

Informationsmanagement und die Rolle von Wissen

Sebastian Padó

Sprachtechnologische Anwendungen

- ▶ **Speech technology**
 - ▶ Spracherkennung
 - ▶ Sprechererkennung
- ▶ **Text technology**
 - ▶ Sprachassistentz
 - ▶ Maschinelle Übersetzung
 - ▶ **Informationsmanagement**

Informationsmanagement: Was ist das?

Große Datenmengen
zugänglich und nutzbar machen

- ▶ Konkret:
 - ▶ Dokumente klassifizieren
 - ▶ Dokumente zusammenfassen
 - ▶ Relevante Informationen identifizieren
 - ▶ **Relevante Dokumente für Anfragen finden**

▶ 2

Anfragen

- ▶ Im Internet
 - ▶ „Wie starb Sokrates?“
- ▶ Firmen-Intranet
 - ▶ „Welche Telefonnummer hat Herr Schneider?“
- ▶ Online-Katalog-Recherche
 - ▶ „Was kostet das neue Buch von Neil Gaiman?“

Jede Frage hat ihre besonderen Schwierigkeiten

▶ 3

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?

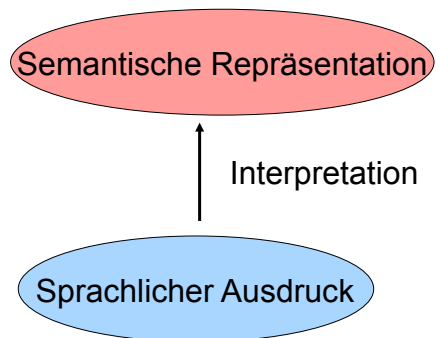
▶ 4

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**

▶ 5

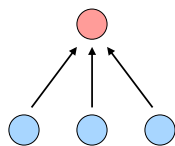
Interpretation



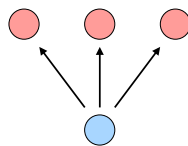
▶ 6

Interpretation: Drei grundlegende Aspekte

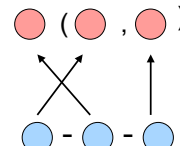
Abstraktion



Disambiguierung

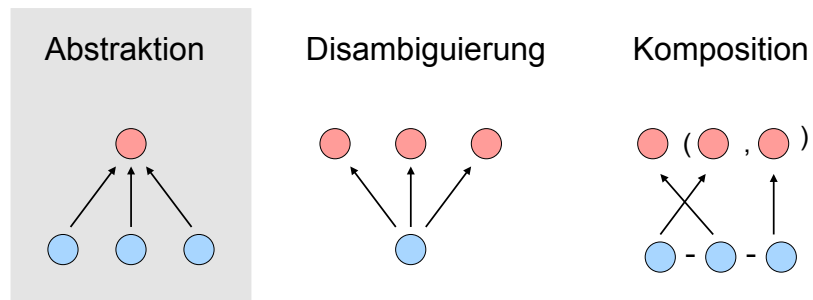


Komposition



▶ 7

Interpretation



▶ 8

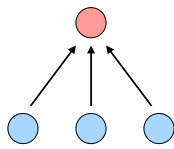
Abstraktion von der Oberfläche

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

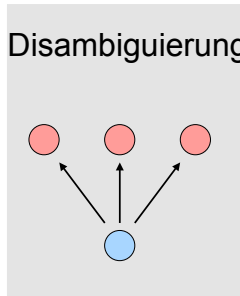
▶ 9

Interpretation

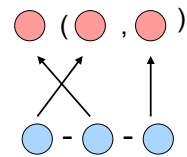
Abstraktion



Disambiguierung



Komposition



▶ 10

Disambiguierung von Bedeutung

▶ Homonymie

- ▶ innersprachlich: Bank, Zug, ...
- ▶ sprachübergreifend: Aller, Porto, ...

▶ Polysemie

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!")

