Einführung in die Computerlinguistik

Statistische Modellierung

WS 2014/2015
Vera Demberg

Fragebogenaktion Bachelor-StudentInnen

http://www.coli.uni-saarland.de/bsc/page.php?id=fragebogen

Klausuranmeldung nicht vergessen!

Mehrdeutigkeit

Strukturelle Mehrdeutigkeit:

Peter sieht den Mann mit dem Teleskop (Anbindungsambiguität)
Zwei Fremdsprachen spricht jeder Linguist (Skopusambiguität)

Referenzielle Mehrdeutigkeit:

er, sie, es, dort, damals, der Präsident, die Vorlesung
Hans mag seinen Hund, obwohl er ihn manchmal beißt

Lexikalische Mehrdeutigkeit:

Bank, Absatz, Baum

Wortart-Mehrdeutigkeit

laute

finites Verb, Adjektiv

Laute

Gattungssubstantiv (2x), finites Verb, Adjektiv

zu

Adverb, Präposition, Gradpartikel, Verbpartikel

der

Artikel, Demonstrativpronomen, Relativpronomen

Wortart-Disambiguierung

- Einführung von Wortart-Alternativen im Lexikon (alternative lexikalische Ersetzungsregeln bzw. alternative Merkmalsstrukturen).
- Die Grammatik filtert syntaktisch unzulässige Wortartvarianten heraus.
- Wo liegt also das Problem?
- In normalen Texten (z.B. Zeitungstexten) kommen extrem viele "neue" Wörter vor, für die es gar keine Wortartinformation gibt.
- Für viele Sprachen/ Fach- und Sondersprachen gibt es keine Grammatiken; für viele Anwendungen sind große Grammatiken zu langsam. – Es wäre gut, trotzdem Wortartinformation zu haben.

Wortart-Tagging

- Wortartinformation lässt sich glücklicherweise auf der Grundlage „flacher“ linguistischer Information (d.h., ohne syntaktische Analyse) mit großer Sicherheit bereitstellen.
- Wortartinformation wird durch „Wortart-Tagger“ oder „POS-Tagger“ bereitgestellt (POS für „part of speech“, engl. „tag“ ist die Marke/ das Etikett).
- Wortart-Tagger sind heute gut funktionierende Standardwerkzeuge der Sprachverarbeitung, genau wie Morphologie-Systeme. – Sie funktionieren allerdings grundsätzlich anders.

Beispielaufgabe: Adjektiverkennung

- Wortart-Tagger für das Deutsche müssen aus einer von ca. 50 Kategorien wählen, anders ausgedrückt: Sie müssen Textwörter einer von 50 Klassen zuweisen.

- Wir betrachten hier eine einfachere Teilaufgabe: Die Beantwortung der Frage, ob es sich bei einem Vorkommen eines Wortes in einem Text um ein Adjektiv handelt (also eine binäre Klassifikationsaufgabe).

Informative Merkmale

- Woran erkenne ich, dass ein Wortvorkommen ein Adjektiv ist – ohne Lexikon und volle syntaktische Analyse?

  die laute Musik

  das allutivistische Übungsblatt

- Beispiele:
  - Kleinschreibung des aktuellen Wortes w_i
  - Großschreibung des Folgewortes w_{i+1}
  - Vorgängerwort w_{i-1} ist Artikel
  - w_i hat Komparativ-/ Superlativendung
  - w_i hat adjektivspezifisches Derivations-Suffix (-ig, -lich, -isch, -sam)
  - w_{i-1} ist Gradpartikel (sehr, besonders, ziemlich)

