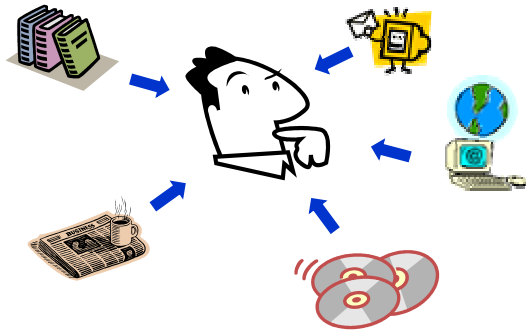
Informationsextraktion aus Textdokumenten: Herausforderung, Technologie, Anwendungen

Dr. Roland Stuckardt
IT-Beratung ● Sprachtechnologie
Frankfurt am Main

roland@stuckardt.de
www.stuckardt.de

Veranstaltungsreihe des VDI – Frankfurt/Darmstadt, AK Kommunikationstechnik

Das informationssuchende Individuum im 21. Jahrhundert



Überflutung mit **digitalen Texten**

- Die Wissensproduktion nimmt stetig zu
- Der textuellen Informationsvermittlung kommt nach wie vor zentrale Bedeutung zu.
- Viele Texte liegen (auch) als Computerdatei bzw. online vor:
  - Internet: WWW, E-Mail, News
  - CD-ROM-Ausgaben klassischer Printmedien
  - digitale Bibliotheken
  - …

Software-Lösung für Informationssuchende

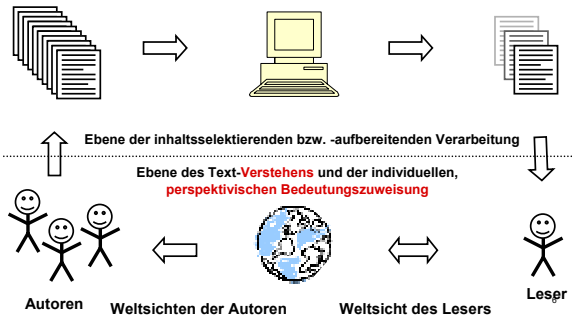
- informationssuchende Individuen verlangen nach
  - Informationen, die auf Ihre persönlichen Bedürfnisse zugeschnitten sind
  - effektivem Schutz vor informationeller Überflutung

… und somit nach Software-Lösungen zur inhaltsorientierten Erschließung von Texten

Gliederung

1) Der Mensch als Informationssuchendes Individuum:
   - Herausforderung: Software-Lösungen zur Bewältigung der (textuellen) Informationsflut
2) Basisszenarien der inhaltsorientierten Texterschließung:
   - Text Retrieval
   - Textkategorisierung
   - Textzusammenfassung
   - Informationsextraktion
3) Informationsextraktion: Basisszenario und Basistechnologie
   - Aufgabenspezifikation und Gütekriterien
   - Leistungsfähigkeit aktueller Technologie
   - Software-Architektur und Inhaltserschließungstechniken
   - Beispiel: das FASTUS-System
4) Anwendungsszenario: Telefonüberwachung
   - IE-Technologie im Workflow einer forensischen Anwendung

Rolle von Software-Lösungen für die inhaltsorientierte Erschließung von Texten



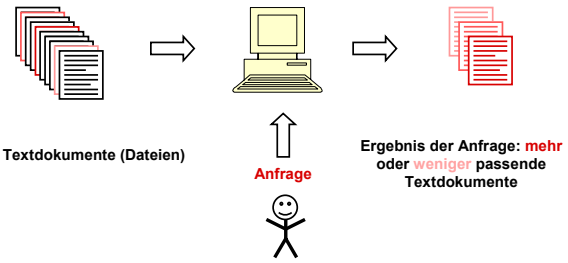
Rolle von Software-Lösungen für die inhaltsorientierte Erschließung von Texten

- **Computer** selektiert und kategorisiert Texte bzw. strukturiert Textinhalte auf der Grundlage bedeutungserhaltender Operationen
- **Mensch** versteht Texte bzw. aufbereitete Inhalte vor dem Hintergrund seines perspektivischen Weltbezugs und weist individuell Bedeutung zu

