Measuring Distributional Similarity in Context

by Georgiana Dinu and Mirella Lapata

Vortrag von L. Heuschkel
Distributionelle Semantik 11/12

Bisher: Vektormodelle

- Weit verbreiteter Einsatz im Gebiet des Natural Language Processing
- Beliebtheit
  - Keine Überwachung notwendig
  - Einfache Berechnung
- Bedeutungen eines Wortes sind Punkte im Raum
- Jede Komponente gehört zu einem gleichzeitig auftretenden kontextabhängigen Element
- Ähnlichkeit der Wortbedeutungen durch Kosinuswinkel

Bisher: Vektormodelle

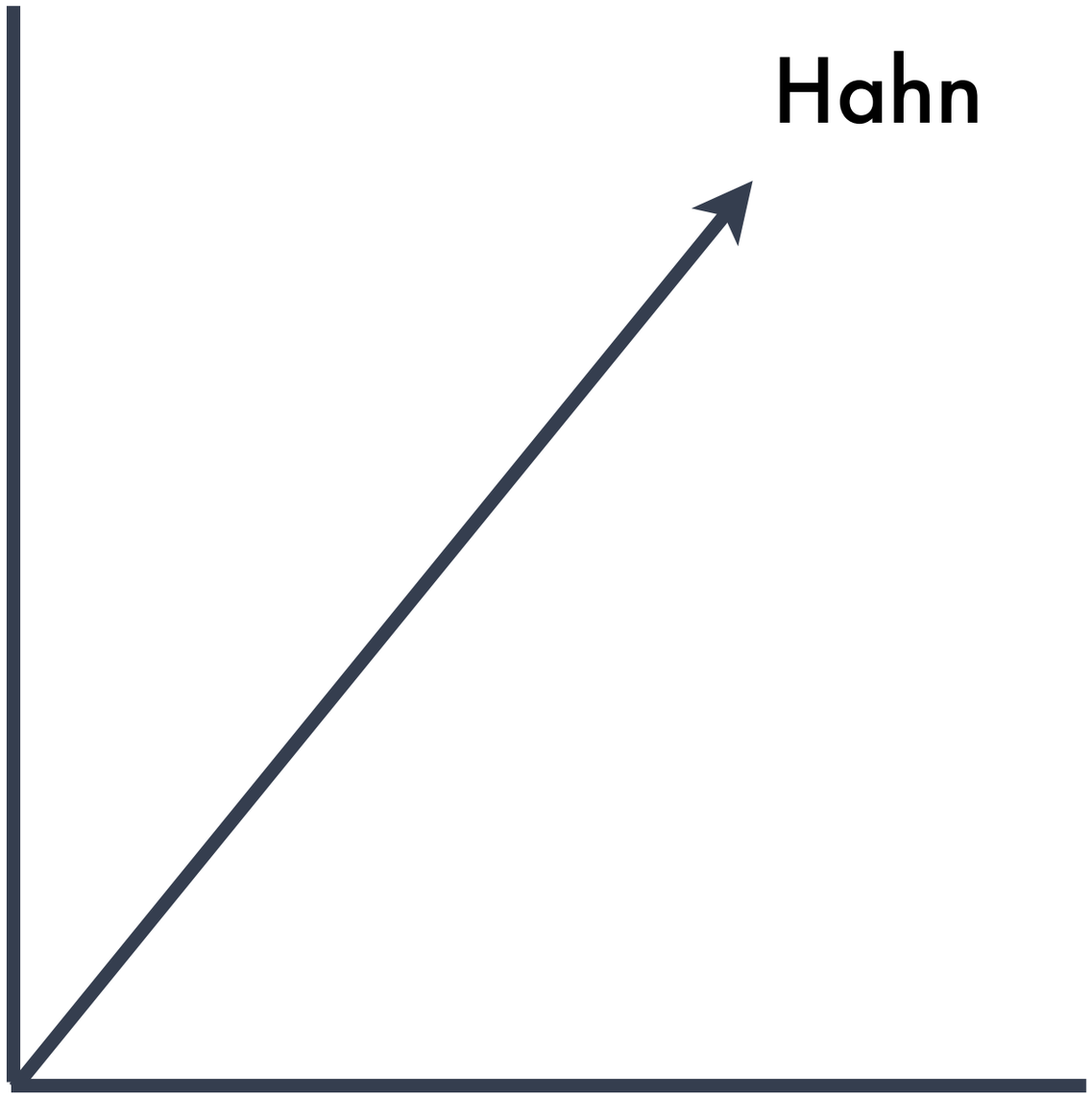
- Erkennen nicht explizit die verschiedenen Bedeutungen von Wörtern
- Stellen dementsprechend ihre Bedeutung unabhängig vom gleichzeitig auftretendem Kontext dar
- Daher: Benutzt zur Repräsentation isolierter Wörter
    - Bedeutungsähnlichkeit außerhalb des Kontextes

