Mathematische Grundlagen III

Informationstheorie in der Computerlinguistik: Kollokationen, Subkategorisierungsrahmen, semantische Präferenzen

Vera Demberg

Universität des Saarlandes

2. Juli 2012

Inhaltsverzeichnis

1. Wiederholung Informationstheorie
2. Kollokationsextraktion
   - Frequenzen
   - Mittelwert und Varianz
   - Hypothesentests
   - Pointwise Mutual Information
3. Subkategorisierungsrahmen
4. Selektionspräferenzen und Selektionsrestriktionen

Inhaltsverzeichnis

1. Wiederholung Informationstheorie
2. Kollokationsextraktion
   - Frequenzen
   - Mittelwert und Varianz
   - Hypothesentests
   - Pointwise Mutual Information
3. Subkategorisierungsrahmen
4. Selektionspräferenzen und Selektionsrestriktionen

Entropie

- **Entropie:** Durchschnittliche Unsicherheit darüber, welches Ereignis als nächstes stattfindet

$$H(x) = -\sum_x p(x) \log_2 p(x)$$

- **Gemeinsame Entropie:** Wieviel Information wird benötigt um den Wert zweier Zufallsvariablen anzugeben?

$$H(X, Y) = H(p(x, y)) = -\sum_x \sum_y p(x, y) \log_2 p(x, y)$$

- **Bedingte Entropie:** Wieviel Information wird benötigt um Y mitzuteilen, wenn X schon bekannt ist?

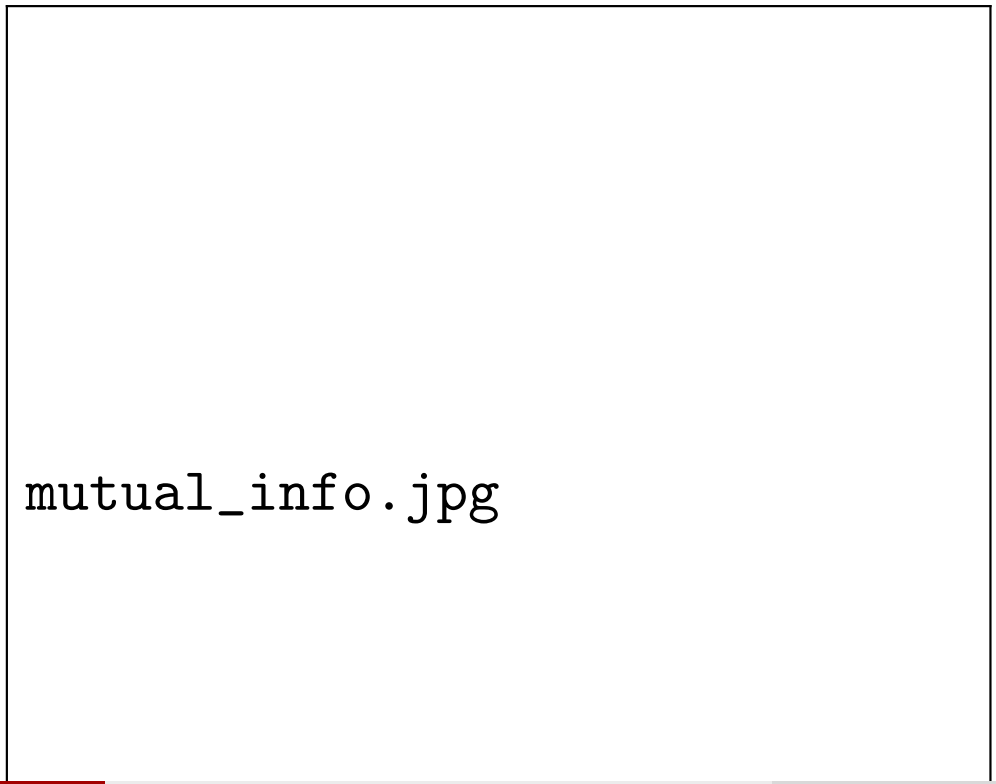
$$H(Y|X) = -\sum_x \sum_y p(x, y) \log_2 p(y|x)$$

Entropie

- **Gegenseitige Information (Mutual Information):** Wie viel Information ist in einer Variable über die andere Variable enthalten?

$$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) = H(Y) + H(X) - H(X,Y)$$

$$I(X;Y) = \sum_{xy} p(x,y) \log_2 \frac{p(x,y)}{p(x)p(y)}$$



Entropie

- **Relative Entropie (Kullback-Leibler Divergenz):**
  Durchschnittliche Anzahl von Bits, die verschwendet werden, wenn eine Verteilung p mit einem für q entwickelten Code enkodiert wird.

$$D(p||q) = \sum_x p(x) \log_2 \frac{p(x)}{q(x)}$$

- **Kreuz-Entropie:** Wenn wir ein Modell m eines Phänomens p (z.B. "englische Sprache") bauen, wollen wir $D(p||m)$ niedrig halten

$$H(X, m) = H(X) + D(p||m) = \sum_x p(x) \log_2 \frac{1}{m(x)}$$

Inhaltsverzeichnis

1 Wiederholung Informationstheorie

2 Kollokationsextraktion
- Frequenzen
- Mittelwert und Varianz
- Hypothesentests
- Pointwise Mutual Information

3 Subkategorisierungsrahmen

4 Selektionspräferenzen und Selektionsrestriktionen

Was ist eine Kollokation?

