Einführung in die Computerlinguistik

Statistische Modellierung und Evaluation

WS 2008/2009

Manfred Pinkal

Mehrdeutigkeit der Wortart

- *Sie haben in Moskau liebe genossen*
- *Sie haben in Moskau liebe Genossen*
- *Sie haben in Moskau Liebe genossen*

liebe: V oder A?

Liebe am Satzbeginn: N oder V oder A?

- *I made her duck*

Mehrdeutigkeit der Wortart

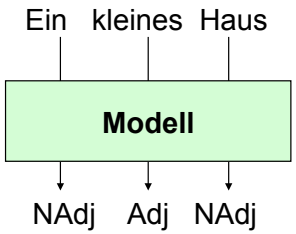
- Mögliche Wortarten der Wortform *zu*?
  - Die Tür ist zu.
  - Er ist nett zu mir.
  - Er ist mir zu nett.
  - Er plant, seine Freunde zu besuchen.

Wortart-Tagger

- Die Wortartinformation ist für viele sprachtechnologische Anwendungen wichtig: Sie ist Vorausetzung für die grammatische Analyse und hilfreich bei Lemmatisierung und morphologischer Analyse.
- Wortartinformation wird durch „Wortart-Tagger" oder „POS-Tagger" bereitgestellt. „POS" wegen engl. „part of speech" für die Wortart. „Tagger", weil die zentrale Funktion der Systeme darin besteht, Textwörtern in Sätzen und Dokumenten eine Wortart-Markierung (engl. „tag") zuzuordnen.
- Wortart-Tagger sind heute Standard-Werkzeuge der Sprachverarbeitung. Sie arbeiten teils mit regelbasierter, teils mit statistischer Modellierung. Die Zuverlässigkeit ist sehr hoch (>99,5%).

Beispielaufgabe: Adjektiverkennung

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



Eingabe

Ausgabe

Wortarterkennung

- **Lexikon**basiertes Modell
  - Prüfe (mithilfe von Lemmatisierer/ Morphologie), ob die Wortform zu einem Lemma im Lexikon gehört, entnehme Wortartinformation aus dem Lexikon
    - Problem1: Wortart kann mehrdeutig sein
    - Problem2: N, V, A sind offene Wortklassen
- **Muster**basierte Modelle, die den Kontext einbeziehen:
  - Option 1: Regelbasierte Modelle: Schreibe explizite Regeln der Form
    `Wenn <Muster>, dann <Wortart>`
    - **Symbolisches** Modell
  - Option 2: Datenbasierte Modelle: Lerne Abhängigkeiten zwischen Mustern und Wortart aus Korpus
    - **Statistisches** Modell

Datenanlyse: Informative Muster für die Adjektiverkennung

Intuitive Frage: **Woran erkenne ich ein Adjektiv**?

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

Datenanlyse: Informative Muster für die Adjektiverkennung

Intuitive Frage: **Woran erkenne ich ein Adjektiv**?

- Nächstes Wort ist großgeschrieben (kapitalisiert)
  - `Der **siebente** Bericht`
- Vorheriges Wort ist Artikel
  - `Der **siebente** Bericht`
- Wort selbst ist nicht kapitalisiert
  - `Der **siebente** Bericht`
- Wort selbst ist kein Artikel

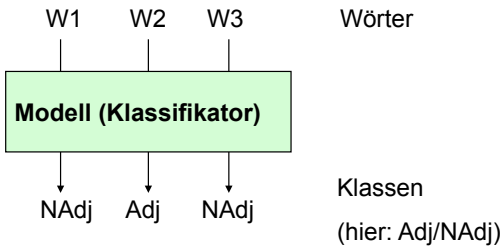
Symbolische Regeln

- Welche expliziten Regeln können wir aufstellen?
  - Nach Artikel, vor groß geschriebenem Wort, aktuelles Wort kein Artikel ⇒ Adj
    - das meiste Geld (Korrektheitsproblem)
  - Nächstes Wort nicht kapitalisiert oder kein vorangehender Artikel ⇒ NAdj
    - der fleißige, aber nicht erfolgreiche Student
    - große Bedenken (Vollständigkeitsproblem)

