

Statistische Modellierung in der Computerlinguistik

Sebastian Pado

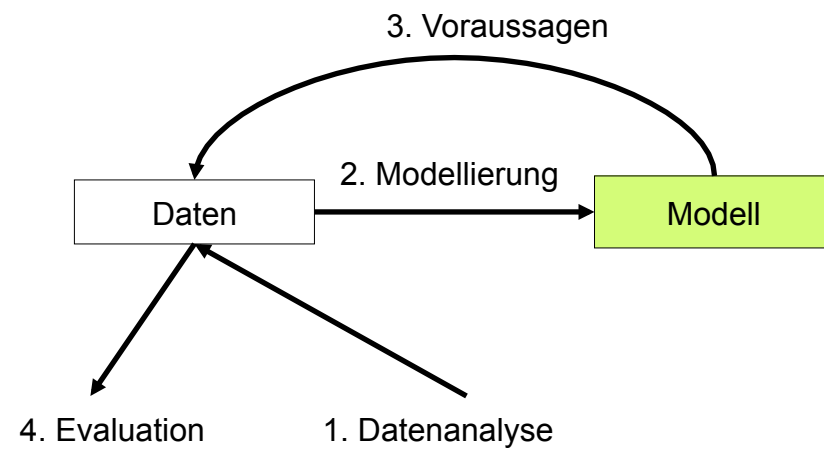
1

Computerlinguistik 2009

- ▶ **Ziel:** Modelle zur automatischen linguistischen Analyse von Text
 - ▶ Robust
 - ▶ Mit geringem Aufwand zu konstruieren
- ▶ **Methode:**
 - ▶ Ein Korpus mit der entsprechenden linguistischen Ebene annotieren
 - ▶ **Muster** aus diesem Korpus ablesen und verwenden, um für neue Korpora automatisch eine linguistische Analyse zu produzieren

▶ 2

Empirische Modellierung



▶ 3

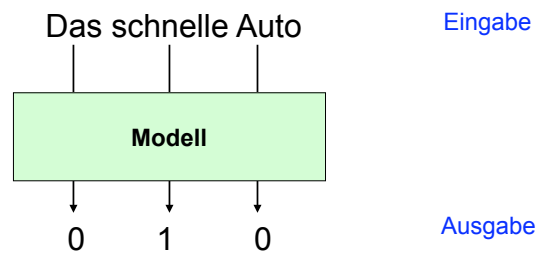
Fragen für heute

1. **Wie sehen diese Muster aus?**
 - ▶ **Features**
2. **Auf welche Weise werden Features zur Vorhersage verwendet?**
 - ▶ **Wahrscheinlichkeiten**
3. **Wie evaluiert man Modelle?**
4. **Welchen Einfluß hat Annotation auf die Modellierung?**

▶ 4

Beispielaufgabe: Wortarterkennung

- ▶ Handelt es sich bei einem Wort (in einem fortlaufenden Text) um ein Adjektiv?
 - ▶ Binäre Entscheidung (Klassifikation)



- ▶ Wie konkret modellieren?

▶ 5

Option 1

- ▶ **Lexikonbasiertes Modell**
 - ▶ Prüfe, ob Wort in Adjektivliste ist
- ▶ **Problem: Woher eine Adjektivliste nehmen?**
 - ▶ Adjektive sind offene Wortklasse
 - ▶ Neue Adjektive entstehen ständig (z.B. durch Derivation)
- ▶ **Problem: Ambiguität**
 - ▶ „schnelle“ kann Adjektiv sein, oder Verb (1.sg.präs.ind oder 3.sg.präs.konj von schnellen)
 - ▶ Lexikon beschreibt i.A. Kontext nicht

▶ 6

Option 2: Datenbasierte Modellierung

Woran erkenne ich ein Adjektiv?

Ich moechte Ihnen fuer den Bericht ueber
den **siebenten** Bericht ueber
staatliche Beihilfen in
der **europäischen** Union danken.

Gesucht: Muster

- ▶ die für (möglichst) alle Adjektive wahr und für alle anderen Wörter falsch sind
- ▶ die einfach genug sind, um von einem Computer berechnet zu werden (kein Rückgriff auf Weltwissen etc.)

