

Zwei methodologische Alternativen in der Computerlinguistik/Sprachtechnologie

Regelbasierte Verfahren vs. Datengesteuerte Verfahren

Regelbasiert: auf Grund von Regelsystemen über den Aufbau einer natürlichen Sprache werden Sätze (Texte) analysiert, bis man zu einer Bedeutungsbeschreibung kommt.

Morphologische Prozesse:

- Lemmatisierung (Flexionsform → Grundform)
- Wortklassenbestimmung
- morphosyntaktische Merkmale (Genus, Numerus, Tempus, Kasus, Person, ...)
- Komposita-Analyse

Die Mäuse fürchteten sich vor der bösen Katze.

D-	Maus	fürchten	sich	vor	d-	böse	Katze	.
Art	N f	Vrb	Prn	Prp	Art	Adj	N f	Dot
Pl	Nom Pl	Imp	Akk	Sg	Dat	Dat	Sg	

Syntaktische Prozesse:

- Regeln für Phrasen,
S → NP + VP; NP → Art + N; PP → Prp + NP
- Strukturbaum als Ergebnis

Semantische Prozesse:

$\exists e, f, x \text{ ISA } (e, \text{Telling}) \wedge \text{ISA } (f, \text{Going})$
 $\wedge \text{Teller } (e, \text{Speaker}) \wedge \text{Tellee } (e, \text{Harry})$
 $\wedge \text{Toldthing } (e, f) \wedge \text{Goer } (f, \text{Harry})$
 $\wedge \text{Destination } (f, x)$

Pragmatische Prozesse:

- Sprecher, Hörer, Ort, Zeit ...
- Absicht der Äußerung
- Mögliche Reaktion (verbal oder nonverbal)

Datengesteuerte Verfahren:

Riesige Textkorpora werden herangezogen, die Realwelttexte enthalten (im Gegensatz zu linguistischen Beispielsätzen).

Hieraus werden Regelmäßigkeiten extrahiert, die zu Annotationen mit ausreichender Qualität führen.

Oft werden Lernverfahren eingesetzt:

Ein kleiner Teil der Korpora wird von Menschen annotiert. Auf dieser Grundlage werden Merkmale extrahiert, so dass das Restkorpus automatisch hinreichend gut mit semantischen Kategorien ausgezeichnet werden kann (> als 80% richtige Zuordnungen).

Beispiel:

Eigennamenerkennung (Named Entity Recognition, NER)

- **Personennamen (Peter Berlin)**
- **Ortsnamen (Berlin ist eine Hauptstadt)**
- **Institutionen (Berlin Kulinarisch e.V.)**