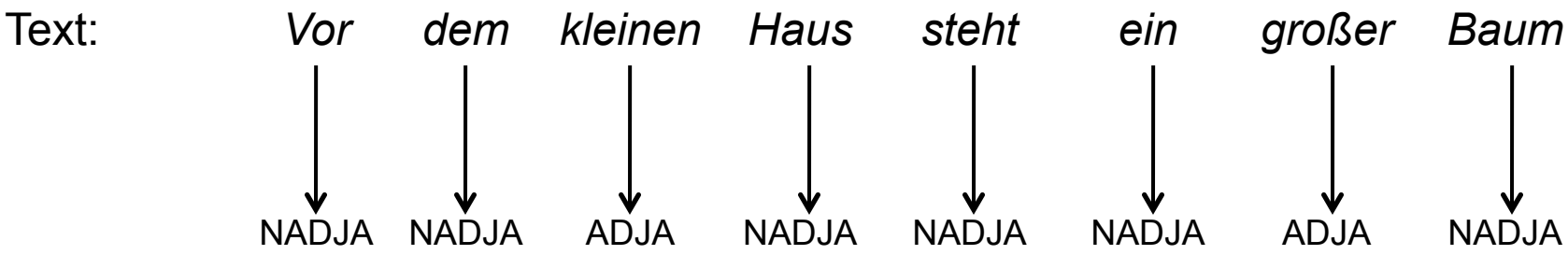
Regelbasierte Wortartzuweisung

- Ein System von wenn-dann-Regeln:

  Wenn `<Merkmal1>, ..., <Merkmaln>` vorliegen, dann weise `<Wortart>` zu.