- ein konventionalisierter Ausdruck aus zwei oder mehr Wörtern
- Beispiele:
  - ins Gras beißen
  - den Löffel abgeben
  - steife Brise
  - Foto schießen
  - Entscheidung fällen
- Fachausdrücke können auch als Kollokationen betrachtet werden
- manchmal werden auch Eigennamen wie Kollokationen behandelt (e.g. New York)
- unterschiedliche Definitionen:
  - nur aufeinanderfolgende Wörter (vorallem in Arbeiten mit eher technischem / statistischen Hintergrund)
  - auch weiter entfernte Wörter (vorallem in linguistischen Arbeiten)
  - letztendlich kommt es darauf an, wofür man die Kollokationen braucht.

Wofür braucht man Kollokationen?

- Leute, die Wörterbücher schreiben und sammeln, in welchen speziellen Kontexten welche Wörter auftauchen
  z.B., Verwendung von “steif” in “steife Brise” aber nicht “steifer Wind”.
- (maschinelle) Übersetzung
  z.B., “prendre une photo” nicht zu übersetzen als “ein Foto nehmen”
- Text Generierung
  korrektes Verb wählen, auch wenn das Verb selbst nicht viel Inhalt hat, z.B. ”eine Entscheidung fällen” statt “eine Entscheidung machen”
- Textverstehen
  Fälle von Wörtern, deren Bedeutung nicht kompositionell ist, erkennen, z.B. “den Löffel abgeben”
- Information Retrieval & automatische Fragebeantwortung
  Bessere Ergebnisse wenn Kollokationen für Retrieval benutzt anstelle von Wörtern.

Fragestellung

Können wir Kollokationen automatisch aus einem Korpus extrahieren?

- Verwendung linguistischer Information: in welchen Konfigurationen treten Kollokationen auf?
- Wie finden wir Kollokationen von Wörtern, die direkt nebeneinander auftauchen vs. Kollokationen mit variierender Entfernung?
- Wie finden wir seltene Kollokationen?

Inhaltsverzeichnis

1 Wiederholung Informationstheorie

2 Kollokationsextraktion
- Frequenzen
- Mittelwert und Varianz
- Hypothesentests
- Pointwise Mutual Information

3 Subkategorisierungsrahmen

4 Selektionspräferenzen und Selektionsrestriktionen

Frequenzen

Einfachste Idee zur Kollokationsextraktion: Bigram-Frequenzen

- Kollokationen sollten häufiger zusammen vorkommen als zufällige Wörter.
- Berechne alle Wortbigramme aus Korpus, zähle, sortiere.

Beispiel

$C(w_1\ w_2)$	w_1	w_2
80871	of	the
58841	in	the
26430	to	the
21842	on	the
18568	and	the
16121	that	the
15630	at	the
15494	to	be
13899	in	a
11428	New	York
10007	he	said

Frequenzen

Einfachste Idee zur Kollokationsextraktion: Bigram-Frequenzen

- Kollokationen sollten häufiger zusammen vorkommen als zufällige Wörter.
- Berechne alle Wortbigramme aus Korpus, zähle, sortiere.

Beispiel

$C(w_1\ w_2)$	w_1	w_2
80871	of	the
58841	in	the
26430	to	the
21842	on	the
18568	and	the
16121	that	the
15630	at	the
15494	to	be
13899	in	a
11428	New	York
10007	he	said

Frequenzen

Einfachste Idee zur Kollokationsextraktion: Bigram-Frequenzen

- Kollokationen sollten häufiger zusammen vorkommen als zufällige Wörter.
- Berechne alle Wortbigramme aus Korpus, zähle, sortiere.
- Wir wollen keine solchen Funktionswörter – baue einen POS-tag Filter ein: A-N, N-N,...
- mit POS-tag Filter bekommen wir schon deutlich bessere Resultate

Beispiel

$C(w_1\ w_2)$	w_1	w_2
80871	of	the
58841	in	the
26430	to	the
21842	on	the
18568	and	the
16121	that	the
15630	at	the
15494	to	be
13899	in	a
11428	New	York
10007	he	said

Frequenzen

Einfachste Idee zur Kollokationsextraktion: Bigram-Frequenzen

- Kollokationen sollten häufiger zusammen vorkommen als zufällige Wörter.
- Berechne alle Wortbigramme aus Korpus, zähle, sortiere.
- Wir wollen keine solchen Funktionswörter – baue einen POS-tag Filter ein: A-N, N-N,...
- mit POS-tag Filter bekommen wir schon deutlich bessere Resultate

Beispiel

$C(w_1\ w_2)$	w_1	w_2
11487	New	York
7261	United	States
5412	Los	Angeles
3301	last	year
3191	Saudi	Arabia
2699	last	week
2514	vice	president
2378	Persian	Gulf
2161	San	Francisco
2001	President	Bush
1942	Saddam	Hussein

Inhaltsverzeichnis

1 Wiederholung Informationstheorie

2 Kollokationsextraktion
- Frequenzen
- Mittelwert und Varianz
- Hypothesentests
- Pointwise Mutual Information

3 Subkategorisierungsrahmen

4 Selektionspräferenzen und Selektionsrestriktionen

Variierende Distanz zwischen Wörtern

- Die vorher genannte Bigram-Frequenz-mit-Filter Methode kann keine Kollokationen finden, bei denen die Kollokationsteile weiter auseinander sind
  ("Die endgültige **Entscheidung** wurde gestern **gefällt**.")
- Kollokationsfenster: 3-4 Worte um das Wort herum

Beispiel

$Wort_1$	$Wort_2$	Distanz
Die	Entscheidung	-2
endgültige	Entscheidung	-1
Enscheidung	wurde	1
Entscheidung	gestern	2
Entscheidung	gefällt	3

Variierende Distanz zwischen Wörtern

- Die vorher genannte Bigram-Frequenz-mit-Filter Methode kann keine Kollokationen finden, bei denen die Kollokationsteile weiter auseinander sind
  ("Die endgültige **Entscheidung** wurde gestern **gefällt**.")
- Kollokationsfenster: 3-4 Worte um das Wort herum
- Mittelwert und Varianz: Idee – gute Kollokationen stehen in stabilem Verhältnis, geringe Varianz

Beispiel

var	mean	count	$Wort_1$	$Wort_2$
0.43	0.97	11657	New	York
0.48	1.83	24	previous	games
0.15	2.98	46	minus	points
4.03	0.44	36	editorial	Atlanta
3.96	0.29	106	subscribers	by

Distanz und Varianz

So ganz toll funktioniert die Distanz und Varianz-Methode nicht.

