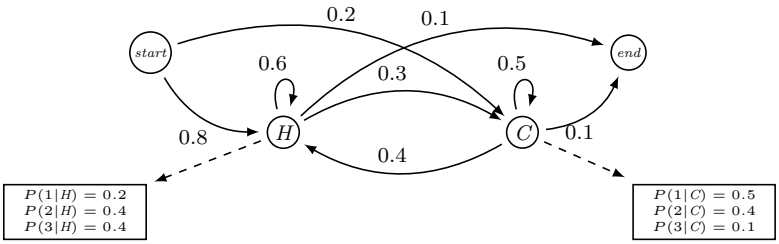
Einführung in die Computerlinguistik
Hausaufgabe 5, Abgabe 17.05.2022. 8.30 Uhr

Laura Kallmeyer

SoSe 2022, 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 probability that, when observing the weather on three different subsequent days, of encountering at least two subsequent cold days, i.e., twice the event C, one following the other?*

3. *What is the probability of an output* 1 3 *given this model?*

Solution:

1. This is independent from the output, only transition probabilities matter: $0.2 \cdot 0.5 \cdot 0.1 = 0.01 = 10^{-2}$

2. There are three possibilities for the paths, the probabilities have to be added.

   $0.2 \cdot 0.5 \cdot 0.5 \cdot 0.1 +$ (path *start C C C end*)
   $0.2 \cdot 0.5 \cdot 0.4 \cdot 0.1 +$ (path *start C C H end*)
   $0.8 \cdot 0.3 \cdot 0.5 \cdot 0.1$ (path *start H C C end*)
   $= (50 + 40 + 120) \cdot 10^{-4} = 21 \cdot 10^{-3} = 0.021$

3. We have to sum over the joint probabilities of path and output for all possible paths and the output 1 3:

   (path *start C C end*) $10^{-2} \cdot 0.5 \cdot 0.1$
   (path *start C H end*) $+0.2 \cdot 0.4 \cdot 0.1 \cdot 0.5 \cdot 0.4$
   (path *start H C end*) $+0.8 \cdot 0.3 \cdot 0.1 \cdot 0.2 \cdot 0.1$
   (path *start H H end*) $+0.8 \cdot 0.6 \cdot 0.1 \cdot 0.2 \cdot 0.4$
   $= (50 + 160 + 48 + 384)10^{-5} = 642 \cdot 10^{-5} = 0.00642$

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

```
Satz 1: (it PRO) (was AUX) (the OP) (telescreen N) (. .)
Satz 2: (the OP) (Thought N) (Police N) (used V) (the OP) (telescreen N) (. .)
Satz 3: (the OP) (Thought N) (Police N) (mattered V) (for P) (them PRO) (. .)
Satz 4: (Winston N) (kept V) (his OP) (back N) (turned V) (to P) (the OP) (telescreen N) (. .)
Satz 5: (Winston N) (wanted V) (to CLM) (turn V) (. .)
```

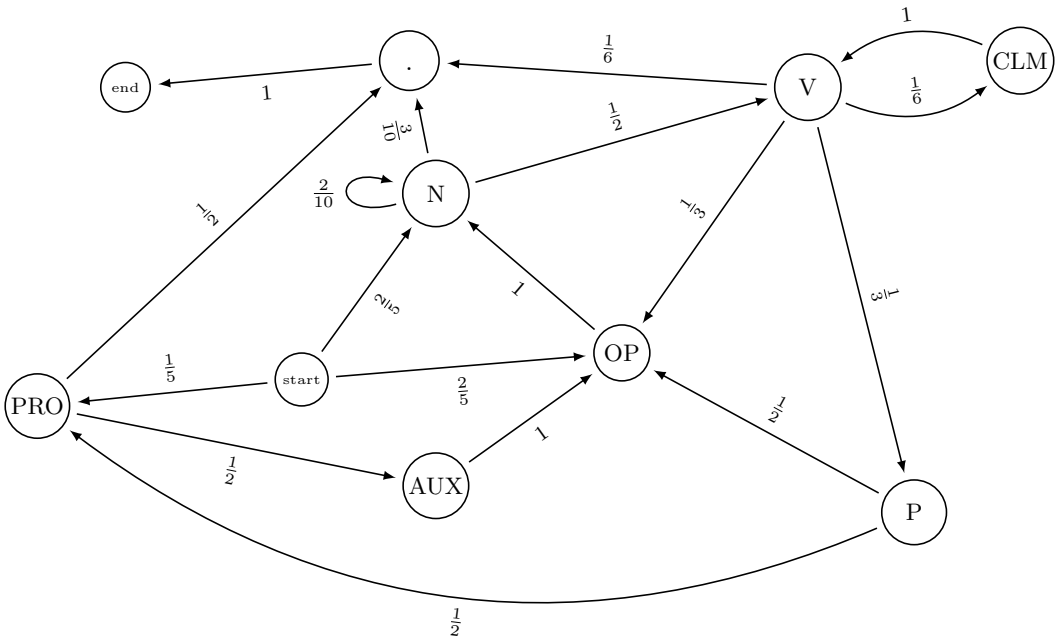
[1] Adaptiert von: RRG Baumbank, basierend auf George Orwell's "1984", https://rrgparbank.phil.hhu.de/

