

Einführung in die Computerlinguistik

12: Informationsmanagement

WS 2006/2007

Manfred Pinkal



Information Management - Was ist das?

Große Datenmengen zugänglich und **nutzbar machen**

- Konkret:
 - Dokumente klassifizieren (Document Classification)
 - Dokumente zusammenfassen (Summarisation)
 - Dokumente mit relevanter Information finden (Information Retrieval)
 - Relevante Information in Dokumenten finden
 - (Information Extraction, Question Answering)



Naiver Algorithmus zur Beantwortung von Fragen

- Benutzer gibt Schlüsselwörter q („Query“) ein
- Gehe durch alle Dokumente d
 - Wenn q in d vorkommen, ist d für q relevant
- Warum ist dies naiv?



Einige Fragen

- When was Bill Clinton born?

[Google](#)

[Language Computer](#)

- Who was American president at the end of World War II?
- Who was American president when Bill Clinton was born?

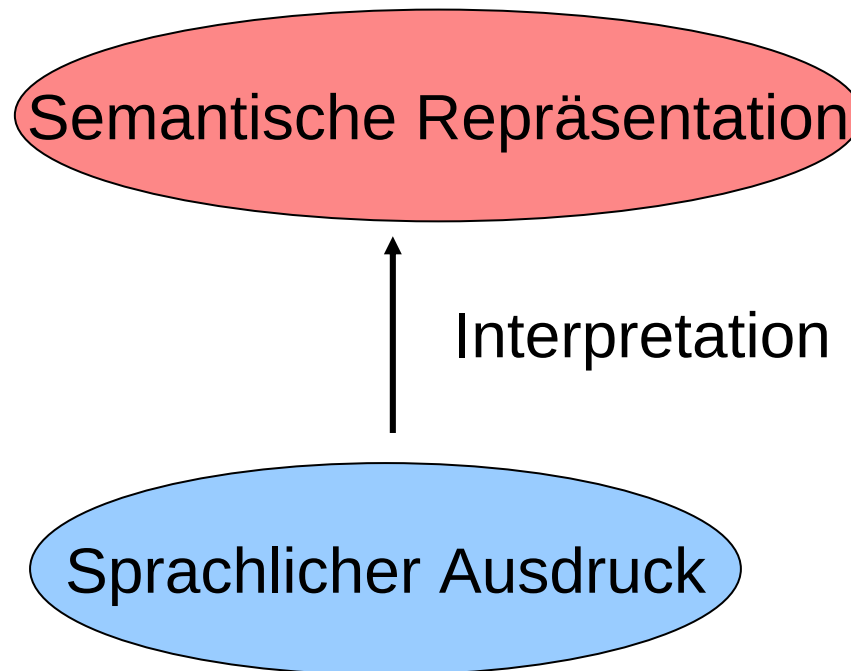


Informationssuche: Probleme

- Information liegt typischerweise in Textform vor: „Semi-strukturierte Daten“
- Typische Anwendungen müssen in riesigen Datenbeständen suchen (Google!)
- Die relevante Ebene ist nicht der Text, sondern die Bedeutungsinformation