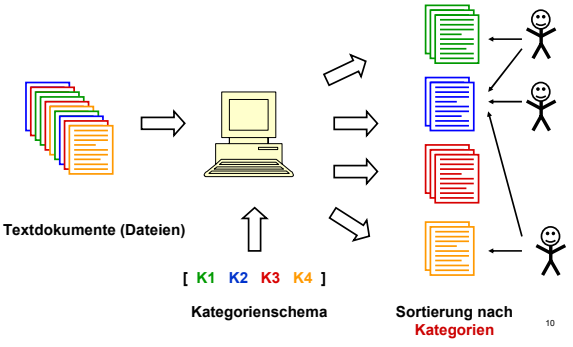
Basisszenarien vs. Anwendungsszenarien

- **Basisszenarien:**
  - Finden relevanter Texte in der Informationsflut
  - Ordnung von Texten gemäß Relevanzgrad
  - Kategorisierung von Texten nach inhaltlicher Ähnlichkeit
  - Zusammenfassung von Texten
  - Strukturierung von Textinhalten
- **Anwendungsszenarien**:
  - Unterstützung von Workflows
  - Verfügbarkeit geeigneter Benutzeroberflächen
  - Abdeckung der jeweils relevanten Textdateiformate

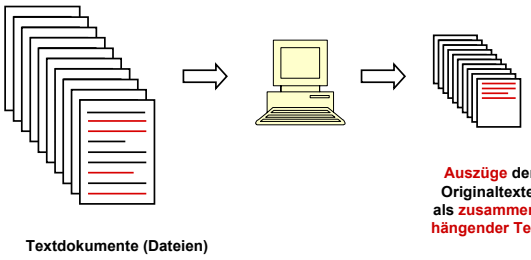
Basisszenario: Text Retrieval



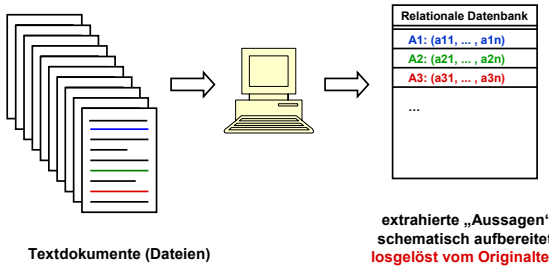
Basisszenario: Textkategorisierung



Basisszenario: Textzusammenfassung



Basisszenario: Informationsextraktion



Basisszenarien – Resümee:

	Ergebnis	Techniken	Status
Text Retrieval	relevanzsortierte Auswahl an Originaltexten	Analyse auf Wortebene, Statistik	Anwendungs-reife
Textkate-gorisierung	Originaltexte kategorisiert	Wortebene, Statistik, Neuronetze	Anwendungs-reife
Textzusam-menfassung	Kernaussagen als zusammen-hängender Text	inhaltliche Tiefenanalyse	Forschungs-gegenstand
Informations-extraktion	Kernaussagen schematisch	**beschränkte** Analyse v. Satz- u.Textbedeutung	Anwendungs-reife

Fokus auf: Informationsextraktion

Behrens, Peter, *1868 in Hamburg, +1940 in Berlin. Behrens entwickelte als einer der ersten Architekten des 20. Jahrhunderts eine architektonische Konzeption, die den Anforderungen der industrialisierten Zivilisation gerecht wurde - zu einer Zeit, in der die Gesellschaft noch in archaischen Vorstellungen dachte, gleichzeitig aber blind auf die überwältigenden Fortschritte der Technik vertraute. Behrens stand am Beginn der modernen Architektur in Deutschland, auf die er zwischen 1900 und 1914 einen entscheidenden Einfluss ausübte.

Ziel: Extraktion von Aussagen über künstlerische Aktivitäten, Schema:

[Künstler (K), erschafft (S), Objekt (O)]

Relationaler Charakter:

entwickeln(Peter Behrens, architektonische Konzeption)

Warum das Problem schwierig ist

- sprachliche Ausdrucksvielfalt

Behrens errichtete die Turbinenhalle der AEG

Die Turbinenhalle der AEG wurde von Behrens errichtet

Die Errichtung der Turbinenhalle durch Behrens erfolgte 1911.

Die Turbinenhalle steht in Berlin. Behrens errichtete >sie< 1911.

