Seminar: Distributionelle Semantik WS 2011/12

“Word Meaning in Context: A Simple and Effective Model“

S. Thater, H. Fürstenau, M. Pinkal

Referent: Simon Ostermann

5.12.2011

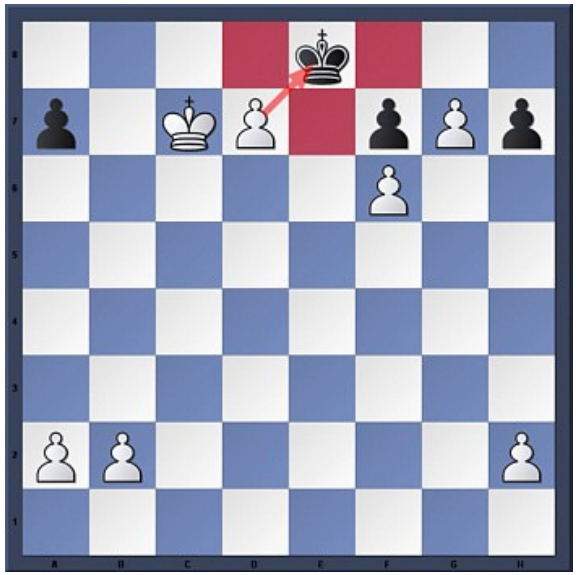
Problemstellung

Problem beim Verarbeiten von natürlicher Sprache: ambige Wörter

Der Zug kommt auf Gleis 4.

Mit dem Zug gewinnen.

Mit einem Zug gewinnt er das Spiel.

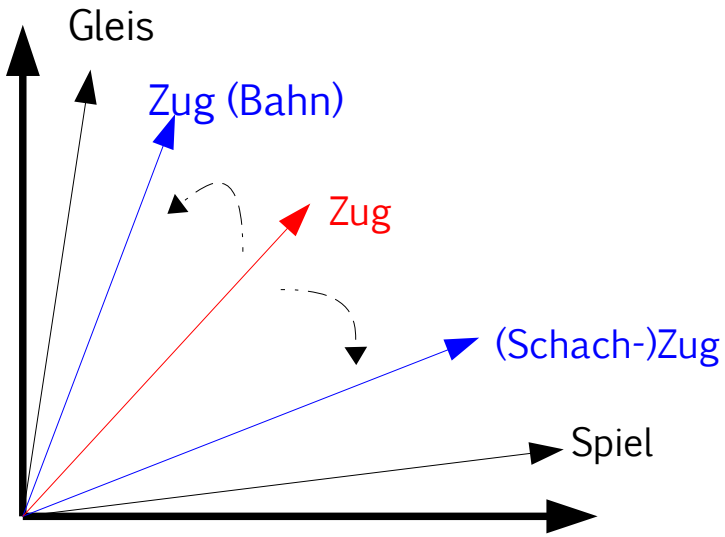


Aber: *Der (Schach-)Zug kommt auf Gleis 4??

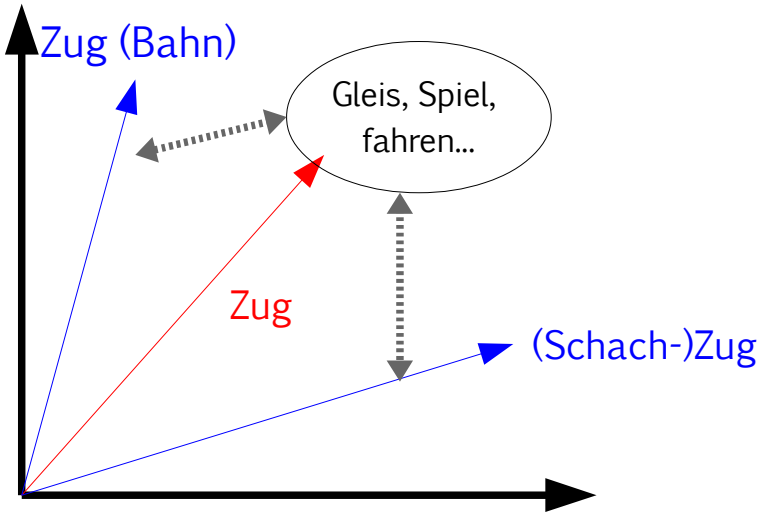
Problemstellung

Lösung: Vektor-basierte semantische Beschreibung eines Wortes – kurzer Rückblick

z.B. M&L '08



Schon sehr gute Ergebnisse, aber geht es besser?

