
Information Management und die Rolle von Wissen

Sebastian Padó

Sprachtechnologie

- Speech technology

- Spracherkennung
- Sprechererkennung

- Text technology

- Sprachassistentz
- Maschinelle Übersetzung
- Informationsmanagement

Nötiges Wissen

Phonetisches Lexikon

Phonetisches Profil

Morphologie, Syntax

Alles

??

Information Management - Was ist das?

**Große Datenmengen
zugänglich und nutzbar machen**

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

Fragen über Fragen

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

Welche Frage ist am schwierigsten?

Information Management

- Probleme des IM
 - Sprache und Bedeutung
 - Klassifikation von Methoden des Sprachverstehens
 - Ansätze zu IM
 - Information Extraction
 - Information Retrieval
 - Question Answering
-

Probleme des IM

- Technische Probleme:
 - Immense Datenmengen
 - ständig wachsend (Internet!)
 - Computerlinguistische Probleme:
 - Formale Probleme
 - Daten nicht strukturiert (i. Vgl. zu Datenbank)
 - Inhaltliche Probleme
 - Daten in verschiedenen Sprachen
 - Linguistisches Material ist nicht gleich Bedeutung
-

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



Bedeutung und das Lexikon

Es gibt keine bijektive Abbildung
zwischen Worten und Konzepten

- Ein Wort, mehrere Konzepte

- Deutsch: Bank, Roller
- Verschiedene Sprachen: Porto, Aller, Bad

Homonymie,
Polysemie

- Ungenauigkeit von Worten

- Blau, groß

Vagheit

- Ein Konzept, mehrere Worte

- {Auto(mobil), Fahrzeug, Wagen, ...}

Synonymie,
Hyponomie

Bedeutung und Syntax

Die Bedeutung eines Ausdrucks ergibt sich aus der Bedeutung der einzelnen Wörter **plus** ihrer syntaktischen Beziehung (**Kompositionalität**)

■ 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“
-

Bedeutung und Kontext

Die Bedeutung hängt vom linguistischem und extralinguistischen Kontext ab

- Linguistischer Kontext

- D: „Der BP hat eine Amtszeit von vier Jahren. Anaphern
 Er wird von der Bundesversammlung gewählt“
- D: „The proof of the pudding is in the eating“ Metaphern

