Probabilistic Context Free Grammars, Part II

Prof Dr. Matthew Crocker

Universität des Saarlandes

16. Juli 2015

Themen heute:

1. Wiederholung: PCFG
2. Formeln und Beispiel: Inside-Outside Algorithmus
3. Unabhängigkeitsannahmen
4. Unabhängigkeitsannahmen abschwächen
   - Lexikalisierte Grammatiken
   - Regeln in Abhängigkeit vom Kontext
5. Evaluation von Parsern
6. Effizient Parsen

Table of Contents

Wiederholung: Probabilistische kontextfreie Grammatiken (PCFGs)

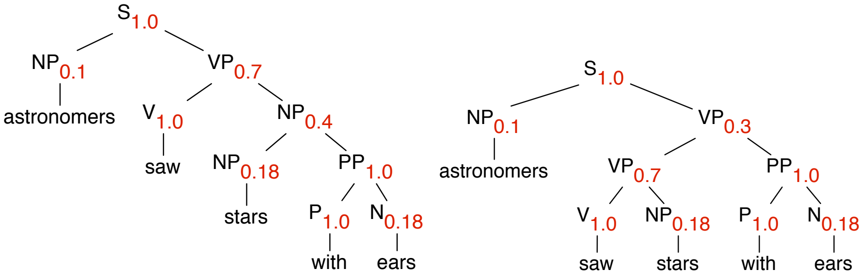
- Eine PCFG ist eine kontextfreie Grammatik, in der jede Regel mit einer Wahrscheinlichkeit versehen ist
- Die Summe der Wahrscheinlichkeiten aller Regeln mit dem selben Symbol auf der *linken* Seite muss 1 betragen

Beispiel

S	→	NP VP	1.0
VP	→	V NP	0.7
VP	→	VP PP	0.3
NP	→	NP PP	0.4
PP	→	P NP	1.0

NP	→	astronomers	0.1
NP	→	telescopes	0.1
NP	→	saw	0.04
NP	→	stars	0.18
NP	→	ears	0.18
P	→	with	1.0
V	→	saw	1.0

Wahrscheinlichkeit eines Satzes



$$P(t1) = 1.0 * 0.18 * 1.0 * 0.18 * 0.4 * 1.0 * 0.7 * 0.1 * 1.0 = 0.009072$$

$$P(t2) = 1.0 * 0.1 * 0.3 * 0.7 * 1.0 * 0.18 * 1.0 * 1.0 * 0.18 = 0.000680$$

Table of Contents

Expectation Maximization Algorithmus

Erinnerung:

- Frage: Woher bekommen wir die Regelwahrscheinlichkeiten?
- Ziel: Abschätzung von P(N→ X Y) als $\frac{C(N \to X\ \ Y)}{C(N)}$
- Wenn wir ein annotiertes Korpus haben, koennen wir $C(N \to X\ \ Y)$ und $C(N)$ einfach auszählen.
- Wenn nicht, können wir mit zufälligen Regelwahrscheinlichkeiten anfangen und die Regelanwendungen dabei auszählen.

Idee bei Inside-Outside Algorithmus (Expectation-Maximization):

- Da wir unseren Ableitungen aber nicht so ganz vertrauen können, da wir ja die Regelwahrscheinlichkeiten noch nicht gut kennen, zählen wir alle Regelanwendungen, die zu einer vollstaendigen Ableitung führen, und gewichten mit der Wahrscheinlichkeit des Satzes.

Inside-Outside Algorithmus (Expectation Maximization)

Wir wollen die Regelwahrscheinlichkeiten lernen, die die Wahrscheinlichkeit unserer Sprache maximieren.

