Token, Types, Häufigkeiten und automatische Wortartenerkennung (Statistik-basiertes Tagging)

Morphologieanalyse und Lexikonaufbau (6. Vorlesung)

Dozent: Gerold Schneider

Übersicht

-

Was ist ein Token? Was ist ein Type?

Token: (nach [Bußmann 83]: einzelne sprachliche Äusserung)

in einem Text vorkommende Wortformen

Bsp.: "Die Frau jagt die Katze." enthält 5 Token (oder 6 Token, wenn man den Satzendepunkt eigens zählt)

Type: (nach Bußmann: die den sprachlichen Äusserungen zugrundeliegenden abstrakten Einheiten)

in einem Text vorkommende unterschiedliche Wortformen

Bsp.: "Die Frau sah das Mädchen, aber das Mädchen hat sie nicht gesehen." enthält 10 Types ('das' und 'Mädchen' werden nur einmal gezählt; oder 12 Types, wenn man die Satzzeichen eigens zählt; 'sah' und 'gesehen' können auch als zwei Instanzen des selben Lemma-Types gesehen werden.)

-

Wie ist das numerische Verhältnis von Token zu Types?

In Smith "Computers and Human Language" (für Englisch)

> Auf 13.000 Token (=Kap. 3 des Buches) kommen 2427 Types (T/T= 5,35).

Im Scanworx-Manual (für das Deutsche)

> Auf 48.000 Token kommen 3700 Types (T/T= 12,97).

Im Brown-Corpus (für Englisch; nach Smith S.79)

> Auf 1 Mio Token kommen ca. 50.000 Types (T/T = 20)

Zeitungscorpus "Die Welt" (für das Deutsche)

> Auf 2.5 Mio Token kommen 166.484 Types (T/T= 15,01)

Nach Smith

> Type/Token Verhältnis ist niedriger in gesprochener Sprache als in geschriebener Sprache

Anwendung in der Lexikographie:

nur die häufigsten Types werden für die Lexikonerstellung berücksichtigt.

Wie ist das numerische Verhältnis von Funktionswörtern (Präpositionen, Artikel, Konjunktionen, Pronomen) zu Inhaltswörtern (Substantive, Adjektive, Verben) in einem gegebenen Text?

> ungefähr gleich

Wieviel Einträge hat ein grosses Lexikon?

- unabridged, general-purpose Dictionary for English: 450.000 (nach Smith S.70)
- Englisch hat 60.000 Inhaltswörter im allgemeinen Gebrauch (nach Smith S.85) und es kommen jährlich 500 neue hinzu. Die produktivsten Quellen für das Englische sind heute Wissenschaft, Geschäftswelt und American Slang (früher war es die Literatur).
- Oxford English Dictionary: 252.000 Haupteinträge; 1,8 Mio Belegzitate (nach Ufert S.80)
- DUDEN Deutsches Universalwörterbuch A-Z: über 120.000 Stichwortartikel
- Celex ~50.000 Lemmas für das Deutsche

-

Häufigkeits-Verteilung der Wörter

Zipfs Gesetz

"Das Verhältnis der Häufigkeit des Auftretens eines Tokens ist invers proportional zu seiner Position in der Häufigkeitsliste."

`frequency * rank = constant`

Bsp.: (für Englisch; aus Crystal S.87)

rank	*	frequency	constant
35	very	836	29.260
45	see	674	30.330
55	which	563	30.965
65	get	469	30.485
75	out	422	31.650

Problem: Gesetzmässigkeit stimmt nicht ganz am Anfang und am Ende der Liste.

Anmerk.: Anderes Gesetz von Zipf: Die Häufigkeit des Auftretens eines Wortes ist umgekehrt proportional zu seiner Länge.

Effizienz ⇔ Expressivität

-

Die häufigsten Wortformen des Deutschen

nach Meier, H.: Deutsche Sprachstatistik. Hildesheim: Georg Olms. 1964

1	die	349.553	
2	der	342.522	
3	und	320.072	1. Konjunktion
4	in	188.078	1. Präposition
5	zu	172.625	
6	den	138.664	
7	das	124.232	
8	nicht	114.518	1. Adverb
9	von	113.201	
10	sie	102.212	1. Personalpronomen
11	ist	96.970	1. Hilfsverb
12	des	96.190	
13	sich	92.945	
14	mit	91.552	
15	dem	89.109	
90	Zeit	14.529	1. Substantiv
127	machen	8.929	1. Vollverb

Vgl. Liste für das Englische und für das Französische (s. Alexejew et al. "Sprachstatistik", Fink, 1973 S. 218 u. S.223-224).

