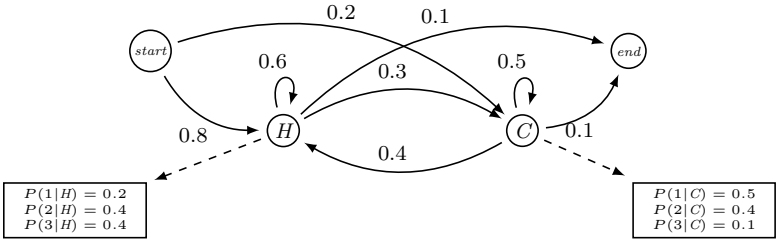
Einführung in die Computerlinguistik
Hausaufgabe 5, Abgabe 25.05.2020

Laura Kallmeyer

SoSe 2020, Heinrich-Heine-Universität Düsseldorf

Aufgabe 1 *Consider the following HMM (the "Jason's ice cream" HMM from the course slides)*



1. *What is the probability of traversing a path*

   $$start\ C\ C\ end$$

   according to this HMM?

2. *What is the most probable output when traversing*

   $$start\ C\ H\ end$$

   according to this HMM?

3. *What is the probability of an output* 3 3 *given this model?*

4. *Given that we see an output* 3 3*, what is the most probable path that has been traversed? (You don't need the viterbi algorithm here, it is enough to compare the different path probabilities for the entire paths.)*

Solution:

1. This is independent from the output, only transition probabilities matter: $0.2 \cdot 0.5 \cdot 0.1 = 10 \cdot 10^{-3} = 10^{-2}$

2. There are two sequences that are equally probably, 1 2 and 1 3.

3. We have to sum over the joint probabilities of path and output for all possible paths and the output 3 3:

(path *start C C end*)	$10^{-3} \cdot 0.1 \cdot 0.1$
(path *start C H end*)	$+0.2 \cdot 0.4 \cdot 0.1 \cdot 0.1 \cdot 0.4$
(path *start H C end*)	$+0.8 \cdot 0.3 \cdot 0.1 \cdot 0.4 \cdot 0.1$
(path *start H H end*)	$+0.8 \cdot 0.6 \cdot 0.1 \cdot 0.4 \cdot 0.4$
	$= (10 + 32 + 96 + 768)10^{-5} = 906 \cdot 10^{-5} = 0.00906$

4. Here we need the maximum over the joint probabilities of path and output for an output 3 3, i.e., the maximum of the four elements of the sum we calculated in 3. This is 0.00768, therefore the most probable path is the path *start H H end*.

Aufgabe 2 *Angenommen, Sie haben folgendes getaggtes Korpus (je ein Wort gefolgt von seinem POS Tag, in einer Klammer):*[1]

[1] Adaptiert von: RRG Baumbank, basierend auf George Orwell's "1984", https://rrgparbank.phil.hhu.de/

```
Satz 1: (Winston N) (made V) (for P) (the OP) (stairs N) (. .)
Satz 2: (it PRO) (was AUX) (no OP) (use N) (trying V) (the OP) (lift N) (. .)
Satz 3: (the OP) (actual A) (writing N) (would OP) (be AUX) (easy A) (. .)
```

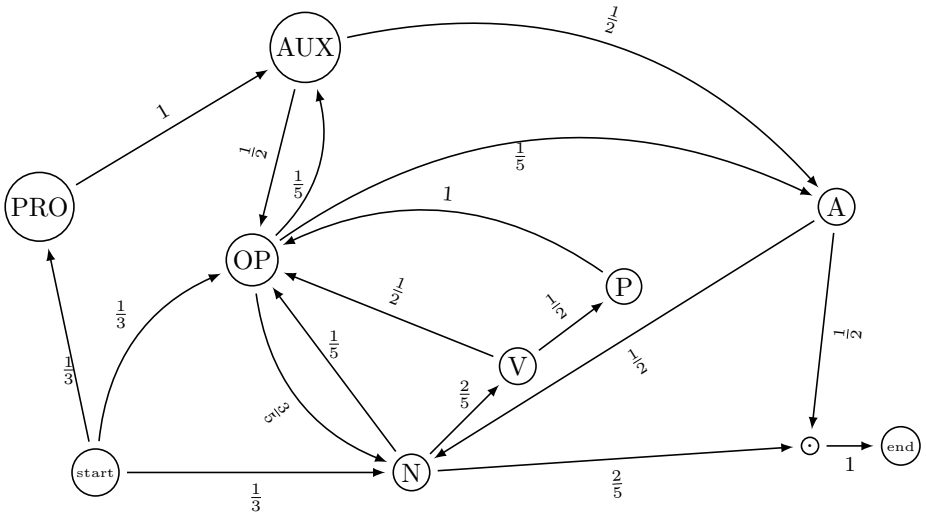
(`Winston`, `made`*, etc. sind die Token, die getaggt werden sollen, und* `N`, `V` *etc. sind die möglichen POS-Tags.)*

1. *Geben Sie das die Übergangswahrscheinlichkeiten des HMM an, das sich aus diesem Korpus durch Abzählen von Auftreten, also als MLE ergibt. Es reicht, wenn Sie den entsprechenden Graphen zeichnen.*

   Geben Sie außerdem die Emissionswahrscheinlichkeiten für die Zustände OP und N an. Listen Sie nur die Wahrscheinlichkeiten, die nicht 0 sind.