Expectation Schritt

$$E(N \text{ used}) = \frac{\sum_{p=1}^{m} \sum_{q=p}^{m} \alpha_N(p,q)\beta_N(p,q)}{\text{Satzwahrscheinlichkeit}}$$

$$E(N \to X\ Y, N \text{ used}) = \frac{\sum_{p=1}^{m} \sum_{q=p}^{m} \sum_{d=p}^{q-1} \alpha_N(p,q)P(N \to X\ \ Y)\beta_X(p,d)\beta_Y(d+1,q)}{\text{Satzwahrscheinlichkeit}}$$

$$E(N \to w, N \text{ used}) = \frac{\sum_{h=1}^{m} \alpha_N(h,h)P(w=w_h)\beta_N(h,h)}{\text{Satzwahrscheinlichkeit}}$$

Inside-Outside Algorithmus (Expectation Maximization)

Wir wollen die Regelwahrscheinlichkeiten lernen, die die Wahrscheinlichkeit unserer Sprache maximieren.

Expectation Schritt

(Inside probability for β_N ohne Summe über X,Y)

$$E(N \text{ used}) = \frac{\sum_{p=1}^{m}\sum_{q=p}^{m}\alpha_N(p,q)\beta_N(p,q)}{\text{Satzwahrscheinlichkeit}}$$

$$E(N \to X\ Y, N \text{ used}) = \frac{\sum_{p=1}^{m}\sum_{q=p}^{m}\sum_{d=p}^{q-1}\alpha_N(p,q)P(N\to X\ \ Y)\beta_X(p,d)\beta_Y(d+1,q)}{\text{Satzwahrscheinlichkeit}}$$

$$E(N \to w, N \text{ used}) = \frac{\sum_{h=1}^{m}\alpha_N(h,h)P(w=w_h)\beta_N(h,h)}{\text{Satzwahrscheinlichkeit}}$$

Inside-Outside Algorithmus (Expectation Maximization)

Wir wollen die Regelwahrscheinlichkeiten lernen, die die Wahrscheinlichkeit unserer Sprache maximieren.

Expectation Schritt

$$E(N \text{ used}) = \frac{\sum_{p=1}^{m} \sum_{q=p}^{m} \alpha_N(p,q) \beta_N(p,q)}{\text{Satzwahrscheinlichkeit}}$$

$$E(N \to X\ Y, N \text{ used}) = \frac{\sum_{p=1}^{m} \sum_{q=p}^{m} \sum_{d=p}^{q-1} \alpha_N(p,q) P(N \to X\ Y) \beta_X(p,d) \beta_Y(d+1,q)}{\text{Satzwahrscheinlichkeit}}$$

$$E(N \to w, N \text{ used}) = \frac{\sum_{h=1}^{m} \alpha_N(h,h) P(w=w_h) \beta_N(h,h)}{\text{Satzwahrscheinlichkeit}}$$

Inside-Outside Algorithmus (Expectation Maximization)

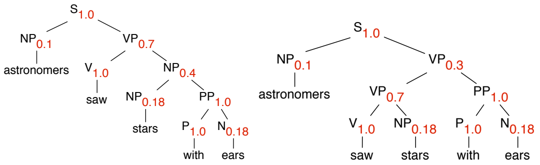
Maximization Schritt

$$\hat{P}(N \rightarrow X\ \ Y) = \frac{E(N \rightarrow X\ \ Y,\ N \text{ used})}{E(N \text{ used})}$$

$$= \frac{\sum_{p=1}^{m} \sum_{q=p}^{m} \sum_{d=p}^{q-1} \alpha_N(p,q) P(N \rightarrow X\ \ Y) \beta_X(p,d) \beta_Y(d+1,q)}{\sum_{p=1}^{m} \sum_{q=p}^{m} \alpha_N(p,q) \beta_N(p,q)}$$

$$\hat{P}(N \rightarrow w) = \frac{E(N \rightarrow w,\ N \text{ used})}{E(N \text{ used})}$$

$$= \frac{\sum_{h=1}^{m} \alpha_N(h,h) P(w = w_h) \beta_N(h,h)}{\sum_{p=1}^{m} \sum_{q=p}^{m} \alpha_N(p,q) \beta_N(p,q)}$$



