Information Management und die Rolle von Wissen

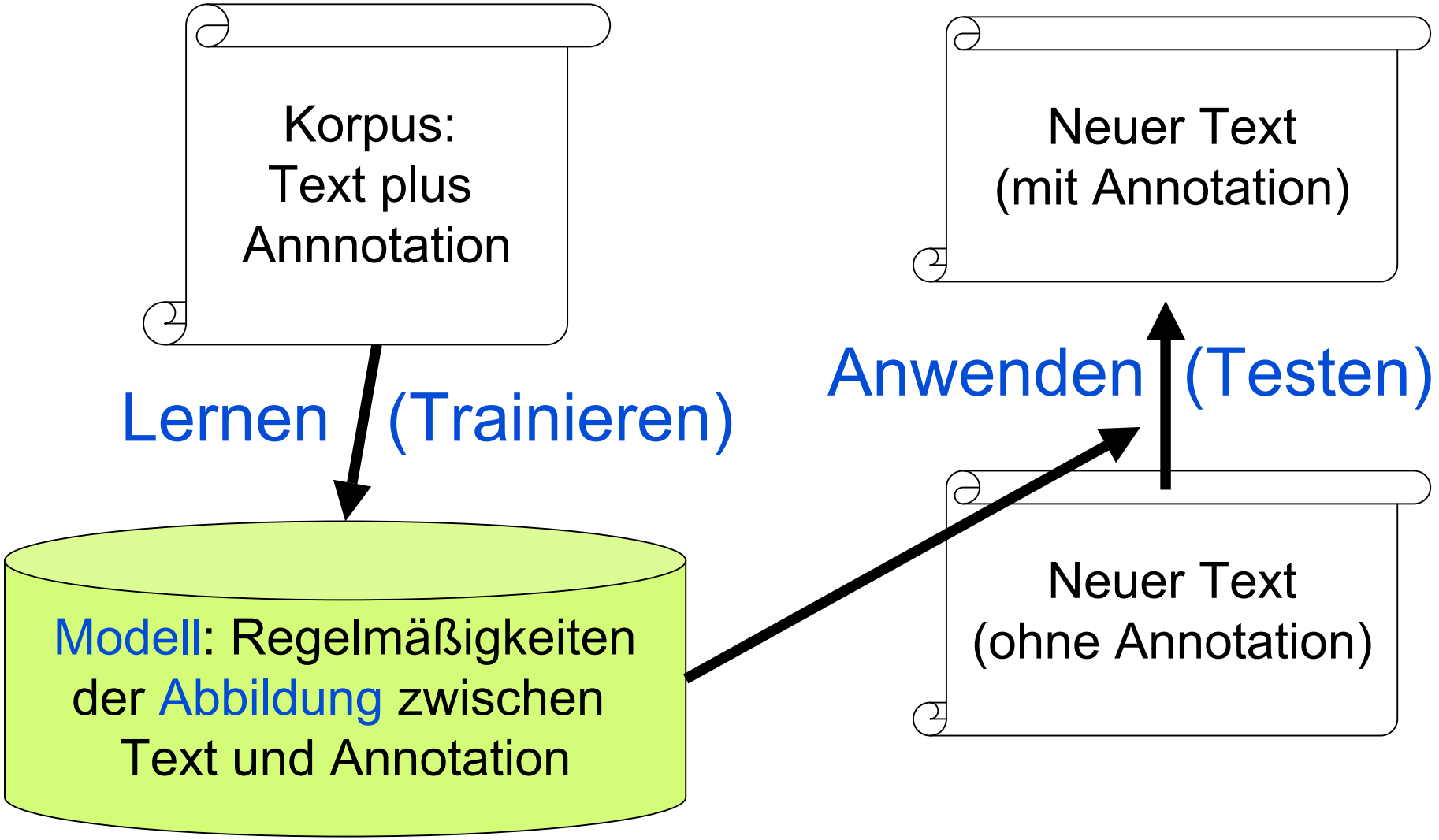
Sebastian Padó

Einschub: Nachtrag zu letztem Mal

- Thema: Korpora und Annotation

- Hauptfragen:
  - Was ist Korpuslinguistik (im Gegensatz zu theoretischer Linguistik)?
  - Wofür Annotation?
  - Was ist ein Annotationsschema?

Wiederholung: Maschinelles Lernen



“Stand der Kunst” im Lernen aus Korpora

- Vorverarbeitung:
  - für sehr viele Sprachen verfügbar
  - sehr hohe Korrektheit (>95%)
- Syntaktische Analyse
  - für viele Sprachen verfügbar
  - ausreichend gut (75-85%)
- Semantische Analyse
  - Teilweise lernbar, für einzelne Sprachen verfügbar
  - Noch nicht sehr gut (60-75%)

