

**Schriftliche Prüfung zur Übung**  
**Statistische Methoden in der maschinellen Sprachverarbeitung**  
**WS 2020/21**  
**Dozent: Helmut Schmid**

Sie haben **60 Minuten** Zeit plus 5 Minuten zum Absenden Ihrer Lösungen per Email.

Wenn Sie einen Fehler in einer der Aufgaben entdecken sollten, dann schicken Sie mir bitte sofort eine Nachricht an `schmid@cis.lmu.de` oder melden Sie sich per Zoom.

---

In dieser Klausur geht es um die Implementierung eines **LC-CRF-Taggers**. Dabei können die Musterlösungen zu der Perzeptron-Aufgabe und zu der Testklausur Übung (die in der letzten Übungsstunde behandelt wurde) nützlich sein.

**Aufgabe 1)** Schreiben Sie eine Methode **localScore**(self, prevtag, tag, words, i), welche die Methode `self.featureVec(prevtag, tag, words, i)` aufruft und das Produkt  $s(t', t, w_1^n, i)$  aus Merkmalsvektor und Gewichtsvektor berechnet und zum Exponenten nimmt:

$$s(t', t, w_1^n, i) = e^{\theta \cdot f(t', t, w_1^n, i)} = e^{\sum_k \theta_k \cdot f_k(t', t, w_1^n, i)}$$

Die Methode `self.featureVec`(prevtag, tag, words, i) müssen Sie nicht schreiben. Sie liefert den Merkmalsvektor als eine Liste von Merkmal-Wert-Paaren (wobei die Werte nicht alle 1 sein müssen). Das Dictionary `self.weight` ist ebenfalls gegeben und liefert das Gewicht eines Merkmals. Die Merkmale sind Strings. (4 Punkte)

**Aufgabe 2)** Schreiben Sie eine Methode **forward**(self, words), welche die Forward-Werte für die Wortfolge *words* nach der folgenden Formel berechnet:

$$\begin{aligned}\alpha_{\langle s \rangle}(0) &= 1 \\ \alpha_t(i) &= \sum_{t'} \alpha_{t'}(i-1) s(t', t, w_1^n, i)\end{aligned}$$

(7 Punkte)

**Aufgabe 3)** Schreiben Sie eine Methode **backward**(self, words), welche die Backward-Werte für die Wortfolge *words* nach der folgenden Formel berechnet:

$$\begin{aligned}\beta_{\langle s \rangle}(n+1) &= 1 \\ \beta_t(i-1) &= \sum_{t'} \beta_{t'}(i) s(t, t', w_1^n, i)\end{aligned}$$

(7 Punkte)

**Aufgabe 4)** Schreiben Sie eine Methode **EStep**(self, words, estFeatValue), welche erst die forward- und die backward-Methode aufruft und dann die Aposteriori-Wahrscheinlichkeiten aller Tagpaare (t, t') für alle Positionen i nach der folgenden Formel berechnet:

$$\gamma_{tt'}(i) = \frac{\alpha_t(i-1) s(t, t', w_1^n, i) \beta_{t'}(i)}{\alpha_{\langle s \rangle}(n+1)}$$

Gewichten Sie jeden möglichen Merkmalsvektor  $f(t, t', w_1^n, i)$  mit  $\gamma_{tt'}(i)$  und summieren Sie die gewichteten Merkmalswerte für jedes Merkmal über alle Positionen in dem Dictionary *estFeatValue*. (12 Punkte)

(30 Punkte insgesamt)

Viel Erfolg!