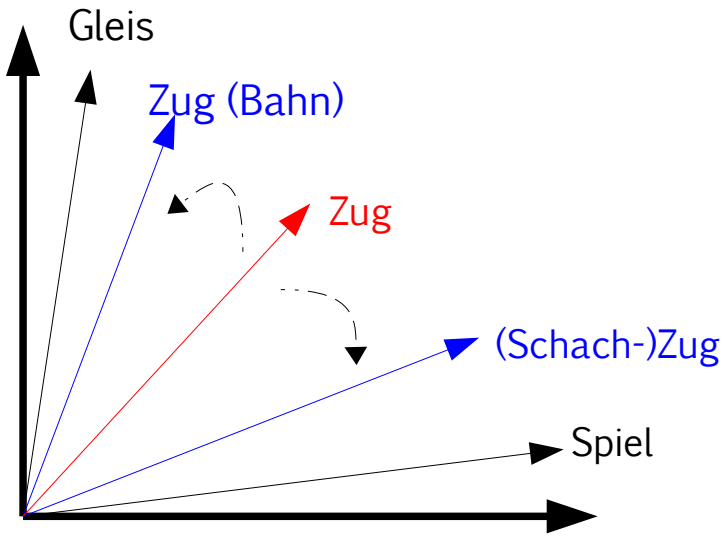


Ja! z.B. Erk & Pado '08, Selektionspräferenzen

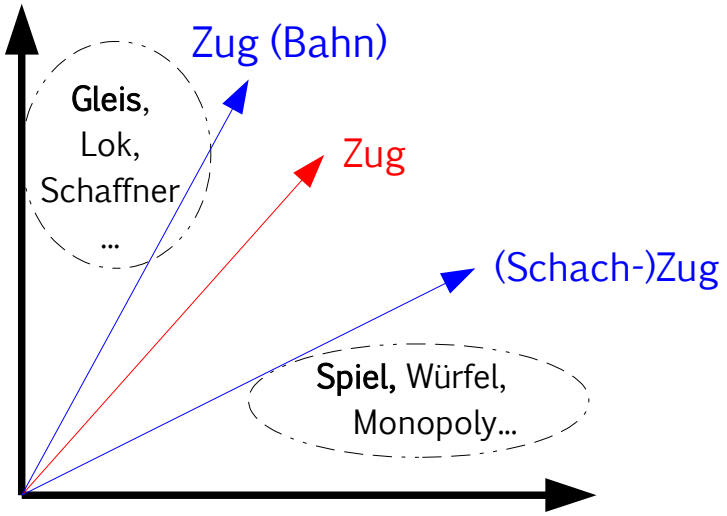
Problemstellung

Lösung: Vektor-basierte semantische Beschreibung eines Wortes – neues Modell

z.B. M&L '08



Schon sehr gute Ergebnisse, aber geht es besser?



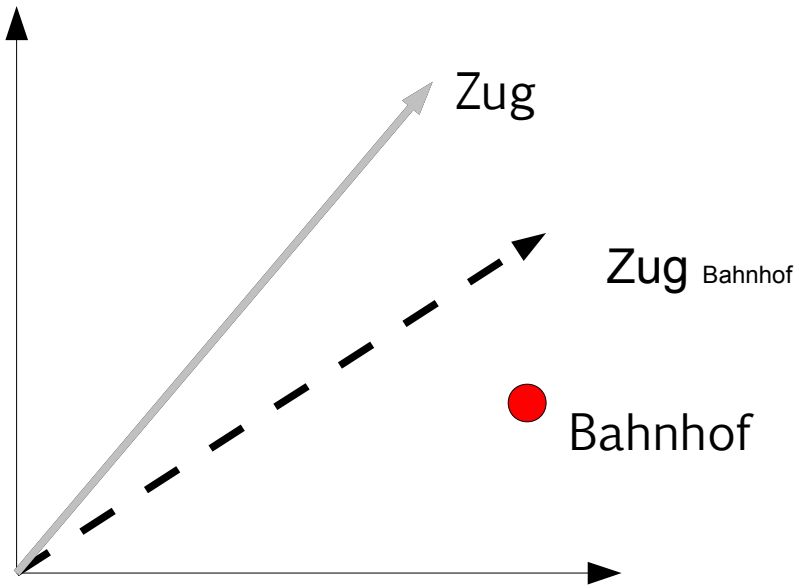
Ja! → „Semantische Ähnlichkeitswolken“

Gliederung und Inhalt

- **Allgemeines** – *Was bedeutet „kontextualisieren“?*
- **Das Modell**
  - → Grundlagen – *Übertragung auf das Modell*
  - → Formen der Kontextualisierung – *Welche Situationen gibt es zu beachten?*
  - → Modelltheorie und Formales - *Wie funktioniert das Modell?*
- **Evaluierung des Modells –** *Der Feldtest*
  - → Ordnen von Paraphrasen
  - → Wortbedeutungsdisambiguierung
- **Ausblick in die Zukunft –** *Was ist noch möglich?*

Allgemeines

- Zunächst: Bedeutungsvektor eines Wortes wird gewonnen durch Zählung über **alle** Bedeutungen des Wortes → 1 **Worttyp** pro Wort, 1 **Token**
- Jetzt: Versuch, **spezielle** Bedeutung eines Wortes in bestimmtem Kontext durch einen Vektor möglichst genau zu beschreiben
  - Realisierung durch Modifizierung des Hauptvektors
    → **mehrere Worttokens** pro Wort



Das Modell: Grundlagen

Vorgehensweise im Modell:

I. Neugewichtung durch Modifikation des Hauptvektors

II. Kontextualisierung durch Miteinberechnen semantischer Ähnlichkeit zwischen den Wörtern der Zielvektordimensionen und den tatsächlich im Kontext auftauchenden Wörtern

Das Modell: Grundlagen

- Einfach gesagt: Wort wird dargestellt durch abgeleiteten Vektor (sog. Hauptvektor), der auf Kontextbasis neu gewichtet wird, wobei syntaktische Relationen beachtet werden.