▶ 7

Option 2a: Symbolische Regeln

- ▶ Welche Regeln können wir aufstellen?
 1. Nächstes Wort kapitalisiert ⇒ Adj
Sonst NAdj

Ich **möchte** Ihnen für **den** Bericht über
den **siebenten** Bericht über
staatliche Beihilfen in der
europäischen Union danken.

▶ 8

Option 2a: Symbolische Regeln

▶ Welche Regeln können wir aufstellen?

1. Nächstes Wort kapitalisiert ⇒ Adj
2. Nächstes Wort kapitalisiert und Wort kein Artikel ⇒ Adj
Sonst NAdj

Ich möchte Ihnen für den Bericht über
den siebenten Bericht über
staatliche Beihilfen in der
europäischen Union danken.

▶ 9

Option 2a: Symbolische Regeln

▶ Welche Regeln können wir aufstellen?

1. Nächstes Wort kapitalisiert ⇒ Adj
2. Nächstes Wort kapitalisiert und Wort kein Artikel ⇒ Adj
3. Nächstes Wort kapitalisiert und Wort kein Artikel und
vorheriges Wort Artikel ⇒ Adj
Sonst NAdj

Ich möchte Ihnen für den Bericht über
den siebenten Bericht über
staatliche Beihilfen in der
europäischen Union danken.

▶ 10

Option 2a: Symbolische Regeln

- ▶ Es ist schwer, **korrekte** und **vollständige** Regeln zu schreiben.
 - ▶ Regel 2 ist zu liberal
 - ▶ Regel 3 ist zu streng
- ▶ Das System trifft eine harte Entscheidung für jede Instanz
 - ▶ Keine Möglichkeit, über „Konfidenz“ zu sprechen

▶ 11

Option 2b: Statistische Modelle

- ▶ Wir extrahieren einzelne Bedingungen aus den Regeln: **Features**

Nachfolgendes Wort ist kapitalisiert
Wort ist kein Artikel
Vorhergehendes Wort ist Artikel
Vorhergehendes Wort ist Gradpartikel

- ▶ ...und versehen jede Bedingung mit einer Zahl

<i>Nachfolgendes Wort ist kapitalisiert</i>	45%
<i>Wort ist kein Artikel</i>	10%
<i>Vorhergehendes Wort ist Artikel</i>	60%
<i>Vorhergehendes Wort ist Gradpartikel</i>	85%

- ▶ Interpretation der Zahl: Wahrscheinlichkeit für Klasse “Adjektiv” unter allen Wörter mit diesem Feature
 - ▶ Gradierter (“weicher”) Zusammenhang zwischen Features und Klasse

▶ 12

Statistische Modellierung

- ▶ Fast alle sprachtechnologischen Anwendungen (und sehr viele Forschungsmodelle) in der CL haben heute eine probabilistische (wahrscheinlichkeitsbasierte) Komponente
- ▶ Grund: Wahrscheinlichkeitstheorie bietet eine feste mathematische Grundlage für die Manipulation von quantitativem Wissen
- ▶ Preis: Entscheidung in den Modellen läuft auf einer abstrakten (von Menschen nicht gut nachvollziehbaren) Ebene ab
 - ▶ Präsenz/Fehlen von einzelnen Features

▶ 13

Fragen für heute

1. Wie sehen diese Muster aus?
 - ▶ Features
2. Auf welche Weise werden Features zur Vorhersage verwendet?
 - ▶ Wahrscheinlichkeiten
3. Wie evaluiert man Modelle?
4. Welchen Einfluß hat Annotation auf die Modellierung?

▶ 14

Wahrscheinlichkeiten

- ▶ **Wahrscheinlichkeiten reden über Ereignisse**
 - ▶ Der Würfel zeigt eine 6, Es regnet, Das Wort ist ein Nomen
- ▶ **Wahrscheinlichkeiten sind Zahlen zwischen 0 und 1**
 - ▶ $P(\text{Der Würfel zeigt eine 6}) = 1/6$
 - ▶ $P(\text{Es regnet}) = ??$
- ▶ **Wahrscheinlichkeiten können als relative Frequenzen verstanden werden**
 - ▶ $P(\text{Es regnet}) = \frac{f(\text{Anzahl Tage mit Regen})}{f(\text{Anzahl beobachtete Tage})}$
 - ▶ $P(\text{Der Würfel zeigt eine 6}) = \frac{f(\text{Anzahl Würfe mit 6})}{f(\text{Anzahl Würfe})}$