-

im Brown-Corpus (nach Smith S.79)

> 1 Mio Token (ca. 50.000 Types), davon machen 6 Types (*the, of, and, to, a, in*) 205.961 Token aus

die Spitzenreiter in der Häufigkeit sind unterschiedlich für geschriebene und gesprochene Sprache (in gesprochener Sprache häufiger als in geschriebener ist z.B. '*I* ')

Nutzen dieses Wissens in der CL:

- bei Indices oder Konkordanzen (häufigste Weglassen)
- bei Textspeicherung (häufigste durch Ein-Byte Code ersetzen)
- bei Lexikon-Lookup entsprechend schnelle Zugriffe einsetzen

Beobachtung: Die häufigsten Wörter sind Funktionswörter, vor allem Determiner und Präpositionen. Sie haben normalerweise keine Synonyme und sind syntaxspezifisch.

Was ist Tagging?

Allgemein: Die Zuweisung eines 'Tags' (Markierungssymbol) an eine Texteinheit.

Meist: Die Zuweisung eines eindeutigen Wortartsymbols an eine Wortform im Kontext.

Tagging folgt meist auf die morphologische Analyse oder ist selbst lexikonbasiert. Es kann entweder statistisch oder regelbasiert ablaufen. Beispiel:

	Morphologische Analyse:	Tagger:
Junge	[Adj, N]	Adj
Männer	[N]	N
gehen	[finV, infV]	finV
zu	[Präp, Adv, iKonj, Adj]	Präp
ihr.	[Pron, Det]	Pron

nach Smith (S.86): 5% der Types sind ambig. Da diese jedoch sehr häufig sind, entspricht das bis zu 20% der Token.

nach Charniak (S.49): Im Brown-Corpus sind 11% der Types ambig. Das entspricht jedoch 40% der Token.

-

Tag-Sets

Das Tag-Set umfasst die Menge der Tags, die von einem Tagger vergeben werden.

Tag-Set	Number of Tags
Brown Corpus	87
Lancaster-Oslo/Bergen	135
Lancaster UCREL	165
London-Lund Corpus of Spoken English	197
Penn Treebank	36 + 12

-

Bsp.: Der Xerox Part-of-Speech-Tagger

Basiert auf dem LOB (Lancester-Oslo-Bergen) Tag-Set. Dieses enthält rund 120 Tags für die Wortarten plus Tags für die Satzzeichen.

Ausgangstext:

```
You can drink from a can of beer and fly home like a fly.
You live your lives as a man would do time and again.
Do you think that a buffalo can buffalo a buffalo?
```

Der analysierte Text:

```
You/PPSS  can/MD  drink/VB  from/IN  a/AT  can/NN  of/IN  beer/NN  and/CC
fly/NN  home/NN  like/CS  a/AT  fly/NN  ./SENT

You/PPSS  live/VB  your/PP$  lives/NNS  as/RBC  a/AT  man/NN  would/MD  do/DO
time/NN  and/CC  again/RB  ./SENT

Do/DO  you/PPSS  think/VB  that/CS  a/AT  buffalo/NN  can/MD  buffalo/VB  a/AT
buffalo/NN  ?/SENT
```

Der Xerox-Tagger kann über das WWW getestet werden:
http://www.rxrc.xerox.com/research/mltt/demos/ .

Es gibt Versionen für DE, FR, NL, EN, ES, PT, IT, RU, Ungarisch (weitere in Entwicklung)

-

Anforderungen an einen Tagger

(nach [Cutting et al. 92] S.133)

Robustheit

Der Tagger kann beliebigen Input verarbeiten (incl. unbekannte Wörter, Sonderzeichen).

Effizienz

Der Tagger arbeitet schnell.

Genauigkeit

Der Tagger arbeitet mit einer geringen Fehlerrate (< 5%).

Anpassbarkeit

Der Tagger kann an besondere Anforderungen eines Texttyps angepasst werden.

Wiederverwertbarkeit

Der Tagger kann leicht für neue Aufgabengebiete eingesetzt werden.

-

Statistisches Tagging

Annahme: die Wahrscheinlichkeit der Aufeinanderfolge von Wortarten ist unterschiedlich. Ausgangspunkt ist einmal die Wahrscheinlichkeit, dass ein gegebenes Wort mit Wahrscheinlichkeit P die Wortart POS1 hat. Die Wahrscheinlichkeit der Wortartenübergänge werden dann berechnet (z.B. indem man manuell disambiguiert ODER abwechselnd manuell disambiguiert, tagged und korrigiert) und über mehrere Wörter hinweg (Tri-Tupel, Quad-Tupel) die maximale Wahrscheinlichkeit der Übergänge ermittelt.