Representativität

Aus Daten gelernte Modelle modellieren Eigenschaften des Korpus

- Korpora sind im Idealfall repräsentativ (balanciert):
  - Alle Genres, Sprachebenen, Gegenstandsbereiche (Domänen)
- Die meisten sind es nicht!
  - Balancierte Korpora: BNC, Brown-Korpus
  - Zeitungskorpora unbalanciert (TIGER, Penn Treebank)
  - Politische Korpora auch nicht
- Resultat: viele Modelle sind **korpusspezifisch**
  - Funktionieren schlechter auf anderen Daten
- Konkrete Beispiele:
  - Häufige syntaktische Strukture in NEGRA: „Wie X heute bekanntgab, ...“
  - Häufigstes Subjekt von „steigen“ in der Penn Treebank: **Aktien**

Größe von Korpora

- Wie groß **sollten** Korpora sein?
  - Groß für flachere sprachliche Ebenen
  - Größer für tiefere sprachliche Ebenen
    - Abbildung Text -- Annotation ist schwieriger

- Aber: tiefe Annotation sehr aufwendig
  - Roher Text: mehrere G Wörter verfügbar
  - POS-Tags: BNC (100M Wörter)
  - Syntax/Semantik: 1-10M Wörter

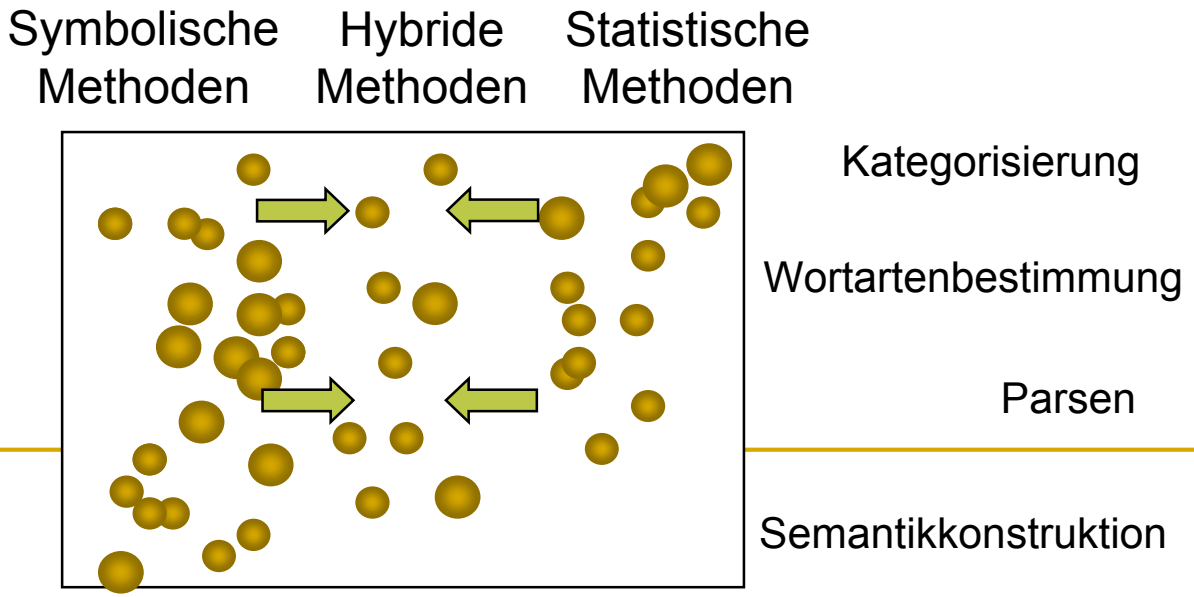
Es gibt nie genug Daten, um alles zu lernen

„The Web as Corpus“

- Vorschlag: Nutzen von **Internet-Daten**
  - Problem 1: Repräsentativität
    - Bedeutung von „amazon“
  - Problem 2: Automatische Annotation nötig
    - Korrektheit der Daten zweifelhaft
- Empirisches Ergebnis: Nutzen hängt von linguistischer Ebene ab
  - Flache Analyse: Zusatzdaten vorteilhaft trotz Fehlern
  - Tiefe Analyse: Fehler überwiegen Vorteil

Perspektive für tiefe Verarbeitung

- Symbolische Methoden: zu wenig flexibel
- Statistische Methoden: zu wenig Daten
- Idee: Hybride Systeme
  - Statistische **plus** symbolischer Verarbeitung
  - Empirik **plus** Rationalismus



Informationsmanagement

Sprachtechnologische Anwendungen

- Speech technology
  - Spracherkennung
  - Sprechererkennung
- Text technology
  - Sprachassistenz
  - Maschinelle Übersetzung
  - Informationsmanagement

Information Management - Was ist das?

Große Datenmengen
zugänglich und nutzbar machen

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

Probleme bei der Suche nach relevanter Information

- Daten oft in Textform, nicht strukturiert
  - Klassische KI-Verfahren zur Wissensrepraesentation nicht gut anwendbar
  - Computerlinguistische Verfahren noetig
- Immense Mengen an Daten
  - Viel „Rauschen“
  - Effiziente Verfahren noetig
- Die Bedeutung entspricht oft nicht direkt dem sprachlichen Material

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?“

Jede Frage hat ihre besonderen Schwierigkeiten

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?

