Automatic Word Sense Discrimination

Hinrich Schütze

Proseminar „Word Sense Disambiguation“
Stefan Thater

Carolyn Ladda
13. Dezember 2010

Word Sense Disambiguation

They buried him in his best suit.

The family brought suit against the landlord.

Word Sense Disambiguation

They buried him in his best suit.

(n) suit, suit of clothes
(a set of garments (usually including a jacket and trousers or skirt) for outerwear all of the same fabric and color)

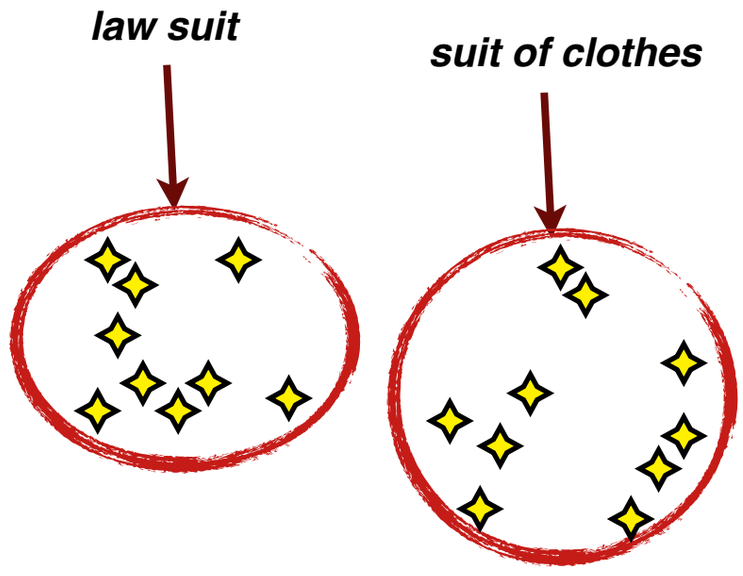
Word Sense Disambiguation

The family brought suit against the landlord.

(n) lawsuit, suit, case, cause, causa
(a comprehensive term for any proceeding in a court of law whereby an individual seeks a legal remedy)

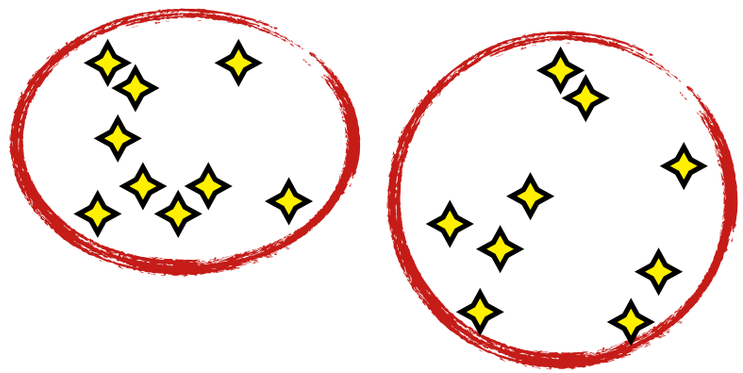
Word Sense Disambiguation

- 2 Teilschritte:
  - Sense Discrimination
  - Sense Labeling



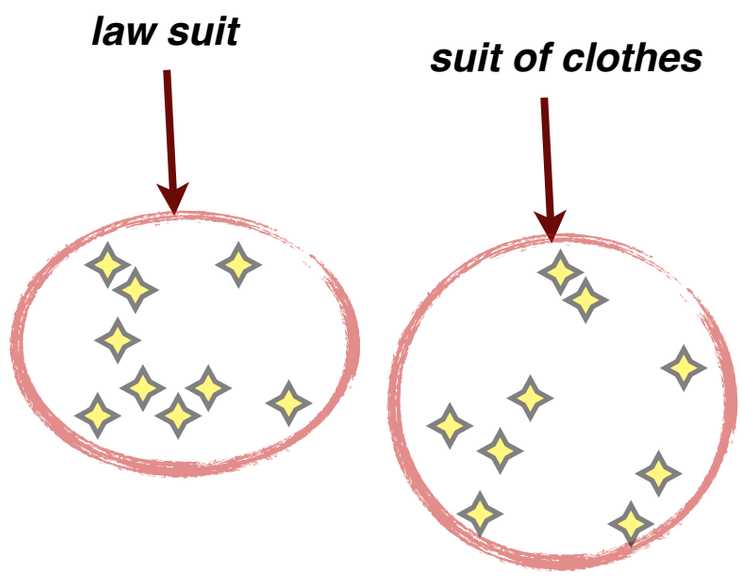
Sense Discrimination

- Aufteilung der Auftreten ambiger Wörter
- Bildung von Bedeutungsclustern



Sense Labeling

- Zuweisung einer Bedeutung für Auftreten ambiger Wörter
- Externe Wissensquelle notwendig