Inside

Regelwahrscheinlichkeiten

S	$\rightarrow$ NP VP	1.0
VP	$\rightarrow$ V NP	0.7
VP	$\rightarrow$ VP PP	0.3
NP	$\rightarrow$ NP PP	0.4
PP	$\rightarrow$ P NP	1.0
NP	$\rightarrow$ astronomers	0.1
NP	$\rightarrow$ telescopes	0.1
NP	$\rightarrow$ saw	0.04
NP	$\rightarrow$ stars	0.18
NP	$\rightarrow$ ears	0.18
P	$\rightarrow$ with	1.0
V	$\rightarrow$ saw	1.0

Wahrscheinlichkeiten

	1	2	3	4	5
1	$\beta_{NP} = 0.1$		$\beta_S = 0.0126$		$\beta_S = 0.0015876$
2		$\beta_V = 1.0$ $\beta_{NP} = 0.04$	$\beta_{VP} = 0.126$		$\beta_{VP} = 0.015876$
3			$\beta_{NP} = 0.18$		$\beta_{NP} = 0.01296$
4				$\beta_P = 1.0$	$\beta_{PP} = 0.18$
5					$\beta_{NP} = 0.18$
	astronomers	saw	stars	with	ears

Outside Wahrscheinlichkeiten

	1	2	3	4	5
1	$\alpha_{NP} = 0.015876$		$\alpha_S = 0$		$\alpha_S = 1$
2		$\alpha_V = 0.0015876$ $\alpha_{NP} = 0$	$\alpha_{VP} = 0.0054$		$\alpha_{VP} = 0.1$
3			$\alpha_{NP} = 0.00882$		$\alpha_{NP} = 0.07$
4				$\alpha_P = 0.0015876$	$\alpha_{PP} = 0.00882$
5					$\alpha_{NP} = 0.00882$
	0.0015876 astronomers	0.0015876 saw	0.0015876 stars	0.0015876 with	0.0015876 ears

Expectation für NP:

$$E(NP) = \frac{\sum_{p=1}^{m} \sum_{q=p}^{m} \alpha_N(p,q)\beta_N(p,q)}{\text{Satzwahrscheinlichkeit}}$$

$$E(NP \to X\ Y) = \frac{\sum_{p=1}^{m} \sum_{q=p}^{m} \sum_{d=p}^{q-1} \alpha_N(p,q)P(N \to X\ \ Y)\beta_X(p,d)\beta_Y(d+1,q)}{\text{Satzwahrscheinlichkeit}}$$

$$E(NP \to w) = \frac{\sum_{h=1}^{m} \alpha_N(h,h)P(w=w_h)\beta_N(h,h)}{\text{Satzwahrscheinlichkeit}}$$

Regelwahrscheinlichkeiten

S	$\to$	NP VP	1.0
VP	$\to$	V NP	0.7
VP	$\to$	VP PP	0.3
NP	$\to$	NP PP	0.4
PP	$\to$	P NP	1.0
NP	$\to$	astronomers	0.1
NP	$\to$	telescopes	0.1
NP	$\to$	saw	0.04
NP	$\to$	stars	0.18
NP	$\to$	ears	0.18
P	$\to$	with	1.0
V	$\to$	saw	1.0

Inside Wahrscheinlichkeiten

	1	2	3	4	5
1	$\beta_{NP} = 0.1$		$\beta_S = 0.0126$		$\beta_S = 0.0015876$
2		$\beta_V = 1.0$ $\beta_{NP} = 0.04$	$\beta_{VP} = 0.126$		$\beta_{VP} = 0.015876$
3			$\beta_{NP} = 0.18$		$\beta_{NP} = 0.01296$
4				$\beta_P = 1.0$	$\beta_{PP} = 0.18$
5					$\beta_{NP} = 0.18$
	astronomers	saw	stars	with	ears