Eingrenzung der Aufgabenstellung „Textuelle Informationsextraktion“

Message Understanding Competitions (MUC), US - Department of Defense / DARPA, 1989-98:

- Definition der zu extrahierenden Inhalte (sog. **Tasks**)
- Definition entsprechender **Gütekriterien**, um Systeme **evaluieren** und vergleichen zu können
- durch den Menschen erstellte **Referenzdaten** dienen als Vergleichsmaßstab

Named Entity Task

Behrens, Peter, *1868 in Hamburg, +1940 in Berlin. Behrens entwickelte als einer der ersten Architekten des 20. Jahrhunderts eine architektonische Konzeption, die den Anforderungen der industrialisierten Zivilisation gerecht wurde - zu einer Zeit, in der die Gesellschaft noch in archaischen Vorstellungen dachte, gleichzeitig aber blind auf die überwältigenden Fortschritte der Technik vertraute. Behrens stand am Beginn der modernen Architektur in Deutschland, auf die er zwischen 1900 und 1914 einen entscheidenden Einfluss ausübte.

Ergebnis:

```
Behrens, Peter  1, 14
Hamburg        27, 32
Berlin         44, 49
Behrens        51, 57
Behrens       399, 405
```


Scenario Template Task
= Kernaufgabe der Informationsextraktion

Behrens, Peter, *1868 in Hamburg, +1940 in Berlin. Behrens entwickelte als einer der ersten Architekten des 20. Jahrhunderts eine architektonische Konzeption, die den Anforderungen der industrialisierten Zivilisation gerecht wurde - zu einer Zeit, in der die Gesellschaft noch in archaischen Vorstellungen dachte, gleichzeitig aber blind auf die überwältigenden Fortschritte der Technik vertraute. Behrens stand am Beginn der modernen Architektur in Deutschland, auf die er zwischen 1900 und 1914 einen entscheidenden Einfluss ausübte.

Ergebnis:

```
[ Schaffens-Akt: entwickeln, 60, 70
  Künstler: Behrens, 51, 57
  Gegenstand: architekton. Konzeption, 131, 157 ]

[ Schaffens-Akt: Einfluss ausüben, 519, 534
  Künstler: Behrens, 399, 405
  Gegenstand: moderne Architektur..., 427, 461
```

Coreference Task

Behrens, Peter, *1868 in Hamburg, +1940 in Berlin. Behrens entwickelte als einer der ersten Architekten des 20. Jahrhunderts eine architektonische Konzeption, die den Anforderungen der industrialisierten Zivilisation gerecht wurde - zu einer Zeit, in der die Gesellschaft noch in archaischen Vorstellungen dachte, gleichzeitig aber blind auf die überwältigenden Fortschritte der Technik vertraute. Behrens stand am Beginn der modernen Architektur in Deutschland, auf die er zwischen 1900 und 1914 einen entscheidenden Einfluss ausübte.

Ergebnis: Koreferenz-Klassen

Behrens, Peter
Behrens
e. d. ersten Architekten ...
Behrens
er

architekton. Konzeption
die

Zeit
die

moderne Architektur in ...
die

Gütekriterien

Die Leistung der Softwaresysteme wird in Bezug auf **Qualität** und **Quantität** evaluiert:

$$Precision = \frac{\#(System \cap Referenz)}{\#System}$$

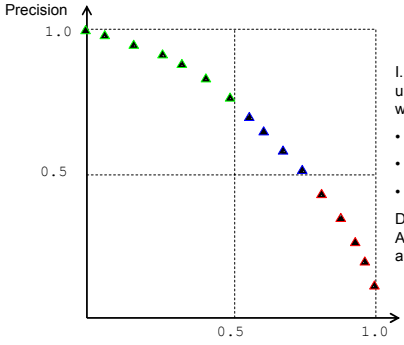
$$Recall = \frac{\#(System \cap Referenz)}{\#Referenz}$$

Qualität, Quantität für Named Entity Task

System		Referenz	
Turbinenhalle	30,43		
AEG	52,55	AEG	52,55
Behrens	71,78	Peter Behrens	65,78
		Berlin	96,103
		Hennigsdorf	114,125

