MENSCH UND SPRACHE

Modelle der Computerlinguistik

Katie McCann, Lukas Gienapp, Raphaela Fietta

May 3, 2016

GLIEDERUNG

1. Einführung in die Computerlinguistik
2. Syntax
3. Semantik
4. Schlussbetrachtung

ANLIEGEN DES VORTRAGS

Fragestellung

- Welcher Modelle bedienen sich Computer, um natürliche Sprache abzubilden?

- wir möchten einen kurzen, allgemeinen Einblick in die Arbeitsweisen der Computerlinguistik geben
- durch den begrenzten zeitlichen Rahmen und das unterschiedliche Vorwissen werden wir uns auf die Analyse sprachlicher Audrücke in Aussagesatzform beschränken

EINFÜHRUNG IN DIE COMPUTERLINGUISTIK

Was ist Computerlinguistik?

- Computerlinguistik (CL) ist ein interdisziplinäres Fachgebiet zwischen den Gebieten Informatik und Linguistik
- CL beschäftigt sich mit der maschinellen (algorithmischen) Verarbeitung natürlicher Sprache und natürlichem Text

CL ALS TEILDISZIPLIN

Computerlinguistik ist ein sehr allgemeiner Begriff. Er umfasst:

- Entwicklung von Programmen
- Realisierung von sprachlichen Phänomenen
- Sprachtechnologie

PRAKTISCHE VS. THEORETISCHE CL

Verschiedene Fragestellungen, die es zu beantworten gilt.

- Wie programmiert man Software, um natürliche Sprache zu verarbeiten?
- Welche Rechnerleistung benötige ich dafür?

ANWENDUNGEN DER CL

- **Maschinelle Übersetzung** - Übersetzung eines Textes aus einer Quellsprache in eine Zielsprache
- **Korrekturprogramme** - Erkennung und Korrektur von Fehlern
- **Volltextsuche** - Auffinden relevanter Keywords in einem Text
- **Klassifikation von Texten** - Einordnung in Kategorien
- **Text-Zusammenfassung** - Erstellen eines *Abstracts*
- **Spracherkennungssysteme** - Übertragung von gesprochener Sprache auf die Maschine
- **Informationsextraktion** - Domänenspezifische Suche in Texten
- **Dialogsysteme** - Kommunikation mit der Maschine
- etc.

METHODEN DER COMPUTERLINGUISTIK

Symbolische Methoden :

- eng mit Linguistik verbunden
- **Parsing**: Analyse sprachlicher Ausdruck mit allen Formen (Grammatik und Wortarten)

Statistische Methoden :

- stochastische Verfahren mit trainierbaren Modellen
- erstellen von Korpora (Text-, Sprach-, Multimodale Korpora, Baumbanken)
- **Tagging**: ein Korpus wird mit grammatischen Informationen Versehen, Klassifizierung Wortarten

Hybride Methoden :

- nutzen sowohl symbolische als auch statistische Verfahren
- **statistisches Parsing** nutzt beide Methoden

SYNTAX

SYNTAX

- Syntax ist im allgemeinen ein Regelsystem zur Kombination sprachlicher Zeichen zu komplexen Ausdrücken
- Syntax erfasst also die Struktur sprachlicher Ausdrücke

Beispiel

(1) Der Hund bellt.

(2) * Hund bellt der.

- Sprecher einer Sprache können über die Wohlgeformtheit (Grammatikalität) eines sprachlichen Ausdrucks entscheiden.
- im Beispiel ist der Grund für Ungrammatikalität die Struktur des Satzes
- die Syntax einer Sprache trifft Aussagen über die Struktur und Wohlgeformtheit eines sprachlichen Ausdrucks
- im folgenden soll die Syntaxanalyse anhand des Modells der Phrasenstrukturgrammatik näher erläutert werden

PHRASENSTRUKTURGRAMMATIK

Eine Grammatik besteht aus vier Elementen:

- einer Menge **terminaler Symbole** (Wörter)
- einer Menge **nicht-terminaler Symbole** (Kategorien von Satzbestandteilen)
- einem **Startsymbol** S
- einer Menge an **Ersetzungsregeln**

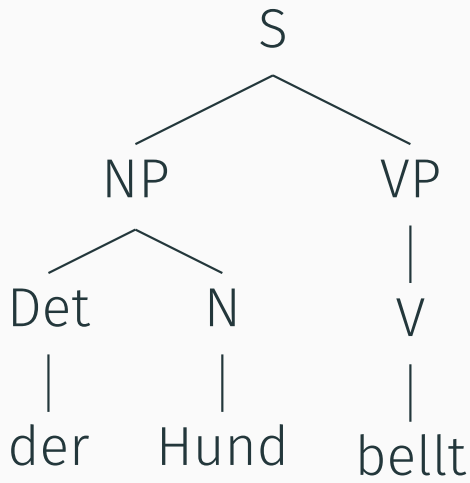
BEISPIEL

Nicht-terminale Symbole
$V_N = \{S, NP, VP, Det, N, V\}$

Terminale Symbole
$V_T = \{der, Hund, bellt\}$

Ersetzungsregeln

$$R = \left\{ \begin{array}{r} S \rightarrow NP\ VP \\ NP \rightarrow Det\ N \\ VP \rightarrow V \\ Det \rightarrow der \\ N \rightarrow Hund \\ V \rightarrow bellt \end{array} \right\}$$