Outside Wahrscheinlichkeiten

	1	2	3	4	5
1	$\alpha_{NP} = 0.015876$		$\alpha_S = 0$		$\alpha_S = 1$
2		$\alpha_V = 0.0015876$ $\alpha_{NP} = 0$	$\alpha_{VP} = 0.0054$		$\alpha_{VP} = 0.1$
3			$\alpha_{NP} = 0.00882$		$\alpha_{NP} = 0.07$
4				$\alpha_P = 0.0015876$	$\alpha_{PP} = 0.00882$
5					$\alpha_{NP} = 0.00882$

	0.0015876 astronomers	0.0015876 saw	0.0015876 stars	0.0015876 with	0.0015876 ears

Expectation für NP:

$$E(NP) = \frac{\sum_{p=1}^{m} \sum_{q=p}^{m} \alpha_N(p,q)\beta_N(p,q)}{\text{Satzwahrscheinlichkeit}}$$

$$E(NP \to X\ Y) = \frac{\sum_{p=1}^{m} \sum_{q=p}^{m} \sum_{d=p}^{q-1} \alpha_N(p,q)P(N \to X\ \ Y)\beta_X(p,d)\beta_Y(d+1,q)}{\text{Satzwahrscheinlichkeit}}$$

$$E(NP \to w) = \frac{\sum_{h=1}^{m} \alpha_N(h,h)P(w=w_h)\beta_N(h,h)}{\text{Satzwahrscheinlichkeit}}$$

Regelwahrscheinlichkeiten

S	→ NP VP	1.0
VP	→ V NP	0.7
VP	→ VP PP	0.3
NP	→ NP PP	0.4
PP	→ P NP	1.0
NP	→ astronomers	0.1
NP	→ telescopes	0.1
NP	→ saw	0.04
NP	→ stars	0.18
NP	→ ears	0.18
P	→ with	1.0
V	→ saw	1.0

Inside Wahrscheinlichkeiten

	1	2	3	4	5
1	$\beta_{NP} = 0.1$		$\beta_S = 0.0126$		$\beta_S = 0.0015876$
2		$\beta_V = 1.0$ $\beta_{NP} = 0.04$	$\beta_{VP} = 0.126$		$\beta_{VP} = 0.015876$
3			$\beta_{NP} = 0.18$		$\beta_{NP} = 0.01296$
4				$\beta_P = 1.0$	$\beta_{PP} = 0.18$
5					$\beta_{NP} = 0.18$
	astronomers	saw	stars	with	ears

Outside Wahrscheinlichkeiten

	1	2	3	4	5
1	$\alpha_{NP} = 0.015876$		$\alpha_S = 0$		$\alpha_S = 1$
2		$\alpha_V = 0.0015876$ $\alpha_{NP} = 0$	$\alpha_{VP} = 0.0054$		$\alpha_{VP} = 0.1$
3			$\alpha_{NP} = 0.00882$		$\alpha_{NP} = 0.07$
4				$\alpha_P = 0.0015876$	$\alpha_{PP} = 0.00882$
5					$\alpha_{NP} = 0.00882$
	0.0015876 astronomers	0.0015876 saw	0.0015876 stars	0.0015876 with	0.0015876 ears

Maximization für NP:

$$\hat{P}(NP \to X\ Y) = \frac{E(N \to X\ Y)}{E(N)}$$
$$\hat{P}(NP \to w) = \frac{E(N \to w)}{E(N)}$$

Regelwahrscheinlichkeiten

S	$\to$	NP VP	1.0
VP	$\to$	V NP	0.7
VP	$\to$	VP PP	0.3
NP	$\to$	NP PP	0.4
PP	$\to$	P NP	1.0
NP	$\to$	astronomers	0.1
NP	$\to$	telescopes	0.1
NP	$\to$	saw	0.04
NP	$\to$	stars	0.18
NP	$\to$	ears	0.18
P	$\to$	with	1.0
V	$\to$	saw	1.0