Automatic Word Sense Discrimination

- Benötigt keine:
  - externen Wissensquellen
  - Bedeutungsdefinitionen
- Anwendung im Bereich Information Access

Algorithmus

- Context-Group Discrimination
- Wortauftreten werden durch Kontextvektoren dargestellt
- Bedeutungen von Wortauftreten werden nach kontextueller Ähnlichkeit gruppiert

Hintergrund

Strong Contextual Hypothesis:

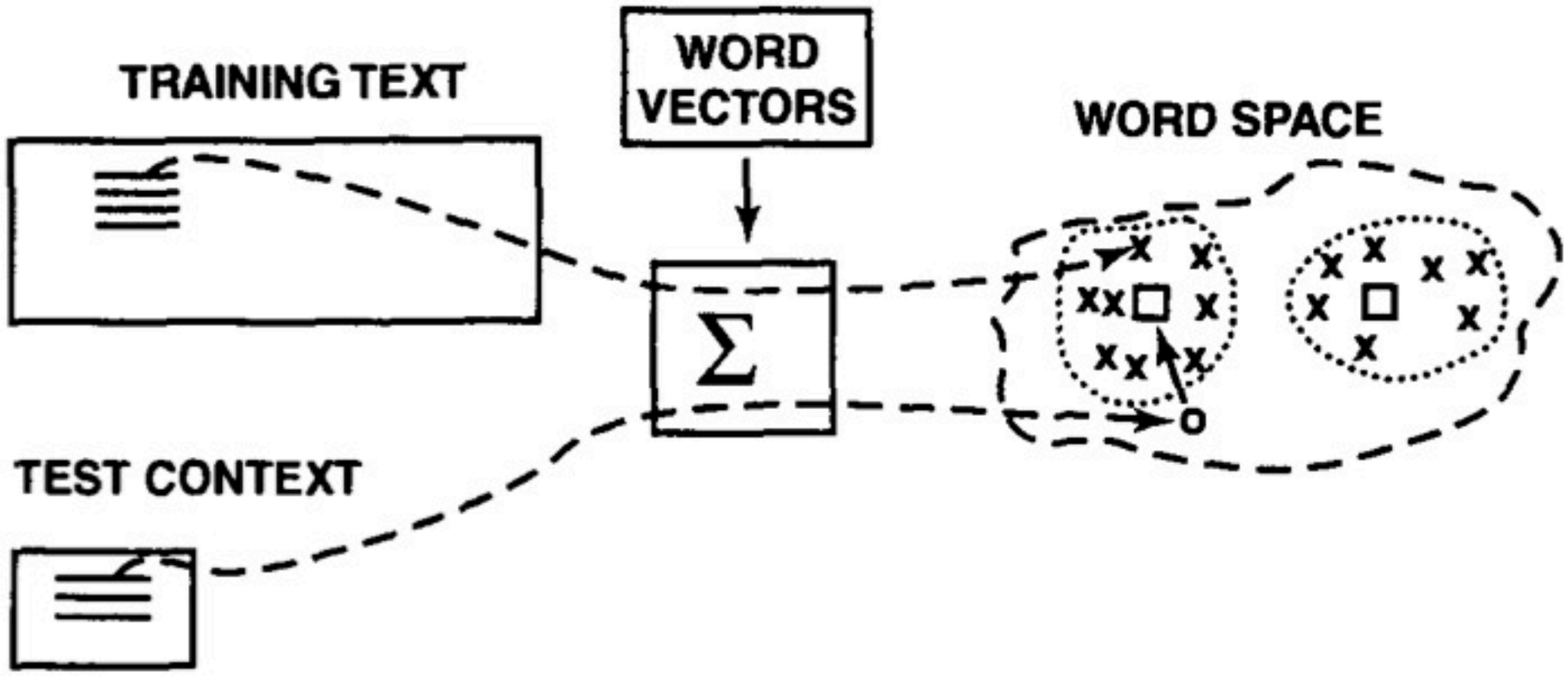
Two words are semantically similar to the extent that their contextual representations are similar.

Hintergrund

Contextual Hypothesis for Senses:

Two occurrences of an ambiguous word belong to the same sense to the extent that their contextual representations are similar.