- Extralinguistischer Kontext

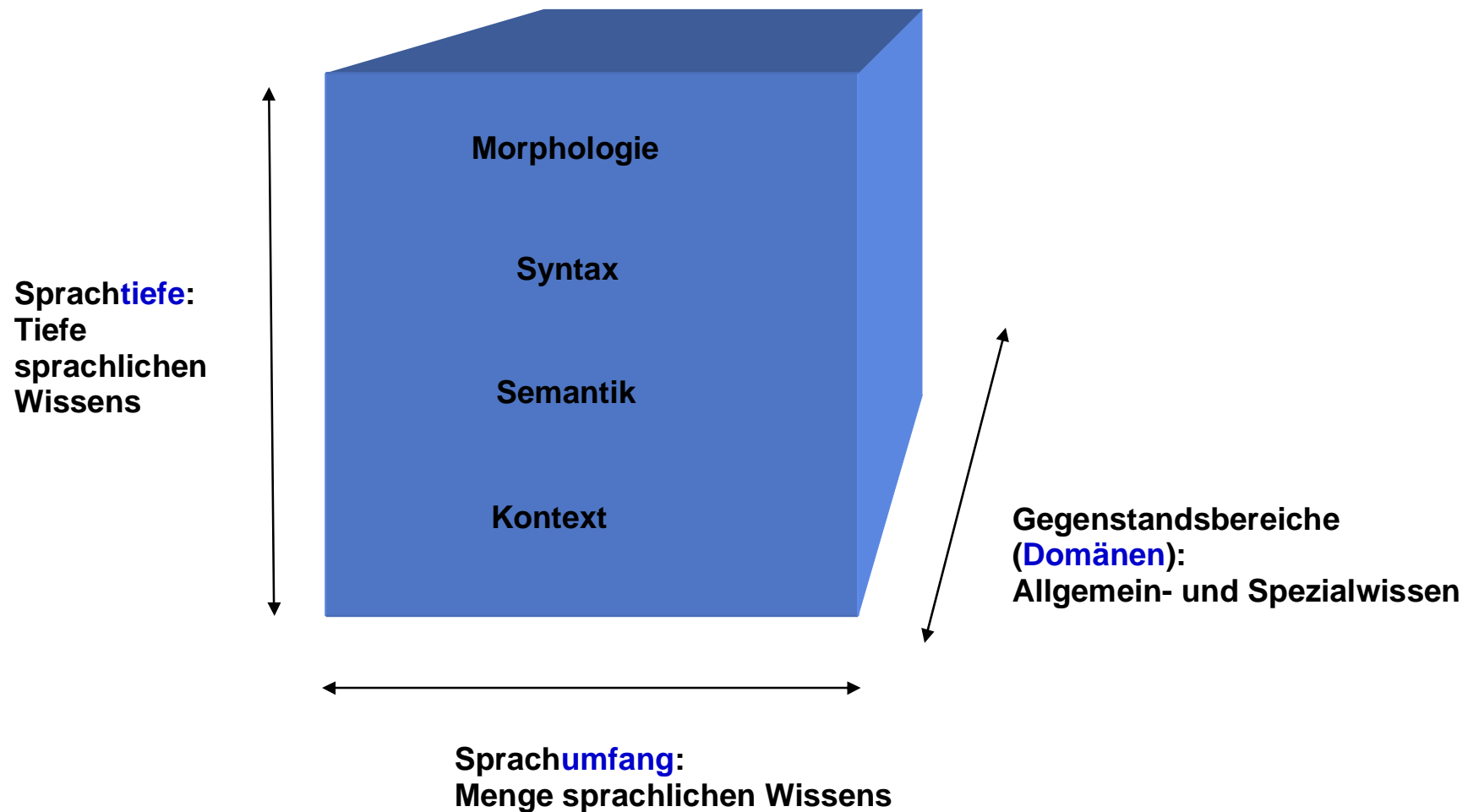
- D: „Wetter **morgen**“ Deixis
-

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, Hyponomie, Anaphern, ...

 - Alternativansatz: Volles Sprachverstehen
-

Sprachverstehen



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 ist einfach zu formalisieren
 - Allgemeinwissen ist sehr schwer zu formalisieren