▶ 15

Wahrscheinlichkeiten

- ▶ **Komplexe Ereignisse:**
 - ▶ **Gemeinsame** Ereignisse: Der Würfel zeigt gerade Zahl **und** zeigt eine Zahl ≥ 4
 - ▶ **Abhängige** Ereignisse: **Wenn** es Dezember ist, **dann** regnet es
- ▶ **Berechnung von gemeinsamen Wahrscheinlichkeiten**
 - ▶ $P(A \text{ und } B) = \frac{f(A \text{ und } B)}{N}$
 - ▶ $P(\text{Würfelzahl ist gerade und } \geq 4) = 2/6 = 1/3$
- ▶ **Berechnung von bedingten Wahrscheinlichkeiten**
 - ▶ $P(A | B) = \frac{f(A \text{ und } B)}{f(B)}$
 - ▶ $P(\text{Regen} | \text{Dezember}) = \frac{\text{Anzahl der Regentage im Dezember}}{\text{Anzahl Dezembertage}}$

▶ 16

Berechnung von Wahrscheinlichkeiten

- ▶ Alle Zahlen können in **Frequenzmatrix** gesammelt werden
 - ▶ Zeilen: Eingabeereignisse (Wörter)
 - ▶ Spalten: Ausgabeereignisse (Klassen)
 - ▶ Zellen: Frequenzen des **gemeinsamen Auftretens von zwei Ereignissen** im Korpus
 - ▶ Alle Frequenzen für Paare von Eingaben und Ausgaben $f(E1 \text{ und } E2)$, z.B. $f(\text{Das und Adj})$

Ereignis	Adj	Nadj
Das	0	25500
schnelle	4000	500
Auto	0	10000
...

▶ 17

Wahrscheinlichkeiten für Ereignisse

- ▶ Was uns interessiert: Beziehung zwischen Eingabe- und Ausgabeereignissen
 - ▶ **Bedingte Wahrscheinlichkeiten**, z.B. $P(\text{Adj} \mid \text{Das})$
 - ▶ Wenn das Wort "das" ist, wie wahrscheinlich ist es dann ein Adjektiv?
 - ▶ Wir wissen: $P(A \mid B) = f(A \text{ und } B) / f(B) = \text{Zelle} / \text{Zeile}$
 - ▶ z.B. $P(\text{Adj} \mid \text{schnelle}) = f(\text{Adj und schnelle}) / f(\text{schnelle}) = 4000/4500 = 8/9 = 89\%$

Ereignis	Adj	Nadj
Das	0	25500
schnelle	4000	500
Auto	0	10000
...

▶ 18

Vorhersagen

- ▶ Wenn eine binäre Vorhersage gewünscht ist:
 - ▶ Für jede Zeile sage die Spalte (Klasse) mit der höchsten Wahrscheinlichkeit vorher.

Ereignis	Adj	Nadj
Das	0	25500
schnelle	4000	500
Auto	0	10000
...

▶ 19

Grenzen dieser Matrix

Ereignis	Adj	Nadj
Das	0	25500
schnelle	4000	500
Auto	0	10000
Xylophon	0	0

- ▶ Keine Vorhersagen für neue Wörter
 - ▶ $P(\text{Nadj} \mid \text{Xylophon}) = f(\text{Nadj und Xyl.}) / f(\text{Xyl.}) = 0 / 0 = \text{n.d}$
- ▶ Kein Einfluß des Kontext auf Vorhersagen
 - ▶ “schnelle vom Bogen” und “schnelle Strecke”: gleiche Vorhersage

▶ 20

“Featurisierung”

- ▶ Idee: Zeilen (Eingabeereignisse) in Matrix sind nicht Wörter, sondern **Kombinationen von Features**

„Das“	„schnelle“	„Auto“
1 nein	1 ja	1 nein
2 nein	2 ja	2 ja
3 nein	3 ja	3 ja
4 ja	4 ja	4 ja

1. Nachfolgendes Wort ist kapitalisiert
2. Wort ist kein Artikel
3. Vorhergehendes Wort ist Artikel
4. Folgendes Wort ist keine Präposition