Precision = 2/3 *Recall = 2/4*

Austauschverhältnis Qualität, Quantität

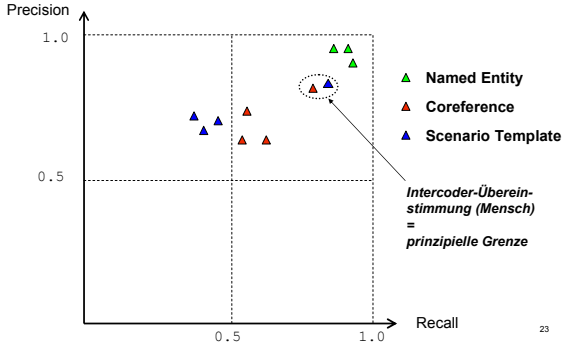


I. d. R. können IE-Systeme unterschiedlich konfiguriert werden:

- Qualität vor Quantität
- Quantität vor Qualität
- „guter Kompromiss“

Die Anforderungen des Anwendungsszenrios sind ausschlaggebend.

Ergebnisse aktueller Technologie



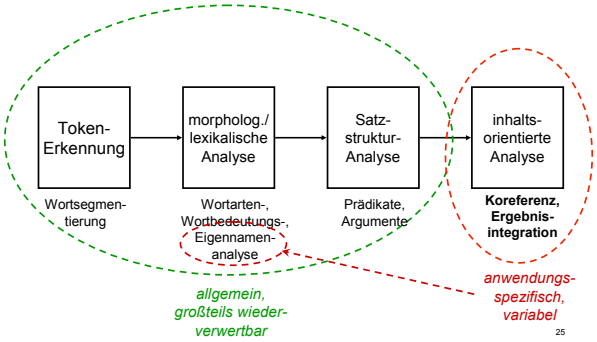
Scenario Template Task, MUC-3:
Terroraktivitäten in Lateinamerika

BOGOTA, 7 SEP 89 (INRAVISION TELEVISION CADENA 1) -- [REPORT] MARIBEL OSORIO] [TEXT] MEDELLIN CONTINUES TO LIVE THROUGH A WAVE OF TERROR. FOLLOWING LAST NIGHT'S ATTACK ON A BANK, WHICH CAUSED A LOT OF DAMAGE, A LOAD OF DYNAMITE WAS HURLED AGAINST A POLICE STATION. FORTUNATELY NO ONE WAS HURT. HOWEVER, AT APPROXIMATELY 1700 TODAY A BOMB EXPLODED INSIDE A FAST-FOOD RESTAURANT. A MEDIUM-SIZED BOMB EXPLODED SHORTLY BEFORE 1700 AT THE PRESTO INSTALLATIONS LOCATED ON [WORDS INDISTINCT] AND PLAYA AVENUE. [...]

... sehr komplex, bis zu 18 (!) Attribute pro Vorfall – u. a.:

DATE OF INCIDENT
TYPE OF INCIDENT
INSTRUMENT: TYPE(S)
LOCATION OF INCIDENT
PERPETRATOR: ID OF INDIV(S)
PERPETRATOR: ID OF ORG(S)

Architektur von IE-Systemen



Morphologische Analyse

Künste	**->**	**Kunst**
Glasmalerei	**->**	**Glas+Malerei**
eingeführt	**->**	**einführen**
führte … ein	**->**	**einführen**

- fürs Deutsche: **schwer**
  - zusammengesetzte Nomen
  - deshalb jedoch gerade essenziell
  - Softwarelösung: GerTwoL, LingSoft Helsinki
- fürs Englische: **leicht**
  - vollständige Auflistung aller Formen machbar

Wortarten, Wortbedeutungen

- in Fällen von Mehrdeutigkeit ist der **Kontext** entscheidend:

schönen — **Wortart:** *Verb* **oder** *Adjektiv*?

```
An den Zahlen gibt es nichts zu schönen.
Am Morgen danach hatte er einen schönen Kater.
```

Kater — **Bedeutung:** *Tier* **oder** *Brummschädel*?

```
Ein schwarzer Kater kreuzte Stefans Weg.
Am Morgen danach hatte er einen schönen Kater.
```


Statistische Wortart-Disambiguierung: N-Gram-Modell des Kontexts

- z. B. N=3, Gruppen von 3 Wörtern anschauen (Trigramme):

```
An den Zahlen gibt es nichts zu schönen.
```

Wsk(„schönen$_V$“ nach „nichts zu“) = 0.03
Wsk(„schönen$_{Adj}$“ nach „nichts zu“) = 0.0005

```
Am Morgen danach hatte er einen schönen Kater.
```

