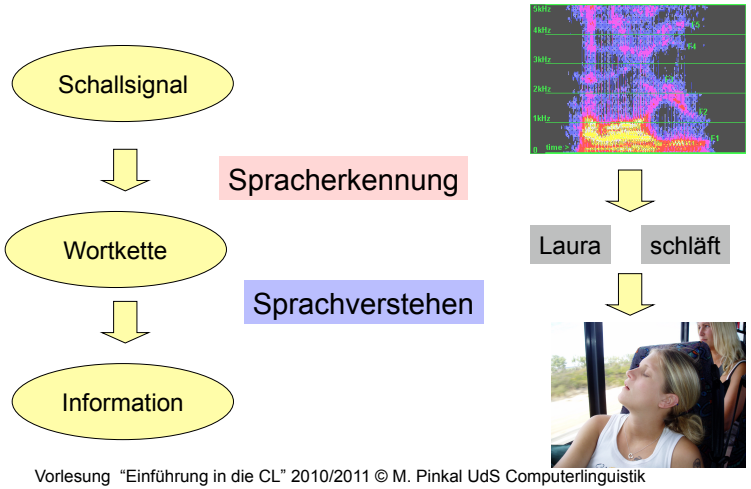
Einführung in die Computerlinguistik

Semantik

WS 2010/2011

Manfred Pinkal



Semantik

- Semantik ist der Teilbereich der (Computer-)Linguistik, der sich mit sprachlicher Bedeutung befasst.
- Wir unterscheiden nach den Ebenen der sprachlichen Gliederung zwischen Wortsemantik, Satzsemantik und Diskurssemantik
- Semantik hat einen Doppelcharakter: Sie ist Teil der Grammatik, und sie ist Schnittstelle zwischen Sprache und Welt.

- Semantik ist Teil der Grammatik: Wie berechnen wir, auf der Grundlage von Lexikon und Satzstruktur, die Bedeutung von Sätzen und Texten?
- Semantische Verarbeitung modelliert die Art und Weise, wie Information über die Welt kodiert, intern verarbeitet und (in Texten und Dialogen) kommuniziert wird.

Semantik

- Semantik ist der Teilbereich der (Computer-)Linguistik, der sich mit sprachlicher Bedeutung befasst.
- Wir unterscheiden nach den Ebenen der sprachlichen Gliederung zwischen Wortsemantik, Satzsemantik und Diskurssemantik.
- Semantik hat einen Doppelcharakter: Sie ist Teil der Grammatik, und sie ist Schnittstelle zwischen Sprache und Welt.
- Natürliche Sprache ist fast durchgehend mehrdeutig und ambivalent.

Mehrdeutigkeit und Vagheit

Lexikalische Mehrdeutigkeit:
Homonymie (*Bank, Absatz, ergeben*)
Polysemie (*Baum*, *beginnen, schnell*)
Vagheit

Referenzielle Mehrdeutigkeit:
er, sie, es, dort, damals, der Präsident, die Vorlesung

Mehrdeutigkeit auf Satzebene:
Peter sieht den Mann mit dem Teleskop
Zwei Fremdsprachen spricht jeder Linguist
Zwei Teilnehmer halten ein Referat
1,2 Millionen Besucher tranken 800000 Tassen Kaffee

Semantik

- Semantik ist der Teilbereich der (Computer-)Linguistik, der sich mit sprachlicher Bedeutung befasst.
- Wir unterscheiden nach den Ebenen der sprachlichen Gliederung zwischen Wortsemantik, Satzsemantik und Diskurssemantik.
- Semantik hat einen Doppelcharakter: Sie ist Teil der Grammatik, und sie ist Schnittstelle zwischen Sprache und Welt.
- Natürliche Sprache ist extrem mehrdeutig und ambivalent.
- Die Bedeutung eines natürlich-sprachlichen Ausdrucks hängt massiv vom Kontext seiner Äußerung ab.

Kontextabhängigkeit

Jeder Student kennt die Prädikatenlogik.
Überall grünt und blüht es.
Peter kommt immer zu spät.

Peter hat sich einen teuren Wagen gekauft.

Bitte noch eins!

Hans hat Peter nicht gegrüßt. Er ist beleidigt.
Hans mag seinen Hund, obwohl er ihn manchmal beißt.