▶ 21

Feature-Matrix

- ▶ Zeilen sind Klassen von Wörtern, die sich ähnlich verhalten

1. Nachfolgendes Wort ist kapitalisiert
2. Wort ist kein Artikel
3. Vorhergehendes Wort ist Artikel
4. Folgendes Wort ist keine Präposition

Ereignis	Adj	Nadj
ja, ja, ja, ja	10000	5
ja, ja, ja, nein	0	200
nein, ja, nein, nein	4000	5000
nein, nein, nein, nein	0	0

“der **schnelle** Wagen”“das **reicht**.Von”“und **schnell** in”, “und **auch** in”“auch **der** Mann”

▶ 22

Vorhersage

- Sage die Klasse mit der höchsten Wahrscheinlichkeit vorher

1. Nachfolgendes Wort ist kapitalisiert
2. Wort ist kein Artikel
3. Vorhergehendes Wort ist Artikel
4. Folgendes Wort ist keine Präposition

Ereignis	Adj	Nadj
ja, ja, ja, ja	10000	5
ja, ja, ja, nein	0	200
nein, ja, nein, nein	4000	5000
nein, nein, nein, nein	0	0

“der schnelle Wagen”

“das reicht.Von”

“nur frisches in”, “und auch in”

“auch der Mann”

► 23

Einfluß von Kontext auf Vorhersage

- Jetzt kann dasselbe Wort in unterschiedlichen Kontexten unterschiedliche Klassen zugewiesen bekommen

Ereignis	Adj	Nadj
ja, ja, ja, ja	10000	5
ja, ja, ja, nein	0	200
nein, ja, nein, nein	4000	5000
nein, nein, nein, nein	0	0

“der schnelle Wagen”: Adj

“Pfeil schnelle vom Bogen”: NAdj

1. Nachfolgendes Wort ist kapitalisiert
2. Wort ist kein Artikel
3. Vorhergehendes Wort ist Artikel
4. Folgendes Wort ist keine Präposition

► 24

Vorhersage für unbekannte Wörter

▶ Jetzt kann neuen Wörter eine Klasse zugewiesen werden!

▶ Schritt I: Berechnung von Features

„Das“	„rüfelige“	„Auto“
1 nein	1 ja	1 nein
2 nein	2 ja	2 ja
3 nein	3 ja	3 ja
4 ja	4 ja	4 ja

1. Nachfolgendes Wort ist kapitalisiert
2. Wort ist kein Artikel
3. Vorhergehendes Wort ist Artikel
4. Folgendes Wort ist keine Präposition

▶ Schritt 2: Ablesen der Klasse für diese Features aus der Matrix

Ereignis	Adj	Nadj
ja, ja, ja, ja	10000	5
ja, ja, ja, nein	0	200
...
nein, ja, nein, nein	4000	5000
nein, nein, nein, nein	0	0

▶ 25

Zusammenfassung

▶ Training:

- ▶ Für jedes Ereignis mit bekannter Klasse:
 - ▶ Berechne Features
 - ▶ Erhöhe $f(\text{Features}, \text{Klasse})$
- ▶ Speichere Matrix

▶ Vorhersage:

- ▶ Für jedes Ereignis mit unbekannter Klasse:
 - ▶ Berechne Features
 - ▶ Für jede Klasse: berechne $P(\text{Klasse} | \text{Features}) = \frac{f(\text{Klasse}, \text{Features})}{f(\text{Features})} = \text{Zelle} / \text{Zeile}$ aus Matrix
- ▶ Vorhersage ist Klasse mit der höchsten Wahrscheinlichkeit

▶ 26

Wo steckt die Schwierigkeit bei der Entwicklung statistischer Modelle?

