Einführung in die Computerlinguistik

Statistische Modellierung I

WS 2010/2011

Manfred Pinkal

Wortartinformation

- Wortartinformation ist eine wichtige Voraussetzung für die syntaktische Analyse. Woher kommt sie?
- Erste Option: Wortartinformation durch das Lexikon
- Ein kleineres Problem: Mehrdeutigkeit der Wortart
  - *die laute Musik*
    V, ADJ
  - *Laute Musik*
    V, ADJ, (2x) N
  - Die Partikel *zu*:
    Adverb, Präposition, Gradpartikel, Infinitivpartikel
- Wortartalternativen werden durch das Lexikon eingeführt, durch die Syntaxregeln disambiguiert (Earley-Algorithmus, Scanner!)

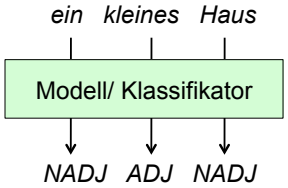
Wortart-Tagger

Ein hartes Problem: Fehlende Abdeckung des Lexikons

- Das Lexikon natürlicher Sprachen ist nie vollständig. Jeder normale Zeitungstext enthält neue Wörter, für die keine Wortartinformation vorliegen kann.
- Wortartinformation lässt sich glücklicherweise auf der Grundlage „flacher" linguistischer Information (d.h., ohne volle 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 eine kpmplexe Klasifikationsaufgabe leisten: Zuweisung einer von ca. 50 Kategorien wählen.
- Wir betrachten 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?

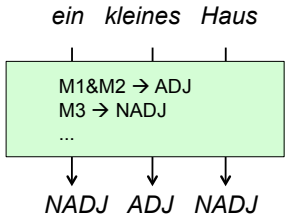
  die laute Musik

  das siebente 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 Adjektiv-Suffix (-ig, -lich, -isch, -sam)
  - w_{i-1} ist Gradpartikel (sehr, besonders, ziemlich)
- Geeignete Merkmale sollten direkt abgelesen oder ohne Aufwand automatisch ermittelt werden können; sie müssen informativ in Bezug auf die Klassifikationsaufgabe sein.

Regelbasiertes Modell:

- 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 → ADJ

Text:	*Vor*	*dem*	*kleinen*	*Haus*	*steht*	*ein*	*großer*	*Baum*
Merkmale:								
w_i groß	+	-	-	+	-	-	-	+
w_{i+1} groß	-	-	+	-	-	-	+	-
w_{i-1} Artikel	-	-	+	-	-	-	+	-
POS-Tags:	NADJ	NADJ	ADJ	NADJ	NADJ	NADJ	ADJ	NADJ

Vollständigkeitsproblem!

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

Text:	*Vor*	*dem*	*kleinen*	*Haus*	*stehen*	*große*	*Bäume*
Merkmale:							
w_i groß	+	-	-	+	-	-	+
w_{i+1} groß	-	-	+	-	-	+	-
w_{i-1} Artikel	-	-	+	-	-	-	-
POS-Tags:	NADJ	NADJ	ADJ	NADJ	NADJ	NADJ	NADJ

Korrigiertes Modell

w_i klein & w_{i+1} groß → ADJ

Text:	*Vor*	*dem*	*kleinen*	*Haus*	*stehen*	*große*	*Bäume*
Merkmale:							
w_i groß	+	-	-	+	-	-	+
w_{i+1} groß	-	-	+	-	-	+	-
w_{i-1} Artikel	-	-	+	-	-	-	-
POS-Tags:	NADJ	NADJ	ADJ	NADJ	NADJ	ADJ	NADJ

Korrektheitsproblem!

w_i klein & w_{i+1} groß → ADJ

Text:	*Vor*	*dem*	*kleinen*	*Haus*	*stehen*	*Bäume*
Merkmale:						
w_i groß	+	-	-	+	-	+
w_{i+1} groß	-	-	+	-	+	-
w_{i-1} Artikel	-	-	+	-	-	-
POS-Tags:	NADJ	NADJ	ADJ	NADJ	ADJ	NADJ

Regelbasierte vs. statistische Modellierung

- Es ist schwer, Regeln zu schreiben, die die Abhängigkeit der Wortart von Merkmalsmustern korrekt und vollständig erfassen.
- Alternative: Lernen des Zusammenhangs von Merkmalsmustern und Wortarten aus Korpora!
- Wir spezifizieren eine Menge von geeigneten Merkmalen („features")-
- Wir wählen ein Textkorpus aus („Trainingskorpus") und annotieren die Daten manuell mit Wortarttags.
- Wir extrahieren für jede Instanz (Textwort) das zugehörige Merkmalsmuster („feature extraction").

Trainingskorpus

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

Statistisches Modell

- Wir wählen einen Klassifikator (die einfachste Form eines maschinellen Lernsystems), und „trainieren“ ihn auf dem Trainingskorpus.
- Der Klassifikator „lernt“ den statistischen Zusammenhang zwischen Merkmalsmustern und Klassen.
- Nach dem Training weist er jeder neuen Instanz die Klasse zu, die er (auf der Grundlage des Merkmalsmusters) für die wahrscheinlichste hält (und ggf. die zugehörige Wahrscheinlichkeit als „Konfidenzwert“).

Wahrscheinlichkeit und Frequenz

- Das einfachste Verfahren für die Wahrscheinlichkeitsschätzung:
  - Wir zählen die Häufigkeit, mit der Merkmalsmuster und Klassen gemeinsam auftreten.
  - 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*.
- Die unterschiedlichen Merkmalsmuster betrachten wir als „Ereignisse“ im Sinne der Wahrscheinlichkeitstheorie. Die Merkmale mit ihren alternativ möglichen Werten spannen den „Ereignisraum“ auf.
- Der Klassifikator weist einer Instanz das wahrscheinlichste Wortart-Tag zu. „Training“ des Klassifikators besteht in diesem einfachsten Fall im Grunde nur im Auszählen des Korpus.

