Einführung in die Computerlinguistik

Informationszugriff, Lexikalische Semantik, Statistische Modelle

WS 2010/2011

Manfred Pinkal

Computerlinguistik und Textdokumente

- Das 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?

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
...	...	...

Mehrdeutigkeit

- Ein Wort ist in der Regel mit verschiedenen Wörtern assoziiert : Lexikalische Ambiguität
  - *Bank: Geldinstitut /Bank: Sitzmöbel*
  - *Maschine: Flugzeug / Maschine: Motorrad*
  - *Sitz im Auto / im Bundestag*
  - *einen Plan / einen Koffer aufgeben*
- Wenn wir Information in Textdokumenten suchen, sind wir an Konzepten interessiert, haben aber unmittelbar nur die Wörter zur Verfügung, die die Konzepte (und zwar möglicherweise viele alternative Konzepte) ausdrücken.
- Zentrale Aufgabe: Disambiguierung

WSD

- Disambiguierung der Wortbedeutung/ Word-Sense-Disambiguation/ WSD
- Wichtige und schwierige Aufgabe
- Wissensbasierte Verfahren sind praktisch nicht anwendbar: Zehntausende von Mehrdeutigkeiten, die keinem gemeinsamen Muster folgen.
- Statistische WSD

WSD

- Sammle Instanzen/ bilde ein Trainingskorpus für ein Zielwort (z.B. Bank").
- Verwende die in Wörterbüchern oder in WordNet gelisteten Lesarten als Zielklassen.
- Annotiere das Trainingskorpus.
- Extrahiere geeignete Merkmale.
- Trainiere mit einem maschinellen Lernsystem/ Klassifikator.

Trainings-Korpus

Für diejenigen, denen Komfort wichtig ist, haben wir eine Bank mit leicht schwingender Rückenlehne entwickelt.

Ich suche noch eine Bank für meinen Garten und sondiere deshalb gerade Angebote.

Habe im März 2000 einen höheren Betrag bei einer Bank angelegt. Es handelte sich dabei um eine aktiv gemanagte Anlage ...

Beim Test Anlageberatung der Banken löste kein Institut die einfache Frage nach einer sicheren Anlage wirklich gut.

Trainings-Korpus - Annotation

Für diejenigen, denen Komfort wichtig ist, haben wir eine $Bank_1$ mit leicht schwingender Rückenlehne entwickelt.

Ich suche noch eine $Bank_1$ für meinen Garten und sondiere deshalb gerade Angebote.

Habe im März 2000 einen höheren Betrag bei einer $Bank_2$ angelegt. Es handelte sich dabei um eine aktiv gemanagte Anlage ...

Beim Test Anlageberatung der $Banken_2$ löste kein Institut die einfache Frage nach einer sicheren Anlage wirklich gut.

Trainingskorpus - Merkmalsextraktion

Für diejenigen, denen Komfort wichtig ist, haben wir eine Bank$_1$ mit leicht schwingender Rückenlehne entwickelt.

Ich suche noch eine Bank$_1$ für meinen Garten und sondiere deshalb gerade Angebote.

Habe im März 2000 einen höheren Betrag bei einer Bank$_2$ angelegt. Es handelte sich dabei um eine aktiv gemanagte Anlage ...

Beim Test Anlageberatung der Banken$_2$ löste kein Institut die einfache Frage nach einer sicheren Anlage wirklich gut.

Merkmalsextraktion

	Anlage	Angebot	Betrag	Frage	Garten	Institut	Komfort	Rückenlehne	Test	⋮
Bank1	0	0	0	0	0	0	1	1	0	...
Bank1	0	1	0	0	1	0	0	0	0	...
Bank2	1	0	1	0	0	0	0	0	0	...
Bank2	1	0	0	1	0	1	0	0	1	...
...	...									

Kontextmodellierung mit „Bag of Words“

„Stoppwörter“ werden nicht berücksichtigt.

Klassifikator-Training und Anwendung

	Anlage	Angebot	Betrag	Frage	Garten	Institut	Komfort	Rückenlehne	Test	⋮
Bank1	0	0	0	0	0	0	1	1	0	...
Bank1	0	1	0	0	1	0	0	0	0	...
Bank2	1	0	1	0	0	0	0	0	0	...
Bank2	1	0	0	1	0	1	0	0	1	...
...	...									