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

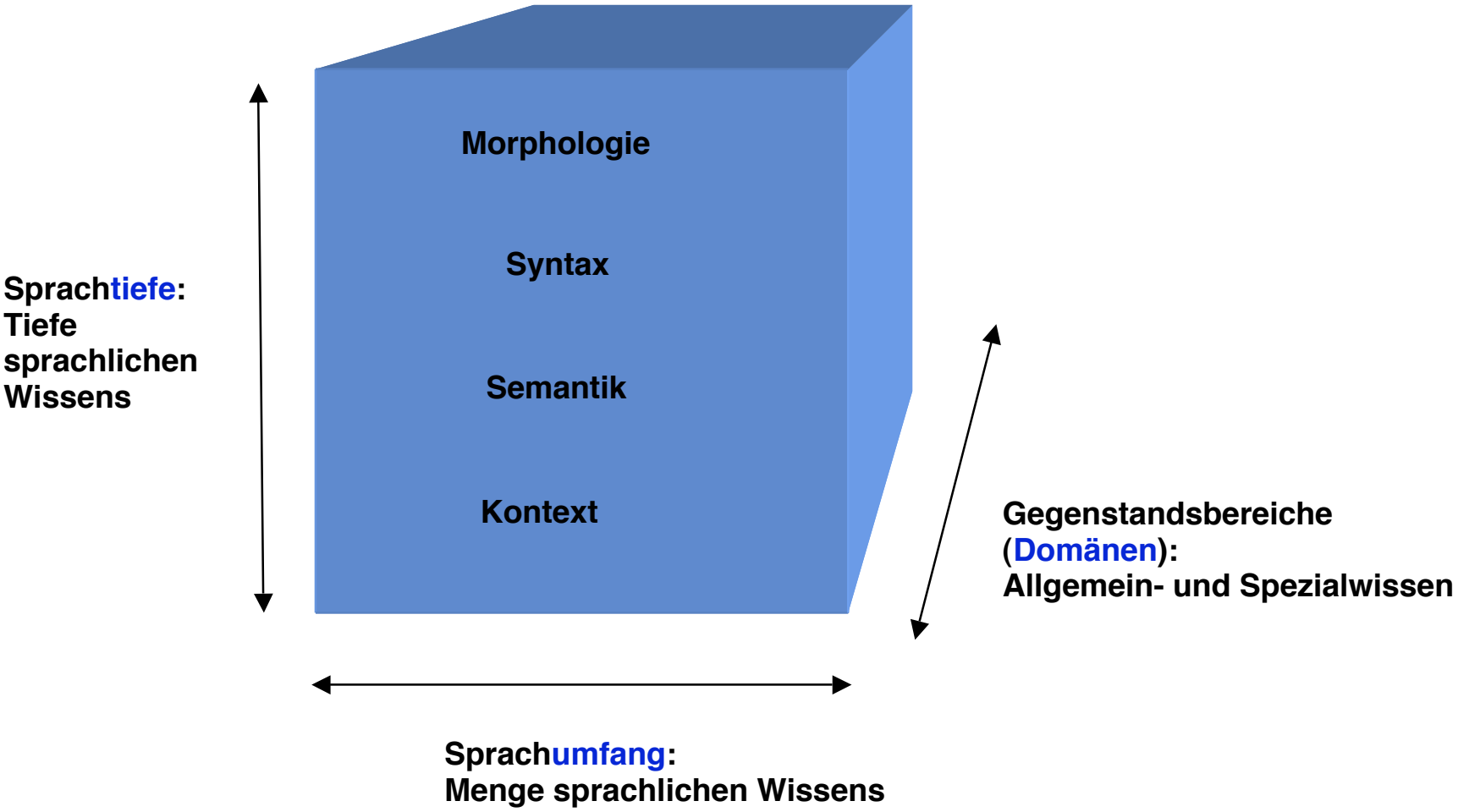
Die Bedeutung hängt vom linguistischem und extralinguistischen Kontext ab

- Linguistischer Kontext
  - D: „Der BP hat eine Amtszeit von vier Jahren. 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

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

Sprachverstehen



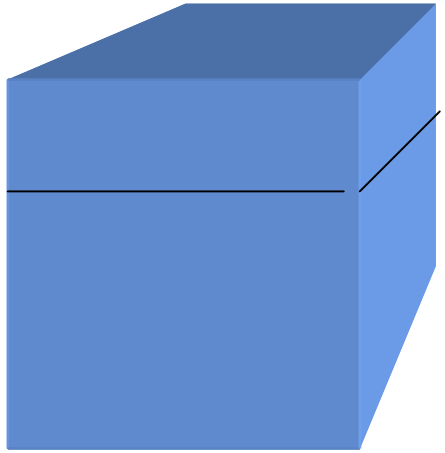
Sprachverstehen - Abstufungen

- 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 / Allgemeinwissen

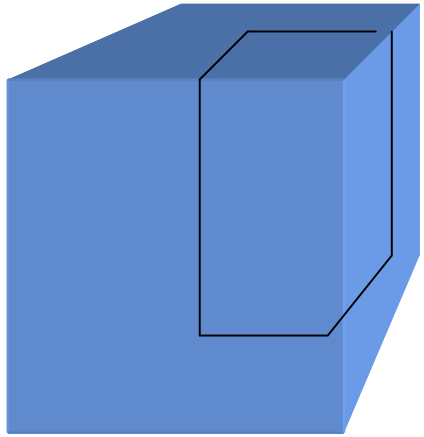
Je eingeschränkter die Aufgabe, desto machbarer

