Seminar
Distributionelle Semantik

Stefan Thater
FR 4.7 Allgemeine Linguistik (Computerlinguistik)
Universität des Saarlandes

Wintersemester 2011/12

Semantische Ähnlichkeit

- Fundamentale Aufgabe für semantische Modelle:
  - Wie ähnlich sind zwei Wörter (Bedeutungen) w und w’?
- Einigen Anwendungen (Turney & Pantel, 2010):
  - Automatische Erzeugung von Thesauri
  - Disambiguierung mehrdeutiger Wörter
  - Semantic Role Labelling
  - Query Expansion
  - …

Ähnlichkeit vs. Relatedness

- Zwei Wörter sind **semantisch ähnlich**, wenn die von den Wörtern bezeichneten Objekte ähnlich sind.
  - *Tasse – Becher*
- Semantische „**Relatedness**“ ist eine weniger strikte Beziehung als semantische Ähnlichkeit:
  - *Tasse – Kaffee*

Distributionelle Hypothese

- Der Kontext eines sprachlichen Ausdrucks enthält Informationen über die Bedeutung des Ausdrucks.
- **Distributionelle Hypothese:**

  If we consider words or morphemes A and B to be more different in meaning than A and C, then we will often find that the distributions of A and B are more different than the distributions of A and C. (Harris, 1954)

  You shall know a word by the company it keeps

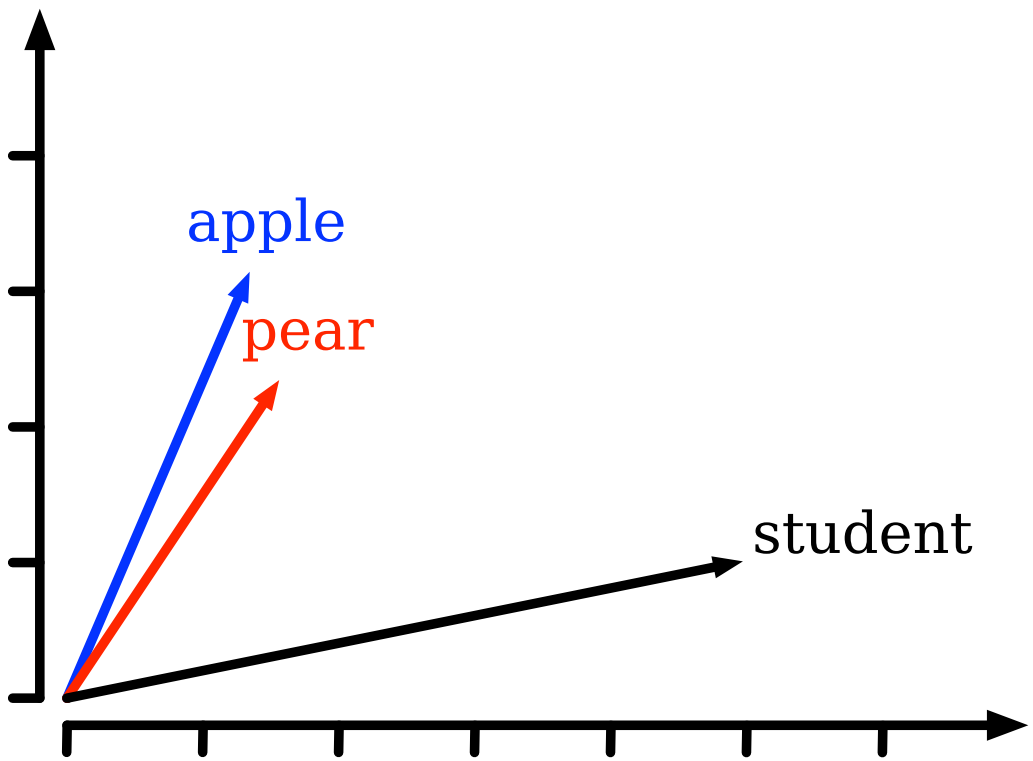
  (Firth, 1957)

Vektorraum-Modell

- Wortbedeutung wird als Vektor repräsentiert.
- Vektoren kodieren die statistische Verteilung des Wortes über relevante sprachliche Kontexte.
- Vektoren = Punkte im „semantischen Raum“
- Semantische Ähnlichkeit ≈ Distanz zwischen Vektoren

Vektorraum-Modell

- Kontextvektoren als Punkte im „semantischen Raum“



Kontext

- Kontext ≈ Kookkurrenz
- Verschiedene Arten von Kookkurrenz:
  - Wörter im Satz, Absatz, Dokument
  - Wörter in einem festen Wortfenster
  - Wörter in bestimmten syntaktischen Beziehungen
  - Muster-basiert
  - etc.

