Sprachorientierte KI: Syntax und Parsing

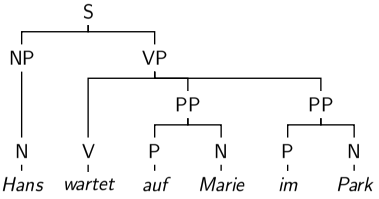
- Syntax als Untersuchungsgegenstand
- Wortartendisambiguierung
- Phrasenstrukturgrammatiken
- Parsing mit Phrasenstrukturgrammatiken
- Restringierte Phrasenstrukturgrammatiken
- Unifikationsgrammatiken
- Constraint-basierte Grammatiken
- Robustes Parsing

Wortartendisambiguierung

- Lexikalische Kategorien
- Regelbasierte Tagger
- Stochastische Tagger
- Transformationsbasierte Tagger
- Anwendungen

Lexikalische Kategorien

- Ziel: syntaktische Regeln sollen generalisierbare Zusammenhänge beschreiben
- Zusammenfassen der terminalen Symbole zu Klassen mit äquivalentem syntaktischen Verhalten.
  $\sim$ Wortarten



Lexikalische Kategorien

- Kriterien für Wortkategorien (Radford 1988)
  - Phonologische Evidenz: Erklärung systematischer Aussprachevarianten

    *We need to in**crease** productivity.*
    *We need an **in**crease in productivity.*
    *Why do you tor**ment** me?*
    *Why do you leave me in **tor**ment?*
    *We might trans**fer** him to another club.*
    *He's asked for a **trans**fer.*

Lexikalische Kategorien

- Kriterien für Wortkategorien (2)
  - Semantische Evidenz: Erklärung struktureller Ambiguitäten

    Mistrust wounds.
    ..., wo die wilden tiere jagen.
    Er hat liebe genossen.

    Semantische Eigenschaften sind irrelevant:

Verben	Aktionen	laufen, tragen, lachen, ...
Nomen	Objekte	Tisch, Pferd, Hans, ...
Adjektive	Zustände	krank, glücklich, ...
...	...	...

Lexikalische Kategorien

- Kriterien für Wortkategorien (3)
  - Morphologische Evidenz
    - verschiedene Flexionsmuster für Verben, Substantive, Adjektive
      - aber: irreguläre Flexion: starke Verben, *ist*
    - unterschiedliche Wortbildungsmuster
      - Steigerungsformen für Adjektive
      - Deverbalisierung: -ung
      - Denominalisierung: -ier, -eln (?)
      - Deadjektivierung: -heit, -keit
      - keine Ableitungen für Präpositionen und Hilfsverben

Lexikalische Kategorien

- Kriterien für Wortkategorien (4)
  - Syntaktische Evidenz: Distributionsklassen
    - Nomen

      Linguistics can be a pain in the neck.
      John can be a pain in the neck.
      Girls can be a pain in the neck.
      Television can be a pain in the neck.
      ** Went can be a pain in the neck.*
      ** For can be a pain in the neck.*
      ** Older can be a pain in the neck.*
      ** Conscientiously can be a pain in the neck.*
      ** The can be a pain in the neck.*

Lexikalische Kategorien

- Kriterien für Wortkategorien (4)
  - Distributionsklassen (2)
    - Im Deutschen: Abstraktion von flexivischen Anpassungen erforderlich
    - analoge Analysen für Verben, Modalverben, Adjektive, Präpositionen, ...

Lexikalische Kategorien

- Inventar lexikalischer Kategorien:

N	Nomen	Haus, Hund, Lehrer, ...
V	Verb	suchen, fragen, werden, sein, ...
P	Präposition	auf, unter, zwischen, nach, ...
A	Adjektiv	schön, gut, rot, ...
ADV	Adverb	abends, anders, vielleicht, ganz, ...
M	Modalverben	wollen, dürfen, sollen, ...
D	Determiner	der, diese, ihr, alle, genug, ...