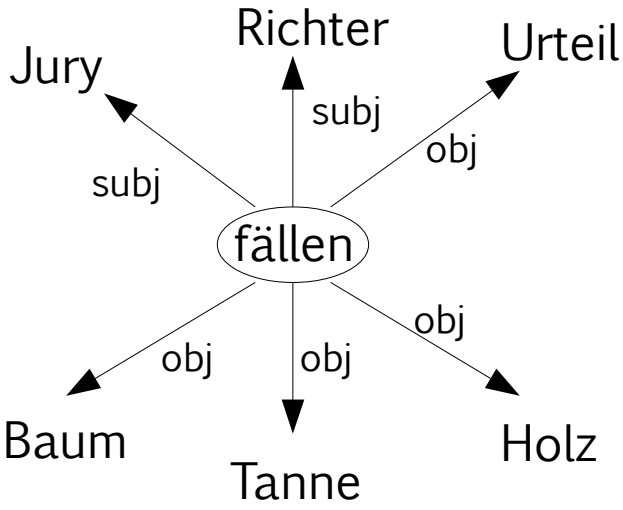
- Dimensionen des Vektors = alle möglichen Wörter, die jemals in Kookurrenz mit dem Zielwort auftreten (im Rahmen eines Korpus)

- im Folgenden zur Vereinfachung: Neugewichtung durch triviale absolute Anzahl der gemeinsamen Auftreten von Zielwort und Kontextwort, außerdem Betrachtung nur eines Kontextwortes

Das Modell : Formen der Kontextualisierung

1. Keine Kontextualisierung

nur Auszählung der Kontextwörter - gerade vorhandener Kontext spielt keine Rolle (Gewichtung = Auszählung)



	Jury subj	Richter subj	Urteil obj	Telefon subj	Tanne obj	Baum obj	Holz obj	Krieger obj
v(fällen)	5	11	23	0	13	30	12	2

Das Modell : Formen der Kontextualisierung

Einfache Formel:

$$v(w) := \sum_{r \in R, w' \in W} f(w, r, w') * e_{(r, w')}$$

Die Häufigkeit, mit der das beliebige Kontextwort w' im Kontext von w mit der Relation r vorkommt. Zur Verbesserung des Algorithmus' wird hier später die Pointwise Mutual Information als Gewichtungsmaß genutzt.

Der Basisvektor vom Wort w' in der Relation r.

→ alle Dimensionen bis auf die von w' haben den Wert 0, w' hat den Wert 1.

Das Modell : Formen der Kontextualisierung

Nachteil:

- „einen Baum fällen“ versus „eine Entscheidung fällen“ : fällen hat nicht die gleiche Bedeutung

- aber: Realisierung durch einen einzigen Vektor

  z.B. $v(\text{fällen}_{\text{Baum}}) =$ (5 11 23 0 13 30 12 2)

  $v(\text{fällen}_{\text{Entscheidung}}) =$ (5 11 23 0 13 30 12 2)

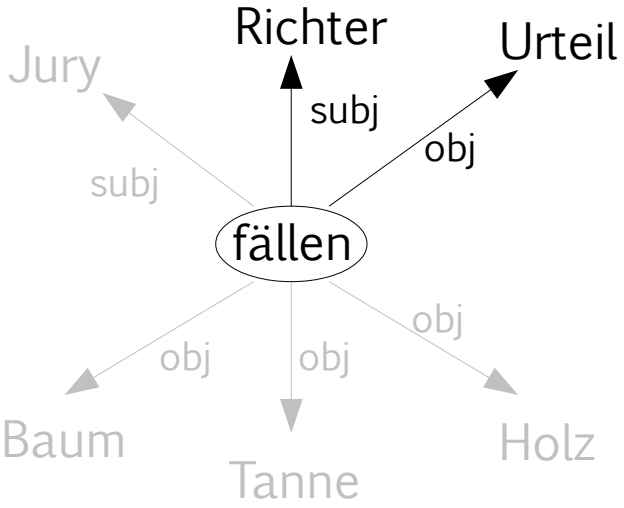
  $v(\text{fällen}_{\text{Baum}}) = (\text{fällen}_{\text{Entscheidung}})$!!

→ **Verbesserung durch Miteinbeziehen des Kontext**

Das Modell : Formen der Kontextualisierung

2. Strikte Kontextualisierung

Kontext:
„Nach reichlicher Überlegung fällte der **Richter** ein **Urteil**...“

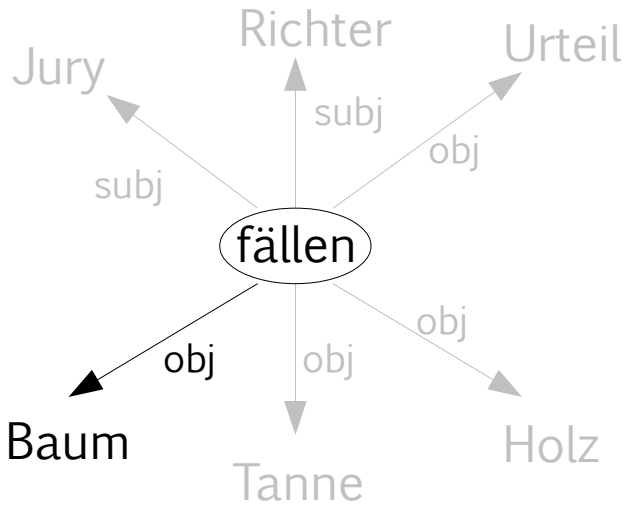


	Jury subj	Richter subj	Urteil obj	Telefon subj	Tanne obj	Baum obj	Holz obj	Krieger obj
v(fällen)	5 * 0	11 * 1	23 * 1	0 * 0	13 * 0	30 * 0	12 * 0	2 * 0