Wsk(„schönen$_V$“ nach „er einen“) = 0.00001
Wsk(„schönen$_{Adj}$“ nach „er einen“) = 0.05

- Grundgedanke: Anwendung **partieller** Kontextinformation
  betrachtetes Wort: w_k
  bisher gelesener Text: $T = w_1\, w_2 \ldots w_{k-1}$ **= Kontext**
  es werden aber nur die vorherigen N-1 Wörter betrachtet
- Kontextwahrscheinlichkeiten lassen sich per statistischer Analyse großer Textsammlungen **automatisch** ermitteln

Eigennamenerkennung

- wichtiger Bestandteil der **anwendungsspezifischen** Analyse
- vollständige Auflistung oft unmöglich
- **Beispiel: Firmennamen** in Wirtschaftstexten
  - Neugründungen
  - Fusionen
  - Änderung der Rechtsform

- **Erkennungsregeln:**

```
Firma := Token+ { „AG“ | „GmbH“ | „KG“ | … | „OHG“ }
```

- nicht immer derart einfach – **Produktnamen**:
  „*I can't believe it's not butter*“ (amerikan. Margarine)

Analyse der Satzstruktur

- = **syntaktische** Analyse

```
Peter Behrens, der wie Gropius ein berühmter Architekt
der Moderne war, errichtete die Turbinenhalle der AEG.

Die Turbinenhalle der AEG wurde von dem berühmten
Architekten 1911 erbaut.
```

- erkennt
  - eingeschobenen Relativsatz
  - Aktiv- und Passivsätze
  - …
- bringt **Prädikate** (Verben) und zugehörige **Argumente** (Subjekt, Objekte, …) zusammen

Koreferenz- bzw. Pronomenresolution

- Argumente können **pronominal** realisiert sein:

```
errichten ( Subjekt: er,
            Objekt: Turbinenhalle der AEG)
```

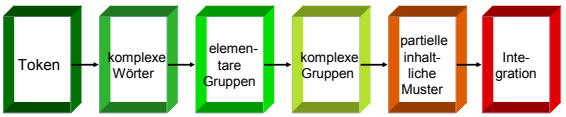
- Ziel: Ermittlung informativerer Ausdrücke, die dasselbe Objekt bezeichnen (**Koreferenz**)

```
errichten ( Subjekt: Peter Behrens,
            Objekt: Turbinenhalle der AEG)
```

- **Wie?** *Hierzu später mehr …*

Das FASTUS-System

- Kaskade sog. **Endlicher Automaten**:
  - einheitliches Analysemodell für (fast) alle Phasen
  - seichte Satzstruktur-Analyse
  - schnell, robust, relativ leicht zu pflegen und zu portieren
- **Sechs Phasen**:



Die Phasen von FASTUS

- Beispiel: Analyse von **Wirtschaftsmeldungen**

```
Bridgestone Sports Co. said Friday it has set up a joint
venture in Taiwan with a local concern and a Japanese trading
house to produce golf clubs to be shipped to Japan. The joint
venture, Bridgestone Sports Taiwan Co., capitalized at 20
million new Taiwan dollars, will start production in January
1990 with production of 20,000 iron and metal wood clubs a
month.
```

- Ziel: Aussagen über **Joint Ventures** extrahieren, Attribute
  - *beteiligte Firmen*
  - *Joint-Venture-Firma, Eigenkapital*
  - *Produkt(e)*
  - *ab wann?*

Die Phasen von FASTUS

- Phase 1: Erkennung von Token

```
<Bridgestone> <Sports> <Co> <.> <said> <Friday> <it> <has>
<set> <up> <a> <joint> <venture> <in> <Taiwan> <with> <a>
<local> <concern> <and> <a> <Japanese> <trading> <house> <to>
<produce> <golf> <clubs> <to> <be> <shipped> <to> <Japan> <.>
```

- Phase 2: „Multiwörter“, Namen, Zahlen

```
<Bridgestone> <Sports> <Co> <.> <said> <Friday> <it> <has>
<set> <up> <a> <joint> <venture> <in> <Taiwan> <with> <a>
<local> <concern> <and> <a> <Japanese> <trading> <house> <to>
<produce> <golf> <clubs> <to> <be> <shipped> <to> <Japan> <.>
```


