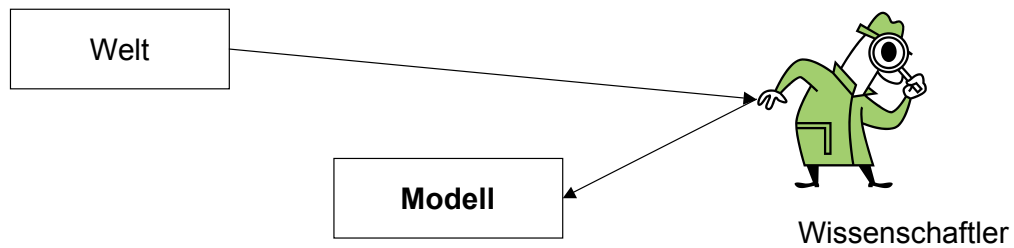

Statistische Sprachverarbeitung

Sebastian Pado
11.01.2005

Übersicht

- Modellierung
- Statistik - Einige grundlegende Konzepte
- Statistische Klassifikation zur Sprachverarbeitung
 - In 5 Schritten zum erfolgreichen Statistiker
 - Konkretes Beispiel: Anbindung von Präpositionalphrasen
- (Nächstes Mal: Probabilistische kontextfreie Grammatiken)

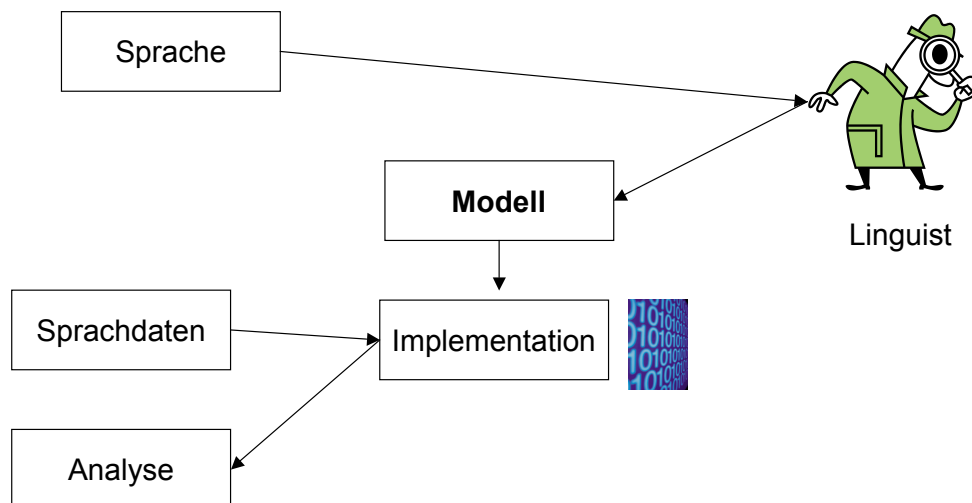
1. Modellierung



Ein Modell ist **vereinfachtes Abbild** der Welt.
Modelle machen i.A. **Aussagen/Vorhersagen** über die Welt.

Einführung in die Coli WS 04/05

Computerlinguistische Modellierung



Einführung in die Coli WS 04/05

Beispiele für linguistische “Modelle”

- Schulgrammatiken
 - “Das Subjekt eines Satzes steht im Nominativ”
- Kontextfreie Grammatiken
 - NP -> Det (Adj)* N (RelS)
- Phonologie
 - Deutsche Silben haben die Form
(K)(K)(K)(K)(V)(K)(K)(K)(K)(K)

Erlauben (mehr oder weniger korrekte) Vorhersagen

Einführung in die Coli WS 04/05

Symbolische Modelle

- Formalisieren Wissen des Autors über Konzepte
 - Traditionell wichtig in Computerlinguistik und theoretischer Informatik
- Arbeiten mit abstrakten Symbolen, die eine Bedeutung (Konzept) tragen
 - Linguistik: Satz, Subjekt, Silbe, etc.
- Oft als Regeln oder Bedingungen formuliert
 - Wenn X, dann Y