Semantik

- Semantik ist der Teilbereich der (Computer-)Linguistik, der sich mit sprachlicher Bedeutung befasst.
- Wir unterscheiden nach den Ebenen der sprachlichen Gliederung zwischen Wortsemantik, Satzsemantik und Diskurssemantik.
- Semantik hat einen Doppelcharakter: Sie ist Teil der Grammatik, und sie ist Schnittstelle zwischen Sprache und Welt.
- Natürliche Sprache ist extrem mehrdeutig und ambivalent.
- Die Bedeutung eines natürlich-sprachlichen Ausdrucks hängt massiv vom Kontext seiner Äußerung ab.
- Semantik hört nicht auf der Satzebene auf: Wir unterscheiden Wortsemantik, Satzsemantik und Diskurssemantik
- Semantik hört nicht mit der Ermittlung der kontextspezifischen Äußerungsinformation auf: Inferenz

Inferenz

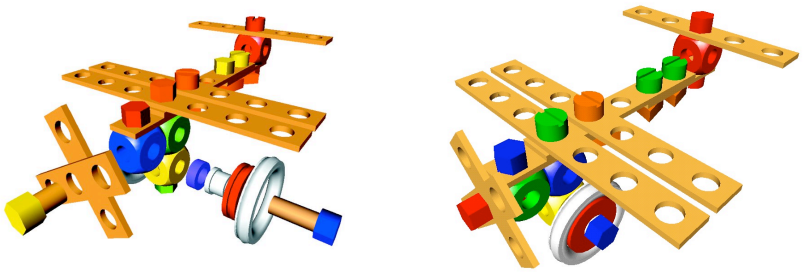
- Semantisches Potenzial oder linguistischer Bedeutungsgehalt durch semantische Komposition oder Semantikkonstruktion.
- Intendierte Äußerungsbedeutung durch Disambiguierung oder Kontexualisierung.
- Relevante Äußerungsinformation durch Inferenz.

Semantik

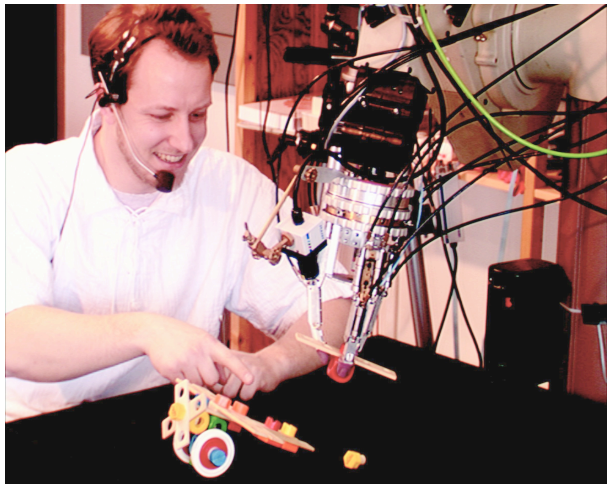
- Semantik ist der Teilbereich der (Computer-)Linguistik, der sich mit sprachlicher Bedeutung befasst.
- Semantik hat einen Doppelcharakter: Sie ist Teil der Grammatik, und sie ist Schnittstelle zwischen Sprache und Welt.
- Natürliche Sprache ist extrem mehrdeutig und ambivalent.
- Die Bedeutung eines natürlich-sprachlichen Ausdrucks hängt massiv vom Kontext seiner Äußerung ab.
- Semantik hört nicht auf der Satzebene auf: Wir unterscheiden Wortsemantik, Satzsemantik und Diskurssemantik
- Semantik hört nicht mit der Ermittlung der kontextspezifischen Äußerungsinformation auf: Inferenz
- Der Gegenstand der Semantik ist abstrakt: Auf Wörter und Sätze können wir zeigen, aber was ist Bedeutung?

Was ist Bedeutung eigentlich?

Eine Robotik-Anwendung



Sonderforschungsbereich
„Situierte Künstliche Kommunikatoren“
Bielefeld



Semantik

- Semantik ist der Teilbereich der (Computer-)Linguistik, der sich mit sprachlicher Bedeutung befasst.
- Semantik hat einen Doppelcharakter: Sie ist Teil der Grammatik, und sie ist Schnittstelle zwischen Sprache und Welt.
- Natürliche Sprache ist extrem mehrdeutig und ambivalent.
- Die Bedeutung eines natürlich-sprachlichen Ausdrucks hängt massiv vom Kontext seiner Äußerung ab.
- Semantik hört nicht auf der Satzebene auf: Wir unterscheiden Wortsemantik, Satzsemantik und Diskurssemantik
- Semantik hört nicht mit der Ermittlung der kontextspezifischen Äußerungsinformation auf: Inferenz
- Der Gegenstand der Semantik ist abstrakt: Auf Wörter und Sätze können wir zeigen, aber was ist Bedeutung?
- Sprachliche Bedeutung ist vielschichtig und heterogen.