▶ 27

Fehler

Peter, der **sieht** Paul

- 1 ja
- 2 ja
- 3 ja
- 4 ja

- 1. Nachfolgendes Wort ist kapitalisiert
- 2. Wort ist kein Artikel
- 3. Vorhergehendes Wort ist Artikel
- 4. Folgendes Wort ist keine Präposition

Ereignis	Adj	Nadj
ja, ja, ja, ja	10000	5
ja, ja, ja, nein	0	200
nein, ja, nein, nein	4000	5000
nein, nein, nein, nein	0	0

▶ 28

Intelligenter Modelle = mehr Features

- ▶ Welche weiteren Muster könnte man ausnutzen, um Adjektive zu identifizieren?
 - ▶ Wortendungen (morphologische Information)
 - ▶ Attributive Adjektive können nicht auf -t enden wie "sieht"
 - ▶ Gradpartikel stehen fast immer vor Adjektiven
 - ▶ "sehr", "besonders", ...
 - ▶ ...
- ▶ Wieso verwendet man nicht einfach alle Features, die einem einfallen?

▶ 29

Größe des Ereignisraumes

- ▶ Wieviele Zeilen hat die Modell (Tabelle)?
 - ▶ Anzahl möglicher verschiedener Ereignisse
 - ▶ Produkt der Anzahl möglicher Werte aller Features
 - Beispiel: **Selbst Artikel?** x **Nächstes Wort kapitalisiert** = $2 \times 2 = 4$
 - Lexikalische Features: pro Feature oft > 10.000 Werte
- ▶ Frequenzen in Trainingskorpus werden auf Zeilen (Ereignisse) verteilt
 - ▶ Wenn Größe des Trainingskorpus $<$ Anzahl möglicher Ereignisse, treten in den Testdaten **ungesehene** Ereignisse auf:
Modell kann keine Vorhersage machen
- ▶ Das "Sparse Data"-Problem

▶ 30

Sparse Data

Auch die **kleine** Fraktur

1 ja
2 Fraktur

1. Wort ist kein Artikel
2. Folgendes Wort

Ereignis	Adj	Nadj
ja, braucht	500	50000
ja, Beinbruch	1000	50
ja, Femularfraktur	0	0

▶ 31

Weitere Strategien

▶ Das Dilemma:

- ▶ Je mehr Features, desto besser die Datenlage für die Entscheidung
- ▶ Je mehr Features, auf desto mehr Ereignisse verteilen sich die Trainingsdaten

▶ Strategie 1: Entwicklung besserer Features

- ▶ Weniger Lexikalisierung
- ▶ Linguistische Einsicht!!
- ▶ Automatische Methoden zur Featureauswahl

▶ Strategie 2: Bessere statistische Modelle

- ▶ Entscheidung nicht direkt aus Frequenzen abschätzen
- ▶ Forschungsthema im Maschinellen Lernen

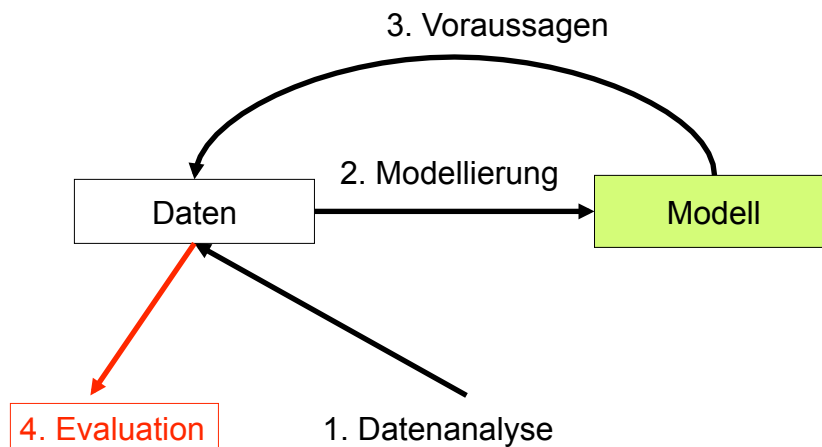
▶ 32

Fragen für heute

1. Wie sehen diese Muster aus?
 - ▶ Features
2. Auf welche Weise werden Features zur Vorhersage verwendet?
 - ▶ Wahrscheinlichkeiten
3. **Wie evaluiert man Modelle?**
4. Welchen Einfluß hat Annotation auf die Modellierung?