- Varianz könnte zufällig gering sein, weil wir so wenige Vorkommen gesehen haben.
- Manche Wortpaare (z.B. new companies) kommen vielleicht häufig vor, einfach weil beide Wörter häufig sind, und nicht weil es eine Kollokation ist.
- Geringe Varianz in Distanz ist nur eine Annäherung daran, was wir eigentlich wollen: Wortpaare mit stabiler syntaktischer Relation.

$\rightarrow$ Hypothesentests und schlauere Filter

Inhaltsverzeichnis

1 Wiederholung Informationstheorie

2 Kollokationsextraktion
- Frequenzen
- Mittelwert und Varianz
- Hypothesentests
- Pointwise Mutual Information

3 Subkategorisierungsrahmen

4 Selektionspräferenzen und Selektionsrestriktionen

Hypothesentests

- Problem mit bisheriger Diskussion:
  Wie unterscheiden wir Wortpaare, die zufällig eine hohe Häufigkeit und geringe Distanz-Varianz von solchen die Kollokationen sind?
- Was wir eigentlich wissen wollen: Kommt ein Wortpaar häufiger zusammen vor, als man durch Zufall erwarten könnte?
- Statistische Antwort: Hypothesentest.
- **Null-Hypothese** H_0: ein Wortpaar kommt nur zufällig zusammen vor.
  H_0: $P(w_1\ w_2) = P(w_1) * P(w_2)$
- Wir wollen H_0 vergleichen mit tatsächlicher Beobachtung

Wir besprechen:

- t-test
- χ^2 test

t Test

- Test ob beobachtete Frequenz w_1 w_2 signifikant häufiger als zu erwarten gegeben die Frequenzen von w_1 und w_2.
- Berechnet Unterschied zwischen erwartetem und beobachteten Vorkommenshäufigkeiten
- Skaliert in Bezug auf Varianz der Daten

Formel:

$$t = \frac{\bar{x} - \mu}{\sqrt{\frac{s^2}{N}}}$$

t Test

Formel:

$$t = \frac{\bar{x} - \mu}{\sqrt{\frac{s^2}{N}}}$$

- $\bar{x} = P(into\ them) = \frac{20}{14307668} = 1.3978^{-6}$
- $\mu = P(into) \times P(them) = \frac{14734}{14307668} \times \frac{13478}{14307668} = 9.7008^{-7}$
- $s^2 = p(1-p) \approx p = 1.3978^{-6}$[a]
- $N = 14307668$

Beispiel

Wortpaar: "into them"

Freq(into)	=	14 734
Freq(them)	=	13 478
Freq(into them)	=	20
total # bigrams	=	14 307 668

[a] Bernouilli Experiment bei dem "into them" mit Wahrscheinlichkeit p gesehen wird.

t Test

Formel:

$$t = \frac{\bar{x} - \mu}{\sqrt{\frac{s^2}{N}}} = \frac{1.3978^{-6} - 9.7008^{-7}}{\sqrt{\frac{1.3978^{-6}}{14307668}}} = 1.368561 \text{ (nicht signifikant)}$$

- $\bar{x} = P(into\ them) = \frac{20}{14307668} = 1.3978^{-6}$
- $\mu = P(into) \times P(them) = \frac{14734}{14307668} \times \frac{13478}{14307668} = 9.7008^{-7}$
- $s^2 = p(1-p) \approx p = 1.3978^{-6}$[a]
- $N = 14307668$

Beispiel

Wortpaar: "into them"

Freq(into)	=	14 734
Freq(them)	=	13 478
Freq(into them)	=	20
total # bigrams	=	14 307 668

[a] Bernouilli Experiment bei dem "into them" mit Wahrscheinlichkeit p gesehen wird.

t Test

Results:

t	C(w_1)	C(w_2)	C(w_1 w_2)	w_1	w_1
4.4721	42	20	20	Ayatollah	Ruhollah
4.4720	77	59	20	videocassette	recorder
4.4720	24	320	20	unsalted	butter
2.3714	14907	9017	20	first	made
1.3685	14734	13478	20	into	them
0.8036	15019	15629	20	time	last

- mit Hilfe des t Tests können wir jetzt die Qualität von Bigrammen mit gleicher Kookkurenzfrequenz (hier 20) bewerten.

Zurück zu unserer Nullhypothese

- Nullhypothese H_0: $P(w_1\ w_2) = P(w_1) * P(w_2)$
- natürlich sind Wörter nicht zufällig verteilt, sondern in Abhängigkeit von Satzstruktur, daher werden wir mit dieser Nullhypothese fast immer finden, dass ein gemeinsames Vorkommen wahrscheinlicher als Zufall ist
- tatsächliche Signifikanzlevel sind in dieser Formulierung also nicht sehr aussagekräftig
- nützlich ist aber das Ranking, das wir durch den t-Test bekommen.
- t-Test: häufig benutzt für Kollokationsextraktion, aber problematisch
  - nimmt Normalverteilung an, die aber nicht gegeben ist.