- Computerlinguistik benutzt Phrasenstrukturgrammatik, um syntaktische Struktur eines sprachlichen Ausdrucks zu analysieren
- mithilfe des **Parsings** kann ein Computer die Grammatikalität eines Satzes bewerten
- es existieren verschiedene Parsing-Algorithmen: Top-Down-Parsing und Bottom-Up-Parsing

Beispiel

Der englische Satz “Book that flight.” soll geparsed werden.

BEISPIELGRAMMATIK

Nicht-terminale Symbole

$V_N = \{S, NP, VP, Det, N, V\}$

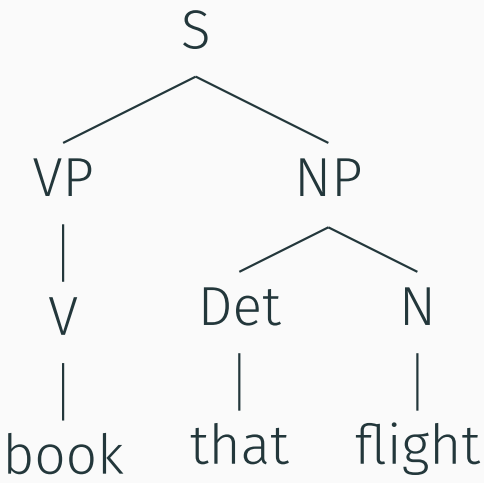
Terminale Symbole

$V_T = \{book, that, flight\}$

Ersetzungsregeln

$$R = \left\{ \begin{array}{r} S \rightarrow VP\ NP \\ VP \rightarrow V \\ NP \rightarrow Det\ N \\ Det \rightarrow that \\ N \rightarrow flight \\ N \rightarrow book \\ V \rightarrow book \end{array} \right\}$$

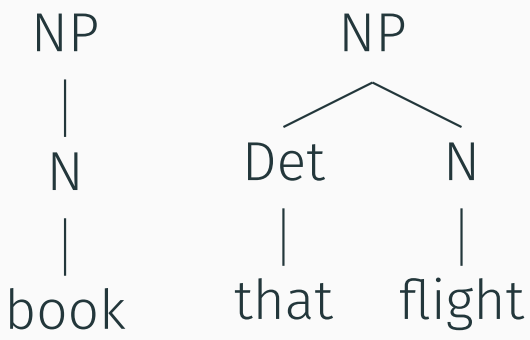
TOP-DOWN PARSING



Ersetzungsregeln

$$R = \left\{ \begin{array}{r} S \rightarrow VP\ NP \\ VP \rightarrow V \\ NP \rightarrow Det\ N \\ Det \rightarrow that \\ N \rightarrow flight \\ N \rightarrow book \\ V \rightarrow book \end{array} \right\}$$

BOTTOM-UP PARSING



- Bottom-Up-Parsing steht also vor einem Problem
- Lösung: es wird mit einem Backtracking-Algorithmus zum letzten möglichen Knoten zurückgekehrt und dort die nächste Möglichkeit gewählt

ZUSAMMENFASSUNG SYNTAX

- Parsing kann Struktur sprachlicher Ausdrücke analysieren und ihre Grammatikalität bewerten
- Parsing kann allerdings keine Aussagen über Bedeutung treffen

SEMANTIK

WAS IST SEMANTIK?

- Teildisziplin der Linguistik, die sich mit der **Bedeutung** sprachlicher Ausdrücke beschäftigt
- Analyse der Prozesse, die Rezipienten eines Ausdrucks in die Lage versetzen, die Äußerung mit Sachverhalten der realen Welt in Verbindung zu setzen
- beschäftigt sich nur mit **literaler Bedeutung**

Beispiel

"Es zieht" hat im semantischen Sinne die Bedeutung, das Durchzug herrscht. Die darin möglicherweise enthaltene Aufforderung, das Fenster zu schließen, ist nicht Teil der Semantik.