Context-Group Discrimination



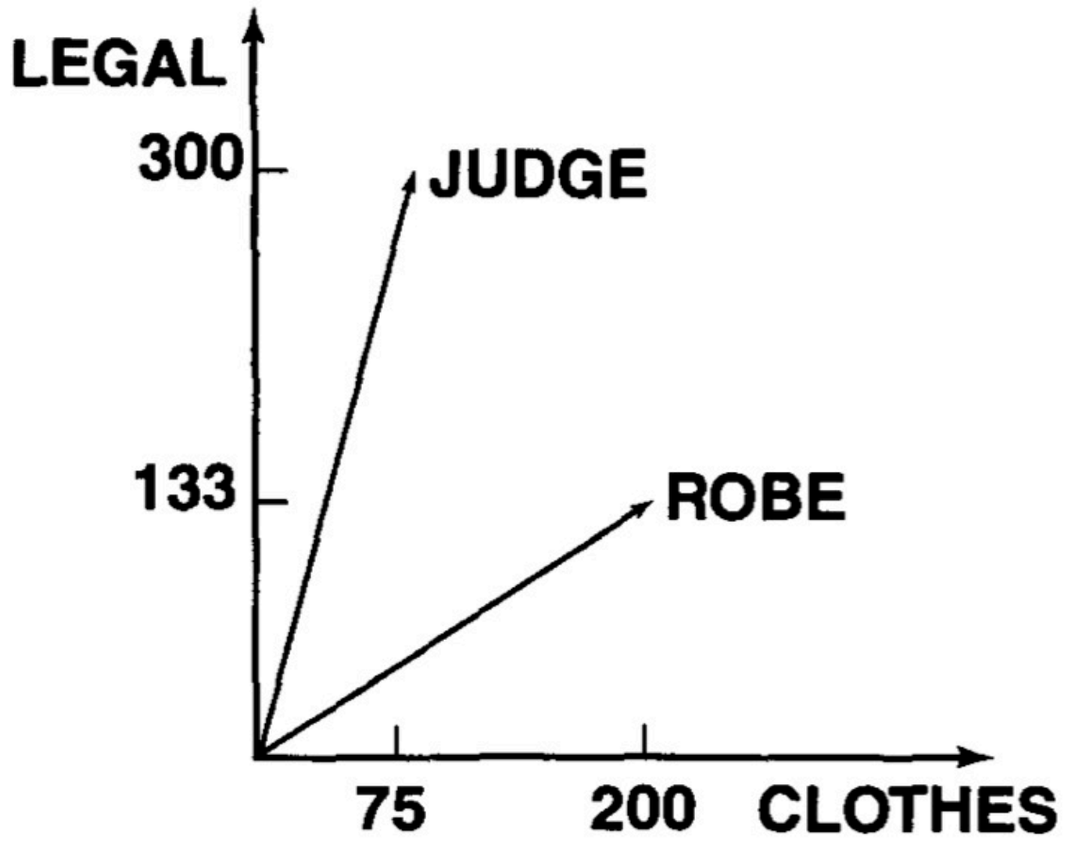
Vektoren für 3 Gebilde

- Worte
- Kontexte
- Bedeutungen

Multidimensionaler Vektorraum

- Übliche Repräsentation im Bereich Information Retrieval
- Dimensionen entsprechen Wörtern

Wortvektor



Wortvektor

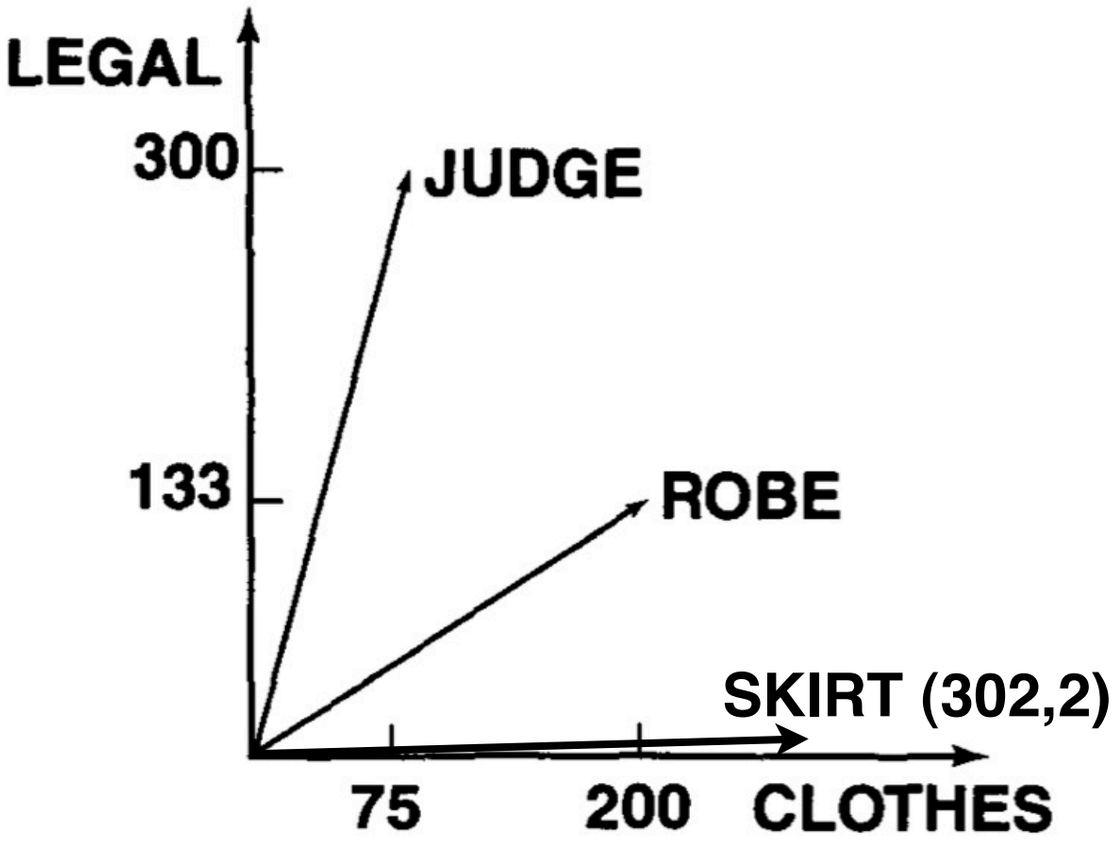
- ergibt sich aus Nachbarn im Korpus
  - im Satz oder in größerem Wortfenster
- Summe der benachbarten Auftreten aller Auftreten eines Wortes

