
Semantische Sprachverarbeitung

Carola Eschenbach
Universität Hamburg, FB Informatik
AB Wissens- und Sprachverarbeitung (WSV)

Sommersemester 2004

Aufgabe: Thematische Rollen und Selektionsrestriktionen

- gib zu den Verben des Textes die thematischen Rollen und die Komplementausdrücke an
 - Wie funktioniert das Linking?
 - Welche Selektionsrestriktionen gelten für die Argumentpositionen?
- Bei welchen anderen Ausdrücken (Nomen, Adjektiven, Präpositionen) sind Selektionsrestriktionen zu beobachten?

Aufgabe: Thematische Rollen und Selektionsrestriktionen

1. Wenn du beim Pförtner **stehst**, dann **siehst** du das höchste Gebäude auf dem Gelände, Haus F
2. Von Haus F **führt** im ersten Stock ein Übergang zu Haus D
3. **Gehe** zuallererst zwischen Haus D und Haus F unter dem Übergang **durch**
4. Auf der Rückseite von Haus D **gehst** du **entlang**, bis auf deiner rechten Seite Haus E **erscheint**
5. Haus E **betrittst** du über eine Rampe

Semantische Sprachverarbeitung

Sitzung 8

- Language Engineering: Robuste Verfahren
 - Überblick
 - Bewertung von Verfahren
 - Template-Ansätze (FASTUS)
 - Flache Ansätze zur Desambiguierung (WSD)
-

Language Engineering

Aufgabenstellungen

- Machine Translation (MT)
 - Übersetze Texte in eine andere Sprache
- Information Retrieval (IR)
 - Finde Texte und präsentiere sie dem Benutzer
- Information Extraction (IE)
 - Analysiere Texte und präsentiere spezifische Information daraus
 - + Reduktion des Leseaufwands
 - viel Wissen erforderlich (knowledge-intensive)
 - Domänenabhängig
 - weniger genau als menschlicher Leser
 - + Basis für mehrsprachige Ausgabe

Information Extraction

Eingabematerial

- Real vorkommende Texte (unrestricted text), i.a. kurze Nachrichtenmeldungen

Teilaufgaben: Extraktion von

- Benannten Entitäten (Named Entity recognition, NE)
- Koreferenz (Coreference resolution, CO)
- deskriptiver Information (Template Element construction, TE)
- Relationen zwischen Entitäten (Template Relation construction, TR)
- Ereignisinformation (Scenario Template production, ST)

Genauere Definition der Aufgabe

- erlaubt quantitative Auswertung von IE Systemen

Beispiel

- The shiny red rocket was fired on Tuesday. It is the brainchild of Dr. Big Head. Dr. Head is a staff scientist at We Build Rockets Inc.

Benannte Entitäten (NE)

- Tuesday, Dr. Big Head, Dr. Head, We Build Rockets Inc.

Koreferenz (CO)

1. the shiny red rocket = it
2. Dr. Big Head = Dr. Head

Deskriptive Information (TE)

1. shiny, red, a rocket
2. a scientist

Beispiel

- The shiny red rocket was fired on Tuesday. It is the brainchild of Dr. Big Head. Dr. Head is a staff scientist at We Build Rockets Inc.

Benannte Entitäten (NE)

Koreferenz (CO)

Deskriptive Information (TE)

Relationen (TR)

- Dr. Big Head's brainchild
- employee of WBRI

Ereignisse (ST)

- 1 was launched [on Tuesday]

Informationsextraktion vs. Textverstehen

Informationsextraktion

- nur ein Bruchteil des Textes ist relevant (z.B. 10 %)
- extrahierte Information wird auf ein einfaches, vordefiniertes Schema abgebildet
- feine Nuancen und Intentionen des Textproduzenten sind irrelevant

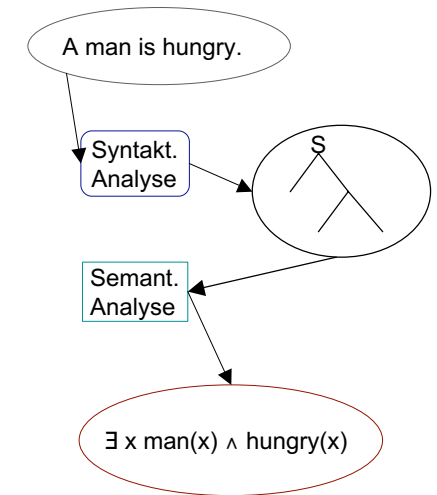
Textverstehen

- Der ganze Text soll interpretiert werden.
- Zielrepräsentation muss die volle Komplexität des Textinhaltes widerspiegeln können
- Nuancen und Sprecherintentionen sollen erkannt werden.