Es ist schwer, explizite Regeln zu schreiben, die sowohl korrekt als auch vollständig sind

Adjektiverkennung als statistische Klassifikationsaufgabe

- **Alternative Idee:** Zusammenhang zwischen Mustern und "Adjektivstatus" automatisch (maschinell) lernen!
- **Klassifikationsaufgabe**: jedem Eingabe-Ereignis wird eine Klasse (aus einer kleinen Menge) zugeordnet



Wahrscheinlichkeiten und relative Häufigkeit

- Die Klassifikation erfolgt wahrscheinlichkeitsbasiert.
- Die Wahrscheinlichkeit wird auf der Grundlage relativer Häufigkeiten geschätzt / approximiert.

Beispiel1

- Wir versuchen herauszufinden, wie hoch die Wahrscheinlichkeit ist, dass wir mit einem Würfel als Nächstes eine „6" würfeln, anders ausgedrückt: Dass als Ergebnis eines Wurfes das Ereignis E= „6 liegt oben" eintritt. Diese Wahrscheinlichkeit nennen wir P(E) bzw. P(„6 liegt oben")
- Wir schätzen P(E), indem wir die relative Häufigkeit ermitteln, mit der in einer Folge von n Versuchen (einer Stichprobe aus n Instanzen) das Ereignis E eintritt:

  $P(E) \approx h_n(E)$ (für hinreichend großes n)
- $h_n(E) = H(E)/n$, wobei H(E) die absolute Häufigkeit ist.
- Am Beispiel: Nach 100 Würfen wurde 19 mal die 6 geworfen:

  $P(E) \approx h_n(E) = 0{,}190$

Wahrscheinlichkeiten und relative Häufigkeit

- Beispiel2:

  Die Wahrscheinlichkeit, dass es morgen in Saarbrücken regnet.
  E= „es regnet“
  Instanzen: Das Wetter in SB in einer Stichprobe von n Kalendertagen.
  Von 365 Instanzen fallen 147 unter E:

  $P(E) \approx h_n(E) = 0{,}403$

Wahrscheinlichkeiten und relative Häufigkeit

- Beispiel3:

  Die Wahrscheinlichkeit, dass das nächste gelesene Textwort ein Adjektiv ist
  E= „Adjektiv“
  Instanzen: 1 000 000 Wörter fortlaufender Text aus einem Korpus, davon 65909 Adjektive:

  $P(E) \approx h_n(E) = 0{,}066$

Bedingte Wahrscheinlichkeiten

- Beispiel:
  Die bedingte Wahrscheinlichkeit, dass es morgen in Saarbrücken regnet (E=„es regnet“), gegeben dass Saarbrücken im Einfluss eines Tiefdruckgebietes ist (F=„Tiefdruck“), schreiben wir P(E |F).

- Wir approximieren: $P(E \mid F) \approx \frac{h_n(E \cap F)}{h_n(F)} = \frac{H(E \cap F)}{H(F)}$

- Am Beispiel: 215 Tage mit Tiefdruck, davon an 122 Tagen Regen:
  P(E |F) = 0,567

Wortart-Tagging als Klassifikation

- Wir lernen den statistischen Zusammenhang zwischen Merkmalsmustern (Ereignissen) und Wortarten (ebenfalls Ereignissen) automatisch.
- Schritt1: Datenanalyse
- Wir spezifizieren geeignete Merkmalsmuster:
  - Merkmale, die mit großer Sicherheit automatisch für die Instanzen (Textwörter) ermittelt werden können
  - Informativ in Bezug auf die Aufgabe sind: möglichst eindeutige Entscheidung bezüglich der Klassifikationsaufgabe ermöglichen.
- Vorbereitung: Wir annotieren ein „Trainingskorpus“ mit Wortart-Tags

Einfachste statistische Klassifikation

- Schritt 2: Modell lernen (“Training”)
- Wir lesen den Zusammenhang zwischen Merkmalsmustern (E1, E2, ...) und Wortart-Tags (K1, K2, ...) aus dem Korpus ab.