Einführung in die Coli WS 04/05

Probleme symbolischer Modelle (1)

- Vollständigkeit ist schwer
 - Autor muss jeden Fall explizit behandeln
 - Modelle für komplexe Phänomene sehr groß
- Konsistenz ist schwer
 - Interaktionen zwischen Regeln schwer zu übersehen
- Modell-Autor zentral
 - Bedeutung der Symbole im Kopf des Autors
 - Anwendung des Modells ist automatisierbar
 - Erstellung und Pflege nicht

Einführung in die Coli WS 04/05

Probleme symbolischer Modelle (2)

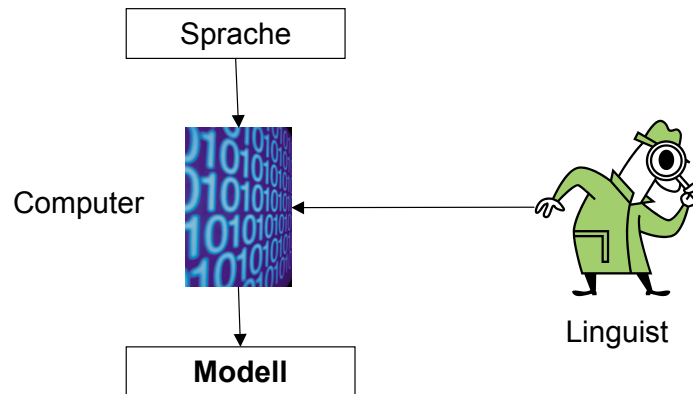
- Produktivität der Sprache:
 - Neologismen (mailen), Komposita
- Sprache entwickelt sich weiter: Orthographie (thut/tut)
- Menschen verwenden Sprache „falsch“ -- und verstehen sich trotzdem
 - „Nur“ das Richtige modellieren reicht nicht
- Sprachliche Kategorien sind oft nicht binär
 - Syntax: Grammatisch / ungrammatisch
 - Ich kenne ihn, weil wir uns letztes Jahr getroffen haben.
 - (?)Ich kenne ihn, weil wir haben uns letztes Jahr getroffen.

Graduierung ist schwierig symbolisch zu modellieren

Einführung in die Coli WS 04/05

Ein Gegenentwurf

- **Automatisches** Lernen von Modellen aus Sprachdaten = Maschinelles Lernen



- Statistik = verbreiteter Ansatz („Framework“) für maschinelles Lernen

Einführung in die Coli WS 04/05

Statistische Modellierung

- **Beobachte Häufigkeiten** von **Ereignissen** in Sprache
 - Betrachte sprachliche Objekte als **Experimente**
- Schätze **Wahrscheinlichkeiten** P aller **Ereignisse** (**Wahrscheinlichkeitsverteilung**)
 - **Statistisches Modell**
- Mache Vorhersagen über neue Ereignisse

Einführung in die Coli WS 04/05

Vorteile maschinell gelernter Modelle

- Vollständigkeit, Konsistenz durch automatisches Lernen garantiert
- Geringerer manueller Aufwand
 - Mensch bestimmt “nur” Form des Modells
- Einfaches Bauen neuer Modelle aus neuen Daten
 - Adaption auf veränderte Sprache, etc.
- Modellierung des “ist”-Zustandes, nicht des “soll”-Zustandes
- Statistische Modelle: “eingebaute” Graduierung

Einführung in die Coli WS 04/05

2. Statistik

- Elementarereignisse
- Ereignisse
- Kombinationen von Ereignissen

Einführung in die Coli WS 04/05

Elementarereignisse

- Statistik beschreibt Ergebnisse vergangener („**gesehener**“) Experimente und sagt damit Ergebnisse neuer („**ungesehener**“) Experimente voraus
- Menge der möglichen **Elementarereignisse** Ω
 - Werfen eines Würfels: $\Omega = \{1,2,3,4,5,6\}$
- Jedes Elementarereignis ω hat eine **Wahrscheinlichkeit** $P(\omega)$:
$$P(\omega) = \frac{f(\omega)}{N}$$
 - Ermittlung z.B. durch
 - Beobachtung von **Frequenzen bisheriger Experimente**
 - Annahmen über Experiment
 - Werfen eines Würfels: $P(\omega) = 1/6$ für alle ω .