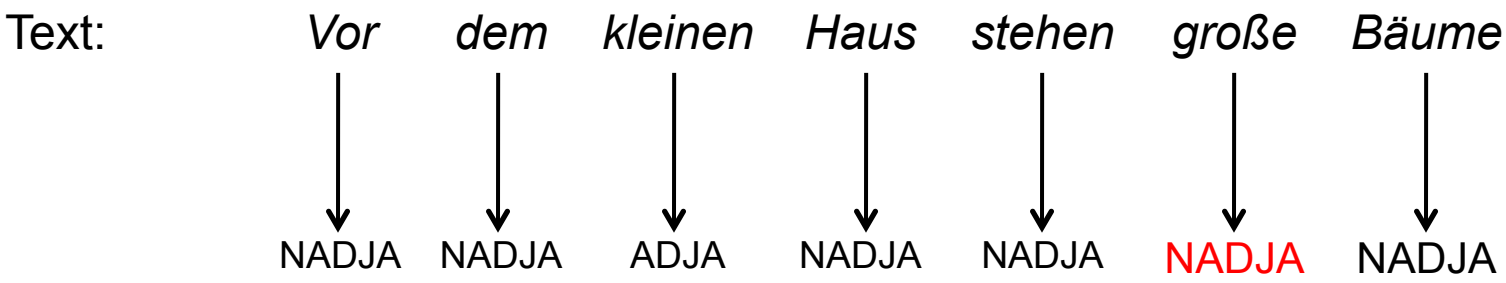
Regelbasiertes Modell

w_i klein & w_{i+1} groß & w_{i-1} Artikel → ADJA



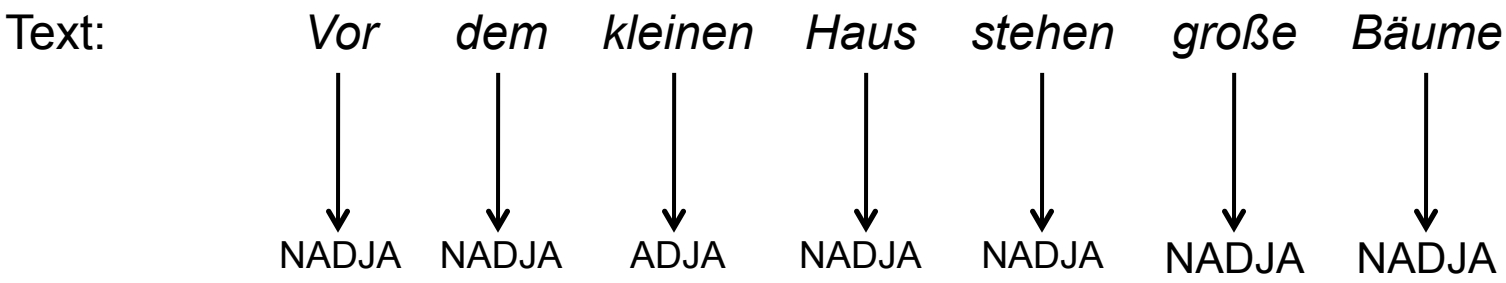
Vollständigkeitsproblem

w_i klein & w_{i+1} groß & w_{i-1} Artikel → ADJ



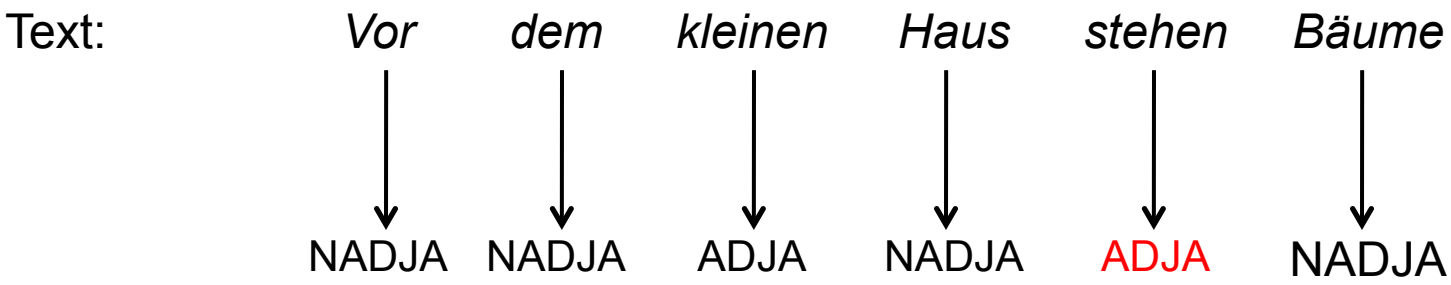
Korrigiertes Modell

w_i klein & w_{i+1} groß → ADJA



Korrektheitsproblem

w_i klein & w_{i+1} groß → ADJA



Regelbasierte und statistische Modellierung

- Regelsysteme, die die Abhängigkeit der Wortart von Merkmalsmustern korrekt und vollständig erfassen sollen, werden schnell sehr komplex und aufwändig zu formulieren.

- Alternative: Wir bauen Systeme, die den Zusammenhang von Merkmalsmustern und Wortarten aus Textkorpora lernen!
  (Singular: das Korpus)

Ein ganz kleiner Korpus-Ausschnitt

Text:

Vor dem kleinen Haus steht ein großer Baum

Ein ganz kleiner Korpus-Ausschnitt

Text:	*Vor*	*dem*	*kleinen*	*Haus*	*steht*	*ein*	*großer*	*Baum*
Manuelle Annotation	NADJA	NADJA	ADJA	NADJA	NADJA	NADJA	ADJA	NADJA

Ein ganz kleiner Korpus-Ausschnitt

Text:	*Vor*	*dem*	*kleinen*	*Haus*	*steht*	*ein*	*großer*	*Baum*
Merkmale:								
w_i groß								
w_{i+1} groß								
w_{i-1} Artikel								
Manuelle Annotation	NADJA	NADJA	ADJA	NADJA	NADJA	NADJA	ADJA	NADJA

Ein ganz kleiner Korpus-Ausschnitt

Text:	*Vor*	*dem*	*kleinen*	*Haus*	*steht*	*ein*	*großer*	*Baum*
Merkmals-extraktion:								
w_i groß	+	-	-	+	-	-	-	+
w_{i+1} groß	-	-	+	-	-	-	+	-
w_{i-1} Artikel	-	-	+	-	-	-	+	-
Manuelle Annotation	NADJA	NADJA	ADJA	NADJA	NADJA	NADJA	ADJA	NADJA

Statistische Modellierung

- Manuelle Korpusannotation:
  - Wir wählen ein Textkorpus und nehmen eine manuelle Annotation mit den Zielklassen (in unserem Fall ∈ {ADJA, NADJA}) vor.
- Merkmalsspezifikation:
  - Wir spezifizieren eine Menge von geeigneten Merkmalen („features“) mit zugehörigen Wertebereichen.
  - In unserem Fall (bisher) 3 Merkmale mit jeweils binärem Wertbereich: {+,–} oder {0,1}: binäre oder Boole'sche Merkmale