Bisher: Vektormodelle

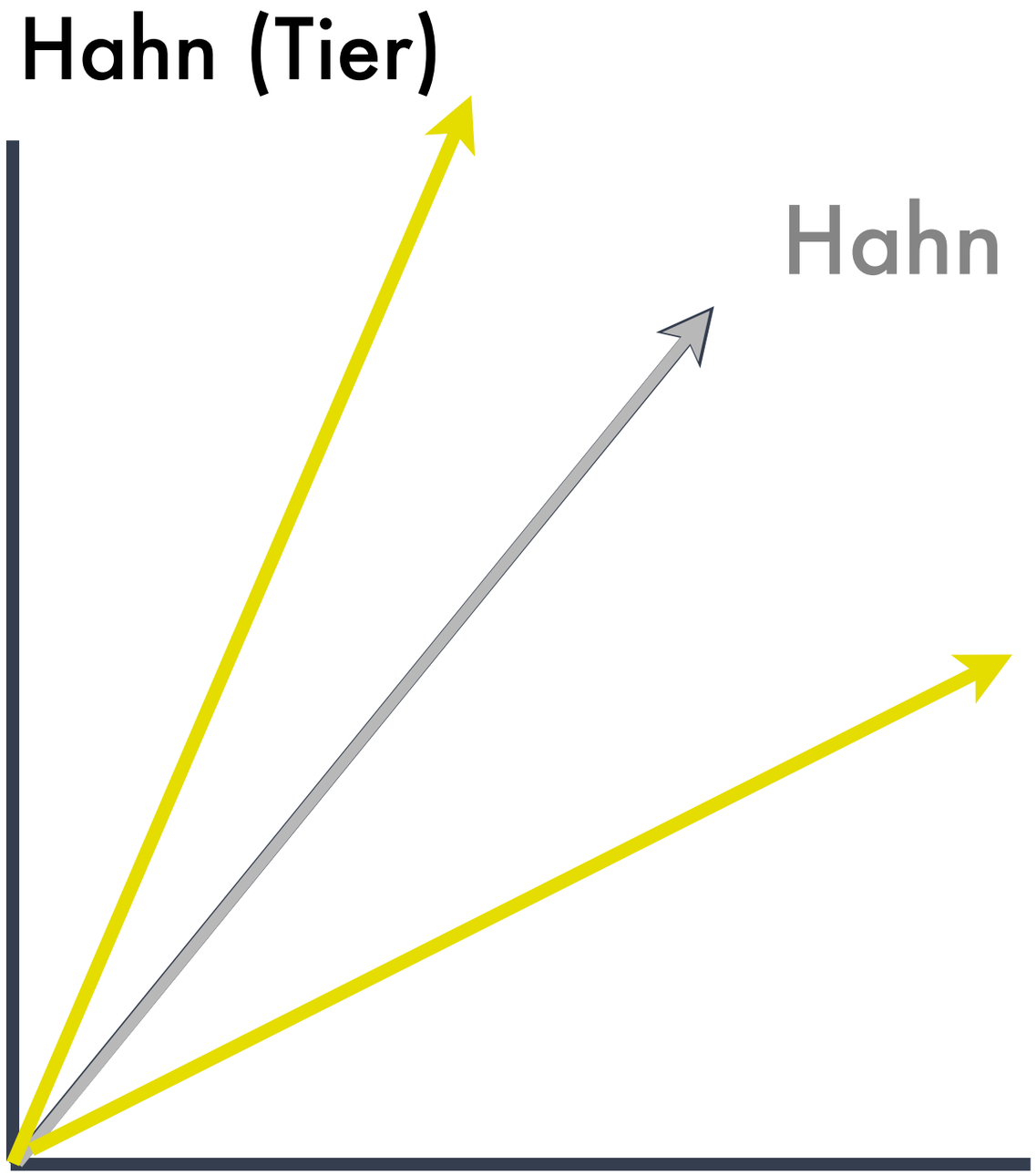
- Entwicklung spezieller Modelle, die Wortbedeutungen im Kontext repräsentieren
    - Mitchell & Lapata, 2008 ; Erk & Padó, 2008 ; Thater et al., 2009

- Problem wird nur indirekt adressiert
- Extrahieren die typischen Kookkurrenzvektoren: Vermischung von Bedeutungen!!
- Benutzung von Vektoroperationen, um entweder kontextualisierte Repräsentationen des Zielwortes/ Repräsentation für eine Reihe von Wörtern zu bekommen

Bisher: Vektormodelle



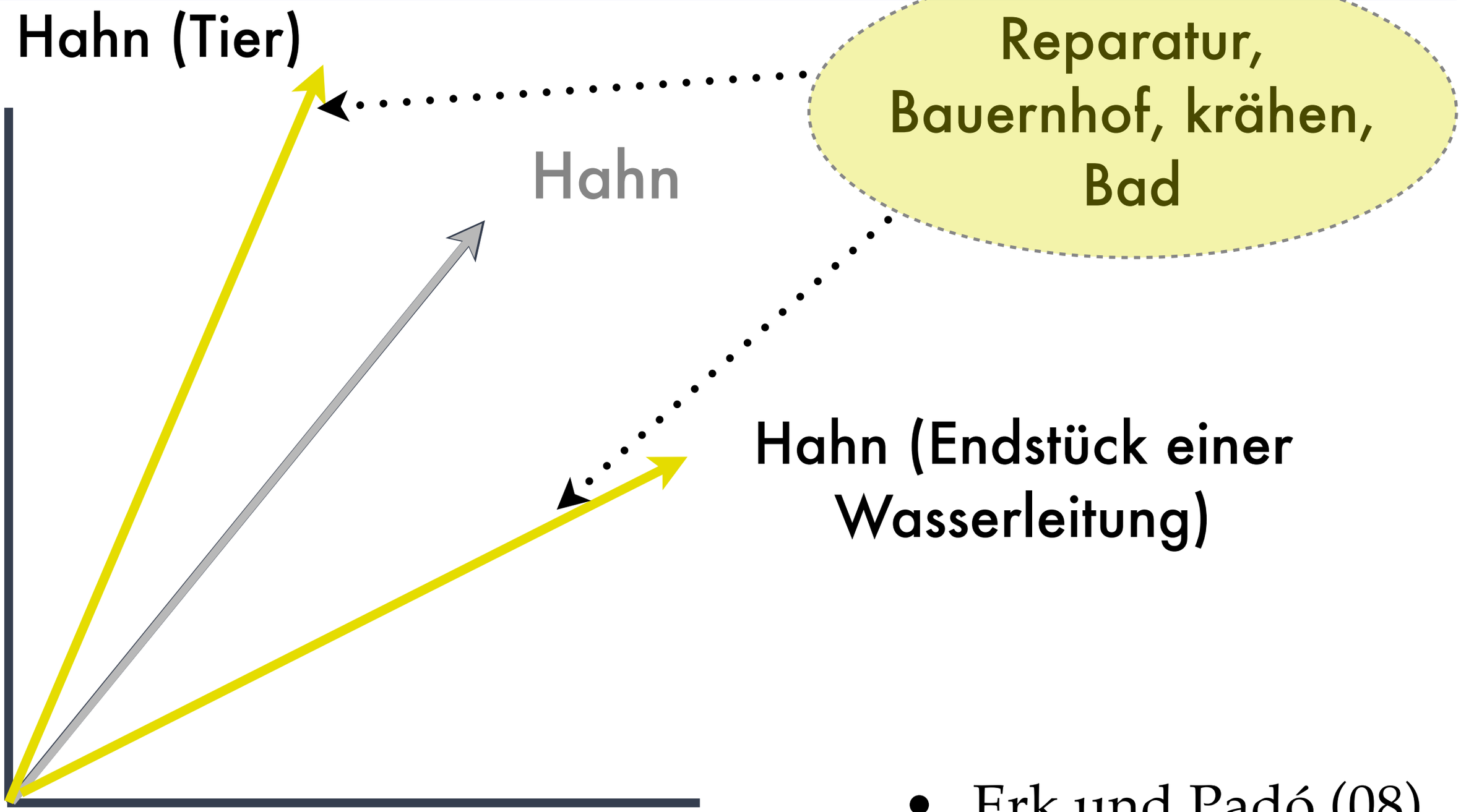
Bisher: Vektormodelle



Hahn (Endstück einer Wasserleitung)