Keine Frage: In einen ordentlichen Garten gehören neben einer Bank auch die richtigen Möbel.

?	0	0	0	1	1	0	0	0	0	...

Das Bayessche Theorem

- Wie schließen wir vom Merkmalsmuster/ in diesem Fall von der der Konfiguration der Kontextwörter v auf die Wahrscheinlichkeit einer bestimmten Lesart/ eines Wortsinns s?
- Wie bestimmen wir *P(s|v)*?
- Sparse-Data-Problem!
- Wir sehen uns ein masschinelles Lernsystem genauer an, der (in bestimmten Fällen) eine Lösung für das Sparse-Data-Problem anbietet:
- Den „Naive-Bayes“-Klassifikator.

Das Bayessche Theorem

- Wie bestimmen wir *P(s|v)*?
- Das Bayessche Theorem oder die Bayes-Regel:

$$P(E \mid F) = \frac{P(F \mid E) \cdot P(E)}{P(F)}$$

- Die Bayes-Regel ist ein elementares Gesetz der Wahrscheinlichkeitstheorie. Sie ist überall da nützlich, wo der Schluss von einer Größe F auf eine andere Größe E bestimmt werden soll (typischerweise von einer Wirkung/ einem Symptom auf die mögliche Ursache), die Abhängigkeit in der anderen Richtung (von der Ursache auf die Wirkung) aber besser zugänglich ist.

Das Bayessche Theorem

- Kontextvektor v : Symptom
- Wortsinn s : Ursache

- Mit Bayes-Regel :

$$P(s \mid v) = \frac{P(v \mid s) \cdot P(s)}{P(v)}$$

- Der wahrscheinlichste Wortsinn:

$$\max_s P(s \mid v) = \max_s \frac{P(v \mid s) \cdot P(s)}{P(v)}$$

$$= \max_s P(v \mid s) \cdot P(s)$$

„Naive Bayes Classifier“

- Sparse-Data-Problem: Es gibt sehr viele unterschiedliche Kontextmuster. Die Auftretenshäufigkeit eines bestimmten Musters mit einer Lesart ist klein (oft 0) und erlaubt deshalb keine verlässliche Abschätzung von *P(v|s)*.
- Unter der („naiven“) Voraussetzung, dass die einzelnen Wortfrequenzen unabhängig voneinander sind, lässt sich approximieren als Produkt der Einzelwahrscheinlichkeiten für die Kontextwörter:

$$P(v \mid s) \approx \prod_{v_i} P(v_i \mid s)$$

Von Merkmalsmustern ...

	Anlage	Angebot	Betrag	Frage	Garten	Institut	Komfort	Rückenlehne	Test	⋮
Bank1	0	0	0	0	0	0	1	1	0	...
Bank1	0	1	0	0	1	0	0	0	0	...
Bank2	1	0	1	0	0	0	0	0	0	...
Bank2	1	0	0	1	0	1	0	0	1	...
...	...									

... zu Kontextwortfrequenzen ...

	Anlage	Angebot	Betrag	Frage	Garten	Institut	Komfort	Rückenlehne	Test
Bank1	3	2	5	11	40	0	7	18	2
Bank2	84	38	41	15	1	19	3	0	3

Absolute Häufigkeit von Kontextwörtern, in „Merkmalsvektoren“ für die verschiedenen Lesarten repräsentiert.

... zur Wahrscheinlichkeitsschätzung

	Anlage	Angebot	Betrag	Frage	Garten	Institut	Komfort	Rückenlehne	Test
Bank1	0,010	0,007	0,017	0,050	0,133	0,000	0,023	0,060	0,007
Bank2	0,280	0,127	0,137	0,037	0,003	0,063	0,010	0,000	0,010

Relative Frequenzen als Wahrscheinlichkeitsschätzung

Anwendung

	Anlage	Angebot	Betrag	Frage	Garten	Institut	Komfort	Rückenlehne	Test
Bank1	0,010	0,007	0,017	0,050	0,133	0,000	0,023	0,060	0,007
Bank2	0,280	0,127	0,137	0,037	0,003	0,063	0,010	0,000	0,010

Keine Frage: In einen ordentlichen Garten gehören neben einer <u>Bank</u> auch die richtigen Möbel.