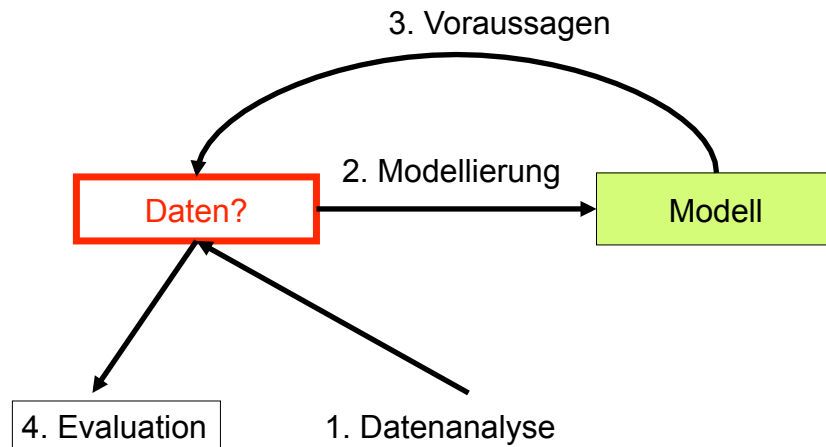
▶ 33

Evaluation



▶ 34

Training vs. Testing



▶ 35

Daten



- Meistens zwischen 70 und 90% der Daten
- Grundlage der Modellierung, Quelle für Features und Frequenzen (Schritt 1, 2)
- Überprüfung des Modells an **unabhängigen Daten** (Schritt 4)

▶ 36

Beispielevaluation

- ▶ 1000 Instanzen in den Testdaten
- ▶ Aufgabe: Klassifikation nach Adjektiv / Nicht-Adjektiv
- ▶ Konfusionsmatrix
 - ▶ Zeilen: Zuweisungen des Modells
 - ▶ Spalten: „Echte“ Klassen
 - ▶ Diagonale: **Richtige Zuweisungen**
 - ▶ Andere Zellen: Fehler

	Echtes Adj	Echtes Nadj
Als Adj klass.	10	80
Als Nadj klass.	10	900

▶ 37

Accuracy / Error

- ▶ Einfachstmögliche Evaluation
- ▶ Accuracy = (# richtige Instanzen) / (# alle Instanzen)
= (# Diagonale) / (# ganze Matrix)
- ▶
$$= \frac{\text{\# Instanzen:Vorhersage = Gold-Klasse}}{\text{\# alle Instanzen}}$$
 - ▶ Hier: $(10 + 900) / 1000 = 910 / 1000 = 91\%$
- ▶ Error = (1 - Accuracy)
 - ▶ Hier: 9%
- ▶ Ist die Evaluation angemessen?

▶ 38

Problem von Accuracy-Evaluation

- ▶ Macht keinen Unterschied zwischen (Gold-)Klassen (Adj vs. Nadj)
 - ▶ Wieso unterscheiden?
 - ▶ In der Computerlinguistik sind die interessanten Klassen oft klein
 - ▶ In Gesamtevaluation geht Qualität der kleinen Klasse unter
- ▶ Nötig: **Klassenspezifische** Fehler
- ▶ I. Approximation: Klassenspezifische Accuracy / Error
 - ▶ 10 / 20 korrekt für Adj: 50%
 - ▶ 900 / 980 korrekt für Nadj: 91.8%
- ▶ Aber: Akkuratheit unterscheidet nicht zwischen Fehlerarten

▶ 39

Fehlerarten

- ▶ 80 Instanzen wurden falsch klassifiziert. Was stellen Sie für eine Art von Fehler dar?
 - ▶ Hängt von der betrachteten Klasse ab

	Echtes Adj	Echtes Nadj
Als Adj klass.	10	80
Als Nadj klass.	10	900

▶ 40

Fehlerarten

▶ Fehlerart 1: Korrektheitsfehler von X (“false positive”):

- ▶ Eine Instanz ist kein X, wird aber vom Modell als X klassifiziert
- ▶ Hier: 80 Instanzen sind kein Adj, werden aber vom Modell als Adj klassifiziert
= **Korrekttheitsfehler für Klasse Adj**

	Echtes Adj	Echtes Nadj
Als Adj klass.	10	80
Als Nadj klass.	10	900

▶ 41

Fehlerarten

▶ Fehlerart 2: Vollständigkeitsfehler von X (“false negative”):

- ▶ Eine Instanz ist ein X, wird aber vom Modell nicht als X klassifiziert
- ▶ Hier: 80 Instanzen sind Nadj, werden aber vom Modell nicht als Nadj klassifiziert
= **Vollständigkeitsfehler für Klasse Nadj**

