Einführung in die Computerlinguistik

Sprachtechnologie für das Informationsmanagement

WS 2009/2010

Manfred Pinkal

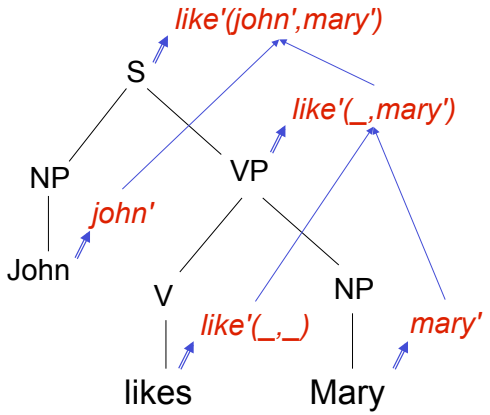
Computerlinguistik und Textdokumente

- Die Informationstechnologie, insbesondere as Internet haben riesige Mengen von Wissen digital verfügbar gemacht.
- Der mit Abstand größte Anteil dieses Wissens ist in „semi-strukturierter" Form in Textdokumenten verfügbar.
- Wie kann die Computerlinguistik dazu beitragen, dass dies Wissen erschlossen und nutzbar gemacht wird?

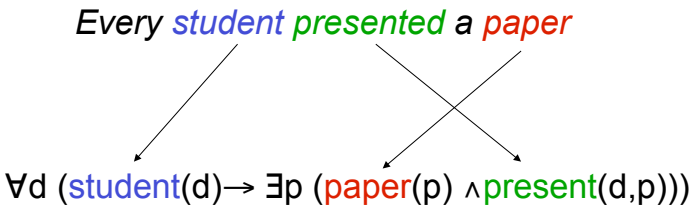
Semantische Verarbeitung von Texten

- Die Kodierung von Information in Sprache ist Gegenstand der Semantik.
- Zugang zur Information erfolgt natürlicherweise durch semantische Analyse:
  - Wortsemantik (vgl. WordNet)
  - Satzsemantik fügt Wortbedeutungen zu satzsemantischen Repräsentationen zusammen
  - Diskurssemantik disambiguiert Sätze im Kontext und fügt sie zur Diskursrepräsentation zusammen
- Logisches Framework für die Repräsentationsformat für Satz- und Textbedeutungen und für Verfahren zur Bedeutungskomposition

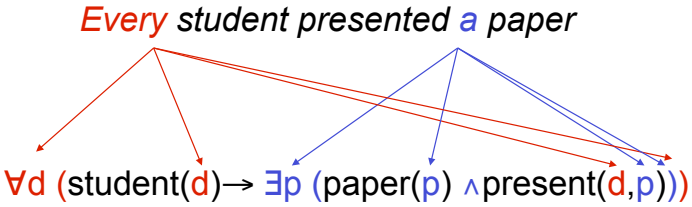
Bedeutungskomposition: Ein Beispiel



Bedeutungskomposition: Eine Herausforderung

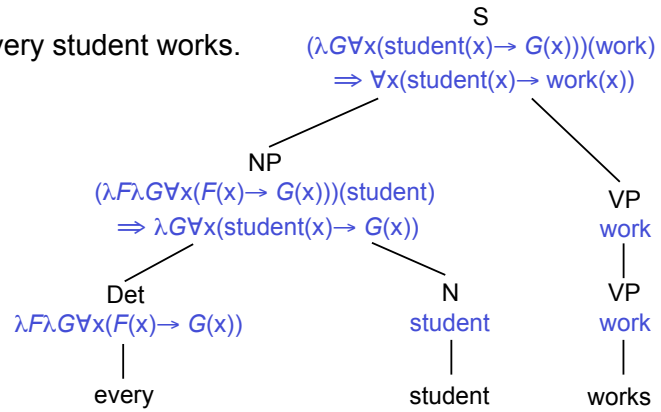


Eine Herausforderung für die semantische Komposition



Semantik-Konstruktion

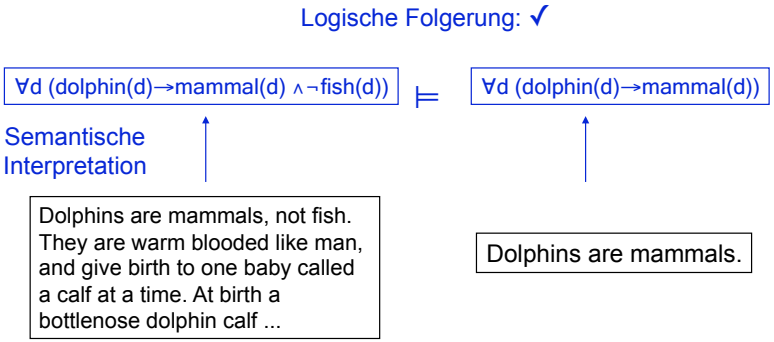
- Every student works.