Volles Sprachverstehen zur Zeit nicht möglich

Erinnerung: TTS und CTS



TTS: voller Sprachumfang,
ziemlich viele Domänen,
kann also nicht tief sein



CTS: begrenzter Sprachumfang,
begrenzte Domänen,
kann daher tiefer sein

Erinnerung: Flache und tiefe Methoden

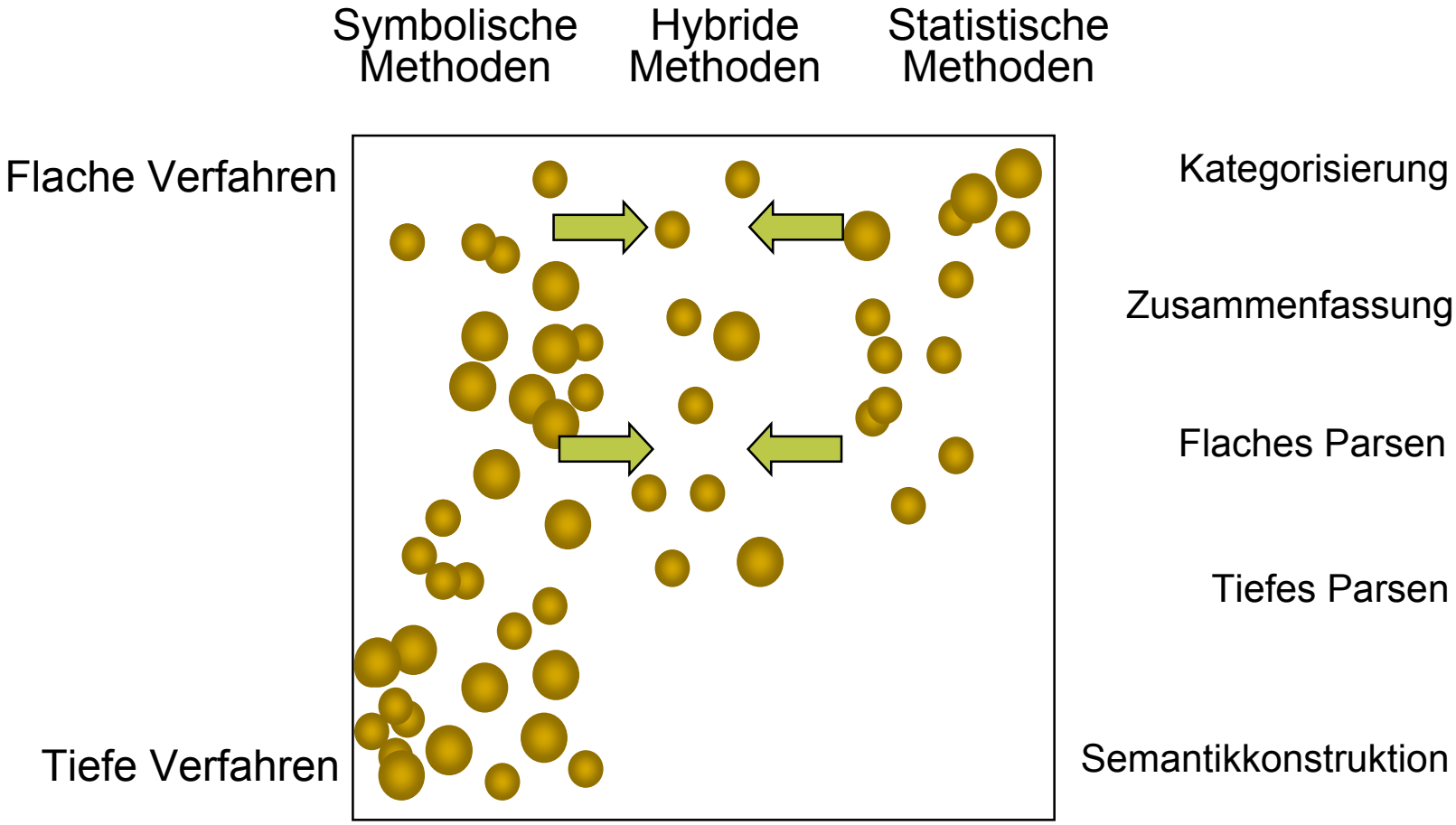
- Tiefe Methoden (deep processing - SHRDLU)
  - Vollständiges Verstehen
  - Vorteile: sicheres, informatives Resultat
  - Nachteile: Komplexität, störanfällig, **spezifisch**

- Flache Methoden (shallow processing - ELIZA)
  - 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
  - Entspricht oft introspektiven Modellen
- Statistische Methoden: analytische Mathematik
  - Statistische Klassifikation
  - Lernen von Funktionen
  - Entspricht oft empirischen Modellen

Methoden und Anwendungen



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 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

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“]].

