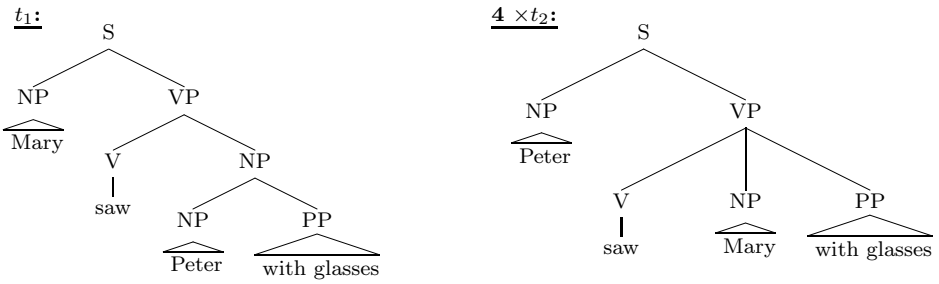
Einführung in die statistische Sprachverarbeitung

Detlef Prescher

May 29, 2007

Hausaufgabe 8

Es sei die folgende einfache Baumbank gegeben:



Ferner sei:

- $f_{\texttt{TB}}(x)$ die Vorkommenshäufigkeit eines Baumes x in der Baumbank
- $\tilde{p}(x)$ die empirische Wahrscheilichkeit eines Baumes x in der Baumbank
- G die von der Baumbank abgelesene kontextfreie Grammatik
- $\mathcal{X}$ die Menge aller Bäume der Grammatik G
- $\mathcal{M}_G$ das Wahrscheinlichkeitsmodell der Grammatik G
- $f(r)$ die Vorkommenshäufigkeit einer Regel r in der Baumbank
- $f_r(x)$ die Vorkommenshäufigkeit einer Regel r im Baum x

(i) Zeige, dass $\tilde{p} \notin \mathcal{M}_G$. (Nimm an, dass $\tilde{p} \in \mathcal{M}_G$ und zeige, dass dies zu einem Widerspruch führt. Tip: Zeige zunächst, dass aus der Annahme folgt, dass alle Regelwahrscheinlichkeiten positiv sein müssen. Betrachte dann einen Baum von G, der NICHT in der Baumbank ist...)

(ii) Ist $\tilde{p}$ ein MLE von $\mathcal{M}_G$ auf $f_{\texttt{TB}}$? Begründe Deine Antwort.

(iii) Berechne für die gegebene Baumbank alle Werte von $f(r), f_{\texttt{TB}}(x)$ und $f_r(x)$. Zeige, dass für alle Regeln gilt, dass:

$$f(r) = \sum_{x \in \mathcal{X}} f_{\texttt{TB}}(x) \cdot f_r(x)$$

(iv) Seien f_A die Korpora aller Regeln mit jeweils der linken Seite A. Wähle eine beliebige Modellinstanz $p \in \mathcal{M}_G$ aus, und zeige, dass:

$$L(f_{\texttt{TB}}, p) = \prod_A L(f_A, p)$$

Diskutiere, ob dies für jede Instanz $p \in \mathcal{M}_G$ so sein wird.

(v) Beschreibe nun in Deinen eigenen Worten, warum Charniaks Baumbankgrammatik das eindeutige(!) MLE von $\mathcal{M}_G$ auf f_{TB} sein wird.

Abgabetermin: Montag, 11. Juni

Besprechung: Mittwoch, 13. Juni