Lösungen für Aufgabenblatt 7 — 5.12.2019

Einführung in die Computerlinguistik (WS19/20) — Abgabe bis 14.12.2019

Aufgabe 1) Manuelles POS Tagging **Punkte: 6**

1. Taggen Sie folgenden Textauszug mithilfe des Penn Treebank Tagsets[1]. Berücksichtigen Sie auch die erforderliche Tokenisierung. (2 Punkte)

   `My mama always said 'Life was like a box of chocolates; you never know what you're gonna get.'`
   aus: *Forrest Gump*

2. Taggen Sie folgenden Textauszug mithilfe des Stuttgart-Tübingen-Tagsets[2]. Berücksichtigen Sie auch die erforderliche Tokenisierung. (2 Punkte)

   `Die Hummel hat eine Flügelfläche von 0,7 Quadratzentimeter, bei 1,2 Gramm Gewicht. Nach den bekannten Gesetzen der Aerodynamik ist es unmöglich, bei diesen Verhältnissen zu fliegen.`
   Arthur Lassen, Motivationstrainer

3. Nennen Sie zwei Unterschiede im Tagging von Eigennamen zwischen STTS und Penn Treebank Tagset (2 Punkte).

Lösung zu 1)

1. `My/PRP$ mama/NN always/RB said/VBD '/`` Life/NN was/VBD like/IN a/DT box/NN of/IN chocolates/NNS ;/: you/PRP never/RB know/VBP what/WP you/PRP 're/VBP gon/VBG na/TO get/VB ./. '/''`

2. `Die/ART Hummel/NN hat/VAFIN eine/ART Flügelfläche/NN von/APPR 0,7/CARD Quadratzentimeter/NN ,/$, bei/APPR 1,2/CARD Gramm/NN Gewicht/NN ./$. Nach/APPR den/ART bekannten/ADJA Gesetzen/NN der/ART Aerodynamik/NN ist/VAFIN es/PPER unmöglich/ADJD ,/$, bei/APPR diesen/PDAT Verhältnissen/NN zu/PTKZU fliegen/VVINF ./$.`

3. Unterscheidung zwischen Plural- und Singulareigennamen in der Penn Treebank. Wochentage als Eigennamen in Penn Treebank.

[1] `https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/Penn-Treebank-Tagset.pdf`
[2] `http://www.sfs.uni-tuebingen.de/resources/stts-1999.pdf`

Aufgabe 2) HMM Tagging Simulation **Punkte: 14**

Gegeben ist das folgende deutsche Korpus. Groß- und Kleinschreibung haben wir ignoriert, sonst wäre es zu einfach. Getaggt wurde es mit dem Universal Tagset. Das Tagging enthält Punktuation. Nehmen Sie dennoch auch Satzanfangszustände und Satzendzustände an.

```
regenwuermer/NOUN regen/VERB sich/PRON im/ADP regen/NOUN ./PUNCT
im/ADP regen/ADJ Verkehr/NOUN regen/VERB sich/PRON menschen/NOUN auf/PART ./PUNCT
im/ADP november/NOUN folgt/VERB auf/ADP regen/NOUN regen/NOUN ?/PUNCT
```

a) Trainieren des HMMs *Punkte: 4*

Berechnen Sie für das Korpus die Übergangs- und Emissionswahrscheinlichkeiten eines Hidden Markov Modells, welches die POS-Tag-Folge als unbeobachtete Zustände annimmt. Vergessen Sie bei den Übergangswahrscheinlichkeiten nicht die Satzanfänge und Satzenden. Smoothing ist nicht nötig.

Lösung zu 2a)

t_{i-1}	t_i	prob
PUNCT	END	1
START	NOUN	0.33
START	ADP	0.67
VERB	PRON	0.67
VERB	ADP	0.33
NOUN	PUNCT	0.29
NOUN	PART	0.14
NOUN	NOUN	0.14
NOUN	VERB	0.43
PRON	NOUN	0.5
PRON	ADP	0.5
ADP	NOUN	0.75
ADP	ADJ	0.25
PART	PUNCT	1
ADJ	NOUN	1

(a) Übergangswahrscheinlichkeiten $p(t_i|t_{i-1})$

PUNCT	.	0.67
PUNCT	?	0.33
VERB	folgt	0.33
VERB	regen	0.67
NOUN	regenwuermer	0.14
NOUN	verkehr	0.14
NOUN	november	0.14
NOUN	regen	0.43
NOUN	menschen	0.14
PRON	sich	1
ADP	auf	0.25
ADP	im	0.75
PART	auf	1
ADJ	regen	1

(b) Emissionswahrscheinlichkeiten $p(w_i|t_j)$

Nullwahrscheinlichkeiten sind nicht angegeben bzw. alle anderen, nicht gegebenen Wahrscheinlichkeiten sind Null.

b) Tagging *Punkte: 10*

Gegeben ist nun folgender Testsatz:

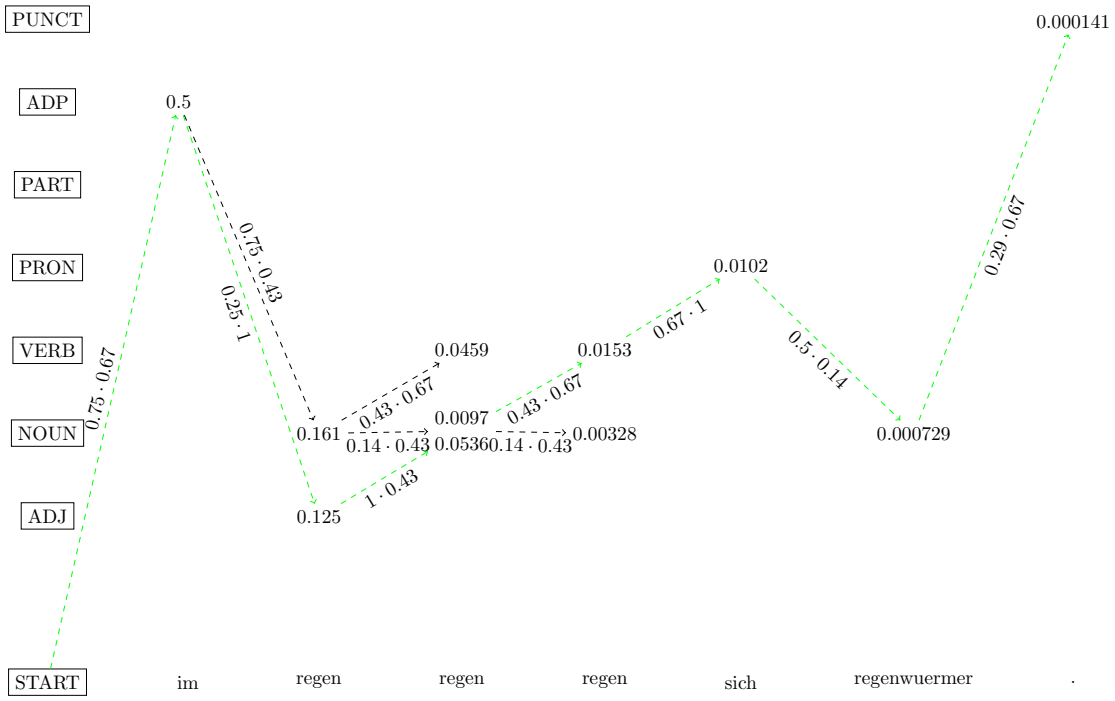
```
im regen regen regen sich regenwuermer.
```

1. Ermitteln Sie die wahrscheinlichste POS-Tag Sequenz für den Testsatz durch Anwendung Ihres trainierten HMM-Modells. Benutzen Sie hierfür den Viterbi-Algorithmus und zeigen Sie Ihre Trellis. (Es gibt sehr viele Nullen, was das ganze stark vereinfacht). (7 Punkte)

2. Welche Wahrscheinlichkeit und welche Taggingsequenz weist Ihr Modell dem Satz zu? Ist die zugewiesene Sequenz richtig? (1 Punkt)

3. Wie würde ein greedy-Vorgehen den Satz taggen? Welche Entscheidungen sind dann richtig, welche falsch? (2 Punkte)

Lösung zu 2b)



	im	regen	regen	regen	sich	regenwuermer	.
ADJ	0.00e+00	**1.25e-01**	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00
VERB	0.00e+00	0.00e+00	4.59e-02	**1.53e-02**	0.00e+00	0.00e+00	0.00e+00
PART	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00
ADP	**5.00e-01**	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00
PRON	0.00e+00	0.00e+00	0.00e+00	0.00e+00	**1.02e-02**	0.00e+00	0.00e+00
NOUN	0.00e+00	1.61e-01	**5.36e-02**	3.28e-03	0.00e+00	**7.29e-04**	0.00e+00
PUNCT	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	**1.41e-04**

Wahrscheinlichkeit: $1.41 \cdot 10^{-4}$; Sequenz: ADP ADJ NOUN VERB PRON NOUN PUNCT

Ein Greedy Verfahren würde nicht richtig taggen. Hier würde schon das erste *regen* fälschlicherweise als Nomen taggen, da Nomen häufiger auf Präpositionen folgen als Adjektive (im Trainingskorpus). Dann würde es mit ADP NOUN VERB beginnen und danach keine weitere Möglichkeit zum Tagging mit dem ungesmootheden Verfahren finden, die nicht zu einer Nullwahrscheinlichkeit führt.

Aufgabe 3) Bonus: HMM Fragen **Punkte: 4**

Ein weiteres geläufiges Beispiel, um HMMs zu erklären, besteht aus dem Zusammenhang von Wetter und Tätigkeiten. Dabei sind Wetterfeatures (`warm`, `regnerisch`, `kalt` ...) die *(hidden) states*, und Tätigkeiten (*kochen*, *schwimmen*, ...) die *observations/Beobachtungen*. Aus einer Beobachtungssequenz mit je einer Beobachtung pro Tag wie *kochen schwimmen schwimmen Fußball lesen Lego* soll man dann das Wetter an den (im Beispiel 6) Tagen dekodieren.

1. Welche Wahrscheinlichkeiten braucht man, um solch ein Modell zu trainieren? (2 Punkte)

2. In wiefern werden sich die Übergangswahrscheinlichkeiten in diesem Modell von denen für POS Tags unterscheiden? Begünden Sie Ihre Antwort linguistisch? (2 Punkte)

Lösung zu 3)

1. Emission: $p(aktion_i|wetter_j)$, bedeutend, wie wahrscheinlich ist eine Aktivität bei einer bestimmten Wetterlage. Übergang $p(wetter_i|wetter_{i-1})$, bedeutend wie wahrscheinlich ist das Wetter an Tag i bei einem bestimmtem Wetter am Vortag.

2. Das Wetter kann öfters in gleichen states verweilen, während das Aufeinanderfolgen von Sequenzen wie "VERB VERB VERB" durch die der Sprache zugrunde liegenden Syntax unterbunden wird.