

Nominalausdrücke [1]

- In unseren Automaten für Wortartmuster (s. Übungen zu Automaten für Wortartmuster und kontextfreie Grammatik für NP) sind wir davon ausgegangen, dass Nominalausdrücke entweder aus Artikel und NN bestehen, eventuell mit einem oder mehreren Adjektiven davor oder einer oder mehreren PP dahinter, oder aus einem einzelnen Pronomen oder Eigennamen. Beispiele:

das Haus

das große Haus

das Haus an der Ecke

das Haus an der Ecke mit dem roten Dach

Peter

sie

Nominalausdrücke [2]

- Es gibt offenbar NPs, die ein Gattungssubstantiv und keinen Artikel enthalten, Eigennamen, die mit Adjektiven und PPs ergänzt werden, und NPs, bei denen das Adjektiv hinter dem Substantiv kommt.

Gold und Silber

neue Besen

der schwarze Peter

Hans im Glück

sie mit der roten Kappe

Röslein rot

Augen schwarz wie die Nacht

Nominalausdrücke [3]

- Ein Problem: Unsere erste Analyse war offenbar viel zu restriktiv. Wir müssen viele zusätzliche Regeln vorsehen, um die Abdeckung (engl. coverage) unserer Grammatik zu verbessern.
- Dabei reicht es im allgemeinen nicht, einfach neue Regeln hinzuzufügen. Z.B. ist $NP \rightarrow (ART)(ADJ) NN$ viel zu liberal:
 - Sie hat Augen schwarz wie die Nacht
 - *Sie hat Auge schwarz wie die Nacht
 - Rotkäppchen hat Kuchen gebracht.
 - *Rotkäppchen hat Wolf getroffen.Regeln und Kategorien müssen spezifischer sein: Z.B. gehen ohne Artikel nur sog. Stoffbegriffe (Kuchen, Geld, Zeit,...) und Plurale.

Nominalausdrücke [4]

- Ein zweites Problem: Es gibt viele Konstruktionen, die möglich, aber ungewöhnlich, unwahrscheinlich, "linguistisch markiert" sind - beispielsweise eine NP, die aus Pronomen mit nachfolgender PP besteht. Wenn wir die Regel NP → PPRO PP uneingeschränkt zulassen, bekommen wir die folgenden Analyse-Alternativen:
 - (1) *Peter hat* [_{NP} *ihm*] [_{PP} *mit dem Knüppel*] *gedroht*
Peter hat [_{NP} *ihm* [_{PP} *mit dem Knüppel*]] *gedroht*
 - (2) *Peter hat* [_{NP} *ihm*] [_{PP} *mit dem Knüppel*] *geholffen*
Peter hat [_{NP} *ihm* [_{PP} *mit dem Knüppel*]] *geholffen*
 - (3) *Peter hat* [_{NP} *ihr*] [_{PP} *mit der roten Kappe*] *gedroht*
Peter hat [_{NP} *ihr* [_{PP} *mit der roten Kappe*]] *gedroht*
- Dabei ist die zweite Analyse immer die unwahrscheinlichere, auch wenn die erste "mit dem Knüppel helfen", "mit der roten Kappe drohen" eigentlich unplausibel ist.

Nominalausdrücke [5]

- Das Problem: Wenn wir Texte mit einer restriktiven Grammatik analysieren, werden bestimmte NPs gar nicht erkannt: Der "Recall" ist schlecht.
 - Recall ist die Anzahl der korrekten Beurteilungen (als NP) geteilt durch die Anzahl aller tatsächlich vorhandenen Fälle (von NPs). Wenn in einem Text von 100 vorkommenden NPs 77 korrekt als NPs analysiert wurden, ist der Recall 0,77 oder 77%.
- Wenn wir Regeln für die markierten, unwahrscheinlichen Konstruktionen durchgehend zulassen, werden in vielen Fällen fälschlich NPs erkannt: Die "Präzision" ist schlecht.
 - Präzision ist die Anzahl der korrekten Beurteilungen (als NP) geteilt durch die Anzahl aller Beurteilungen (als NP). Wenn im obigen Beispiel 110x eine NP erkannt wurde (und 77x davon korrekt), ist die Präzision 0.7 oder 70%.

Präferenzen und statistische Häufigkeit

- Ein endlicher Automat oder eine Grammatik, die zwischen guten und schlechten Fällen scharf unterscheiden, geben uns normalerweise zu wenig:
 - die Beschreibung ist restriktiv, die Präzision hoch, aber der Recall niedrigoder zu viel:
 - die Beschreibung ist zu liberal, der Recall ist hoch, die Präzision niedrig.
- Wir benötigen zusätzlich Information darüber, wie markiert oder wahrscheinlich eine bestimmte Interpretation ist.
- Die Information würde uns auch grundsätzlich erlauben, in Fällen von Mehrdeutigkeit die wahrscheinlichere oder präferierte Lesart zu ermitteln (s. Beispiel von Folie 4).
- Information über Normalität oder Wahrscheinlichkeit oder erhalten wir vor allem durch die statistische Auswertung von Textkorpora.