Ereignis	K1	K2	K3
E1	10	20	0
E2	2	0	15
E3	...	...	...

- Schritt 3: Modell anwenden (Analyse neuer Daten)
  - Jedem Ereignis in den neuen Daten wird die häufigste (≈ auf der Grundlage des bekannten Merkmalsmusters wahrscheinlichste) Klase aus den Trainingsdaten zugewiesen.

Was ist ein Ereignis? (I)

- Im einfachstes Fall: ein Wort

Trainingskorpus

Ereignis	Ein kleines Haus
Klasse	Nadj Adj Nadj

Modell

Ereignis	Nadj	Adj
ein	1	0
kleines	0	1
Haus	1	0

- Keine gute Idee!
  - Klassifikation eines neuen Satzes:
    - Wortliste!

Ereignis	Ein großes Haus
Klasse	Nadj ?? Nadj

Was ist ein Ereignis? (II)

- Besser: Ereignis ist Kombination von informativen Mustern (“Features”)

Ein kleines Haus

Ereignis	Ja	Nein	Nein
	Nein	Ja	Nein
Klasse	Nadj	Adj	Nadj

Trainingskorpus

Selbst Artikel?

Nächstes Wort kap.?

Ereignis	Nadj	Adj
ja, nein	1	0
nein, ja	0	1
nein, nein	1	0

Modell

Klassifikation eines neuen Satzes

Ein großes Haus

Testkorpus

Ereignis	Ja	Nein	Nein	Selbst Artikel?
	Nein	Ja	Nein	Nächstes Wort kap.?

Ereignis	Nadj	Adj
ja, nein	1	0
nein, ja	0	1
nein, nein	1	0

Modell

Klasse	Nadj	Adj	Nadj

Was "bedeutet" das Modell?

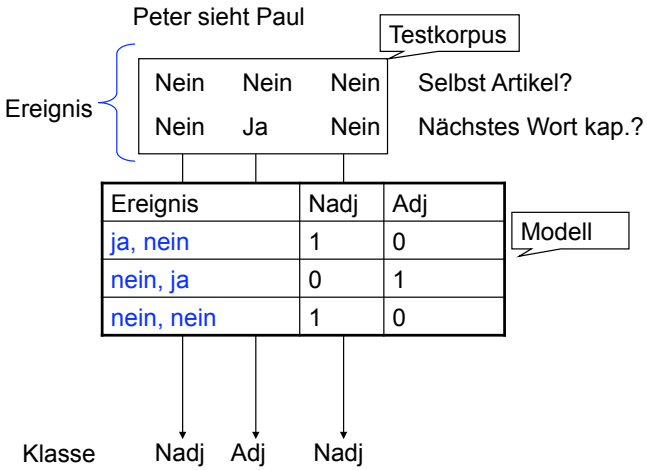
- Features:
  - Selbst Artikel?
  - Nächstes Wort kap.?

Ereignis	Nadj	Adj
ja, nein	1	0
nein, ja	0	1
nein, nein	1	0

- Zeile 1: Artikel, nächstes Wort nicht kapitalisiert: Kein Adjektiv
- Zeile 2: Kein Artikel, nächstes Wort kapitalisiert: Adjektiv
- Zeile 3: Kein Artikel, nächstes Wort nicht kap.: Kein Adjektiv

- Also doch Regeln: nur eben automatisch gelernte...