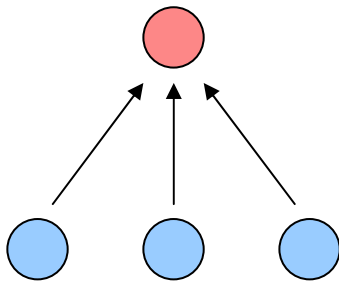


Interpretation

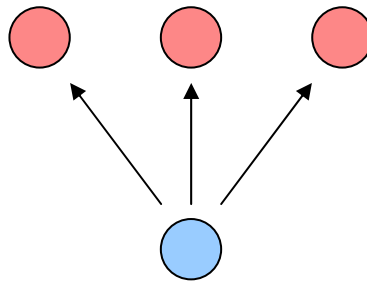


Interpretation: Drei grundlegende Aspekte

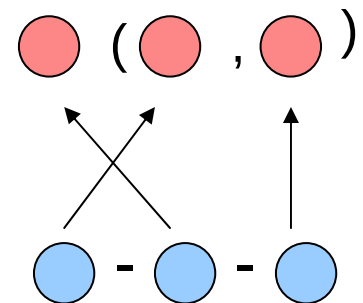
Abstraktion



Disambiguierung

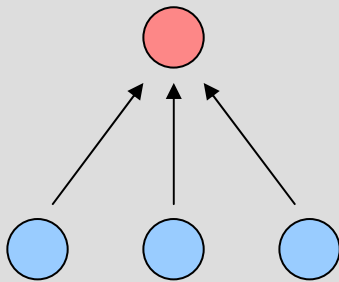


Komposition

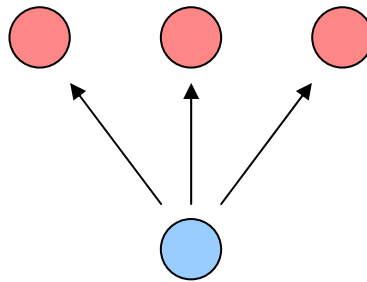


Interpretation

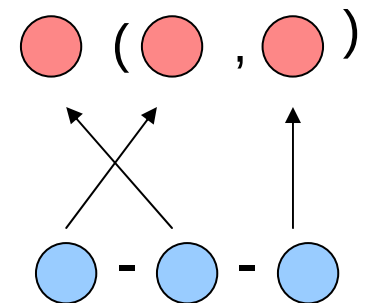
Abstraktion



Disambiguierung



Komposition



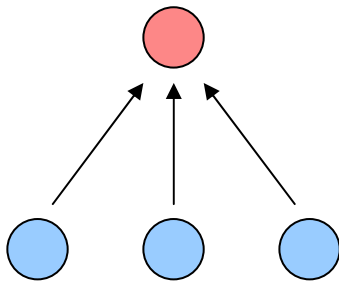
Abstraktion von der sprachlichen Oberfläche

- Wortformen → Stämme/Lemmata
 - *end, ends, ended*
 - *beschreiben, beschreibt, beschreibe, beschrieb, beschrieben*
- Synonyme → Konzepte
 - *production, manufacture*
 - *Herstellung, Produktion, Fertigung*

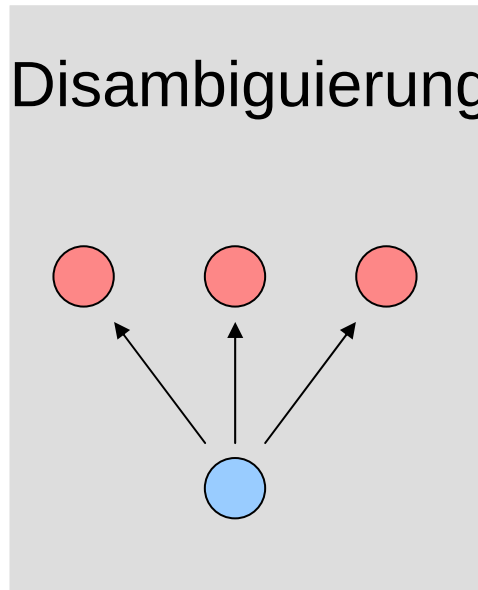


Interpretation

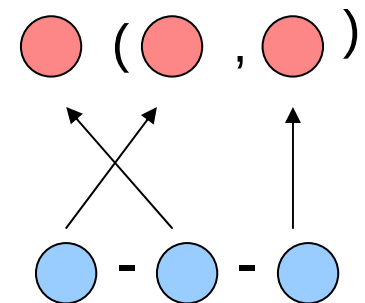
Abstraktion



Disambiguierung



Komposition



WordNet – call, 28 senses

1. **name, call** -- (assign a specified, proper name to;
"They named their son David"; ...)
2. **call, telephone, call up, phone, ring** -- (get or try to get into
communication (with someone) by telephone;
"I tried to call you all night"; ...)
3. **call** -- (ascribe a quality to or give a name of a common noun that
reflects a quality; *"He called me a bastard"; ...)*
4. **call, send for** -- (order, request, or command to come;
"She was called into the director's office"; "Call the police!")



Disambiguierung

Disambiguierung erfordert Kontext:

- Kookkurenz-Information (Dokument, n-Wort-Fenster)
- Strukturierter linguistischer Kontext

Disambiguierung nach:

- Wortart
- Wortbedeutung
- Syntaktischer Struktur
- Semantischer Struktur
- ...