Die Phasen von FASTUS

- Phase 3: elementare Gruppen

```
<Bridgestone> <Sports> <Co> <.> <said> <Friday> <it> <has> <set>
<up> <a> <joint> <venture> <in> <Taiwan> <with> <a> <local>
<concern> <and> <a> <Japanese> <trading> <house> <to> <produce>
<golf> <clubs> <to> <be> <shipped> <to> <Japan> <.>
```

Unternehmen
Ortsangaben
Verbgruppen
Nominalgruppen

Präpositionen
Konjunktionen

Die Phasen von FASTUS

- Erkennung von Adjektiv-Gruppen / -Sequenzen

Grammatikregeln als *reguläre Ausdrücke:*

```
AdjP -> Ordinal
        | [ { Q-er | Q-est } ] { Adj | Vparticle }+
        | N<sing,!Time-NP> - Vparticle
        | Number [-] { month | day | year } [-] old

Adjs -> AdjP [ { , | [,] Conj } { AdjP | Vpart } ]*
```

Beispiele:

```
<fifth> <Symphony>
<most> <hungry> <guys>
<five> <-> <year> <-> <old> <boy>
<big> <,> <bad>, <and> <ugly> <car>
```

Die Phasen von FASTUS

- Phase 4: Erkennung komplexer Nominal- und Verbgruppen

z.B. koordinierte Nominalgruppen:

… <The> <joint> <venture> <,> <Bridgestone> <Sports> <Taiwan> <Co> <.> <,> <capitalized> <at> <20> <million> <new> <Taiwan> <dollars>, <will> <start> <production> <in> <January> <1990> <with> <production> <of> <20,000> <iron> <and> <metal> <wood> <clubs> <a> <month> <.>

Die Phasen von FASTUS

- Phase 5: partielle inhaltliche Muster

(1)	Relationship:	TIE-UP
	Entities:	„Bridgestone Sports Co." „a local concern" „a Japanese trading house"
(2)	Activity:	PRODUCTION
	Product:	„golf clubs"
(3)	Relationship:	TIE-UP
	JV Company:	„Bridgestone Sports Taiwan Co."
	Amount:	NT$20000000
(4)	Activity:	PRODUCTION
	Company:	„Bridgestone Sports Taiwan Co."
	Start Date:	DURING: January 1990
(5)	Activity:	PRODUCTION
	Product:	„iron and metal wood clubs"

Die Phasen von FASTUS

- Phase 6: Integration in IE-Zielstruktur

TIE-UP-1:	
Relationship:	TIE-UP
Entities:	„Bridgestone Sports Co." „a local concern" „a Japanese trading house"
JV Company:	„Bridgestone Sports Taiwan Co."
Activity:	ACTIVITY-1
Amount:	NT$20000000
ACTIVITY-1:	
Company:	„Bridgestone Sports Taiwan Co."
Product:	„iron and metal wood clubs"
Start Date:	DURING: January 1990

Koreferenz- bzw. Pronomenresolution

Behrens ist berühmt. Er baute die Turbinenhalle. Der Architekt ...

- Facetten:
  - Ermittlung der Klassen **koreferenter** Ausdrücke
  - IE-Ziel: Ermittlung möglichst **informativer** Ausdrücke
  - aus algorithmischer Sicht: zwei Schritte
    - Identifikation von Ausdrücken, die Objekte referenzieren (<Behrens>, <er>, <die Turbinenhalle>, <der Architekt>)
    - Identifikation koreferenter **Antezedens**-Ausdrücke (insbesondere für Pronomen)
      (<Behrens> ← <er>, <Behrens> ← <der Architekt>)

Strategien zur Koreferenzresolution

- Ermittlung koreferenter Antezedenten für **Pronomen**:
  - Restriktion: Kongruenz in Numerus und Genus

    Behrens sagte seiner Kollegin, dass er sie schätze.

  - Restriktion: syntaktisch-konfigurationale Zugänglichkeit

    Behrens verlangt, dass der Friseur sich rasiert.
    Behrens verlangt, dass der Friseur ihn rasiert.

  - Präferenzkriterium: Bevorzugung des Subjekts

    Der Friseur betrat den Salon. Er bediente den Kunden.

  - …

Anwendungsfall: Telefonüberwachung

Wie können aufgezeichnete Gespräche sinnvoll inhaltlich exploriert werden?