Volles Sprachverstehen ist zur Zeit nicht möglich

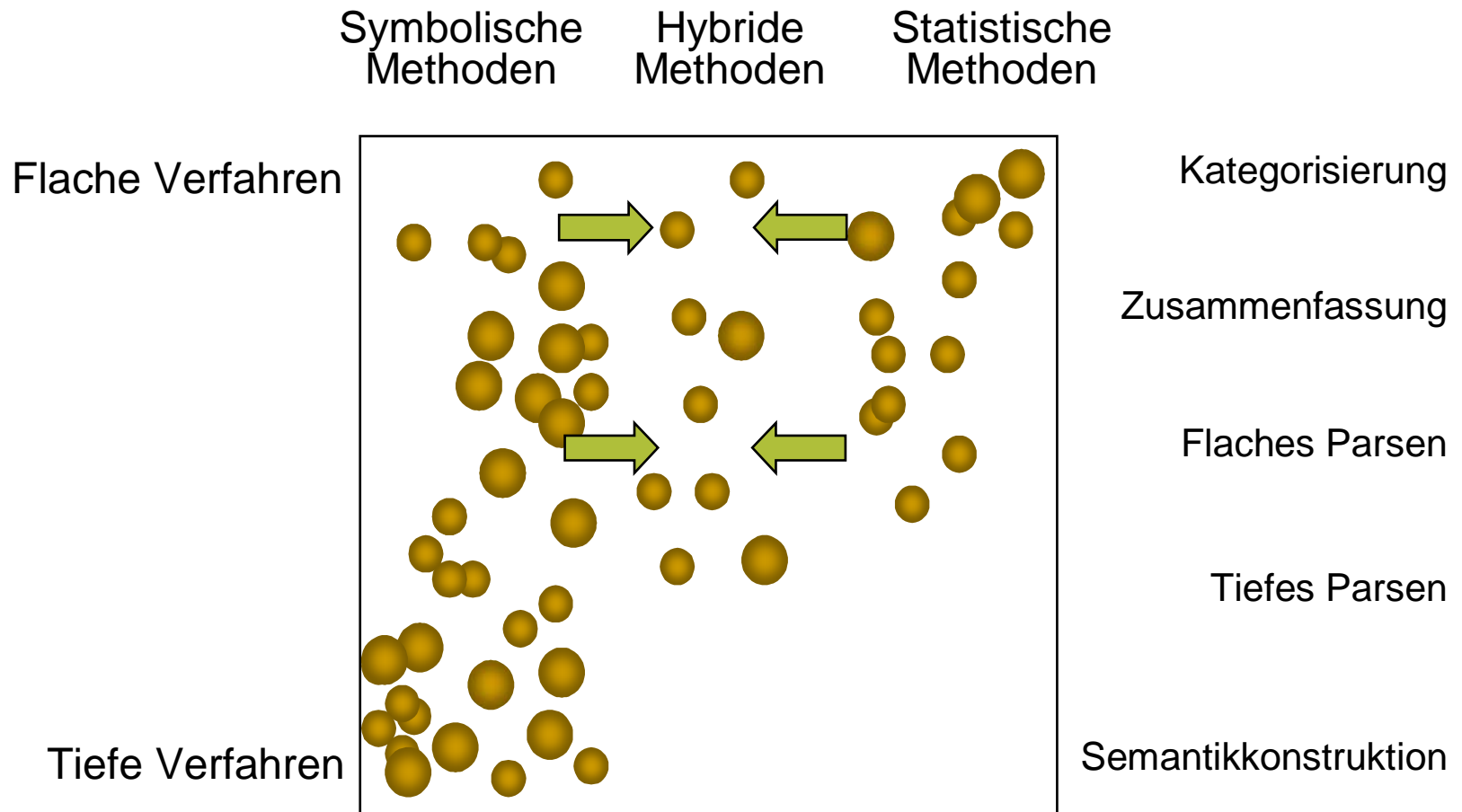
Flache und tiefe Methoden

- **Tiefe** Methoden (deep processing)
 - ❑ Vollständiges Verstehen
 - ❑ Vorteile: sicheres, informatives Resultat
 - ❑ Nachteile: Komplexität, störanfällig, spezifisch
 - **Flache** Methoden (shallow processing)
 - ❑ Nur so viel verstehen, wie nötig oder möglich
 - ❑ Vorteile: Schnell, Robust, Flexibel
 - ❑ Nachteile: Unvollständig, unsicher
-

Symbolische und statistische Methoden

- **Symbolische** Methoden: algebraische Mathematik und Logik
 - Grammatiken, reguläre Ausdrücke
 - Regeln, Ableitungssysteme
 - **Statistische** Methoden: analytische Mathematik
 - Statistische Klassifikation
 - Lernen von Funktionen
-

Methoden und Anwendungen



Ansätze im Information Management

Erinnerung: volles Sprachverstehen ist nicht realistisch

- Information Extraction (www.gate.ac.uk/annie)
 - domänenspezifisch
 - strukturierte Ausgabe
 - Information Retrieval (www.google.com)
 - domänenunspezifisch
 - Ausgabe: Liste von Dokumenten
 - Question Answering (answerbus.coli.uni-sb.de)
 - Wie IR. Unterschied: Ausgabe ist (kurzer) Antworttext.
-

Information Extraction

Who did what to whom?

- Fülle Rollen in Template mit Information
 - Ignoriere Rest des Textes
 - 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

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

- Named Entity Recognition

- PERSON, ORGANISATION, TIME, DATE, QUANTITY...