Sprachliche Bedeutung ist vielschichtig und heterogen

- Zur Bedeutung können gehören:
  - Propositionale/ konzeptuelle Information – begrifflich erfassbar, in einem logischen Framework darstellbar
  - Visuelle (und andere perzeptuelle) prototypische Information
  - Handlungsbezogene Information
  - Sterotypische Information – nur im Regelfall gültig (Default-Information)

- Es gibt keine scharfe Grenze zwischen sprachlicher Bedeutung und nicht-sprachlichem Wissen

Semantik in der Computerlinguistik

- Vollständige Beschreibung der sprachlichen Bedeutung ist nicht möglich.
- Bedeutungsbeschreibung ist notwendigerweise partiell - auf eine Schicht/ bestimmte Schichten semantischer Information eingeschränkt.
- Die Computerlinguistik hat sich bisher hauptsächlich mit propositionaler Bedeutung befasst, und hat sie mit Logik repräsentiert.
- Multimodale semantische Verarbeitung ist ein relativ neues und spannendes Thema.

Logik-basierter Ansatz

Logische Folgerung: ✓

$\forall d\ (dolphin(d) \rightarrow mammal(d) \wedge \neg fish(d)) \models \forall d\ (dolphin(d) \rightarrow mammal(d))$

Semantische Interpretation

Dolphins are mammals, not fish. They are warm blooded like man, and give birth to one baby called a calf at a time. At birth a bottlenose dolphin calf ...

Dolphins are mammals.

Ist Flipper ein Säugetier?

$\forall d\ (dolphin(d) \rightarrow mammal(d) \wedge \neg fish(d))$
$dolphin(flipper) \rightarrow mammal(flipper) \wedge \neg fish(flipper)$ $dolphin(flipper)$
$mammal(flipper) \wedge \neg fish(flipper)$
$mammal(flipper)$

Logical Entailment: ✓

$\forall d\ (dolphin(d) \rightarrow mammal(d) \wedge \neg fish(d)) \models mammal(flipper)$

Semantic Interpretation

Dolphins are mammals, not fish. They are warm blooded like man, and give birth to one baby called a calf at a time. At birth a bottlenose dolphin calf ...

Flipper is a mammal.

Lexikalische Semantik und Inferenz

Einige Automobilproduzenten haben sich zur wirtschaftlichen Entwicklung 2010 positiv geäußert.

Haben sich Industriefirmen zur wirtschaftlichen Entwicklung 2010 positiv geäußert?

Lexikalische Semantik und Inferenz

Einige Automobilproduzenten haben sich zur wirtschaftlichen Entwicklung 2010 positiv geäußert.

Haben sich Industriefirmen zur wirtschaftlichen Entwicklung 2010 positiv geäußert?

-

Lexikalische Semantik und Inferenz

Der US-Flugzeughersteller Lockhead hat von Grossbritannien den Auftrag fuer 25 Flugzeuge des Typs Hercules C130J erhalten. Vertreter des Verteidigungsministeriums bezifferten den Wert des Auftrags mit umgerechnet 2,5 Mrd. DM.

Wieviel kosten die Maschinen, die Lockhead an Großbritannien verkauft hat?

Lexikalische Semantik und Inferenz

Der US-Flugzeughersteller Lockhead hat von Grossbritannien den Auftrag fuer 25 Flugzeuge des Typs Hercules C130J erhalten. Vertreter des Verteidigungsministeriums bezifferten den Wert des Auftrags mit umgerechnet 2,5 Mrd. DM.

- *Wieviel kosten die Maschinen, die Lockhead an Großbritannien verkauft hat?*

Elementare semantische Relationen

- Synonymie: Flugzeug - Maschine - Flieger
- Hypernymie / Hyponymie, die Ober-/Unterbegriffsrelation:
  - *Flugzeug - Propellermaschine*
  - *Tier - Hund*
  - *töten - umbringen*
- Meronymie/ Holonymie die Teil-von-Relation:
  - Ast - Baum
  - Holz - Baum
  - Baum - Wald
- Antonymie: Kontrastrelation
  - gut - schlecht
  - teuer - billig
  - Freund - Feind

Die Wort-Bedeutungs-Relation

- Genau genommen bestehen semantische Relationen nicht zwischen Wörtern, sondern zwischen Konzepten bzw. Wortbedeutungen (“word senses”): Die Abbildung zwischen phonologischen/ orthographischen Wörtern und Konzepten ist in beiden Richtungen nicht eindeutig.
- Ein Konzept kann in verschiedenen Wörtern kodiert sein: Synonymie
- Ein Wort ist in der Regel mit verschiedenen Wörtern assoziiert : Lexikalische Ambiguität
- Ambiguität zwischen nicht-verwandten Konzepten heißt Homonymie
  - *Bank: Geldinstitut /Bank: Sitzmöbel*
- Ambiguität zwischen semantisch verwandten Konzepten heißt Polysemie
  - *Maschine: Flugzeug / Maschine: Motorrad*
  - *Sitz im Auto / im Bundestag*
  - *einen Plan / einen Koffer aufgeben*