Bedeutung und Kontext

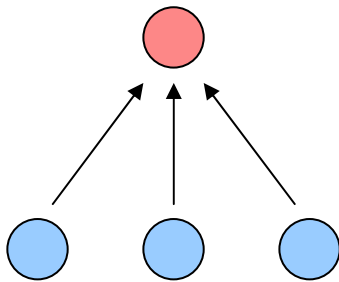
Die Bedeutung hängt vom linguistischem und extralinguistischen Kontext ab

- Linguistischer Kontext
 - D: „Der BP hat eine Amtszeit von vier Jahren. **Er** wird von der Bundesversammlung gewählt“ **Anaphern**
 - D: „The proof of the pudding is in the eating“ **Metaphern**
- Extralinguistischer Kontext
 - D: „Wetter **morgen**“ **Deixis**
 - D: „Herr Schneider hat die Telefonnummer 4315“ **Referenz**

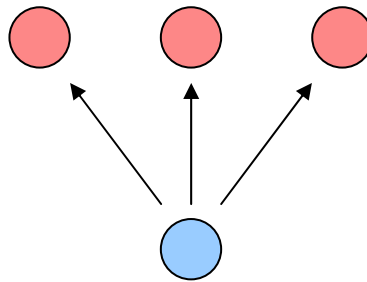


Interpretation

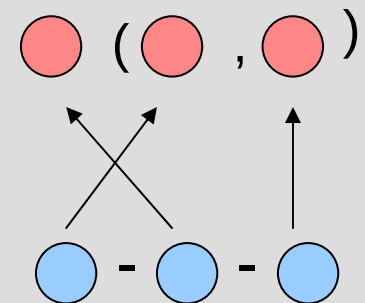
Abstraktion



Disambiguierung



Komposition



Ein spezielles Problem

- Operatoren wie Negation oder Ausdrücke des Glaubens oder Zweifels nehmen einem eingebetteten Satz die Kraft des Faktischen, die Faktizität. Wenn sie nicht berücksichtigt werden, gibt es viele falsche Treffer bei der Informationssuche.
- **Negation**
 - Q: „discover America“
 - D: „The Italians did **not** discover America“
- **Einbettung**
 - Q: „Wahl Bundespräsident“
 - D: „50% der Deutschen **glauben**, daß der Bundespräsident direkt vom Volk gewählt wird“



Informationssuche: Anforderungen

- Präzision
- Abdeckung
- Effizienz



Sprachverstehen - Abstufungen

- Wie tief muß man analysieren?
 - Zweck des Sprachverstehens
- Wie groß muß die Abdeckung sein?
 - Fester Text (feste Textsorte) oder Internet
- Wie domänenspezifisch ist die Anwendung?
 - Domänenwissen / Allgemeinwissen



Aktuelle Ansätze zur Suche nach relevanter Information

- Information Retrieval (www.google.com)
 - Domänenunspezifisch, sehr flach
 - Ausgabe: Liste von Dokumenten
- Question Answering (answerbus.coli.uni-sb.de)
 - Wie IR. Unterschied: Ausgabe ist (kurzer) Antworttext.
 - flach, aber mit zusätzlicher Verarbeitungsstufe
- Information Extraction (www.gate.ac.uk/annie)
 - Domänenspezifisch, tief(er)
 - strukturierte Ausgabe



Information Retrieval

- Gegeben: Anfrage (Query)
- Gesucht: **Relevante** Dokumente
- Weitverbreitete Methode: **Semantischer Raum**
 - Nähe im semantischen Raum **modelliert** Relevanz



Beispiel: Vorlesungsankündigung 1

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 .

Term

Term-
frequenz (tf)

die: 3
Veranstaltung: 1
werden: 2
als: 1
Ringvorlesung: 1
durchführen: 1
...
Morphologie: 1
Syntax: 1
....



Beispiel: Vorlesungsankündigung 2

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.

Ziel: 1
die: 4
Veranstaltung: 1
sein: 1
es: 1
Teilnehmer: 1
...
Syntax: 1
Morphologie: 1
...



Beispiel: FAZ-Politik-Artikel

Gegen den Widerstand von
Arbeitsminister Clement haben
sich Bundeskanzler Schröder
und die SPD- Spitze für eine
Ausbildungsabgabe
ausgesprochen. Ein
entsprechender Beschluß der
Bundestagsfraktion wird für
Montag erwartet

gegen: 1
der: 1
Widerstand: 1
von: 1
Arbeitsminister: 1
Clement: 1
haben: 1
...
die: 2
...



Query

„Welche Veranstaltung
behandelt Morphologie
und Syntax?“

welche: 1
Veranstaltung: 1
behandeln: 1
Morphologie: 1
und: 1
Syntax: 1



Vektoren

die: 3
Veranstaltung: 1
werden: 2
als: 1
Morphologie: 1
Syntax: 1
Widerstand: 0
Arbeitsminister: 0
Clement: 0
...