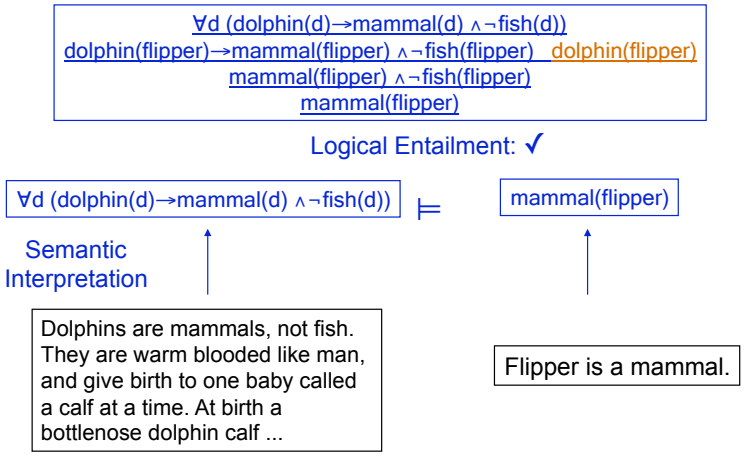
Textverstehen

- Klassische Anwendung der Computerlinguistik:
  Textverstehen:
  - Eingabetext wird in logische Repräsentation überführt.
  - Anfrage wird in logische Repräsentation überführt.
  - „Inferenzmaschine"/ Theorembeweiser stellt fest, ob sich eine sinnvolle Antwort auf die Anfrage aus der Textrepräsentation ableiten lässt.

Logik-basierter Ansatz



Ist Flipper ein Säugetier?



Textverstehen

- Klassische Anwendung der Computerlinguistik:
  Textverstehen:
  - Eingabetext wird in logische Repräsentation überführt.
  - Anfrage wird in logische Repräsentation überführt.
  - „Inferenzmaschine"/ Theorembeweiser stellt fest, ob sich eine sinnvolle Antwort auf die Anfrage aus der Textrepräsentation ableiten lässt.
- Problem: Die „Vollübersetzung" von Texten in Logikrepräsentationen ist extrem schwierig
- ... und für viele Anwendungen auch nicht nötig.

Informationsmanagement

Konkrete Aufgaben für die Computerlinguistik:

- Textdokumente klassifizieren: Document Classification
- Textdokumente zusammenfassen: Summarisation
- Relevante Information in Textdokumenten/ in Textdatenbanken/ im Web auffinden:
  - Information Retrieval
  - Question Answering
  - Informationsextraktion

Information Retrieval

- Gegeben: Suchanfrage (Query)
  - Im einfachsten Fall eine Menge von Suchbegriffen (Termen)
- Gesucht: Relevante Dokumente
  - Liste von Dokumenten, die relevante Information zu den Termen der Suchanfrage enthalten.
  - Beispiel: Web-Suchmaschine (Google), aber auch: Anfragen in Fachdatenbanken, Firmen-IntraNets etc.

Question Answering

- Gegeben: Umgangssprachliche Frage
- Gesucht: Dokument mit dem relevantem Satz, der eine plausible Antwort darstellt
- Beispiele:
  - Wer war im Jahr 2002 deutscher Fußball-Meister?
  - Wann wurde Barack Obama geboren?
  - Wer war amerikanischer Präsident, als Barack Obama geboren wurde?

Informations-Extraktion

- Suche nach Instanzen bestimmter Ereignisse/ Sachverhalte in Textdatenbanken
- Ausgabe geht als Information in eine relationale Datenbank
- Beispiel:
  - Wechsel im Vorstandvorsitz von Industriefirmen
  - Wer ist wann bei welcher Firma Chef geworden?

Name	Jahr	Firma
Josef Ackermann	2002	Deutsche Bank
Rüdiger Grube	2009	Deutsche Bahn
Norbert Reithofer	2006	BMW
...	...	...

Information Retrieval

- Informationswunsch des Nutzers:
  - *Welche Lehrveranstaltung behandelt Syntax?*
- Kodiert in eine Suchanfrage/ Query:
  - {*Veranstaltung, Syntax*}
- Relevant Dokumente: Signifikante Wortüberlappung mit der Suchanfrage.