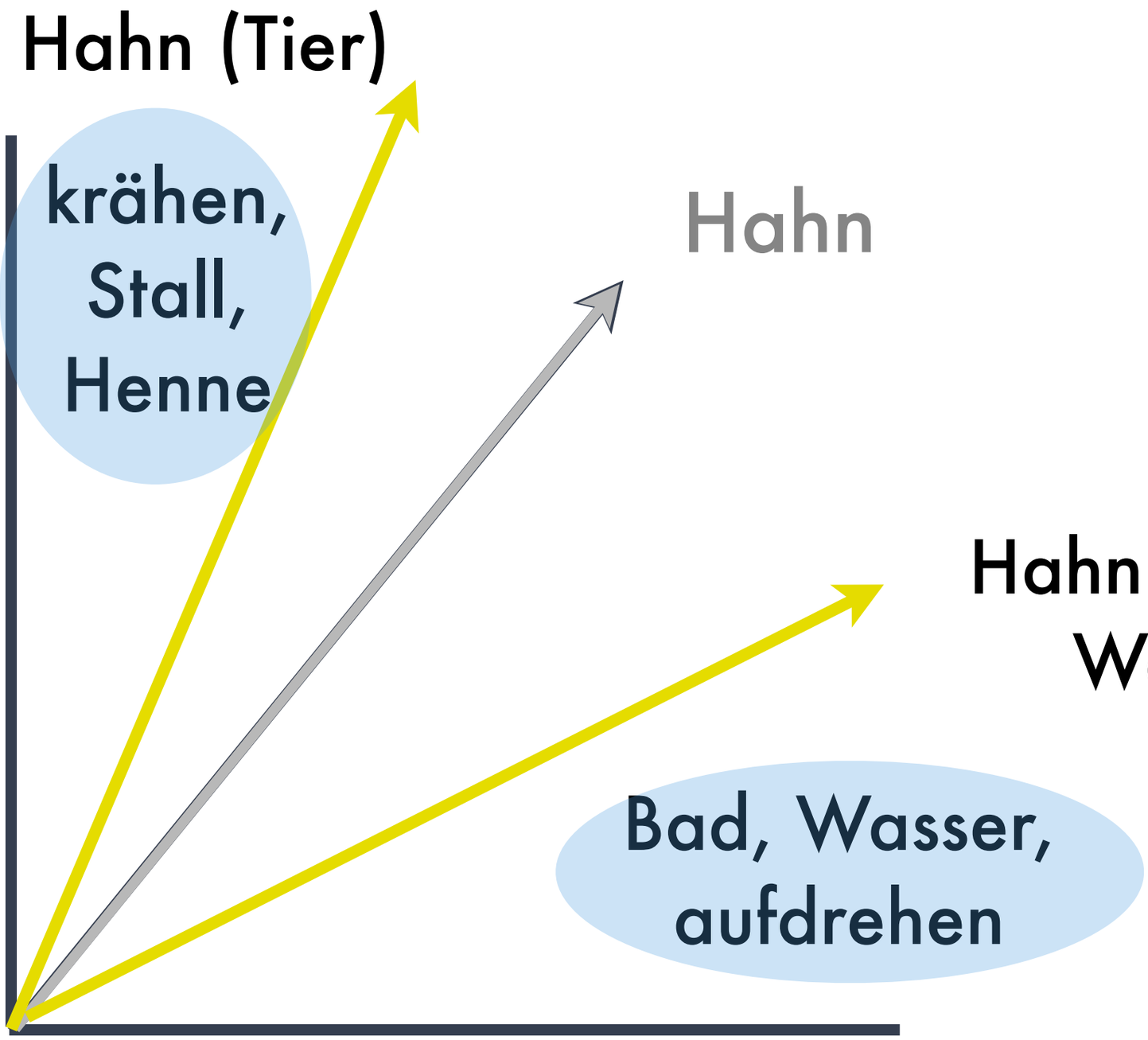
- Mitchell und Lapata (08)

Bisher: Vektormodelle



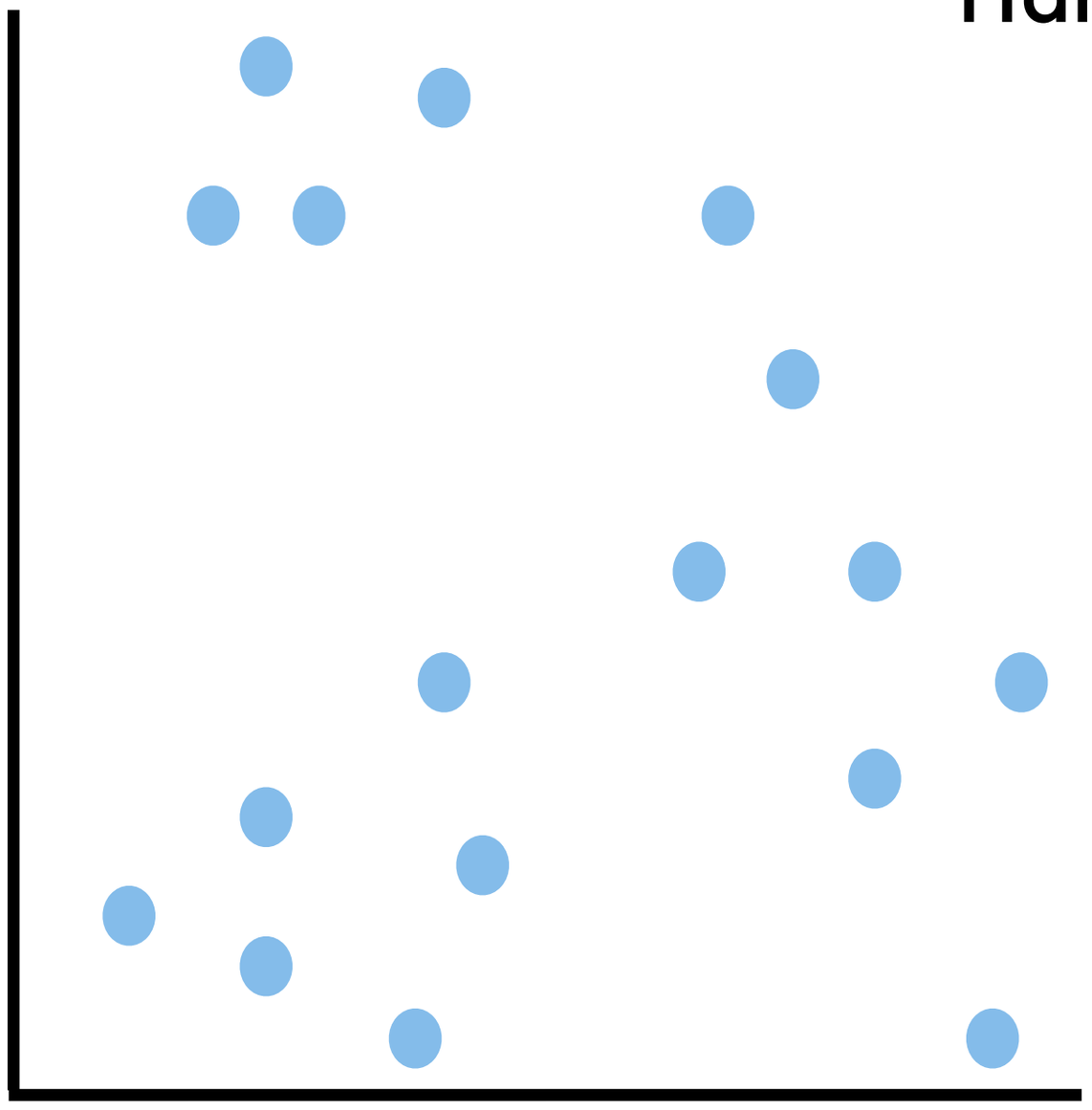
- Erk und Padó (08)