Einführung in die Coli WS 04/05

Ereignisse

- **Ereignis E**: Menge von Elementarereignissen
 - Können alle interessanten Eigenschaften von Experimenten beschreiben
 - Augenzahl ist gerade: $E = \{2,4,6\}$
 - Augenzahl ist größer als 4: $E = \{5,6\}$
- Wahrscheinlichkeiten für Ereignisse:
 - Frequenzen
 - Nur möglich, wenn beobachtet

$$P(\text{Ereignis } E) = \frac{f(E)}{N}$$

Einführung in die Coli WS 04/05

Wahrscheinlichkeitsverteilungen

- Eine **Wahrscheinlichkeitsverteilung** ordnet jedem Ereignis E eine Wahrscheinlichkeit zu
 - Nutzung als statistisches Modell
- Der Ereignisraum ist sehr gross
 - Potenzmenge der Elementarereignisse
 - Wir werden nie alle Ereignisse sehen
 - Nötig: Berechnung der Wahrscheinlichkeit **komplexer Ereignisse** aus einfachen Ereignissen („Rechenregeln“)

Einführung in die Coli WS 04/05

Kombination von Ereignissen

- Welche komplexen Ereignisse sind interessant?
 - **Gemeinsame Wahrscheinlichkeit** $P(A, B)$:
 - Die Wahrscheinlichkeit, daß sowohl Ereignis A als auch Ereignis B eintreten („A und B“)
 - $P(\text{Augenzahl gerade, Augenzahl} > 3)$
 - Gemeinsame Wahrscheinlichkeiten sind oft beobachtbar
 - **Bedingte Wahrscheinlichkeit** $P(A | B)$:
 - Die Wahrscheinlichkeit, A zu sehen, wenn wir schon B gesehen haben („A gegeben B“)
 - $P(\text{Augenzahl gerade} | \text{Augenzahl} > 3)$
 - Setzt zwei Ereignisse in Beziehung
 - Typischerweise: was wir gesehen haben und was wir sehen werden

Einführung in die Coli WS 04/05

Gemeinsame Wahrscheinlichkeit

■ Geometrische Deutung:

- $P(A,B)$ ist das Verhältnis der Schnittfläche von A und B zur **Gesamtfläche**
- $P(A|B)$ ist das Verhältnis der Schnittfläche von A und B zur **Fläche von B**

Augen
zahl
> 4

Augenzahl gerade

1	3	5
2	4	6

■ Vorschlag für Rechenregeln:

- $P(A,B) = P(A) * P(B)$
- $P(A|B) = P(A,B) / P(B)$

■ Beispiel 1:

- Augenzahl gerade: $P(A) = 1/2$
- Augenzahl > 4: $P(B) = 1/3$

■ Geometrisch (siehe Grafik)

- $P(A,B) = 1/6$
- $P(A|B) = 1/2$

■ Rechenregel

- $P(A,B) = 1/2 * 1/3 = 1/6$
- $P(A|B) = 1/6 / 1/3 = 1/2$

Scheint zu klappen!