χ^2 Test

- χ^2 Test nimmt keine Normalverteilung an
- also besser geeignet für unsere Daten.
- vergleicht auch beobachtete gemeinsame Frequenz mit erwarteter Frequenz bei Unabhängigkeit

χ^2 Test

Formel:

$$\chi^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = 2.704 \text{ (nicht signifikant)}$$

O_{ij}	w_1 =into	$w_1 \neq$into
w_2 =them	20	13458
$w_2 \neq$them	14714	14279476

$$E_{ij} = \frac{O_i}{N} \times \frac{O_j}{N} \times N$$

$$E_{11} = \frac{14734}{14307668} \times \frac{13478}{14307668} \times 14307668 = 13.87$$

E_{ij}	w_1 =into	$w_1 \neq$into
w_2 =them	13.87	13464.10
$w_2 \neq$them	14720.1	14279430

Beispiel

Wortpaar: "into them"

Freq(into)	=	14 734
Freq(them)	=	13 478
Freq(into them)	=	20
total # bigrams	=	14 307 668

χ^2 Test

- In der Praxis, Resultate ähnlich zu t Test
- Funktioniert besser bei großen Wahrscheinlichkeiten, für die die Normalverteilung nicht gilt.
- Aber: χ^2 Test problematisch in Fällen wo weniger Beobachtungen vorliegen (gesamt < 20 oder gesamt $20 - 40$ und ein Eintrag < 5.

Inhaltsverzeichnis

1 Wiederholung Informationstheorie

2 Kollokationsextraktion
- Frequenzen
- Mittelwert und Varianz
- Hypothesentests
- Pointwise Mutual Information

3 Subkategorisierungsrahmen

4 Selektionspräferenzen und Selektionsrestriktionen

Pointwise Mutual Information

$$
\begin{aligned}
I(x,y) &= \log_2 \frac{P(x\ y)}{P(x)P(y)} \\
&= \log_2 \frac{P(x|y)}{P(x)} \\
&= \log_2 \frac{P(y|x)}{P(y)}
\end{aligned}
$$

- Interpretation: Menge an Information über eine Variable, die in einer anderen Variable enthalten ist.

Pointwise Mutual Information

Formel:

$$I(x, y) = \log_2 \frac{P(x\ y)}{P(x)P(y)}$$

Mutual Information for "into them"

- $P(into\ them) = \frac{20}{14397688}$
- $P(into) = \frac{14734}{14397688}$
- $P(them) = \frac{13478}{14397688}$
- $I(into, them) = 0.527$

MI for "Ayatollah Ruhollah"

- $P(Ayatollah\ Ruhollah) = \frac{20}{14397688}$
- $P(Ayatollah) = \frac{42}{14397688}$
- $P(Ruhollah) = \frac{20}{14397688}$
- $I(Ayatollah, Ruhollah) = 18.38$

Beispiel

Wortpaar: "into them"

Freq(into)	=	14 734
Freq(them)	=	13 478
Freq(into them)	=	20
total # bigrams	=	14 307 668

Wortpaar: "Ayatollah Ruhollah"

Freq(Ayatollah)	=	42
Freq(Ruhollah)	=	20
Freq(Ayat.Ruho.)	=	20
total # bigrams	=	14 307 668

Pointwise Mutual Information

Formel:

$$I(x, y) = \log_2 \frac{P(x\ y)}{P(x)P(y)}$$

Unsicherheit bzgl. Ruhollah um 18,38 bits reduziert gegeben Ayatollah.

$I(w_1, w_2)$	t	C(w_1)	C(w_2)	C($w_1\ w_2$)	w_1	w_1
18.38	4.4721	42	20	20	Ayatollah	Ruhollah
15.94	4.4720	77	59	20	videocassette	recorder
15.19	4.4720	24	320	20	unsalted	butter
1.09	2.3714	14907	9017	20	first	made
0.527	1.3685	14734	13478	20	into	them
11.25	0.999	43	267	1	fewest	visits
0.29	0.8036	15019	15629	20	time	last

Problem mit Abschätzungen bei seltenen Ereignissen.

Pointwise Mutual Information

Formel:

$$I(x, y) = \log_2 \frac{P(x\ y)}{P(x)P(y)}$$

Unsicherheit bzgl. Ruhollah um 18,38 bits reduziert gegeben Ayatollah.

$I(w_1, w_2)$	t	C(w_1)	C(w_2)	C($w_1\ w_2$)	w_1	w_1
18.38	4.4721	42	20	20	Ayatollah	Ruhollah
15.94	4.4720	77	59	20	videocassette	recorder
15.19	4.4720	24	320	20	unsalted	butter
1.09	2.3714	14907	9017	20	first	made
0.527	1.3685	14734	13478	20	into	them
11.25	0.999	43	267	1	fewest	visits
0.29	0.8036	15019	15629	20	time	last

Problem mit Abschätzungen bei seltenen Ereignissen.

