Extracting Opinion Targets in a Single- and Cross-Domain Setting with Conditional Random Fields

Niklas Jakob und Iryna Gurevych

Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing

Überblick

- Einführung und Begriffsklärung
- Modelle
  - Zhuang et. al, 2006 (Baseline)
  - Jakob und Gurevych, 2010
- Datensets
- Experimente und Ergebnisse
  - Single-Domain
  - Cross-Domain
  - Vergleich
- Zusammenfassung

Opinion Targets und Opinion Expressions

Target

Etwas, worüber eine Meinung geäußert wird

Expression

Meinung, die geäußert wird

It gets great gas mileage, has a powerful engine, and a nice interior!

Aufgabe: Finde Targets und Expressions!

Opinion Targets und Opinion Expressions

Target

Etwas, worüber eine Meinung geäußert wird

Expression

Meinung, die geäußert wird

It gets **great** **gas mileage**, has a **powerful** **engine**, and a **nice** **interior**!

Aufgabe: Finde **Targets** und **Expressions**!

Jetzt: Target Extraction

Anwendungen

Opinion question answering

Was gefällt Nutzern an diesem Auto?

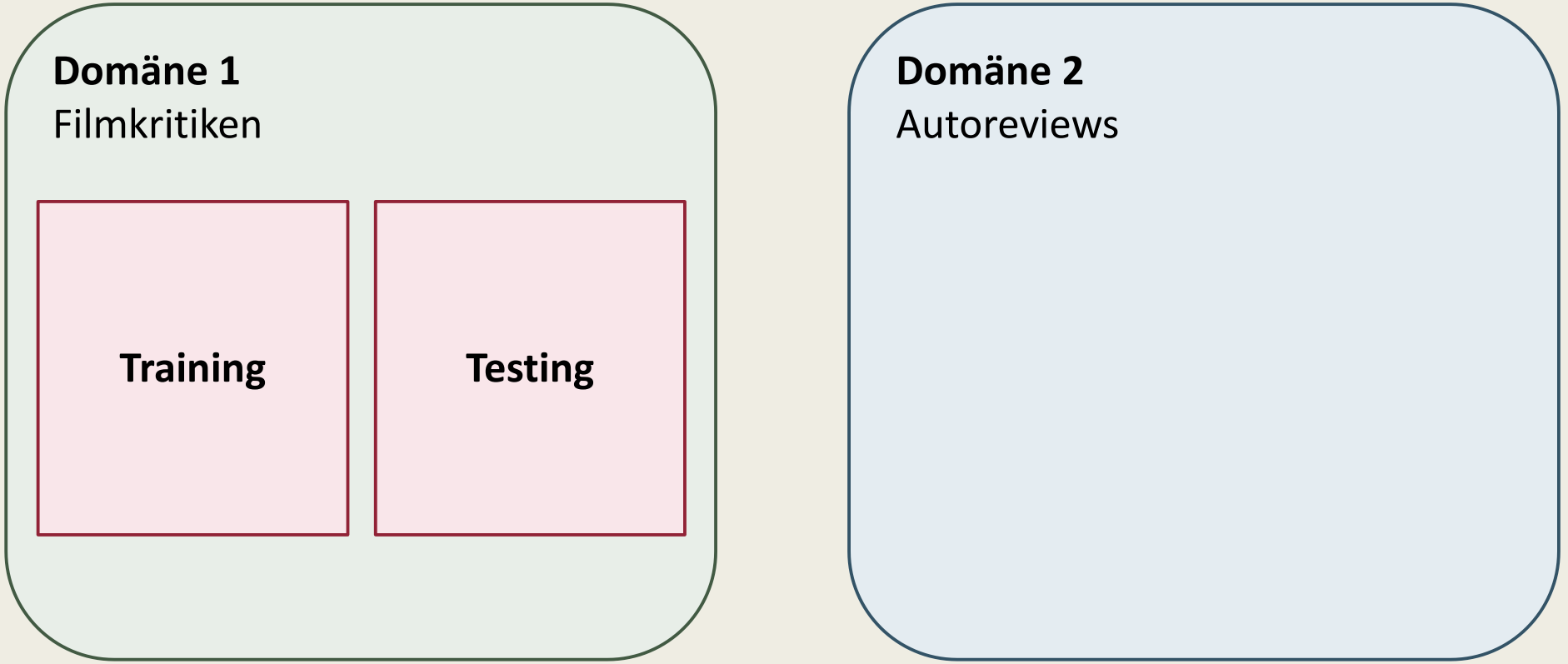
Opinion summarization

Überblick über Reviews von Spritverbrauch eines Autos

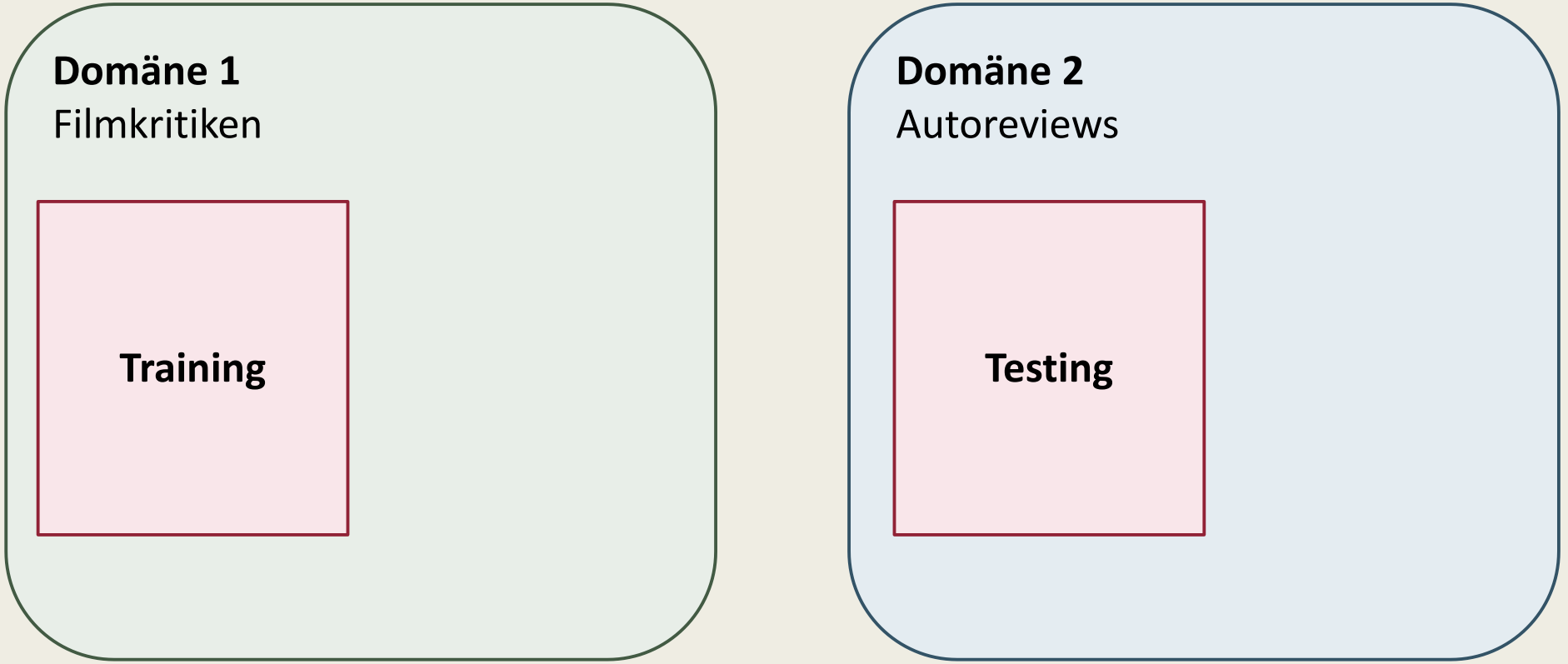
Recommender systems

Nur Empfehlungen von Autos, die einen guten Spritverbrauch haben

Single-Domain Settings



Cross-Domain Settings



Cross-Domain: Motivation und Probleme

The story is **simple** and boring.
Super **simple** to use!

- Gleiche Expression hat unterschiedliche Polarität in verschiedenen Domänen

Filme
story, actor, screenplay, soundtrack, …

Autos
gas mileage, speed, interior, …

- Verschiedene Domänen reden über verschiedene Targets!