Wortvektor

- verwandte Wörter haben ähnliche Kontexte
- Cosinus als semantisches Ähnlichkeitsmaß:

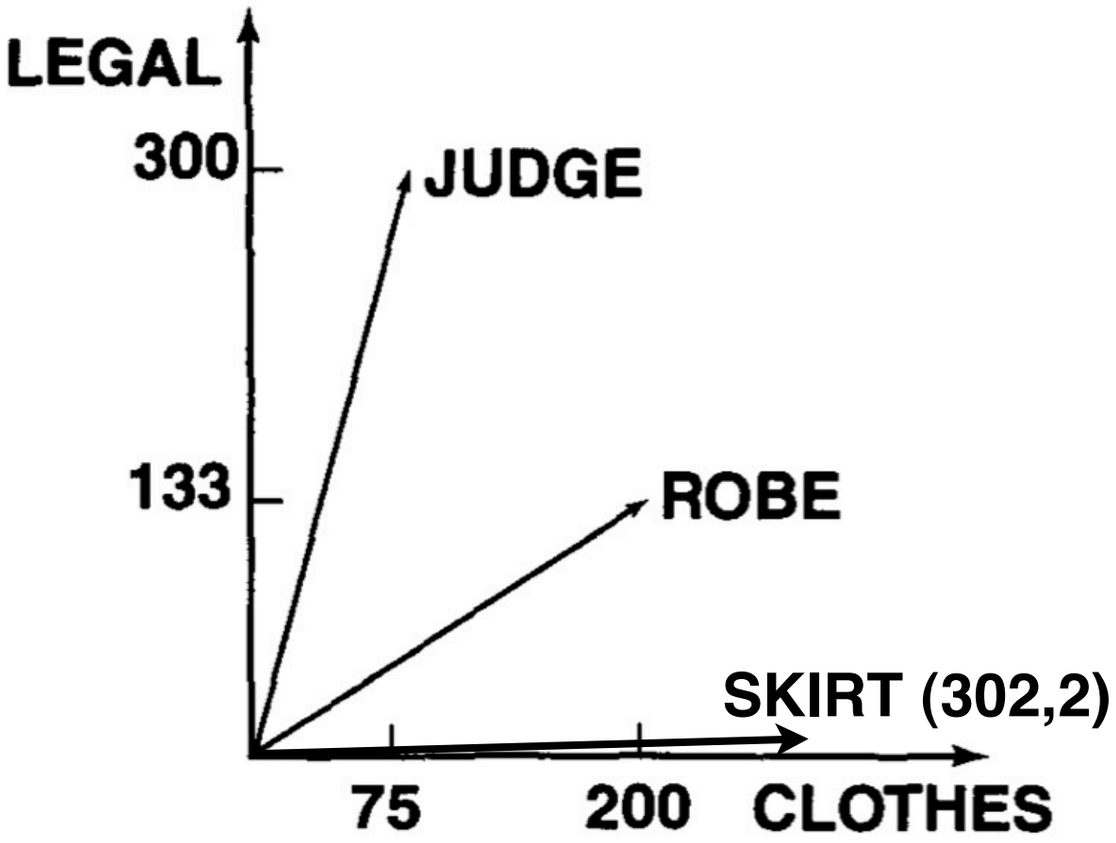
$$corr(\vec{v}, \vec{w}) = \frac{\sum_{i=1}^{N} v_i w_i}{\sqrt{\sum_{i=1}^{N} v_i^2 \sum_{i=1}^{N} w_i^2}}$$