Das Modell : Formen der Kontextualisierung

2. Strikte Kontextualisierung

Anderer Kontext:
„Die Kreissäge fällt den **Baum**...“



	Jury subj	Richter subj	Urteil obj	Telefon subj	Tanne obj	Baum obj	Holz obj	Krieger obj
v(fällen)	5 * 0	11 * 0	23 * 0	0 * 0	13 * 0	30 * 1	12 * 0	2 * 0

Das Modell : Formales

Formel für strikte Kontextualisierung:

$$v(w) := \sum_{r \in R, w' \in W} f(w, r, w') * e_{(r, w')} * \alpha_{w', w_c}$$

w: Zielwort w': Wort der Vektordimension r: Relation zwischen w und w'
w_c: gerade auftretendes Kontextwort
α: **ist 1 oder 0**

→ **α = 1**, wenn w' = w_c (nur die Dimension, die **exakt** dem Kontextwort entspricht, spielt eine Rolle)
→ **α = 0** sonst

Das Modell : Formen der Kontextualisierung

Nachteil:

- „einen Baum fällen“ versus „eine Tanne fällen“ : fällen hat die gleiche Bedeutung

- im Modell aber Realisierung durch unterschiedliche, semantisch unähnliche Vektoren

  z.B. $v(\text{fällen}_{\text{Baum}}) = (0\ 0\ 0\ 0\ 0\ 30\ 0\ 0)$

  $v(\text{fällen}_{\text{Tanne}}) = (0\ 0\ 0\ 0\ 13\ 0\ 0\ 0)$

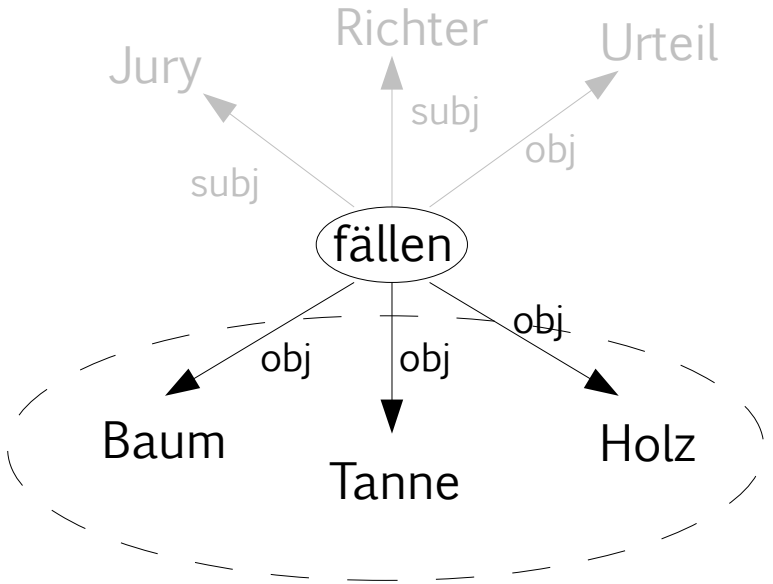
  $v(\text{fällen}_{\text{Baum}}) \neq (\text{fällen}_{\text{Tanne}})$!!

→ **Verbesserung durch Miteinberechnen semantischer Ähnlichkeit**

Das Modell : Formen der Kontextualisierung

3. Ähnlichkeitsbasierte Kontextualisierung

Kontext: „Die Kreissäge fällt den **Baum**...“



Umsetzung der informellen Regel: „fällen“ im Kontext

mit einem Wort ähnlich zu „Baum“

- was in gleicher Relation zu ihm steht

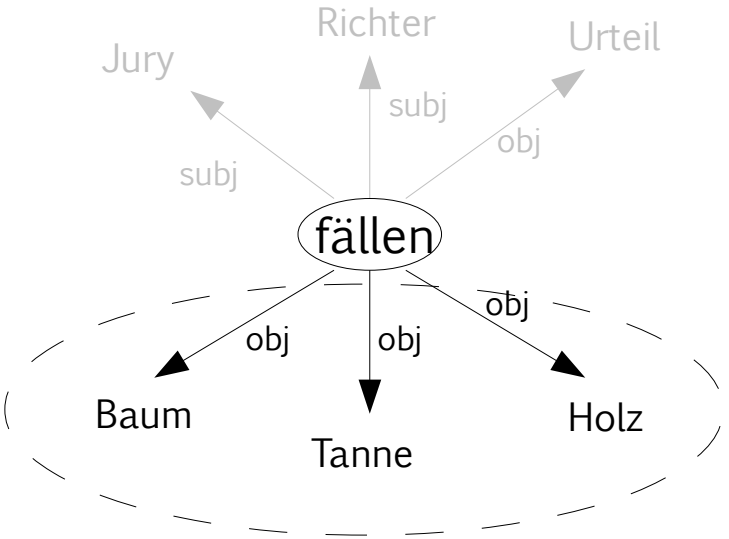
→ Jury, Richter fallen weg (subj)

Das Modell : Formen der Kontextualisierung

3. Ähnlichkeitsbasierte Kontextualisierung

Kontext: „Die Kreissäge fällt den **Baum**...“

Idee: Multiplikation jeder Vektordimension mit der Ähnlichkeit zwischen Dimension und Kontextwort - wenn Relation stimmt!



