Können Computer Texte verstehen?

-

Maschinelle Sprachverarbeitung aus den Perspektiven von Philosophischer Hermeneutik, Kognitionswissenschaften und Künstlicher Intelligenz

Roland Stuckardt

inm - numerical magic gesellschaft für neue medien mbh
Abteilung LT

roland@stuckardt.de

11. Mai 1999

Warum Computerunterstützung für die inhaltsorientierte Textverarbeitung?

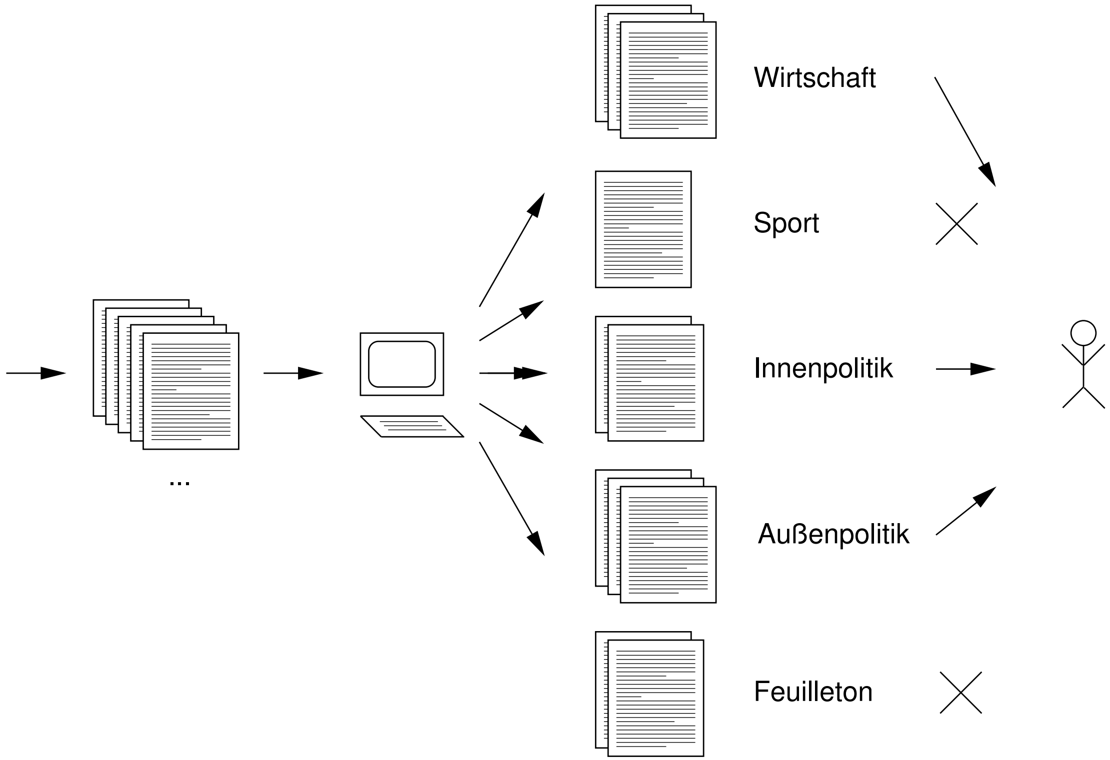
- zentrale Rolle der textuellen Informationsvermittlung
- wachsende Bedeutung elektronischer Medien:
  - Internet: www, email, news, ...
  - CDROM-Ausgaben klassischer Printmedien
  - elektronische Nachrichten-Ticker
  - digitale Bibliotheken
  - ...
- globale Vernetzung
- stetig zunehmende Wissensproduktion

Anwendungsgebiete einer verfeinerten Computer-Texttechnologie

- Archivierung und Retrieval
- Text-Klassifikation
- Text-Filterung
- Text-Kondensierung
- Fakten-Extraktion

$\Rightarrow$

- Unterstützung und Entlastung
  - in Archivierung und Dokumentation,
  - in der Postverteilung,
  - **von Entscheidungsträgern!**



Beispiel: Klassifikation und Filterung von Presse-Meldungen

Beispiel: Informationskondensierung

aus Künstler-Biographien

*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.*

Extraktion von Fakten des Typs

Künstler (K) erschafft (S) Objekt (O)

d.h. **Relationen** der Form

$$S(K, O)$$

mögliche Ergebnisse

entwickeln(Behrens,architektonische Konzeption)
bauen(Behrens,Verwaltungsgebäude der Höchst AG)

konstruieren(Wright,Präriehäuser)

planen(Utzon,Opernhaus in Sydney)
entwerfen(Utzon,Parlamentsgebäude in Kuweit)

bauen(Gropius,Erweiterungsbau der Fagus-Werke)

...