Lexikalische Kategorien

- Distributionsanalyse lässt Gestaltungsspielraum
  - Engl.: Partikel und Konjunktionen als Präpositionen
  - Engl.: Adjektive und Adverbien als positionelle Varianten der gleichen Kategorie
    - Adjektive modifizieren Nomen
    - Adverbien modifizieren Adjektive, Adverbien, Präpositionen und Verben
      - *There is a real crisis.*
      - *He is really nice.*
      - *He walks really slowly.*
      - *He is really down.*
      - *He must really squirm.*

Lexikalische Kategorien

- Hauptkategorien (major categories): N, V, A, P
  - Merkmalsrepräsentation für die Hauptkategorien: ±V, ±N

	[V +]	[V –]
[N +]	Adjektiv	Nomen
[N –]	Verb	Präposition

Lexikalische Kategorien

- Ziel: Formulierung kategorienübergreifender Generalisierungen
  - Engl.: nur [N –]-Wortformen erlauben Nominalgruppenkomplemente
    - *John* **loves** *[Mary]* (V + NP)
    - *John bought a present* **for** *[Mary]* (P + NP)
    - ** John's* **admiration** *[Mary]* (N + NP)
    - ** John is* **fond** *[Mary]* (A + NP)
  - Ital.: [N +] flektiert nach dem Genus, [N –] nicht
    - *bravo ragazzo (guter Junge)*
    - *brava ragazza (gutes Mädchen)*
    - *bravi ragazzi (gute Jungen)*
    - *brave ragazze (gute Mädchen)*

Lexikalische Kategorien

- feinere Unterteilung der Verben

[AUX –]	[AUX +]	
	[M +]	[M –]
schlafen	wollen	haben
gehen	können	sein
sagen	dürfen	
...	...	

Lexikalische Kategorien

- offene Wortklassen: produktiv, Neubildungen möglich
  - Nomen, Verben, Adjektive, Adverbien
- geschlossene Wortklassen: relativ fester Bestand, Funkionswörter
  - Präpositionen, Artikel, Pronomen, Konjunktionen, Hilfsverben, Partikel, Zahlwörter

Lexikalische Kategorien

- Tagsets
  - Kategorieninventare zur Annotation von Korpora
  - teilweise auch morphosyntaktische Subkategorisierung
  - "technische" Tags
    - Fremdwörter, Symbole, Interpunktion, ...

Lexikalische Kategorien

- Tagsets

Penn-Treebank	Marcus et al. (1993)	45
British National Corpus (C5)	Garside et al. (1997)	61
British National Corpus (C7)	Leech et al. (1994)	146
Tiger (STTS)	Schiller, Teufel (1995)	54
Prague Treebank	Hajic (1998)	3000/1000

Lexikalische Kategorien

- Penn-Treebank (Marcus, Santorini, Marcinkiewicz 1993)

CC	Coordinating conjunction	*and,but,or, ...*
CD	Cardinal Number	*one, two, three, ...*
DT	Determiner	*a, the*
EX	Existential *there*	*there*
FW	Foreign Word	*a priori*
IN	Preposition or subord. conjunction	*of, in by, ...*
JJ	Adjective	*big, green, ...*
JJR	Adjective, comparative	*bigger, worse*
JJS	Adjective, superlative	*lowest, best*
LS	List Item Marker	*1, 2, One, ...*
MD	Modal	*can, could, might, ...*
NN	Noun, singular or mass	*bed, money, ...*
NNP	Proper Noun, singular	*Mary, Seattle, GM, ...*
NNPS	Proper Noun, plural	*Koreas, Germanies, ...*
NNS	Noun, plural	*monsters, children, ...*

Lexikalische Kategorien

- Penn-Treebank (2)