Inside Wahrscheinlichkeiten

	1	2	3	4	5
1	$\beta_{NP} = 0.1$		$\beta_S = 0.0126$		$\beta_S = 0.0015876$
2		$\beta_V = 1.0$ $\beta_{NP} = 0.04$	$\beta_{VP} = 0.126$		$\beta_{VP} = 0.015876$
3			$\beta_{NP} = 0.18$		$\beta_{NP} = 0.01296$
4				$\beta_P = 1.0$	$\beta_{PP} = 0.18$
5					$\beta_{NP} = 0.18$
	astronomers	saw	stars	with	ears

Outside Wahrscheinlichkeiten

	1	2	3	4	5
1	$\alpha_{NP} = 0.015876$		$\alpha_S = 0$		$\alpha_S = 1$
2		$\alpha_V = 0.0015876$ $\alpha_{NP} = 0$	$\alpha_{VP} = 0.0054$		$\alpha_{VP} = 0.1$
3			$\alpha_{NP} = 0.00882$		$\alpha_{NP} = 0.07$
4				$\alpha_P = 0.0015876$	$\alpha_{PP} = 0.00882$
5					$\alpha_{NP} = 0.00882$
	0.0015876 astronomers	0.0015876 saw	0.0015876 stars	0.0015876 with	0.0015876 ears

Table of Contents

Wiederholung: Annahmen bei PCFGs

PCFGs haben folgende zugrundeliegend sind folgende Annahmen:

- **Positionsunabhängigkeit**: Die Wahrscheinlichkeit eines Teilbaums ist unabhängig davon, wo im Satz die entsprechende Wortfolge vorkommt (vgl. Zeitunabhängigkeit bei HMMs)
- **Kontextunabhängigkeit**: Die Wahrscheinlichkeit eines Teilbaums ist unabhängig von Wörtern, die er nicht dominiert
- **Vorfahrenunabhängigkeit**: Die Wahrscheinlichkeit eines Teilbaums ist unabhängig von Vorgängerknoten im Baum

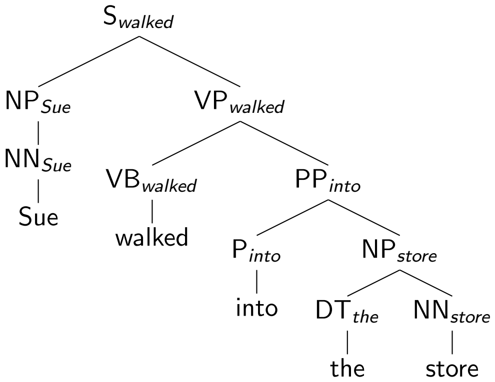
Implikationen der Unabhängigkeitsannahmen

- Vorteil: weniger Datenspärlichkeit durch Unabhängigkeitssannahmen.
- Nachteil: Wahrscheinlichkeiten rein strukturell, daher manchmal unintuitiv – Probleme mit PCFGs:
    - Kontext spielt keine Rolle, aber wir wissen, dass Personalpronomen in Subjektposition häufiger als in Objektposition sind
      (*"NP $\rightarrow$ Pronoun"* müsste zwei unterschiedliche, kontextabhängige Wahrscheinlichkeiten haben)
    - einfache PCFGs modellieren keine Subkategorisierung oder Selektionseinschränkungen.
      Beispiel: ob keine, eine oder zwei NPs auf ein Verb folgen, ist bei einer einfachen PCFG unabhängig vom Verb!
    - Globale strukturelle Präferenzen haben keine Auswirkung (z.B. in Bezug auf die Anbindung von PPs, Relativsätze, Adverbien, usw.)

Table of Contents

Mögliche Lösungen

- **Lexikalisierung** der Nicht-Terminalen



- **Parentisierung**: Ähnliches Prozess, nur hier wird ein bestimmter Vorgängerknoten zur Vorbedingung einer Regel gemacht
- Weitere "Tricks": Subkategorisierung einführen, Traces, Interpunktion, Clustering, usw.