- Es liegen zwei Kategorien von Informationen vor
  - **Strukturierte Daten**: beteiligte Anschlüsse/ Teilnehmer, Datum/Uhrzeit, Gesprächsdauer, Provider
  - **Audiodaten**: das eigentliche Gespräch
- sehr große Menge an Information – über 1000 Gespräche pro Maßnahme sind die Regel

Audiodaten: Analyse problematisch

Fast immer in fremden Sprachen

- **mehrsprachiges** Anwendungsszenario
- **Verschriftung durch Übersetzer** zwingend

- Auch verschriftete Gespräche sind noch nicht ausreichend strukturiert

⇒ **Informationsextraktion**

Ziel: Strukturierte Repräsentation der Gesprächstexte

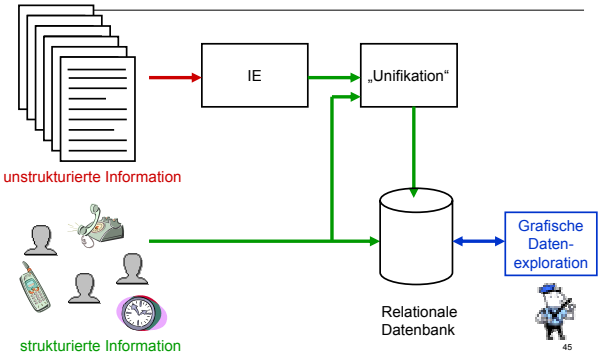
Identifizierung bestimmter Klassen besonders relevanter Objekte:

- Personen, Firmen, Orte, Datum, Waren, Transaktionen etc

Überführung erkannter Objekte in eine relationale Datenbank:

- **Zusammenführung mit strukturierten Daten**
- einheitliches grafisches Explorationswerkzeug für beide Informationskategorien

Workflow der IE-Anwendung



Zusammenfassung

- Informationsextraktion stellt eine zentrale Herausforderung im Zeitalter der digitalen Informationsflut dar
- flexible Software-Basistechnologie - zuschneidbar auf unterschiedliche Anwendungsszenarien - ist verfügbar
- Zur Entwicklung von Anwendungen textueller Informations-extraktion bedarf es Expertenwissen in den Gebieten
  - Informatik / Software Engineering
  - Linguistik
  - Computerlinguistik / Natural Language Engineering
- Projektgeschäft – (noch) kein fertiges Produkt verfügbar
- Kompetente Beratung in der Analyse potenzieller Anwen-dungsszenarien, in der Auswahl geeigneter Technologien und in der Projektdurchführung ist essenziell!

Folien mit ergänzender Information

Basisszenarien: unterscheidbar hinsichtlich

- **Input**: Dokumentenbestand typischerweise
  - statisch (weitgehend unveränderlich)?
  - dynamisch (fortwährend neue Dokumente)?
- **Output**: Welche Ergebnisse erhält der Nutzer?
  - Originaltexte?
  - Auszüge von Originaltexten?
  - strukturiert aufbereitete „Inhalte“?
- **Techniken** zur Inhaltserschließung?
- **Beispiele** möglicher Anwendungsszenarien?

Basisszenario: Text Retrieval

- **Input**: sehr großer Bestand an Textdokumenten, großteils statisch
- **Output**: Auswahl an relevanten Textdokumenten, geordnet nach Relevanz
- **Anwender**: erhält vollständige Texte
- **Techniken**: Schlüsselwortsuche, statistische Häufigkeitsanalysen, Synonymlexika
- **Anwendungen:** Internet-Suchmaschinen, Suchwerkzeuge für CD-ROM-Textarchive

Basisszenario: Textkategorisierung

- **Input:** große Anzahl von Textdokumenten, typischerweise dynamisch
- **Output:** Originaldokumente, geordnet nach Kategorien
- **Anwender**: erhält vollständige Texte
- **Techniken**: wie Information Retrieval; auch Neuronale Netze
- Anwendungen: Spamfilter;
  Verteiler für E-Mails, Pressemeldungen; Wissensmanagement in Unternehmen

