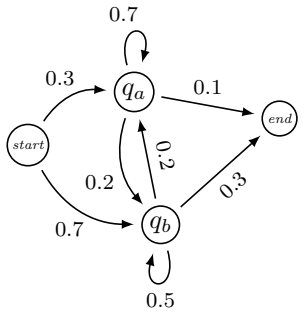
Einführung in die Computerlinguistik
Hausaufgabe 5, Abgabe 13.05.2019

Laura Kallmeyer

SoSe 2019, Heinrich-Heine-Universität Düsseldorf

Aufgabe 1 *Consider the following HMM:*



in q_a, *a is emitted with probability* 0.5
in q_a, *b is emitted with probability* 0.5
in q_b, *b is emitted with probability* 0.9
in q_b, *a is emitted with probability* 0.1

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

   $$start\ q_a\ q_b\ end$$

   according to this HMM?

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

   $$start\ q_b\ q_b\ end$$

   according to this HMM?

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

4. *Given that we see an output ba, what is the most probable path that has been traversed?*

Solution:

1. This is independent from the output, only transition probabilities matter: $0.3 \cdot 0.2 \cdot 0.3 = 18^{-3}$

2. In q_b, b has a higher probability as an emission.

   Most probable output for *start* $q_b q_b$ *end* is therefore *bb*.

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

(path *start* $q_a q_b$ *end*)	$18^{-3} \cdot 0.5 \cdot 0.1$
(path *start* $q_a q_a$ *end*)	$+0.3 \cdot 0.7 \cdot 0.1 \cdot 0.5 \cdot 0.5$
(path *start* $q_b q_a$ *end*)	$+0.7 \cdot 0.2 \cdot 0.1 \cdot 0.9 \cdot 0.5$
(path *start* $q_b q_b$ *end*)	$+0.7 \cdot 0.5 \cdot 0.3 \cdot 0.9 \cdot 0.1$
	$= (90 + 525 + 630 + 945)10^{-5} = 219 \cdot 10^{-4} = 0.0219$

4. Here we need the maximum over the joint probabilities of path and output for an output *ba*, i.e., the maximum of the four elements of the sum we calculated in 3. This is 0.0945, therefore the most probable path is the path *start* $q_b q_b$ *end*.

Aufgabe 2 *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') ('b', 'Y') ('.', '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 .`*"?*

Lösung

1. - $Q = \{X, Y, P, q_0, q_F\}$, wobei q_0 Start- und q_F Endzustand sind.
   - Übergangswahrscheinlichkeiten A:
     $A(q_0, X) = \frac{2}{3}$, $A(q_0, Y) = \frac{1}{3}$, $A(q_0, P) = 0$.
     $A(X, X) = \frac{1}{3}$, $A(X, Y) = \frac{2}{3}$, $A(X, P) = 0$, $A(X, q_F) = 0$.
     $A(Y, X) = \frac{0}{5} = 0$, $A(Y, Y) = \frac{2}{5} = 0.4$, $A(Y, P) = \frac{3}{5} = 0.6$, $A(Y, q_F) = 0$.
     $A(P, X) = 0$, $A(P, Y) = 0$, $A(P, P) = 0$, $A(P, q_F) = 1$.
   - $O = \{a, b, .\}$
   - Emissionswahrscheinlichkeiten B:
     $b_X(a) = \frac{2}{3}$, $b_X(b) = \frac{1}{3}$, $b_X(.) = 0$.
     $b_Y(a) = \frac{2}{5} = 0.4$, $b_Y(b) = \frac{3}{5} = 0.6$, $b_Y(.) = 0$.
     $b_P(a) = 0$, $b_P(b) = 0$, $b_P(.) = 1$.
2. Für "." kommt als einziges Tag P in Frage, aufgrund der Wahrscheinlichkeiten $A(X, q_F) = A(Y, q_F) = 0$ und $A(P, q_F) = 1$.

   Für "a" am Satzanfang kommt aufgrund von $A(q_0, P) = 0$ nur X oder Y in Frage.

   Außerdem schließt $A(X, P) = 0$ ein Aufeinanderfolgen von X und P aus, damit bleibt nur die Tagfolge YP.

   Wahrscheinlichkeit von YP als Sequenz für "a .": $\frac{1}{3} \cdot \frac{2}{5} \cdot \frac{3}{5} \cdot 1 \cdot 1 = \frac{6}{75} = \frac{2}{25}$

   Alle anderen Kombinationen führen, wie erläutert, zu einer Wahrscheinlichkeit 0, damit ist also YP die beste Tagsequenz.

Aufgabe 3 *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:

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

Alle anderen Emissionswahrscheinlichkeiten für unsere Eingabewörter seien 0*.*

Relevante Übergangswahrscheinlichkeiten:

$P(N|TO) = 8 \cdot 10^{-1}$ $P(TO|V) = 2 \cdot 10^{-1}$ $P(V|V) = 1 \cdot 10^{-1}$ $P(N|N) = 1 \cdot 10^{-1}$
$P(V|TO) = 1 \cdot 10^{-1}$ $P(TO|N) = 2 \cdot 10^{-1}$ $P(N|V) = 3 \cdot 10^{-1}$ $P(V|N) = 3 \cdot 10^{-1}$

Angenommen, die Wahrscheinlichkeit, dass ein N am Satzanfang steht ist $1 \cdot 10^{-1}$*, die, dass ein V oder ein TO am Satzanfang steht, ebenso. Die, dass ein Satzende auf N oder V folgt, ist auch jeweils* $1 \cdot 10^{-1}$*, die Wahrscheinlichkeit, dass ein Satz mit einem TO endet, ist* $1 \cdot 10^{-10}$*.*

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?*

Lösung:

1.

q_F				$144 \cdot 10^{-10}$, N
TO		$6 \cdot 10^{-5}$, N		
V	$2 \cdot 10^{-4}$, q_0		$12 \cdot 10^{-9}$, TO	
N	$3 \cdot 10^{-4}$, q_0		$144 \cdot 10^{-9}$, TO	
	1 start	2 to	3 study	

2. Die beste POS-TAG Folge ist N TO N.