▶ 11

Disambiguierung

Disambiguierung erfordert **Kontext**:

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

Disambiguierung nach:

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

▶ 12

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

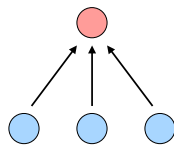
▶ Extralinguistischer Kontext

- ▶ D: „Wetter **morgen**“ Deixis
- ▶ D: „Herr Schneider hat die Telefonnummer 4315“ Referenz

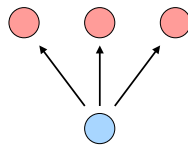
▶ 13

Interpretation

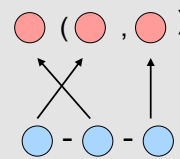
Abstraktion



Disambiguierung



Komposition



▶ 14

Bedeutung und Syntax

Operatoren wie Negation oder Ausdrücke des Glaubens oder Zweifels nehmen einem eingebetteten Satz die Kraft des Faktischen (**Faktizität**).

▶ 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“

▶ 15

Konsequenzen für den naiven Algorithmus

- ▶ **Reine Wort-für-Wort-Suche nicht möglich**
 - ▶ Man findet irrelevante Dokumente
 - ▶ Einbettung, Homonymie, Negation, Metaphern, ...
 - ▶ Man findet nicht alle relevanten Dokumente
 - ▶ Synonymie, Hyponymie, Anaphern, ...

- ▶ **Alternativansatz: Mehr Sprachverstehen**

▶ 16

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.
 - ▶ zunächst flach, dann tiefe, typischerweise domänenspezifische linguistische Verarbeitung

- ▶ **Information Extraction (www.gate.ac.uk/ie/ie_example.html)**
 - ▶ domänenspezifisch, tief(er)
 - ▶ strukturierte Ausgabe („templates“)

▶ 17

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)

▶ 18

Vortragsankündigung

Am Donnerstag, den 13.11.2007, redet Martha Palmer (University of Pennsylvania) um 16:15 in der Azenbergstrasse 12.21 zum Thema „Putting Meaning into your Trees“.

Redner: ?
 Zeit: ?
 Datum: ?
 Ort: ?
 Titel: ?

▶ 19

Schritt 1: Datenaufbereitung

- ▶ **POS-Tagging**
 - ▶ um, am, im: PRP
 - ▶ redet: VFIN

- ▶ **Named Entity Recognition**
 - ▶ PERSON, ORGANISATION, TIME, DATE, QUANTITY...

- ▶ **Flache Grammatik**
 - ▶ Phrasen erkennen
 - ▶ PRP + TIME → Präpositionalphrase (PP)

▶ 20

Schritt 2: Scenario oder Event Patterns

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

- ▶ **Analyseregeln** kodieren Wissen darüber, 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**

▶ 21

Vortragsankündigung

Am [pp **Donnerstag, den 13.11.2007**], redet **Martha Palmer**
(University of Pennsylvania) [pp um **16:15**] [pp **in der**
Azenbergstrasse 12.21] [pp zum Thema [„Putting Meaning
into your Trees“]].

Redner: Martha Palmer

Zeit: 16:15

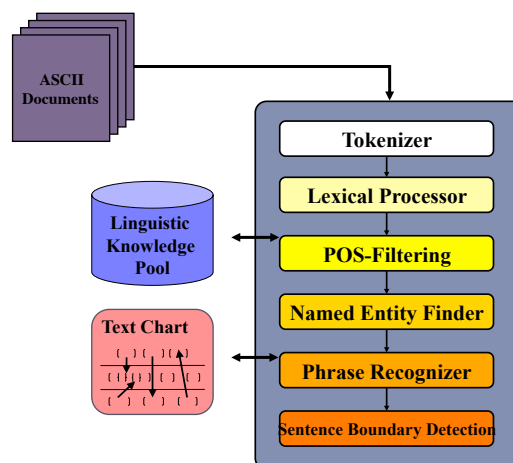
Datum: **Donnerstag, den 13.11.2007**

Ort: **Azenbergstrasse 12.21**

Titel: „Putting Meaning into your Trees“

▶ 22

Beispiel: SPPC-System (DFKI)