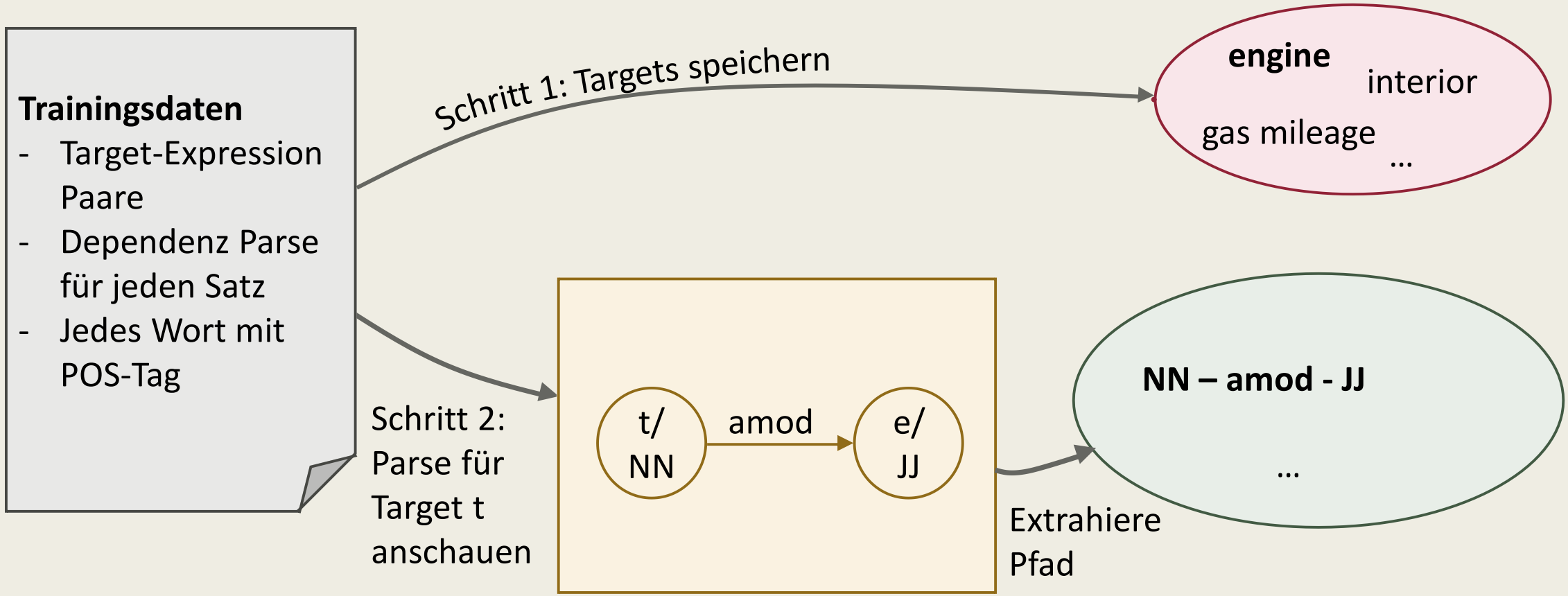
- Domänen mit weniger vorhandenen Daten

→ Ziel: „universelle“ Features von Targets finden

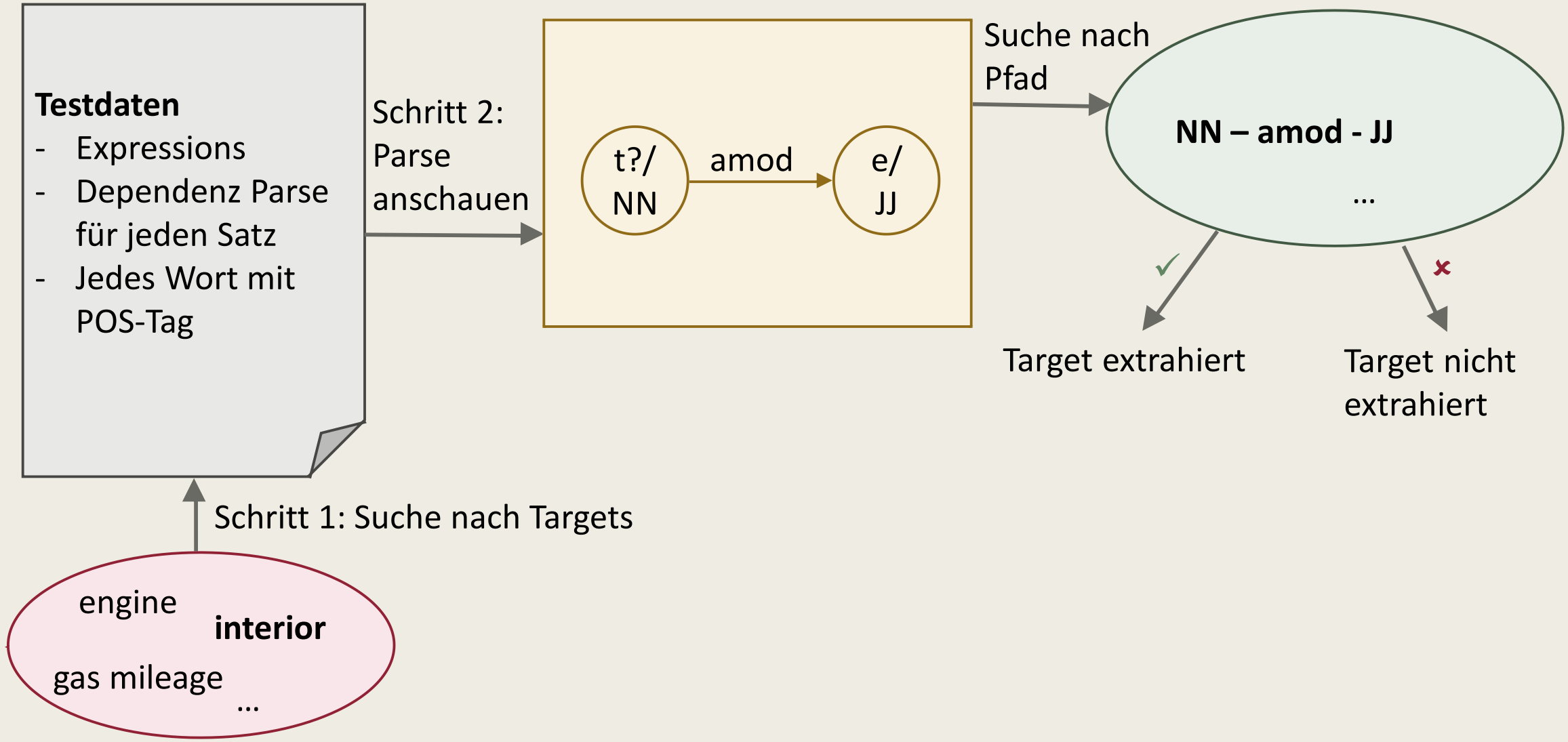
Baseline

- 2006 von Zhuang et. al vorgestellt
- Überwachter Algorithmus
- State-of-the-art auf movies Datenset
- Lernt
  1) Menge an Targets
  2) Menge an Dependenzpfaden zwischen Target und Expression

Baseline: Training

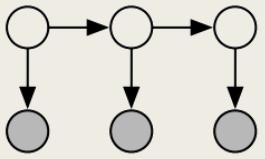


Baseline: Testing



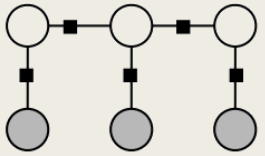
Conditional Random Fields (CRF)

- Linear-chain CRF
- Für uns: besseres HMM



HMMs

CONDITIONAL



Linear-chain CRFs

- Gegeben Beobachtungen/Features: Weise Tags zu
- IOB-Schema

Input:	It	has	a	*powerful*	engine
Output:	O	O	O	O	B

Features 1/2

Token String
Token als Feature

It

has

a

powerful

engine

.

POS-Tag
Part-of-speech Tag des Tokens

PRP_It

VBZ_has

DT_a

JJ_powerful

NN_engine

._.

Opinion Sentence
Token wird markiert, wenn es in einem Satz auftaucht, in dem eine Meinung geäußert wird

op_It

op_has

op_ a

op_powerful

op_engine

op_.