- Flache Grammatik

- Phrasen erkennen
 - PRP + TIME → Präpositionalphrase (PP)

Schritt 2: Scenario oder Event Patterns

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

■ Pattern Matching

- Wenn [_{PP} um **TIME**], dann Zeit → **TIME**
 - Wenn [_{PP} zum Thema **S**], dann Titel → **S**
 - etc.
-

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

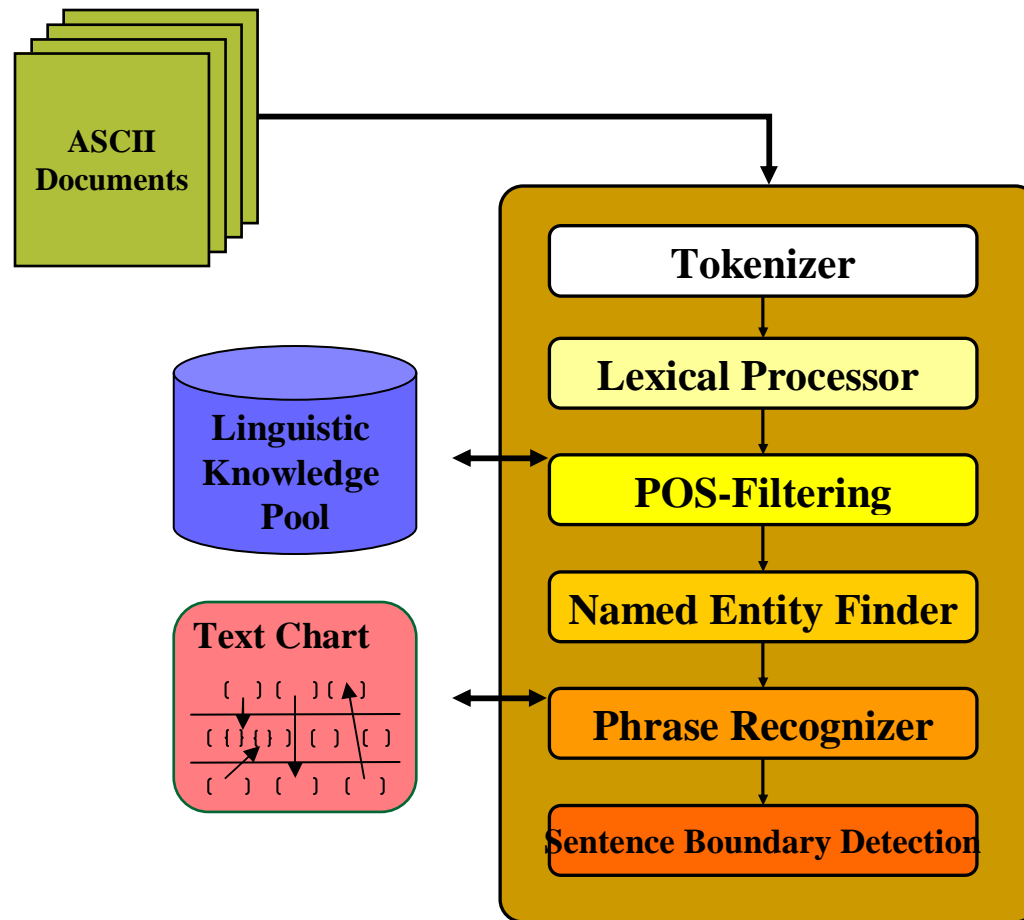
Zeit: 16:15

Datum: Donnerstag, den 13.01.2003

Ort: Seminarraum (Geb. 17.1)

Titel: „Putting Meaning into your Trees“

Beispiel: SPPC-System (DFKI)



Beurteilung von Information Extraction

- Gut für Suche nach spezieller Information
 - Templates gut zur Weiterverarbeitung
 - Relativ sicheres Wissen
 - 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
 - Rolle von Wissen:
 - Domänenstruktur: Rollen
 - Redewendungen und Fachvokabular in Regeln
-