	Jury subj	Richter subj	Urteil obj	Telefon subj	Tanne obj	Baum obj	Holz obj	Krieger obj
v(fällen)	5 * 0	11 * 0	23 * 0	0 * 0	13 * 0,7	30 * 1	12 * 0,6	2 * 0

Das Modell : Formen der Kontextualisierung

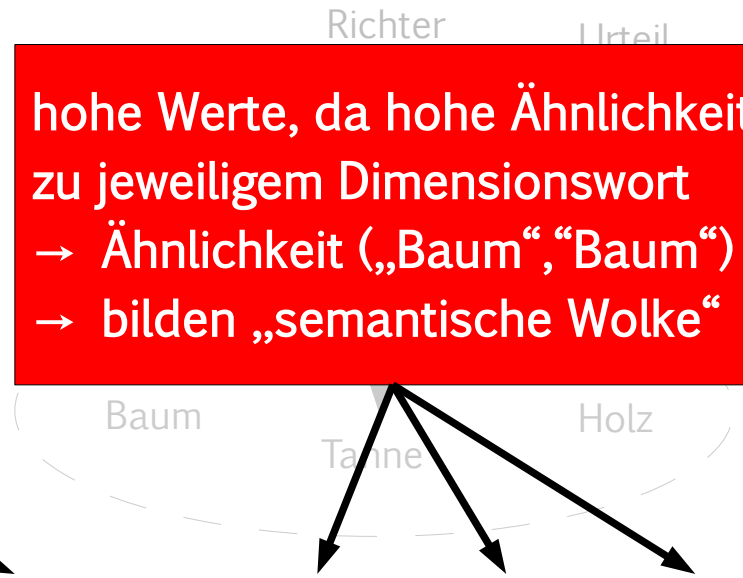
3. Ähnlichkeitsbasierte Kontextualisierung

Kontext: „Die Kreissäge fällt den Ba

Idee: Mu
Vektord
Ähnlichkeit zwischen Dimension und Kontextwort
- wenn Relation stimmt!

Nullwerte, da falsche syntaktische Relation

hohe Werte, da hohe Ähnlichkeit von „Baum“ zu jeweiligem Dimensionswort
→ Ähnlichkeit („Baum“,“Baum“) ist maximal!
→ bilden „semantische Wolke“

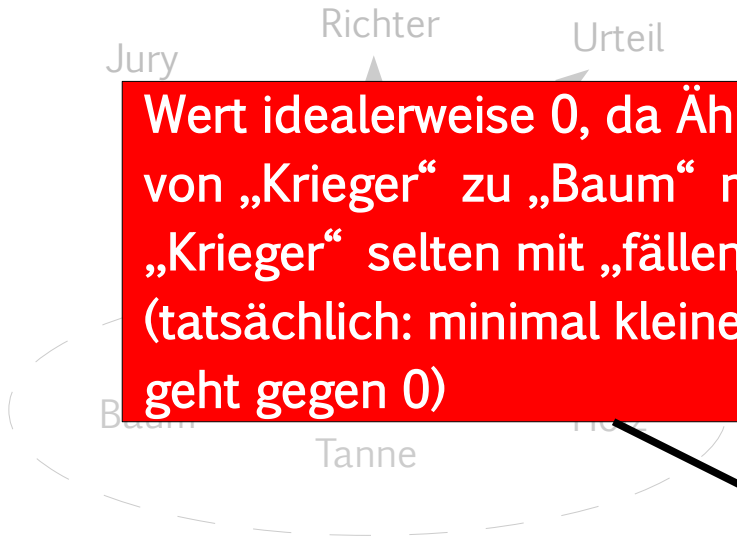


	Jury subj	Richter subj	Urteil obj	Telefon subj	Tanne obj	Baum obj	Holz obj	Krieger obj
v(fällen)	5 * 0	11 * 0	23 * 0	0 * 0	13 * 0,7	30 * 1	12 * 0,6	2 * 0

Das Modell : Formen der Kontextualisierung

3. Ähnlichkeitsbasierte Kontextualisierung

Kontext: „Die Kreissäge fällt den Baum.“



geringe Ähnlichkeit zwischen „Urteil“ und „Baum“, idealerweise 0 (tatsächlich: relativ geringer Wert)

Wert idealerweise 0, da Ähnlichkeit von „Krieger“ zu „Baum“ minimal und „Krieger“ selten mit „fällen“ auftaucht (tatsächlich: minimal kleiner Wert, geht gegen 0)

Dimension und Kontextwort - wenn Relation stimmt!

	Jury subj	Richter subj	Urteil obj	Telefon subj	Tanne obj	Baum obj	Holz obj	Krieger obj
v(fällen)	5 * 0	11 * 0	23 * 0	0 * 0	13 * 0,7	30 * 1	12 * 0,6	2 * 0

Das Modell : Formales

Volle Formel:

$$v(w) := \sum_{r \in R,\, w' \in W} f(w, r, w') * e_{(r, w')} * \alpha_{r_c, w_c, r, w'}$$

w: Zielwort w': Wort der Vektordimension r: Relation zwischen w und w'
w_c : auftretendes Kontextwort r_c : Relation zwischen w und w_c
α: Ähnlichkeit zwischen w_c und w'.