Seltene Ereignisse

- Alle Maße, die wir bislang gesehen haben sind problematisch bei seltenen Ereignissen.
- Besonders schlimm bei Mutual Information

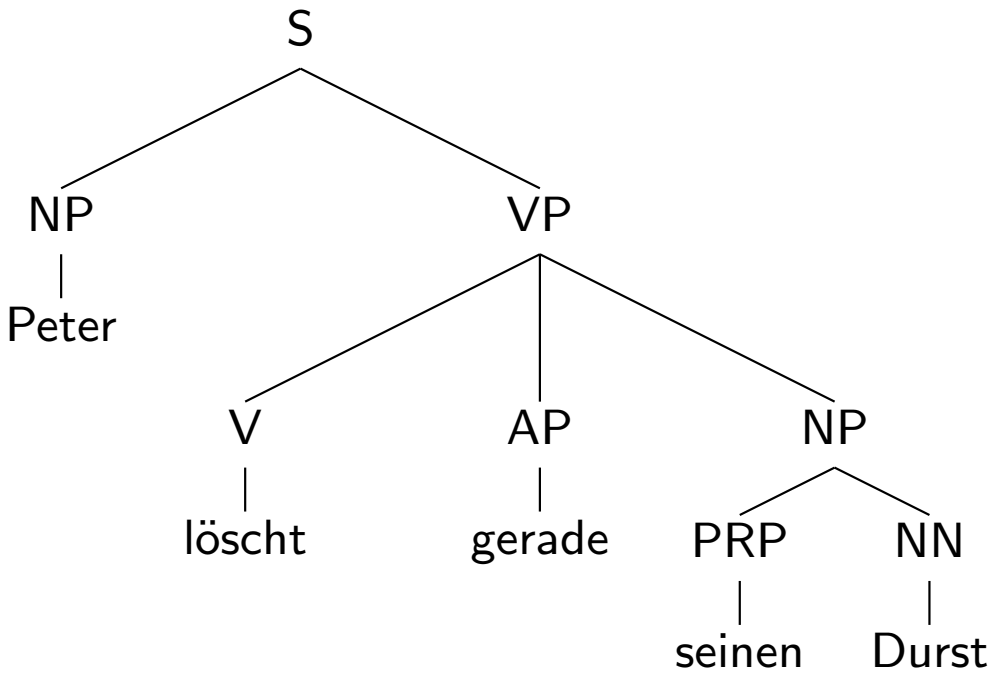
MI für 100% miteinander korrelierte Ereignisse

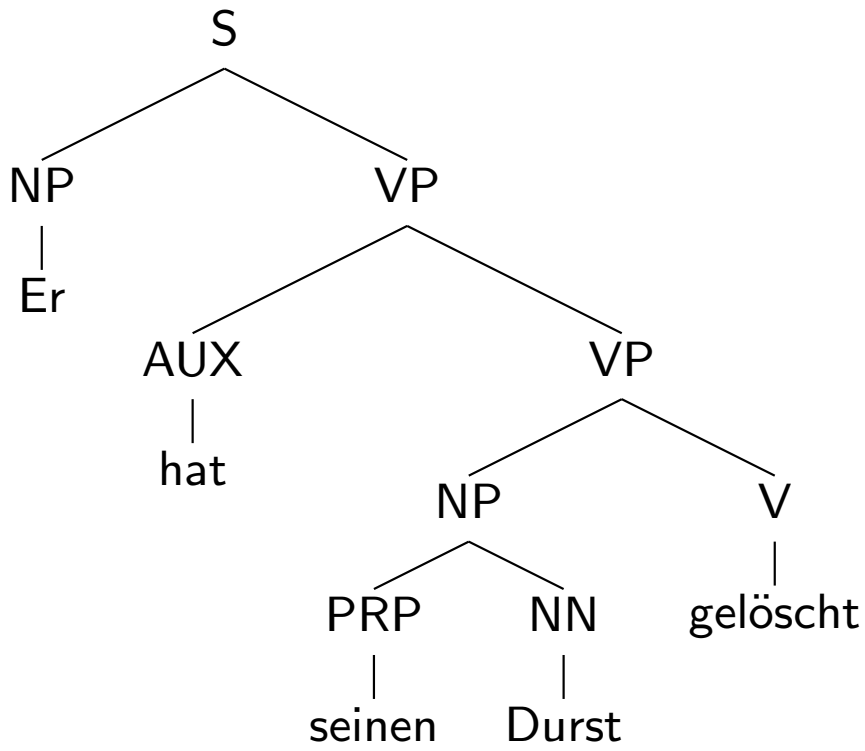
$$I(x,y) = \log \frac{P(x,y)}{P(x)P(y)} = \log \frac{P(x)}{P(x)P(y)} = \log \frac{1}{P(y)}$$

- Je seltener das Ereignis (d.h. je kleiner P(y)), desto größer ist MI!
- Das ist das Gegenteil von dem was wir wollen (je mehr Beobachtungen, desto können wir in unserer Bewertung sein).
- MI: gutes Maß für Unabhängigkeit, schlechtes Maß für Abhängigkeit!
- Möglichkeiten, MI für Kollokationsextraktion zu nutzen:
  - Cutoff: seltene Ereignisse nicht bewerten
  - Gewichtung durch Frequenz von w_1 w_2.