(Text aus Wikipedia)

Beispiel (Wortfenster)

The apple is the pomaceous fruit of the apple tree, species Malus domestica in the rose family (Rosaceae). […] There are more than 7,500 known cultivars of apples, resulting in a range of desired characteristics. […] The apple forms a tree that is small and deciduous, reaching 3 to 12 metres (9.8 to 39 ft) tall, with a broad, often densely twiggy crown. […] The apple tree was perhaps the earliest tree to be cultivated, and its fruits have been improved through selection over thousands of years. […]

(Text aus Wikipedia)

Beispiel (Wortfenster)

The **apple** is the pomaceous fruit of the apple tree, species Malus domestica in the rose family (Rosaceae). […] There are more than 7,500 known cultivars of apples, resulting in a range of desired characteristics. […] The apple forms a tree that is small and deciduous, reaching 3 to 12 metres (9.8 to 39 ft) tall, with a broad, often densely twiggy crown. […] The apple tree was perhaps the earliest tree to be cultivated, and its fruits have been improved through selection over thousands of years. […]

(Text aus Wikipedia)

Beispiel (Wortfenster)

The apple is the pomaceous fruit of the **apple** tree, species Malus domestica in the rose family (Rosaceae). […] There are more than 7,500 known cultivars of apples, resulting in a range of desired characteristics. […] The apple forms a tree that is small and deciduous, reaching 3 to 12 metres (9.8 to 39 ft) tall, with a broad, often densely twiggy crown. […] The apple tree was perhaps the earliest tree to be cultivated, and its fruits have been improved through selection over thousands of years. […]

(Text aus Wikipedia)

Beispiel (Wortfenster)

The apple is the pomaceous fruit of the apple fruit, species Malus domestica in the rose family (Rosaceae). […] There are more than 7,500 known cultivars of **apples**, resulting in a range of desired characteristics. […] The apple forms a tree that is small and deciduous, reaching 3 to 12 metres (9.8 to 39 ft) tall, with a broad, often densely twiggy crown. […] The apple tree was perhaps the earliest tree to be cultivated, and its fruits have been improved through selection over thousands of years. […]

(Text aus Wikipedia)

Beispiel (Wortfenster)

The apple is the pomaceous fruit of the apple tree, species Malus domestica in the rose family (Rosaceae). […] There are more than 7,500 known cultivars of apples, resulting in a range of desired characteristics. […] The apple forms a tree that is small and deciduous, reaching 3 to 12 metres (9.8 to 39 ft) tall, with a broad, often densely twiggy crown
be cul
select

	tree	fruit	forms	perhaps	apple	…
…						
apple	3	2	1	1	0	…
tree	0	1	1	1	3	…
…						

(Text aus Wikipedia)

Beispiel (Wortfenster)

The apple is the pomaceous fruit of the apple tree, species
). […] There
s, resulting in
ple forms a
o 12 metres
(9.8 to 39 ft) tall, with a broad, often densely twiggy
crown
be cul
select

Zeilen

das Vokabular, an dem wir interessiert sind

Spalten

„Eigenschaften“ des Vokabulars

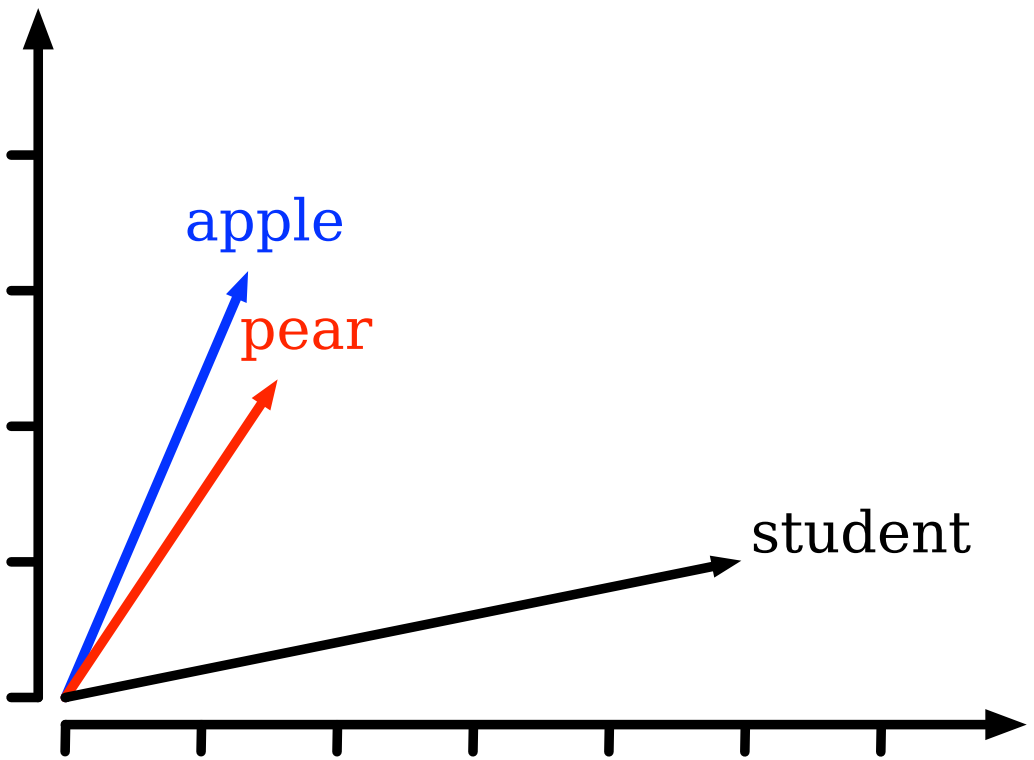
	tree	fruit	forms	perhaps	apple	…
…						
apple	3	2	1	1	0	…
tree	0	1	1	1	3	…
…						

