Die “Message Understanding”-Konferenzen (MUCs) - Eine Institution zur Förderung der Entwicklung anwendungstauglicher inhaltserschließender Textanalysesysteme

Roland Stuckardt

D-60433 Frankfurt am Main

roland@stuckardt.de

“Intellektuelles” Textverstehen durch den Menschen

*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.*

- Wahrnehmung von
  - semantischen Entitäten
  - pragmatischen Relationen
- Bezugnahme auf sowie Einbettung in implizites Hintergrundwissen

Klassische Computergestützte Textanalyse

- Kategoriendefinition per Wortliste:

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"
  - komplexe (relationale) inhaltliche Entitäten:
    Künstler (K) erschafft (S) Objekt (O)
    $S(K, O)$

Klassische Computergestützte Textanalyse

- zusätzliche Verknüpfungs-Operatoren:

Schaffens-Akte künstlerischer Beruf
VOR
(wirken ODER vorantreiben)

- problematisch: sprachliche Realisierungsvarianten:

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 der Inhaltserschließung:
  - fehlende syntaktische / semantische Generalisierung
  - Textthemen-Analyse auf einer rein lexikalischen Ebene (Isotopie)

Computerlinguistische Textanalyse-Modelle

- für die identifizierten Teilprobleme:
  - morphologische (Flektions-) Analyse: $\surd$
  - lexikalische Disambiguierung: $(\surd)$
  - syntaktische Analyse: ?
  - Ermittlung pronominaler Bezüge: ?
  - Ermittlung komplexer inhaltlicher Entitäten: ?
- Anforderungen im Lichte der Anwendungs-Zielsetzung:
  - für anwendungsrelevante Texte beliebiger Domänen
  - Algorithmen-Eigenschaft
  - Analysegeschwindigkeit
  - Robustheit
  - **Relevanz der erzeugten Beschreibungen!**

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:

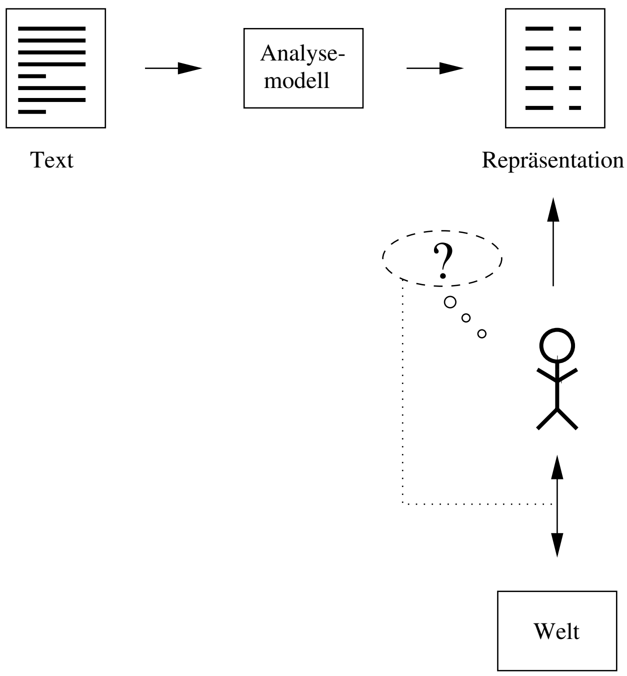
Behrens erbaute <u>id</u>e Turbinenhalle der AEG.

Behrens erbaute<u>n</u> die Turbinenhalle der AEG.

  - gegenüber fehlerbehafteter Texte,
  - gegenüber unvollständiger Analysemodelle (!),
  - Ziel: “möglichst gutes” Ergebnis.

Computerlinguistische Textanalyse-Modelle

- Relevanz der erzeugten Beschreibungen:



$\longrightarrow$

Ist eine verfeinerte computergestützte Inhaltsanalyse, 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

- Sekundär-Zielsetzungen:
  - Domänenunabhängigkeit bzw. **Portabilität**
  - Komponenten-Technologie / modulare Systemarchitektur
- **formale Evaluation** von Textanalyse-Systemen
  - Definition geeigneter Evaluations-Maße
  - Erstellung von Ergebnis-Schlüsseln
  - computergestützte vergleichende Evaluation
  - Analyse der statistischen Relevanz
- grundlegende Evaluationsmaße:

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

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

Information Extraction

- Beispiel: Eigennamen-Extraktion

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

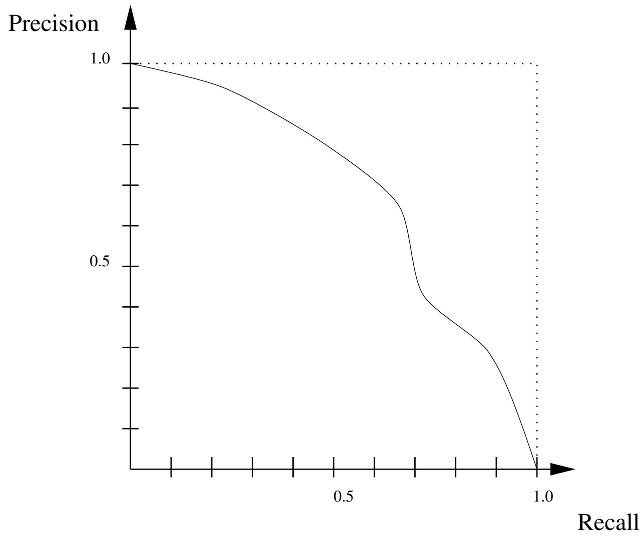
$\longrightarrow$

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

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

Information Extraction

- Austauschverhältnis Precision ↔ Recall:



- Was kann überhaupt erreicht werden?