Grundlage: Hidden Markov Modelle (HMM)

- basieren auf Markov-Ketten. Markov-Ketten entsprechen endlichen Automaten, bei denen jede Kante mit einer Wahrscheinlichkeit versehen ist. Die Summe aller Kanten, die den selben Knoten verlassen, muss 1 ergeben.
- sind eine Verallgemeinerung von Markov-Ketten, in denen ein gegebener Knoten (Zustand) mehrere ausgehende Kanten hat, die das gleiche Symbol tragen. Deshalb ist es nicht möglich, aufgrund der Ausgabe zu erschliessen, welche Zustände berührt wurden (-> *hidden*).
- die Wahrscheinlichkeit jedes Zustands hängt ab vom Vorgängerzustand.
- die Hidden Markov Modelle sind ein stochastischer Prozess, der über den Markov-Ketten liegt und Sequenzen von Zuständen berechnet.
- Aus all diesen Sequenzen findet man mit Hilfe spezieller Algorithmen die beste Sequenz (z.B. mit Viterbi-Algorithmus).

-

Ein Beispiel aus: [Feldweg 96].

die auf der Bank sitzende Frau.

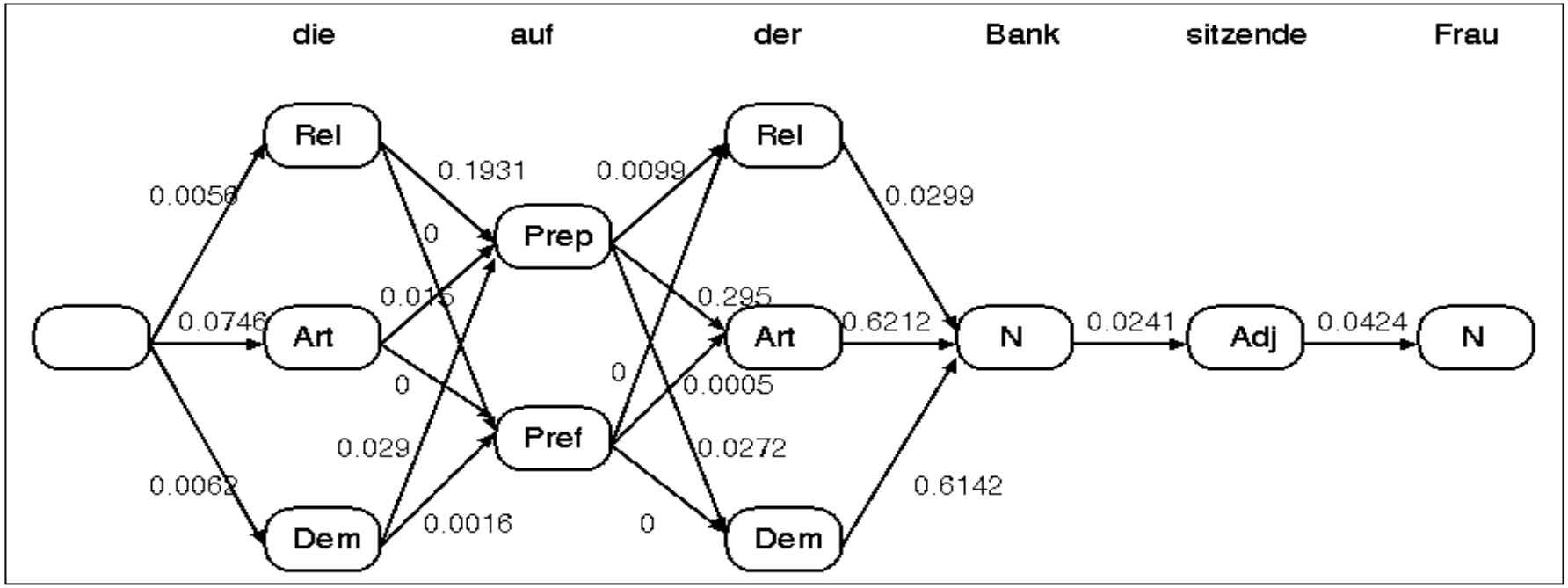
folgende Mehrdeutigkeiten:

(.)	die	auf	der	Bank	sitzende	Frau
(.)	REL ART DEM	PREP PREF	REL ART DEM	N	ADJ	N

PREF = Verbpräfix, entsprechend PTKVZ (Partikel Verbzusatz) im STTS Tagset

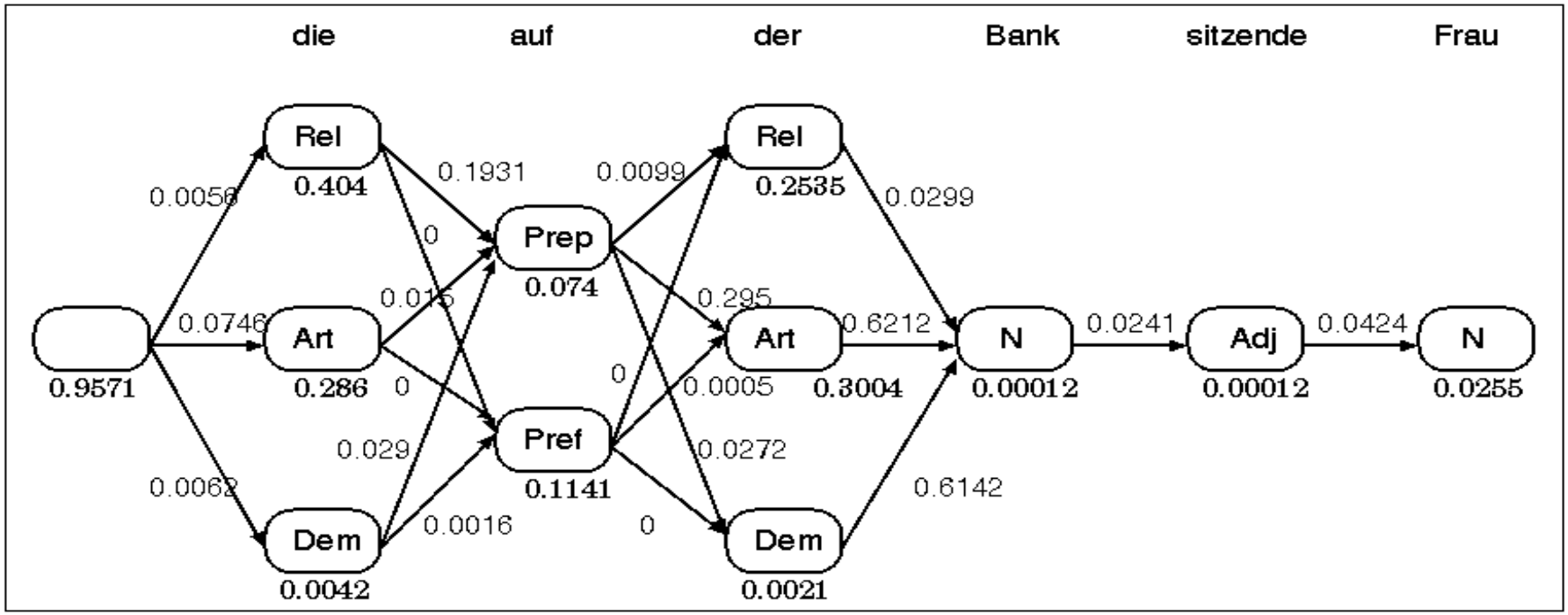
-

Man ordnet den Übergängen zwischen den Wortarten Wahrscheinlichkeiten zu.



-

Schliesslich wird jeder Wortform-Wortart-Kombination eine Wahrscheinlichkeit zugeordnet:



Für jeden möglichen Pfad durch ein solches Netz lässt sich durch Multiplikation der auf dem Pfad liegenden Werte eine Gesamtwahrscheinlichkeit berechnen.

-

Formal betrachtet handelt es sich bei diesem Verfahren um ein Hidden-Markov-Modell erster Ordnung. Ein solches Modell ist definiert über:

1. die Menge von *n* Symbolen
   $V = \{w_1, ..., w_n\}$
   In unserem Beispiel bildet das Vokabular (= die Menge der Wortformen) die Menge V.
2. die Menge von *m* Zuständen
   $S = \{s_1, ..., s_m\}$
   Diese Menge entspricht den möglichen Wortarten.
3. einer Menge von m^2 Übergangswahrscheinlichkeiten zwischen Zuständen
   $P = \{p(s_1|s_1), ..., p(s_i|s_j)\}, 1 < i,j, < m$
   Hier die Übergangswahrscheinlichkeiten zwischen Wortarten.
4. einer Menge von Observationswahrscheinlichkeiten
   $L = \{p(w_1|s_1), ..., p(w_k|s_l)\}, 1 < k < n, 1 < l < m$
   Dies entspricht der Menge der lexikalischen Wahrscheinlichkeiten: gegeben die Wortart *s*, wie hoch ist die Wahrscheinlichkeit von Wortform *w*?
5. Für eine gegebene Folge von *i* Symbolen, kann mit Hilfe dieses Modells die wahrscheinlichste Folge bestimmt werden durch:
   max(s) vom produkt(j=1 bis i) von $p(s_j|s_{j-1}) \_ p(w_j|s_j)$