Beispiel

d1: Vorlesung Einführung in die Sprachwissenschaft:

Die Veranstaltung wird als Ringvorlesung durchgeführt. Die jeweilige Lehrkraft für verschiedene Bereiche der Sprachwissenschaft führt in die Ziele und Begriffe des Bereiches ein. Behandelt werden Phonetik und Phonologie, Morphologie und Syntax, Semantik, Pragmatik und Psycholinguistik .

d2: Vorlesung Syntax und Morphologie:

Ziel der Veranstaltung ist es, die Teilnehmer mit Grundbegriffen und Grundproblemen der deskriptiven wie theoretischen Syntax und Morphologie vertraut zu machen. Im Vordergrund steht dabei die Syntax des Deutschen, aber auch Phänomene im Englischen oder anderen Sprachen werden diskutiert.

d3: Regierung befürwortet Ausbildungsabgabe:

Gegen den Widerstand von Arbeitsminister Clement haben sich Bundeskanzler Schröder und die SPD- Spitze bei einer Veranstaltung des DGB für eine Ausbildungsabgabe ausgesprochen. Eine entsprechende Vorlage wird Montag in der Bundestagsfraktion behandelt.

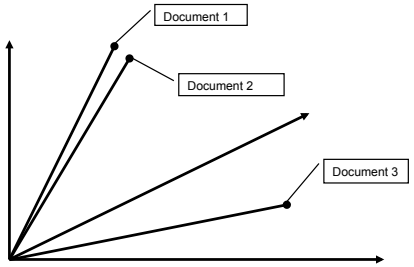
Termfrequenz

- Die Frequenz der Wörter in einem Dokument (Termfrequenz) ist ein Indikator für die inhaltliche Ausrichtung des Dokuments.
- Dokumentinformation wird als Muster von Termfrequenzen dargestellt: als Vektor, dessen Dimensionen Wörter sind, mit den jeweiligen Worthäufigkeiten als Werten.
- Ein Dokument wird repräsentiert als Vektor im vieldimensionalen semantischen Raum, dessen Dimensionen Wörtern entsprechen ("Wortraum" / "word space")
- Informationelle/ semantische Ähnlichkeit von Dokumenten untereinander wird durch den Vergleich ihrer Vektoren modelliert.
- Die Suchanfrage wird ebenfalls als Vektor bestimmt.
- Die Relevanz eines Dokuments für die Suchanfrage wird durch den Vergleich der jeweiligen Vektoren bestimmt.

Term-Dokument-Matrix

	d1	d2	d3	…
Veranstaltung	1	1	1	…
Teilnehmer	0	1	0	…
behandelt	1	0	1	…
Gesetz	0	0	1	…
Arbeitsminister	0	0	1	
Clement	0	0	1	…
Syntax	1	3	0	…
Morphologie	1	1	0	…
…	…	…	…	…

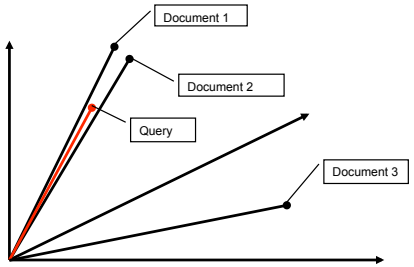
Semantischer Raum



Term-Dokument-Matrix

	d1	d2	d3	q
Veranstaltung	1	1	1	1
Teilnehmer	0	1	0	0
behandelt	1	0	1	0
Widerstand	0	0	1	0
Arbeitsminister	0	0	1	0
Clement	0	0	1	0
Syntax	1	3	0	1
Morphologie	1	1	0	0

Semantischer Raum



Distanz zwischen Vektoren als (inverses) Ähnlichkeitsmaß

- Euklidische Distanz:

$$dist(\vec{x},\vec{y}) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

- Unser Beispiel:

$$dist(\vec{d}_1,\vec{d}_2) = 1{,}73$$
$$dist(\vec{d}_1,\vec{d}_3) = 2{,}45$$
$$dist(\vec{d}_2,\vec{d}_3) = 3{,}00$$

- Problem: abhängig von der absoluten Häufigkeit der Terme, und damit von der Größe der Dokumente.