Einführung in die Coli WS 04/05

Gemeinsame Wahrscheinlichkeit

■ Geometrische Deutung:

- $P(A,B)$ ist das Verhältnis der Schnittfläche von A und B zur **Gesamtfläche**
- $P(A|B)$ ist das Verhältnis der Schnittfläche von A und B zur **Fläche von B**

Augen
zahl
> 4

Augenzahl gerade

1	3	5
2	4	6

■ Vorschlag für Rechenregeln:

- $P(A,B) = P(A) * P(B)$
- $P(A|B) = P(A,B) / P(B)$

■ Beispiel 2:

- Augenzahl gerade: $P(A) = 1/2$
- Augenzahl > 3: $P(B) = 1/2$

■ Geometrisch (siehe Grafik)

- $P(A,B) = 1/3$
- $P(A|B) = 2/3$

■ Rechenregel

- $P(A,B) = 1/2 * 1/2 = 1/4$
- $P(A|B) = 1/4 / 1/2 = 1/2$

Geht schief!

Einführung in die Coli WS 04/05

Das Problem: Abhängigkeit

- Zwei Ereignisse sind voneinander unabhängig, wenn sie sich **gegenseitig nicht beeinflussen**
 - Intuition: „ob B eintritt oder nicht, ändert nichts an $P(A)$ und umgekehrt“
- Unabhängige Ereignisse:
 - Sonntag, es regnet
 - Augenzahl gerade, Augenzahl > 4
- Abhängige Ereignisse:
 - Es regnet, ich werde nass
 - Augenzahl gerade, Augenzahl > 3

Einführung in die Coli WS 04/05

Abhängigkeit (II)

- Nur wenn zwei Ereignisse unabhängig sind, gilt $P(A,B) = P(A) \cdot P(B)$
 - Intuition “Zuerst passiert A und dann B”
- Bei Unabhängigkeit gilt auch $P(A|B) = P(A)$
 - Wenn B eintritt, sagt uns das nichts über A
 - $P(\text{Augenzahl gerade} \mid \text{Augenzahl} > 4) = 1/2$
 - $P(\text{Augenzahl gerade}) = 1/2$
- Bei Abhängigkeit ist entweder $P(A|B) > P(A)$ oder $P(A|B) < P(A)$
 - Dann ist B ein positiver / negativer **Indikator** für A
 - Wenn wir B sehen, ist es (un)wahrscheinlicher, auch A zu sehen
 - $P(\text{Augenzahl gerade}) = 1/2$
 - $P(\text{Augenzahl gerade} \mid \text{Augenzahl} > 3) = 2/3$

Einführung in die Coli WS 04/05

3. Klassifikation

- Aufgabe: Klassifiziere Daten (Instanzen) in **Klassen**
 - Klassen sind nicht direkt beobachtbare Ereignisse
 - Deshalb wollen wir, dass das Modell sie entscheidet!
 - Beispiel: Phoneme bei Spracherkennung
 - Beispiel: Wortarten bei Worten
- Identifiziere hilfreiche **beobachtbare** Ereignisse
 - „Features“
 - **Hilfreich** heisst **abhängig** von Klassen
 - Positive oder negative Indikatoren für Klassen (Evidenz)
 - Features bilden Grundlage für Entscheidung zwischen Klassen
 - Beispiel: verschiedene Formanten bei Vokalerkennung

Einführung in die Coli WS 04/05

Beispiele für Klassifikatoren

- Parsing: $P(\text{syntaktische Analyse} \mid \text{Wörter})$
- Tagging: $P(\text{Wortart} \mid \text{Wort})$
- Spracherkennung:
 $P(\text{Buchstabe} \mid \text{Frequenzen})$

Einführung in die Coli WS 04/05

Klassifikation: Details

- **Konditionaler** statistischer Klassifikator:

$$P(\text{Klasse } K \mid \text{Instanz } I) = \\ P(\text{Klasse } K \mid \text{Features } f_1, f_2, f_3, \dots)$$

„Wahrscheinlichkeit für Klasse K gegeben die Kombination der Ereignisse (Features) f_1, f_2, f_3, \dots “

- Entscheidungsalgorithmus: Wähle die wahrscheinlichste Klasse K:

$$K = \operatorname{argmax}_K P(K \mid f_1, f_2, f_3, \dots)$$

- Beispiel: Wahrscheinlichster Vokal für Kombination der Formanten F1 und F2?

