,59	0,70
	F_1	**0,80**	0,62	0,76	**0,73**	0,55	0,71	**0,71**	0,57	0,70
LM	G.	**0,81**	0,63	0,72	**0,77**	0,56	0,71	**0,72**	0,60	0,65
	F_1	**0,79**	0,57	0,66	**0,76**	0,53	0,67	**0,69**	0,57	0,58

Ergebnisse verschiedener Gewichtungen

Kl.	M.	T.-T.-Teilung 0,3	T.-T.-Teilung 0,5	T.-T.-Teilung 0,7	Zehnf. Kreuzv. 0,3	Zehnf. Kreuzv. 0,5	Zehnf. Kreuzv. 0,7	Projektsp. Kreuzv. 0,3	Projektsp. Kreuzv. 0,5	Projektsp. Kreuzv. 0,7
ZW	G.	0,65	0,68	**0,71**	0,55	0,61	**0,65**	0,60	**0,66**	**0,66**
	F_1	0,58	0,61	**0,65**	0,50	0,55	**0,59**	0,53	**0,58**	**0,58**
SV	G.	0,74	0,77	**0,78**	0,67	0,69	**0,70**	0,64	0,69	**0,70**
	F_1	0,73	0,77	**0,78**	0,67	0,69	**0,70**	0,65	0,69	**0,70**
LR	G.	0,80	**0,81**	0,78	0,68	**0,74**	0,73	0,69	0,71	**0,72**
	F_1	0,80	**0,81**	0,78	0,68	**0,73**	**0,73**	0,69	0,71	**0,72**
LM	G.	0,80	0,81	**0,83**	0,71	0,77	**0,78**	0,70	0,72	**0,75**
	F_1	0,77	0,79	**0,81**	0,70	0,76	**0,78**	0,68	0,69	**0,73**

Generisch und Domänenadaptiert

Kl.	M.	T.-T.-Teilung gen.	T.-T.-Teilung ad.	Zehnf. Kreuzv. gen.	Zehnf. Kreuzv. ad.	Projektsp. Kreuzv. gen.	Projektsp. Kreuzv. ad.
ZW	G.	0,68	**0,71**	0,61	**0,65**	0,62	**0,66**
	F_1	0,62	**0,65**	0,55	**0,59**	0,55	**0,58**
SV	G.	0,74	**0,78**	**0,70**	**0,70**	0,69	**0,70**
	F_1	0,73	**0,78**	**0,70**	**0,70**	0,69	**0,70**
LR	G.	0,77	**0,78**	0,71	**0,73**	0,70	**0,72**
	F_1	0,76	**0,77**	0,71	**0,73**	0,70	**0,72**
LM	G.	0,72	**0,83**	0,71	**0,78**	0,65	**0,75**
	F_1	0,66	**0,81**	0,67	**0,78**	0,58	**0,73**

Fazit

- Ziel: Bildung von Worteinbettungen für die Anforderungsdomäne
- Ansatz
  - Aufbau eines Korpus aus Anforderungsdokumenten
  - Trainieren von `fastText` auf Korpus
  - Kombination mit generischen Worteinbettungen durch Normalisierung und gewichteten Durchschnitt
- Verbesserung (F_1, LSTM):
  - T.-T.-Teilung: 0,15
  - Zehnf. Kreuzv.: 0,11
  - Projektsp. Kreuzv.: 0,15
- Ausblick
  - Extrinsische Evaluation auf weiteren Aufgaben
  - Erweiterung des Textkorpus
  - Feineinstellung der Gewichtungen

Literatur I

Bakarov, Amir (Jan. 2018). “A Survey of Word Embeddings Evaluation Methods”. In:

Coates, Joshua und Danushka Bollegala (Juni 2018). “Frustratingly Easy Meta-Embedding – Computing Meta-Embeddings by Averaging Source Word Embeddings”. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, S. 194–198. DOI: 10.18653/v1/N18-2031.