Dokument 1

die: 4
Veranstaltung: 1
werden: 0
als: 0
Syntax: 1
Morphologie: 1
Widerstand: 0
Arbeitsminister: 0
Clement: 0
...

Dokument 2

die: 2
Veranstaltung: 0
werden: 1
als: 0
Syntax: 0
Morphologie: 0
Widerstand: 1
Arbeitsminister: 1
Clement: 1
...

Dokument 3

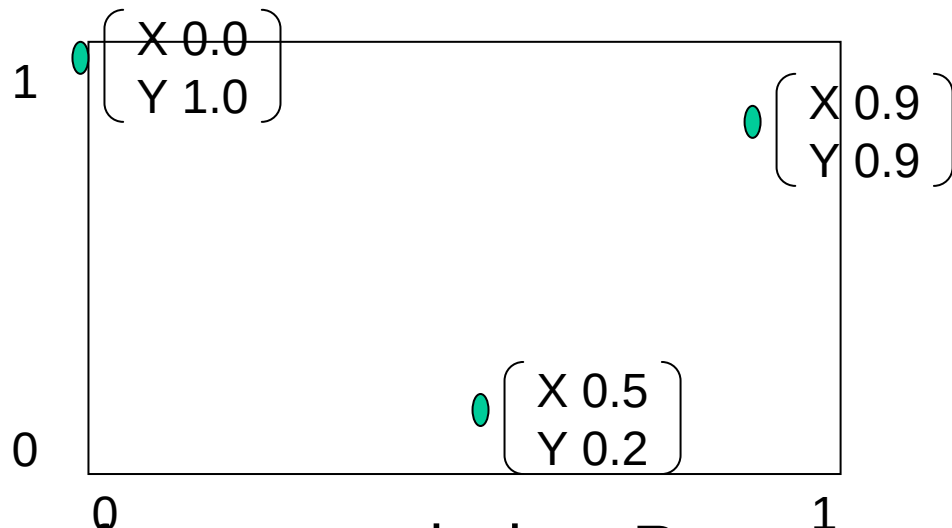
die: 0
Veranstaltung: 1
werden: 0
als: 0
Syntax: 1
Morphologie: 1
Widerstand: 0
Arbeitsminister: 0
Clement: 0
...