- Dazu: Lernen der Wahrscheinlichkeitsverteilung $P(K \mid f_1, f_2, f_3, \dots)$
-

Einführung in die Coli WS 04/05

Die fünf Schritte der Klassifikation

- Schritt 1: Klassen festlegen und Daten vorbereiten
 - Schritt 2: Informative Features **identifizieren**
 - Schritt 3: Modell **wählen und** auf Trainingsdaten **trainieren**
 - Schritt 4: Modell auf Testdaten **anwenden**
 - Schritt 5: Modell **evaluieren**
-

Einführung in die Coli WS 04/05

Ablauf statistischer Modellierung

- Schritt 1: Klassen festlegen und Daten vorbereiten
 - Schritt 2: Informative Features identifizieren
 - Schritt 3: Modell wählen und auf Trainingsdaten trainieren
 - Schritt 4: Modell auf Testdaten anwenden
 - Schritt 5: Modell evaluieren
-

Einführung in die Coli WS 04/05

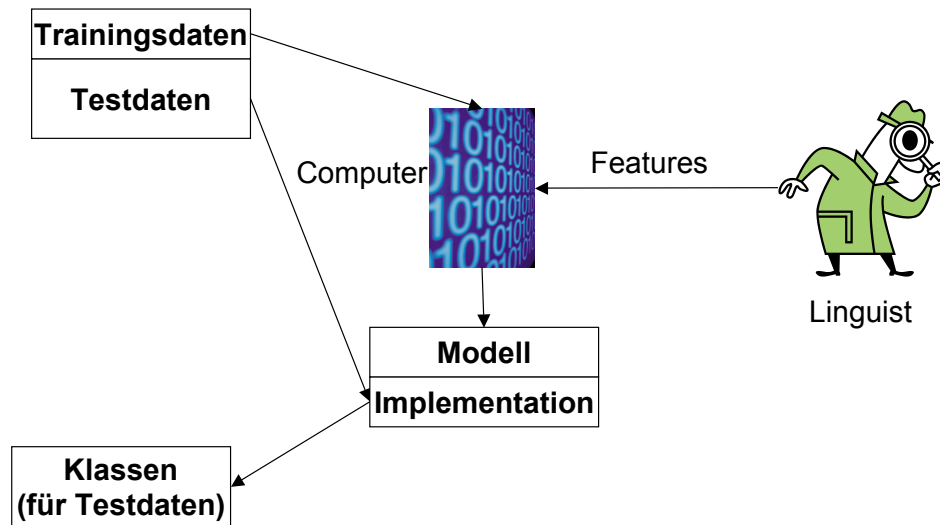
Beispiel: Klassen für Anbindung von Präpositionalphrasen (PPs)

Friedrich sieht den Mann mit dem Fernrohr

- Friedrich hat das Fernrohr:
 - Friedrich [_{VP} sieht [_{NP} den Mann] [_{PP} mit dem Fernrohr]]
 - Klasse 1: PP modifiziert VP (das Sehen)
 - Der Mann hat das Fernrohr:
 - Friedrich [_{VP} sieht [_{NP} den Mann [_{PP} mit dem Fernrohr]]]
 - Klasse 2: PP modifiziert NP (den Mann)
-

Einführung in die Coli WS 04/05

Modell und Daten



Einführung in die Coli WS 04/05

Trainings- und Testdaten

- Statistische Modelle werden auf einem Korpus entwickelt (trainiert) und auf einem anderen getestet
 - Man darf nicht auf genau denselben Daten trainieren und testen!
- **Overfitting:** Modell kann nicht zwischen allgemeinen Regelmäßigkeiten (die wir wollen) und speziellen Eigenheiten des Trainingskorpus unterscheiden
 - Modell ist besser auf Trainingskorpus
 - Modell ist schlechter auf allen anderen Daten
 - (Intuition: Training ist Hypothesen bilden – müssen an unabhängigen Daten verifiziert werden)