Kontext-Vektoren

- $v_{apple} = \langle 3, 2, 1, 1, 0, \ldots \rangle$
- $v_{apple} = \sum_w f(apple, w)\, e_w$
  - $e_{tree} = \langle 1, 0, 0, 0, \ldots \rangle$
  - $e_{fruit} = \langle 0, 1, 0, 0, \ldots \rangle$
  - …

	tree	fruit	forms	perhaps	apple	…
…						
apple	3	2	1	1	0	…
tree	0	1	1	1	3	…
…						

Semantische Ähnlichkeit



Ähnlichkeitsmaße

- Ein Standardmaß für die semantische Ähnlichkeit ist der Kosinus des Winkels zwischen den Vektoren
- Kosinus = 1

  ⇒ Vektoren zeigen in die gleiche Richtung
- Kosinus = 0

  ⇒ Vektoren orthogonal (rechtwinklig)

$$cos(\vec{x}, \vec{y}) = \frac{\sum_{i=1}^{k} x_i y_i}{\sqrt{\sum_{i=1}^{k} x_i^2} \sqrt{\sum_{i=1}^{k} y_i^2}}$$

Weitere Ähnlichkeitsmaße

Distanzmaße (kleiner Wert = große Ähnlichkeit)

- Euklidische Distanz

Korrelationsmaße (großer Wert = große Ähnlichkeit)

- Kosinus
- Skalarprodukt
  (für normalisierte Vektoren äquivalent zu Kosinus)
- Dice-Koeffizient
- Jaccard-Koeffizient
- Overlap-Koeffizient
- ...

(Text aus Wikipedia)

Varianten

The apple is the pomaceous fruit of the apple tree, species Malus domestica in the rose family (Rosaceae). […] There are more than 7,500 known cultivars of apples, resulting in a range of desired characteristics. […] The apple forms a tree that is small and deciduous, reaching 3 to 12 metres (9.8 to 39 ft) tall, with a broad, often densely twiggy crown
be cul
select

	tree	fruit	forms	perhaps	apple	…
…						
apple	3	2	1	1	0	…
tree	0	1	1	1	3	…
…						

Varianten

- Vorverarbeitung:
  - Vollformen vs. Lemmata
  - Mit oder ohne Funktionswörter („Stop words“)
  - …
- Verschiedene Arten von Kontexten
  - Wortfenster vs. syntaktische Struktur vs. …
- Gewichte:
  - Häufigkeiten vs. Wahrscheinlichkeiten vs. Pointwise Mutual Information vs. …

Problem: Mehrdeutigkeit

- Vektoren kodieren alle Kontexte eines Wortes ohne seine Bedeutung im jeweiligen Kontext zu berücksichtigen.
  - *Die in einer **Batterie** gespeicherte elektrische Ladung wird umgangssprachlich als Kapazität bezeichnet, das ist nicht zu verwechseln mit der elektrischen Kapazität.*
  - *Die **Batterie** ist bei der Artillerie der Bundeswehr normalerweise in zwei schießende Züge zu vier Geschützen oder Werfern […] gegliedert.*
  - *In Deutschland werden ca. 90 Prozent der Hühner in **Batterien** gehalten.*
- ⇒ „unsaubere“ Vektoren

Problem: Mehrdeutigkeit

- Vektoren kodieren alle Kontexte eines Wortes ohne seine Bedeutung im jeweiligen Kontext zu berücksichtigen.
- ⇒ Wie kann man Vektor-Repräsentationen „kontextualisieren“ (disambiguiren), so dass nur die „richtige“ Bedeutung kodiert wird?

Problem: Kompositionalität

- Vektoren kodieren nur die Kontexte einer endlichen Menge festgelegter Ausdrücken (typischerweise Wörter)
- **Aber:** Bedeutung wird typischerweise auf Satzebene kodiert.
- ⇒ Wie kann man geeignete Vektoren für komplexe Ausdrücke aus ihren Teilausdrücken berechnen? Geht das überhaupt?