Information Retrieval

- Gegeben: Anfrage (Query)
- Gesucht: **Relevante** Dokumente
- Weitverbreitete Methode: **semantischer Raum**
 - Jedes Dokument ist ein Punkt
 - Query ist auch ein Punkt
 - Ähnliche Dokumente = Nachbarpunkte

Nähe im semantischen Raum = 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)

Bag of
words

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

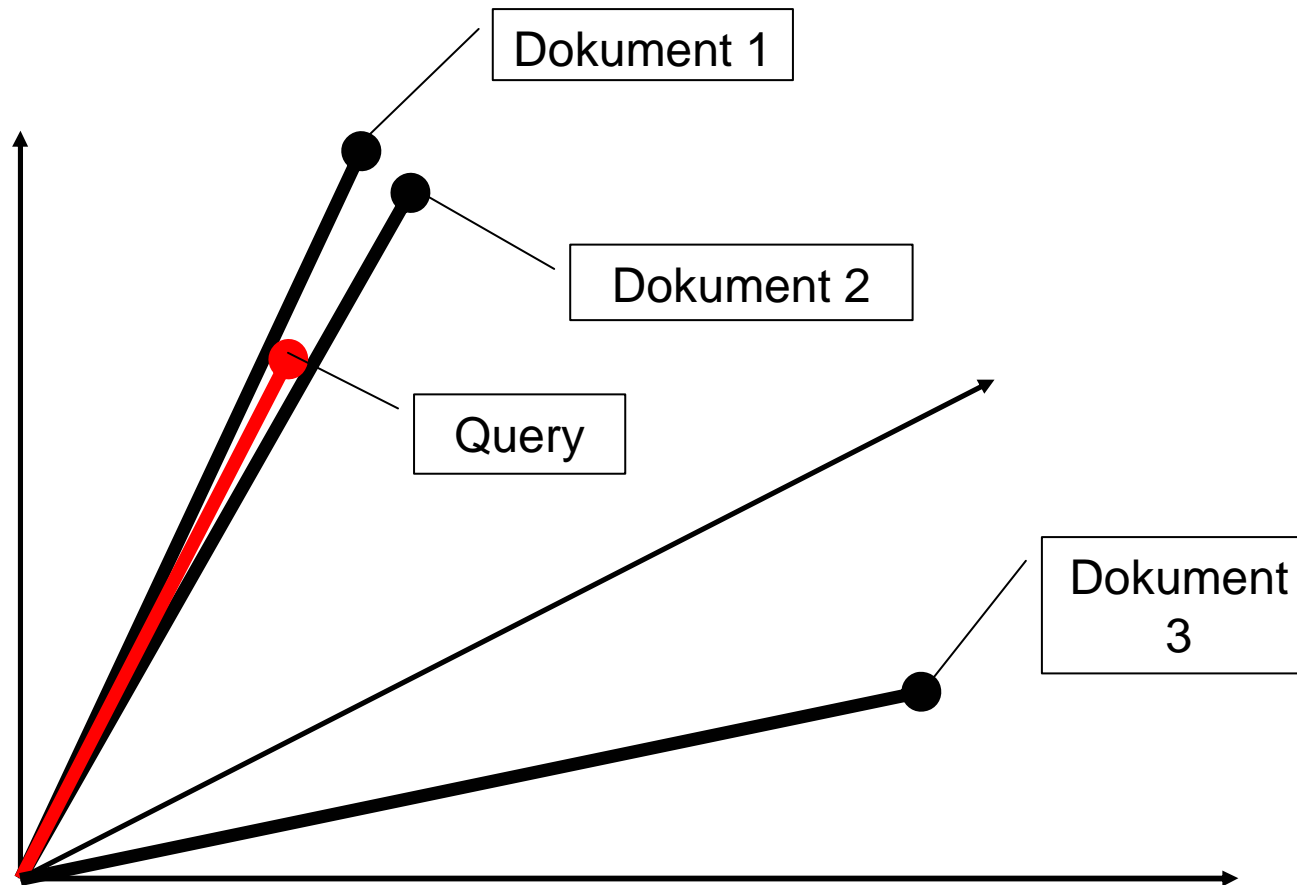
Arbeitsminister: 0

Clement: 0

...