Kontrast zu tiefer Verarbeitung

Bedeutungsanalyse basierend auf Syntax-Semantik-Schnittstelle

- Probleme bei der syntaktischen Analyse:
 - Fehlende Lexikoneinträge
 - Ungrammatische Eingabe
 - Tippfehler
- semantische Analyse hängt von der syntaktischen Analyse ab:
 - Ungleichgewicht zwischen Syntax und Semantik



Evaluation von IE Systemen

Metrik statt Grammatikalität / Erfüllbarkeit

- Gemessene Werte (in einem Beispieltext)
 - N = # der Textstellen die erkannt werden sollten.
 - E = # der extrahierten Textstellen
 - K = # der davon korrekt erkannten Textstellen
- Recall: $R = K/N$
 - Wieviel relevante Information hat das System extrahiert?
- Precision: $P = K/E$
 - Wieviel der extrahierten Information ist korrekt?
- F-measure
 - kombiniertes Maß aus Recall und Precision
 - Parameter β für Gewichtung

$$F = \frac{(\beta^2 + 1)PR}{\beta^2 P + R}$$

Beispiel: Güte eines NP chunker

Test-File

- enthält (N =) 554 NP chunks, die erkannt werden sollten.

Leistung des Verfahrens

- (E =) 503 NP chunks werden extrahiert.
- davon waren (K =) 420 korrekt erkannte NP chunks.

Recall

- $R = K / N = 420 / 554 = 0.76$

Precision

- $P = K / E = 420 / 503 = 0.84$

F-measure (für $\beta=1$):

- $2 * P * R / (P + R) = 2 * 0.76 * 0.84 / (0.76 + 0.84) = 0.80$

$$F = \frac{(\beta^2 + 1)PR}{\beta^2 P + R}$$

State-of-the-art (MUC-7, 1998)

Aufgabe	Recall	Precision	F-Measure
Named Entity	92	95	93,93
Coreference	56,1	68,8	61,8
Template Element	86	87	86,76
Template Relation	67	86	75,63
Scenario Template	42	65	50,79

Nachrichtenartikel analysieren

MUC (Message Understanding Conference)

- Beispielaufgabe (ST): extrahiere aus Nachrichtentexten Infos über joint ventures (MUC-5)

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.

Beispiel: Ergebnis

TIE-UP-1:	
Relationship:	TIE-UP
Entities:	„Bridgestone Sports Co.“ „a local concern“ „a Japanese trading house“
Joint Venture Company:	„Bridgestone Sports Taiwan Co.“
Activity:	ACTIVITY-1
Amount:	NT\$20 000 000
ACTIVITY-1:	
Activity:	PRODUCTION
Company:	„Bridgestone Sports Taiwan Co.“
Product:	"iron and 'metal wood' clubs"
Start Date:	DURING: January 1990

Template-Verfahren (IE)

Ähnlich zu regelbasierten Ansätzen der semantischen Sprachverarbeitung, aber einfacher und flacher

- Verarbeitung semantischer Information und Füllen von einfachen Mustern „templates“
- Klassifikation einer kleinen Teilmenge an relevanter Information: entities, relations, scenarios

Beispiel

- Fastus: Hobbs et al. 1997

Kaskaden von endlichen Automaten

Einfache und effiziente Methode, Templates aufzufüllen

- Endliche Automaten (EA/FSTA) werden für jede Ebene eingesetzt:
- Tokens: Eingabe in Einzelworte zerlegen
 - Complex words: Erkennen von Mehr-Wort-Lexemen, Zahlen und Eigennamen
 - Basic phrases: Segmentierung in NP, VP chunks und Partikeln
 - Complex phrases: Erkennen von komplexen NP,VP chunks
 - Semantic patterns: Extraktion semantischer Entitäten
- **Merging: Referenzauflösung**