Eine Grammatik lexikalisieren

Lexikalisierung der erste VP-Regel, VP → V NP (0.7)

$$
\begin{array}{lcll}
VP_{saw} & \rightarrow & V_{saw}\ NP_{astronomers} & 0.1 \\
VP_{saw} & \rightarrow & V_{saw}\ NP_{ears} & 0.15 \\
VP_{saw} & \rightarrow & V_{saw}\ NP_{saw} & 0.05 \\
VP_{saw} & \rightarrow & V_{saw}\ NP_{stars} & 0.3 \\
VP_{saw} & \rightarrow & V_{saw}\ NP_{telescopes} & 0.1
\end{array}
$$

- Jetzt können wir abschätzen, ob “Sterne” oder “Sägen” bessere Objekte von “sehen” sind.
- Aber: aus einer Regel werden fünf! Gilt für alle Regeln mit Nichtterminalen ($S_{saw} \rightarrow NP_{astronomers}\ VP_{saw}$).
- Extrem viele Parameter müssen geschätzt werden, besonders bei realistischeren Lexika und Regelmengen (Sparse-Data)
- Typischerweise wird nur den lexikalischen Kopf als Bedingung genommen: $VP_{saw} \rightarrow V_{saw}$ NP 0.7

Struktureller Kontext

Strukturelle Unabhängigkeitsannahme

PCFGs nehmen an, dass eine Kategorie mit gleichen Wahrscheinlichkeiten zu anderen Kategorien expandiert, egal, wo in der Satzstruktur sie sich befindet (z.B. hat die Regel NP→NP PP nur eine Wahrscheinlichkeit).

Echte Daten

Regel	als Subj	als Obj
NP → PRP	13.7%	2.1%
NP → DT NN	5.6%	4.6%
NP → NP PP	5.6%	14.1%

- Um diese Verteilung zu erfassen, müssen wir Kontexte finden, auf denen wir konditionieren können, um entsprechende separate Wahrscheinlichkeiten zu testen.

Beispiel für Features auf denen konditioniert wird

(Top-down PCFG parser, Roark, 2001)

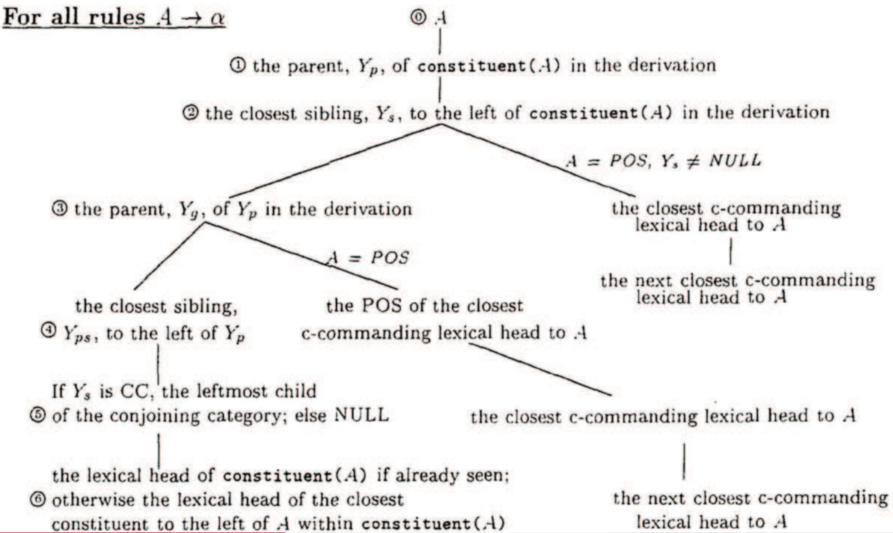


Table of Contents

Evaluation von Parsern

Standard: PARSEVAL

Wir wollen die Ausgabe eines Parsers mit einem "Gold Standard" (z.B. annotierte Baumbank) vergleichen.