Basisszenario: Textzusammenfassung

- **Input**: als relevant bekannte, oft umfangreiche Textdokumente; statisch oder dynamisch
- **Output**: Kernaussagen der Originaldokumente als zusammenhängender Text
- **Anwender:** erhält Ausschnitte der Texte
- **Techniken:** inhaltliche Tiefenanalyse
- **Anwendungen:** nachgeschalteter Schritt in Text-Retrieval-Anwendungen, E-Mail-Verteilern

Basisszenario: Informationsextraktion

- **Input:** größerer Bestand an Textdokumenten, statisch oder dynamisch
- **Output**: inhaltlich relevante „Aussagen“, schematisch aufbereitet (keine Textform)
- **Anwender**: erhält schematisch dargestellte Text-inhalte (zunächst) losgelöst vom Originaltext
- **Techniken**: Analyse v. Satzstruktur u.Textbedeutung Erschließung relationaler inhaltlicher Strukturen
- **Anwendungen:** E-Mail-Antwortmanagement (CRM), forensische Analyse von Telefonatverschriftungen

Methodologische Probleme der Entwicklung von IE-Software

- erschließungsrelevante **“Aussagen“**?
- **unklare Erfolgskriterien**
- **mangelnde Vergleichbarkeit** unterschiedlicher Verfahren

- bis in die 90er Jahre hinein wenig Fortschritt

Stochastische Wortart-Disambiguierung: N-Gram-Modell des Kontexts

- z. B. N=3, Gruppen von 3 Wörtern anschauen (Trigramme):

  `An den Zahlen gibt es nichts zu schönen.`

  $P(schönen_V \mid nichts\ zu) = 0.03$
  $P(schönen_{Adj} \mid nichts\ zu) = 0.0005$

  `Am Morgen danach hatte er einen schönen Kater.`

  $P(schönen_V \mid er\ einen) = 0.00001$
  $P(schönen_{Adj} \mid er\ einen) = 0.05$

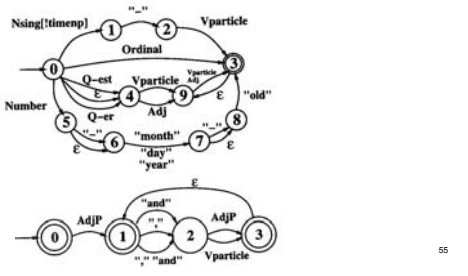
- Grundgedanke: Anwendung **partieller** Kontextinformation
  aktuell betrachtetes Wort: w_k
  bisher gelesener Text: $T = w_1 w_2 \ldots w_{k-1}$ **= Kontext**
  Approximation der bedingten Wahrscheinlichkeit der Wortart (WA) von w_k per **Begrenzung** des berücksichtigten Kontexts:
  $P(WA(w_k) \mid WA(w_1) \ldots WA(w_{k-1})) \sim P(WA(w_k) \mid WA(w_{k-N+1}) \ldots WA(w_{k-1}))$

- bedingte Wahrscheinlichkeiten lassen sich per statistischer Analyse großer Textsammlungen **automatisch** ermitteln

Die Phasen von FASTUS

- Erkennung von Adjektiv-Gruppen / -Sequenzen

aus den Grammatikregeln erzeugte *Finite State Transducer*:



Strategien zur Koreferenzresolution

- Strategien für **Nichtpronomen**?
- vielfältige anaphorische Relationen, nicht unbedingt Koreferenz

```
Behrens hielt seinen Wagen an. Das Kühlwasser kochte.
```

- schwer zu
  - definieren / standardisieren
  - evaluieren
  - implementieren
- Ansatz:
  - **WordNet**: große Datenbank lexikalischer Beziehungen (*Teil-Ganzes*, *Element-von*, *Über- und Unterbegriffe* …)
  - Suche in WordNet nach Beziehungen jenseits Koreferenz

Strategien zur Koreferenzresolution

- Implementierung: **das ROSANA-System**

  ROSANA := RObuste Syntaxbasierte ANAphern-Interpretation

- Evaluation in vier Teildisziplinen (Precision, Recall):
  - referenzierende Ausdrücke: (0.94,0.96)
  - unmittelbare Antezedenten:* (0.71,0.71), (0.76,0.76)
  - informative Substitute:* (0.68,0.67), (0.66,0.66)
  - Koreferenz-Klassen (0.81,0.68)

* = für Dritte-Person-Pronomen (Nichtpossessiva bzw. Possessiva)