Einführung in die Computerlinguistik

Statistische Verfahren in der lexikalischen Semantik

WS 2011/2012
Manfred Pinkal

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

Evaluation

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

- Akkuratheit: Welcher Anteil der Instanzen wurde korrekt klassifiziert?
  - Akkuratheit der ADJ/NADJ-Klassifikation = 900/1000 = 0,9
- Recall: Welcher Anteil der echten X wurde tatsächlich gefunden (als X klassifiziert)?
  - Recall für ADJA = 20/(20+20) = 0,5
- Precision: Welcher Anteil der als X klassifizierten Instanzen ist tatsächlich ein X?
  - 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

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

Lexikalische Mehrdeutigkeit

- Äußerungs- und Textverstehen impliziert die Erkennung der korrekten, im Kontext intendierten Äußerungsbedeutung.
- Wörter sind vielfach mehrdeutig:
  - *Bank: Geldinstitut / Sitzmöbel*
  - *Maschine: Flugzeug / Motorrad/ Technisches Gerät*
  - *Absatz: Schuh/ Treppe/ Text/ Verkauf*
  - *aufgeben: einen Plan / einen Koffer aufgeben*
- Die Disambiguierung der Wortbedeutung (engl. "Word-sense disambiguation": WSD) ist eine zentrale Aufgabe der Computerlinguistik.

WSD

Ich will heute einkaufen, und ich muss vorher zur Bank fahren.

- Wissensbasierte Disambiguierung durch Inferenz mit Weltwissen:
  - Praktisch nicht machbar: Riesige Mengen an handkodiertem Weltwissen wären nötig.
- Alternative: Statistische Modellierung
  - Identifikation von wortspezifischen Merkmalen ("in Objektposition des Verbs *fahren*", "*einkaufen* als Kontextwort") nicht sinnvoll, da es tausende von mehrdeutigen Wörtern gibt, die keinem gemeinsamen Muster folgen.

Statistische Modellierung

- Merkmalsspezifikation
- Automatische Merkmalsextraktion
- Manuelle Korpusannotation
- Training eines statistischen Modells

Statistische Modellierung: WSD

- Korpusannotation:
  - Spezifikation des "Annotationsschemas": Übernahme von Wortbedeutungen aus einem Wörterbuch oder Thesaurus (Standard: WordNet)
  - Annotation aller Instanzen mit der Wortbedeutung

Trainings-Korpus

...

(A237) ... Für diejenigen, denen Komfort wichtig ist, haben wir eine Bank mit leicht schwingender Rückenlehne entwickelt. ...

(A295) ... Ich suche noch eine Bank für meinen Garten und sondiere deshalb gerade Angebote. ...

(A303) ... Habe im März 2000 einen höheren Betrag bei einer Bank angelegt. ...

(A452) ... Beim Test Anlageberatung der Banken löste kein Institut die einfache Frage nach einer sicheren Anlage wirklich gut. ...

...

Trainings-Korpus: Annotation mit WSD-Information

...

(A237) ... Für diejenigen, denen Komfort wichtig ist, haben wir eine Bank <bank1> mit leicht schwingender Rückenlehne entwickelt. ...

(A295) ... Ich suche noch eine Bank <bank1> für meinen Garten und sondiere deshalb gerade Angebote. ...

(A303) ... Habe im März 2000 einen höheren Betrag bei einer Bank<bank2> angelegt. ...

(A452) ... Beim Test Anlageberatung der Banken <bank2> löste kein Institut die einfache Frage nach einer sicheren Anlage wirklich gut. ...

...

Inhaltswörter

...

(A237) ... Für diejenigen, denen Komfort wichtig ist, haben wir eine Bank <bank1> mit leicht schwingender Rückenlehne entwickelt. ...

(A295) ... Ich suche noch eine Bank <bank1> für meinen Garten und sondiere deshalb gerade Angebote. ...