Maerkmalsmuster für das Target *Bank* in diesem Satz:

?	0	0	0	1	1	0	0	0	0

Beispiel

$$P(v \mid Bank1) \approx \prod_{v_i} P(v_i \mid Bank1) = 0{,}00588$$

$$P(v \mid Bank2) \approx \prod_{v_i} P(v_i \mid Bank2) = 0{,}00006$$

Bei einer Apriori-Wahrscheinlichkeit von
$P(Bank1)=0{,}29$ / $P(Bank2)=0{,}71$

ergibt sich

$$P(v \mid Bank1) \cdot P(Bank1) = 0{,}00170$$

$$P(v \mid Bank2) \cdot P(Bank2) = 0{,}00004$$

Bank1	0	0	0	1	1	0	0	0	0

WSD

- Eine der schwierigsten Aufgaben in der Computerlinguistik:
- Sehr viele Wörter sind auf sehr unterschiedliche Weise mehrdeutig. Man benötigt riesige Mengen von Trainingsmaterial.
- Alle bisher vorgestellten Lernverfahren sind „überwachte“ (supervised) Lernvefahren: Sie erfordern die manuelle Annotation eines Trainingskorpus.
- Attraktiver sind „halbüberwachte“ (semi-supervised) Verfahren, bei denen ein großes Trainingskorpus (teil-)automatisch auf der Grundlage einer kleinen Menge von handannotierten „Seed-Daten“ erzeugt wird.
- Noch attraktiver sind „unüberwachte“ statistische Verfahren, die Resultate ohne jedes Training erzielen. Ein klassisches Beispiel: Information Retrieval.

Information Retrieval

- Informationswunsch des Nutzers:
  - *Welche Lehrveranstaltung behandelt Syntax?*
- Kodiert in eine Suchanfrage/ Query:
  - {*Veranstaltung, Syntax*}
- Relevante 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.

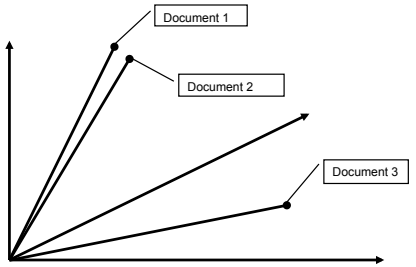
Termfrequenz

- 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 „Bag of Words“, 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	…
…	…	…	…	…

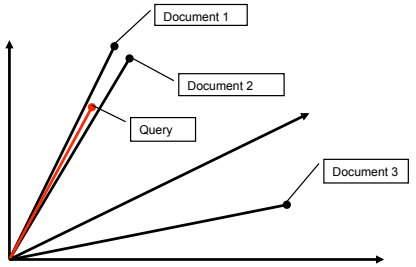
Semantischer Raum



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:

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

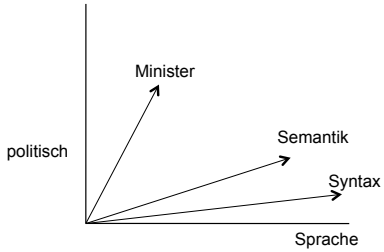
Distributionelle Maße für Ähnlichkeit zwischen Wortbedeutungen

- Mit den Verfahren, mit denen wir Dokumente auf inhaltliche Ähnlichkeit vergleichen, können wir auch Wörter auf semantische Ähnlichkeit vergleichen.
- 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.
- 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 Hypothese:
  Semantisch ähnliche Wörter kommen in ähnlichen Kontexten vor.

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.*

Distributionelle Semantik

- Mit den beschriebenen Verfahren erhalten wir semantische Informationen über Wörter (bzw. im IR über Dokumente) vollständig automatisch („unüberwacht“), mit sehr großer Abdeckung und geringem Aufwand (außer CPU-Zeit).
- Ein Problem für die distributionelle Semantik ist, dass semantische Ähnlichkeit auf der Ebene des Wortes, nicht des Konzepts/ des Wortsinns berechnet wird. Der Bedeutungsvektor von „Bank“ enthält die Kontextwörter zu beiden Lesarten.
- Gebraucht werden unüberwachte Verfahren zur WSD – ein hoch aktuelles Thema in der Computerlinguistik.