Wortvektor



Dimension	Vector *judge*	*robe*
legal	300	133
clothes	75	200

Wortvektor



Dimension	Vector *judge*	*robe*
legal	300	133
clothes	75	200

$\cos(w_{judge}, w_{robe}) \approx 0{,}74$
$\cos(w_{robe}, w_{skirt}) \approx 0{,}84$
$\cos(w_{judge}, w_{skirt}) \approx 0{,}25$

Dimensionen der Wortvektoren

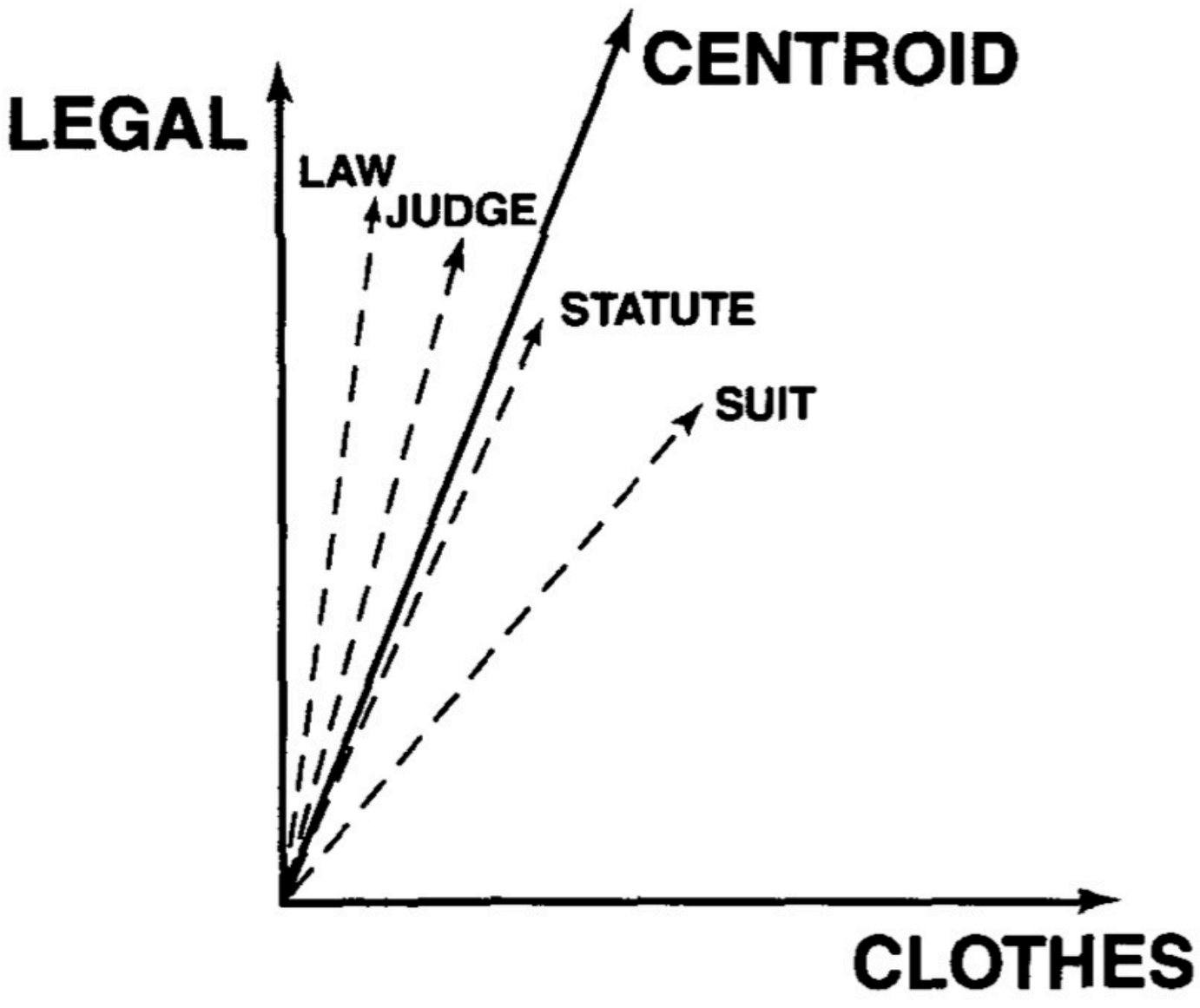
Die Wörter, die Dimensionen des Vektorraums sind, können auf unterschiedliche Arten bestimmt werden:

aus dem lokalen Kontext oder aus dem Gesamtkorpus unter Berücksichtigung der Anzahl ihrer Auftreten.