Beispiel: Wortarterkennung

- Viele Wörter sind mehrdeutig zwischen verschiedenen Wortarten. Unterschiedliche Wortartkategorisierung führt zu ganz unterschiedlichen syntaktischen Analysen.
- Ein Beispiel: *achten*, *weißen*, *einen* sind gleichzeitig Verbformen und Formen eines Zahlworts, Adjektivs bzw. Artikels. Während *achten* als Verb und als Zahlwort üblich und häufig ist, ist der unbestimmte Artikel *einen* extrem häufig und das Verb *einen* sehr selten. (der Fall *weißen* liegt irgendwo dazwischen). Eine Wortartzuweisung, die jedes Vorkommen von *einen* gleichwertig als Artikel und als Verb klassifiziert, führt zu vielen abstrusen syntaktischen Analysen.
- Häufigkeitsverteilungen der Lesarten im Korpus geben uns eine Möglichkeit, die wahrscheinlichste Lesart zu ermitteln.

Einige Lesartenfrequenzen im Wahrig-Korpus

Die folgende Statistik gibt an, wie oft die Wortformen achten, einen, weißen, und zusätzlich weiß, im Korpus insgesamt vorkommen, und wie viele Vorkommen jeweils auf die beiden Wortarten entfallen.

	gesamt	Verb	Nicht-Verb
achten	12881	7218	5663
einen	1429533	13	1429520
weißen	24707	43	24664
weiß	110405	107205	3200

Relative Häufigkeit einzelner Lesarten

	gesamt	Verb	Nicht-Verb
achten	1,00000000	0,56036022	0,43963978
einen	1,00000000	0,00000909	0,99999091
weiß en	1,00000000	0,00174040	0,99825960
weiß	1,00000000	0,97101581	0,02898419

$F(\text{VERB}|\text{achten}) = 0,56036022 \approx 56\%$

$F(\text{ADJ}|\text{achten}) \approx 54\%$

$F(\text{VERB}|\text{weiß en}) \approx 0,2\%$

$F(\text{ADJ}|\text{weiß en}) \approx 99,8 \%$

$F(\text{ADJ}|\text{achten})$ gibt an, welcher Anteil von achten-Vorkommen im Korpus Verb-Vorkommen sind.

Häufigste und wahrscheinlichste Lesart

- Wenn wir uns bei der Wortartzuweisung nur an der Häufigkeit der jeweiligen Lesart orientieren, wird ohne Rücksicht auf den Kontext immer die gleiche Wortart ausgewählt. Das ist offensichtlich nicht angemessen:
Siehst du den weißen Wagen?
Die Rückwand brauchst du nicht zu weißen.
- Wir müssen zusätzlich den Kontext einbeziehen. Manchmal muss man, um zu einem sicheren Urteil zu kommen, den ganzen Satz betrachten. In vielen Fällen liefert aber schon die Wortart des vorhergehenden Wortes einigermaßen verlässliche Information. Nach der Infinitiv-Partikel *zu* folgt fast mit Sicherheit ein Verb. Nach einem Artikel ist die Verblesart nicht ausgeschlossen, aber ziemlich unwahrscheinlich.

Wortart-Tagger

- Wortart-Tagger (engl. POS-Tagger; "POS" für "Part of Speech", engl. für "Wortart") sind statistische Programme, die Wörtern im Kontext ein "Wortart-Tag" aus einem vorgegebenen Tag-Set (z.B. STTS) zuweisen, das ihre (vermutliche) Wortart in diesem Kontext angibt.
- Wortart-Tagger weisen auch dann eine Lesart zu, wenn die Wortform zwischen mehreren Wortarten mehrdeutig ist (s.o.), oder wenn die Wortform gar nicht bekannt ist.
- Tagger lernen die Wortartzuweisung, indem sie an einem handannotierten ("getaggten") Korpus trainiert werden (s.u.)
- Gute Tagger haben eine Akkuratheit zwischen 96% und 98%, sind also bereits sehr verlässlich.