Features 2/2

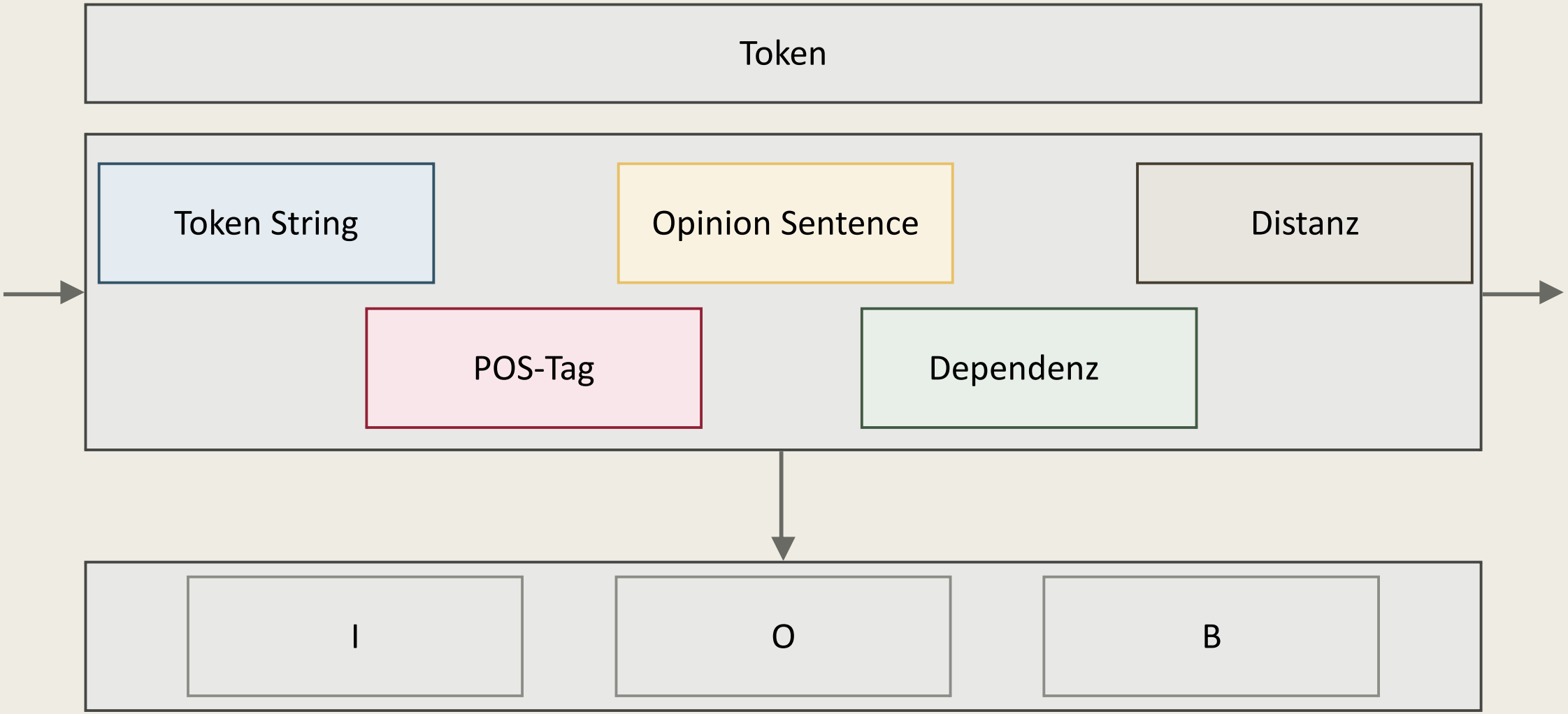
Dependenzpfad
Token wird markiert, wenn es eine direkte Dependenzrelation zur Expression hat

dep_engine

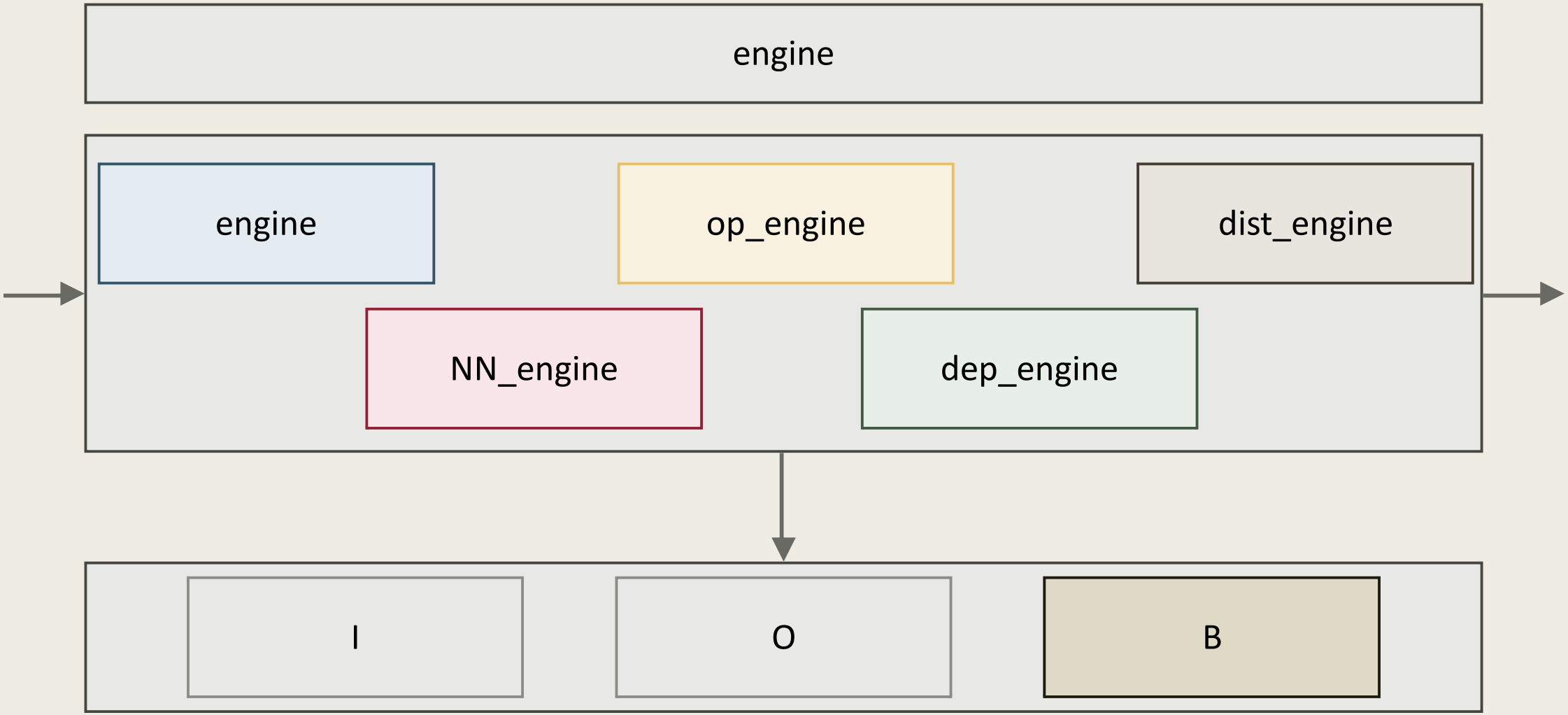
Distanz (Heuristik/Backoff)
NP, die am nächsten an der Expression steht, wird markiert