Bisher: Vektormodelle



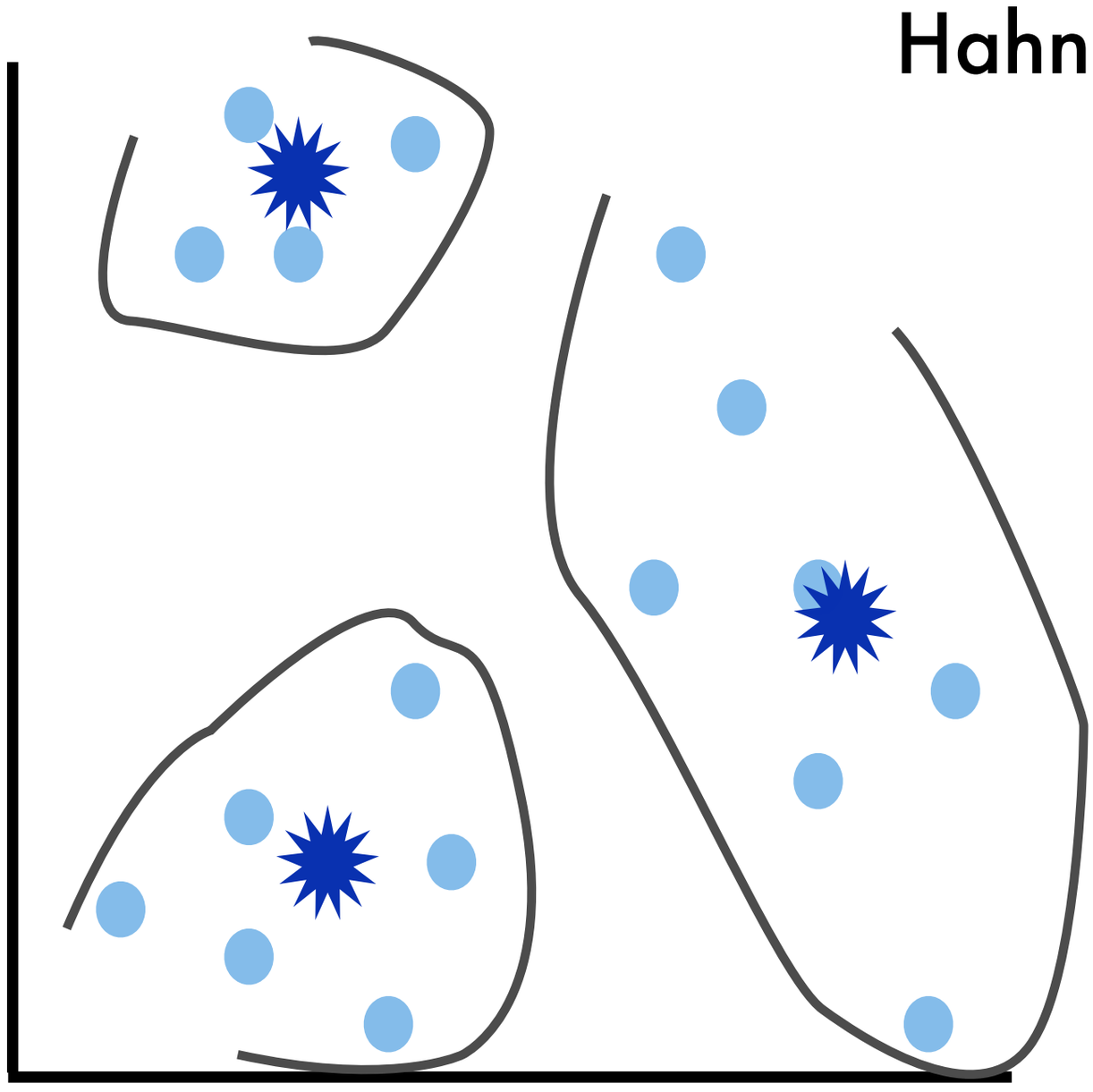
- Thater et al. (2011)

Bisher: Vektormodelle



- Reisinger und Mooney (2010)

Bisher: Vektormodelle



- Reisinger und Mooney (2010)

Neues Modell

$$\mathbf{v(t_i, c_j) = (P(z_1|t_i, c_j), ..., P(z_K|t_i, c_j))}$$

- Wahrscheinlichkeitssystem für die Repräsentation von Wortbedeutung und das Messen von Ähnlichkeit im Kontext
- Bedeutung isolierter Wörter: **Wahrscheinlichkeitsverteilung** über eine Reihe von verborgenen Bedeutungen

Neues Modell

- Verteilung gibt die Wahrscheinlichkeit jeder Bedeutung außerhalb des Kontextes wieder
- Bedeutungsambiguität wird **direkt** beim Konstruktionsprozess des Vektor berücksichtigt
- Kontextualisierte Bedeutung wird auf natürliche Weise als Veränderung in der originalen Bedeutungsverteilung modelliert

Bedeutungsdarstellung im Kontext

- Modell nimmt gleiche Art von Input Data an wie Vektormodelle

  --> Kookkurrenzmatrix

Bedeutungsdarstellung im Kontext

- Modell nimmt gleiche Art von Input Data an wie Vektormodelle

  --> Kookkurrenzmatrix

Hahn: Tier/Wasserhahn

		Kontextmerkmale: Nachbarwörter c_j					
		c_1: Wasser	c_2: Stall	c_3: Henne	c_4: putzen	c_5: aufdrehen	c_6: krähen
Ziel-wörter t_j	t_1: Hahn	12	7	5	5	8	8
	...						