Cosinus als Ähnlichkeitsmaß

- Standardmaß für die Ähnlichkeit ist der Cosinus

$$sim_{\cos}(\vec{x},\vec{y}) = \frac{\vec{x}\cdot\vec{y}}{|\vec{x}||\vec{y}|} = \frac{\sum_{i=1}^{n} x_i \times y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}}$$

- Wenn Vektoren identische Richtung haben, ist Cosinus 1($cos(0^0)$=1); wenn Vektoren rechtwinklig aufeinander stehen, ist der Cosinus 0 ($cos(90^0)$=0).
- Unser Beispiel:

$$\cos(\vec{q},\vec{d}_1) = 0.65$$
$$\cos(\vec{q},\vec{d}_2) = 0.77$$
$$\cos(\vec{q},\vec{d}_3) = 0.29$$

Vorteile von distributionellen Maßen

- Nutzung von Frequenzinformation
  - Konzeptuell einfach, effizient
  - Dokumente sind ähnlich, wenn Begriffe gleich häufig vorkommen
- Formalisierung
  - Mathematische Standardverfahren zur Berechnung von Ähnlichkeit / Relevanz (z.B.: euklidische Distanz, Cosinus)
- Erweiterungsmöglichkeiten
  - Nicht-sprachliche Information, z.B. Verlinkungsstruktur
  - Linguistische Verfahren: genauere Modellierung der Terme

Beispiel

d1: Vorlesung Einführung in die Sprachwissenschaft: Die Veranstaltung wird als Ringvorlesung durchgeführt. Die jeweilige Lehrkraft für verschiedene Bereiche der Sprachwissenschaft führt in die Ziele und Begriffe des Bereiches ein. Behandelt werden Phonetik und Phonologie, Morphologie und Syntax, Semantik, Pragmatik und Psycholinguistik .
d2: Vorlesung Syntax und Morphologie: Ziel der Veranstaltung ist es, die Teilnehmer mit Grundbegriffen und Grundproblemen der deskriptiven wie theoretischen Syntax und Morphologie vertraut zu machen. Im Vordergrund steht dabei die Syntax des Deutschen, aber auch Phänomene im Englischen oder anderen Sprachen werden diskutiert.
d3: Regierung befürwortet Ausbildungsabgabe: Gegen den Widerstand von Arbeitsminister Clement haben sich Bundeskanzler Schröder und die SPD- Spitze bei einer Veranstaltung des DGB für eine Ausbildungsabgabe ausgesprochen. Eine entsprechende Vorlage wird Montag in der Bundestagsfraktion behandelt.

Genauere Modellierung von Termen (1)

Nicht alle Worte sind gleich

- „Stoppworte" komplett entfernen
  - Artikel, Hilfsverben, Präpositionen:
  - Sind semantisch nicht ergiebig, kommen in ähnlicher Verteilung überall vor

Vorlesung Einführung in die Sprachwissenschaft

Die Veranstaltung wird als Ringvorlesung durchgeführt. Die jeweilige Lehrkraft für verschiedene Bereiche der Sprachwissenschaft führt in die Ziele und Begriffe des Bereiches ein. Behandelt werden Phonetik und Phonologie, Morphologie und Syntax, Semantik, Pragmatik und Psycholinguistik .

Genauere Modellierung von Termen (2)

- Informative Worte stärker gewichten:
  - Dokumentfrequenz n_i von Wort i: Anzahl der Dokumente im Korpus, die das Wort i (mindestens einmal) enthalten
  - Gesamtzahl der Dokumente: *N*
  - idf (inverse Dokumentfrequenz) $idf_i = \log\left(\frac{N}{n_i}\right)$
  - idf ist ein Maß für die Informativität von Suchtermen
  - **tf * idf**: Termfrequenz * Inverse Dokumentfrequenz

Die Wort-Dokument-Matrix

	idf	d1	d2	d3	q
Veranstaltung	5	1	1	1	1
Teilnehmer	5	0	1	0	0
behandelt	3	1	0	1	0
Widerstand	6	0	0	1	0
Arbeitsminister	10	0	0	1	0
Clement	14	0	0	1	0
Syntax	10	1	3	0	1
Morphologie	11	1	1	0	0

Die Wort-Dokument-Matrix

	d1	d2	d3	q
Veranstaltung	5	5	5	5
Teilnehmer	0	5	0	0
behandelt	3	0	3	0
Widerstand	0	0	6	0
Arbeitsminister	0	0	10	0
Clement	0	0	14	0
Syntax	10	30	0	10
Morphologie	11	11	0	0