-

Beim beschriebenen Modell handelt es sich um ein Hidden-Markov-Modell erster Ordnung. Übergangswahrscheinlichkeiten werden dabei nur für direkt benachbarte Zustände berücksichtigt (Bigram-Modell). Es sind jedoch auch Modelle höherer Ordnung möglich.

Parametergewinnung

Das Tagging mittels HMM ist prinzipiell sprachunabhängig. Voraussetzung ist jedoch, dass die in den Gleichungen 1-4 aufgeführten Parameter bekannt sind. Die Gewinnung der Parameter ist jedoch das eigentliche Problem.

1. Einfach: Bestimmung des Vokabulars
2. Schwieriger: Festlegung des Tag-Sets (Wortartenmenge und -abgrenzung). Siehe Diskussion über die Wortartenklassifikation in der 1. Vorlesung.

-

3. Sehr komplex: Gewinnung von Übergangswahrscheinlichkeiten und Observationswahrscheinlichkeiten. Präzise allgemeingültige Werte kennt man nicht . Diese Parameter müssen geschätzt werden (i. A. berechnet aus bereits getaggten Korpora).

    1. für die meisten Sprachen sind keine hinreichend grossen, getaggten Textkorpora verfügbar. Möglichkeiten:

        - Verringerung der Anzahl der zu schätzenden Parameter (z.B. individuelle Werte nur für hochfrequente Wortformen; ansonsten Werte für die Ambiguitätsklasse (= Menge aller Wortformen mit den gleichen Tags) berechnen)

        - Gewinnung von Werten aus ungetaggten Textkorpora: zufällige Auswahl von Wahrscheinlichkeiten und Abgleich der gewonnenen Tags mit einem Vollformenlexikon; sich wiederholende Taggingvorgänge

    2. vermeintliche Nullübergänge müssen abgefangen werden, denn Nullübergänge haben grosse Auswirkungen. Lösung: Ersetzen der Nullübergänge durch sehr kleine Wahrscheinlichkeiten

    3. kein Korpus kann wirklich allgemeingültig und ausgewogen sein

-

Einsatzgebiete für Tagger

(nach [Church 93] S.7)

- Sprachsynthese
- Spracherkennung
- Information Retrieval
- Bedeutungsdisambiguierung
- Lexikographie

-

Genauigkeit statistischer Tagger

Statistische Tagger erreichen eine Genauigkeit von rund 94-97% bei einem Tag-Set wie dem STTS mit rund 50 Tags. Besonderes Handicap für Tagger sind:

- unbekannte Wörter (d.h. Wortformen, die im Trainingskorpus nicht vorkamen); insbesondere auch fremdsprachige Einschübe
- weite Abhängigkeiten im Satz, die über das Bigram- bzw. Trigram-Fenster hinausgehen.

```
... weil wir diese Probleme schon kennen/VVFIN.
Wir sollten diese Probleme schon kennen/VVINF.
Die Frauen, die/ART Kinder und alte Männer wurden evakuiert.
Die Frauen, die/PRELS Kinder und alte Männer evakuierten, wurden
geehrt.
```

- Aufzählungen, die keine vollständigen Sätze bilden

Vorsicht! Bei Sätzen, die im Durchschnitt ~20 Wörter lang sind, bedeutet eine Fehlerrate von 4%, dass 56% aller Sätze (also jeder zweite Satz) ein falsch getaggtes Wort enthalten. Für eine Fehlerrate von 4% bei den Sätzen müsste die Fehlerrate bei den Wörtern auf 0,2% sinken.

-

Tagger zum Testen

Ein frei verfügbarer Tagger für das Deutsche, der mit statistischen Methoden arbeitet, findet sich im Morphy-System der Universität Paderborn. Dieser Tagger läuft unter MS-DOS und Windows95 auf PCs.

Ein weiterer verfügbarer Tagger für das Deutsche wurde an der Universität Stuttgart entwickelt. Er nennt sich TreeTagger und läuft unter SunOS und Linux. Neuerdings gibt es auch eine Windows Demoversion.

Gerold Schneider, Martin Volk

-