- 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

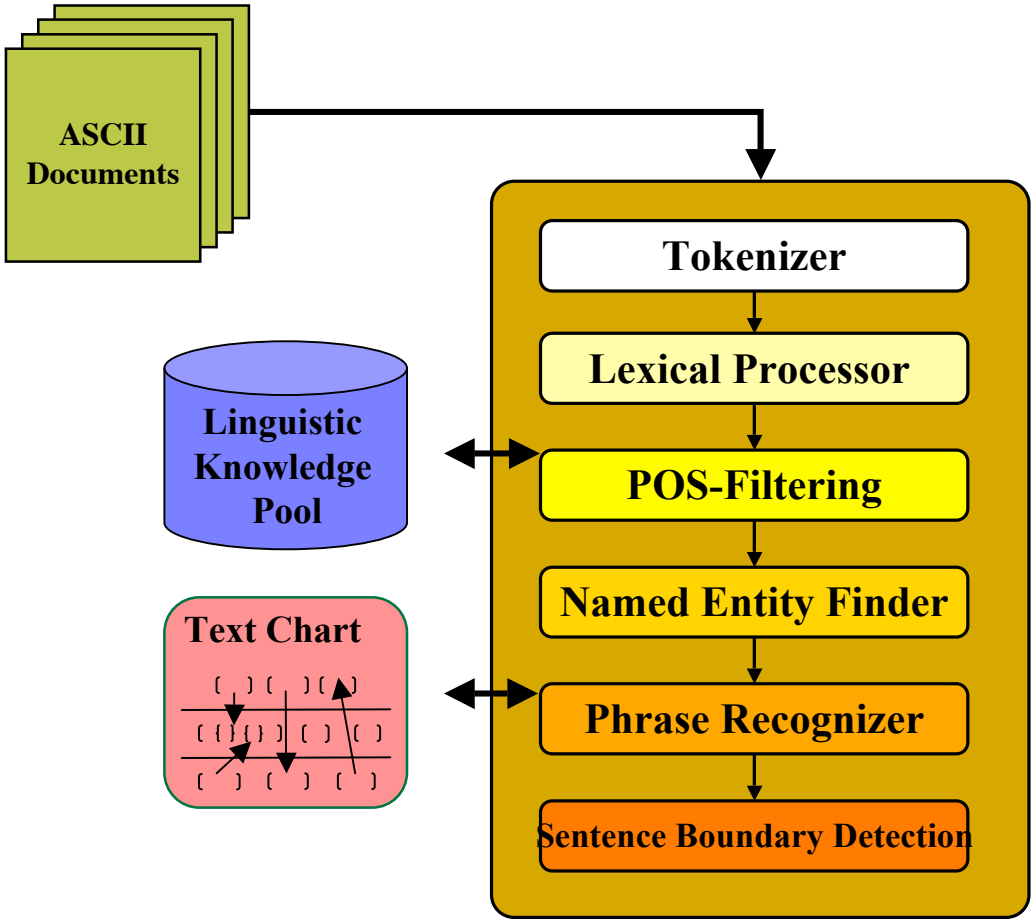
Zeit: 16:15

Datum: Donnerstag, den 13.01.2003

Ort: Seminarraum (Geb. 17.1)

Titel: „Putting Meaning into your Trees“

Beispiel: SPPC-System (DFKI)



Grafik © Feiyu Xu, Jakub Piskorski

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

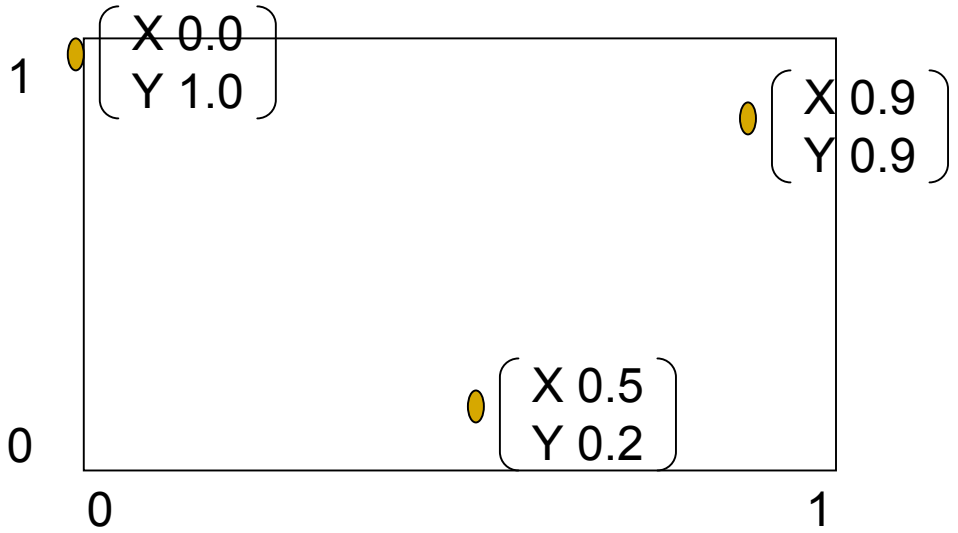
Information Retrieval

- Gegeben: Anfrage (Query)
- Gesucht: Relevante Dokumente

- Weitverbreitete Methode: Semantischer Raum
  - Jedes Dokument ist ein Punkt
  - Query ist auch ein Punkt
  - Nähe im semantischen Raum modelliert Relevanz