▶ 23

Information Extraction: Beurteilung

- ▶ 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

▶ 24

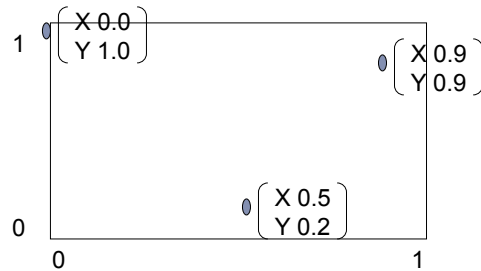
Information Retrieval

- ▶ Gegeben: Anfrage (Query)
- ▶ Gesucht: **Relevante** Dokumente
- ▶ Verbreitetste Methode: **Semantischer Raum**
 - ▶ Jedes Dokument ist ein Punkt
 - ▶ Query ist auch ein Punkt
 - ▶ Nähe im semantischen Raum **modelliert** Relevanz

▶ 25

Punkte und Vektoren

- ▶ Jeder Punkt kann als Vektor verstanden werden



- ▶ Dimensionen fuer semantischen Raum:
Woerter in Dokument ("Terme")
- ▶ Dokumente werden dargestellt als „Bags of words“

▶ 26

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
....	

▶ 27

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
 ...

▶ 28

Beispiel: FAZ-Politik-Artikel

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

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

▶ 29

Query

„Welche Veranstaltung
behandelt Morphologie
und Syntax?“

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

▶ 30

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

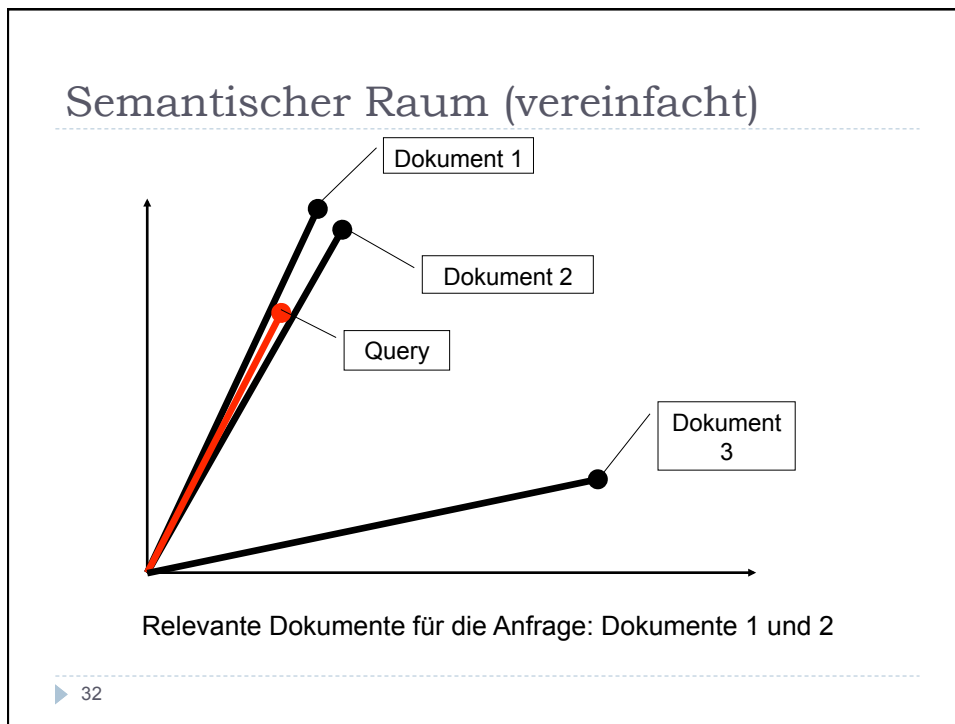
die: 2
Veranstaltung: 0
werden: 1
als: 0
Syntax: 0
Morphologie: 0
Widerstand: 1
Arbeitsminister: 1
Clement: 1
...