Schlauere Filter

- viele interessante Kollokationen sind Verb-Objekt Paare (z.B. “Durst löschen”).
- diese können unterschiedlich weit von einander entfernt sein
- sogar in unterschiedlicher Reihenfolge auftauchen
- Lösung: Distanz auf Syntax definieren anstelle von Wortfolge





Zusammenfassung Kollokationen

- Kollokationen: Wörter, die häufig zusammen auftreten und in besonders engem Verhältnis stehen
- unterschiedlich strenge Definitionen von Kollokation, je nach Anwendung
- Einfache Ansätze: Frequenzen und Varianz in Distanz zwischen Wörtern
- Hypothesentests für zusammen auftretende Wörter (zum Ranking von möglichen Kollokationen): t Test und χ^2 Test
- Pointwise Mutual Information als Kriterion für Kollokationsextraktion
- Schwierigkeit der Bewertung von selten auftauchenden Ereignissen

Weitere Beispiele

Kollokationen sind nur ein Beispiel von Mustern, die wir aus Text lernen wollen. Jetzt: allgemeinere syntaktische und semantische Eigenschaften von Wörtern

- Subkategorisierungsrahmen: welche Verben nehmen welche Argumente (direktes / indirektes / präpositional Objekt)?
  $\rightarrow$ Hypothesentests
- Selektionspräferenzen: welche Verben haben nehmen was für Argumente? (Belebte Wesen / essbare Dinge / Werkzeuge...)
  $\rightarrow$ Relative Entropie (Kullback-Leibler Divergenz)

Inhaltsverzeichnis

1 Wiederholung Informationstheorie

2 Kollokationsextraktion
- Frequenzen
- Mittelwert und Varianz
- Hypothesentests
- Pointwise Mutual Information

3 Subkategorisierungsrahmen

4 Selektionspräferenzen und Selektionsrestriktionen

Subkategorisierungsrahmen für Verben

Beispiele für Subkategorisierungsrahmen

Rahmen	Funktion	Verb	Beispiel
NP NP	Subject, Object	lesen	Er liest ein Buch.
NP S	Subject, Nebensatz	hoffen	Er hofft, dass er gewinnt.
NP INF	Subject, Infinitiv	hoffen	Er hofft zu gewinnen.
NP PP NP	Subj, Präp-Obj	hoffen	Er hofft auf besseres Wetter.
NP NP S	Subj, Obj, Nebensatz	sagen	Er hat ihr gesagt dass er gewinnt.
NP NP INF	Subj, Obj, Infinitiv	helfen	Er hat ihr geholfen zu gewinnen.
NP NP NP	Subj, Obj, Dat-Obj	geben	Er hat mir das Buch gegeben.

Subkategorisierungsrahmen sind wichtig um z.B. korrekt parsen zu können.

Beispiel

Sie	erzählte	dem Mann	wo Peter aufgewachsen war.
Sie	fand	den Ort	wo Peter aufgewachsen war.

Subkategorisierungsrahmen für Verben

Beispiele für Subkategorisierungsrahmen

Rahmen	Funktion	Verb	Beispiel
NP NP	Subject, Object	lesen	Er liest ein Buch.
NP S	Subject, Nebensatz	hoffen	Er hofft, dass er gewinnt.
NP INF	Subject, Infinitiv	hoffen	Er hofft zu gewinnen.
NP PP NP	Subj, Präp-Obj	hoffen	Er hofft auf besseres Wetter.
NP NP S	Subj, Obj, Nebensatz	sagen	Er hat ihr gesagt dass er gewinnt.
NP NP INF	Subj, Obj, Infinitiv	helfen	Er hat ihr geholfen zu gewinnen.
NP NP NP	Subj, Obj, Dat-Obj	geben	Er hat mir das Buch gegeben.

Subkategorisierungsrahmen sind wichtig um z.B. korrekt parsen zu können.

Beispiel

Sie erzählte (dem Mann) (wo Peter aufgewachsen war).
Sie fand (den Ort (wo Peter aufgewachsen war)).

Warum Subkategorisierungsrahmen lernen?

- Fehlende Subkategorisierungsinformation sehr häufige Fehlerquelle beim Parsen.
- Viele Verben erlauben mehr als einen Subkategorisierungsrahmen.
- Subkategorisierungsrahmen sind nicht leicht zu erkennen, da auch viele Modifikatoren wie Argumente aussehen (insbesondere wenn wir Präpositionalobjekte lernen wollen).

Beispiele

- Ich habe Donnerstag einen Unfall gesehen.
- Ich habe Peter ein Bild gezeigt.
- Ich warte auf meine Mutter.
- Ich schlafe auf meinem Bett.

Hypothesentests

- Um Subkategorisierungsrahmen zu lernen, können wir Hypothesentests verwenden.
- Idee: Modifikatoren kommen zufällig zusammen mit einem Verb vor, Argumente kommen systematisch vor.
- Null-Hypothese H_0: Verb kommt mit bestimmten Rahmen nicht vor.
- Binomialtest: vergleiche relative Häufigkeit von beobachteten Rahmen mit Wahrscheinlichkeit, dass es nur Modifikatoren sind (Details siehe Brent 1993).
- Gute Präzision, aber leider finden wir viele, vor allem seltene Rahmen mit dieser Methode nicht.

Inhaltsverzeichnis

1 Wiederholung Informationstheorie

2 Kollokationsextraktion
- Frequenzen
- Mittelwert und Varianz
- Hypothesentests
- Pointwise Mutual Information

3 Subkategorisierungsrahmen

4 Selektionspräferenzen und Selektionsrestriktionen

Selektionspräferenzen und Selektionsrestriktionen

Gerade haben wir uns über syntaktische Rahmen von Verben unterhalten. Verben haben aber auch semantische Präferenzen und Beschränkungen auf ihre Argumente.

Beispiel

- "fressen" $\rightarrow$ Subjekt: Tier, Object: Nahrung
- "bellen" $\rightarrow$ Subject: Hund
- "denken" $\rightarrow$ Subject: Mensch
- diese sind Präferenzen: "Computerprogramm frisst Speicher."

Beachte: auch andere Wortarten (z.B. Adjektive) können Selektionspräferenzen oder -restriktionen haben.