Efstathiou, Vasiliki, Christos Chatzilenas und Diomidis Spinellis (2018). “Word Embeddings for the Software Engineering Domain”. en. In: *Proceedings of the 15th International Conference on Mining Software Repositories - MSR ’18*. Gothenburg, Sweden: ACM Press, S. 38–41. ISBN: 978-1-4503-5716-6. DOI: 10.1145/3196398.3196448.

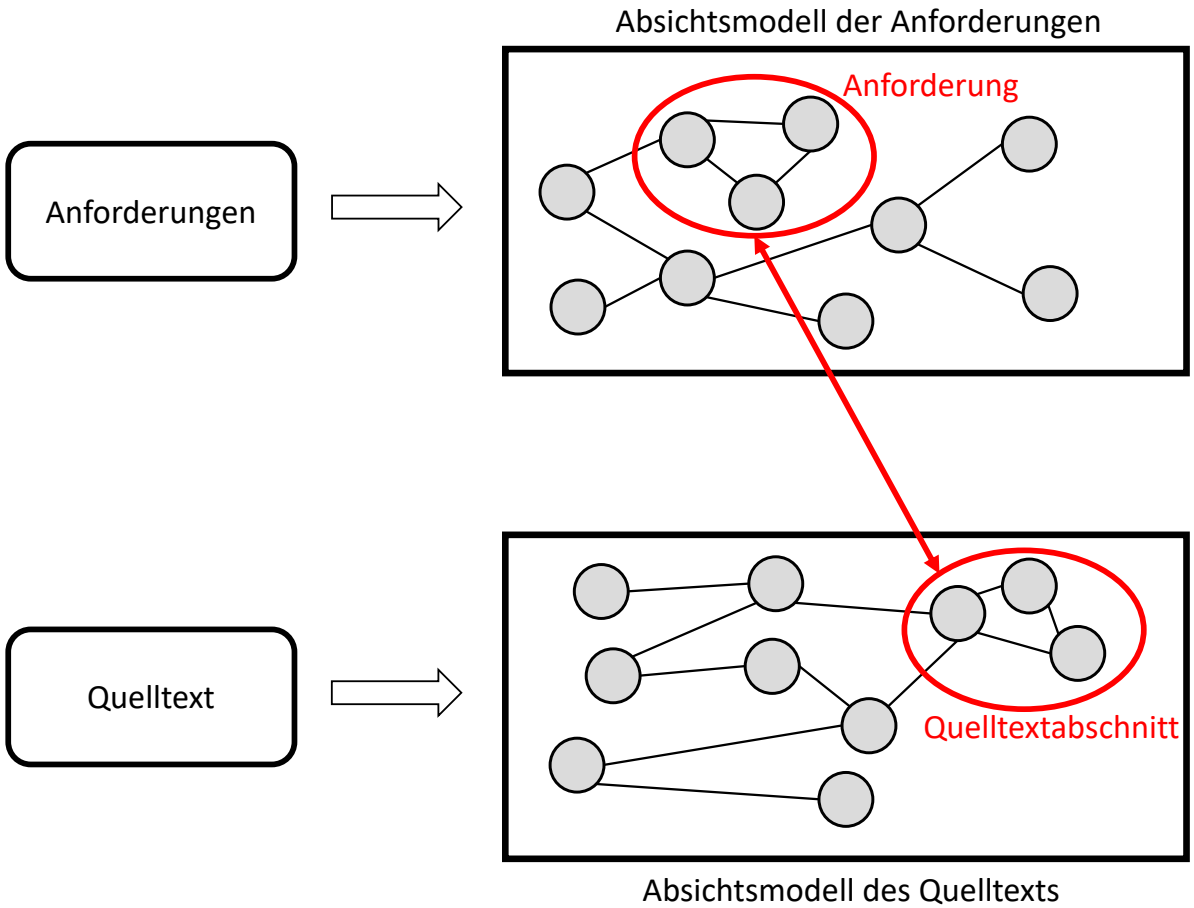
Literatur II

Ferrari, Alessio, Giorgio Oronzo Spagnolo und Stefania Gnesi (Sep. 2017). “PURE: A Dataset of Public Requirements Documents”. In: *2017 IEEE 25th International Requirements Engineering Conference (RE)*. Lisbon, Portugal: IEEE, S. 502–505. ISBN: 978-1-5386-3191-1. DOI: 10.1109/RE.2017.29.