Query



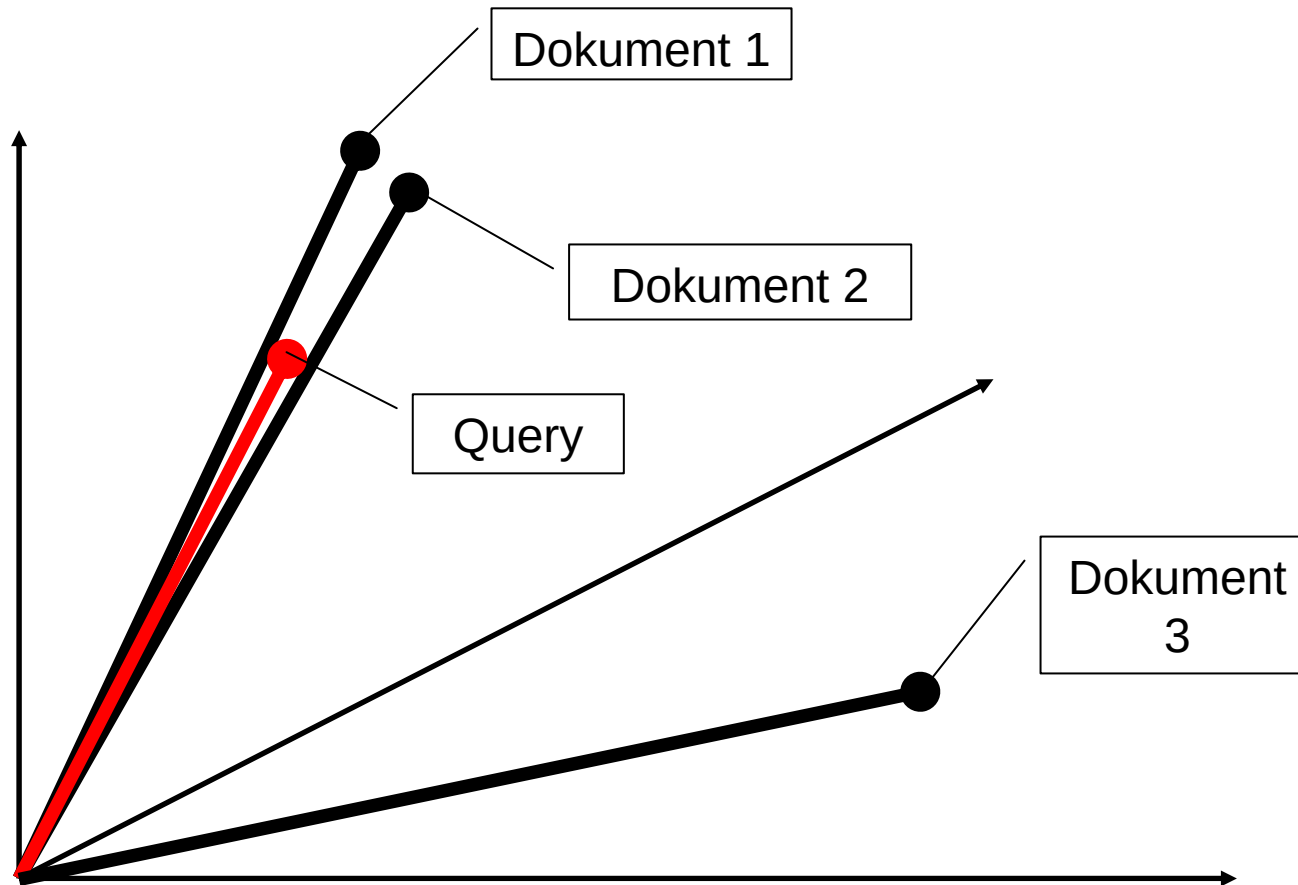
Punkte und Vektoren

- Jeder Punkt kann als Vektor verstanden werden



- Dimensionen fuer semantischen Raum:
Woerter in Dokument ("Terme") und ihre Frequenzen
 - Ergibt hochdimensionalen Raum

Semantischer Raum (vereinfacht)



Relevante Dokumente für die Anfrage: Dokumente 1 und 2

Vorteile von semantischen Räumen

- Nutzung von Frequenzinformation
 - Konzeptuell einfach, effizient
 - Dokumente sind ähnlich, wenn Begriffe **gleich oft** vorkommen
- Formalisierung
 - Mathematische Standardverfahren zur Berechnung von Ähnlichkeit / Relevanz
- Leichte Erweiterbarkeit mit mehr Wissen
 - Mathematische / statistische Methoden (z.B. Googles PageRanking)
 - Linguistische Verfahren: genauere Modellierung der Terme



Genauere Modellierung von Termen

- Nicht alle Worte sind gleich
 - **Stoppworte** komplett entfernen
 - Sehr häufig (sein, werden, ...)
 - Funktionswörter (Präpositionen, Konjunktionen, ...)
 - **Informative Worte** stärker werten
 - „Worte in wenigen Dokumenten sind informativ“
 - **tf * idf**: Termfrequenz * Inverse Dokumentfrequenz



Query Expansion mit Wortformen und Synonymen

- Query Expansion:
 - Anfrage mit Suchwörtern *<aircraft, manufacture>* wird erweitert zu
 - *< aircraft, manufacture, production, industry>* (Synonyme)
 - *< aircraft, airplane, plane, helicopter, vehicle, ...>* (Hyponyme, Hyperonyme)



Beurteilung von Information Retrieval

- Gut zur Suche von Dokumenten aus großen Datenmengen
 - einfach zu realisieren
 - schnell
- Problem: Niedrige Präzision
 - Falsche Treffer
 - Ergebnis nur Liste von Dokumenten
- Rolle von sprachlichem Wissen
 - Wenig Wissen nötig
 - Kann zur Optimierung des semantischen Raumes dienen
 - Stoppwörter, Kombination verwandter Wörter



Information Extraction

Who did what to whom?

- Fülle **Rollen** in **Template** mit Information
 - Ignoriere Rest des Textes
 - Information muß als Template darstellbar sein
 - Information muss mithilfe einfacher Regeln im Text identifizierbar sein
- Beispiele:
 - Vortragsankündigung (wer, wann, wo, worüber)
 - Wetterbericht (wann, wo, wie)
 - Wirtschaftsmeldungen (wer, wen, was)



Vortragsankündigung

Am Donnerstag, den 13.11.2003, redet Martha Palmer (University of Pennsylvania) um 16:15 im Seminarraum (Geb. 17.1) zum Thema „Putting Meaning into your Trees“.