Einführung in die Coli WS 04/05

Trainings- und Testdaten

- Korpus unterteilt sich in:
 - Trainingsdaten (training set)
 - Testdaten (test set)
- Auf Trainingsdaten trainieren
- Auf Testdaten anwenden und evaluieren
- Präpositionalphrasen-Experiment:
 - Korpus: Computerhandbücher von IBM
 - 80% Trainingsdaten, 20% Testdaten

Einführung in die Coli WS 04/05

Ablauf statistischer Modellierung

- Schritt 1: Klassen festlegen und Daten vorbereiten
- Schritt 2: Informative Features identifizieren
- Schritt 3: Modell wählen und auf Trainingsdaten trainieren
- Schritt 4: Modell auf Testdaten anwenden
- Schritt 5: Modell evaluieren

Einführung in die Coli WS 04/05

Schritt 2: Features

- Welche beobachtbaren Ereignisse können die einzelnen Klassen am besten vorhersagen?
 - Wahl von Features ist sehr wichtig
 - Welche Information hilft, die Entscheidung zu treffen?
- Hauptfrage: wieviele Feature?
 - In Prinzip gilt: mehr Features, bessere Vorhersage
 - Aber: Manche statistischen Modelle funktionieren nur mit wenigen Features gut
 - Aber: Oft wenig Trainingsdaten vorhanden

Einführung in die Coli WS 04/05

Features für PP-Anbindung (I)

Friedrich sieht den **Astronomen** mit dem **Fernrohr** (NP)

Friedrich sieht den **Stern** mit dem **Fernrohr** (VP)

- Feature 1: Kopf der NP (n_1)

Einführung in die Coli WS 04/05

Features für PP-Anbindung (II)

Friedrich **sieht** den Astronomen mit der Gitarre (NP)

Friedrich **stört** den Astronomen mit der Gitarre (VP)

- Feature 2: Kopf der VP (v)

Features für PP-Anbindung (III)

Friedrich stört den Mann mit der **Sonnenbrille** (NP)

Friedrich stört den Mann mit der **Gitarre** (VP)

- Feature 3: Kopf der NP in der PP (n_2)

Ablauf statistischer Modellierung

- Schritt 1: Klassen festlegen und Daten vorbereiten
 - Schritt 2: Informative Features identifizieren
 - Schritt 3: Modell wählen und auf Trainingsdaten trainieren
 - Schritt 4: Modell auf Testdaten anwenden
 - Schritt 5: Modell evaluieren
-

Einführung in die Coli WS 04/05

Schritt 3: Training von Modellen

- Aus den Trainingsdaten muss die Wahrscheinlichkeit $P(c|f_1, \dots, f_n)$ für jede Klasse c und jede Kombination von Features f_1, \dots, f_n abgeschätzt werden (**Lernen / Estimation**)
- Einfachste Möglichkeit: Über **Frequenz**
 - Sogenannte Maximum Likelihood Estimation (MLE):

$$P(c|f_1, \dots, f_n) = \frac{P(c, f_1, \dots, f_n)}{P(f_1, \dots, f_n)} = \frac{f(c, f_1, \dots, f_n)}{f(f_1, \dots, f_n)}$$

Größtes Problem: Kombination (f_1, \dots, f_n) ist oft nie gesehen worden

Einführung in die Coli WS 04/05

Das „Sparse Data“-Problem

- Viele Features haben große Wertebereiche
 - Besonders „lexikalische“ Features
 - 10,000 Verben
 - 50,000 Nomen
 - Größe des **Featureraums** explodiert (Multiplikation)
 - PP-Anbindung: 3 Features, (n_1, v, n_2) hat $50\,000 \times 10\,000 \times 50\,000 = 25$ Billionen Featurekombinationen
 - Kann nie alle Kombinationen aller Features sehen!
 - **$P(\text{ungesehene Featurekombination}) = 0/0$**
(nicht definiert)
-