dist_engine

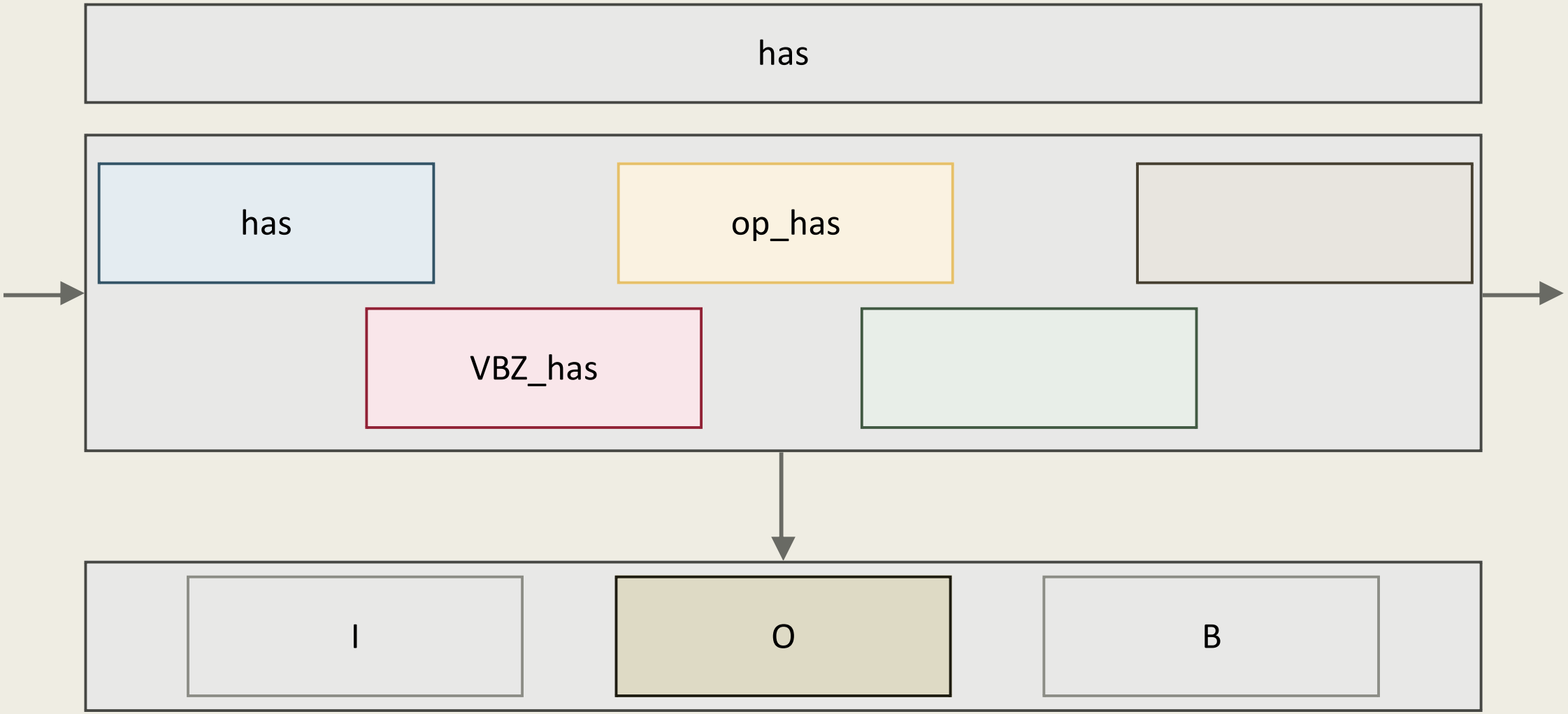
Linear-Chain CRF



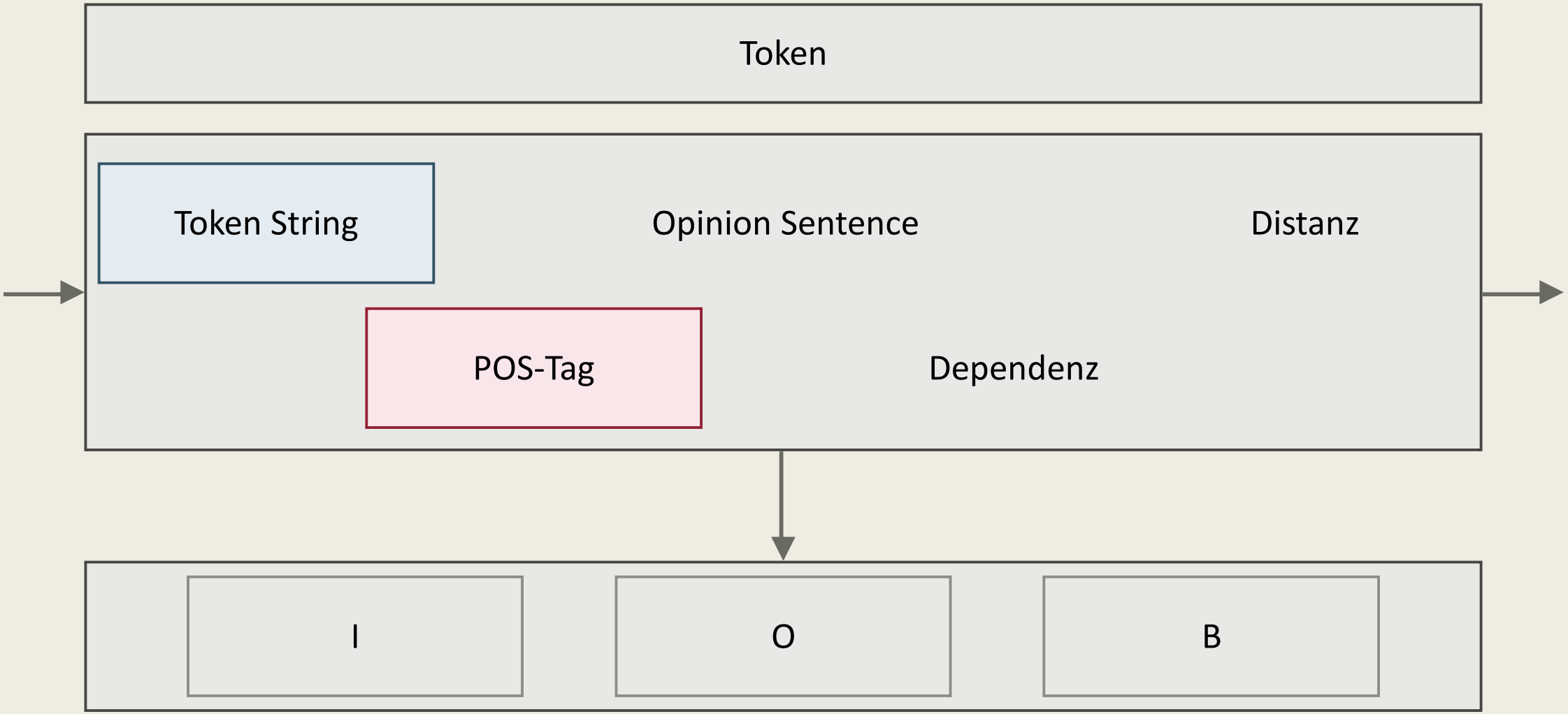
Linear-Chain CRF



Linear-Chain CRF



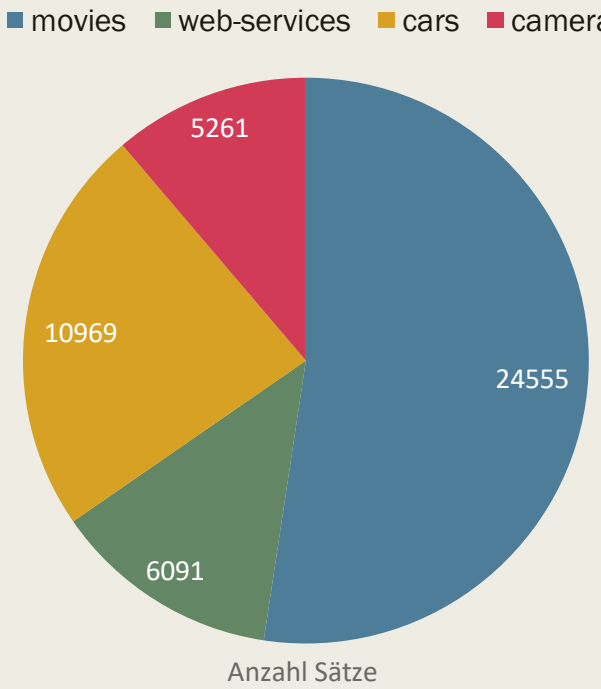
Linear-Chain CRF

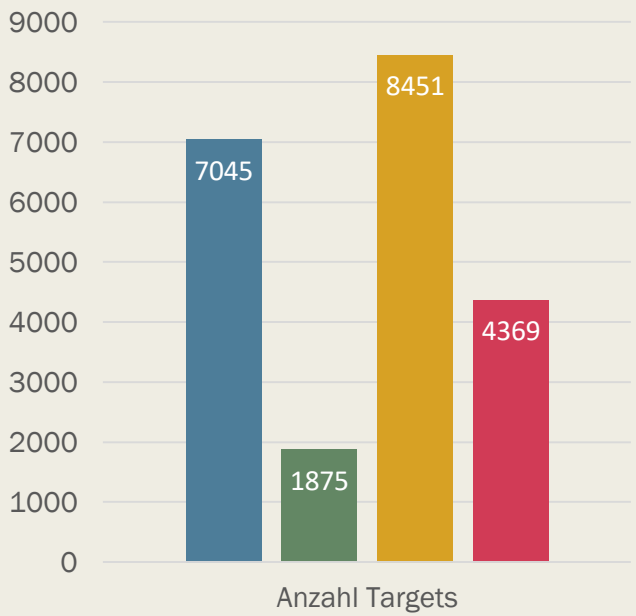


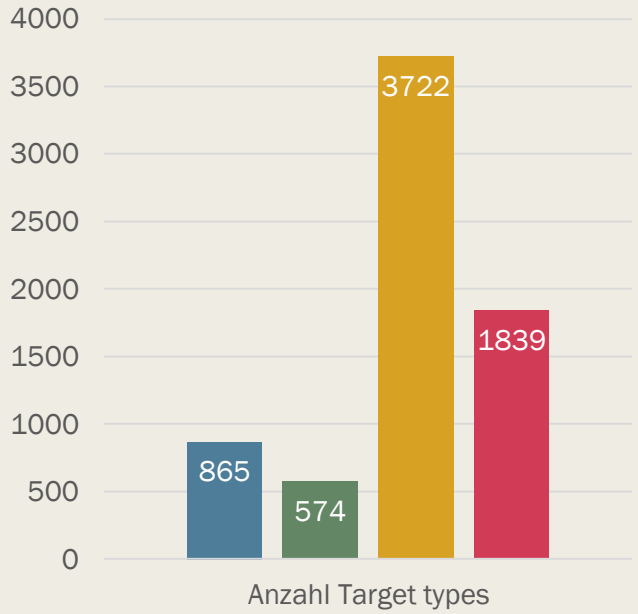
Datensets

movies	web-services	cars + cameras
www.IMDb.com	www.epinions.com	Blogbeiträge

Datensets: Statistiken







- In movies und web-services werden die gleichen Targets wiederholt
- In cars und cameras gibt es viele Targets sehr selten!

Setting

- 10-fache Kreuzvalidierung
- Single-Domain Ergebnisse: macro-averaged
- Cross-Domain Ergebnisse: micro-averaged

Single-Domain Ergebnisse (F-Measure)

	movies	web-services	cars	cameras
Baseline	0.625	0.483	0.322	0.426
Token, POS, Dist	0.271	0.339	0.436	0.446
Token, POS, Dep	0.595	0.475	0.460	0.453
Token, POS, OpSen	0.653	0.476	0.257	0.238
Alle Features	**0.702**	**0.609**	**0.497**	**0.500**
Alle Features außer Token	0.532	0.422	0.460	**0.500**