Dokument 3

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

Query

▶ 31



Vorteile von semantischen Räumen

- ▶ **Ähnlich zu „naiver Suche“**
 - ▶ Konzeptuell einfach, effizient
- ▶ **Nutzung von Frequenzinformation**
 - ▶ Dokumente sind ähnlich, wenn Begriffe **gleich oft** vorkommen
- ▶ **Formalisierung**
 - ▶ Mathematische Standardverfahren zur Berechnung von Ähnlichkeit / Relevanz
- ▶ **Erweiterbarkeit mit mehr Wissen**
 - ▶ Mathematische / statistische Methoden (z.B. Googles Link-Rating)
 - ▶ 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
 - ▶ „die“, „ist“ sind uninformativ

- ▶ Anfragen werden mit verwandten Wörtern erweitert (**Query Expansion**)
 - ▶ Anfrage <aircraft, manufacture> wird erweitert zu
 - ▶ < aircraft, manufacture, **production, industry**> (Synonyme)
 - ▶ < aircraft, **airplane, plane, helicopter, vehicle, ...**> (Hyponyme, Hyperonyme)

▶ 34

Beurteilung von Information Retrieval

- ▶ Gut zur Suche von Dokumenten aus großen Datenmengen
 - ▶ einfach zu realisieren
 - ▶ schnell

- ▶ Problem: Qualität der Ergebnisse
 - ▶ **Beruhet auf Redundanz**
 - ▶ Falsche Treffer

- ▶ Rolle von sprachlichem Wissen
 - ▶ Wenig Wissen nötig
 - ▶ Kann zur Optimierung des semantischen Raumes dienen
 - ▶ Stopwörter, Kombination verwandter Wörter

▶ 35

Question Answering

- ▶ Gegeben: Query
- ▶ 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!

▶ 36

Welche Stellen sind relevant?

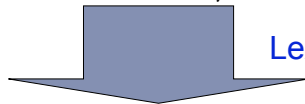
- ▶ **Zentrale Idee: Relevante Stellen treffen Aussage über das gefragte Objekt**
 - ▶ Überlappung mit Frage in Worten ist zu unspezifisch
 - ▶ **Semantische** Repräsentation nötig
- ▶ **Fragenklassifikation: Nach welchem semantischen Typ wird gefragt?**
 - ▶ „Wie viele Sechsecke sind auf einem Fußball?“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

▶ 37

Beispiel 1

Auf einem Fußball befinden sich 20 Sechsecke

befinden(20 Sechsecke, auf Fußball)



Lexikon: Synonymie

sein(20 Sechsecke, auf Fußball)

Anfrage: sein(<?Zahl> Sechsecke, auf Fußball)

Antwort: 20

▶ 38

Beispiel 2

Bill Gates, einst Harvard-Abbrecher, ist heute einer der reichsten Männer Amerikas.

Abbrecher(Bill Gates, Harvard)



Lexikon / Grammatik: (De-)Nominalisierung

abbrechen(Bill Gates, Harvard)



Weltwissen: Harvard ist eine Universität

abbrechen(Bill Gates, Harvard_University)



Lexikon/Weltwissen: „abbrechen“ impliziert vorheriges „gehen“

gehen(Bill Gates, Harvard_University)



Lexikon: Universität ist Synonym zu College

gehen(Bill Gates, <?College>)

Antwort: auf die Harvard University

▶ 39

Beurteilung von Question Answering

- ▶ **Gibt relevanten Satz zurück**
 - ▶ Benutzerfreundlichster Ansatz
- ▶ **Question Answering ist schwer**
 - ▶ Aufwändig
 - ▶ Robustheit großes Problem
 - ▶ Oft für begrenzte Domänen realisiert
 - ▶ Richtung „Expertensysteme“
- ▶ **Rolle von sprachliches Wissen**
 - ▶ Normalisierung – kann auch Anfragen beantworten, die nicht wörtlich im Text beantwortet werden
 - ▶ Braucht **deutlich mehr** Wissen als reines Information Retrieval

▶ 40

Zusammenfassung und Ausblick

- ▶ **Information Management ist schwierig**
 - ▶ Wenig Wissen: erstaunlich gute Ergebnisse (IR) aber abhängig von Redundanz
 - ▶ 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)

▶ 41