  Geeignete Merkmale sind
  - informativ in Bezug auf die Klassifikationsaufgabe
  - einfach zugänglich: direkt ablesbar oder ohne Aufwand automatisch zu ermitteln
- Automatische Merkmalsextraktion:
  - Wir stellen ein Verfahren bereit, das für jede Instanz (hier: für jedes Textwort) automatisch das zugehörige Merkmalsmuster bestimmt.

Statistische Modellierung

- Wir „trainieren“ ein maschinelles Lernsystem auf dem Korpus („Trainingskorpus“).

- Das System „lernt“ ein statistisches Modell, das neuen, nicht annotierten Instanzen (auf der Grundlage des Merkmalsmusters) die wahrscheinlichste Klasse zuweisen kann.

- Das einfachste Verfahren für das Erlernen eines Klassifikations-modells besteht im Auszählen der Häufigkeit, mit der Klassen im Zusammenhang mit bestimmten Merkmalsmustern auftreten.

Beispiel: Adjektive im Wahrig-Korpus

- Frequenzen in einem kleinen Teilkorpus:

n groß	-	-	-	-	+	+	+	+
n+1 groß	-	-	+	+	-	-	+	+
n-1 Art.	-	+	-	+	-	+	-	+
ADJA	31	12	140	84	1	1	8	2
NADJA	1827	58	738	18	730	249	98	3

- Relative Frequenz als geschätzte Wahrscheinlichkeit:
  Ein einfaches statistisches Modell

n groß	-	-	-	-	+	+	+	+
n+1 groß	-	-	+	+	-	-	+	+
n-1 Art.	-	+	-	+	-	+	-	+
ADJA	0,017	0,171	0,159	0,824	0,001	0,004	0,075	0,400
NADJA	0,983	0,829	0,841	0,176	0,999	0,996	0,925	0,600

Wahrscheinlichkeit und Frequenz

- Wir nehmen die relative Häufigkeit, mit der eine Klasse *k* im Kontext eines Merkmalsmusters *e* auftritt, als Schätzung der bedingten Wahrscheinlichkeit, dass *k* vorliegt, gegeben *e*.

- Beispiel: ADJA kommt mit dem Merkmalsmuster <-,+,+>, also "n klein, n+1 groß, n-1 Artikel" 738mal (von insgesamt 878) vor; die relative Frequenz ist ≈ 0,824, wir nehmen also die Wahrscheinlichkeit, dass in dieser Konstellation ein Adjektiv vorliegt, ebenfalls mit 0,824, also 82,4% an.

Etwas Terminologie zur Wahrscheinlichkeitstheorie

- Beobachtung:
  - Einzelvorkommen oder Instanz
  - Beispiel: ein Wurf mit zwei Würfeln, ein Textwort
- Ereignis:
  - Klasse von Beobachtungen mit gleichen Merkmalen
  - Beispiele: "eine 7 würfeln", "ein groß geschriebenes Wort"
  - Die unterschiedlichen Merkmalsmuster spezifizieren „Ereignisse“ im Sinne der Wahrscheinlichkeitstheorie. Die Merkmale in unserem Beispiel spannen den „Ereignisraum“ auf (hier mit 2*2*2=8 Elementen).
- Wahrscheinlichkeit eines Ereignisses: $P(e) \in [0,1]$
- Gemeinsame Wahrscheinlichkeit, Wahrscheinlichkeit, dass zwei Ereignisse gleichzeitig vorliegen: $P(e, e')$
- Bedingte Wahrscheinlichkeit *(e gegeben e')*:

$$P(e \mid e') = \frac{P(e,e')}{P(e')}$$

Wahrscheinlichkeit und Frequenz

- Wir sind an der Wahrscheinlichkeit einer Klasse k, gegeben ein Merkmalsmuster f, interessiert:

$$P(k \mid f) = \frac{P(k,f)}{P(f)}$$

- Wir schätzen die Wahrscheinlichkeiten über Korpusfrequenzen:

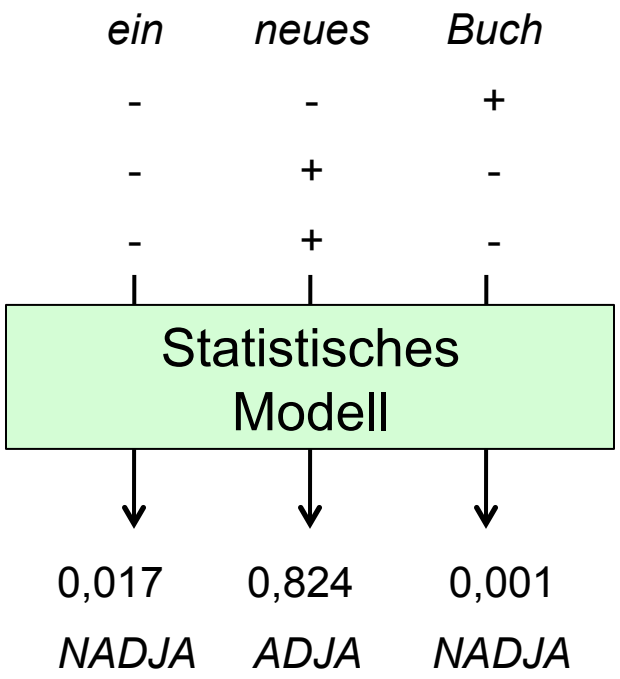
$$P(k \mid f) = \frac{P(k,f)}{P(f)} \approx \frac{Fr(k,f)}{Fr(f)}$$

- Beispiel:

$$P(ADJA \mid \langle -,+,+ \rangle) \approx \frac{Fr(ADJA, \langle -,+,+ \rangle)}{Fr(\langle -,+,+ \rangle)} = \frac{84}{102} = 0{,}824$$

Anwendung des statistischen Modells

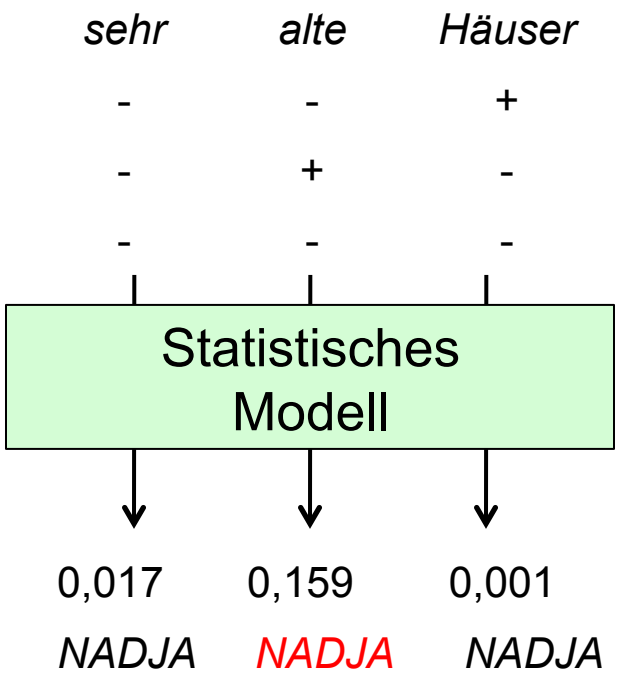
- Neuer Text: Merkmalsextraktion; auf der Grundlage der Merkmale Bestimmung (eigentlich "Ablesen") der geschätzten Wahrscheinlichkeit.
- Da wir an der Zuweisung der im Kontext angemessenen Wortart interessiert sind, verwenden wir das Modell als Klassifikator: Es weist die jeweils aufgrund des Merkmals-musters wahrscheinlichste Klasse zu.