Definition Finite State Transducer

Ein FST (finite state transducer)

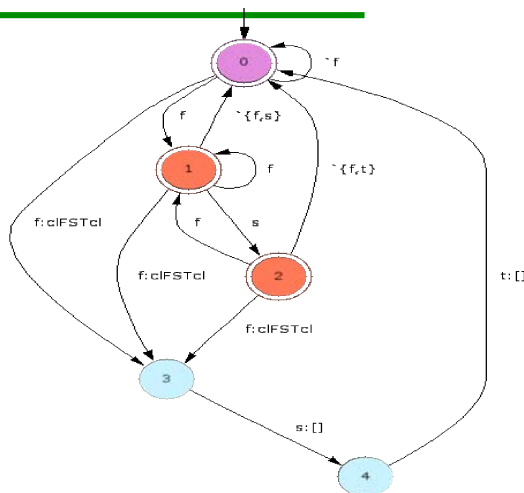
- ist ein endlicher Automat mit einem spezifischen Alphabet.

Formal

- ein 5-Tupel $(Q, \Sigma, q_0, F, \delta)$
 - Q ist eine endliche Menge von n Zuständen q_0, \dots, q_n .
 - Σ ist ein endliches Alphabet von komplexen Symbolen. Jedes Symbol besteht aus einem Ein- und Ausgabepaar.
 - q_0 ist der Anfangszustand.
 - F ist die Menge der Endzustände.
 - δ ist die Übergangsfunktion.

fst:cIFSTcl

	0	1	2	3	4
-f	0				
f:cIFSTcl	3	3	3		
s:[]				4	
t:[]			1		0
f	1	1			
s:[]		2	0		
-(f,t)			0		
-(f,s)	0				



Erkennen von Namen (NE)

Eigennamen von

- Personen (Peter, Prof. Dr.-Ing. H. Siegfried Stiehl)
- Firmen (Apple)
- Organisationen (UNO)
- Zeiträumen (Montag, Januar, 2004)
- Orten (Hamburg, Australien)

Verschiedene Methoden

- FST
- Gazetteers (Liste von Eigennamen)
- Statistische Methoden
- Maschinelles Lernen

Firmennamen erkennen

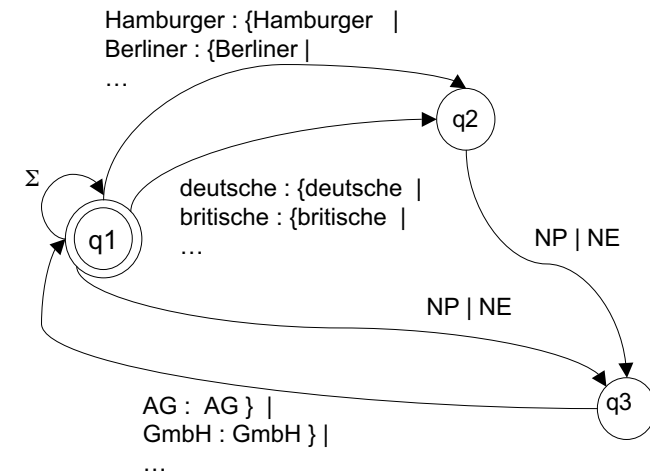
Einfache Regeln erfassen bestimmte Eigenschaften von Firmennamen

- Comp-Name -> (pre-location) NP Comp-Suffix
- pre-location -> locname | nationality
- locname -> Hamburger | Berliner | ...
- nationality -> deutsche | britische | ...
- Comp-Suffix -> AG | GmbH | ...

Regeln werden in FST übertragen

- Approximation mit Übergeneralisierung

FST



Haupteigenschaften von Template-Ansätzen / FST

- Templates sind vordefiniert und können nicht verändert werden
- Die extrahierte Information wird wortwörtlich wiedergegeben, fakten-zentriert.