Redner: ?

Zeit: ?

Datum: ?

Ort: ?

Titel: ?



Schritt 1: Datenaufbereitung

- POS-Tagging
 - um, am, im: PRP
 - redet: VVFIN
- Named Entity Recognition
 - PERSON, ORGANISATION, TIME, DATE, QUANTITY...
- Flache Grammatik
 - Phrasen erkennen
 - PRP + TIME → Präpositionalphrase (PP)

Einzelne Module
entweder **symbolisch**
oder **statistisch**



Schritt 2: Scenario oder Event Patterns

[Am DATE] redet PERSON (ORGANISATION) [um TIME] [im PLACE] [zum Thema [„Putting Meaning into your Trees“]].

- Regeln kodieren Wissen darueber, wie Information aus Template sprachlich ausgedrueckt wird („Abbildung Sprache nach Bedeutung“)
 - Wenn [_{PP} um TIME], dann Zeit → TIME
 - Wenn [_{PP} zum Thema S], dann Titel → S



Vortragsankündigung

Am [PP Donnerstag, den 13.11.2003], redet Martha Palmer
(University of Pennsylvania) [PP um 16:15] [PP im
Seminarraum (Geb. 17.1)] [PP zum Thema [„Putting
Meaning into your Trees“]].

Redner: Martha Palmer

Zeit: 16:15

Datum: Donnerstag, den 13.01.2003

Ort: Seminarraum (Geb. 17.1)

Titel: „Putting Meaning into your Trees“



Beurteilung von Information Extraction

- Gut für Suche nach spezieller Information
 - Templates gut zur Weiterverarbeitung
 - Relativ sicheres Wissen
 - Einigermassen gut automatisierbar
- Problem: **Flexibilität**
 - Wortwahl: „über“ vs. „zum Thema“
 - Satzbau: „XY redet am 01.11.“. vs. „Am 01.11. redet XY“
 - Abdeckung der Regeln?
 - Übertragbarkeit auf andere Domänen problematisch
- Rolle von sprachlichem Wissen:
 - Wissen ueber Domänenstruktur: Definition der Rollen
 - Sprachliche Realisierung von Rollen in Event Pattern



Question Answering

- Gegeben: Query (als umgangssprachliche Frage)
- Gesucht: Relevanter Satz (aus Dokument)
- Typische QA-Systeme machen nur **Extraktion**
 - Schritt 1: IR → Liste von Dokumenten
 - Schritt 2: Extraktion der relevanten Stellen

Zur Extraktion ist **tiefe(re) Verarbeitung** nötig!



Welche Stellen sind relevant?

- Zentrale Idee: Relevante Stellen treffen **Aussage** über das **gefragte Objekt**
 - Überlappung mit Frage in Worten ist zu unspezifisch
 - **Semantische** Repräsentation nötig
- Fragenklassifikation: Wonach wird gefragt?
 - „Wie viele Sechsecke sind auf einem Fußball?“ (Zahl)
 - sein: ?<Zahl> Sechsecke, auf Fußball
 - „Wo ging Bill Gates auf College?“ (Ort)
 - gehen: Bill Gates, College, ?<Ort>
- Strategie: Finde Aussage in Dokument, **die die Lücke in der Anfrage füllen kann**



Beurteilung von Question Answering

- Gibt relevanten Satz zurück
 - Benutzerfreundlichster Ansatz
- Question Answering ist schwer
 - Aufwändig
 - Robustheit großes Problem
 - Oft für begrenzte Domänen untersucht
 - Richtung „Expertensysteme“
- Rolle von sprachliches Wissen
 - Braucht **deutlich mehr** Wissen als reines Information Retrieval



Zusammenfassung und Ausblick

- Information Management ist schwierig
 - Wenig Wissen: erstaunlich gute Ergebnisse (IR)
 - Qualitativer Sprung (QA) erfordert viel Wissen
- Verschiedene Verfahren für verschiedene Aufgaben
 - Homogene Daten, kleine Domäne: Information Extraction
 - Domänenunabhängige Suche: Information Retrieval
 - Mit viel Wissen: Question Answering
- Sehr aktives Gebiet
 - Text Retrieval Conference (TREC)
 - Message Understanding Conference (MUC)