Wahrscheinlichkeit und Frequenz

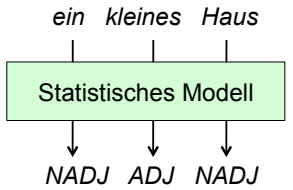
Frequenz im Korpus

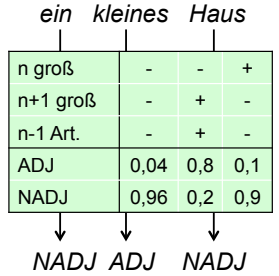
n groß	-	-	+
n+1 groß	-	+	-
n-1 Art.	-	+	-
ADJ	5	40	12
NADJ	120	10	108

Relative Frequenz/
Geschätzte Wahrscheinlichkeit

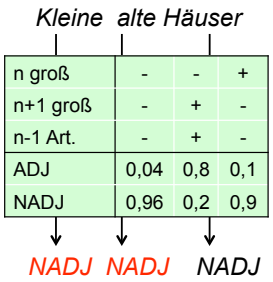
n groß	-	-	+
n+1 groß	-	+	-
n-1 Art.	-	+	-
ADJ	0,04	0,8	0,1
NADJ	0,96	0,2	0,9

Einfaches statistisches Modell





Klassifikationsfehler



Größe des Merkmalsraums

- Wieso verwendet man nicht alle Merkmale, die irgendwie erfolgversprechend sind?
- Produkt der Anzahl möglicher Werte aller Merkmale:
  - Wir haben im Beispiel 3 binäre Merkmale verwendet, es gibt also 2*2*2=8 Muster.
  - Wenn wir 10 binäre Merkmal verwenden, haben wir über 1000 Muster.
  - Wenn wir z.B. noch als zwei Merkmale das Vorgängerwort und das Nachfolgerwort selbst hinzunehmen, kommen wir auf Milliarden von Kombinationen.
- 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.
  - „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.
- Techniken, um dem Problem fehlender Daten vorzubeugen sind als „Smoothing", zu deutsch: Glättung bekannt.
  - Einfachstes Verfahren: „Add one". Man zählt die absolute Frequenz jeweils „um eins" hoch.
  - „Back-Off-Strategie": Wenn bei einer feinkörnigen Merkmalsauswahl ungesehene Ereignisse Vorkommen, approximiert man mit einer gröberen Merkmalsauswahl.

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") gemessen.
  - Testkorpus und Trainingskorpus müssen disjunkt sein, um Effekte aus den Besonderheiten der Korpustexte auszuschließen („overfitting")
- Automatische Annotation des Testkorpus mit statistischem Modell/ Klassifikator
- Messung der Performanz durch Vergleich von automatischer Annotation mit Goldstandard

Akkuratheit

- Akkuratheit (accuracy) ist das einfachste Maß:

  Akkuratheit = korrekt klassifizierte Instanzen/alle Instanzen

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

  Fehlerrate = 1 – Akkuratheit

- Das Akkuratheitsmaß verdeckt oft die tatsächliche Performanz eines Verfahrens.
- Grundlage für eine feinere Evaluation des Klassifikators ist die Konfusionsmatrix.

Konfusionsmatrix

- Konfusionsmatrix (Verwechslungstabelle) für binäre Klassifikation:

	Echtes ADJ	Echtes NADJ
Klassifiziert als ADJ	**ok**	**falsch**
Klassifiziert als NADJ	**falsch**	**ok**

Konfusionsmatrix

- Konfusionsmatrix (Verwechslungstabelle) für binäre Klassifikation:

	Echtes ADJ	Echtes NADJ
Klassifiziert als ADJ	**ok**	Korrektheits-fehler
Klassifiziert als NADJ	Vollständigkeits-fehler	**ok**

Konfusionsmatrix

- Konfusionsmatrix (Verwechslungstabelle) für binäre Klassifikation:

	Echtes ADJ	Echtes NADJ
Klassifiziert als ADJ	true positive	false positive
Klassifiziert als NADJ	false negative	true negative

Konfusionsmatrix

- (Fiktives) Beispiel:

	Echtes ADJ	Echtes NADJ
Klassifiziert als ADJ	20	80
Klassifiziert als NADJ	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 ADJ klassifizierten Instanzen ist nur eine korrekt.
- Recall und Precision als klassenspezifische Maße, die Vollständigkeits- und Korrektheitsfehler separat messen.

Recall

	Echtes ADJ	Echtes NADJ
Klassifiziert als ADJ	True positive	False positive
Klassifiziert als NADJ	False negative	True negative

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

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

	Echtes ADJ	Echtes NADJ
Klassifiziert als ADJ	20	80
Klassifiziert als NADJ	20	880

Recall für ADJ = 20/(20+20) = 0,5

Precision

	Echtes ADJ	Echtes NADJ
Klassifiziert als ADJ	True positive	False positive
Klassifiziert als NADJ	False negative	True negative

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

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

	Echtes ADJ	Echtes NADJ
Klassifiziert als ADJ	20	80
Klassifiziert als NADJ	20	880

Precision für ADJ = 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

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 ADJ im Beispiel:

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

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 klassifzert nichts als X: Recall 0%, Precision nicht definiert

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, beste Wahrscheinlichkeiten für Sequenzen zu bestimmen, ist auch in der Verarbeitung gesprochener Sprache wichtig („HMMs: „Hidden Markov Models“)