Klassische Computergestützte Textthemen-Analyse

- wortorientierte Kategoriendefinition:

Architektur	"Architekt", "architektonisch", "Architektur"
künstlerischer Beruf	"Architekt", "Maler", "Formgestalter"
wirken	"entwickeln", "Einfluß", "ausüben", "Tätigkeit", "erschließen", "Arbeit", "Wirkung", "Entwicklung", "Verwirklichung"
vorantreiben	"entwickeln", "Einfluß", "erschließen", "Entwicklung", "Verwirklichung"

- problematisch:
  - Flektion:
    {"Architekt", "Architekten"}; {"Mann", "Männer", ...}.
  - Homonymie/Polysemie: "Bauer"; "Würde"
  - **relationale inhaltliche Entitäten?**

Klassische Computergestützte Textthemen-Analyse

- zusätzliche Operatoren

Schaffens-Akte künstlerischer Beruf

VOR

(wirken ODER vorantreiben)

- problematisch: **Ausdrucksvarianten**:

Behrens erbaute die Turbinenhalle der AEG.

Die Turbinenhalle der AEG wurde von Behrens erbaut.

Die Erbauung der Turbinenhalle der AEG durch Behrens ...

Die Turbinenhalle der AEG steht in Berlin. Sie wurde von Peter Behrens erbaut.

$\longrightarrow$

- Grenzen:
  - Textthemen-Analyse bislang rein **lexikalisch**
  - **fehlende syntaktische / semantische Generalisierung**

Zwischenfazit

Computer müssen über lexikalisches, syntaktisches, semantisches und pragmatisches Wissen verfügen, um den typischen Anforderungen der Erschließung von Textinhalten gerecht werden zu können.

$\Rightarrow$

Fragen:

- Wieviel Vorwissen wird im Einzelfall benötigt?
- Wie bringt man Computern das notwendige Vorwissen bei?
- **Ist eine solche Zielsetzung überhaupt realistisch?**

Philosophische Hermeneutik

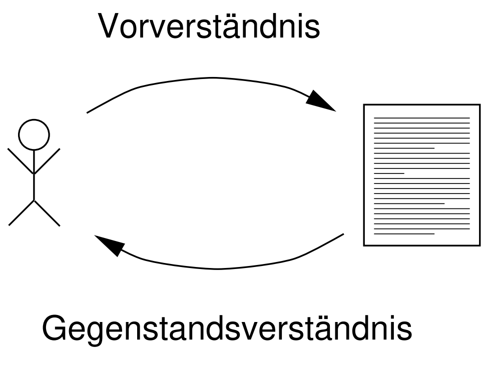
- Textexegese, Begriff des **Verstehens**
- Schleiermacher: Systematische Hermeneutik

  Verstehen als Reproduktion des Produktions-prozesses

- Dilthey: Hermeneutik als “Kunstlehre des Verstehens” schriftlich fixierter Lebensäußerungen,

  als methodologische Grundlegung der **verstehenden** Geisteswissenschaften

- Heidegger, Gadamer: **hermeneutischer Zirkel**



“Vorverständnis ist nicht vollständig explizierbar”

Soziale Dimension des Verstehens

- textuelle **Bedeutung** bzw. **Sinn** sind determiniert durch:

  - soziales und kulturelles Umfeld
  - Rollen
  - Handlungskontexte
  - $\Leftrightarrow$ **Sozialisation**

  des textproduzierenden bzw. rezipierenden **Individuums**

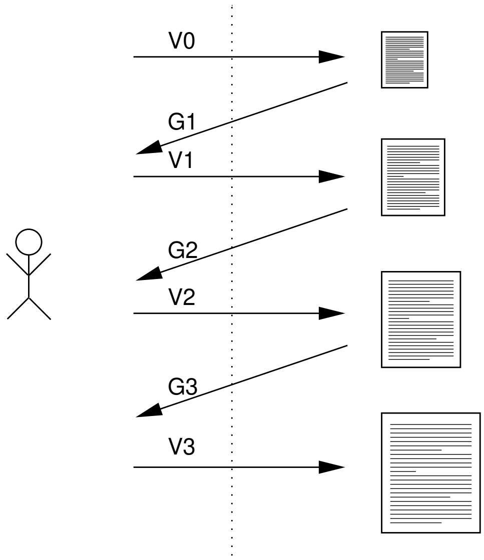
- Text-Verstehen heißt somit insbesondere:
  Rekonstruktion der sozialen Entstehensbedingungen

$\Rightarrow$

Nach dem Diktum der Hermeneutik ist das Vorwissen gesellschaftlich determiniert. Ein textinhaltserschließender Computer sei somit in geeigneter Weise zu "sozialisieren".

genauere Analyse

- hermeneutische **Spirale**



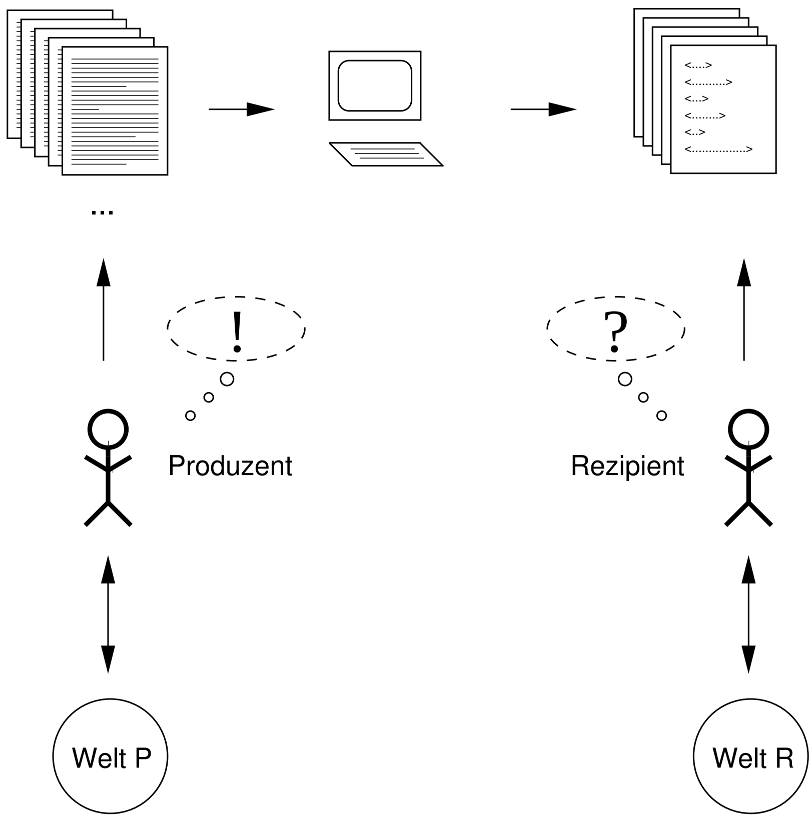
- **keine prinzipielle, formal belegbare Schranke** für die algorithmische Explizierbarkeit des Vorverständnisses

- **wieviel** Vorverständnis zu explizieren ist, hängt von der spezifischen Inhaltserschließungs-Aufgabe ab!

- **Lernansätze** sind möglich - "**Implementierung der hermeneutischen Spirale**"

Wechsel der Perspektive

- Computer als Werkzeug für **bedeutungserhaltende Symboltransformationen**:



- **Bedeutungszuweisung durch Rezipienten**

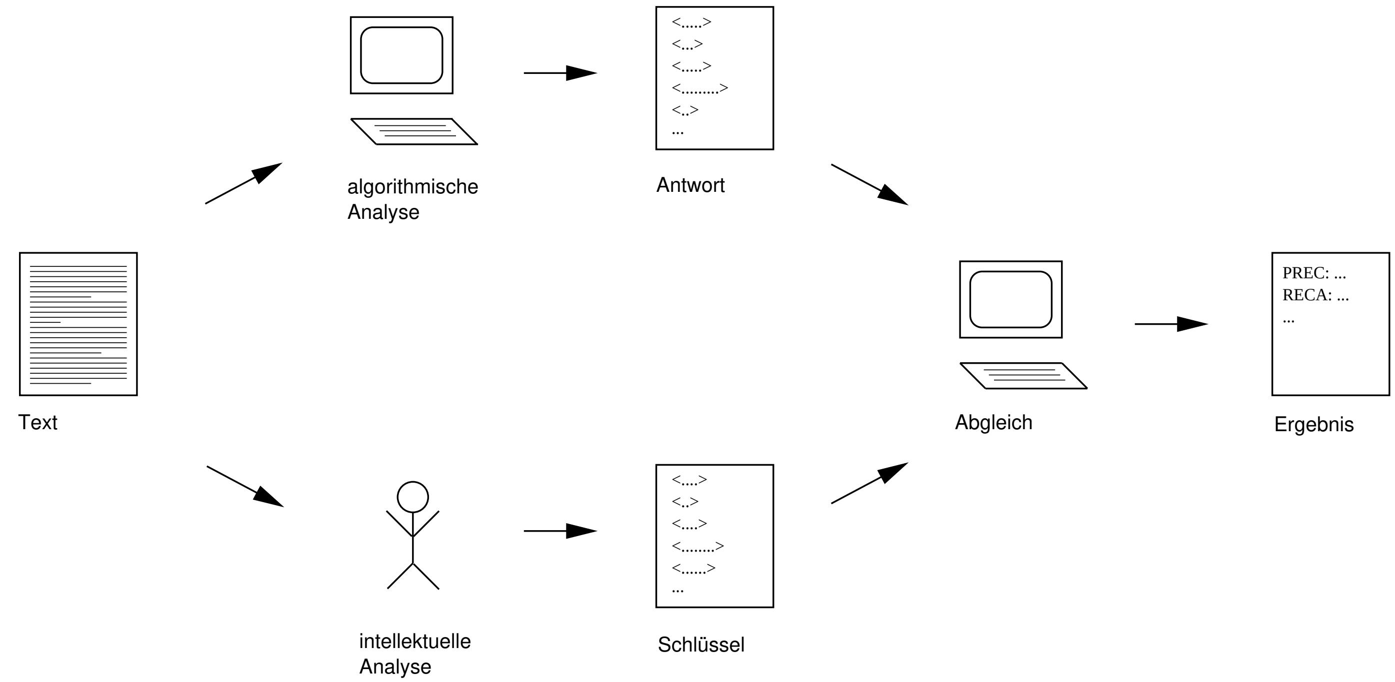
- ⇔ Kern des **Verstehensprozesses computer-extern**!

Ingenieurwissenschaftliche Sichtweise

- Perspektive einer **ergebnisorientierten** Künstlichen Intelligenz
- $\Leftrightarrow$ Reproduktion der menschlichen Sprach-**Performanz**
- These:

Bestimmte Aufgaben der symboltransformatorischen Textinhaltserschließung können auf der Basis einer partiellen, auf dem heutigen Stand der Technik mit wirtschaftlich vertretbarem Aufwand bewerkstelligbaren algorithmischen Explikation des menschlichen Vorverständnisses an den Computer delegiert werden!

- Nachweis über **formale Evaluationen**



Szenario der formalen Evaluation von Textanalyse-Systemen

Kognitionswissenschaftliche Perspektive

- Ziel: Reproduktion der menschlichen Sprach-**Kompetenz**, Identifikation allgemeingültiger **Prinzipien** der menschlichen Sprachverarbeitung

- Computer-Modelle dienen zur empirischen Überprüfung von Theorien der Sprachkompetenz - nach Miller:

Wenn eine Maschine M eines I-Verhaltens ("intelligenten" Verhaltens) fähig ist, dann unterliegt M einem System P von Prinzipien, dem jede Maschine M', die des I-Verhaltens fähig ist, ebenfalls unterliegt.

Sind "bedeutungserhaltende Symboltransformationen" über Texten überhaupt algorithmisierbar?