2. *Was ergibt sich mit diesem Modell als beste Tagsequenz für den Satz "*`The stairs .`*"?*

Lösung

1. 

   Emissionswahrscheinlichkeiten für OP: $P(the\,|\,OP) = \frac{1}{2}$, $P(no\,|\,OP) = \frac{1}{4}$, $P(would\,|\,OP) = \frac{1}{4}$.

   Emissionswahrscheinlichkeiten für N: $P(Winston\,|\,N) = \frac{1}{5}$, $P(stairs\,|\,N) = \frac{1}{5}$, $P(use\,|\,N) = \frac{1}{5}$, $P(lift\,|\,N) = \frac{1}{5}$, $P(writing\,|\,N) = \frac{1}{5}$.

2. Aufgrund der Emissionswahrscheinlichkeiten gibt es nur eine Tagfolge, für die sich ein Wert $\neq 0$ ergibt: `(the OP) (stairs N) (. .)`

Aufgabe 3 *Nehmen Sie an, Sie haben einen HMM-POS Tagger, unter anderem mit folgenden Wahrscheinlichkeiten:*

Emissionswahrscheinlichkeiten:

$P(the\|Det) = 1$	$P(bear\|N) = 3 \cdot 10^{-3}$	$P(duck\|N) = 3 \cdot 10^{-3}$
	$P(bear\|V) = 1 \cdot 10^{-3}$	$P(duck\|V) = 2 \cdot 10^{-3}$

Alle anderen Emissionswahrscheinlichkeiten für bear, the *und* duck *seien* 0.

Übergangswahrscheinlichkeiten:

$P(N\|Det) = 5 \cdot 10^{-1}$	$P(N\|N) = 1 \cdot 10^{-1}$	$P(N\|V) = 3 \cdot 10^{-1}$
$P(V\|Det) = 1 \cdot 10^{-1}$	$P(V\|N) = 4 \cdot 10^{-1}$	$P(V\|V) = 1 \cdot 10^{-1}$
	$P(Det\|N) = 1 \cdot 10^{-1}$	$P(Det\|V) = 1 \cdot 10^{-1}$

Angenommen, die Wahrscheinlichkeit, dass ein V am Satzanfang steht, ist $1 \cdot 10^{-1}$*, die, dass ein N am Satzanfang steht, ist* $4 \cdot 10^{-1}$ *und die, dass ein Det am Satzanfang steht, ist* $4 \cdot 10^{-1}$*.*

Die Wahrscheinlichkeit, dass auf ein N oder V ein Satzende folgt, ist jeweils $0.1 = 1 \cdot 10^{-1}$*.*

1. *Geben Sie die Viterbi Matrix an, die sich bei diesen Wahrscheinlichkeiten für die Eingabe* bear the duck *ergibt. Es reicht, die Einträge anzugeben, die* $\neq 0$ *sind. Geben Sie für jedes Feld Ihren Rechenweg an.*
2. *Was ist die beste POS-Tag Sequenz, die sich als Ergebnis für* bear the duck *ergibt?*

Lösung:

1.

q_F				$18 \cdot 10^{-9}$
N	$12 \cdot 10^{-4}$,q_0		$18 \cdot 10^{-8}$, Det	
V	$1 \cdot 10^{-4}$, q_0		$24 \cdot 10^{-9}$, Det	
Det		$12 \cdot 10^{-5}$, N		
	1 bear	2 the	3 duck	

bear, N: $P(\text{bear}|\text{N}) \cdot 4 \cdot 10^{-1} = 3 \cdot 10^{-3} \cdot 4 \cdot 10^{-1} = 12 \cdot 10^{-4}$

bear, V: $P(\text{bear}|\text{V}) \cdot 1 \cdot 10^{-1} = 1 \cdot 10^{-3} \cdot 1 \cdot 10^{-1} = 1 \cdot 10^{-4}$

the, Det: $\max\{12 \cdot 10^{-4} \cdot 1 \cdot 10^{-1} \cdot 1 \text{ Vorgänger N}, 1 \cdot 10^{-4} \cdot 1 \cdot 10^{-1} \cdot 1 \text{ Vorgänger V}\}$

duck, N: $12 \cdot 10^{-5} \cdot 5 \cdot 10^{-1} \cdot 3 \cdot 10^{-3} = 18 \cdot 10^{-8}$

duck, V: $12 \cdot 10^{-5} \cdot 1 \cdot 10^{-1} \cdot 2 \cdot 10^{-3} = 24 \cdot 10^{-9}$

q_F: $\max\{1 \cdot 10^{-1} \cdot 18 \cdot 10^{-8} \text{ Vorgänger N}, 1 \cdot 10^{-1} \cdot 24 \cdot 10^{-9} \text{ Vorgänger V}\}$

2. Die beste POS-Tag Folge ist demnach N Det N.