Einführung in die Coli WS 04/05

Sparse Data (II)

- Strategie 1: Modelle nicht direkt aus Frequenzen abschätzen
 - Komplexere statistische Modelle
 - Strategie 2a: Einfachere Modelle (weniger Features)
 - Strategie 2b: Lerne einfachere Modelle (weniger Features) und komplexere Modelle (mehr Features)
 - Kombination bei Anwendung
-

Einführung in die Coli WS 04/05

Ablauf statistischer Modellierung

- Schritt 1: Klassen festlegen und Daten vorbereiten
 - Schritt 2: Informative Features identifizieren
 - Schritt 3: Modell wählen und auf Trainingsdaten trainieren
 - Schritt 4: Modell auf Testdaten anwenden
 - Schritt 5: Modell evaluieren
-

Einführung in die Coli WS 04/05

Entscheidungsalgorithmus (Wiederholung)

- Konditionaler statistischer Klassifikator

$$P(\text{Klasse } K \mid \text{Instanz } I) = \\ P(\text{Klasse } K \mid \text{Features } f_1, \dots, f_n)$$

- Klassifiziere Instanz I mit Feature-Repräsentation f_i mit Klasse

$$K(I) = \operatorname{argmax}_K P(K \mid f_1, \dots, f_n)$$

Einführung in die Coli WS 04/05

Entscheidung bei PP-Anbindung

- 95% der 4-Tupel (v, n_1, n_2, p) in den Testdaten kommen nicht in den Trainingsdaten vor
- Lösung: „back-off“
 - Wenn du komplexen Modell vertrauen kannst, nimm es; sonst nimm einfacheres Modell
- Aufgabe: Klassifiziere (stört, Mann, Gitarre)
 - Wenn (stört, Mann, Gitarre) gesehen, klassifiziere Tripel
 - Sonst versuche Paare (stört, Mann), (stört, Gitarre), (Mann, Gitarre)
 - Wenn auch nicht gesehen, versuche (stört), (Mann), (Gitarre)

Einführung in die Coli WS 04/05

PP-Anbindung: Back-off

1. **If** $f(v, n_1, p, n_2) > 0$ Falls 4-Tupel gesehen

$$\hat{p}(1|v, n_1, p, n_2) = \frac{f(1, v, n_1, p, n_2)}{f(v, n_1, p, n_2)}$$

2. **Else if** $f(v, n_1, p) + f(v, p, n_2) + f(n_1, p, n_2) > 0$ Falls 3-Tupel gesehen

$$\hat{p}(1|v, n_1, p, n_2) = \frac{f(1, v, n_1, p) + f(1, v, p, n_2) + f(1, n_1, p, n_2)}{f(v, n_1, p) + f(v, p, n_2) + f(n_1, p, n_2)}$$

3. **Else if** $f(v, p) + f(n_1, p) + f(p, n_2) > 0$ Falls Paare gesehen

$$\hat{p}(1|v, n_1, p, n_2) = \frac{f(1, v, p) + f(1, n_1, p) + f(1, p, n_2)}{f(v, p) + f(n_1, p) + f(p, n_2)}$$

4. **Else if** $f(p) > 0$ Falls Präposition gesehen

$$\hat{p}(1|v, n_1, p, n_2) = \frac{f(1, p)}{f(p)}$$

5. **Else** $\hat{p}(1|v, n_1, p, n_2) = 1.0$ (default is noun attachment). Default

Einführung in die Coli WS 04/05

Ablauf statistischer Modellierung

- Schritt 1: Klassen festlegen und Daten vorbereiten
 - Schritt 2: Informative Features identifizieren
 - Schritt 3: Modell wählen und auf Trainingsdaten trainieren
 - Schritt 4: Modell auf Testdaten anwenden
 - Schritt 5: Modell evaluieren
-

