Tutorium zu „Einführung in die Computerlinguistik“

POS-Tagging + Hidden Markov Model + Viterbi

Brown POS-Tags

AT	–	Artikel
BEZ	–	‚is‘
IN	–	Präposition
JJ	–	Adjektiv
JJR	–	Komparativ Adj.
MD	–	Modalverb
NN	–	Nomen Sing./nicht zählbar
NNP	–	Sing. Name (Propernoun)
NNS	–	Nomen Plural
PERIOD	–	Satzzeichen
PN	–	Pronomen
RB	–	Adverb
RBR	–	Komparativ Adv.
TO	–	‚to‘
VB	–	Verb infinitiv
VBD	–	V. Vergangenheit
VBG	–	V. Gerund Gegenw. (-ing)
VBN	–	V. Gerund Vergang. (-ing)
VBZ	–	V. 3. Pers. sing. Gegenw.
WDT	–	Interrogativartikel (which, what,...)

nur die für die Klausur relevanten Brown Tags; Übersicht über alle Tags z.B. hier: https://en.wikipedia.org/wiki/Brown_Corpus#Part-of-speech_tags_used

Bert	kaufte	im	Laden	leckere	Schokolade	für	seine	Freunde	.
NNP	VBD	IN	NN	JJ	NN	IN	AT	NNS	PERIOD

It	is	good	taking	a	break	sometimes	.
PN	BEZ	JJ	VBG	AT	NN	RB	PERIOD

Markov Model n-ter Ordnung

0. Ordnung:	Wort/Tag hängt nicht von Vorgänger ab	$\frac{\text{Count Vorkommen Wort/Tag}}{\text{Count alle Wörter (auch Satzzeichen \& ggf. <s>)}}$
1. Ordnung:	Wort/Tag hängt vom direkten Vorgänger ab	$\frac{\text{Count Vorkommen Vorgänger \& Wort/Tag}}{\text{Count Vorkommen Vorgänger}}$
2. Ordnung:	Wort/Tag hängt von 2 Vorgängern ab	$\frac{\text{Count Vorkommen 2. Vorg. \& 1. Vorg. \& Wort/Tag}}{\text{Count Vorkommen 2. \& 1. Vorgänger}}$
3. Ordnung:	Wort/Tag hängt von 3 Vorgängern ab	$\frac{\text{Count Vorkommen 3. \& 2. \& 1. Vorgänger \& Wort/Tag}}{\text{Count Vorkommen 3. \& 2. \& 1. Vorgänger}}$

...

Übung

<s> Bert gönnt sich morgens gerne eine schöne heiße Tasse Kaffee. Susi dagegen mag lieber eine schöne heiße Tasse Tee. Susi mag keinen Kaffee, nicht mal eine Tasse Kaffee. <s>

<s> Wenn Fliegen hinter Fliegen fliegen, fliegen Fliegen Fliegen nach. <s>

Viterbi

Übergangswahrscheinlichkeit:

$P(t^k|t^{k-1})$ >> Wahrscheinlichkeit, dass Tag k auf Tag k-1 folgt („P für t^k gegeben t^{k-1}“)

$$P_{ml}(t^k|t^{k-1}) = \frac{C(t^{k-1}\, t^k)}{C(t^{k-1})} = \frac{\text{Count Vorkommen Tag k-1 dann Tag k}}{\text{Count Vorkommen Tag k-1}}$$

k-1 = Vorgänger von k

Lexikalische Wahrscheinlichkeit:

$$P(w|t) = \frac{C(w : t)}{C(t)} = \frac{\text{Count w als t getagged}}{\text{Count Tag t}}$$

$vtrb_k$(Tag k) = P(Tag Position k-1) Wahrsch. Tag * P(Tag Pos. k | Tag Pos. k-1) Übergangswahrscheinlichkeit * P(Wort Position k | Tag Position k) Lexikalische Wahrscheinlichkeit

>> Immer bei der wahrscheinlichsten Möglichkeit weiter machen

>> beim letzten Wort wird nur der wahrscheinlichste Tag mit der Übergangswahrscheinlichkeit von diesem zum End-Tag multipliziert (eine lexikalische Wahrscheinlichkeit gibt es hier nicht)

Übung

Viterbi Algorithmus für „The birds fly high“

Übergangswahrscheinlichkeiten

t \ t‘	S	NN	DT	VB	JJ
E	0	0,15	0,05	0,35	0,25
NN	0,3	0,3	0,6	0,2	0,3
DT	0,4	0,05	0,05	0,2	0,05
VB	0,2	0,4	0,05	0,05	0,2
JJ	0,1	0,1	0,25	0,2	0,2

Lexikalische Wahrscheinlichkeiten

w	the	birds	fly	high
P(w|NN)	0,04	0,6	0,45	0,4
P(w|DT)	0,7	0,05	0,03	0,03
P(w|VB)	0,05	0,1	0,6	0,1
P(w|JJ)	0,02	0,05	0,04	0,6