PDT	Predeterminer	*all, both, ... (of the)*
POS	Possessive Ending	*'s*
PRP	Personal Pronoun	*I, me, you, he, ...*
PRP$	Possessive Pronoun	*my, your, mine, ...*
RB	Adverb	*quite, very, quickly, ...*
RBR	Adverb, comparative	*faster, ...*
RBS	Adverb, superlative	*fastest, ...*
RP	Particle	*up, off, ...*
SYM	Symbol	*+, %, & ...*
TO	*to*	*to*
UH	Interjection	*uh, well, yes, my, ...*
VB	Verb, base form	*write, ...*
VBD	Verb, past tense	*wrote, ...*
VBG	Verb, gerund	*writing*
VBN	Verb, past participle	*written, ...*

Lexikalische Kategorien

- Penn-Treebank (2)

VBP	Verb, non-3rd singular present	*write, ...*
VBZ	Verb, 3rd person singular present	*writes, ...*
WDT	Wh-determiner	e.g. *which, that*
WP	Wh-pronoun	e.g. *what, whom, ...*
WP$	Possessive wh-pronoun	*whose, ...*
WRB	Wh-adverb	e.g. *how, where, why*
$	Dollar sign	*$*
#	Pound sign	*#*
"	left quote	*"*
´´	right quote	*´´*
(	left parantheses	*(*
)	right parantheses	*)*
,	comma	*,*
.	sentence final punct.	*., !, ?*
:	mid-sentence punct.	*:, ;, —, ...*

Lexikalische Kategorien

- Beispiele

  Book/NN/VB that/DT/WDT flight/NN ./.

  Book/VB that/DT flight/NN ./.

Lexikalische Kategorien

- das Stuttgart-Tübinger Tagset (STTS)
  (Schiller und Teufel 1995)

ADJA	attributives Adjektiv	das *große* Haus
ADJD	adverbiales oder prädikatives Adjektiv	er fährt/ist *schnell*
ADV	Adverb	*schon, bald, doch*
APPR	Präposition; Zirkumposition links	*in* der Stadt, *ohne* mich
APPRART	Präposition mit Artikel	*im* Haus, *zur* Sache
APPO	Postposition	ihm *zufolge*, der Sache *wegen*
APZR	Zirkumposition rechts	von jetzt *an*
ART	bestimmter oder unbestimmter Artikel	*der, die, das, ein, eine, ...*
CARD	Kardinalzahl	*zwei* Männer, im Jahre *1994*
FM	Fremdsprachliches Material	Es wird mit *"A big fish"* übersetzt
ITJ	Interjektion	mhm, ach, tja
ORD	Ordinalzahl	[der] neunte [August]

Lexikalische Kategorien

- das Stuttgart-Tübinger Tagset (STTS)(2)

KOUI	unterordn. Konjunktion mit "zu" + Infinitiv	*um/anstatt* zu leben
KOUS	unterordnende Konjunktion mit Satz	*weil, dass, damit, wenn, ob*
KON	nebenordnende Konjunktion	*und, oder, aber*
KOKOM	Vergleichskonjunktion	*als, wie*
NN	normales Nomen	*Tisch, Herr*, das *Reisen*
NE	Eigennamen	*Hans, Hamburg, HSV*
PDS	substituierendes Demonstrativpronomen	*dieser, jener*
PDAT	attribuierendes Demonstrativpronomen	*jener* Mensch
PIS	substituierendes Indefinitpronomen	*keiner, viele, man, niemand*
PIAT	attrib. Indefinitpron. ohne Determiner	*kein/irgendein* Mensch,
PIDAT	attrib. Indefinitpron. mit Determiner	ein *wenig* Bier, *beide* Brüder
PPER	irreflexives Personalpronomen	*ich, er, ihm, mich, dir*
PPOSS	substituierendes Possessivpronomen	*meins, deiner*
PPOSAT	attribuierendes Possessivpronomen	*mein* Buch, *deine* Mutter

Lexikalische Kategorien

- das Stuttgart-Tübinger Tagset (STTS)(3)