(`it`, `was`*, etc. sind die Token, die getaggt werden sollen, und* `PRO`, `AUX` *etc. sind die möglichen POS-Tags.)*

1. *Geben Sie 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 "Winston", "mattered", "to", "them" und "." betreffen, und die nicht* 0 *sind.*
3. *Was ergibt sich mit diesem Modell als beste Tagsequenz für den Satz "*`Winston mattered to them .`*"? 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. $P(Winston \mid N) = \frac{1}{5}$,
   $P(mattered \mid V) = \frac{1}{6}$,
   $P(to \mid P) = \frac{1}{2}$, $P(to \mid CLM) = 1$,
   $P(them \mid PRO) = \frac{1}{2}$,
   $P(. \mid .) = 1$.

3.

q_F						$\frac{2}{25\cdot2\cdot6\cdot6\cdot2\cdot2\cdot2}$, .
.					$\frac{2}{25\cdot2\cdot6\cdot6\cdot2\cdot2\cdot2}$, PRO	
N	$\frac{2}{25}$, start					
V		$\frac{2}{25\cdot2\cdot6}$, N				
CLM			$\frac{2}{25\cdot2\cdot6\cdot6}$, V			
P			$\frac{2}{25\cdot2\cdot6\cdot3\cdot2}$, V			
PRO				$\frac{2}{25\cdot2\cdot6\cdot6\cdot2\cdot2}$, P		
	Winston	mattered	to	them	.	

Rechnung für *them* und PRO

1. Vorgänger CLM: $\frac{2}{25\cdot2\cdot6\cdot6}\cdot 0\cdot\frac{1}{2}=0$
2. Vorgänger P: $\frac{2}{25\cdot2\cdot6\cdot6}\cdot\frac{1}{2}\cdot\frac{1}{2}=\frac{2}{25\cdot2\cdot6\cdot6\cdot2\cdot2}$, max

Beste POS Tag Folge: N V P PRO .

(Es gibt nur einen Pfad mit einer Wahrscheinlichkeit $\neq 0$, trotz der Ambiguität des POS Tags für *to*, da die Übergangswahrscheinlichkeit von CLM zu PRO 0 ist.)

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

Emissionswahrscheinlichkeiten:

$P(es|PRO)=1\cdot10^{-1}$ $P(ist|AUX)=2\cdot10^{-1}$ $P(hart|A)=1\cdot10^{-3}$
$P(ist|V)=1\cdot10^{-3}$ $P(hart|ADV)=6\cdot10^{-3}$

Alle anderen Emissionswahrscheinlichkeiten für es, ist *und* hart *seien* 0.

Übergangswahrscheinlichkeiten:

$P(AUX|PRO)=5\cdot10^{-1}$ $P(A|AUX)=1\cdot10^{-1}$ $P(A|V)=2\cdot10^{-2}$
$P(V|PRO)=1\cdot10^{-1}$ $P(ADV|AUX)=4\cdot10^{-2}$ $P(ADV|V)=2\cdot10^{-1}$

Angenommen, die Wahrscheinlichkeit, dass ein PRO am Satzanfang steht, ist $3\cdot10^{-1}$*, die, dass auf ein A ein Satzende folgt, ist* $0.01=1\cdot10^{-2}$ *und die, dass auf ein ADV ein Satzende folgt,* $0.1=1\cdot10^{-1}$*.*

Berechnen Sie in der folgenden, noch unvollständigen Viterbi Matrix die Einträge für A und ADV in Spalte 3, der sich bei diesen Wahrscheinlichkeiten für die Eingabe es ist hart *ergibt. Geben Sie Ihren Rechenweg an.*

q_F				
ADV				
A				
V		$3\cdot10^{-6}$, *PRO*		
AUX		$3\cdot10^{-3}$, *PRO*		
PRO	$3\cdot10^{-2}$, q_0			
	1 *es*	*2* *ist*	*3* *hart*	

(Spalten 1 und 2 sind schon vollständig.)

Lösung:

q_F				
ADV			$72\cdot10^{-8}$, AUX	
A			$3\cdot10^{-7}$, AUX	
V		$3\cdot10^{-6}$, PRO		
AUX		$3\cdot10^{-3}$, PRO		
PRO	$3\cdot10^{-2}$, q_0			
	1 es	2 ist	3 hart	

hart, ADV: $\max\{3\cdot10^{-3}\cdot P(\text{hart}|\text{ADV})\cdot P(\text{ADV}|\text{AUX})=72\cdot10^{-8}$ Vorgänger AUX,
$3\cdot10^{-6}\cdot P(\text{hart}|\text{ADV})\cdot P(\text{ADV}|\text{V})=36\cdot10^{-10}$ Vorgänger V$\}$

hart, A: $\max\{3\cdot10^{-3}\cdot P(\text{hart}|\text{A})\cdot P(\text{A}|\text{AUX})=3\cdot10^{-7}$ Vorgänger AUX,
$3\cdot10^{-6}\cdot P(\text{hart}|\text{A})\cdot P(\text{A}|\text{V})=6\cdot10^{-11}$ Vorgänger V$\}$