- Wir können die Wahrscheinlichkeitsinformation zusätzlich verwenden, z.B. als "Konfidenz" (Klassifikation wird nur bei einer Wahrscheinlichkeit ≥ 0,8 zugewiesen) oder zur Parsersteuerung (Bottom-Up-Parser probiert die Wortart-Alternativen in der Reihenfolge ihrer Wahrscheinlichkeit aus).

Klassifikationsfehler

- Auch statistische Modelle machen Korrektheits- und Vollständigkeits-fehler.
- Man kann die Modelle verbessern, indem man die Merkmalsinformation verfeinert, beispielsweise durch Einführung eines Merkmals "Vorgängerwort ist Gradpartikel".
- Das Verfahren stößt allerdings an Grenzen.



Größe des Merkmalsraums

- Wieso verwendet man nicht alle Merkmale, die irgendwie erfolgversprechend sind?
- Ereignisraum:
  - Wir haben im Beispiel 3 binäre Merkmale verwendet, es gibt also 2*2*2=8 Ereignisse.
  - Wenn wir 10 binäre Merkmal verwenden, haben wir bereits über 1000 Ereignisse.

- Die Instanzen im Trainingskorpus verteilen sich auf die einzelnen Ereignisse (Merkmalsmuster).
  - Das Trainingskorpus muss deutlich größer sein als der Ereignisraum. Ansonsten treten viele Merkmalsmuster nur wenige Male auf, oder auch gar nicht („ungesehene Ereignisse“): Das Modell kann im ersten Fall nur sehr unzuverlässige Schätzungen machen, im letzteren Fall gar keine.
  - Dies ist das sogenannte „Sparse-Data“-Problem.

Sparse-Data-Problem

- Je mehr Merkmale, umso besser ist grundsätzlich die Datenlage für die Entscheidung, aber:

- Je mehr Merkmale, auf desto mehr Ereignisse verteilen sich die Trainingsdaten. Die Wahrscheinlichkeitsschätzung wird ungenau oder sogar unmöglich.

- Faustregel für die Wahl einer geeigneten Merkmalsmenge:
  - Wenige gute (aussagekräftige) Merkmale sind besser als viele mittelmäßige.
  - Merkmale mit weniger möglichen Werten sind grundsätzlich vorzuziehen.

Evaluation

- Jedes Modell muss evaluiert werden: Stimmt es mit der Realität, die es beschreiben soll, mit der Funktion, die es ausführen soll, überein?

- Dies gilt für wissensbasierte und statistische Modelle grundsätzlich in gleicher Weise.

- Da statistische Verfahren typischerweise auf Probleme angewandt werden, die keine vollständige Korrektheit erreichen können (z.B. Disambiguierung in allen Spielarten), ist es hier besonders wichtig.

Evaluation

- Annotation eines „Goldstandard“: Testkorpus mit der relevanten Zielinformation (z.B. Wortart)
  - Um subjektive Varianz auszuschließen, wird durch mehrere Personen unabhängig annotiert und die Übereinstimmung („Inter-Annotator-Agreement“: IAA) gemessen.
  - Testkorpus und Trainingskorpus müssen disjunkt sein, um Effekte aus individuellen Besonderheiten eines Korpus auszuschließen („overfitting“).
- Automatische Annotation des Testkorpus mit statistischem Modell/ Klassifikator
- Messung der Performanz durch Vergleich von automatischer Annotation mit Goldstandard

Akkuratheit

- Akkuratheit (engl. *accuracy*) ist das einfachste Maß:

> Akkuratheit = korrekt klassifizierte Instanzen/alle Instanzen

- Fehlerrate (engl. *error rate*) ist der Komplementärbegriff zu Akkuratheit:

> Fehlerrate = 1 – Akkuratheit

- Das Akkuratheitsmaß verdeckt oft tatsächlich relevante Eigenschaften eines Modells.