	Echtes Adj	Echtes Nadj
Als Adj klass.	10	80
Als Nadj klass.	10	900

▶ 42

Vollständigkeits- vs. Korrektheitsfehler

- ▶ Warum zwischen Vollständigkeits- und Korrektheitsfehler unterscheiden?
 - ▶ Accuracy behandelt beide Fehlerarten gleich

- ▶ Die Fehler können unterschiedlich wichtig sein
 - ▶ Für manche Klassen sind Vollständigkeitsfehler schlimm
 - ▶ “Hat dieser Patient ein erhöhtes Herzanfallrisiko?”
 - ▶ Für andere Klassen sind Korrektheitsfehler schlimmer
 - ▶ “Ist dieser Patient HIV-positiv?”

▶ 43

Recall

	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)

▶

44

Recall für Klasse Adj

	Echtes Adj	Echtes Nadj
Als Adj klassifiziert	10	80
Als Nadj klass.	10	900

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

- ▶ Hier: $10/(10+10) = 0.5$
- ▶ Interpretation: Die Hälfte aller echten Adjektive wurde durch das Modell richtig erkannt

45

Precision

	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 ist wirklich ein X? (Korrektheit)
- ▶ Werte zwischen 0 und 1 (höher = besser)

46

Precision für Klasse Adj

	Echtes Adj	Echtes Nadj
Als Adj klassifiziert	10	80
Als Nadj klass.	10	900

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

- ▶ Hier: $10 / (10+80) = 11\%$
- ▶ Interpretation: Wenn das Modell behauptet, eine Instanz sei ein Adj, ist das nur 11% der Fälle wahr

47

Precision und Recall

- ▶ Precision und Recall sollten zusammen betrachtet werden
 - ▶ Hohe Precision, hoher Recall: fast perfekte Klassifikation
 - ▶ Niedrige Precision, niedriger Recall: sehr schlechte Klassifikation
 - ▶ Hohe Precision, niedriger Recall: "vorsichtiges Modell"
 - ▶ Findet nicht alle Instanzen von X
 - ▶ Klassifiziert fast keine Nicht-Xe als X
 - ▶ Niedrige Precision, hoher Recall: "mutiges Modell"
 - ▶ Findet fast alle Instanzen von X
 - ▶ Klassifiziert auch Nicht-Xe als X

▶ 48

Extremfälle..und die Kombination

▶ Extremfälle

- ▶ Modell klassifiziert alles als X
 - ▶ Recall für X ist 100%, Precision sehr niedrig
- ▶ Modell klassifiziert nichts als X
 - ▶ Recall für X ist 0%, Precision nicht definiert (0/0)

▶ F-Score: Kombination aus P und R:

- ▶ Ein Maß für „Gesamtgüte“ der Klassifikation
 - ▶ Werte zw. 0 und 1 (höher = besser)
 - ▶ Harmonisches Mittel von P und R
- ▶ Bevorzugt „Balance“ zwischen P und R
 - ▶ immer kleiner als arithmetischer Durchschnitt

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

▶ 49

F-Score für Klasse Adj

	Echtes Adj	Echtes Nadj
Als Adj klassifiziert	10	80
Als Nadj klass.	10	900

- ▶ Precision: $10 / (10+80) = 0.11$
- ▶ Recall: $10 / (10+10) = 0.5$
- ▶ F-Score: $(2 * 0.5 * 0.11) / (0.5 + 0.11) = 0.18$

▶

50

Fragen für heute

1. Wie sehen diese Muster aus?
 - ▶ Features
2. Auf welche Weise werden Features zur Vorhersage verwendet?
 - ▶ Wahrscheinlichkeiten
3. Wie evaluiert man Modelle?
4. Welchen Einfluß hat Annotation auf die Modellierung?

▶ 51

Korpusverarbeitung

Textdateien (Webseiten, Zeitungsarchive, etc.)