Einführung in die Coli WS 04/05

Evaluation

- Wie gut ist eine Klassifikation?
 - Accuracy (Akkuratheit):
Prozent richtiger Klassifikationen
 - Error (Fehler):
Prozent Fehler
 - Precision (Präzision, Genauigkeit)
 - Recall (Vollständigkeit)
- Detaillierter, werden deshalb zunehmend benutzt
-

Einführung in die Coli WS 04/05

Konfusionsmatrix und Evaluationsmaße

	Echtes X	Echtes Nicht X
Als X klassifiziert	True positives	False positives
Als Nicht-X klass.	False negatives	True negatives

$$\text{Precision} = \frac{\text{True positives}}{\text{True positives} + \text{False positives}}$$

- Welcher Anteil der als X klassifizierten Instanzen hat wirklich Klasse X? (Genauigkeit)
- Werte zwischen 0 und 1 (höher = besser)

Einführung in die Coli WS 04/05

Konfusionsmatrix und Evaluationsmaße II

	Echtes X	Echtes Nicht X
Als X klassifiziert	True positives	False positives
Als Nicht-X klass.	False negatives	True negatives

$$\text{Recall} = \frac{\text{True positives}}{\text{True positives} + \text{False negatives}}$$

- Welcher Anteil der echten X wurde als X klassifiziert? (Vollständigkeit)
- Werte zwischen 0 und 1 (höher = besser)

Einführung in die Coli WS 04/05

F-Score

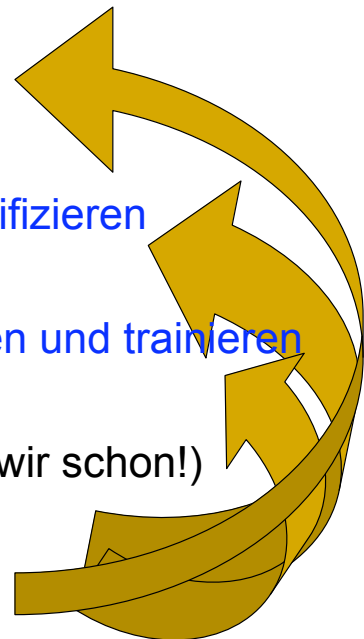
- Nur gute Precision nutzt nichts
- Nur guter Recall nutzt auch nichts
 - Nichts als X klassifiziert: P=100%, R=0%
 - Alles als X klassifiziert: P=0%, R=100%
- **F-Score**: Kombination aus P und R:
 - Ein Maß für „Gesamtgüte“ der Klassifikation
 - Werte zw. 0 und 1 (höher = besser)
 - Bevorzugt „true positives“

$$F = \frac{2PR}{P+R}$$

Einführung in die Coli WS 04/05

Gehen Sie nicht über Los...

- Schritt 1: Klassen festlegen
- Schritt 2: Informative Features **identifizieren**
- Schritt 3: Statistisches Modell **wählen und trainieren**
- Schritt 4: Modell **anwenden** (haben wir schon!)
- Schritt 5: Modell **evaluieren**



Einführung in die Coli WS 04/05

PP-Attachment: Evaluation

- Daten: Computer-Handbücher von IBM
 - 80% Trainingsdaten, 20% Testdaten
- Ergebnis:
 - Modelle mit vielen Features sehr gut, aber schlechte Abdeckung (90%/10%)
 - Modelle mit wenig Features schlecht, aber mit guter Abdeckung (30% / 80%)
 - Backing-off-Modell: 84% Accuracy

Einführung in die Coli WS 04/05

Literatur

- Die Bibel der statistischen Sprachverarbeitung:
 - Manning & Schütze: Foundations of statistical language processing (MIT Press 1999)
- Einführung in die Verwendung statistischer Methoden:
 - Steven Abney: Statistical Methods and Linguistics
 - David Magerman: Everything you wanted to know about probability theory but were afraid to ask

Einführung in die Coli WS 04/05