Nützlich in NLP

- Bedeutung unbekannter Wörter eingrenzen
- besseres Parsing durch Bevorzugung plausibler Sätze

Selektionspräferenzen

- Traditioneller Ansatz: Features

Beispiel

Schenken:	SUBJEKT	Mensch
	OBJEKT	Physisches Objekt
	INDIR. OBJEKT	Lebewesen

- Problem: unmöglich alle zu spezifizieren, unflexibel und zu restriktiv.

Selektionspräferenzen

- Wie stark schränkt ein bestimmtes Verb sein Objekt ein?
- Starke Selektionspräferenz: eine bestimmten Klasse von Argumenten wird wesentlich häufiger als Argument des Verbs beobachtet als andere Wörter anderer Klassen.
- Schwache Selektionspräferenz: Frequenzverteilung über Klassen unterscheidet sich nicht so stark von der durchschnittlich erwarteten Verteilung.

Zutaten

Was ist eine Klasse? Woher wissen wir, welches Wort zu welcher Klasse gehört? z.B. aus einer Taxonomie, wie z.B. WordNet

- Liquid $>$ Beverage $>$ wine
- Animated $>$ Animal $>$ Mammal $>$ Dog $>$ Poodle

Stärke der Selektionspräferenz

Stärke der Selektionspräferenz S(v) [Resnik 1996]:

$$S(v) = D(P(C|v)||P(C)) = \sum_c P(c|v) \log \frac{P(c|v)}{P(c)}$$

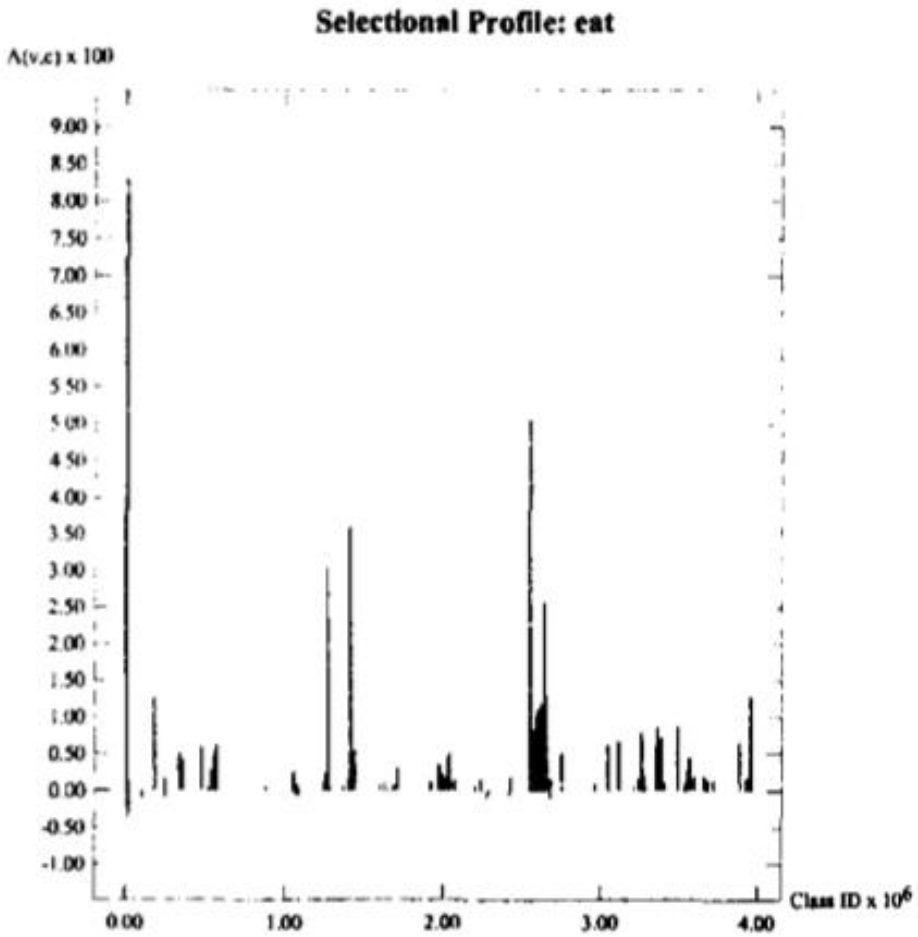
- Intuition: wie unterschiedlich ist die tatsächlich beobachtete Verteilung von Klassen unter den Argumenten eines bestimmten Verbs $P(C|v)$ von der allgemeinen Verteilung von Nominalklassen $P(C)$?
- Dann können wir sagen ob ein Verb stark selektiert oder nicht.