Auswahl

Stichprobe

Vorverarbeitung

“Rohes” Textkorpus

(manuelle) Annotation

Annotiertes Textkorpus

▶ 52

Annotation: Korrektheit

- ▶ Wichtigstes Kriterium: Korrektheit
 - ▶ Falsche Annotation führt zu falschen Modellen
 - ▶ **Manuelle Annotation**
- ▶ Selbst manuelle Annotation ist nie fehlerfrei
 - ▶ Grund 1: Unaufmerksamkeit der Annotatoren
 - ▶ Grund 2: Schwierigkeit der Aufgabe
- ▶ Nötig:
 1. Überprüfung der Korrektheit
 2. Entscheidung über Granularität

▶ 53

Annotation: Qualitätssicherung

- ▶ Annotation muß über die Zeit gleich bleiben (hohes **Intra-Annotator Agreement**)
 - ▶ Denselben Annotator **mehrmals** annotieren lassen (in zeitlichem Abstand)
- ▶ Mehrere Annotatoren müssen gleich annotieren (hohes **Inter-Annotator Agreement**)
 - ▶ Mehrere **unabhängige** Annotatoren annotieren dasselbe

▶ 54

Annotationsschema

- ▶ Definiert die Kategorien, die annotiert werden

- ▶ Definiert die Bedingungen, unter denen jede Kategorie annotiert wird
 - ▶ Richtlinien

▶ 55

Annotationsschema: Nomen-Wortarten

- ▶ Penn Tagset (45 Kategorien)
 - ▶ NN – noun, singular
 - ▶ NNS – noun, plural
 - ▶ NNP – proper noun, singular
 - ▶ NNPS – proper noun, plural

▶ 56

Annotationsschema: Nomen-Wortarten

▶ CLAWS2-Tagset (132 Kategorien)

- ▶ ND1 – singular noun of direction (north, southeast)
- ▶ NN / NNI / NN2 – common noun, neutral / sg / pl (cod / book / books)
- ▶ NNI\$ -- genitive singular common noun (domini)
- ▶ NNJ / NNJ1 / NNJ2 – organization noun (department / assembly)
- ▶>NNL />NNLI />NNL2 – locative noun (Is. / street / roads)
- ▶ NNO / NNO1 / NNO2 – numeral noun (dozen / ? / hundreds)
- ▶ NNS / NNS1 / NNS2 – noun of style (? / president / viscounts)
- ▶ NNSA1 / NNSA2 – following noun of style abbreviation (M.A.)
- ▶ NNSB / NNSB1 / NNSB2 – preceding noun of style abbreviation (Prof.)
- ▶ NNT / NNT1 / NNT2 – temporal noun (? / day / days)
- ▶ NNU – unit of measurement (in., inch / inches)
- ▶ NP / NP1 / NP2 – proper noun (Andes / London / Korea)
- ▶ NPD1 / NPD2 – weekday noun (Sunday / Sundays)
- ▶ NPM1 / NPM2 – month noun (October / Octobers)

▶ 57

Annotationsschemata

▶ Wie detailliert soll die Annotation sein?

- ▶ Detaillierte Annotation
 - ▶ Viele Kategorien, viel Information
 - ▶ Viele Zweifelsfälle (schwer, Qualität zu halten)
- ▶ Grobe Annotation
 - ▶ Wenige Kategorien, wenig Information
 - ▶ Einfacher, Qualität zu halten

▶ 58

Zweifelsfälle

- ▶ Annotationsschema muß Richtlinien für Zweifelsfälle beinhalten
 - ▶ Aber: oft bleiben systematische Zweifelsfälle
- ▶ Produktive Verwendung von Sprache kann Theorien sprengen
- ▶ Problem für tiefere Ebenen (Semantik!): häufige **Vagheit/Ambiguität**

Zwiebel (1): Zwiebelpflanze
 Zwiebel (2): Frucht der Zwiebelpflanze

- ▶ Was ist „Ich habe eine Zwiebel gepflanzt“?

▶ 59

Annotation: Aufwand

- ▶ Annotationsaufwand für ein Wort: 30 Sekunden
- ▶ 1M Worte: 500 000 Minuten = 5 Jahre
- ▶ plus Aufwand fuer Qualitätsicherung
- ▶ Beschleunigung: Annotatoren unterstützen
 - ▶ (Semi)-Automatisierung und manuelle Überprüfung
 - ▶ Zweiseitiges Schwert: Kann zu **systematischen Fehlern** führen

▶ 60