Bedeutungsdarstellung im Kontext

- Zielwörter eines Korpus teilen ein globales Set an Bedeutungen
  $Z = \{z_k \mid k : 1 \ldots K\}$

- Bedeutung der individuellen Zielwörter kann als Verteilung über dieses Set von Bedeutungen beschrieben werden

- Ziel t_i wird durch folgenden Bedeutungsvektor dargestellt:

Bedeutungsdarstellung im Kontext

- Zielwörter eines Korpus teilen ein globales Set an Bedeutungen
  $Z = \{z_k \mid k : 1 \ldots K\}$

- Bedeutung der individuellen Zielwörter kann als Verteilung über dieses Set von Bedeutungen beschrieben werden

- Ziel t_i wird durch folgenden Bedeutungsvektor dargestellt:

$$\mathbf{v}(\mathbf{t_i}) = (\mathbf{P}(\mathbf{z_1}|\mathbf{t_i}), \ldots, \mathbf{P}(\mathbf{z_K}|\mathbf{t_i}))$$

$$v(\text{Hahn}) = (P(\text{Wasserhahn} \mid \text{Hahn}), P(\text{Vogel} \mid \text{Hahn}))$$

Bedeutungsdarstellung im Kontext

- Ein Zielwort wird durch eine Reihe von Kernbedeutungen und der Häufigkeit, mit der diese bestätigt werden, beschrieben
- Die Bedeutungen z_1 ... z_k sind latente Variablen (verborgen)
- Mittel zur Reduzierung der Dimensionalität der ursprünglichen Kookkurrenzmatrix
- Bedeutungen sind nicht wortspezifisch, sondern global und werden entweder im oder außerhalb des Kontextes probabilistisch angepasst

Bedeutungsdarstellung im Kontext

- Bedeutung eines Zielwortes, wenn ein Kontextmerkmal gegeben ist:

$$\mathbf{v(t_i, c_j) = (P(z_1|t_i, c_j), ..., P(z_K|t_i, c_j))}$$

- Ziel t_i ist nun auf einen spezifischen Kontext c_j abgestimmt, der tatsächliche Wortbenutzung widerspiegelt

- Verteilung ist fokussierter

- Kontext hilft die Bedeutung des Zielwortes zu disambiguieren

- Weniger Bedeutungen teilen das Meiste der Wahrscheinlichkeitsmasse

Bedeutungsdarstellung im Kontext

$$\mathbf{v}(\mathbf{t_i}) = (\mathbf{P}(\mathbf{z_1}|\mathbf{t_i}), ..., \mathbf{P}(\mathbf{z_K}|\mathbf{t_i}))$$

v(Hahn) = (P(Wasserhahn | Hahn), P(Vogel | Hahn))

v(Hahn) = (0,1 , 0,9)

$$\mathbf{v}(\mathbf{t_i}, \mathbf{c_j}) = (\mathbf{P}(\mathbf{z_1}|\mathbf{t_i}, \mathbf{c_j}), ..., \mathbf{P}(\mathbf{z_K}|\mathbf{t_i}, \mathbf{c_j}))$$

v(Hahn, Wasser) = (P(Wasserhahn | Hahn, Wasser), P(Vogel | Hahn, Wasser))

v(Hahn, Stall) = (P(Wasserhahn | Hahn, Stall), P(Vogel | Hahn, Stall))

Abschätzung der Whk

$$\mathbf{v}(\mathbf{t_i}, \mathbf{c_j}) = (\mathbf{P}(\mathbf{z_1}|\mathbf{t_i}, \mathbf{c_j}), ..., \mathbf{P}(\mathbf{z_K}|\mathbf{t_i}, \mathbf{c_j}))$$

$$P(z_k|t_i, c_j) = \frac{P(t_i, z_k)P(c_j|z_k, t_i)}{\sum_k P(t_i, z_k)P(c_j|z_k, t_i)}$$

$$P(z_k|t_i, c_j) \approx \frac{P(z_k|t_i)P(c_j|z_k)}{\sum_k P(z_k|t_i)P(c_j|z_k)}$$

vereinfachernde Annahme, dass t_i und c_j bedingt wahrscheinlich sind, wenn z_k gegeben ist

Abschätzung der Whk

$$\mathbf{v(t_i, c_j) = (P(z_1|t_i, c_j), ..., P(z_K|t_i, c_j))}$$

$$P(z_k|t_i, c_j) \approx \frac{P(z_k|t_i)P(c_j|z_k)}{\sum_k P(z_k|t_i)P(c_j|z_k)}$$

v(Hahn, Wasser)

$$= \quad (\mathrm{P(Wasserhahn \mid Hahn,\ Wasser),\ P(Vogel \mid Hahn,\ Wasser)})$$

$$= \left(\frac{\mathrm{P(Wasserhahn \mid Hahn)\ P(Wasser \mid Wasserhahn)}}{\mathrm{P(Wasserhahn \mid Hahn)\ P(Wasser \mid Wasserhahn) + P(Vogel \mid Hahn)\ P(Wasser \mid Vogel)}} \right.$$

$$\left. , \frac{\mathrm{P(Vogel \mid Hahn)\ P(Wasser \mid Vogel)}}{\mathrm{P(Wasserhahn \mid Hahn)\ P(Wasser \mid Wasserhahn) + P(Vogel \mid Hahn)\ P(Wasser \mid Vogel)}} \right)$$