Punkte und Vektoren

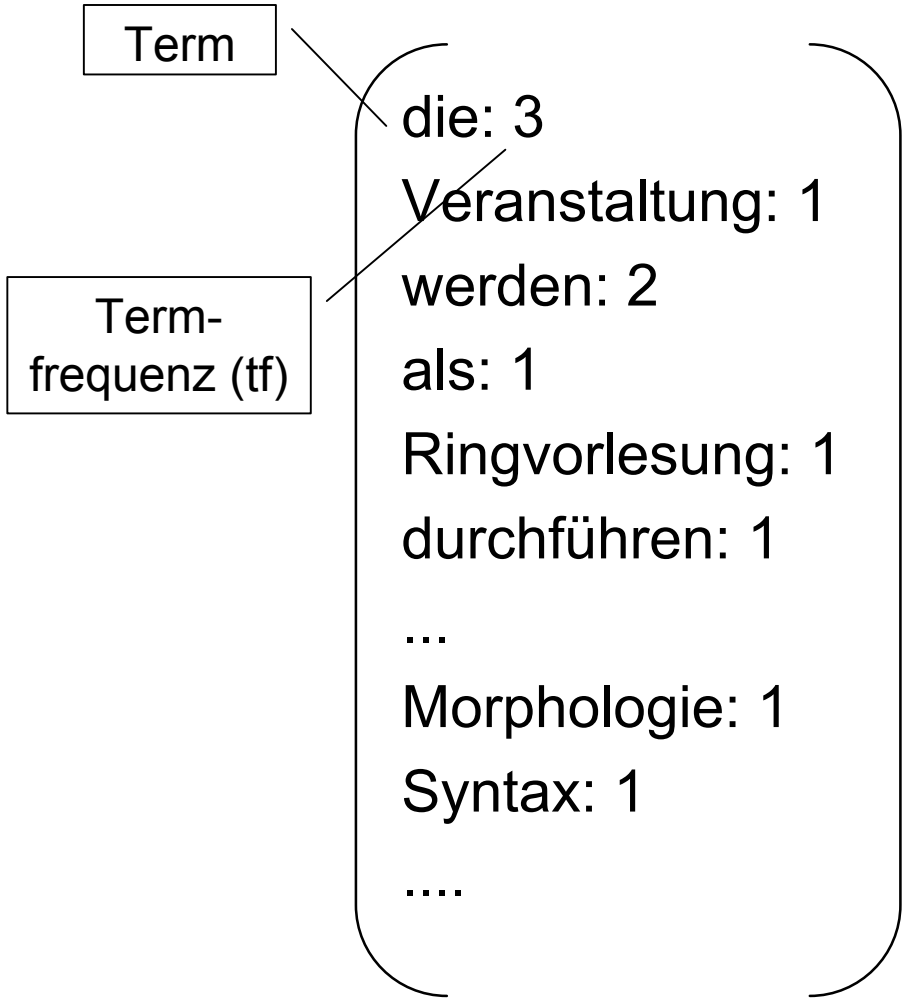
- Jeder Punkt kann als Vektor verstanden werden



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

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 .