Konfusionsmatrix

- Grundlage für eine feinere Evaluation des Klassifikators ist die Konfusionsmatrix.
- Konfusionsmatrix (Verwechslungstabelle) für binäre Klassifikation:

	Echtes ADJA	Echtes NADJA
Klassifiziert als ADJA	**ok**	**falsch**
Klassifiziert als NADJA	**falsch**	**ok**

Konfusionsmatrix

- Fehlertypen für ADJA-Klassifikation:

	Echtes ADJA	Echtes NADJA
Klassifiziert als ADJA	**ok**	Korrektheits-fehler
Klassifiziert als NADJA	Vollständigkeits-fehler	**ok**

Konfusionsmatrix

- Fehlertypen für ADJA-Klassifikation:

	Echtes ADJA	Echtes NADJA
Klassifiziert als ADJA	true positive	false positive
Klassifiziert als NADJA	false negative	true negative

Konfusionsmatrix

- (Fiktives) Beispiel:

	Echtes ADJA	Echtes NADJA
Klassifiziert als ADJA	20	80
Klassifiziert als NADJA	20	880

- Von insgesamt 1000 Fällen sind 900 korrekt (Wahre Positive und wahre Negative): Akkuratheit ist also 90%, Fehlerrate 10%.
- Tatsächlich ist die Adjektiverkennung miserabel: von fünf als ADJA klassifizierten Instanzen ist nur eine korrekt.
- Wir bestimmen Recall und Präzision/ Precision als klassenspezifische Maße, die Vollständigkeits- und Korrektheitsfehler (für eine gegebene Klasse) separat messen.

Recall

	Echtes ADJA	Echtes NADJA
Klassifiziert als ADJA	True positive	False positive
Klassifiziert als NADJA	False negative	True negative

- Welcher Anteil der echten X wurde tatsächlich gefunden (als X klassifiziert)?

Recall = True positives/(True positives + False negatives)

	Echtes ADJA	Echtes NADJA
Klassifiziert als ADJA	20	80
Klassifiziert als NADJA	20	880

Recall für ADJA = 20/(20+20) = 0,5

Präzision

	Echtes ADJA	Echtes NADJA
Klassifiziert als ADJA	True positive	False positive
Klassifiziert als NADJA	False negative	True negative

- Welcher Anteil der als X klassifizierten Instanzen ist tatsächlich ein X?

Precision = True positives/(True positives + False positives)

	Echtes ADJA	Echtes NADJA
Klassifiziert als ADJA	20	80
Klassifiziert als NADJA	20	880

Precision für ADJA = 20/(20+80) = 0,2

Präzision und Recall

- Präzision und Recall sind im allgemeinen nur zusammen aussagekräftig
  - 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 kaum Nicht-Xe als X
  - Niedrige Präzision, hoher Recall: „Mutiges“ Modell
    - Findet fast alle Instanzen von X
    - Klassifiziert viele nicht-Xe fehlerhaft als X
  - Extremfälle
    - Modell klassifiziert alles als X: Recall 100%, Precision niedrig
    - Modell klassifiziert nichts als X: Recall 0%, Precision nicht definiert

F-Score

- Der „F-Score“ ist ein Maß für die „Gesamtgüte“ der Klassifikation, in das Precision und Recall eingehen.

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

- F-Score für die Klasse ADJA im Beispiel:

$$F = \frac{2 * 0{,}2 * 0{,}5}{0{,}2 + 0{,}5} = 0{,}29$$

Noch einmal: Wortart-Tagging

- Standard Wortart-Tagger arbeiten mit ca. 50 Klassen und haben dabei eine Akkuratheit von deutlich über 99%.
- Sie gehen dabei natürlich etwas anders vor, als hier demonstriert: Sie verwenden maschinelle Lernverfahren, die nicht nur die besten POS-Tags für die einzelnen Wörter im Satz, sondern die beste POS-Kette für einen ganzen Satz zu bestimmen versuchen.
- Beispiel: Auch wenn in „*I made her duck*“ die wahrscheinlichste Wortart für *her* Personalpronomen und für *duck* Gattungssubstantiv ist, ist die Kombination der Wortarten sehr unwahrscheinlich.
- Die Methode, Wahrscheinlichkeiten für Sequenzen zu bestimmen, ist auch in der Verarbeitung gesprochener Sprache wichtig („HMMs: „Hidden Markov Models“)