$\cos(\vec{q},\vec{d}_1) = 0.65$ $\cos(\vec{q},\vec{d}_1) = 0.73$

$\cos(\vec{q},\vec{d}_2) = 0.77$ $\cos(\vec{q},\vec{d}_2) = 0.86$

$\cos(\vec{q},\vec{d}_3) = 0.29$ $\cos(\vec{q},\vec{d}_3) = 0.04$

Termüberschneidung vs. Semantische Ähnlichkeit

- Eigentlich brauchen wir nicht gleiche Wörter in Suchanfrage und Dokument, sondern gleiche Konzepte (Wortbedeutungen).
- Problem 1: {Absatz, Pkw} als Query und *„Absatz 2 der STVO"* sollten nicht matchen
- Problem 2: *Pkw* in der Query, *Automobil, Wagen, Auto* im Dokument sollten matchen.
- Lösung für Problem1: „WSD" – Word-Sense Disambiguation (generelle Verfahren funktionieren nur beschränkt, aber Suchterme in einer Query schränken disambiguieren sich in gewissem Umfang gegenseitig)
- Lösung für Problem2: Query Expansion

Query Expansion

- Erweiterung der Suchanfrage mit WordNet-Synonymen (Hyponymen etc.)
- {Absatz, Pkw}
  - → {Absatz, Pkw, Auto, Wagen Automobil}
  - → {..., Cabrio, Kraftfahrzeug, ...}
- Effekt: Erhöhung des Recall, Abnahme der Präzision
- Ausgleich durch Heruntergewichten von semantisch ähnlichen Termen gegenüber voller Identität

Question Answering

- Gegeben: Query (als umgangssprachliche Frage)
- Gesucht: Relevanter Satz (aus Dokument)
- Vorgehen:
  - Schritt 1: IR → Liste von Dokumenten
  - Schritt 2: Identifikation und Extraktion der relevanten Passage: Suche den Satz des Dokumentes, der semantisch am besten zur Anfrage passt.

Question Answering

- Josef Ackermann wurde 2002 zum Vorstandsvorsitzenden der deutschen Bank gewählt.
- Wann wurde Ackermann Chef der deutschen Bank?

WordNet-Ähnlichkeit

- Quantitatives Maß für die Nähe/ Distanz zwischen zwei Termen auf der basis der WordNet-Hierarchie
- Einfaches Distanzmaß: Die Pfadlänge $dist_{WN} = pathlength(s_1, s_2)$
- Einfaches Ähnlichkeitsmaß: Inverse Pfadlänge $sim_{WN} = \frac{1}{pathlength(s_1, s_2) + 1}$

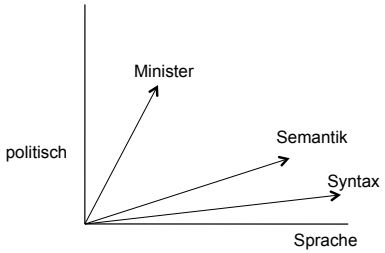
Distributionelle Maße für semantische Ähnlichkeit

- Distributionelle Hypothese:
  Die semantische Ähnlichkeit von Wörtern kann an der Ähnlichkeit ihrer Kontext gemessen werden.
- Was ist der Kontext eines Wortes w?
  - das Dokument, der Absatz, der Satz, in dem das Wort w vorkommt
  - Ein Fenster mit n (5, 10, 30, ...) Wörtern vor und nach w.
- Distributionelle Bedeutungsrepräsentation von w:
  - Wir zählen die Vorkommen der Inhaltsworte in allen Kontexten von w (in einem Korpus).
  - Die Bedeutung von w ist der Frequenzvektor von w im Wortraum.

Einfaches Beispiel

- Häufigkeiten von ‘politisch’ und ‘Sprache’ im Kontext von ‘Syntax’, ‘Semantik’ und ‘Minister’

	Syntax	Semantik	Minister
politisch	2	5	70
Sprache	45	40	15



- Tabelle und graphische Repräsentation zeigen, dass *Semantik* und *Syntax* zur Domäne der Sprache, *Minister* zur Domäne der Politik tendiert.
- Sie illustrieren auch, dass die Ähnlichkeit zwischen *Syntax* und *Semantik* größer ist als die Ähnlichkeit dieser beiden Begriffe zu *Minister.*

Generelle Probleme mit distributionellen Verfahren

- Semantische Ähnlichkeit ist symmetrisch, nicht sensitiv für die Richtung der Implikation:
  - *Auto* im Dokument – *Cabrio* in der Anfrage
  - *Cabrio* im Dokument – *Auto* in der Anfrage
- Wortbasierte Ähnlichkeitsmaße für Sätze und Texte sind nicht sensibel für Wortstellung:
  - *Mann beißt Hund*
  - *Hund beißt Mann*
- Suche mit Ähnlichkeitsmaßen berücksichtigt nicht die satzsemantische Struktur, Negation, Modalitäten etc.
- Lösung: Geeignete Kombination von statistisch-distributioneller Information (für die Abdeckung) mit syntaktischer und logisch-semantischer Information (für die Präzision).