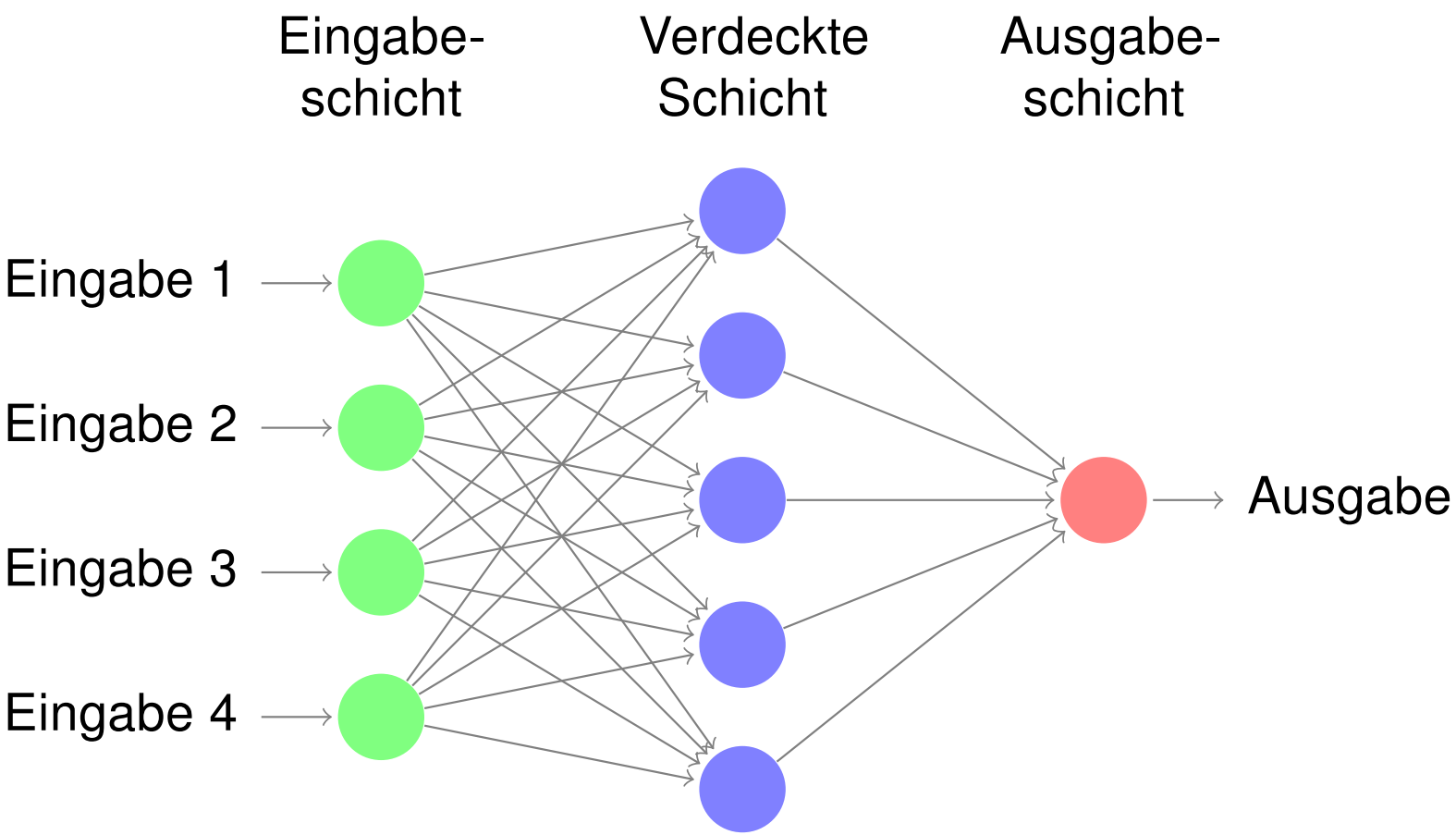
Kameswara Sarma, Prathusha, Yingyu Liang und Bill Sethares (2018). “Domain Adapted Word Embeddings for Improved Sentiment Classification”. en. In: *Proceedings of the Workshop on Deep Learning Approaches for Low-Resource NLP*. Melbourne: Association for Computational Linguistics, S. 51–59. DOI: 10.18653/v1/W18-3407.