→ **α = sim(w_c,w') wenn r_c = r'**

→ **α = 0 sonst**

Als Ähnlichkeitsmaß wird schlicht die Kosinusähnlichkeit zwischen den Hauptvektoren von w_c und w' genommen.

Das Modell : Formales

Das funktioniert schon sehr gut, aber was ist mit mehreren Kontextwörtern?

Mittel der Wahl: Vektoraddition über alle vorhandenen Kontextwörter.

→ Berechne für jedes Kontextwort den semantisch angereicherten Vektor

→ addiere alle diese Vektoren auf

Beispielkontext: Er fällt den Baum und die Tanne, indem er….“

	Jury subj	Richter subj	Urteil obj	Telefon subj	Tanne obj	Baum obj	Holz obj	Krieger obj
$v(\text{fällen}_{\text{Baum}})$	5 * 0	11 * 0	23 * 0	0 * 0	13 * 0,7	30 * 1	12 * 0,6	2 * 0
$v(\text{fällen}_{\text{Tanne}})$	5 * 0	11 * 0	23 * 0	0 * 0	13 * 1	30 * 0,7	12 * 0,5	2 * 0
v(fällen)	**0**	**0**	**0**	**0**	**22,1**	**51**	**13,2**	**0**

Zwischenfazit

- Im Grunde Verschwendung von Ressourcen, wenn immer nur ein Kontextwort betrachtet wird und alle anderen Dimensionen nicht beachtet werden
  → strikte Kontextualisierung, viele Nulldimensionen

- Daher Bildung einer „semantischen Wolke“, die nicht nur das explizite Kontextwort betrachtet, sondern auch alle, die zu ihm ähnlich sind
  → „fällen“ hat im Kontext mit „Baum“ oder „Tanne“ dieselbe Bedeutung, was hierbei zum Ausdruck kommt

- Modell ist trotzdem sehr variabel, kann auch keine oder strikte Kontextualisierung simulieren, wenn benötigt
  → keine Kontextualisierung: **α** sei immer 1
  → strikte Kontextualisierung: **α** sei immer 0 außer in einem Fall, dann 1

Evaluierung – 1. Experiment

Anordnen von Paraphrasen - Erklärung

- Ausgangslage:
  - Wort w in einem Kontext
  - Wörter w1,.....wk : lexikalische Paraphrasen von w in einer seiner Bedeutungen

- Aufgabe: Anordnen von w1,.....wk nach semantischer Ähnlichkeit zu w

- **Beispiel : „Die Axt fällt den Baum.“**
  - **„fällen“ = w**
    „umhauen“ und „entscheiden“ = w1 und w2
  - **„fällen“ = „umhauen“ oder fällen = „entscheiden“?**
    → **„umhauen“ am Ende weiter oben als „entscheiden“**

Evaluierung – 1. Experiment

Anordnen von Paraphrasen – experimentelles Setup

- 197 Zielwörter, 1986 Instanzen der Wörter in unterschiedlichen Kontexten/Bedeutungen (~ 10 Instanzen pro Wort), nach dem Vorbild von SemVal

  → Ursprünglich Zweiteilung der Aufgabe: 1) Paraphrasen finden und 2) Paraphrasen ordnen

- Hier: Vereinfachung. Paraphrasen werden aus Goldstandard entnommen, nur Anordnung wird vorgenommen

- Erstellung des Goldstandards: 5 Versuchspersonen sollten sich Paraphrasen überlegen
  - → gewichtet, je nachdem wie oft eine Paraphrase gewählt wurde (Goldstandard)
  - → pro Instanz ~ 3,5 Paraphrasen
  - → alle Paraphrasen aller Bedeutungen zusammen bilden „Kandidaten“ für ein Wort (~ 17 Kandidaten pro Wort)

Evaluierung – 1. Experiment

Anordnen von Paraphrasen – experimentelles Setup

- In der Evaluierung: Pointwise Mutual Information statt absoluter Häufigkeit
  - → zur Erinnerung: PMI besser als Häufigkeit, da Zufall „rausgerechnet“ wird
  - → PMI = 3 bedeutet z.B. ein Vorkommen, 3 Mal häufiger als es der Zufall erwarten ließe

- Stanford Parser auf English Gigaword Corpus
  - → zur Verbesserung der Laufzeit Modifikation von Vektoren und Parsingbäumen: Streichen von zu seltenen Auftreten oder zu niedrigen PMIs...
  - → resultiert in 888 Mio. Baumtripeln und 28 Mio. Tripeltypen

Evaluierung – 1. Experiment

Anordnen von Paraphrasen – zur Auswertung

- Anordnen der Kandidaten, indem semantisch angereicherter Kontextvektor von Zielwort und Hauptvektor von Paraphrase verglichen werden

- Berechnung der Ähnlichkeit über Skalarprodukt

- Zum Vergleich mit dem Goldstandard: Generalzied Average Precision (Kishida, GAP)

  → nimmt Werte zwischen 0 und 1 an (0% und 100%)
  → hoher Wert: gute Paraphrasen stehen weit oben, bessere vor schlechteren
  → 1 bedeutet, dass alle Paraphrasen korrekt sortiert sind