PRELS	substituierendes Relativpronomen	der Hund, *der*
PRELAT	attribuierendes Relativpronomen	der Mann, *dessen* Hund
PRF	reflexives Personalpronomen	*sich, einander, dich, mir*
PWS	substituierendes Interrogativpronomen	*wer, was*
PWAT	attribuierendes Interrogativpronomen	*welche* Farbe, *wessen* Hut
PWAV	adverbiales Interrogativoder Relativpronomen	*warum, wo, wann, worüber*
PAV	Pronominaladverb	*dafür, deswegen, trotzdem*
PTKZU	"zu" vor Infinitiv	*zu* gehen
PTKNEG	Negationspartikel	*nicht*
PTKVZ	abgetrennter Verbzusatz	er kommt *an*, er fährt *rad*
PTKANT	Antwortpartikel	*ja, nein, danke, bitte*
PTKA	Partikel bei Adjektiv oder Adverb	*am* schönsten, *zu* schnell
SGML	SGML Markup	`<turnid=n022k_TS2004>`
SPELL	Buchstabierfolge	*S-C-H-W-E-I-K-L*

Lexikalische Kategorien

- das Stuttgart-Tübinger Tagset (STTS)(4)

TRUNC	Kompositions-Erstglied	*An-* und Abreise
VVFIN	finites Verb, voll	du *gehst*, wir *kommen* [an]
VVIMP	Imperativ, voll	*komm* !
VVINF	Infinitiv, voll	*gehen, ankommen*
VVIZU	Infinitiv mit "zu", voll	*anzukommen, loszulassen*
VVPP	Partizip Perfekt, voll	*gegangen, angekommen*
VAFIN	finites Verb, aux	du *bist*, wir *werden*
VAIMP	Imperativ, aux	*sei* ruhig !
VAINF	Infinitiv, aux	*werden, sein*
VAPP	Partizip Perfekt, aux	*gewesen*
VMFIN	finites Verb, modal	*dürfen*
VMINF	Infinitiv, modal	*wollen*
VMPP	Partizip Perfekt, modal	*gekonnt*, er hat gehen *können*
XY	Nichtwort, Sonderzeichen enthaltend	*3:7, H2O, D2XW3*

Lexikalische Kategorien

- das Stuttgart-Tübinger Tagset (STTS)(5)

$,	Komma	,
$.	Satzbeendende Interpunktion	. ? ! ; :
$(	sonstige Satzzeichen; satzintern	- [ ] ( )

- Beispiele (Tiger-Korpus)

  Werden/VAFIN sie/PPER diesmal/ADV lachen/VVINF //$.
  kreischen/VVINF ?$.

  Mehr/PIAT Zeit/NN wenden/VVFIN die/ART
  US-Bürger/NN nur/ADV für/APPR Arbeiten/NN und/KON
  Schlafen/NN auf/PTKVZ ./$.

Regelbasierte Tagger

- ENGTWOL, Universität Helsinki (Voutilainen 1995)
- zweistufiger Ansatz
  - Zuweisung von Wortarthypothesen
  - Selektion von Wortarthypothesen
- reichhaltiges Lexikon mit morphosyntaktischen Merkmalen
- zwei Stufen
  - Morphologische Analyse
  - Morphologische Disambiguierung

Regelbasierte Tagger

- Morphologischer Analysator
  - Zweiebenen-Morphologie
  - reichhaltige morphosyntaktische Information verfügbar

```
("<round>"
  ("round" <SVO><SV> V SUBJUNCTIVE VFIN (@+FMAINV))
  ("round" <SVO><SV> V IMP VFIN (@+FMAINV))
  ("round" <SVO><SV> V INF)
  ("round" <SVO><SV> V PRES -SG3 VFIN (@+FMAINV))
  ("round" PREP)
  ("round" N NOM SG)
  ("round" A ABS)
  ("round" ADV ADVL (@ADVL)))
```

Regelbasierte Tagger

- 35-45% der Token sind mehrdeutig: 1.7-2.2 Analysealternativen pro Wortform
- Hypothesenselektion mit Constraints (1100)
  - lineare Abfolge von morphologischen Merkmalen
- Beispiel
  - Eingabesatz: *a reaction* **to** *the ringing of a bell*
  - Lexikoneintrag:

```
("<to>"
  ("to" PREP)
  ("to" INFMARK> (@INFMARK>))
```

Regelbasierte Tagger

- Beispiel
  - Constraint

```
("<to>" =0 (INFMARK>) (NOT 1 INF)
                      (NOT 1 ADV)
                      (NOT 1 QUOTE)
                      (NOT 1 EITHER)
                      (NOT 1 SENT-LIM))
```

  Streiche die Infinitivlesart, wenn unmittelbar rechts von *to* kein Infinitiv, Adverb, Zitat, *either, neither, both* oder Satzende vorkommt.