(A303) ... Habe im März 2000 einen höheren Betrag bei einer Bank<bank2> angelegt. ...

(A452) ... Beim Test Anlageberatung der Banken <bank2> löste kein Institut die einfache Frage nach einer sicheren Anlage wirklich gut. ...

...

Merkmalsspezifikation

- Merkmalsspezifikation:
  - Wir betrachten als Merkmalsinformation die Wörter, die im Kontext einer Instanz vorkommen. Präziser ausgedrückt:
  - Wir nehmen für alle zu disambiguierenden Wörter eine gemeinsame Merkmalsmenge an, nämlich die n (z.B. n=1000) häufigsten Inhaltswörter (Substantive, Verben, Adjektive).
  - Hochdimensionaler Merkmalsraum, alle Merkmale sind Boole'sche Merkmale. Für das spezifische Merkmalsmuster (den Merkmalsvektor) v einer Instanz setzen wir $v_i = 1$, wenn das Wort w_i als Kontextwort auftritt, ansonsten $v_i = 0$.
  - Den Kontext einer Instanz legen wir als den Satz fest, in dem die Instanz vorkommt (alternativ: das Fenster mit fester Länge von k Wörtern rechts und links von der Instanz (z.B. k=5)).
- Merkmalsextraktion: Lemmatisierung

Extrahierte Merkmalsmuster

Instanz Nr	...	A237	A295	A303	A452	...
Annotation	...	bank1	bank1	bank2	bank2	...
Frage	...	0	0	0	1	...
Komfort	...	1	0	0	0	...
anlegen	...	0	0	1	0	...
Betrag	...	0	0	1	0	...
Garten	...	0	1	0	0	...
suchen	...	0	1	0	0	...
fahren	...	0	0	0	0	...
richtig	...	0	0	0	0	...
Test	...	0	0	0	1	...
...	...	...	...	...	...	...

Merkmalsextraktion

Instanz Nr	A237	A295	A303	A452	...
Annotation	bank1	bank1	bank2	bank2	...
Frage	0	0	0	1	...
Komfort	1	0	0	0	...
anlegen	0	0	1	0	...
Betrag	0	0	1	0	...
Garten	0	1	0	0	...
suchen	0	1	0	0	...
fahren	0	0	0	0	...
richtig	0	0	0	0	...
Test	0	0	0	1	...
...	...	...	...	...	...

Neue Instanz:

Keine Frage: In einen ordentlichen Garten gehören neben einer Bank auch die richtigen Möbel.

Neu
?
1
0
0
0
1
0
0
1
0
...

Statistische Modellierung: WSD

- Wie bestimmen wir den Wortsinn einer neuen Instanz von *Bank* auf der Grundlage des Musters von Kontextwörtern v ?

Versuch: Analog zum POS-Modell der letzten Woche:

- Wir zählen zunächst für jedes Merkmalsmuster aus, wie oft im Trainingskorpus das Muster mit *bank1* und *bank2* vorkommt.
- Wir schätzen die bedingten Wahrscheinlichkeiten *P(bank1|v)* und *P(bank2|v)* auf der Grundlage dieser Frequenzen.

$$P(s \mid v) = \frac{P(s,v)}{P(v)} \approx \frac{Fr(s,v)}{Fr(v)}$$

- Wir weisen den wahrscheinlicheren Wortsinn zu.

- Sparse-Data-Problem!

Das Bayessche Theorem

- Das Bayessche Theorem oder die Bayes-Regel:

$$P(E \mid F) = \frac{P(F \mid E) \cdot P(E)}{P(F)}$$

- Die Bayes-Regel ist ein elementares Gesetz der Wahrscheinlichkeitstheorie. Sie ist überall da nützlich, wo der Schluss von einer Größe F auf eine andere Größe E bestimmt werden soll (typischerweise von einem Symptom auf eine relevante Eigenschaft/ die Ursache), die Abhängigkeit in der anderen Richtung (von der Ursache auf das Symptom) aber besser zugänglich ist.