Parametrisierung

- Diese generelle Grundstruktur kann auf die Input Kookkurrenz Matrix und dem Algorithmus, der zum Erzeugen der Latenten Struktur benutzt wird, angepasst werden
- Großer Spielraum beim Erzeugen der Kookkurrenz-Matrix
- Mögliche Spalten (dh. Kontextmerkmale):
  - Nachbarwörter um das Zielwort
  - Ganze Paragraphen, Dokumente
  - Syntaktische Abhängigkeiten

Parametrisierung

- Es können auch eine Anzahl von Wahrscheinlichkeitsmodellen benutzt werden, um die verborgenen Bedeutungen hervorzurufen
  - Non-negative Matrix Factorization
  - Latent Dirichlet Allocation

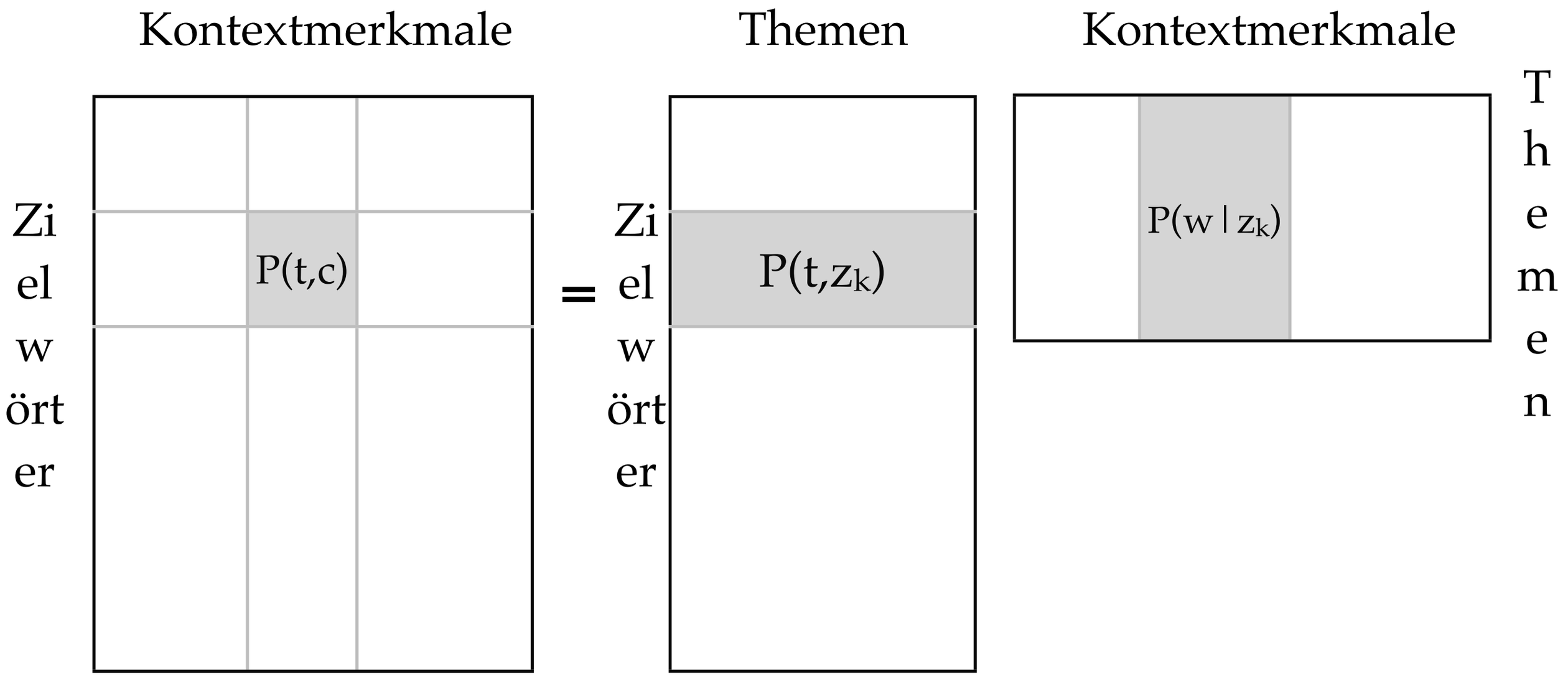
Non-negative Matrix Factorization

- Input Matrix V wird in zwei nicht negative Matrizen W und H zerlegt

$$V_{I,J} \approx W_{I,K} H_{K,J}$$

- W und H sind dimensional reduziert
- Ihr Produkt kann als komprimierte Form der Daten aus V angesehen werden

Non-negative Matrix Factorization



$$V_{I,J} \approx W_{I,K} H_{K,J}$$

W: Basis Vektor Matrix
H: kodierte Matrix der Basisvektoren

Latent Dirichlet Allocation (LDA)

- Wahrscheinlichkeitsmodell für Textgenerierung
- Jedes Dokument d wird als eine Verteilung über K Themen angesehen
- Jedes Wort wiederum ist einem oder mehreren Themen zugeordnet
- Festlegung der Anzahl der Themen zu Beginn

Experimente: Aufgaben

- Aufgaben:
- Test zur Wortähnlichkeit
- Test zur lexikalischen Substitution

Test zur Wortähnlichkeit

- Modell repräsentiert Wörter durch eine Reihe von erzeugten Bedeutungen
- Experimente mit 2 Arten von Semantic Space, basierend auf NMF und LDA und optimierten Parametern für diese Modelle
- Beurteilung der Gleichheit zweier Wörter außerhalb des Kontextes

$$sim(v(t_i), v(t'_i))$$

- Daten aus Finkelstein et al. 2002 (353 Wortpaare und ihre Ähnlichkeitswertung)

Test zur Wortähnlichkeit

possibility	girl	1,94
population	development	3,75
planet	sun	8,02
planet	star	8,45
planet	space	7,92
planet	people	5,75
planet	moon	8,08
planet	galaxy	8,11
planet	constellation	8,06
planet	astronomer	7,94
plane	car	5,77
physics	proton	8,12
physics	chemistry	7,35
phone	equipment	7,13
peace	plan	4,75
peace	insurance	2,94
peace	atmosphere	3,69

Finkelstein et al. 2002

(353 Wortpaare und ihre Ähnlichkeitswertung)

Test zur lexikalischen Substitution

- Systeme bekommen eine Reihe von Substitutionswörtern für Zielwörter aus dem Kontext und müssen dann eine Rangordnung erstellen
- Geeignete Substitutionswörter sollen verglichen mit weniger geeigneten einen höheren Rang bekommen
- Benutzung des SemEval 2007 Lexical Substitution Task benchmark data set
  - 200 Zielwörter aus 10 distinktiven Satzkontexten
  - insgesamt 2.000 Sätze
  - 5 menschliche Annotatoren bestimmen die Ersatzwörter

