Mathematische Grundlagen III

Motivation statistischer Ansätze

Garance PARIS

18. April 2011

Natürliche (menschliche) Sprache

Zitat

...language is **a biological system**, *and biological systems typically are* ***"messy"***, *intricate,* **the result of evolution***ary "tinkering", and* **shaped by accidental circumstances** *and by ... conditions that hold of complex systems...*

(Chomsky, *The minimalist program*)

Natürliche (menschliche) Sprache

Aus dem Verbmobil-Korpus

Spontan-sprachliche Terminabsprache
Deutsch-Englisch-Japanisch:

...bei mir ist die Woche davor schlecht, ***also****, die Woche nach Pfingsten****, und*** *die erste Maiwoche,* ***also,*** *alles andere* ***wäre stünde*** *zur Disposition, dann würde ich mal sagen, daß wir den ersten Termin auf Montag, den neunten Mai legen...*

Kompetenz und Performanz

Kompetenz

- Potenzielle, idealistische (angeborene) Fähigkeit zur Sprache bzw. Wissen um die Sprache
- Endliche Menge von Sprachregeln, die Sprecher verinnerlicht haben und die zum Verstehen und Produzieren von Sprache dienen
- Beschreibt die wohlgeformten Äußerungen einer Sprache
- Kann man nicht direkt beobachten

Kompetenz und Performanz

Performanz

- Anwendung der zur Kompetenz gehörenden Regeln
- Tatsächlich vorkommende Äußerungen
- Zu beobachtendes Verhalten

Zwei Ansätze

The Armchair Linguist

He sits in a deep soft comfortable armchair, with his eyes closed and his hands clasped behind his head. Once in a while he opens his eyes, sits up abruptly shouting "Wow, what a neat fact", grabs his pencil, and writes something down. Then he paces around for a few hours in the excitement of having come still closer to knowing what language is really like.

(Charles Fillmore)

Zwei Ansätze

The Corpus Linguist

He has all the primary facts that he needs, in the form of a corpus of approximately one zillion running words, and he sees his job as that of deriving secondary facts from his primary facts. At the moment he is busy determining the relative frequencies of the eleven parts of speech as the first words of a sentence versus as the second word of a sentence.

(Charles Fillmore)

Regelbasierte Modelle

- Modellierung durch theoretische Überlegung
- Gesucht werden Regeln,
  - die alle Fälle eines Phänomens erfassen,
    aber nicht übergenerieren
  - die einfach genug sind, um von einem Computer
    berechnet zu werden
    (kein Rückgriff auf Weltwissen usw.)

Regelbasierte Modelle

Bsp.: Woran erkenne ich ein Adjektiv?

Wir suchen ein Regelsystem, das für (möglichst) alle Adjektive wahr und für alle anderen Wörter falsch ist

Regelbasierte Modelle

Bsp.: Woran erkenne ich ein Adjektiv?

Wir suchen ein Regelsystem, das für (möglichst) alle Adjektive wahr und für alle anderen Wörter falsch ist

1. Nächstes Wort kapitalisiert: Adj
   Sonst: NAdj

Regelbasierte Modelle

Bsp.: Woran erkenne ich ein Adjektiv?

Wir suchen ein Regelsystem, das für (möglichst) alle Adjektive wahr und für alle anderen Wörter falsch ist

1. Nächstes Wort kapitalisiert: Adj
   Sonst: NAdj
2. Nächstes Wort kapitalisiert und Wort kein Artikel: Adj
   Sonst NAdj

Regelbasierte Modelle

Bsp.: Woran erkenne ich ein Adjektiv?

Wir suchen ein Regelsystem, das für (möglichst) alle Adjektive wahr und für alle anderen Wörter falsch ist

1. Nächstes Wort kapitalisiert: Adj
   Sonst: NAdj
2. Nächstes Wort kapitalisiert und Wort kein Artikel: Adj
   Sonst NAdj
3. Nächstes Wort kapitalisiert und Wort kein Artikel und vorheriges Wort Artikel: Adj
   Sonst NAdj

Regelbasierte Modelle

Bsp.: Woran erkenne ich ein Adjektiv?