Regelbasierte Tagger

- Qualitätsmessung
  - Messung auf einem annotierten Testset ("gold standard")

$$\text{Recall} = \frac{\text{ermittelte korrekte Kategorien}}{\text{gesuchte korrekte Kategorien}}$$

  - Recall $<$ 100%: fehlerhafte Klassifikationen

$$\text{Precision} = \frac{\text{ermittelte korrekte Kategorien}}{\text{Gesamtzahl der ermittelten Kategorien}}$$

  - Recall $<$ Precision: unvollständige Kategorisierung
  - Recall $=$ Precision: vollständig disambiguierter Output
    $\longrightarrow$ accuracy, Genauigkeit
  - Recall $>$ Precision: unvollständige Disambiguierung

Regelbasierte Tagger

- ENGTWOL:
  - Testset: 2167 Wortformtoken
  - Recall: 99.77 %
  - Precision: 95.94 %

  $\longrightarrow$ unvollständige Disambiguierung

Regelbasierte Tagger

- Wie gut sind die Ergebnisse?
  1. Obere Schranke: Wie gut ist die Annotation?
     - 96-97% Übereinstimmung zwischen Annotatoren (Marcus et al. 1993)
     - fast 100% Übereinstimmung bei gegenseitiger Abstimmung (Voutilainen 1995)
     - für deutsche Texte 98.6% (Brants 2000)
  2. Untere Schranke: Wie gut ist die Klassifikation?
     - Baseline:
       z.B. häufigster Tag (Unigram-Wahrscheinlichkeit)
     - Beispiel: $\boldsymbol{P}(\text{NN}|\textit{race}) = \mathbf{0.98}$ $\boldsymbol{P}(\text{VB}|\textit{race}) = \mathbf{0.02}$
     - 90-91% Precision/Recall (Charniak et al. 1993)

Stochastische Tagger

- manuelle Erstellung von Regelsystemen
  - aufwendig
  - fehleranfällig
- Ausweg: trainierbare Verarbeitungskomponenten
  - freie Parameter eines Modells werden aufgrund von Beobachtungsdaten eingestellt
  - Maschinelles Lernen

Stochastische Tagger

- zwei grundlegende Klassen von Lernverfahren
  - überwachtes Lernen (Lernen mit Lehrer)
  - Ausgangspunkt: Beispielsammlung
    - Paare aus Eingabedaten und gewünschten Verarbeitungsresultaten
    - annotierte Korpora
  - unüberwachtes Lernen (Lernen ohne Lehrer)
    - Lernprozess extrahiert eigenständig regelhafte Strukturen aus den Daten
    - Beispiel: Clustern von Daten aufgrund inhärenter Regelmäßigkeiten

Stochastische Tagger

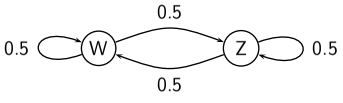
- stochastische Lernverfahren:
  - Schätzen von Wahrscheinlichkeiten in einem stochastischen Modell
- stochastisches Modell versucht die Entstehung der Beobachtungsdaten durch einen stochastischen Prozess zu beschreiben
  - Modellvorstellung: gestörter Kanal