Query

Semantischer Raum



Relevante Dokumente für die Anfrage: Dokumente 1 und 2

Was haben wir gewonnen?

- Ähnlich zu „naiver Suche“??
 - Vorteil des semantischen Raums:
 - Frequenzinformation
 - Ähnlichkeit wird quantifiziert
 - Formalisierung
 - Einsatz mathematischer / statistischer Verfahren
-

Verfeinerungen

- 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
- Worte können verwandt sein
 - Kombination „ähnlicher“ Dimensionen
 - Anreicherung (Paraphrasierung) der Anfrage

Beurteilung von Information Retrieval

- Gut zur Suche von Dokumenten aus großen Datenmengen
 - einfach zu realisieren
 - schnell
 - Problem: Qualität der Ergebnisse
 - Falsche Treffer
 - Ergebnis nur Liste von Dokumenten
 - Rolle von Wissen
 - Wenig Wissen nötig
 - Wissen schwer integrierbar
 - Optimierung des semantischen Raumes
 - Stopwörter, Kombination verwandter Wörter
-

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 **tieferer Verarbeitung** nötig!

Analyse der Frage

- Linguistische Analyse
 - Phrasenerkennung, Named Entity Recognition, ...
 - Einordnung in Fragengruppe
 - Ziel der Suche bestimmt Form der Suche:
 - „Wie viele Sechsecke sind auf einem Fußball?“ (Zahl)
 - „Wo ging Bill Gates auf College?“ (Ort)
 - Repräsentation der Frage als „Wissen mit Lücke“
 - `gehen(Bill Gates, College, ?<Ort>)`
 - `sein(Sechseck, Fußball, ?<Zahl>)`
-

Suche nach der Antwort

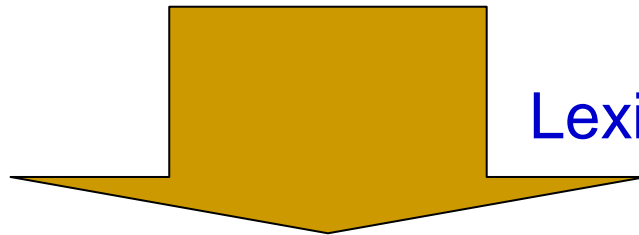
- Repräsentation der Sätze als Wissen
 - Tiefe Analyse
- Matching zwischen Anfrage und Sätzen
 - Wann passen zwei Repräsentationen zusammen?

Beide Vorgänge benötigen viel Wissen!

Beispiel 1

Auf einem Fußball befinden sich 20 Sechsecke

`befinden(Sechseck, Fußball, 20)`



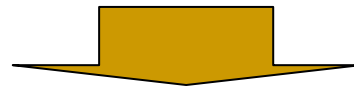
Lexikon: Synonymie

`sein(Sechseck, Fußball, 20)`

Beispiel 2

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

Abbrecher(Bill Gates, Harvard)



Lexikon: (De-)Nominalisierung

abbrechen(Bill Gates, Harvard)



Weltwissen: Harvard ist eine Universität

abbrechen(Bill Gates, Universität, Harvard)



Lexikon/Weltwissen: „abbrechen“ impliziert „gehen“

gehen(Bill Gates, Universität, Harvard)



Lexikon: Universität ist Synonym zu College

gehen (Bill Gates, College, Harvard)

Beurteilung von Question Answering

- Gibt relevanten Satz zurück
 - Benutzerfreundlichster Ansatz
 - Question Answering ist schwer
 - Aufwändig
 - Verlangt sehr viel Sprachtechnologie
 - Robustheit großes Problem
 - Rolle von Wissen
 - Braucht **viel** Wissen
 - Sprachliches Wissen
 - Weltwissen
-

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)
-