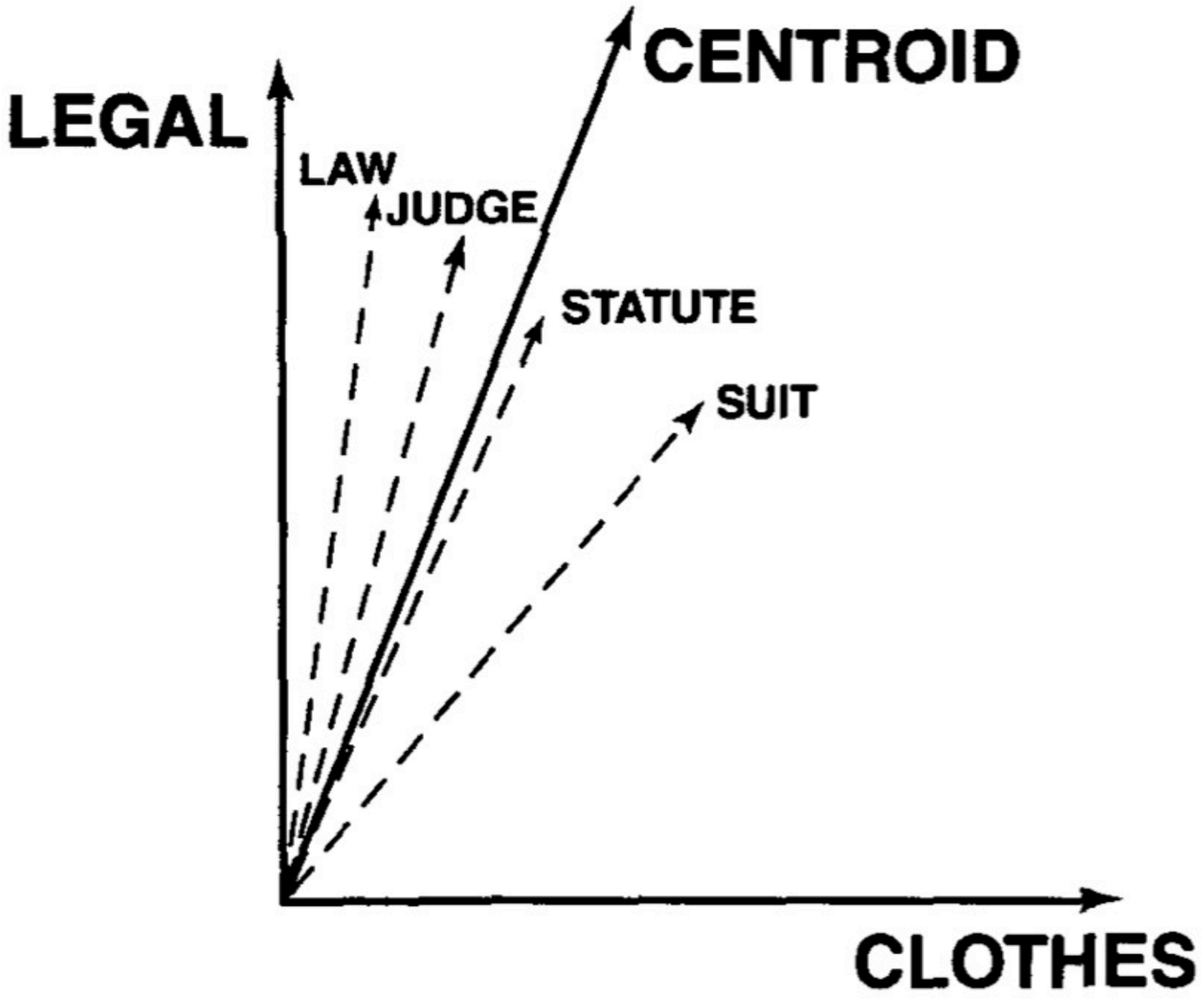
Kontextvektor

- repräsentiert einzelnes Wortauftreten
- ergibt sich aus entsprechenden Vektoren der Wörter im Kontext des Auftretens

Kontextvektor



Kontextvektor



[...]Common *law* systems place great weight on court decisions, which are considered "*law*" with the same force of *law* as *statutes*[...] a *judge* deciding a given case has more freedom to interpret the text of a *statute* independently[...] To initiate a *law suit*, a pleading had to be drafted to meet myriad technical requirements[...]