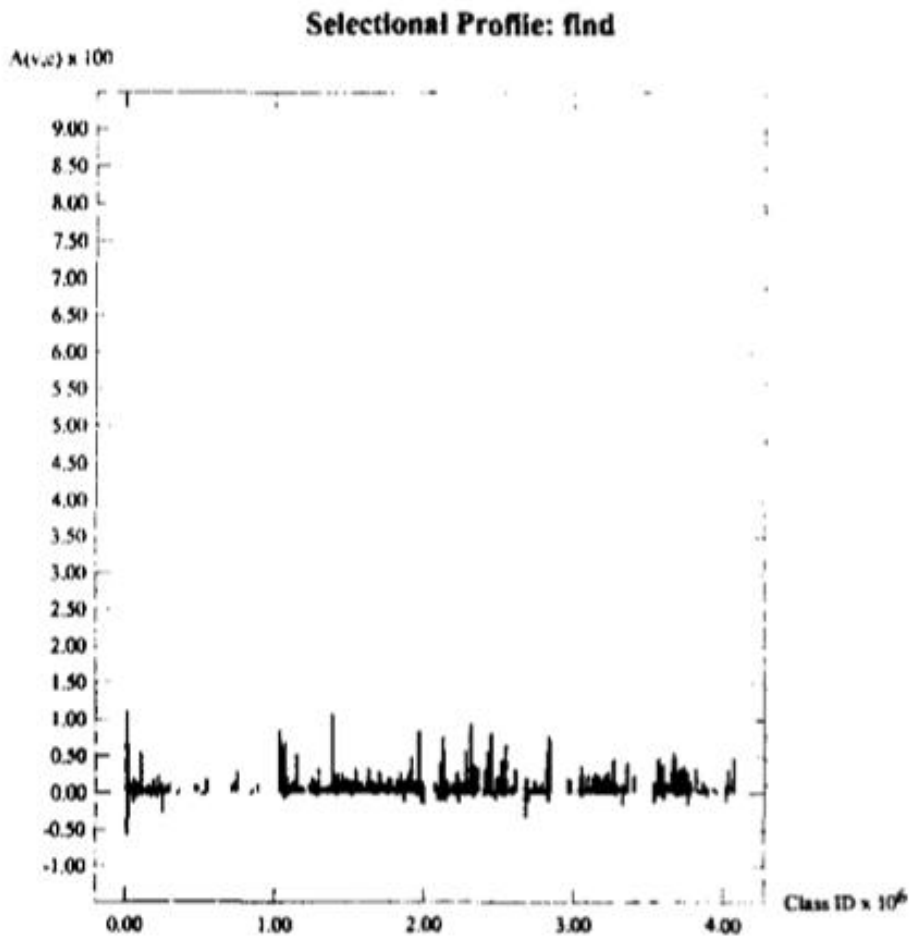
Selektionsstärke von Verben

Strength of selectional preference for direct objects

Verb	Strength			Verb	Strength		
	Brown	CHILDES	Norms		Brown	CHILDES	Norms
pour	4.80	2.30	2.57	explain	2.39	4.41	2.20
drink	4.38	2.38	2.83	read	2.35	2.58	1.81
pack	4.12	3.71	1.75	watch	1.97	1.44	1.86
sing	3.58	3.15	2.63	do	1.84	–	2.21
steal	3.52	2.28	1.34	hear	1.70	1.67	1.71
eat	3.51	1.15	2.47	call	1.52	0.95	2.39
hang	3.35	2.03	1.96	want	1.52	0.70	1.71
wear	3.13	2.02	2.30	show	1.39	1.83	1.42
open	2.93	2.41	1.88	bring	1.33	0.88	1.04
push	2.87	1.77	1.98	put	1.24	0.40	1.34
say	2.82	0.94	2.56	see	1.06	0.48	1.54
pull	2.77	1.55	2.22	find	0.96	0.71	1.30
like	2.59	0.89	1.30	take	0.93	0.74	1.28
write	2.54	2.33	2.18	get	0.82	0.28	1.17
play	2.51	2.13	2.64	give	0.79	1.18	1.81
hit	2.49	1.31	1.91	make	0.72	0.77	1.58
catch	2.47	1.67	1.92	have	0.43	–	1.23

Selektionsprofil von Verben





Stärke der Selektionspräferenz

(Selektionsstärke: $S(v) = D(P(C|v)||P(C)) = \sum_c P(c|v) \log \frac{P(c|v)}{P(c)}$)

Selektionsassoziation:

$$A(v,n) = \frac{P(c|v) \log \frac{P(c|v)}{P(c)}}{\sum_c P(c|v) \log \frac{P(c|v)}{P(c)}}$$

- Intuition: Wie stark trägt eine bestimmte Klasse zu der Selektionspräferenzstärke eines Verbs bei?
- Jetzt können wir sagen, welche Klasse ein Verb selektiert.
- Um zu bewerten, wie gut ein Argument zu einem Verb passt, können wir berechnen zu welcher Klasse ein Nomen gehört und dann wie stark das Verb mit dieser Klasse assoziiert ist.
- das machen wir für jede Bedeutung eines Nomens – falls eine Bedeutung besonders gut zum Verb passt, können wir auch so die Bedeutung des Nomens disambiguieren.

Resultat für Bewertung der Assoziationsstärke zwischen einem Verb und einem Nomen

Verb v	Noun n	$A(v,n)$	Klasse
answer	request	4.49	speech act
answer	tragedy	3.88	communication
hear	story	1.89	communication
hear	issue	1.89	communication
find	label	1.10	abstraction
find	fever	0.22	psych. feature
read	article	6.80	writing
read	fashion	-0.20	activity
see	friend	5.78	entity
see	method	-0.01	method

$$A(v,n) = \frac{P(c|v)\log\frac{P(c|v)}{P(c)}}{\sum_c P(c|v)\log\frac{P(c|v)}{P(c)}}$$

Zusammenfassung

Subkategorisierungsrahmen

- Ziel: automatisch herausfinden, welche Argumente ein Verb nehmen kann
- Nicht trivial: oft haben Modifikatoren gleiche Form wie Argumente
- Idee: valide Argumente sollten mit größerer Häufigkeit beobachtet werden als Modifikatoren
- Hypothesentest mit Binomialtest: akzeptiere Subkategorisierungsrahmen wenn bestimmte Argumente signifikant häufiger mit einem Verb auftreten als erwartet

Selektionspräferenzen

- Welche semantischen Typen nimmt ein Verb als Argument?
- Relative Entropie: Unterscheiden von Verben mit hoher vs. geringer Selektionsstärke
- Assoziationsstärke: mit Hilfe von Entropie bewerten, wie typisch ein Nomen als Argument eines Verbs ist
- Kann so auch zur Wortbedeutungsdisambiguierung genutzt werden.