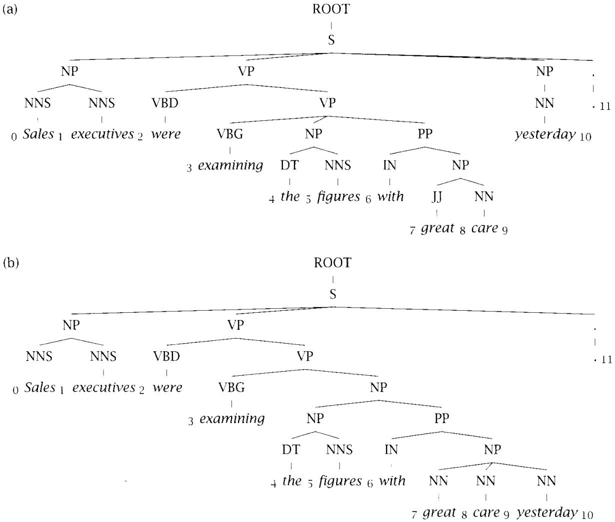
- Eine Konstituente ist **korrekt**, wenn es in der Baumbank für den Satz eine Konstituente mit gleichem Startpunkt, Endpunkt und Nichtterminalsymbol gibt.
- **labeled recall**:

$$\frac{\text{Anzahl korrekter Konstituenten im Parse}}{\text{Anzahl korrekter Konstituenten in der Baumbank}}$$

- **labeled precision**:

$$\frac{\text{Anzahl korrekter Konstituenten im Parse}}{\text{Gesamtanzahl aller Konstituenten im Parse}}$$

- **crossed brackets**: Wie viele Konstituenten haben die überkreuzt? z.B. ((A B) C) anstelle von (A (B C))



(c) Brackets in gold standard tree (a.):
S-(0:11), **NP-(0:2)**, VP-(2:9), VP-(3:9), **NP-(4:6)**, PP-(6-9), NP-(7,9), *NP-(9:10)

(d) Brackets in candidate parse (b.):
S-(0:11), **NP-(0:2)**, VP-(2:10), VP-(3:10), NP-(4:10), **NP-(4:6)**, PP-(6-10), NP-(7,10)

(e)

Precision:	3/8 = 37.5%	Crossing Brackets:	0
Recall:	3/8 = 37.5%	Crossing Accuracy:	100%
Labeled Precision:	3/8 = 37.5%	Tagging Accuracy:	10/11 = 90.9%
Labeled Recall:	3/8 = 37.5%		

Vergleich einfache PCFG vs. PCFG mit mehr Kontext

Vergleich:

Parser	Labeled Recall	Labeled Precision
Standard PCFG	71.7	75.8
Lexicalized PCFG	83.4	84.1
Charniak (2000)	91.1	90.1

Table of Contents

Ambiguität

- PCFGs helfen bei Ambiguität, da sie verschiedene Analysen nach der Wahrscheinlichkeit ranken können.
- Trotzdem gibt es für die meisten Sätze so viele Analysen, dass es sogar mit dem Viterbi Algorithmus (o.ä.) lange dauern würde, sie alle zu berechnen.
- **Beam Search**: wir verfolgen nur die besten Analysen
- Konsequenzen: Parser wird schneller, aber es kann passieren, dass wir die im Endeffekt beste Analyse nicht finden.
- mögliche Definition des Beams:
  - Festgelegte Größe (z.B. 1000 Analysen)
  - Relativ zur Wahrscheinlichkeit der besten Analyse (Analysen mit Wahrscheinlichkeit kleiner 10^{-4} als die beste Analyse werden nicht weiterverfolgt.)

Zusammenfassung

- Unabhängigkeitsannahmen werden der Wirklichkeit nicht gerecht
- Um PCFGs zu besseren Modellen zu machen kann man die Unabhängigkeitsannahmen aufweichen
  - Regeln lexikalisieren
  - strukturelle Information hinzufügen durch Konditionierung auf andere Knoten im Baum
- Evaluation von Parsern: Parseval, Labelled Precision, Labelled Recall
- Um effizienter zu Parsen, können wir Beam Search benutzen und nur die wahrscheinlichsten Analysen verfolgen.