- **Computersemantik** versucht, semantische Analysen algorithmisch umzusetzen, also **maschinelle Bedeutungsbestimmung**

- *Annahmen:*
  - Bedeutung eines Satzes ist bekannt, wenn man die Bedingungen weiß, unter denen er wahr ist
  - Kompositionalitätsprinzip: Bedeutung eines komplexen sprachlichen Ausdruck ergibt sich aus den Bedeutungen seiner Teile und der Art ihrer Kombination
- *Konsequenzen:*
  - keine semantische Analyse ohne strukturelle Analyse
  - Bedeutung ergibt sich allein aus Ausdruck selbst, nicht aus dem Kontext

MONTAGUE-SEMANTIK

- die per Syntaxanalyse generierte Struktur lässt sich nun mit formaler Logik semantisch analysieren, d.h. die Bedeutung wird mit Hilfe von logischen Operatoren und Quantoren abgebildet
- das formale Modell, an dem wir uns orientieren werden ist die Montague-Semantik
- die Montague-Semantik zeigt auf, wie Syntax und Semantik mithilfe der Methoden der mathematischen Logik systematisch verbunden werden können

(3) Lisa singt.

- der Satz hat intuitiv die Bedeutung, das ein Individuum, welches mit Lisa bezeichnet ist die Eigenschaft hat, zu singen
- der Satz ist eine Eigenschaftszuschreibung, eine **Prädikation**

- *Frage:* wie lässt sich die Bedeutung des Satzes formal abbilden?
- da das Kompositionalitätsprinzip erfüllt sein soll, berechnet sich die Bedeutung des Satzes also aus der *VP*-Bedeutung und der *NP*-Bedeutung und der Art ihrer Verknüpfung

VP-Bedeutung

Ein Verb drückt eine Eigenschaft aus, welches in Kombination mit dem Subjekt eine Satzbedeutung ergibt. So kann eine VP als Funktion analysiert werden.

NP-Bedeutung

Eigennamen referieren auf Individuen, d.h. “Lisa” bezeichnet das Individuum namens *Lisa*. So kann eine NP als Argument einer Funktion analysiert werden.

Funktionsapplikation

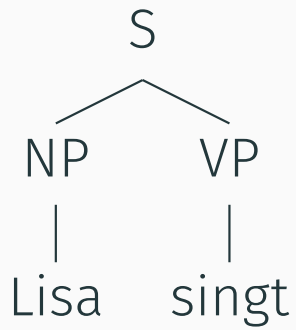
Die Bedeutung des gesamten Satzes lässt sich nun als Andwendung des Arguments in Form der NP auf die Funktion in Form der VP analysieren.

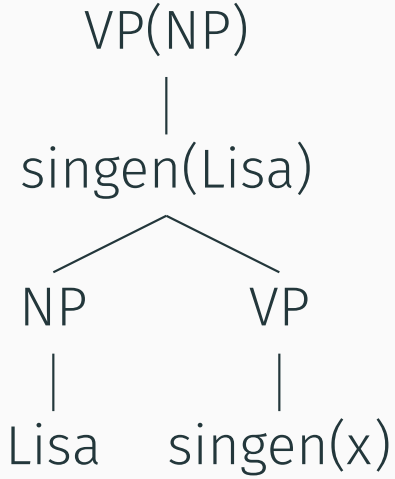
Beispiel

“Lisa singt” $\rightarrow$ singen(Lisa)[1]

[1] Die Funktion würde eigenlich als Lambda-Kalkül abgebildet werden, ist hier aber aus praktischen Gründen schon vereinfacht dargestellt

SYNTAX VS. SEMANTIK





- die Semantik hat also aus der strukturellen Analyse der Syntax eine Funktion gebildet, welche die Bedeutung des Satzes für den Computer verständlich abbildet

AUSBLICK

- der Computer hat nun also Möglichkeiten zur Verfügung, Struktur und Bedeutung natürlicher Sprache zu analysieren
- daraus gewonnen Erkenntnisse können weiter verarbeitet werden
- zB weiß der Computer noch nicht, was *singen* oder *Lisa* genau ist
- zu diesem Zwecke können die Ergebnisse mit Hintergrundwissen in Form von Ontologien / semantischen Netzen verrechnet werden, um weitere Erkenntnisse zu erlangen oder semantische Ambiguitäten aufzulösen

SCHLUSSBETRACHTUNG

Statistische Computerlinguistik

- Statistische Systeme recyclen menschenübersetzte Texte
- Ambiguitäten können statistisch ausgebessert werden (*wooden table* vs. *mathematical table*)

Semantische Netze

- der Computer weiß nicht, was *singen* für eine Tätigkeit ist oder was *Lisa* für Eigenschaften hat
- zu diesem Zwecke können die Ergebnisse mit Hintergrundwissen in Form von Ontologien / semantischen Netzen verrechnet werden, um weitere Erkenntnisse zu erlangen oder semantische Ambiguitäten aufzulösen

FRAGEN

Fragen?

LITERATUR

- H. Langer (Hrsg.): *Computerlinguistik und Sprachtechnologie - Eine Einführung*, Heidelberg 2010
- S. Löbner: *Semantik*, Berlin 2003
- H. Lobin: *Computerlinguistik und Texttechnologie*, Paderborn 2010
- S. Naumann, H.Langer: *Parsing*, Stuttgart 1994
- N. Chomsky: *Syntactic Stuctures*, Paris 1957