http://en.wikipedia.org/wiki/Common_law

Kontextvektor

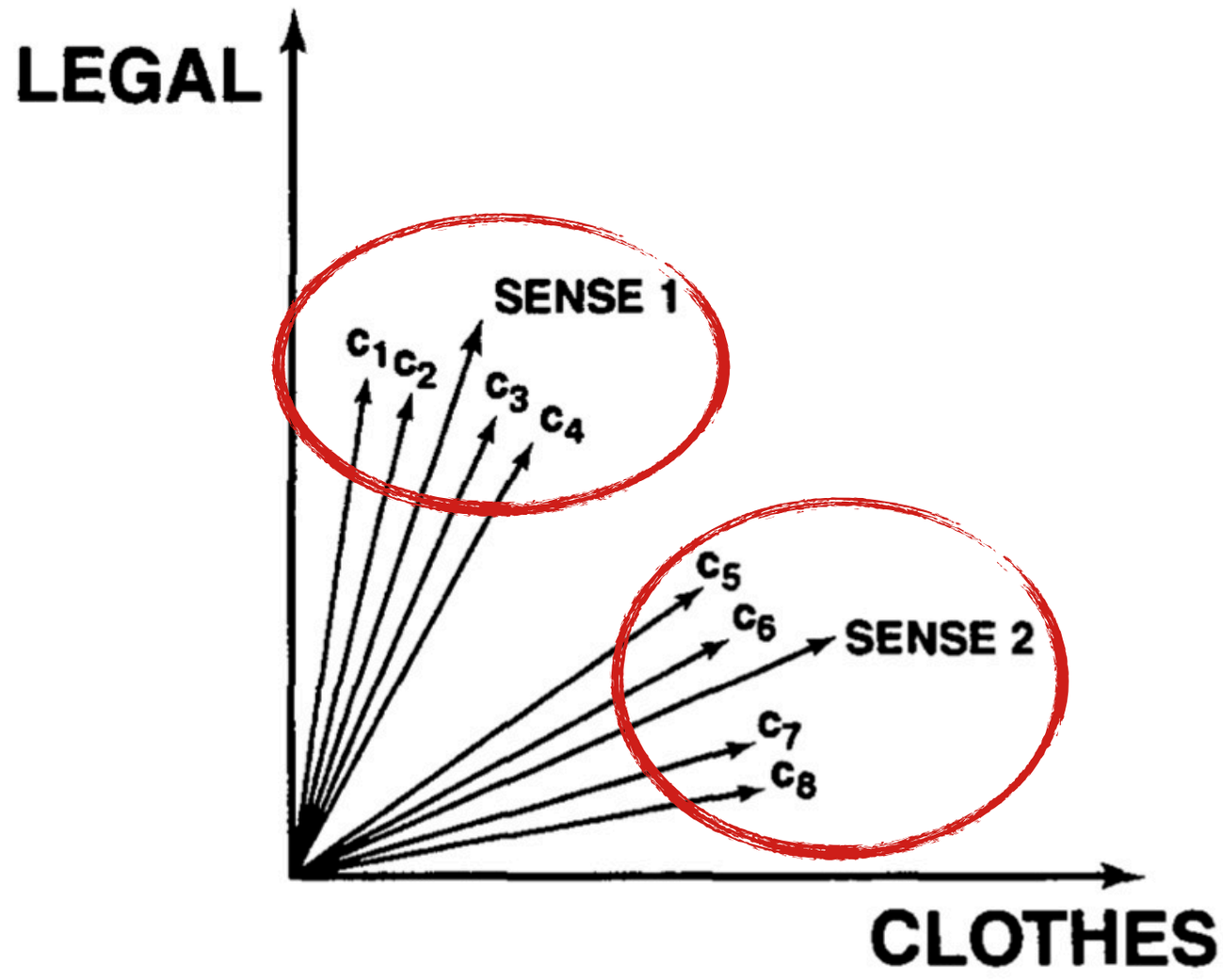
- ist Mittelpunkt der Wortvektoren
- d.h. er mittelt deren Richtung
- Wortvektoren w_i werden nach Diskriminierungspotential gewichtet:

$$a_i = \log\left(\frac{N}{n_i}\right)$$

Bedeutungsvektor

- Alle Auftreten eines ambigen Wortes werden in Bedeutungscluster eingeteilt
- Ein Bedeutungsvektor ist der Mittelpunkt eines solchen Clusters

Bedeutungsvektor



Bedeutungsvektor

Cluster sind optimal verteilt, wenn die Summe der Quadrate der Distanzen zwischen den Kontextvektoren und Bedeutungsvektoren minimal ist.

Bedeutungsvektor

- 2 Clusteringverfahren:
  - Group-Average Agglomerative Clustering für Startkonfiguration
  - Expectation-Maximization-Algorithmus für lokal optimale Lösung

