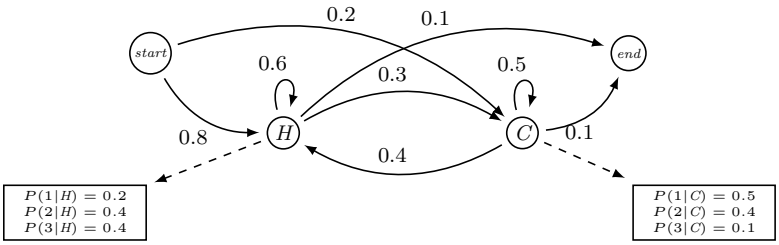
Einführung in die Computerlinguistik
Hausaufgabe 5, Abgabe 25.05.2021 (Dienstag), 35 Pkte

Laura Kallmeyer

SoSe 2021, 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*

   $$start\ C\ H\ C\ end$$

   according to this HMM?

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

   $$start\ C\ C\ end$$

   according to this HMM?

3. *What is the probability of an output* 2 2 *given this model?*

Solution:

1. This is independent from the output, only transition probabilities matter: $0.2 \cdot 0.4 \cdot 0.3 \cdot 0.1 = 24 \cdot 10^{-4} = 0,0024$

2. 1 1.

3. We have to sum over the joint probabilities of path and output for all possible paths and the output 2 2:

(path *start C C end*)	$10^{-2} \cdot 0.4 \cdot 0.4$
(path *start C H end*)	$+0.2 \cdot 0.4 \cdot 0.1 \cdot 0.4 \cdot 0.4$
(path *start H C end*)	$+0.8 \cdot 0.3 \cdot 0.1 \cdot 0.4 \cdot 0.4$
(path *start H H end*)	$+0.8 \cdot 0.6 \cdot 0.1 \cdot 0.4 \cdot 0.4$
	$= (160 + 128 + 384 + 768)10^{-5} = 1440 \cdot 10^{-5} = 0.0144$

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

```
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: (only ADV) (the OP) (Thought N) (Police N) (mattered V) (. .)
Satz 4: (Winston N) (kept V) (his OP) (back N) (turned V) (to P) (the OP) (telescreen N) (. .)
Satz 5: (he PRO) (sat V) (back ADV) (. .)
Satz 6: (we PRO) (use V) (the OP) (telescreen N) (already ADV) (. .)
```

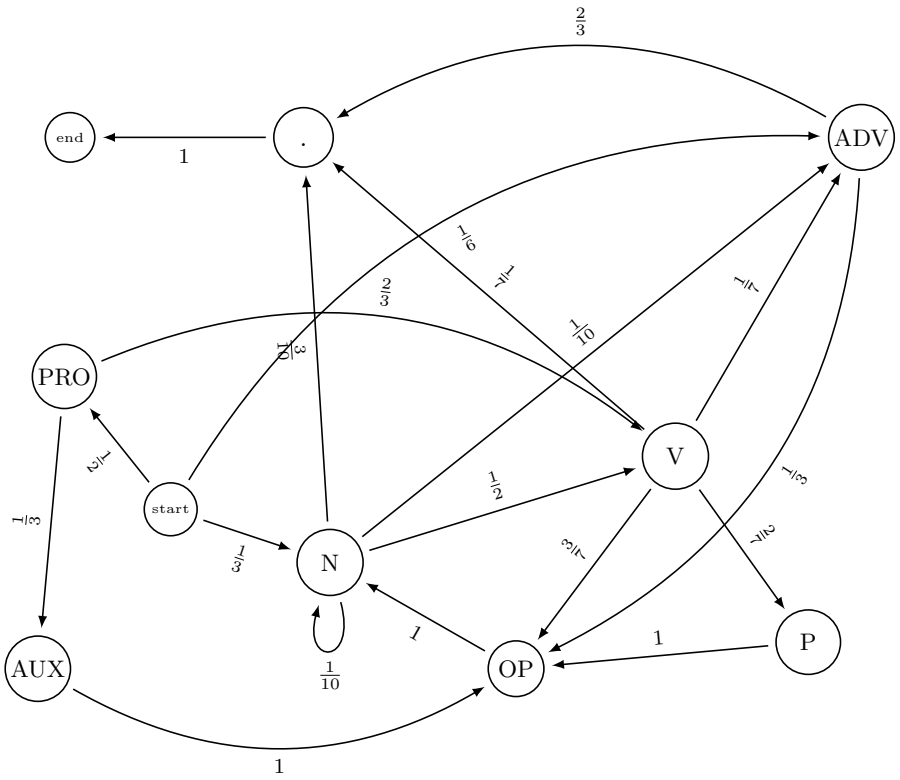
[1] Adaptiert von: RRG Baumbank, basierend auf George Orwell's "1984", https://rrgparbank.phil.hhu.de/