Test zur lexikalischen Substitution

Sentences	Substitutes
It is important to apply the herbicide on a *still* day, because spray drift can kill non-target plants.	calm (5) not-windy (1) windless (1)
A movie is a visual document comprised of a series of *still* images.	motionless (3) unmoving (2) fixed (1) stationary (1) static (1)

Test zur lexikalischen Substitution

Sentences	Substitutes
It is important to apply the herbicide on a *still* day, because spray drift can kill non-target plants.	calm (5) not-windy (1) windless (1)
A movie is a visual document comprised of a series of *still* images.	motionless (3) unmoving (2) fixed (1) stationary (1) static (1)

ring (n)	call (5) telephone call (1) bell (1)
ring (n)	band (2) fob (1) chain (1) hoop (1) holder (1) circle (1)
ring (n)	circle (3) group (2) network (1)

Test zur lexikalischen Substitution

- Für jedes Zielwort: Zusammenlegen aller Ersatzwörter
- Modell muss eine Rangordnung für jede Ersatzmenge erstellen
- Anordnung basiert auf Ähnlichkeit des kontextualisierten Ziels und des Ersatzwortes außerhalb des Kontextes
- Einordnung von nur einem der Wörter im Kontext
    - Modell bekommt höheres Differenzierungspotenzial

Experimente: Modelltraining

- Alle Modelle benutzen die gleichen Input Daten
  - Eine bag-of-words Matrix aus der GigaWord Collection aus Nachrichtentexten

    bag-of-words: ungeordnet, Grammatik nicht berücksichtigt, keine Wortordnung

    - Reihen: Zielwörter
    - Spalten: Nachbarwörter ±5
    - Kontextwörter: 3.000 der am meist vorkommenden Wörter aus dem Korpus
- Abstimmung der Modellvariablen
  - Bestmögliche Instanziierung jedes Modelltyps

Baselines

- Alternative / Vergleich zu eigenen Modellen
- Ähnlichkeit außerhalb des Kontextes:
  - LSA (Latent Semantic Analysis): sucht Hauptkomponenten in Reihe von Dokumenten
  - Simple semantic space: benutzt originale Input Matrix mit mehrere Bewertungsschemen
- Kontextualisierte Ähnlichkeit:
  - Vektoraddition / Vektormultiplikation

Evaluation

- Wortähnlichkeitstest: Ähnlichkeitsanalyse, Vergleich zwischen menschlichen Bewertungen und deren zugehörigen Vektor-basierten Ähnlichkeitswerten (Spearman‘s *p)*

- Lexikalische Substitution: Vergleich Systembewertung / Goldstandard Ranking mit Kendall‘s τ_b rank correlation

- Für alle kontextualisierten Modelle:
  Kontext des Zielwortes = Nachbarwörter ±5

Modelle Überblick

- SVS: simple co-occurrence based vector space model
- LSA: latent semantic analysis
- NMF: non-negative matrix factorization
- LDA: latent Dirichlet allocation
- Mixtures: LSA_{MIX}, NMF_{MIX}, LDA_{MIX}

Ergebnisse: Wortähnlichkeit

Model	Spearman ρ
SVS	38.35
LSA	49.43
NMF	**52.99**
LDA	**53.39**
LSA_{MIX}	49.76
NMF_{MIX}	51.62
LDA_{MIX}	51.97

Am besten mit tf-idf weighting und Kosinusähnlichkeit

Intitial line normalization, Skalarprodukt Ähnlichkeitsmaß

Skalarprodukt

Bessere Ergebnisse bei einer größeren Anzahl ein Bedeutungen

NMF/LDA erbringen signifikant bessere Übereinstimmungen als LSA und SVS

Ergebnisse: Lexikalische Substitution

Model	Kendall's τ_b
SVS	11.05
Add-SVS	12.74
Add-NMF	12.85
Add-LDA	12.33
Mult-SVS	14.41
Mult-NMF	13.20
Mult-LDA	12.90
Cont-NMF	14.95
Cont-LDA	13.71
Cont-NMF$_{\text{MIX}}$	**16.01**
Cont-LDA$_{\text{MIX}}$	**15.53**

Benutzt keine kontextuale Information, gibt das gleiche Ranking der Ersatzwörter für jede Instanz zurück, ausschließlich auf der Ähnlichkeit mit dem Zielwort basierend

pmi weighting, Lin's similarity measure

tf-idf weighting, Kosinusähnlichkeit

Andere Kombinationen lieferten signifikant niedrigere Resultate

Ergebnisse: Lexikalische Substitution

Model	Kendall's τ_b
SVS	11.05
Add-SVS	12.74
Add-NMF	12.85
Add-LDA	12.33
Mult-SVS	14.41
Mult-NMF	13.20
Mult-LDA	12.90
Cont-NMF	14.95
Cont-LDA	13.71
Cont-NMF$_{\text{MIX}}$	**16.01**
Cont-LDA$_{\text{MIX}}$	**15.53**

Alle Modelle liefern signifikant bessere Ergebnisse als SVS

pmi weighting, Lin's similarity measure

Multiplikations Modell ist das best funktionierende Kompositionsmodell --> bestätigt die Resultate von Mitchell und Lapata

Bessere Ergebnisse wahrscheinlich wegen der Wahrscheinlichkeitsfomulierung des kontextualisierten Modells als ganzes, nicht wegen der Benutzung von NMF oder LDA

Ergebnisse: Lexikalische Substitution
Performance der Modelle bei verschiedenen Wortarten

Kendall's τ_b

Model	Adv	Adj	Noun	Verb
SVS	22.47	14.38	09.52	7.98
Add-SVS	22.79	14.56	11.59	10.00
Mult-SVS	22.85	16.37	13.59	11.60
Cont-NMF$_{MIX}$	**26.13**	**17.10**	15.16	**14.18**
Cont-LDA$_{MIX}$	21.21	16.00	**16.31**	13.67

Beste Ergebnisse

Ergebnisse: Lexikalische Substitution

Performance der Modelle bei verschiedenen Wortarten

Am kompliziertesten?