Stochastische Tagger

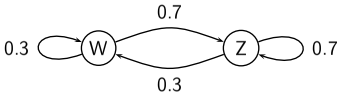
- Kanalmodell
  - Zuordnung von Tags zu Wortformen ist nicht deterministisch
  - "Störung" im Kanal ist vom Kontext abhängig
  - Modell mit Gedächtnis: Markov-Modell
  - Gedächtnis wird durch Zustände modelliert
  - Modellparameter beschreiben die Wahrscheinlichkeit eines Zustandsübergangs
    - Transitionswahrscheinlichkeiten: $P(s_i|s_1 \ldots s_{i-1})$

Stochastische Tagger

- Markov-Modell einer Münze



- Markov-Modell einer gezinkten Münze

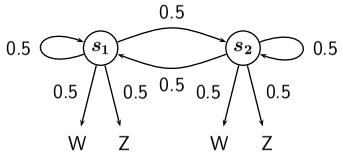


Stochastische Tagger

- Hidden-Markov-Modelle
  - Beobachtungen sind nicht mehr strikt an Zustände gekoppelt
  - Zustandsfolge beeinflusst die Beobachtungsfolge nur stochastisch
    - Emissionswahrscheinlichkeiten: $P(o_i|s_1 \ldots s_{i-1})$

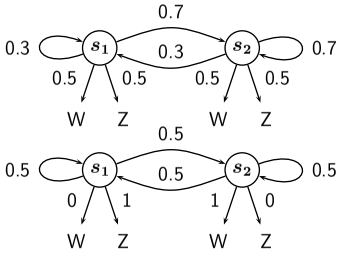
Stochastische Tagger

- Hidden-Markov-Modelle
  - für einen externen Beobachter ist die Zustandsfolge nicht unmittelbar feststellbar



Stochastische Tagger

- verschiedene HMMs können die gleichen Beobachtungsdaten beschreiben



- ⟶ große Flexibilität beim Schätzen der Parameter

Stochastische Tagger

- Vorgehen:
  - Entwurf: Wahl einer geeigneten Modellstruktur: Zustände und Wahrscheinlichkeiten
  - Training: Schätzen der Wahrscheinlichkeiten auf den Beobachtungsdaten (Lernstichprobe)
  - Klassifikation: Ermittlung der wahrscheinlichsten Zustandsfolge
    - durch welche Zustandsfolge hat das Modell die gegebene Beobachtung vermutlich erzeugt?
  - Evaluation: Messen der Verarbeitungsqualität auf einer separaten Teststichprobe

Stochastische Tagger

- Modellstruktur eines HMM-Taggers
  - Beobachtung: Wortformen w_i
  - Zustände: Tags t_i
  - Transitionswahrscheinlichkeiten: $P(t_i|t_1 \dots t_{i-1})$
  - Emissionswahrscheinlichkeiten: $P(w_i|t_1 \dots t_{i-1})$

Stochastische Tagger

- Klassifikation: Auswahl eines Tags t_i, so dass

$$t_i = \arg\max_i P(t_i|t_{i-1} \dots t_1, w_i)$$

- Ermitteln der wahrscheinlichsten Tagsequenz

$$t_i[1,n] = \arg\max_{t[1,n]} P(t[1,n]|w[1,n])$$

- Bayessche Regel

$$t_i[1,n] = \arg\max_{t[1,n]} \frac{P(t[1,n]) \cdot P(w[1,n]|t[1,n])}{p(w[1,n])}$$

- Wahrscheinlichkeit der Wortformenfolge ist für gegebene Beobachtung konstant und beeinflusst daher die Auswahl der Tagfolge nicht

Stochastische Tagger

- Kettenregel für Wahrscheinlichkeiten

$$P(t[1,n]) \cdot P(w[1,n] \mid t[1,n])$$

$$= \prod_{i=1}^{n} P(t_i \mid w_1 t_1 \dots w_{i-1} t_{i-1}) \cdot P(w_i \mid w_1 t_1 \dots w_{i-1} t_{i-1} t_i)$$

$$t_i[1,n] = \arg\max_{t[1,n]} \prod_{i=1}^{n} P(t_i \mid w_1 t_1 \dots w_{i-1} t_{i-1}) \cdot P(w_i \mid w_1 t_1 \dots w_{i-1} t_{i-1} t_i)$$

