Einführung in die Computerlinguistik

Statistische Modellierung

WS 2011/2012

Manfred Pinkal

Wortartinformation

- Wortartinformation ist eine wichtige Voraussetzung für die syntaktische Analyse. Woher kommt sie?
- Erste Option: Wortartinformation durch das Lexikon
- Problem 1: Mehrdeutigkeit der Wortart
  - *die laute Musik*
    V, ADJ
  - *Laute Musik*
    V, ADJ, (2x) N
  - Die Partikel *zu*:
    Adverb, Präposition, Gradpartikel, Infinitivpartikel
- Wir benötigen die im Kontext angemessene Wortart: Disambiguierung!

Wortart-Tagger

- Problem 2: Das Lexikon natürlicher Sprachen ist nie vollständig. Jeder normale Zeitungstext enthält neue Wörter, für die keine Wortartinformation vorliegen kann (beispielsweise jede Menge Eigennamen)
- 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), als Vorverarbeitungsschritt für die syntaktische Analyse.
- Wortart-Tagger sind heute Standardwerkzeuge der Sprachverarbeitung – wie Morphologie-Systeme.

Beispielaufgabe: Adjektiverkennung

- Wortart-Tagger für das Deutsche müssen aus einer von ca. 50 Kategorien wählen.
- Wir betrachten hier eine 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

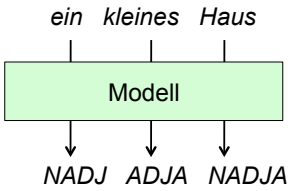
- Woran erkenne ich, dass ein Wortvorkommen ein Adjektiv ist – ohne Lexikon und volle syntaktische Analyse?

  die laute Musik

  das adjudikative 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)

Aufgabe ist Modellbildung

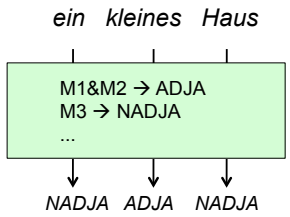
- Wie verwenden wir Merkmale eines Wortvorkommens in einem Modell der Wortartklassifikation?



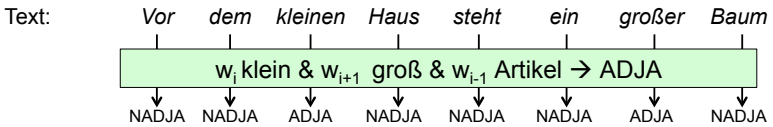
Regelbasiertes Modell:

- Ein System von wenn-dann-Regeln:

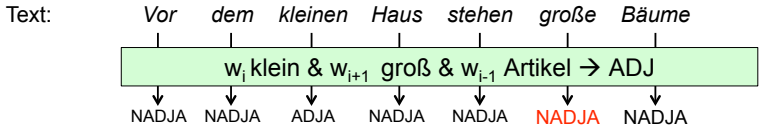
  Wenn `<Merkmal1>, ..., <Merkmaln>` vorliegen, dann weise `<Wortart>` zu.



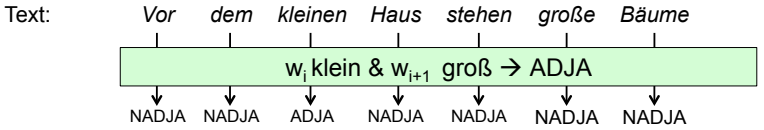
Regelbasiertes Modell



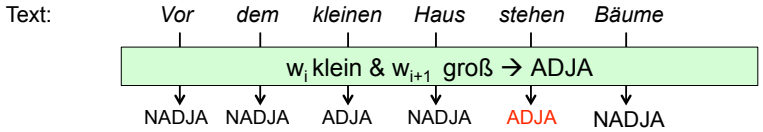
Vollständigkeitsproblem



Korrigiertes Modell



Korrektheitsproblem



Regelbasierte und statistische Modellierung

- Es ist schwer, Regeln zu schreiben, die die Abhängigkeit der Wortart von Merkmalsmustern korrekt und vollständig erfassen.
- Alternative: Wir lernen den Zusammenhang von Merkmalsmustern und Wortarten aus Textkorpora!
  - Ein Korpus (Neutrum!) ist eine endliche Sammlung von konkreten sprachlichen Äußerungen, die als Grundlage für sprachwissenschaftliche Untersuchungen dienen (Bussmann, Lexikon der Sprachwissemschaft)

Statistische Modellierung

- 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.
- Manuelle Korpusannotation:
  - Wir wählen ein Textkorpus, extrahieren die Merkmale und nehmen eine manuelle Annotation mit den Zielklassen (in unserem Fall ∈ {ADJA, NADJA}) vor.

Ein 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

- 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 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 geschätzte Wahrscheinlichkeit: die bedingte 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 Andjektiv vorliegt, ebenfalls mit 0,824, also 82,4% an.

Etwas Terminologie zur Wahrscheinlichkeitstheorie

- Beobachtungen:
  - Einzelvorkommen oder Instanzen
  - Beispiele: ein Wurf mit zwei Würfeln, ein Textwort
- Ereignisse:
  - Klassen von Beobachtungen mit gleichen Merkmalen
  - Beispiele: "eine 7 würfeln", "ein groß geschriebenes Wort"
  - Die unterschiedlichen Merkmalsmuster spezifizieren „Ereignisse" im Sinne der Wahrscheinlichkeitstheorie, deshalb die Bezeichnung „e". Die Merkmale in unserem Beispiel spannen den „Ereignisraum" auf (hier mit 2*2*2=8 Elementen).
- Wahrscheinlichkeit eines Ereignisses: P(e) ∈ [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:

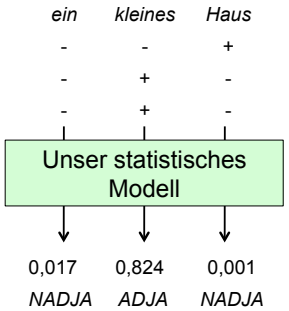
$$P(k \mid f) = \frac{P(k,f)}{P(f)} \approx \frac{Fr(k,f)}{Fr(f)}$$

- Beispiel:

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

Anwendung des statistischen Modells

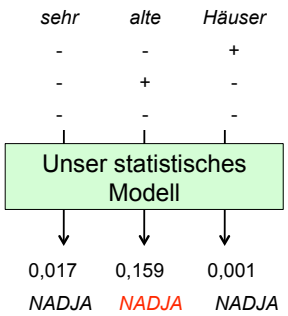
- Neuer Text: Merkmalsextraktion; Eingabe der Merkmale in das statistische Modell; Bestimmung (eigentlich "Ablesen") der geschätzten Wahrscheinlichkeit.
- Da wir an der Zuweisung der im Kontext angemessenen Wortart interessiert sind, verwenden wir das Modell als Klassifikator: Es weist die jeweils aufgrund 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 Vollständigkeitsfehler.
- Man kann die Modelle verbessern, indem man die Merkmalsinformation verfeinert, besipeisweise 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 Merkmalsmuster.
  - Das Trainingskorpus muss deutlich größer sein als der Ereignisraum. Ansonsten treten viele Merkmalsmuster gar nicht auf („ungesehene Ereignisse"): Das Modell kann dafür keine Vorhersage machen.
  - 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, 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

- (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 Precision als klassenspezifische Maße, die Vollständigkeits- und Korrektheitsfehler separat messen.

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

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

Precision

	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

Precision und Recall

- Precision 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 keine 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 klassifziert nichts als X: Recall 0%, Precision nicht definiert