Evaluierung – 1. Experiment

Anordnen von Paraphrasen – Auswertung

POS	Random	no Context	strict Context	sim.-based Context
Verb	27,4	38,4	41,6	48,8
Nomen	30,1	45,2	47,3	52,9
Adjektive	28,4	42,2	45,8	51,1
Adverbien	36,4	51,6	50,6	55,3
Alle	30	43,7	46	51,7

Evaluierung – 1. Experiment

Anordnen von Paraphrasen – Variation

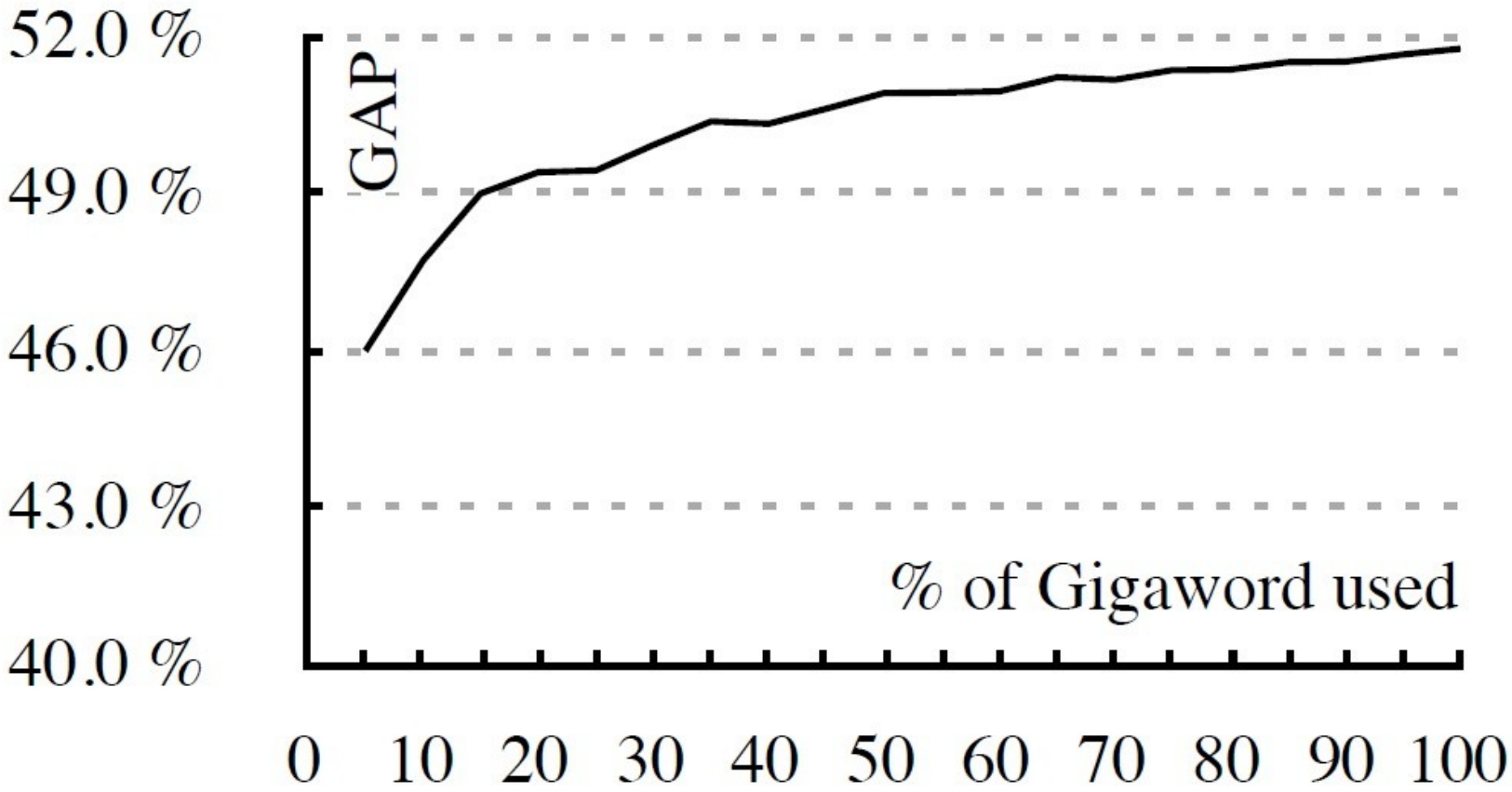
- Vergleich mit anderen Modellen nur bedingt möglich
  → im Experiment: sehr großer Korpus

- neue Evaluationen
  → nur Teile des Corpus werden verwendet
  → zuvor erzeugte Dependenzbäume werden zufällig angeordnet

- zum Vergleich: bei kleinstem Teil (5%) von Gigaword: Größe von ungefähr 2/3 des British National Corpus

Evaluierung – 1. Experiment

Anordnen von Paraphrasen – Lernkurve



Evaluierung – 2. Experiment

Wordsense – Disambiguierung – experimentelles Setup

- Benutzen von WordNet (großes semantisches Lexikon)

- Vorhersagen des korrekten, möglichst vollständigen WordNet-Sinnes eines Wortes im Kontext

Beispiel: Vorhersage für „Zug“ in:
„Der **Zug** kommt auf Gleis 4.“
→ unbelebt
→ Fahrzeug
→ öffentliches Verkehrsmittel...