Kendall's τ_b

Model	Adv	Adj	Noun	Verb
SVS	22.47	14.38	09.52	7.98
Add-SVS	22.79	14.56	11.59	10.00
Mult-SVS	22.8	7	7	11.60
Cont-NMF$_{MIX}$	**26.13**	**17.10**	15.16	**14.18**
Cont-LDA$_{MIX}$	21.21	16.00	**16.31**	13.67

+ ~7%

+ ~6%

Beste Ergebnisse

Alle kontextualisierte Modelle machen kleinere Verbesserungen bei Adjektiven

Nur Cont-NMFMIX macht Verbesserungen bei Adverben

Ergebnisse: Lexikalische Substitution

- Vergleich mit bisherigen Modellen
- Generalized Average Precision (GAP, Kishida (2005)) als Bewertungsmaß
- GAP berücksichtigt die Anordnung der richtig angeordneten Kandidaten durch ein theoretisches System

Modell	GAP
Erk und Padó (2008)	27,4
Erk und Padó (2010)	38,6
Cont-NMF$_{MIX}$	42,7
Cont-LDA$_{MIX}$	42,9
Thater et al.	46,0

Auswirkungen von Kontextwörtern auf Whk-Verteilung

- Test wie Kontextwörter die Wahrscheinlichkeitsverteilungen der Zielwörter beeinflussen
- Beispiele aus dem lexical substitution dataset und dem Output eines individuellen Cont-LDA Modells
- Oft: Zielwort fängt mit Whk-Verteilung über eine große Anzahl von Bedeutungen an
    - Kontextwort lagert diese Verteilung auf eine Hauptbedeutung um

Auswirkungen von Kontextwörtern auf Whk-Verteilung

„With their transcendent, improvisational jams and Mayan-inspired sense of a higher, meta- physical purpose, the band's music delivers a spiritual sustenance that has earned them a very devoted core following."

Kontextunabhängig 5 häufigst assoziierte Wörter

Senses	Word Distributions
TRAFFIC (0.18)	*road, traffic, highway, route, bridge*
MUSIC (0.04)	*music, song, rock, band, dance, play*
FAN (0.04)	*crowd, fan, people, wave, cheer, street*
VEHICLE (0.04)	*car, truck, bus, train, driver, vehicle*

verkehrszugehörig

musikzugehörig

Auswirkungen von Kontextwörtern auf Whk-Verteilung

„With their transcendent, improvisational jams and Mayan-inspired sense of a higher, meta- physical purpose, the **band**'s music delivers a spiritual sustenance that has earned them a very devoted core following."

Senses	Word Distributions
TRAFFIC (0.18)	*road, traffic, highway, route, bridge*
MUSIC (0.04)	*music, song, rock, band, dance, play*
FAN (0.04)	*crowd, fan, people, wave, cheer, street*
VEHICLE (0.04)	*car, truck, bus, train, driver, vehicle*

+

0,88 !

Auswirkungen von Kontextwörtern auf Whk-Verteilung: Probleme

- Oft wird das Zielwort nur teilweise durch ein Kontextwort disambiguiert
    - Spiegelt sich auch in der resultierenden Verteilung wider

bug –> software (0,09)
disease (0,06)

bug –> secret agency (0,34)
software (0,29)

Kontext: **client**

Auswirkungen von Kontextwörtern auf Whk-Verteilung: Probleme

- Oft wird das Zielwort nur teilweise durch ein Kontextwort disambiguiert
    - Spiegelt sich auch in der resultierenden Verteilung wider
- Kontextualisierte Verteilung kann falsch sein
    - Bsp: Bedeutungen sind Domain-spezifisch

Auswirkungen von Kontextwörtern auf Whk-Verteilung: Probleme

- Oft wird das Zielwort nur teilweise durch ein Kontextwort disambiguiert
    - Spiegelt sich auch in der resultierenden Verteilung wider
- Kontextualisierte Verteilung kann falsch sein
    - Bsp: Bedeutungen sind Domain-spezifisch

function (im mathematischen Sinne) mit Kontextwort **distribution**
–> ausgelöste Bedeutungen alle aus „dienstlichen“ Sektor (function = Funktion, Amt, ...)

- Konsequenz des benutzten Nachrichtenkorpus

Auswirkungen von Kontextwörtern auf Whk-Verteilung: Probleme

- Oft wird das Zielwort nur teilweise durch ein Kontextwort disambiguiert
    - Spiegelt sich auch in der resultierenden Verteilung wider
- Kontextualisierte Verteilung kann falsch sein
    - Bsp: Bedeutungen sind Domain-spezifisch
- Zielwort und eines der Kontextwörter werden Bedeutungen zugeteilt, die lokal korrekt sind, aber falsch im größeren Kontext

„Check the shoulders so it hangs well, stops at hips or below, and make sure the pants are long enough.“

–> Injury (0,81) oder Ball-Sports (0,10)

Fazit

- Schlüssel dieses Verfahrens:
  - Fundiert **probabilistisches** Modell
  - Repräsentation von Wortbedeutung als eine Verteilung über eine Reihe von globalen **Bedeutungen (nicht mehr nur Wörter!)**
  - **Kontextualisierte** Bedeutung ist als eine Veränderung in dieser Verteilung modeliert

Fazit

- Nutzung von NMF und LDA zur Herbeiführung der verborgenen Struktur
- Überbieten bisherige Modelle zur Bedeutungsähnlichkeit im Kontext
- Beide profitieren von der Vermischung der Modellvorhersagen über eine Reihe von verschiedenen Parametermöglichkeiten

Fazit

- Viele und verschiedene Richtungen für die Zukunft
- Modell macht Unterschied zwischen Zielwörtern und Kontextmerkmalen
- Benutzt allerdings Vektorrepräsentationen, die nicht unterscheiden
  - Zur Erleichterung für den Vergleich mit gängigen bag-of-words Vektor space Modellen
- Differenzierung zwischen Zielwort und Kontext-repräsentation vorteilhaft
  - Ähnlichkeitsberechnungen werden für andere Aufgaben gebraucht (Aneignung von Paraphrasen, Lexikaaufbau)

Fazit

- Modell kontextualisiert Zielwörter in Bezug auf die individuellen Kontexte
- Idealerweise könnte der kollektive Einfluss von mehreren Wörtern auf das Ziel berechnet werden
- Pläne: weitere Forschung zur Auswahl oder zur besseren Anhäufung aller aus dem Kontext extrahierten Merkmale

Ende

- Fragen?


DAS IST UNSER FRANZÖSISCHES AUSTAUSCH-RENTIER.
ICH WEISS. ABER WARUM HÄNGT ES METTWURST AN DEN BAUM?
VOILA! LA METTA!
WWW.NICHTLUSTIG.DE