- Trend: beste Ergebnisse auf movies, schlechteste bei cars und cameras
- cars und cameras: Token Feature auslassen macht wenig bis keinen Unterschied

Cross-Domain Ergebnisse: Baseline

Training	Testing	F-Measure
cameras + web-services	cars	**0.171**

Training: movies
- The **story** is amazing…
- The worst **acting** ever!

Testing: cars
- The **sound system** is amazing…
- The worst **engine** ever!

- Algorithmus schaut nur nach gelernten Target Strings
- Überschneidung von Target-Vokabular zwischen Domänen sehr klein!
- Dependenzpfade eher domänenübergreifend

Cross-Domain Ergebnisse: CRF

- Featurekombination: Alle außer Token
- Mit Token: sehr niedriger Recall

- Schlägt Baseline in allen Domänen
- cameras liefert beste F-Measure
- cameras ist kleinstes Datenset!

Training	Testing	F-Measure
movies	web-services	0.316
	cars	0.384
	cameras	0.391
cars	movies	0.479
	web-services	0.340
	cameras	**0.475**
cameras	movies	**0.499**
	web-services	**0.345**
	cars	**0.465**

Cross-Domain Ergebnisse: CRF

Training	Testing	F-Measure
cameras	movies	**0.499**
	web-services	**0.345**
cars + cameras	movies	0.489
	web-services	**0.345**

- Zusätzliche cars Trainingsdaten verbessern Ergebnisse nicht weiter

Single- vs. Cross-Domain Ergebnisse

Features	Training	Testing	F-Measure
Alle	movies		**0.702**
Alle außer Token	movies		0.532
	web-services + cameras	movies	0.518

- Gleiche Features: Cross-Domain Ergebnisse nah an Single-Domain Ergebnissen
- Bestes Feature für Single-Domain: Token
- Token Feature in Cross-Domain eher schädlich

Zusammenfassung

- Überwachtes Modell zur Extraktion von Opinion Targets
- Schlägt Baseline in Single- und Cross-Domain Experimenten
- Ohne Token Feature: Cross-Domain Ergebnisse kommen relativ nah an Single-Domain Ergebnisse heran
- Features scheinen domänenbergreifend zu funktionieren
- Je nach Datenset und Setting (SD/CD) sind verschiedene Features hilfreich

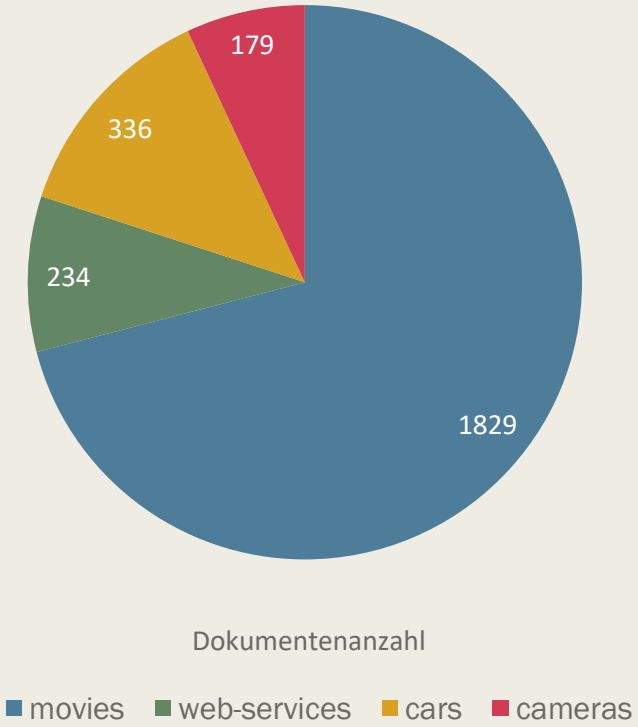
Vielen Dank für Eure Aufmerksamkeit!