Risch, Julian und Ralf Krestel (Feb. 2019). “Domain-Specific Word Embeddings for Patent Classification”. en. In: *Data Technologies and Applications* 53.1, S. 108–122. ISSN: 2514-9288. DOI: 10.1108/DTA-01-2019-0002.

INDIRECT


Absichtsmodell der Anforderungen
Anforderung
Anforderungen
Quelltext
Quelltextabschnitt
Absichtsmodell des Quelltexts

Neuronales Netz



Intrinsische Evaluation

Wort	Ähnl.	Diff.
windowing	0,84	0,33
windows	0,83	0,05
win	0,79	0,50
finish	0,66	0,36
concurrence	0,66	0,44
cancellation	0,63	0,36
wind	0,62	0,22
concurring	0,61	0,46
short	0,60	0,23
expiration	0,59	0,26

Wort	Ähnl.	Diff.
windows	0,78	$-0,05$
window-wall	0,70	$(-0,24)$
door	0,68	$(0,38)$
window-door	0,67	$(-0,28)$
pane	0,66	$(0,35)$
curtainless	0,65	$(0,25)$
uncurtained	0,64	$(0,31)$
window-frame	0,64	$(-0,25)$
non-window	0,63	$(-0,34)$
doorway	0,63	$(0,39)$

Ergebnisse verschiedener Methoden

Kl.	M.	T.-T.-Teilung AVG	T.-T.-Teilung PROJ	Zehnf. Kreuzv. AVG	Zehnf. Kreuzv. PROJ	Projektsp. Kreuzv. AVG	Projektsp. Kreuzv. PROJ
ZW	G.	**0,68**	0,59	**0,61**	0,49	**0,66**	0,56
	F_1	**0,61**	0,53	**0,55**	0,45	**0,58**	0,50
SV	G.	**0,77**	0,68	**0,69**	0,59	**0,69**	0,60
	F_1	**0,77**	0,64	**0,69**	0,56	**0,69**	0,58
LR	G.	**0,81**	0,65	**0,74**	0,58	**0,71**	0,59
	F_1	**0,80**	0,62	**0,73**	0,55	**0,71**	0,57
LM	G.	**0,81**	0,63	**0,77**	0,56	**0,72**	0,60
	F_1	**0,79**	0,57	**0,76**	0,53	**0,69**	0,57

Vergleich der Ergebnisse

Kl.	M.	T.-T.-Teilung gen.	T.-T.-Teilung sp.	T.-T.-Teilung ad.	Zehnf. Kreuzv. gen.	Zehnf. Kreuzv. sp.	Zehnf. Kreuzv. ad.	Projektsp. Kreuzv. gen.	Projektsp. Kreuzv. sp.	Projektsp. Kreuzv. ad.
ZW	G.	0,68	0,59	**0,71**	0,61	0,49	**0,65**	0,62	0,58	**0,66**
	F_1	0,62	0,54	**0,65**	0,55	0,45	**0,59**	0,55	0,51	**0,58**
SV	G.	0,74	0,66	**0,78**	**0,70**	0,57	**0,70**	0,69	0,56	**0,70**
	F_1	0,73	0,65	**0,78**	**0,70**	0,57	**0,70**	0,69	0,58	**0,70**
LR	G.	0,77	0,72	**0,78**	0,71	0,58	**0,73**	0,70	0,58	**0,72**
	F_1	0,76	0,72	**0,77**	0,71	0,59	**0,73**	0,70	0,60	**0,72**
LM	G.	0,72	0,75	**0,83**	0,71	0,67	**0,78**	0,65	0,62	**0,75**
	F_1	0,66	0,74	**0,81**	0,67	0,66	**0,78**	0,58	0,62	**0,73**