- Idee: Ausnutzung der **sprachinhärenten**, nicht der hermeneutischen Weltenrelativität unterliegenden **inhaltlichen Relationen**

- $\Rightarrow$ **Kompetenzmodelle**: Universelle Grammatiken

  - sprachübergreifende **Prinzipien**

  - sprachspezifische **Parameter**

- Beispiel: **Government and Binding Theory** (Chomsky)

$\Rightarrow$

Brückenschlag Kognitionswissenschaften
$\rightarrow$ ergebnisorientierte Künstliche Intelligenz

Textanalyse-Modelle einer ergebnisorientierten Künstlichen Intelligenz

- Zwischenfazit:

  Sicht des Computers als symboltransformatorisches Werkzeug zur Beschleunigung des menschlichen Verstehens-Prozesses bezüglich bestimmter inhaltlicher Teilaspekte von Texten

- Anforderungen im Hinblick auf **Anwendungs-Tauglichkeit**:

  - **Relevanz** der bewerkstelligten Transformationen
  - Verarbeitung von Texten **beliebiger** Domänen
  - **Algorithmen**-Eigenschaft
  - **Robustheit**
  - **Geschwindigkeit** der Analyse

- $\Rightarrow$ Textanalyse-Modelle der **Computerlinguistik**

Computerlinguistische Textanalyse-Modelle

- Negativbeispiel: Diskursrepräsentationstheorie (Kamp, Reyle)

  Jede Bäuerin, die einen Esel$_i$ *besitzt, schlägt ihn*$_i$*.*
  $*$ *Er*$_i$ *steht dabei im Stall.*

- Zwar: semantische Generalisierung, jedoch

  - **von begrenzter Aussagekraft** für Anwendungskorpora,
  - **nicht algorithmisch!**

- **Robustheits-Anforderung**:

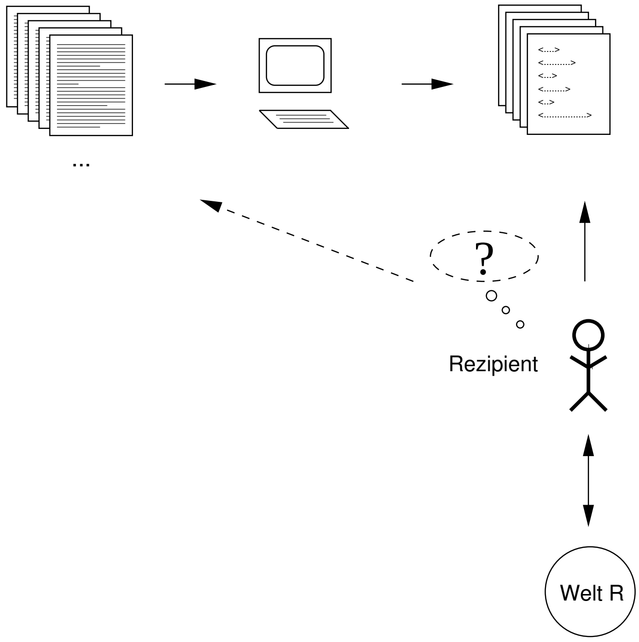
  Behrens erbaute ide Turbinenhalle der AEG.

  Behrens erbauten die Turbinenhalle der AEG.

  - gegenüber fehlerbehafteten Texten,
  - gegenüber unvollständigen Analysemodellen (!),
  - Ziel: “möglichst gutes” Ergebnis.

Computerlinguistische Textanalyse-Modelle

- **Relevanz** der erzeugten Beschreibungen:



$\longrightarrow$

Ist eine symboltransformatorische Texttechnologie, die den Kernanforderungen Algorithmen-Eigenschaft, Robustheit, Domänen-Portierbarkeit und Anwendungs-Relevanz genügt, überhaupt im Bereich des heute Machbaren?

Computerlinguistische Textanalyse-Modelle

- “**Information Extraction**”-Technologie:

  - “Extraktion” komplexer inhaltlicher Entitäten

  - **robuste Analyse** anwendungsrelevanter Texte:

    Zeitungsartikel

    Nachrichtenagentur-Meldungen

    Künstlerbiographien

    ...

  - Verarbeitung **größerer Textmengen**

Nicht das linguistische Einzelphänomen oder die theoretische Eleganz des Lösungsansatzes steht im Zentrum des Interesses - primär entscheidend ist, wie gut das Textanalysesystem den Zielsetzungen der Extraktionsaufgabe gerecht wird!