Fehler



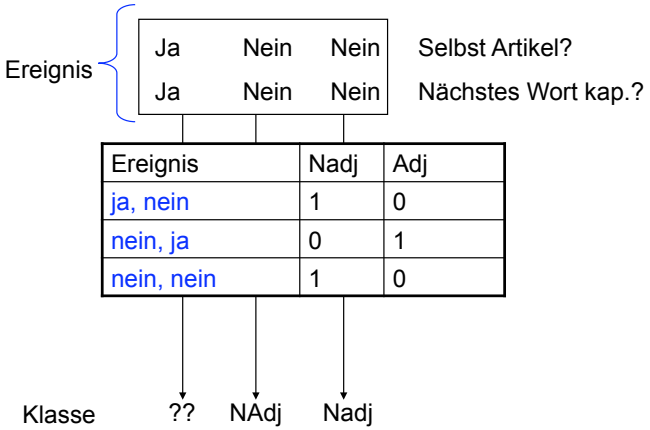
Intelligentere Modelle = mehr Features

- Welche weiteren Muster könnte man ausnutzen, um Adjektive zu identifizieren?
  - Wortendungen (morphologische Information)
    - Wörter /Lemmata auf –ig, -lich, -isch sind mit großer Wahrscheinlichkeit Andjektive
  - Gradpartikel stehen fast immer vor Adjektiven
    - "sehr", "besonders", ...
  - Kombination von existierenden Features
    - Voriges Wort Artikel UND selbst nicht kapitalisiert
  - ...
- Wieso verwendet man nicht einfach alle Features, die einem einfallen?

Größe des Ereignisraumes

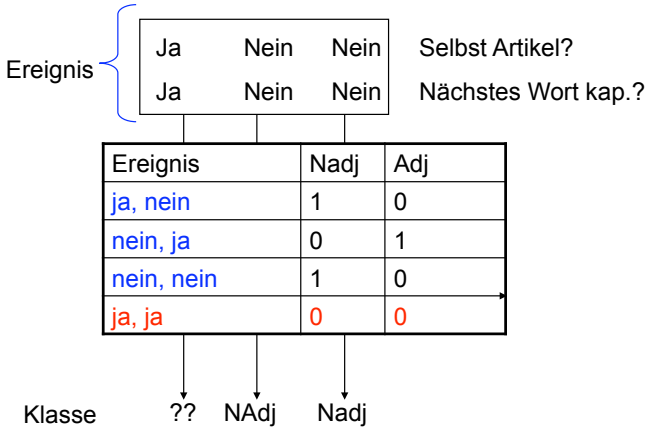
- Wieviele Zeilen hat das Modell (die Tabelle)?
  - Anzahl möglicher verschiedener Ereignisse
    - Produkt der Anzahl möglicher Werte aller Features
      - Beispiel: Selbst Artikel? x Nächstes Wort kapitalisiert = 2 x 2 = 4
      - Lexikalische Features: Wort alleine > 10.000
- Frequenzen in Trainingskorpus werden auf Zeilen (Ereignisse) verteilt
  - Wenn Trainingskorpus nicht deutlich größer ist als die Menge möglicher Ereignisse, können in den Testdaten ungesehene Ereignisse auftreten: Modell kann keine Vorhersage machen
- Das "Sparse Data"-Problem

Sparse Data: Beispiel



Ereignis	Nadj	Adj
ja, nein	1	0
nein, ja	0	1
nein, nein	1	0

Sparse Data: Beispiel



Ereignis	Nadj	Adj
ja, nein	1	0
nein, ja	0	1
nein, nein	1	0
ja, ja	0	0

Das Sparse-Data-Dilemma

- Je mehr Features, desto besser die Datenlage für die Entscheidung
- Je mehr Features, auf desto mehr Ereignisse verteilen sich die Trainingsdaten
- Richtlinien für die Identifikation von Features:
  - Wenige gute Features sind besser als viele mittelmäßige
  - Bevorzuge Features mit wenigen Werten
- Einsatz von „Smoothing"-Techniken, z.B.
  - „Add One"
  - Kombination von Modellen unterschiedlicher Granularität

Smoothing, Beispiel

Ereignis	K1	K2	K3
E1	10	20	0
E2	2	0	15
E3	...	...	...

Ereignis	K1	K2	K3
E1	11	21	1
E2	3	1	16
E3	...	...	...