*Ich möchte Ihnen für den Bericht über den **siebenten** Bericht über **staatliche** Beihilfen in der **europäischen** Union danken.*

(European Parliament Proceedings)

Es ist schwer, korrekte und vollständige Regeln zu schreiben

- Regel 2 ist zu liberal (möchte = Adj)
- Regel 3 ist zu streng (staatliche = NAdj)
- Das System trifft eine harte Entscheidung für jede Instanz
- Keine Möglichkeit, über "Wahrscheinlichkeit" zu sprechen

Regelbasierte Modelle

- Erfolgreich für Morphologie, Grammatiken (Grammatiktheorie), formale semantische Analyse
- Vorteile
  - Erlaubt Modellierung komplexer Phänomene ("tiefe" Analyse)
  - Kann negative Evidenz einbeziehen (=Was **nicht** möglich ist)
  - Ergebnis ist für Menschen verständlich
  - Bietet oft eine Erklärung des Phänomens an

Regelbasierte Modelle

Nachteile regelbasierter Systeme:

- Nicht geeignet für stetige Phänomene
- Können keine Präferenzen ausdrücken
- Häufig präskriptiv statt deskriptiv
- Mangel an Robustheit: Schon bei kleinen Fehlern in der Eingabe bricht die Analyse ab
- Objektivität?
- Hand-Arbeit: Hoher Aufwand
  Die "English Resource Grammar" (ERG) wird seit Mitte der 90er Jahre in mehreren großen CL-Projekten entwickelt, aber es wird noch daran gearbeitet!

Korpuslinguistik und Statistik

- Daten-orientierte Untersuchungen:
  Modellierung durch Sichtung von Beispielen
- Erkennung ähnlicher Muster und
  Regelmäßigkeiten in den Daten
- Vorteile
  - Auf Grund von Daten trainiert: Weniger Handarbeit
    (Einsatz maschineller Lernverfahren)
  - Bestimmung der wahrscheinlichsten Lesart
  - Robust: Können mit fehlerhafter oder unbekannter
    Eingabe umgehen
  - Modelle können Übergenerierung erlauben, um
    Robustheit zu erreichen
  - Zugriff auf in den Daten implizites Weltwissen
  - Schnelle Modellierung neuer Domänen, Sprachen, usw.

Korpuslinguistik und Statistik

Einige Beispiele

- Lexikalische Präferenzen
  - Wortkategorie: *bank* = Substantiv 85 %, Verb 15 %
  - Bedeutung: *bank* (river) = 22 %, *bank* (money) = 78 %
- Syntax:
  - realized + NP = 20 %
  - realized + S = 65 %
  - realized + other = 15 %
- Anaphern: *He* bezieht sich auf Englisch in 63 % der Fälle auf das Subjekt des vorigen Satzes
- Textanalyse: Autor X verwendet das Wort *bezüglich* “signifikant” öfter als Autor Y

Korpuslinguistik und Statistik

- Nachteile
  - Flache Analyse (Engl. „shallow")
  - Modelle nur approximativ richtig
  - Schwierige Probleme können oft nicht zuverlässig modelliert werden
  - Modelle für Menschen schwierig zu verstehen und abzuändern
  - Rein descriptiv, keine Erklärung
  - Abhängigkeit von den Daten
  - Problem mit unbekannten Wörtern/Strukturen (Sparse Data)
- Erfolgreich für:
  - Wortartenanalyse
  - Automatische syntaktische Analyse

Kleine Methodengeschichte der Computerlinguistik

- 1950er-1980er: Theoretische Linguistik
  - Linguistische Grundlagenarbeit (Grammatiktheorien)
  - Weniger Fokus auf praktische Anwendung
- Seit 1990: Verwaltung riesiger Datenmengen wird als zentrale Aufgabe der CL erkannt
  - Maschinelles Lernen als zentrale Methode
  - Hoffnung:
    "Jedes Problem lässt sich durch genügend Daten lösen"
    (Heute zum Teil enttäuscht)

"Every time I fire a linguist, the performance of our speech recognition system goes up." (F. Jelinek, 1988)

Parallel mit Nativismus vs. Empirie

Welche Rolle spielt Spracherfahrung beim Sprachenlernen?

- Nativismus: Sprache ist sehr komplex, daher muss die Fähigkeit dazu und deren Grundprinzipien beim Menschen angeboren sein

  (Vgl. Chomsky's *Principles and Parameters*:
  - Sowohl Prinzipien als auch Parameter sind Sprachuniversalien
  - Menschen kennen die Prinzipien von Geburt an, z. B. dass alle Sätze ein Subjekt haben, auch wenn es in manchen Sprachen overt (=sichtbar) weggelassen werden kann
  - Spracherwerb besteht darin, die Parameter für die eigene Muttersprache zu setzen: SVO oder OVS? usw.)

- Empirizismus: Sprachliches Wissen erwerben Kinder ausschließlich durch das Hören der Sprache ihrer Eltern

Zusammenfassung

- Kompetenz und Performanz
- Stärken und Schwächen regelbasierter Systeme
- Stärken und Schwächen statistischer Systeme
- Unterschiede im Ansatz
- Parallel mit Theorien über Spracherwerb