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

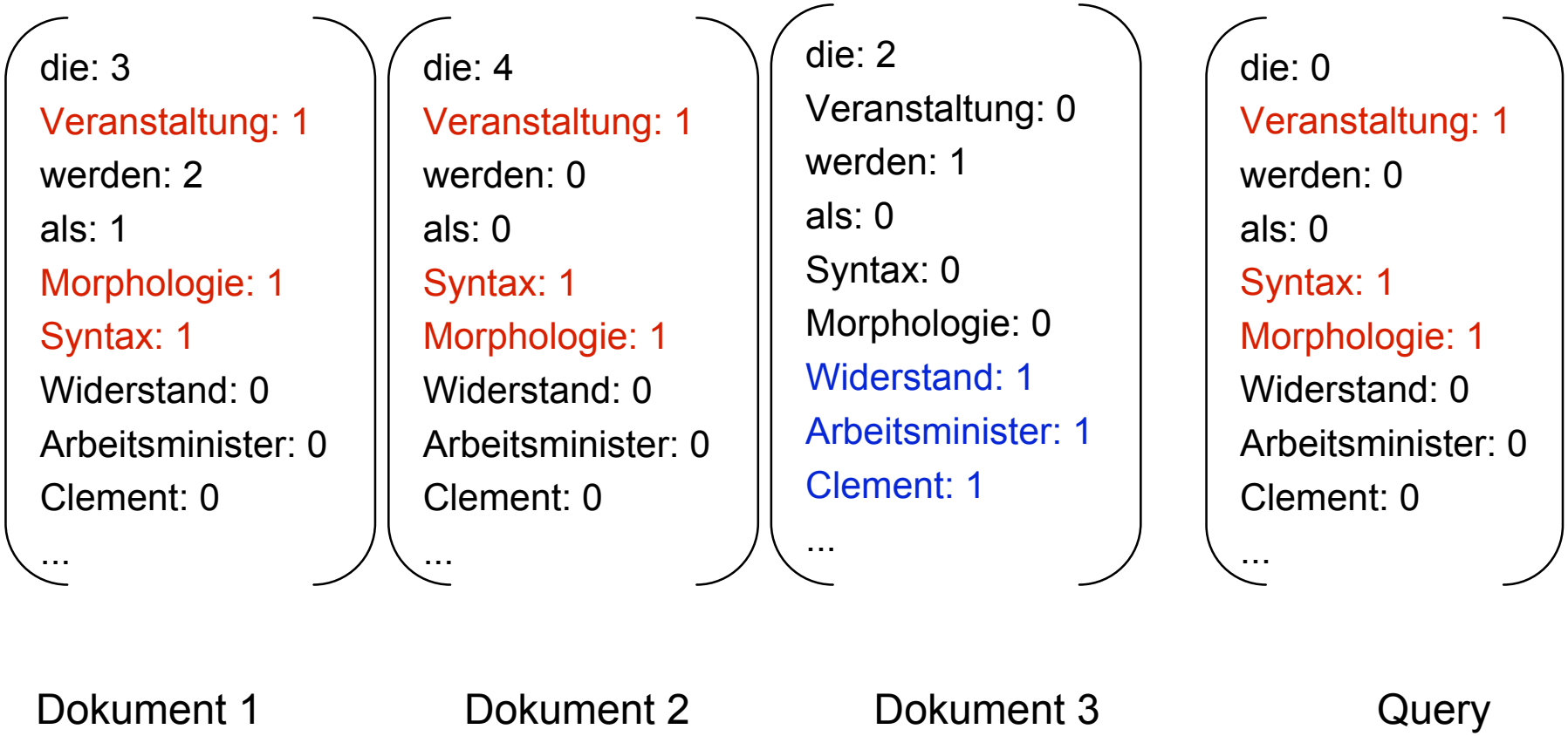
behandeln: 1

Morphologie: 1

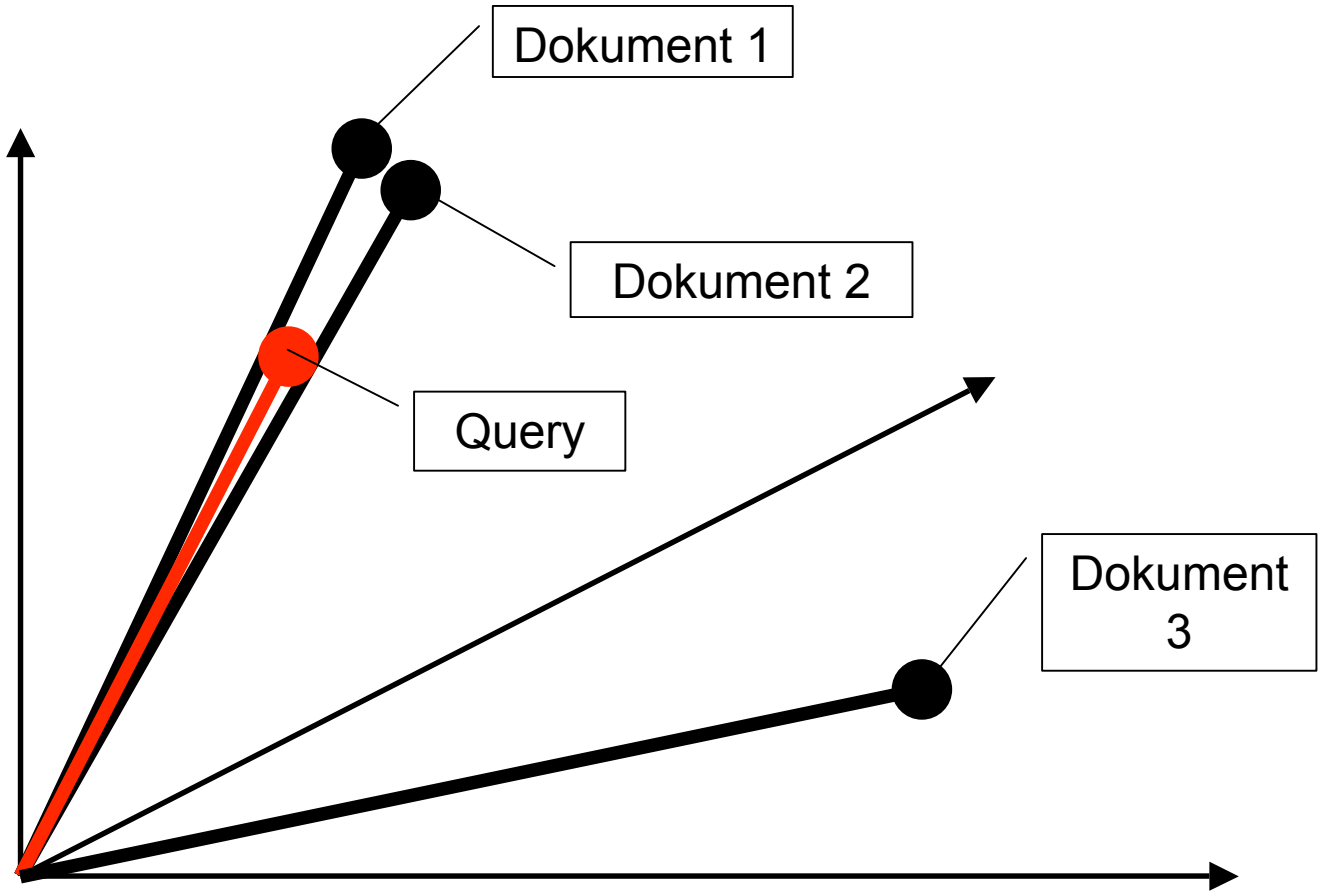
und: 1

Syntax: 1

Vektoren


die: 3
Veranstaltung: 1
werden: 2
als: 1
Morphologie: 1
Syntax: 1
Widerstand: 0
Arbeitsminister: 0
Clement: 0
...
die: 4
Veranstaltung: 1
werden: 0
als: 0
Syntax: 1
Morphologie: 1
Widerstand: 0
Arbeitsminister: 0
Clement: 0
...
die: 2
Veranstaltung: 0
werden: 1
als: 0
Syntax: 0
Morphologie: 0
Widerstand: 1
Arbeitsminister: 1
Clement: 1
...
die: 0
Veranstaltung: 1
werden: 0
als: 0
Syntax: 1
Morphologie: 1
Widerstand: 0
Arbeitsminister: 0
Clement: 0
...
Dokument 1
Dokument 2
Dokument 3
Query

Semantischer Raum (vereinfacht)



Relevante Dokumente für die Anfrage: Dokumente 1 und 2

Vorteile von semantischen Raeumen

- Ähnlich zu „naiver Suche“
  - Konzeptuell einfach, effizient
- Nutzung von Frequenzinformation
  - Dokumente sind aehnlich, wenn Begriffe gleichoft vorkommen
- Formalisierung
  - Mathematische Standardverfahren zur Berechnung von Aehnlichkeit / Relevanz
- Leichte 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“
      - **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 sprachlichem Wissen
  - Wenig Wissen nötig
  - Kann zur Optimierung des semantischen Raumes dienen
    - 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 tiefe(re) Verarbeitung nötig!