  - Vergleich der Leistungsfähigkeit unterschiedlicher Ansätze

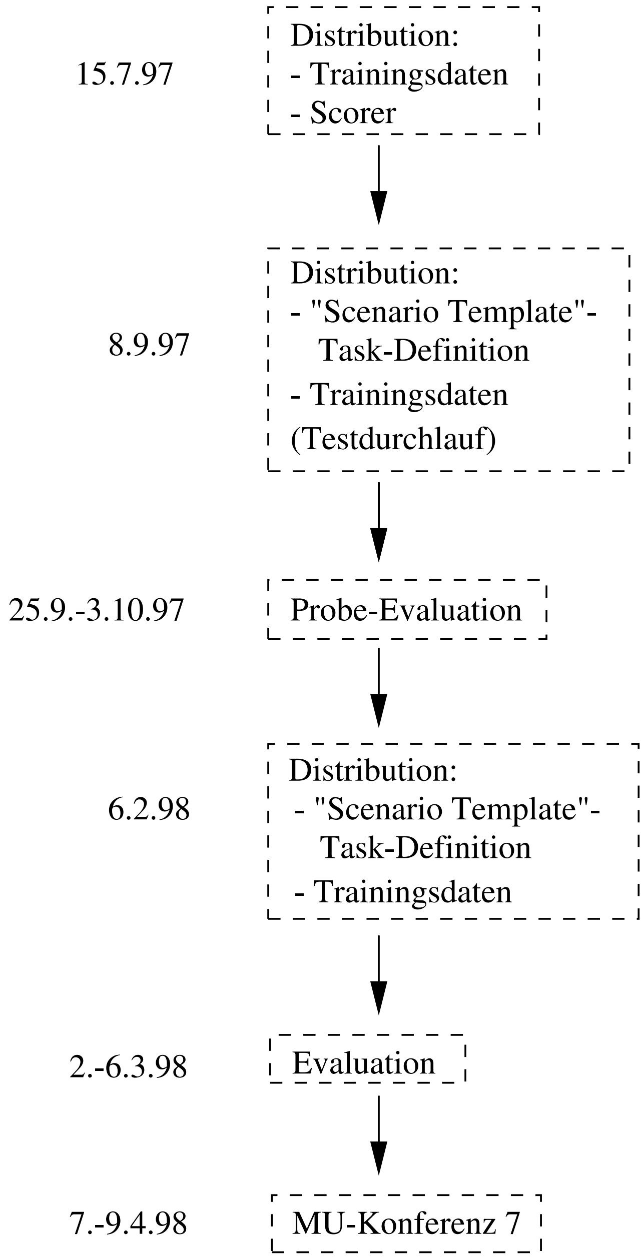
  - "interannotator variability" als obere Schranke des Möglichen - **und überhaupt Meßbaren!**

“Message Understanding”-Konferenzen

- MUCs: etablierte Institution zur Evaluation von “Information Extraction”-Systemen
- Träger:
  - NRaD:
    Naval Command, Control and Ocean Surveillance Center, Research, Development, Test & Evaluation Division
  - DARPA: Defense Advanced Research Projects Agency
- akademische und industrielle Partizipanten
- Historie:
  - MUC-1 (1987): militärische Beobachtungsmeldungen
  - MUC-3 (1991): “Terroranschläge” (18 Slots)
  - MUC-5 (1993): “Joint Ventures” (47 Slots)
  - MUC-6 (1995): “Wallstreet Journal”-Artikel, Personalveränderungen in Unternehmen.
    Komponenten-Technologie; Portabilität, Tiefenanalyse.
  - MUC-7 (1998): “NYT News Service”-Artikel

“Message Understanding”-Konferenzen

- Fahrplan MUC-7:



MUC-6

- *Scenario Template*-Task:
  - Testdurchgang: “Labour Negotiation”-Szenario
  - Evaluation: “Management Succession”-Szenario
- Sub-Tasks: Komponententechnologie, Tiefenanalyse
  - *Named Entity* (NE)
  - *Scenario Template* (ST)
  - *Coreference Resolution* (CO)

$\longrightarrow$

- **Förderung portabler Systemarchitekturen**

MUC-6: CO-Task

- 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!
  "*Was genau heißt Koreferenz?*"

- Definition der Scoring-Maße nicht-trivial!

- "Interannotator Agreement":

$$Precision = 0.82$$

$$Recall = 0.80$$

MUC-6: CO-Task

- koreferenklassenbezogene Scoring-Maße:
  Äquivalenz von:

Behrens ist berühmt. Er baute Fabriken. Der große Architekt

Behrens ist berühmt. Er baute Fabriken. Der große Architekt

Fazit

- “Message Understanding”-Konferenzen:
  - adäquate Definition von I.E.-Tasks
  - Identifikation relevanter Subtasks
  - formale, theoriebasierte Definition von Scoring-Maßen
  - Ressourcen:

    Scoring-Software

    annotierte Korpora
  - Evaluationsrahmen, der domänenportable, modulare Systamarchitekturen fördert

$\longrightarrow$

Die “Information Extraction”-Technik hat Textanalysesysteme hervorgebracht, die den oben identifizierten Anforderungen an eine verfeinerte computergestützte Inhaltsanalyse genügen und zumindest in Teilbereichen bereits sehr gute Ergebnisse liefern.