Semantische Sprachverarbeitung

Disambiguierung von Wortbedeutungen

Robuste Verfahren zur Bedeutungs-Disambiguierung

Eigenschaften robuster Verfahren

- Unabhängig (stand-alone)
- Minimale Annahmen über Informationen von anderen Prozessen

Drei Verfahren des Maschinellen Lernens

- überwacht (menschl. Vorklassifikation)
- Bootstrapping (minimale Vorklassifikation)
- nicht überwacht (keine Vorklassifikation)

Weiterer Ansatz: Lexikon-basierte Verfahren

Maschinelles Lernen

Lernverfahren werden auf Sprachdaten angewendet

- Vorgabe: Anzahl und Art der Bedeutungen
- Unterschiede zwischen Verfahren
 - Art des Lernmaterials
 - Menge des Lernmaterials
 - Menschlicher Eingriff (supervised vs. unsupervised)
 - Linguistisches Wissen
- Resultat: Klassifikator

Offene Frage

- Anwendbarkeit auf kleine oder auch große Datenmengen?

Allgemeine Voraussetzungen

Eigenschaftsvektoren (Feature Vectors)

- Eingabe: Text: Zielwort mit / im Kontext
- Vorverarbeitungsoptionen
 - Part-of-speech-tags (Wortklassenerkennung)
 - Reduktion des Kontextes (n Wörter)
 - Bildung von Stammformen für Kontext
 - partielles Parsen (seltener)
- Verarbeitung: Extraktion der Eigenschaften
- Ausgabe: Feature Vectors
 - Kollokationen (collocations)
 - Häufigste zusammen auftretende Wörter (Co-occurrence)

Beispiel: Collocational Feature Vector

Eingabe mit Zielwort

- *An electric guitar and bass player stand off to one [...]*

Eigenschaftsvektoren

- Reduktion des Kontexts auf je n (=2) Wörter rechts und links des Wortes

[guitar, and, player, stand]

- Möglich: Beibehaltung der Reihenfolge
- Möglich: Zuordnung von Wortklassenmarkierungen

[guitar, NN1, and, CYC, player, NN1, stand, VVB]

Beispiel: Co-Occurrence Feature Vector

Eingabe mit Zielwort

- An electric *guitar* and bass *player* stand off to one [...]

Vorwissen: die häufigsten Kontextwörter des Zielwortes

- für *bass*: [fishing, big, sound, *player*, fly, rod, pound, double, runs, playing, *guitar*, bans]

Eigenschaftsvektoren

- Reduktion des Kontexts auf je 5 Wörter rechts und links des Wortes
- Ignoranz der Reihenfolge und Syntax
- Vergleich mit dem Vektor häufigster Kontextwörter
→ [0,0,0,1,0,0,0,0,0,0,1,0]

Überwachte Lernverfahren

Disambiguierung der Bedeutung

= Auswahl des wahrscheinlichsten Sinns bei gegebenen Eigenschaftsvektor

Lernverfahren mit Eingabe von Trainingsmenge

- Eigenschaftsvektor
- + Klassifikation
- Klassifikator (= Klassifikationsregel)
- **Verschiedene Verfahren**
 - Bayes
 - Entscheidungslisten(-bäume)
 - Neuronale Netzwerke etc.

Naive Bayes'sche Klassifikation

Auswahl des korrekten Sinns s^*

- Gegeben: Menge möglicher Sinne S
- $P(s|V)$: Die Wahrscheinlichkeit für s , gegeben V
- Wähle den Sinn s^* , der bei gegebenem Vektor V am wahrscheinlichsten ist
- $s^* = \operatorname{argmax}_{s \in S} P(s|V)$
- Die Wahrscheinlichkeiten $P(s|V)$ sind schwer direkt abschätzbar

Deshalb wird das Bayes'sche Verfahren angewandt

Bayes'sche Formel für bedingte Wahrscheinlichkeit

$$P(s|V) = \frac{P(s) \cdot P(V|s)}{\sum_{s_i \in S} P(s_i) \cdot P(V|s_i)}$$

Der Nenner

- ist für alle s gleich
- dient der Normierung
- kann für die Bestimmung des Maximums ignoriert werden