Information Extraction

$\Rightarrow$

- **ingenieurwissenschaftliche Vorgehensweise**
- **formale Evaluation** von Textanalyse-Systemen
  - Definition geeigneter Evaluations-Maße
  - intellektuelle Erstellung von Ergebnis-Schlüsseln
  - computergestützte vergleichende Evaluation
  - Analyse der statistischen Relevanz
- grundlegende **Evaluationsmaße**:

$$Precision := \frac{Korrekt}{Gefunden}$$

$$Recall := \frac{Korrekt}{Gesucht}$$

Information Extraction

- Beispiel: Extraktion von **Eigennamen**

Analyse-Ergebnis	Schlüssel
Turbinenhalle	
AEG	*AEG*
Behrens	*Peter Behrens*
	Berlin
	Hennigsdorf

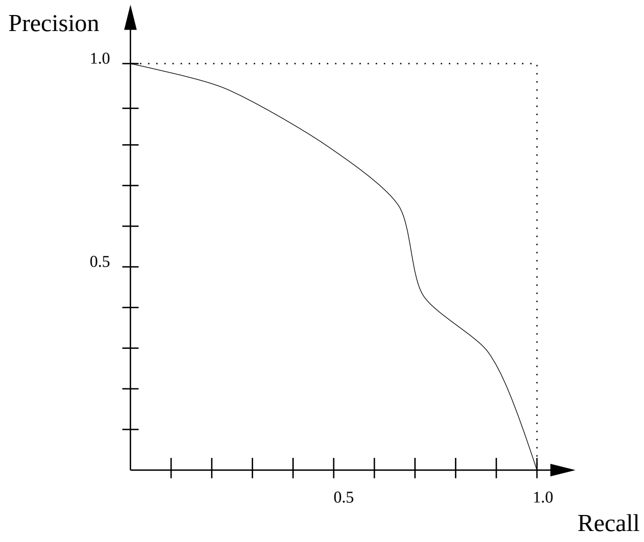
$\longrightarrow$

$$Precision = \frac{2}{3}$$

$$Recall = \frac{2}{4} = \frac{1}{2}$$

Information Extraction

- Austauschverhältnis Precision $\leftrightarrow$ Recall:



- Was kann überhaupt erreicht werden?
  - Vergleich der Leistungsfähigkeit unterschiedlicher Ansätze
  - “interannotator variability” als obere Schranke des Möglichen - **und überhaupt Meßbaren!**

“Message Understanding”-Konferenzen (MUCs)

- **etablierte Institution zur Evaluation von “Information Extraction”-Systemen**

- *Scenario Template*-Task (MUC-6):
  - Testdurchgang: “Labour Negotiation”-Szenario
  - Evaluation: “Management Succession”-Szenario

- Sub-Tasks (MUC-6):
  - *Named Entity*
  - *Template Element*
  - *Coreference Resolution*

$\longrightarrow$

- **Förderung der Entwicklung von**
  - Komponententechnologie,
  - modularer, portabler Systemarchitekturen.

Zum Beispiel: Koreferenz-Resolution

*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.*

- Task-Definition kontrovers:

  Wie kann das intuitive Verständnis von Koreferenz möglichst weitgehend formal beschrieben werden?

- "Interannotator Agreement" (MUC-6): 81 %
- Definition formaler Evaluations-Maße (Precision, Recall) nicht-trivial!

Koreferenz-Resolution: formale Evaluation

- koreferenz**klassen**-bezogene Evaluations-Maße:

  Äquivalenz von:

Behrens ist berühmt. Er baute Fabriken. Der große Architekt

Behrens ist berühmt. Er baute Fabriken. Der große Architekt

Koreferenz-Resolution: Strategien zur Algorithmisierung

- **Restriktionen**:
  - Kongruenz in Numerus und Genus

    Behrens *sagte seiner Kollegin, daß <u>er</u> sie schätze.*

  - syntaktisch-konfigurationale Bedingungen

    Behrens bemerkt, daß der **Friseur** *<u>sich</u> rasiert.*

    Behrens *bemerkt, daß der Friseur <u>ihn</u> rasiert.*

- **Präferenzen**:
  - Trägheit der semantischen Rolle:

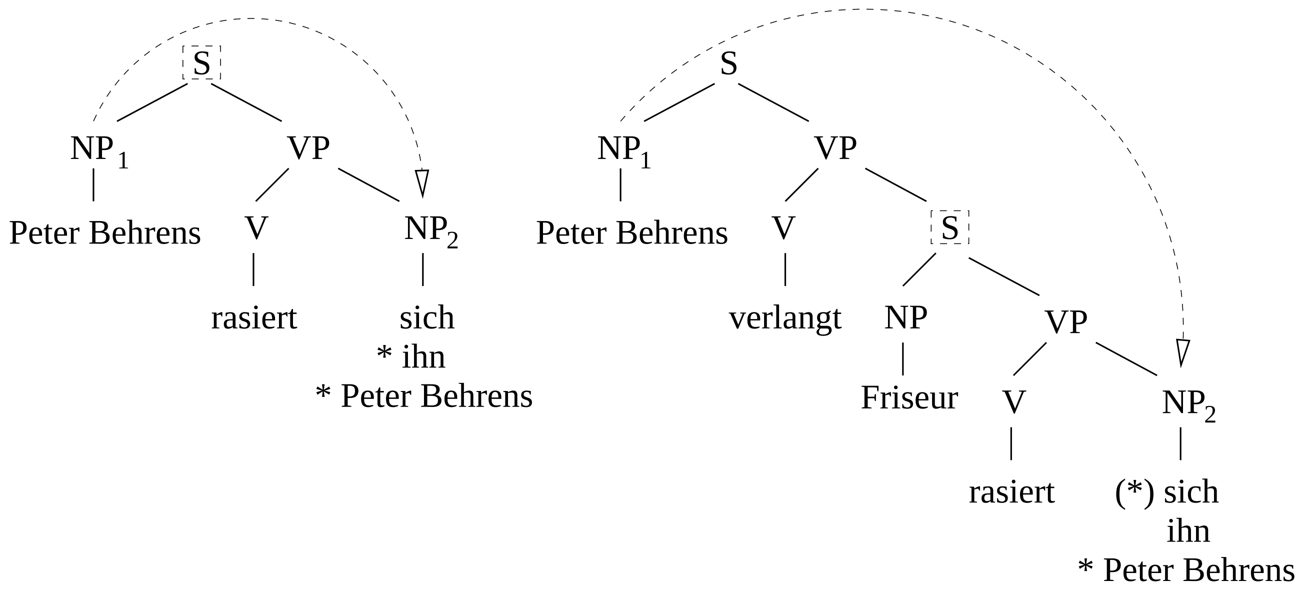
    Der Friseur *betrat den Raum.*
    Behrens ließ sich von <u>ihm</u> rasieren.

  - Subjekt-Präferenz:

    Der Friseur *betrat den Raum.*
    <u>Er</u> bediente den Kunden.

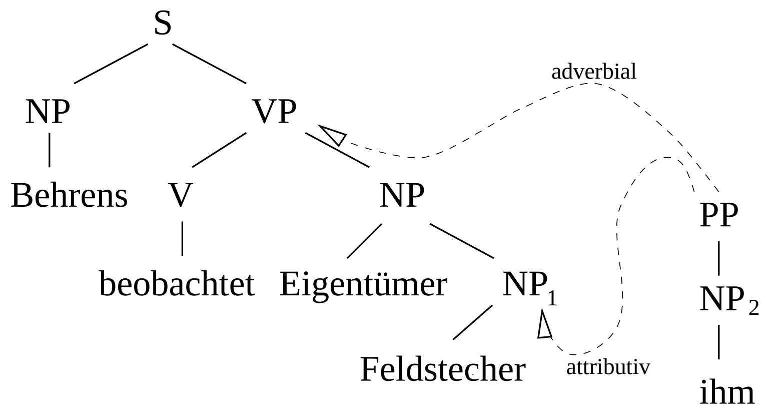
Koreferenz-Resolution: Strategien zur Algorithmisierung

- **Beispiel**: Operationalisierung syntaktisch-konfigurationaler Bedingungen qua kompetenzlinguistischen Konzepten
- drei zentrale Prinzipien der **Bindungstheorie** (Chomsky):
  - Reflexivpronomen sind **lokal gebunden**;
  - sonstige Pronomen sind **nicht lokal gebunden**;
  - Nichtpronomen sind **weder lokal noch nichtlokal gebunden**.
- die **GB-Theorie** leistet die Formalisierung der Begriffe
  - **lokal/nichtlokal**,
  - **Bindung**.



