

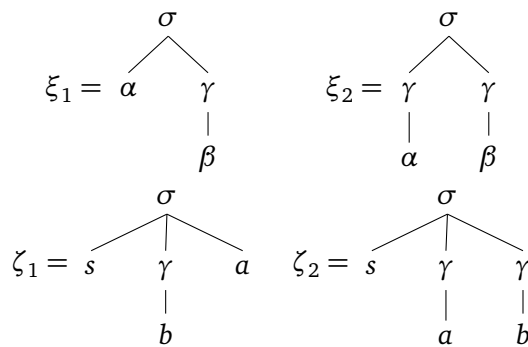
Maschinelles Übersetzen natürlicher Sprachen

Aufgabe 1 (Yamada-Knight Training)

Sei $\Sigma = \{\sigma, \gamma\}$ ein Alphabet. Wir nehmen einen bilingualen Baumkorpus $\mathcal{D} \subseteq U_{\Sigma}(V_E) \times U_{\Sigma}(V_F)$ zwischen zwei künstlichen Sprachen E und F an, deren Vokabular aus $V_E = \{\alpha, \beta\}$ bzw. aus $V_F = \{s, a, b\}$ bestehen soll.¹ Der Korpus ist recht klein, er setzt sich aus den zwei Baumpaaren

$$\mathcal{D} = \{(\xi_1, \zeta_1), (\xi_2, \zeta_2)\}$$

zusammen, dabei sind die Bäume von der folgenden Form:



Nutzen Sie diesen Korpus zum Training der Parameter eines Yamada-Knight-Übersetzungsmodells zwischen E und F nach dem in der Vorlesung vorgestellten Algorithmus! Wenden Sie dabei sinnvolle Vereinfachungen an, um die Anzahl der Parameter überschaubar zu halten.

Anmerkung: Es bietet sich an, die iterative Berechnungsvorschrift des Trainingsalgorithmus erst mathematisch auf das konkrete Beispiel zu instanzieren. Das wiederholte Anwenden dieser vereinfachten Berechnungsvorschrift lässt sich dann in einem kurzen Programm implementieren.

¹Dabei bezeichnet $U_{\Sigma}(A)$, gegeben ein Alphabet Σ und eine Menge A , die Menge der ranglosen Bäume über Σ mit Blättern aus A . Formal definieren wir $U_{\Sigma}(A)$ als die kleinste Obermenge U von A , so dass für alle $\sigma \in \Sigma$, $k \in \mathbb{N}$ und $\xi_1, \dots, \xi_k \in U$ auch $\sigma(\xi_1, \dots, \xi_k) \in U$ gilt. Ranglose Bäume lassen sich genau so wie Rangbäume graphisch darstellen, vergleiche oben.