Evaluierung – 2. Experiment

Wordsense – Disambiguierung – experimentelles Setup

- Durchführung ähnlich zu Experiment 1:

  1) Sammeln aller WordNet - Beschreibungen der **lexikalischen** Paraphrasen
  2) Ähnlichkeitsbestimmung zwischen kontextualisiertem Zielwortvektor und Bedeutungsbeschreibung der Kandidaten

- Variation: Miteinberechnen der a-priori-Wahrscheinlichkeit (+MFS, most frequent sense)

- Evaluierung auf SemEval Coarse-Grained English All-Words test set (2007)

Evaluierung – 2. Experiment

Wordsense – Disambiguierung – Auswertung

Modell	+MFS	-MFS
Random	52,4	52,4
Li et al. (2010)	81,3	78,8
vorgestelltes Modell	80,9	78,7
Kombination der Modelle	82,2	78,9

nahezu komplementär!

Gesamtfazit

- Kontextbasierte Modelle arbeiten besser als Modelle ohne Einbeziehung des Kontextes
- Ähnlichkeitsbasierte Kontextualisierung verschwendet weniger Potenzial als strikte Kontextualisierung durch Bilden „semantischer Wolken“
- Modell sehr flexibel, kann jede Form von Kontextualisierung simulieren
- Evaluierung bestätigt o.g. Hypothesen
- Größe des Corpus hat Einfluss auf Gesamtergebnis, aber Modell erweist sich trotzdem als am effektivsten

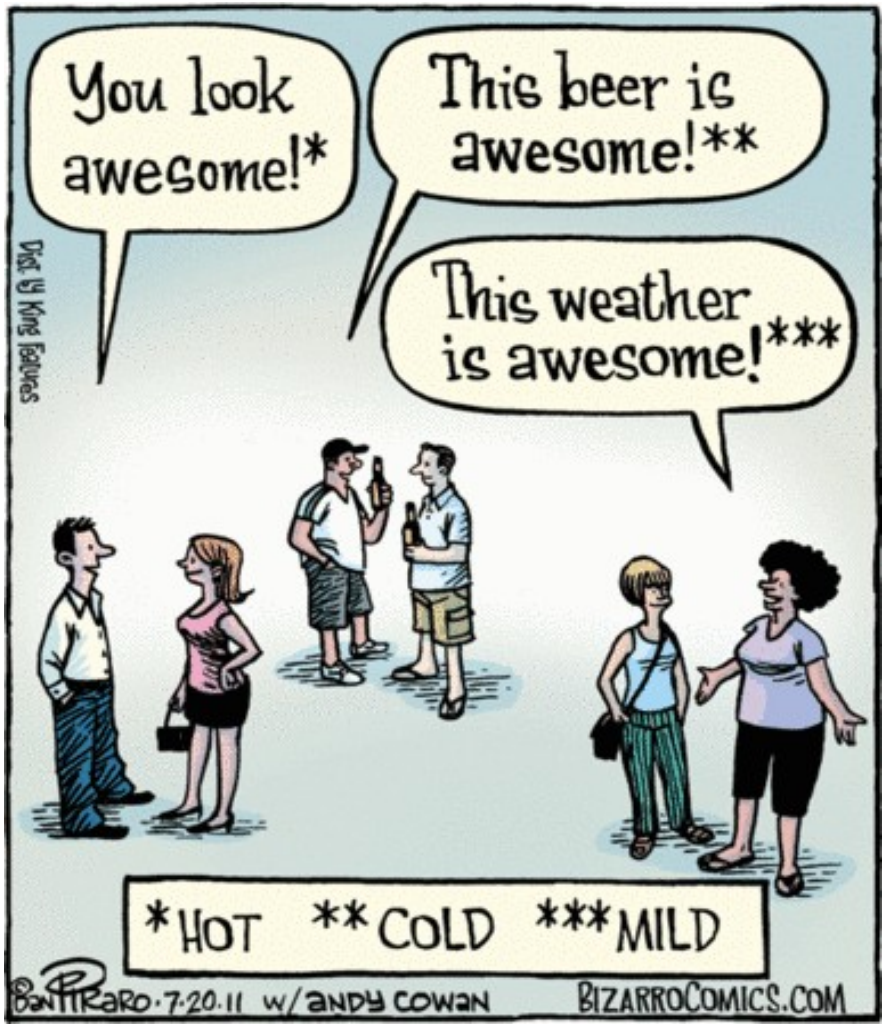
Ausblick in die Zukunft

- Im Modell: Einbeziehen des lokalen Kontextes eines Wortes
  → „*fällen* im Kontext von einem Wort wie *Baum*“

- Idee für die Zukunft: Rekursive Bestimmung von beliebigen Kontextwörtern, die im Kontext von anderen Wörtern stehen
  → „*fällen* im Kontext von einem Wort wie *Baum*, das im Kontext von einem Wort wie *Wald* steht, das in einem Kontext von einem Wort.....“

- Miteinbeziehen Syntaktischer Relationen verbessert das Modell – aber wäre nicht gleich eine semantische Information noch besser??
  → Miteinbeziehen semantischer Rollen von Wörtern in der Zukunft?

Fragen??


You look awesome!*
This beer is awesome!**
This weather is awesome!***
*HOT **COLD ***MILD
BIZARROCOMICS.COM