Algorithmische Verifikation syntaktisch-konfigurationaler Bedingungen

[idealisiert]



Algorithmische Verifikation syntaktisch-konfigurationaler Bedingungen

[! Robustheits-Anforderungen !]

Problem partieller syntaktischer Analysen

- Mehrdeutigkeiten in der Zuordnung bestimmter syntaktischer Elemente

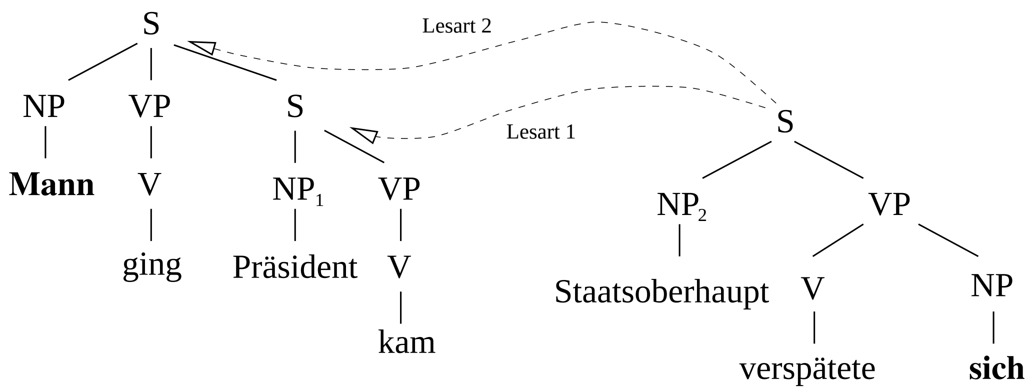
- Hintergrund:

> **Syntaktische und semantische Analyseebenen <u>interagieren</u>. Eine sequentielles Analysemodell, in dem das Ergebnis der syntaktischen Analyse den Input der semantischen Interpretation bildet, ist aus theoretischer Sicht inadäquat.**

$\Rightarrow$

- Bedarf an **Modellen <u>robuster</u> Verarbeitung**

Robuste Verifikation syntaktisch-konfigurationaler Bedingungen



Der Mann ging, bevor der Präsident kam, weil das Staatsoberhaupt sich verspätete.

Regelschema

* *Fd = [... k<typ A/B/C> ...]*   *Fe = [... bk(a)(... a<typ A> ...) ...]*

erlaubt definitiven Ausschluß von **“Mann”** als unmittelbares, bindendes Antezedens des Reflexiv-Pronomens **“sich”** !

Der ROSANA-Algorithmus

ROSANA :=

robuste **s**yntaxbasierte **Ana**phern-Interpretation

Evaluation von ROSANA

- drei **Teildisziplinen**:

  - Ermittlung referenzierender Ausdrücke,
  - Ermittlung der Koreferenz-Klassen,
  - Ermittlung nichtpronominaler Substitute.

- Beispiel:

Behrens ist berühmt. Er baute Fabriken. Seine Arbeiten ...

Ref. Ausdrücke	Koreferenzklassen	nichtpron. Substitute
Behrens Er Fabriken Seine Arbeiten	Behrens Er Seine	Er -> Behrens Seine -> Behrens
	Fabriken	
	Arbeiten	

Fazit

- für bestimmte Inhaltserschließungs-Aufgaben reicht die Qualität von Computer-Analysen nahe an die menschliche Performanz heran
- sprachinhärente inhaltliche Relationen als Basis

Sieht man die Rolle des Computers in der Durchführung bedeutungserhaltender Symbol-transformationen, so ist die Frage nach der Machbarkeit einer maschinellen Textinhalts-Erschließung zumindest partiell zu bejahen.

- **problematisch**: außersprachliches, der hermeneutischen Spirale unterliegendes kontextuelles Wissen

$\Rightarrow$

- **sozialisatorisches Modell**:

Forschungsgebiet der Arbeitsgruppe LT am Institut für Neue Medien: lernende Knowbots