Das Bayessche Theorem

- Merkmalsmuster v : Symptom
- Wortsinn s : Ursache

- Mit Bayes-Regel : $P(s \mid v) = \frac{P(v \mid s) \cdot P(s)}{P(v)}$

- Der wahrscheinlichste Wortsinn: $\max_s P(s \mid v) = \max_s \frac{P(v \mid s) \cdot P(s)}{P(v)}$

$$= \max_s P(v \mid s) \cdot P(s)$$

- $P(s)$ ist die globale, "a priori"-Wahrscheinlichkeit des Wortsinns s.
- $P(v)$, die Wahrscheinlichkeit des Merkmalsmusters, wird nicht mehr benötigt.
- Wie ermitteln wir $P(v \mid s)$? – Sparse-Data-Problem!

Unabhängigkeitsannahme

- Es gibt sehr viele unterschiedliche Kontextmuster. Die Auftretenshäufigkeit eines bestimmten Musters mit einer Lesart ist typischerweise klein (meistens sogar 0) und erlaubt deshalb keine verlässliche Abschätzung von *P(v|s)*.

- Unter der („naiven") Voraussetzung, dass die einzelnen Wortfrequenzen unabhängig voneinander sind, lässt sich approximieren als Produkt der Einzelwahrscheinlichkeiten für die Komponenten der Kontextmuster:

$$P(v \mid s) \approx \prod_{v_i} P(v_i \mid s)$$

- Maschinelle Lernverfahren, die diese Annahme nutzen, um Wahrscheinlichkeiten trotz geringer Datenmengen zu approximieren, heißen "Naive Bayes Classifier"

Von binären Merkmalsmustern ...

Instanz Nr	...	A237	A295	A303	A452	...
Annotation	...	bank1	bank1	bank2	bank2	...
Frage	...	0	0	0	1	...
Komfort	...	1	0	0	0	...
anlegen	...	0	0	1	0	...
Betrag	...	0	0	1	0	...
Garten	...	0	1	0	0	...
suchen	...	0	1	0	0	...
fahren	...	0	0	0	0	...
richtig	...	0	0	0	0	...
Test	...	0	0	0	1	...
...	...	...	...	...	...	...

... zu wortsinnspezifischen Kontextwort-Frequenzen

	bank1	bank2
Frage	11	15
Komfort	7	3
anlegen	3	84
Betrag	5	41
Garten	40	1
suchen	7	32
fahren	4	24
richtig	12	21
Test	2	5
...	...	...

... zu Wahrscheinlichkeitsschätzungen von Kontextmerkmalen

- Wir gehen von insgesamt 500 Instanzen von "Bank" im Trainingskorpus aus, davon 200 als bank1 und 300 als bank2 annotiert.

	bank1	bank2
Frage	11	15
Komfort	7	3
anlegen	3	84
Betrag	5	41
Garten	40	1
suchen	7	32
fahren	4	24
richtig	12	21
Test	2	5
...	...	...

	$P(1\|bank1)$	$P(0\|bank1)$
Frage	0,055	0,945
Komfort	0,035	0,965
anlegen	0,015	0,985
Betrag	0,025	0,975
Garten	0,200	0,800
suchen	0,035	0,965
fahren	0,020	0,980
richtig	0,060	0,940
Test	0,010	0,990
...	...	...

... zu Wahrscheinlichkeitsschätzungen von Kontextmerkmalen

- Wir gehen von insgesamt 500 Instanzen von "Bank" im Trainingskorpus aus, davon 200 als bank1 und 300 als bank2 annotiert.

	bank1	bank2
Frage	11	15
Komfort	7	3
anlegen	3	84
Betrag	5	41
Garten	40	1
suchen	7	32
fahren	4	24
richtig	12	21
Test	2	5
...	...	...

