Einführung in die Computerlinguistik
Hausaufgabe 5, Abgabe 14.05.2012

Laura Kallmeyer

SS 2012, Heinrich-Heine-Universität Düsseldorf

Aufgabe 1 *Gegeben sei das folgende POS-getaggte Mini-Korpus:*

Satz 1:	*light*	*the*	*bright*	*fires*	.
	V	*Det*	*A*	*N*	*PUNCT*
Satz 2:	*fires*	*are*	*nice*	.	
	N	*V*	*A*	*PUNCT*	
Satz 3:	*the*	*light*	*is*	*bright*	.
	Det	*N*	*V*	*A*	*PUNCT*

1. *Wie sehen Zustands- und Ausgabemenge für ein aus diesem Korpus gelerntes Bigram-HMM aus?*
2. *Geben Sie für ein aus diesem Korpus gelerntes Bigram-HMM die Übergangswahrscheinlichkeiten und die Ausgabewahrscheinlichkeiten an. (Es reichen die Brüche, die Dezimalzahlen müssen nicht angegeben werden.)*

Lösung:

1. $Q = \{q_0, q_f, \text{V}, \text{Det}, \text{A}, \text{N}, \text{PUNCT}\}$, $O = \{\text{light}, \text{the}, \text{bright}, \text{fires}, \text{are}, \text{nice}, \text{is}, .\}$.
2. Überganswahrscheinlichkeiten:

$P(\text{Det}\|\text{V}) = \frac{1}{3}$	$P(\text{Det}\|\text{Det}) = 0$	$P(\text{Det}\|\text{A}) = 0$	$P(\text{Det}\|\text{N}) = 0$
$P(\text{A}\|\text{V}) = \frac{2}{3}$	$P(\text{A}\|\text{Det}) = \frac{1}{2}$	$P(\text{A}\|\text{A}) = 0$	$P(\text{A}\|\text{N}) = 0$
$P(\text{N}\|\text{V}) = 0$	$P(\text{N}\|\text{Det}) = \frac{1}{2}$	$P(\text{N}\|\text{A}) = \frac{1}{3}$	$P(\text{N}\|\text{N}) = 0$
$P(\text{V}\|\text{V}) = 0$	$P(\text{V}\|\text{Det}) = 0$	$P(\text{V}\|\text{A}) = 0$	$P(\text{V}\|\text{N}) = \frac{2}{3}$
$P(\text{PUNCT}\|\text{V}) = 0$	$P(\text{PUNCT}\|\text{Det}) = 0$	$P(\text{PUNCT}\|\text{A}) = \frac{2}{3}$	$P(\text{PUNCT}\|\text{N}) = \frac{1}{3}$

   PUNCT steht mit Wahrscheinlichkeit 1 am Satzende, alle anderen Nachfolgetags für PUNCT haben Wahrscheinlichkeit 0.

   Für alle anderen Tags ist die Wahrscheinlichkeit, am Satzende zu stehen, 0.

   N, V und Det stehen jeweils mit Wahrscheinlichkeit $\frac{1}{3}$ am Satzanfang, alle anderen Tags mit Wahrscheinlichkeit 0.

   Ausgabewahrscheinlichkeiten:

	light	the	bright	fires	are	nice	is	.
Det	0	1	0	0	0	0	0	0
N	$\frac{1}{3}$	0	0	$\frac{2}{3}$	0	0	0	0
A	0	0	$\frac{2}{3}$	0	0	$\frac{1}{3}$	0	0
V	$\frac{1}{3}$	0	0	0	$\frac{1}{3}$	0	$\frac{1}{3}$	0
PUNCT	0	0	0	0	0	0	0	1

Aufgabe 2 *Nehmen Sie an, Sie haben einen HMM-POS Tagger, unter anderem mit folgenden Wahrscheinlichkeiten:*

Emissionswahrscheinlichkeiten:

$P(\textit{the}\|\textit{Det}) = 1$	$P(\textit{light}\|\textit{N}) = 3 \cdot 10^{-3}$	$P(\textit{fires}\|\textit{N}) = 5 \cdot 10^{-3}$
	$P(\textit{light}\|\textit{Adj}) = 3 \cdot 10^{-3}$	$P(\textit{fires}\|\textit{V}) = 3 \cdot 10^{-3}$
	$P(\textit{light}\|\textit{V}) = 2 \cdot 10^{-3}$	

Alle anderen Emissionswahrscheinlichkeiten für light*,* the *und* fires *seien* 0.

Übergangswahrscheinlichkeiten:

$P(N|Det) = 5 \cdot 10^{-1}$ $P(N|N) = 1 \cdot 10^{-1}$ $P(N|Adj) = 5 \cdot 10^{-1}$ $P(N|V) = 2 \cdot 10^{-1}$
$P(Adj|Det) = 3 \cdot 10^{-1}$ $P(V|N) = 4 \cdot 10^{-1}$ $P(V|Adj) = 1 \cdot 10^{-1}$ $P(V|V) = 2 \cdot 10^{-1}$
$P(V|Det) = 1 \cdot 10^{-1}$

Angenommen, die Wahrscheinlichkeit, dass ein Det am Satzanfang steht, ist 1*, die, 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* the light fires *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* the light fires *ergibt?*

Lösung:

1.

q_F				$225 \cdot 10^{-9}$, N
N		$15 \cdot 10^{-4}$, Det	$225 \cdot 10^{-8}$, Adj	
V		$2 \cdot 10^{-4}$, Det	$180 \cdot 10^{-8}$, N	
Adj		$9 \cdot 10^{-4}$, Det		
Det	1, q_0			
	1 the	2 light	3 fires	

light, Adj: $1 \cdot P(\text{Adj}|\text{Det}) \cdot P(\text{light}|\text{N}) = 1 \cdot 3 \cdot 10^{-1} \cdot 3 \cdot 10^{-3}$

light, N: $1 \cdot 5 \cdot 10^{-1} \cdot 3 \cdot 10^{-3}$

light, V: $1 \cdot 1 \cdot 10^{-1} \cdot 2 \cdot 10^{-3}$

fires, N: $\max\{15 \cdot 10^{-4} \cdot 1 \cdot 10^{-1} \cdot 5 \cdot 10^{-3}$ Vorgänger N $, 9 \cdot 10^{-4} \cdot 5 \cdot 10^{-1} \cdot 5 \cdot 10^{-3}$ Vorgänger Adj$, 2 \cdot 10^{-4} \cdot 2 \cdot 10^{-1} \cdot 5 \cdot 10^{-3}$ Vorgänger V$\}$

fires, V: $\max\{15 \cdot 10^{-4} \cdot 4 \cdot 10^{-1} \cdot 3 \cdot 10^{-3}$ Vorgänger N $, 9 \cdot 10^{-4} \cdot 1 \cdot 10^{-1} \cdot 3 \cdot 10^{-3}$ Vorgänger Adj$, 2 \cdot 10^{-4} \cdot 2 \cdot 10^{-1} \cdot 3 \cdot 10^{-3}$ Vorgänger V$\}$

2. Die beste POS-TAG Folge ist demnach Det Adj N.