Wortart-Tagger [2]

- Tagger "lernen" aus einem annotierten Korpus, indem sie, vereinfacht dargestellt, zwei Arten statistischer Information kombinieren:
 - die relative Häufigkeit, mit der eine Wortform w im Korpus einer bestimmten Wortart t angehört
 - die relative Häufigkeit, mit der das Wortart-Tag t auf das Tag t' des vorausgehenden Wortes (oder die Tags t' und t'' der beiden vorausgegangenen Wörter) folgt.
 - Weil dabei Paare von Tags oder Tripel (Dreiergruppen) von Tags ausgezählt werden, reden wir von "Bigramm" bzw. "Trigramm"-Wahrscheinlichkeiten.
- Damit approximieren Tagger die Wahrscheinlichkeit, mit der eine Wortform w in einem Kontext tatsächlich die Wortart t besitzt: Sehr stark vereinfacht ausgedrückt, ist eine Wortart für eine Wortform w umso wahrscheinlicher, je häufiger die Wortform insgesamt in dieser Wortart auftritt, und je besser sie zu der Wortart des Vorgängerwortes passt.

Die folgenden Folien sind kein Lernstoff für die
Klausur!

Theoretische Überlegungen zum POS-Tagging

Was wir eigentlich brauchen:

Die wahrscheinlichste Folge $t_1 \dots t_n$ von Wortart-Tags, gegeben eine Wortkette $w_1 \dots w_n$:

$$\arg \max_{t_1 \dots t_n \in TAGSEQ} (P(t_1 \dots t_n \mid w_1 \dots w_n))$$

Problem: Wir können $P(t_1 \dots t_n \mid w_1 \dots w_n)$ nicht mit $F(t_1 \dots t_n \mid w_1 \dots w_n)$ approximieren: Die meisten möglichen Sätze kommen in einem noch so großen Korpus gar nicht oder höchstens einmal vor:
Das „sparse data“-Problem (Spärliche-Daten-Problem(?))

Das Bayessche Gesetz

$$P(A | B) = \frac{P(A) \cdot P(B | A)}{P(B)}$$

Angewandt auf das Wortart-Tag-Problem:

$$\begin{aligned} &P(t_1 \cdots t_n | w_1 \cdots w_n) \\ &= \frac{P(t_1 \cdots t_n) \cdot P(w_1 \cdots w_n | t_1 \cdots t_n)}{P(w_1 \cdots w_n)} \end{aligned}$$

$$\begin{aligned}
& \arg \max_{t_1 \cdots t_n \in TAGSEQ} (P(t_1 \cdots t_n \mid w_1 \cdots w_n)) \\
&= \arg \max_{t_1 \cdots t_n \in TAGSEQ} \left(\frac{P(t_1 \cdots t_n) \cdot P(w_1 \cdots w_n \mid t_1 \cdots t_n)}{P(w_1 \cdots w_n)} \right) \\
&= \arg \max_{t_1 \cdots t_n \in TAGSEQ} (P(t_1 \cdots t_n) \cdot P(w_1 \cdots w_n \mid t_1 \cdots t_n))
\end{aligned}$$

Erste vereinfachende Annahme: n-Gramm-Technik

Kettenregel:

$$\begin{aligned} P(t_1 \cdots t_n) &= P(t_1) \cdot P(t_2 | t_1) \cdots P(t_n | t_1 \cdots t_{n-1}) \\ &= \prod_{i=1}^n P(t_i | t_1 \cdots t_{i-1}) \end{aligned}$$

Vereinfachende Annahme: Die Wahrscheinlichkeit eines Wortart-Tags hängt nur vom letzten vorhergehenden/ von den letzten beiden vorhergehenden Tags ab (Bigramm/Trigramm)

$$P(t_1 \cdots t_n) \approx P(t_1) \cdot P(t_2 | t_1) \cdot P(t_3 | t_1 t_2) \cdots P(t_n | t_{n-2} t_{n-1})$$

Zweite vereinfachende Annahme

Die Vorkommenswahrscheinlichkeit eines Wortes wird nur durch die Wortart bestimmt.

$$\begin{aligned} P(w_1 \cdots w_n \mid t_1 \cdots t_n) \\ &\approx P(w_1 \mid t_1) \cdot P(w_2 \mid t_2) \cdots P(w_n \mid t_n) \\ &= \prod_{i=1}^n P(w_i \mid t_i) \end{aligned}$$

Statistischer Wortart-Tagger

Die Basisformel für den Trigramm-Tagger:
Das wahrscheinlichste Wortartmuster für eine Kette/
einen Satz $w_1 \dots w_n$:

$$\begin{aligned} & \arg \max_{t_1 \dots t_n \in TAGSEQ} (P(t_1 \dots t_n \mid w_1 \dots w_n)) \\ & \approx P(t_1) \cdot P(t_2 \mid t_1) \cdot \prod_{i=3}^n P(t_i \mid t_{i-2} t_{i-1}) \cdot \prod_{i=1}^n P(w_i \mid t_i) \\ & \approx F(t_1) \cdot F(t_2 \mid t_1) \cdot \prod_{i=3}^n F(t_i \mid t_{i-2} t_{i-1}) \cdot \prod_{i=1}^n F(w_i \mid t_i) \end{aligned}$$