	$P(1\|bank2)$	$P(0\|bank2)$
Frage	0,050	0,950
Komfort	0,010	0,990
anlegen	0,280	0,720
Betrag	0,137	0,863
Garten	0,003	0,997
suchen	0,107	0,893
fahren	0,080	0,920
richtig	0,070	0,930
Test	0,017	0,983
...	...	...

Beispiel

Keine Frage: In einen ordentlichen Garten gehören neben einer Bank auch die richtigen Möbel.

$$\max_s P(s \mid v) = \max_s P(v \mid s) \cdot P(s)$$

$$s \in \{bank1, bank2\}$$

$$P(v \mid bank1) \approx \prod_{v_i} P(v_i \mid bank1)$$

$$P(v \mid Bank2) \approx \prod_{v_i} P(v_i \mid Bank2)$$

$P(v_i \mid bank1)$

	$P(1\|bank1)$	$P(0\|bank1)$
Frage	0,055	0,945
Komfort	0,035	0,965
anlegen	0,015	0,985
Betrag	0,025	0,975
Garten	0,200	0,800
suchen	0,035	0,965
fahren	0,020	0,980
richtig	0,060	0,940
Test	0,010	0,990
...	...	...

v_i	$P(v_i\|bank1)$
1	0,055
0	0,965
0	0,985
0	0,975
1	0,200
0	0,965
0	0,980
1	0,060
0	0,990
...	...

$$P(v \mid bank1) \approx \prod_{v_i} P(v_i \mid bank1) = 0{,}000572$$

$P(v_i \mid bank2)$

	$P(1\|bank2)$	$P(0\|bank2)$
Frage	0,050	0,950
Komfort	0,010	0,990
anlegen	0,280	0,720
Betrag	0,137	0,863
Garten	0,003	0,997
suchen	0,107	0,893
fahren	0,080	0,920
richtig	0,070	0,930
Test	0,017	0,983
...	...	...

v_i	$P(v_i\|bank2)$
1	0,050
0	0,990
0	0,720
0	0,863
1	0,003
0	0,893
0	0,920
1	0,070
0	0,983
...	...

$$P(v \mid Bank2) \approx \prod_{v_i} P(v_i \mid Bank2) = 0{,}000006$$

Beispiel

Keine Frage: In einen ordentlichen Garten gehören neben einer Bank auch die richtigen Möbel.

$$\max_s P(s \mid v) = \max_s P(v \mid s) \cdot P(s)$$

$$s \in \{bank1, bank2\}$$

$$P(v \mid bank1) \approx \prod_{v_i} P(v_i \mid bank1) = 0{,}000572 \qquad P(bank1) = 0{,}4$$

$$P(v \mid bank2) \approx \prod_{v_i} P(v_i \mid bank2) = 0{,}000006 \qquad P(bank2) = 0{,}6$$

$$P(v \mid bank1) \cdot P(bank1) = 0{,}000228$$

$$P(v \mid bank2) \cdot P(bank2) = 0{,}000004$$

$$\max_s P(s \mid v) = bank1$$

WSD

- Eine der schwierigsten Aufgaben in der Computerlinguistik:
- Sehr viele Wörter sind auf sehr unterschiedliche Weise mehrdeutig. Man benötigt riesige Mengen von Trainingsmaterial.
- Alle bisher vorgestellten Lernverfahren sind „überwachte“ (supervised) Lernvefahren: Sie erfordern die manuelle Annotation eines Trainingskorpus.
- Attraktiver sind „halbüberwachte“ (semi-supervised) Verfahren, bei denen ein großes Trainingskorpus (teil-)automatisch auf der Grundlage einer kleinen Menge von handannotierten „Seed-Daten“ erzeugt wird.
- Noch attraktiver sind „unüberwachte“ statistische Verfahren, die Resultate ohne jede Annotation erzielen. Dazu kommen wir leider in der Einführungsvorlesung nicht.