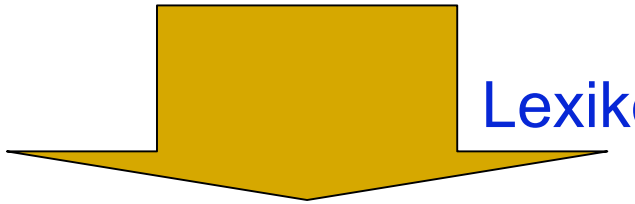
Welche Stellen sind relevant?

- 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)
    - `gehen: Bill Gates, College, ?<Ort>`
  - „Wo ging Bill Gates auf College?“ (Ort)
    - `sein: ?<Zahl> Sechsecke, auf Fußball`
- Strategie: Finde Aussage in Dokument, die die Luecke in der Anfrage fuellen kann

Beispiel 1

Auf einem Fußball befinden sich 20 Sechsecke

```
befinden: 20 Sechsecke, auf Fußball
```



Lexikon: Synonymie

```
sein: 20 Sechsecke, auf Fußball
```

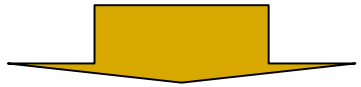
```
sein: ?<Zahl> Sechsecke, auf Fußball (Anfrage)
```

```
Antwort: 20
```

Beispiel 2

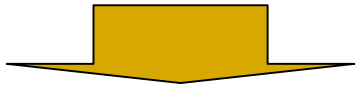
Bill Gates, einst Harvard-Abbrecher, ist heute einer der reichsten Männer Amerikas.

```
Abbrecher: Bill Gates, Harvard
```



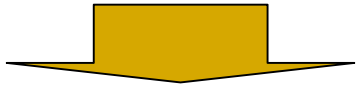
Lexikon / Grammatik: (De-)Nominalisierung

```
abbrechen: Bill Gates, Harvard
```



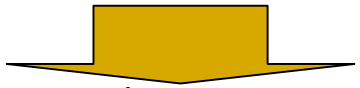
Weltwissen: Harvard ist eine Universität und ein Ort

```
abbrechen: Bill Gates, Universität, in Harvard
```



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

```
gehen: Bill Gates, Universität, in Harvard
```



Lexikon: Universität ist Synonym zu College

```
gehen: Bill Gates, College, in Harvard
gehen: Bill Gates, College, ?<Ort> (Anfrage)
```

Antwort: in Harvard

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)