Evaluation: Die Rolle von Korpora

- Datenanalyse: Auswahl von informativen Merkmalen für die (Klassifikations-) Aufgabe
- Annotation: Manuelle Annotation eines hinreichend großen Korpus mit der relevanten Zielinformation (z.B. Wortart) – Annotation ist subjektiv, deshalb: Unabhängige Annotation durch mehrere Personen, Ermittlung der Übereinstimmung („Inter-Annotator-Agreement“, IAA)
- Training eines statistischen Modells
- Test / Evaluation auf einem Korpus(„Gold-Standard“), das
  - Manuell annotiert
  - Und unabhängig ist, d.h., nicht für Analyse und training verwendet wurde

Daten

Trainingsdaten (Training set)	Testdaten (Test set)

- Meistens zwischen 70 und 90% der Daten
- Grundlage der Modellierung
- Quelle fuer Frequenzen

- Überprüfung des Modells an unab -hängigen Daten

Konfusionsmatrix

	Echtes X	Echtes nicht-X
Als X klassifiziert	✓	✘
Als nicht-X klass.	✘	✓

- Klassenspezifische Evaluation
- Setzt die „tatsächliche“ oder „echte“ Beispielklasse zu den vom Modell zugewiesenen Klassen in Beziehung.

Konfusionsmatrix, Beispiel

	Echtes Adj	Echtes NAdj
Als Adj klassifiziert	✓	Korrektheits-fehler
Als NAdj klassifiziert	Vollständig-keitsfehler	✓

- Vier Fälle:
  - Ist Adjektiv, wird als Adj klassifiziert („true positive“)
  - Ist kein Adjektiv, wird als NAdj klassifiziert („true negative“)
  - Ist kein Adjektiv, wird als Adj klassifiziert („false positive“: Korrektheitsfehler)
  - Ist Adjektiv, wird als NAdj klassifiziert („false negative“: Vollständigkeitsfehler)

Konfusionsmatrix, Beispiel

	Echtes Adj	Echtes NAdj
Als Adj klassifiziert	10	80
Als NAdj klassifiziert	10	900

- Gesamtfrequenzen der wirklichen und der zugewiesenen Klassen sind als Summen von Spalten bzw. Zeilen ablesbar

Akkuratheit und Fehlerrate

- Einfachstmögliche Evaluation ist die Berechnung der Akkuratheit: Die relative Häufigkeit der korrekten Klassifikationen (wahre Positive und wahre Negative) in der Gesamtmenge.

	Echtes Adj	Echtes NAdj
Als Adj klassifiziert	10	80
Als NAdj klassifiziert	10	900

$$\frac{10+900}{1000} = \frac{910}{1000} = 0{,}91$$

- Fehlerrate ist definiert als 1-Akkuratheit (im Beispiel: 0,09)

Probleme

- Problem 1: 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
  - Klassenspezifische Akkuratheit / Fehlerrate
    - 10 / 20 korrekt für Adj: 50%
    - 900 / 980 korrekt für Nadj: 91.8%
- Problem 2: keine Unterscheidung zwischen Korrektheits- und Vollständigkeitsfehlern (falschen Positiven und falschen Negativen)

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)

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

Präzision

	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)

Präzision 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

Präzision und Recall

- Präzision und Recall können i.A. nur zusammen betrachtet werden
  - Hohe Präzision, hoher Recall: gutes Modell
  - Niedrige Präzision, niedriger Recall: schlechtes Modell
  - Hohe Präzision, niedriger Recall: "vorsichtiges" Modell
    - Findet nicht alle Instanzen von X
    - Klassifiziert fast keine Nicht-Xe als X
  - Niedrige Präzision, hoher Recall: "mutiges" Modell
    - Findet fast alle Instanzen von X
    - Klassifiziert auch Nicht-Xe als X

Extremfälle..und die Kombination

- Extremfälle
  - Modell klassifiziert (fast) alles als X
    - Recall gegen 100%, Präzision beliebig niedrig
  - Modell klassifiziert (fast) nichts als X
    - Recall beliebig niedrig, Präzision geht gegen 100% (bei Recall 0% ist Präzision nicht definiert)

- F-Score: Kombination aus P und R: $F = \frac{2PR}{P+R}$
  - Ein Maß für „Gesamtgüte“ der Klassifikation
    - Werte zw. 0 und 1 (höher = besser)

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