Stochastische Tagger

- 1. vereinfachende Annahme: Wortform ist nur vom aktuellen Tag abhängig

$$t_i[1,n] = \arg\max_{t[1,n]} \prod_{i=1}^{n} P(t_i \mid w_1 t_1 \dots w_{i-1} t_{i-1}) \cdot P(w_i \mid t_i)$$

- 2. vereinfachende Annahme: aktuelles Tag ist nur von seinen Vorgängern (nicht den Wortformen!) abhängig

$$t_i[1,n] = \arg\max_{t[1,n]} \prod_{i=1}^{n} P(t_i \mid t_1 \dots t_{i-1}) \cdot P(w_i \mid t_i)$$

Stochastische Tagger

- 3. vereinfachende Annahme: aktuelles Tag ist nur von den zwei unmittelbaren Vorgängern abhängig
  - begrenztes Gedächtnis (Markov-Annahme): Trigram-Modell

$$t_i[1,n] = \arg\max_{t[1,n]} \prod_{i=1}^{n} P(t_i \mid t_{i-1} t_{i-2}) \cdot P(w_i \mid t_i)$$

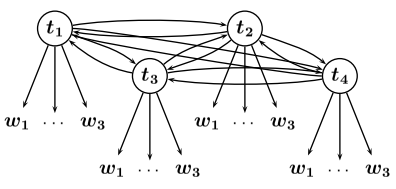
→ Markov-Prozess 2. Ordnung

Stochastische Tagger

- weitere Vereinfachung zum Bigram-Modell
  - stochastische Abhängigkeit ist auf den unmittelbaren Vorgänger beschränkt

$$t_i[1,n] = \arg\max_{t[1,n]} \prod_{i=1}^{n} P(t_i \mid t_{i-1}) \cdot P(w_i \mid t_i)$$

→ Markov-Prozess 1. Ordnung



Stochastische Tagger

- iterative Berechnungsvorschrift ist nicht praktikabel

$$|T[1,n]| = |T|^n$$

$T[1,n]$ Menge aller Tagsequenzen, T Tagset, n Satzlänge

- rekursive Reformulierung: Viterbi, Bellmann-Ford Prinzip der "dynamischen Programmierung"

$$\alpha_n = \max_{t[1,n]} \prod_{i=1}^{n} P(t_i \mid t_{i-1}) \cdot P(w_i \mid t_i)$$

$$\alpha_n = \max_{t_{n-1}} P(t_n \mid t_{n-1}) \cdot P(w_n \mid t_n) \cdot \alpha_{n-1}$$

- Scores können als Konfidenzabschätzung ausgegeben werden

Stochastische Tagger

- Schätzen der Wahrscheinlichkeiten:
  - Transitionswahrscheinlichkeiten

$$P(t_i \mid t_{i-2}t_{i-1}) = \frac{c(t_{i-2}t_{i-1}t_i)}{c(t_{i-2}t_{i-1})}$$

  - Emissionswahrscheinlichkeiten

$$P(w_i \mid t_i) = \frac{c(w_i, t_i)}{c(t_i)}$$

Stochastische Tagger

- ungesehene Übergangswahrscheinlichkeiten
  - Backoff: Rückgriff auf Bigram- bzw. Unigram-Wahrscheinlichkeiten

$$P(t_i|t_{i-2}t_{i-1}) = \begin{cases} P(t_i|t_{i-2}t_{i-1}) & \text{falls } c(t_{i-2}t_{i-1}t_i) > 0 \\ P(t_i|t_{i-1}) & \text{falls } c(t_{i-2}t_{i-1}t_i) = 0 \text{ und } c(t_{i-1}t_i) > 0 \\ P(t_i) & \text{sonst} \end{cases}$$

Stochastische Tagger

- ungesehene Übergangswahrscheinlichkeiten
  - Interpolation: Mischen der Trigram mit den Bigram- und Unigram-Wahrscheinlichkeiten

$$P(t_i|t_{i-2}t_{i-1}) = \lambda_1 P(t_i|t_{i-2}t_{i-1}) \cdot \lambda_2 P(t_i|t_{i-1}) \cdot \lambda_3 P(t_i)$$