Erforderlich für die Klassifikation für jedes s und V

- Abschätzung von $P(s)$
- Abschätzung von $P(V|s)$

Abschätzung von $P(V|s)$

Annahme

- Für die verschiedenen Positionen v_j des Eigenschaftsvektors sind die $P(v_j|s)$ unabhängig
- Wäre dieses tatsächlich der Fall, dann würde gelten
 - $P((v_1, \dots, v_n)|s) = \prod_{j \in \{1, \dots, n\}} P(v_j|s)$
- Da diese aber nicht gelten muss, haben wir nur

$$P((v_1, \dots, v_n)|s) \approx \prod_{j \in \{1, \dots, n\}} P(v_j|s)$$

Erforderlich für die Klassifikation für jedes s und v_j

- Abschätzung von $P(s)$
- Abschätzung von $P(v_j|s)$

Relative Häufigkeit (RH) im Trainingsmaterial

Bass Beispiel (alles fiktiv)

$P(s) \approx$ RH des Sinnes s im Lernmaterial

- $RH(\text{bass}(\text{fish})) = 0.23$
- $RH(\text{bass}(\text{music})) = 0.77$

$P(v_j|s) \approx$ RH von v_j wenn der Sinn s ist im Lernmaterial

- Co-Occurrence-Vektor [0,0,0,1,0,0,0,0,0,1,0]
[fishing, big, sound, player, fly, rod, pound, double, runs, playing, guitar, bans]
- $RH(x_4 = 1 | \text{bass}(\text{music})) = 0.02$
- $RH(x_4 = 1 | \text{bass}(\text{fish})) = 0.002$
- Colocation-Vektor
[guitar, and, player, stand]
- $RH(\text{player} | \text{bass}(\text{music})) = 0.003$
- $RH(\text{player} | \text{bass}(\text{fish})) = 0.00001$

Bayes-Klassifikator

Gegeben

Lernmaterial LM

mit Klassifikation: $LM(s)$: Teil von LM mit Sinn s

Bestimmt wird

$$s^* = \underset{s \in S}{\operatorname{argmax}} (RH(s, LM) * \prod_{j \in \{1, \dots, n\}} RH(v_j, LM(s)))$$

- Bestimmung des ‚wahrscheinlichsten‘ Sinnes

Colocation-Vektor [guitar, and, player, stand]

- $RH(\text{guitar} | \text{bass}(\text{music})) * RH(\text{and} | \text{bass}(\text{music})) * RH(\text{player} | \text{bass}(\text{music})) * RH(\text{stand} | \text{bass}(\text{music})) * RH(\text{bass}(\text{music}))$

sollte größer sein als

- $RH(\text{guitar} | \text{bass}(\text{fish})) * RH(\text{and} | \text{bass}(\text{fish})) * RH(\text{player} | \text{bass}(\text{fish})) * RH(\text{stand} | \text{bass}(\text{fish})) * RH(\text{bass}(\text{fish}))$

Bootstrapping Verfahren

Überwachte Lernverfahren

- benötigen große Menge annotierter Daten

Bootstrapping-Verfahren

- arbeiten mit weniger Lernmaterial
 - handverlesen
 - möglichst typische Beispiele
 - klassifiziert
- Iteration
 - unklassifizierte Daten werden auf dieser Basis klassifiziert
 - Lernmaterial wird um das gewonnene Material ergänzt
- bis die Fehlerrate niedrig genug ist.

Nicht überwachte Lernverfahren

Lernverfahren ohne Klassifizierung der Daten

Input

- Eigenschaftsvektoren von nicht-klassifizierten Instanzen

Verfahren

- Gruppierung entsprechend einer Ähnlichkeitsmetrik
- gut erforschtes Problem (Duda and Hart, 1973)

Output

- Cluster von Wörtern

Nachbereitung von Hand

- Zuweisung passender Bedeutung zum Cluster

Lexikon-basierte Ansätze

Alle besprochenen Ansätze sind nur für kleinere Datensätze praktikabel

- realisierte Ansätze meist nur 2-12 Lexeme

Für eine größere Datenmenge werden *Maschinenlesbare Lexika* verwendet:

- Definitionen werden aus dem Lexikon extrahiert
- Übereinstimmungen zwischen Wortdefinitionen von Zielwort und Kontextwort selegiert Wortbedeutung

Drei Lexikon-basierte Ansätze

Die Überlappung der Wortbedeutung erfolgt aufgrund von

- Bedeutungsdefinition im Lexikon (z.B. Lesk, 1986)
- Relationen in Thesauri (z.B. Walker, 1987)
- Übersetzung in eine andere Sprache (Dagan & Itai, 1991)

Bedeutungsdefinitionen

Definitionen der Wortbedeutungen

- *bags-of-words*

Der Kontext eines ambigen Wortes

- *bag-of-words*

Korrekte Wortbedeutung

- wird durch die höchste Überlappung bestimmt.

Beispiel

Bedeutung von pine und cone in *pine cone*

pine

- 1 kinds of evergreen tree with needle-shaped leaves
- 2 waste away through sorrow or illness

cone

- 1 solid body which narrows to a point
- 2 something of this shape whether solid or hollow
- 3 fruit of certain evergreen trees

Auswahl: pine1, cone3

bass player (WordNet)

1. bass -- (the lowest part of the musical range)
 2. bass -- (the lowest part in polyphonic music)
 3. bass -- (an adult male singer with the lowest voice)
 4. bass -- (flesh of lean-fleshed saltwater fish of the family Serranidae)
 5. bass -- (any of various North American lean-fleshed freshwater fishes especially of the genus Micropterus)
 6. bass -- (the lowest adult male singing voice)
 7. bass -- (the member with the lowest range of a family of musical instruments)
 8. bass -- (nontechnical name for any of numerous edible marine and freshwater spiny-finned fishes)
 9. bass -- (having or denoting a low vocal or instrumental range)
-
1. player -- (a person who participates in or is skilled at some game)
 2. player -- (someone who plays a musical instrument (as a profession))
 3. player -- (a theatrical performer)

Thesaurus-basiert

Jeder Bedeutung eines Wortes

- wird ein *subject code* zugeordnet
- bass1 → music; bass4 → animal

Aus dem Kontext des ambigen Wortes

- werden für jedes Wort ebenfalls die *subject codes* extrahiert
- player → music, sports, theater

Übereinstimmung im *subject code*

- selektiert die Bedeutung

Mehrsprachige Lexika

Die Übersetzung von *interest* (engl.) ins Deutsche

- *Beteiligung* (an einem Unternehmen)
- *Interesse* (an einer Sache)

mehrsprachiges Lexikon

- Abbildung der Bedeutungen von *interest*
- auf die Wortbedeutung der deutschen Übersetzungen

Zusammenfassung

Language Engineering

- Echte Texte, keine linguistischen Einschränkungen aber Textsorten- und Domänen-Beschränkungen
- Kein volles Verstehen sondern Fokussierung auf Extraktion wesentlicher Teile
- Bewertung über Recall, Precision und F-Measure
- Nutzung bestehender elektronischer Ressourcen
- Teilaufgabe WSD
 - Maschinelle Lernverfahren:
 - überwacht/bootstrapping/unüberwacht
 - Lexikon-basierte Verfahren

Literatur

- Hobbs, Jerry R., Douglas E. Appelt, John Bear, David Israel, Megumi Kameyama, Mark Stickel & Mabry Tyson (1997). FASTUS: A cascaded finite-state transducer for extracting information from natural-language text. In E. Roche & Y. Schabes (eds.) Finite State Devices for Natural Language Processing (pp. 383-406). MIT Press: Cambridge, MA.
<http://www.isi.edu/~hobbs/fastus-schabes-jul95.pdf>
- Jurafsky, Daniel & James H. Martin (2000). Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Prentice Hall: Upper Saddle River, NJ. ch. 17.1-17.2
- Cunningham, Hamish (1999). Information Extraction – a User Guide. Institute for Language, Speech and Hearing (ILASH) and Department of Computer Science, Report CS – 99 – 07. University of Sheffield: UK.