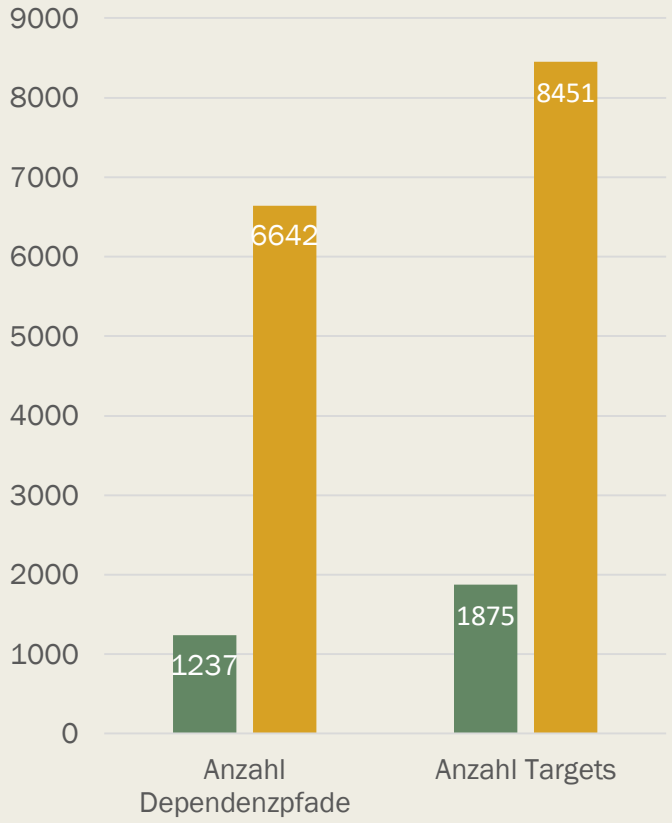
Fragen & Diskussion

Referenzen

- Hu & Liu, 2004. Mining and Summarizing Customer Reviews. *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining.* [pdf] abgerufen am 14.10.2019.
- Jakob & Gurevych, 2010. Extracting Opinion Targets in a Single- and Cross-Domain Setting with Conditional Random Fields. *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing.* [pdf] abgerufen am 10.10.2019.
- Kessler & Nicolov, 2009. Targeting Sentiment Expressions through Supervised Ranking of Linguistic Configurations. *Proceedings of the Third International ICWSM Conference (2009).* [pdf] abgerufen am 10.10.2019.
- Stanford CoreNLP für Dependenzparse. Abgerufen am 14.10.2019
- Sutton & McCallum, 2006. An Introduction to Conditional Random Fields for Relational Learning. *Book chapter in Introduction to Statistical Relational Learning*. MIT Press. [pdf] abgerufen am 11.10.2019.
- Zhuang et. al, 2006. Movie Review Mining and Summarization. *CIKM '06 Proceedings of the 15th ACM international conference on Information and knowledge management.* [pdf] abgerufen am 11.10.2019.

Appendix: Fehleranalyse Baseline Single-Domain





- cars und web-services enthalten ähnlich viele Dokumente
- Nach Training: cars hat 8 mal so viele Targets wie web-services
- Führt laut Autoren zu vielen False Positives
- Aber: cars enthält zwar ähnlich viele Dokumente, aber weitaus mehr Targets als web-services

Appendix: Fehleranalyse CRF Single-Domain

- Untersuchten Recall Fehler: Targets, die nicht gefunden wurden
- B-Targets werden fälschlicherweise als O (other) klassifiziert
- Großteil davon weist weder Dependenz- noch Distanzfeature auf

> A lens cap and strap may not sound very important, but it makes a *huge difference* in the speed and usability of the camera.

- CRF: „speed“ als target → weist Dependenz- und Distanzfeature auf

> Komplexere Strukturen werden von Features nicht abgedeckt!

Appendix: gelernte Expressions

- Bisher: Expressions aus dem Goldstandard
- Jetzt: Finde Expressions vorher über Subjektivitätslexikon (Session 2)
- Nur für Single-Domain angegeben, CRF mit allen Features (?)

	F-Measure
movies	0.309
web-services	0.234
cars	0.192
cameras	0.198

- F-Measure sehr viel schlechter
- Drei Features (Dependenzpfad, Distanz, Opinion Sentence) basieren auf (korrekten) Expressions!

Appendix: Overfitting im Cross-Domain Setting?

- Bestes Cross-Domain Trainingsdatenset (isoliert): cameras
- Gleichzeitig auch kleinstes Datenset!

→ Overfittet der Algorithmus auf die anderen Datensets?

- Autoren haben andere Datensets auf Größe des cameras Datensets reduziert
- Lieferte keine Verbesserung im Hinblick auf F-Measure

Wieso ist cameras so gut fürs Training?

→ Hohe Anzahl von Targets
→ Hohe Anzahl von Target types
→ Hohe Anzahl von subjektiven Sätzen