

Aufgabe 5) Geben Sie an, wie bei einer probabilistischen kontextfreien Grammatik folgende Werte definiert sind:

?

- die Wahrscheinlichkeit eines Parsebaumes
- die Wahrscheinlichkeit eines Satzes (= Menge von Parsebäumen)
- die Wahrscheinlichkeit eines Korpus (= Folge von Sätzen) (2 Punkte)

Aufgabe 5) Geben Sie an, wie bei einer probabilistischen kontextfreien Grammatik folgende Werte definiert sind:

?

- die Wahrscheinlichkeit eines Parsebaumes
- die Wahrscheinlichkeit eines Satzes (= Menge von Parsebäumen)
- die Wahrscheinlichkeit eines Korpus (= Folge von Sätzen) (2 Punkte)

Korpuswahrscheinlichkeit $p(C) = \prod_{s \in C} p(s)$

Satzwahrscheinlichkeit $p(s) = \sum_{t \in T(s)} p(t)$

Parsewahrscheinlichkeit $p(t) = p(r_1, \dots, r_n) = \prod_{i=1}^n p(r_i)$

$T(s)$ sind die Analysen des Satzes s .

r_1, \dots, r_n sind die Regeln der Linksableitung von t .

PCFG-Zusammenfassung

- PCFG = CFG + Regelwahrscheinlichkeiten
- Die Wahrscheinlichkeiten aller Regeln mit derselben linken Seite summieren zu 1, d.h.

$$\sum_{\alpha: A \rightarrow \alpha \in P} p(A \rightarrow \alpha) = 1 \text{ für alle } A \in V$$

- Parsebaum-Wahrscheinlichkeit = Produkt der Regelwahrscheinlichkeiten
- syntaktische **Desambiguierung** durch Auswahl der wahrscheinlichsten Analyse
- Die Wahrscheinlichkeit eines Satzes S definieren wir als Summe der Wahrscheinlichkeiten all seiner Parsebäume.

Aufgabe 8) Wie wird bei PCFGs die Wahrscheinlichkeit eines Parsebaumes, die Wahrscheinlichkeit eines Satzes und die Wahrscheinlichkeit einer Folge von Sätzen (=Korpus) definiert? (3 Punkte)

Aufgabe 6) Wie ist die Wahrscheinlichkeit eines Parsebaumes bei einer probabilistischen kontextfreien Grammatik (PCFG) definiert? (2 Punkte)

Aufgabe 7) Erläutern Sie den EM-Algorithmus am Beispiel des unüberwachten Trainings von PCFGs (also Training auf Roh⁰texten). Welche Daten benötigen Sie? Welche Berechnungsschritte führt der EM-Algorithmus aus? (4 Punkte)

Aufgabe 9) Wie berechnet der Viterbi-Algorithmus den besten Parse für einen gegebenen Parsewald? Wie werden die Viterbi-Wahrscheinlichkeiten berechnet (mit Formeln)?
(5 Punkte)

Aufgabe 10) Wie arbeitet der Inside-Outside-Algorithmus und wie trainiert man damit eine PCFG? (4 Punkte)

Aufgabe 11) Wie wird eine Grammatik **markowisiert**. Was ist der Vorteil der Markowisierung?
(2 Punkte)

Aufgabe 5) Warum wird beim PCFG-Parsen die Grammatik oft markowisiert? Welches Problem soll dadurch gelöst werden? (3 Punkte)

Aufgabe 9) Erklären Sie die Grundidee der Berkeley-Parsers. Mit welcher Methode wird er trainiert? (2 Punkte)

Aufgabe 9) Erklären Sie die Grundidee des Berkeley-Parsers von Petrov und Klein. Wie wird der Parser trainiert? (3 Punkte)

Aufgabe 5) Erklären Sie die Grundidee des Berkeley-Parsers von Petrov und Klein. Wie wird der Parser trainiert? (3 Punkte)

Aufgabe 6) Welche Maße werden üblicherweise verwendet bei der Evaluierung von

- Sprachmodellen
- Wortart-Taggern
- Parsern

(2 Punkte)

Aufgabe 10) Wie wird üblicherweise die Genauigkeit von (Konstituenten-)Parsern gemessen? Erklären Sie die Methode und wie das Maß berechnet wird. (3 Punkte)

Aufgabe 6) Wie wird der F-Score zur Messung der Genauigkeit von Parseern berechnet?
(3 Punkte)

7

Aufgabe 3) Geben Sie die Formeln zur Berechnung des besten Parsebaumes aus einer Menge von Parsebäumen in Parsewaldrepräsentation an. Wie wird initialisiert? Wie wird der beste Parse am Ende ausgegeben?

Sie können davon ausgehen, dass w_1, \dots, w_n die Wörter des Eingabesatzes sind, N die Menge der nichtterminalen Knoten im Parsewald und P die Menge der Parsewaldregeln. Die Funktion $p(r)$ liefert die Wahrscheinlichkeit der PCFG-Regel, welche der Parsewaldregel $r \in P$ entspricht. (5 Punkte)

Aufgabe 4) Beim unüberwachten Training (Training ohne annotierte Daten) eines HMM-Taggers mit dem EM-Algorithmus wird der Forward-Backward-Algorithmus verwendet.

- Erklären Sie wozu der Forward-Backward-Algorithmus hier dient. Wird er im E-Schritt oder im M-Schritt benutzt?
- Begründen Sie, warum der Viterbi-Algorithmus hier nicht verwendet werden kann.
- Bei welchem Modell wird der Forward-Backward-Algorithmus beim *überwachten* Training eingesetzt? (Es handelt sich dabei nicht um EM-Training.)
- Welcher Algorithmus entspricht dem Forward-Backward-Algorithmus beim unüberwachten Training von PCFGs?
- Welcher Parser setzt diesen Algorithmus beim überwachten Training ein?