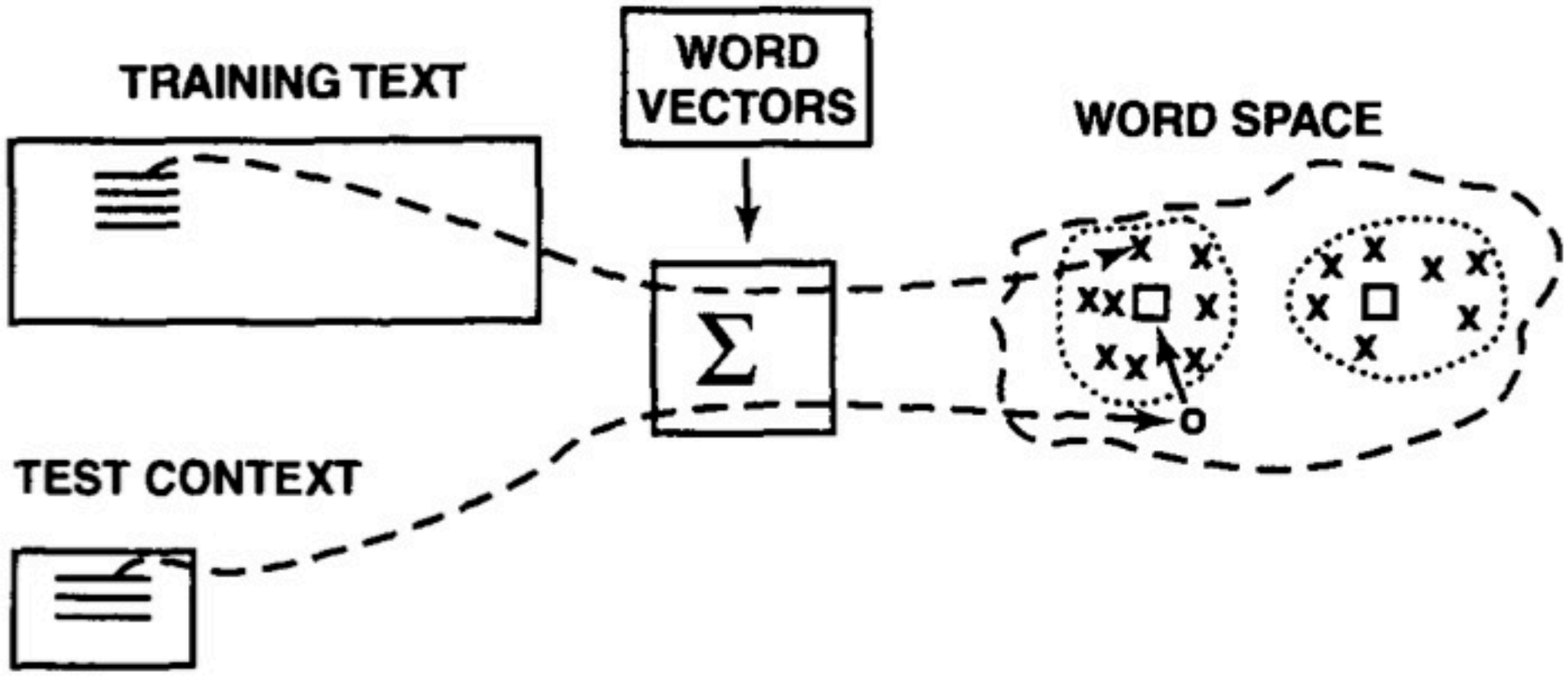
Optimierung

- Reduktion der Dimensionalität
  - Vektoroperationen schneller berechenbar
  - Hervorhebung grundlegender Eigenschaften
- Singular Value Decomposition

Anwendung

- Berechnung des Kontextvektors mit Hilfe von Wordvektoren
- Zuweisung zu einem der Bedeutungsvektoren des ambigen Wortes

Anwendung



Evaluation

- Training und Evaluation auf NYT-Korpus
- Trainingsdaten: 60,5 Millionen Wörter
- Testdaten: 5,4 Millionen Wörter

Evaluation

Word	Training	Test	Rare Senses	Baseline	Frequent Senses
ruling	5,482	200	3.5%	60%	an authoritative decision to exert control, or influence
space	9,136	200	0%	56%	area, volume outer space
suit/s	7,467	200	12.5%	57%	an action or process in a court a set of garments

Evaluation

Word	Training	Test.	Rare Senses	Baseline	Frequent Senses
wide range / consulting firm	1,422	149	0%	62%	wide range consulting firm
heart disease / reserve board	1,197	115	0%	54%	heart disease reserve board
urban development / cease fire	1,582	101	0%	50%	urban development cease fire

Evaluation

Level	Average of $2 \times sin^{-1}(\sqrt{X})$	Difference from Closest	Corresponding Accuracy
local, χ^2, terms	2.11	0.13	76%
local, frequency, terms	2.24	0.13	81%
local, frequency, SVD	2.44	0.06	88%
local, χ^2, SVD	2.50	0.06	90%
global, frequency, SVD	2.66	0.16	94%

Zusammenfassung

- Word Sense Discrimination basierend auf second-order Kookkurrenz
- Modellierung von Kontext mit Vektoren
- Gute Ergebnisse für ambige Wörter mit zwei Hauptbedeutungen aus unterschiedlichen Themengebieten

Offene Fragen

- Wörter mit:
  - mehr als zwei Hauptbedeutungen
  - Nebenbedeutungen
- Kontextübergreifende Wörter

Vielen Dank für Eure Aufmerksamkeit!

Noch Fragen?