Organisatorisches

Organisatorisches

- **Prüfungsleistungen**
  - Vortrag (etwa 45 Minuten)
  - Seminararbeit (etwa 15 Seiten)
- **Weitere Prüfungsleistungen**
  - Aktive Teilnahme (Diskussionsbeiträge)
  - 1x vorbereitete Fragen
- **Gewichtung**
  - Vortrag und Seminararbeit je 50%
  - Liegt der Durchschnitt zwischen zwei Noten, geben die weiteren Prüfungsleistungen den Ausschlag

Organisatorisches

- **Mündliche Prüfung:**
  - wird mit 20% gewichtet
  - (Vortrag und Hausarbeit dann entsprechend je 40%)
- Beachte: In drei Seminaren sind mündliche Prüfungen zusätzlich zu Vortrag und Hausarbeit abzulegen.

Themen

- Jeder Teilnehmer wählt eine Hauptquelle
  - überwiegend Konferenzpapiere, 8 Seiten, Englisch
- Diese Hauptquelle ist Gegenstand des Vortrags und der Seminararbeit
- Soweit inhaltlich erforderlich sollte auch weitere Literatur diskutiert werden
  - eigenständige (!) Literaturrecherche

Zeitplan

- Zwei Wochen vor dem Vortrag
  - Vorbesprechung zur Klärung inhaltlicher Fragen
- Eine Woche vor dem Vortrag
  - Feedback zu den Folien

Literatur

- Jeff Mitchell and Mirella Lapata (2008). Vector-based Models of Semantic Composition.
- Katrin Erk and Sebastian Padó (2008). A Structured Vector Space Model for Word Meaning in Context.
- Stefan Thater, Hagen Fürstenau and Manfred Pinkal (2010). Contextualizing Semantic Representations Using Syntactically Enriched Vector Models.
- Stefan Thater, Hagen Fürstenau and Manfred Pinkal (2011). Word Meaning in Context: A Simple and Effective Vector Model.

Literatur

- Katrin Erk and Sebastian Padó (2010). Exemplar-Based Models for Word Meaning in Context.
- Joseph Reisinger and Raymond J. Mooney (2010). Multi-Prototype Vector-Space Models of Word Meaning.

Literatur

- Georgiana Dinu and Mirella Lapata (2010). Measuring Distributional Similarity in Context.
- Diarmuid Ó Séaghdha and Anna Korhonen (2011). Probabilistic models of similarity in syntactic context.
- Tim Van de Cruys, Thierry Poibeau and Anna Korhonen (2011). Latent Vector Weighting for Word Meaning in Context.

Literatur

- Matthias Hartung and Anette Frank (2010). A Structured Vector Space Model for Hidden Attribute Meaning in Adjective-Noun Phrases.
- Marco Baroni and Roberto Zamparelli (2010). Nouns are Vectors, Adjectives are Matrices: Representing Adjective-Noun Constructions in Semantic Space.

Literatur

- Edward Grefenstette and Mehrnoosh Sadrzadeh (2011). Experimental Support for a Categorical Compositional Distributional Model of Meaning.
- Edward Grefenstette, Mehrnoosh Sadrzadeh, Stephen Clark, Bob Coecke and Stephen Pulman (2011). Concrete Sentence Spaces for Compositional Distributional Models of Meaning.

Zeitplan

2011-10-24	Einführung	Thater
2011-10-31	Themenvergabe	Thater
2011-11-07	– Konferenz –	–
2011-11-14	Mitchell & Lapata (2008)	
2011-11-21	Erk & Padó (2008)	
2011-11-28	Thater, Fürstenau & Pinkal (2010, 2011) [eins davon]	
2011-12-05	Reisinger & Mooney (2010)	
2011-12-12	Dinu & Lapata (2010)	
2011-12-19	Ó Séaghdha & Korhonen (2011)	
2012-01-09	Van de Cruys &al. (2011)	
2012-01-16	Baroni & Zamparelli (2010)	
2012-01-23	Hartung & Frank (2010)	
2012-01-30	Grevenstette & Sadrzadeh (2011)	
2012-02-06	Abschlussdiskussion	Thater

Weitere Literatur

- Sebastian Padó and Mirella Lapata (2007). Dependency-Based Construction of Semantic Space Models.
- Walter Kintsch (2001). Predication.
- Stephen Wu and William Schuler (2011). Structured Composition of Semantic Vectors. IWCS 2011.

Nächste Sitzung

- Themenvergabe
- Wie halte ich einen guten Vortrag?

Danksagung

- Einzelne Folien sind inspiriert durch:
- Diarmuid Ó Séaghdha. Distributional approaches to semantic analysis. HIT-MSRA Summer Workshop on Human Language Technology. 2011. [PDF]