(`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.*
2. *Geben Sie außerdem alle Emissionswahrscheinlichkeiten an, die die Wörter "we", "use", "the" "back" und "." betreffen, und die nicht* 0 *sind.*
3. *Was ergibt sich mit diesem Modell als beste Tagsequenz für den Satz "*`we use the back .`*"? Zur Beantwortung dieser Frage berechnen Sie die Viterbi-Matrix, die sich mit dem Modell und dieser Eingabe ergibt. Es reicht, die Einträge in der Matrix anzugeben, die $\neq 0$ sind.*

Lösung

1.



2. $P(we \mid PRO) = \frac{1}{3}$,
   $P(use \mid V) = \frac{1}{7}$, $P(use \mid N) = \frac{1}{10}$,
   $P(the \mid OP) = \frac{5}{7}$,
   $P(back \mid N) = \frac{1}{10}$, $P(back \mid ADV) = \frac{1}{3}$,
   $P(. \mid .) = 1$.

3.

q_F						$\frac{3\cdot5\cdot3}{63\cdot7\cdot7\cdot10\cdot10}$, .
.					$\frac{3\cdot5\cdot3}{63\cdot7\cdot7\cdot10\cdot10}$, N	
N				$\frac{3\cdot5}{63\cdot7\cdot7\cdot10}$, OP		
V		$\frac{1}{63}$, PRO				
OP			$\frac{3\cdot5}{63\cdot7\cdot7}$, V			
PRO	$\frac{1}{6}$, start					
	we	use	the	back	.	

Beste POS Tag Folge: PRO V OP N .

(Es gibt nur einen Pfad, trotz der Ambiguität der POS Tags für use und back, da die Übergangswahrscheinlichkeiten bei den alternativen Tags zum Teil 0 sind.)

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

Emissionswahrscheinlichkeiten:

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

Alle anderen Emissionswahrscheinlichkeiten für use, the *und* light *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* $2 \cdot 10^{-1}$ *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* use the light *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* use the light *ergibt?*

Lösung:

1.

q_F				$9 \cdot 10^{-9}$, N
N	$6 \cdot 10^{-4}$,q_0		$9 \cdot 10^{-8}$, Det	
V	$1 \cdot 10^{-4}$, q_0		$12 \cdot 10^{-9}$, Det	
Det		$6 \cdot 10^{-5}$, N		
	1 use	2 the	3 light	

use, N: $P(\text{use}|\text{N}) \cdot 2 \cdot 10^{-1} = 3 \cdot 10^{-3} \cdot 2 \cdot 10^{-1} = 6 \cdot 10^{-4}$

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

the, Det: $\max\{6 \cdot 10^{-4} \cdot 1 \cdot 10^{-1} \cdot 1$ Vorgänger N$, 1 \cdot 10^{-4} \cdot 1 \cdot 10^{-1} \cdot 1$ Vorgänger V$\}$

light, N: $6 \cdot 10^{-5} \cdot 5 \cdot 10^{-1} \cdot 3 \cdot 10^{-3} = 9 \cdot 10^{-8}$

light, V: $6 \cdot 10^{-5} \cdot 1 \cdot 10^{-1} \cdot 2 \cdot 10^{-3} = 12 \cdot 10^{-9}$

q_F: $\max\{1 \cdot 10^{-1} \cdot 9 \cdot 10^{-8}$ Vorgänger N$, 1 \cdot 10^{-1} \cdot 12 \cdot 10^{-9}$ Vorgänger V$\}$

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