λ_1, λ_2 und λ_3 sind kontextabhängige Parameter und werden auf einer separaten Datenmenge trainiert

Stochastische Tagger

- ungesehene Wortformen
  - Schätzen der Tagwahrscheinlichkeiten aus Suffixen (ggf. auch Präfixen)
- ungesehene Wortartzuordnung
  - Glättung (smoothing)
  - Umverteilen von Wahrscheinlichkeitsmasse von den gesehenen auf die ungesehenen Ereignisse (discounting)
  - z.B. Witten-Bell discounting (Witten-Bell 1991)
    - Wahrscheinlichkeitsmasse der einmal gesehenen Ereignisse wird auf alle ungesehenen Ereignisse verteilt

Stochastische Tagger

- Beispiel: TnT (Brants 2000)

Korpus	Anteil unbek. Wortf.	Genauigkeit bekannte Wortformen	unbekannte Wortformen	Gesamt
PennTB (engl.)	2.9%	97.0%	85.5%	96.7%
Negra (dt.)	11.9%	97.7%	89%	96.7%
Heise (dt.)*)				92.3%

*) Trainingskorpus $\neq$ Testkorpus

- Maximum entropy tagger (Ratnaparkhi 1996): 96.6%

Transformations-basierte Tagger

- Idee: schrittweise Korrektur von fehlerhaften Zwischenresultaten (Brill 1995)
  - kontextsensitive Regeln, z.B. Change NN to VB when the previous tag is TO
- Regeln werden aus einem Korpus gelernt
  1. Initialisierung: Wähle die Tagsequenz mit der höchsten Unigram-Wahrscheinlichkeit
  2. Vergleiche das Ergebnis mit der Korpusannotation
  3. Generiere eine Regel, die die meisten Fehler beseitigt
  4. erneutes Tagging und weiter mit 2.
- Abbruch, falls Verbesserung nur noch unbedeutend

Transformations-basierte Tagger

- Regelgenerierung wird durch Templates gesteuert
  - Change tag a to tag b when ...
    - ...the preceding/following word is tagged z.
    - ...the word two before/after is tagged z.
    - ...one of the two preceding/following words is tagged z.
    - ...one of the three preceding/following words is tagged z.
    - ...the preceding word is tagged z and the following word is tagged w.
    - ...the preceding/following word is tagged z and the word two before/after is tagged w.

Transformations-basierte Tagger

- Ergebnis des Trainings: geordnete Liste von Transformationsregeln

von	nach	Bedingung	Beispiel
NN	VB	vorheriges Tag ist TO	to/TO race/NN → VB
VBP	VB	eines der 3 vorh. Tags ist MD	might/MD vanish/VBP → VB
NN	VB	eines der 2 vorh. Tags is MD	might/MD not reply/NN → VB
VB	NN	eines der 2 vorh. Tags ist DT	
VBD	VBN	einer der 3 vorh. Tags ist VBZ	

Transformations-basierte Tagger

- 97.0% Genauigkeit, wenn nur die ersten 200 Regeln verwendet werden
- 96.8% Genauigkeit mit den ersten 100 Regeln
- Qualität eines stochastischen Parsers auf den gleichen Daten (96.7%) wird mit 82 Regeln erreicht
- extrem lange Trainingszeiten
  $\approx \mathbf{10^6}$-fache eines HMM-Taggers

Anwendungen

- Wortbetonung in der Sprachsynthese
  'content/NN con'tent/JJ
  'object/NN ob'ject/VB
  'discount/NN dis'count/VB
- Ermittlung des Wortstamms (z.B. Textrecherche)
- Klassenbasierte Sprachmodelle für die Spracherkennung
- "flache" Analyse, z.B. zur Informationsextraktion
- Vorstufe zum Parsing, insbesondere für stochastische Parser