Einführung in die Computerlinguistik
Hausaufgabe 5 (HMM-POS-Tagging), Abgabe 01.06.2015

Laura Kallmeyer

Sommersemester 2015, Heinrich-Heine-Universität Düsseldorf

Aufgabe 1 *Angenommen, Sie haben folgendes getaggtes Korpus:*

```
Satz 1: ('a', 'X') ('a', 'Y') ('b', 'Y')  ('.', 'P')
Satz 2: ('b', 'Y') ('a', 'Y') ('.', 'P')
Satz 3: ('a', 'X') ('b', 'X') ('.', 'P')
```

(`a`*,* `b` *und* `.` *sind die Token, die getaggt werden sollen, und* `X`*,* `Y` *und* `P` *sind die möglichen POS-Tags.)*

1. *Geben Sie das HMM an, das sich aus diesem Korpus durch Abzählen von Auftreten, also als MLE ergibt. Geben Sie das vollständige Tupel $\langle Q, A, O, B, q_0, q_F\rangle$ an.*
2. *Was ergibt sich mit diesem Modell als beste Tagsequenz für den Satz "*`a .`*"?*

Aufgabe 2 *Nehmen Sie an, Sie haben einen HMM-POS Tagger, mit dem folgende Eingabe getaggt werden soll:* start to study*. Dem Tagger liegen folgende Wahrscheinlichkeiten zugrunde:*

Emissionswahrscheinlichkeiten:

$P(start|V) = 2 \cdot 10^{-3}$ $P(study|V) = 2 \cdot 10^{-3}$ $P(to|P) = 1$
$P(start|N) = 3 \cdot 10^{-3}$ $P(study|N) = 3 \cdot 10^{-3}$

Alle anderen Emissionswahrscheinlichkeiten für unsere Eingabewörter seien 0*.*

Relevante Übergangswahrscheinlichkeiten:

$P(N|P) = 5 \cdot 10^{-1}$ $P(P|V) = 2 \cdot 10^{-1}$ $P(V|V) = 2 \cdot 10^{-1}$ $P(N|N) = 1 \cdot 10^{-1}$
$P(V|P) = 1 \cdot 10^{-1}$ $P(P|N) = 1 \cdot 10^{-1}$ $P(N|V) = 2 \cdot 10^{-1}$ $P(V|N) = 3 \cdot 10^{-1}$

Angenommen, die Wahrscheinlichkeit, dass ein N am Satzanfang steht ist $1 \cdot 10^{-1}$*, die, dass ein V am Satzanfang steht* $2 \cdot 10^{-1}$*. Die, dass ein Satzende auf N oder V folgt, ist jeweils* $1 \cdot 10^{-1}$*.*

1. *Geben Sie die Viterbi Matrix an, die sich bei diesen Wahrscheinlichkeiten für die Eingabe* start to study *ergibt. Es reicht, die Einträge anzugeben, die* $\neq 0$ *sind.*
2. *Was ist die beste POS Tag Sequenz, die sich aufgrund dieser Matrix für die Eingabe ergibt?*