WordNet

- WordNet ist eine große lexikalisch-semantische Ressource: Ein Netzwerk aus semantischen Relationen zwischen Konzepten, mit der Hyponymie-Relation als Kern.
- Konzepte werden als „synsets“ repräsentiert: Mengen von synonymen Wörtern, die sich gegenseitig disambiguieren.
- Synsets liefern direkte Information zur Synonymierelation und zur Wort-Konzept-Abbildung: Ein orthographisches Wort hat genau die Synsets als Lesarten, in denen es als Element vorkommt.
- Dies gilt im Prinzip. praktisch sind meist keine oder zu wenige Synonyme vorhanden, um Konzepte eindeutig zu charakterisieren. Deshalb enthält WordNet für alle Synsets „Glossen“ (Umschreibungen) und Anwendungsbeispiele.

Beispiel: Lesarten von *car*

- S: (n) **car**, auto, automobile, machine, motorcar
- S: (n) **car**, railcar, railway car, railroad car
- S: (n) **car**, gondola
- S: (n) **car**, elevator car
- S: (n) cable car, **car**

Beispiel: Synset, Glosse, Sprachbeispiel

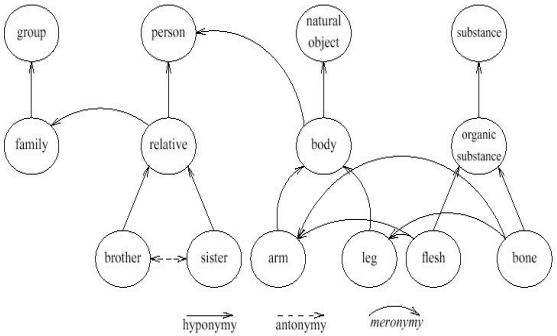
- **car**
  - { car, auto, automobile, machine, motorcar }
  - a motor vehicle with four wheels; usually propelled by an internal combustion engine
  - *"he needs a car to get to work"*

Hyponyme von *motor vehicle*

- S: (n) **motor vehicle**, automotive vehicle (a self-propelled wheeled vehicle that does not run on rails)
- ***direct hyponym*** / *full hyponym*
  - S: (n) amphibian, amphibious vehicle (a flat-bottomed motor vehicle that can travel on land or water)
  - S: (n) bloodmobile (a motor vehicle equipped to collect blood donations)
  - S: (n) car, auto, automobile, machine, motorcar (a motor vehicle with four wheels; usually propelled by an internal combustion engine) *"he needs a car to get to work"*
  - S: (n) doodlebug (a small motor vehicle)
  - S: (n) four-wheel drive, 4WD (a motor vehicle with a four-wheel drive transmission system)
  - S: (n) go-kart (a small low motor vehicle with four wheels and an open framework; used for racing)
  - S: (n) golfcart, golf cart (a small motor vehicle in which golfers can ride between shots)
  - S: (n) hearse (a vehicle for carrying a coffin to a church or a cemetery; formerly drawn by horses but now usually a motor vehicle)
  - S: (n) motorcycle, bike (a motor vehicle with two wheels and a strong frame)
  - S: (n) snowplow, snowplough (a vehicle used to push snow from roads)
  - S: (n) truck, motortruck (an automotive vehicle suitable for hauling)

Ein kleines Fragment von WordNet

Figure 2. Network representation of three semantic relations among an illustrative variety of lexical concepts



WordNet-Daten

- Englisches WordNet hat zurzeit einen Umfang von
  - 150.000 lexikalischen Einträgen (Wörtern)
  - 120.000 Synsets
- WordNet-Versionen gibt es für etwa 45 Sprachen
- WordNet wird in vielen sprach- und informationstechnologischen Anwendungen erfolgreich genutzt.
- "GermaNet": eine deutsche WordNet-Version mit etwa 100.000 lexikalischen Einträgen
- Englisches WordNet
  - Web Interface: http://wordnet.princeton.edu/perl/webwn
  - Allgemeine Information: http://wordnet.princeton.edu

WordNet und Inferenz

- WordNet-Relationen können als logische Formeln reformuliert werde:
  $\forall x(family(x) \rightarrow group(x))$
  $\forall x(person(x) \rightarrow \exists y(substance\_m(y,x) \wedge body(y))$
  $\forall x(body(x) \rightarrow \exists y(part\_m(y,x) \wedge leg(y))$
  $\forall x(body(x) \rightarrow \exists y(part\_m(y,x) \wedge arm(y))$

- Für die semantische Verarbeitung in einem logischen Framework kann WordNet als große Datenbasis verwendet werden, die zusätzliche Axiome für die Inferenz bereitstellt.