Tagging and HMMs
Jonathan Stewart

1)

i. I/PRP need/VBP a/DT flight/NN from/IN Atlanta/NN
Atlanta is a proper noun, so it should be coded as NNP.

ii. Does/VBZ this/DT flight/NN serve/VB dinner/NNS

Dinner is not a plural noun. Instead, it should be coded as NN.

iii. I/PRP have/VB a/DT friend/NN living/VBG in/IN Denver/NNP

Have is a past tense verb in this case, and should be coded as VBD.

iv. Can/VBP you/PRP list/VB the/DT nonstop/JJ afternoon/NN flights/NNS

Afternoon is a descriptor/adjective in this case, and should be coded as JJ.

Tag	Description	Example	Tag	Description	Example	Tag	Description	Example
CC	coordinating conjunction	*and, but, or*	PDT	predeterminer	*all, both*	VBP	verb non-3sg present	*eat*
CD	cardinal number	*one, two*	POS	possessive ending	*'s*	VBZ	verb 3sg pres	*eats*
DT	determiner	*a, the*	PRP	personal pronoun	*I, you, he*	WDT	wh-determ.	*which, that*
EX	existential 'there'	*there*	PRP$	possess. pronoun	*your, one's*	WP	wh-pronoun	*what, who*
FW	foreign word	*mea culpa*	RB	adverb	*quickly*	WP$	wh-possess.	*whose*
IN	preposition/ subordin-conj	*of, in, by*	RBR	comparative adverb	*faster*	WRB	wh-adverb	*how, where*
JJ	adjective	*yellow*	RBS	superlatv. adverb	*fastest*	$	dollar sign	*$*
JJR	comparative adj	*bigger*	RP	particle	*up, off*	#	pound sign	*#*
JJS	superlative adj	*wildest*	SYM	symbol	+,%, &	"	left quote	' or "
LS	list item marker	*1, 2, One*	TO	"to"	*to*	"	right quote	' or "
MD	modal	*can, should*	UH	interjection	*ah, oops*	(	left paren	[, (, {, <
NN	sing or mass noun	*llama*	VB	verb base form	*eat*	)	right paren	],), }, >
NNS	noun, plural	*llamas*	VBD	verb past tense	*ate*	,	comma	,
NNP	proper noun, sing.	*IBM*	VBG	verb gerund	*eating*	.	sent-end punc	. ! ?
NNPS	proper noun, plu.	*Carolinas*	VBN	verb past part.	*eaten*	:	sent-mid punc	: ; ... – -

2)

Back given MD::
Probability of the prior word:

(3E-8)*P(VB|MD)*P(back|VB) = (3E-8)*(.7968)*(.00067)=**1.6E-11**
(2.3E-13)*P(VB|VB)*P(back|VB)=(2.3E-13)*(.005)*(.00067)=7.7E-19

(1.1e-10)*P(VB|NN)*(P(back|VB)=(1.1e-10)*(.0014)*(.00067)

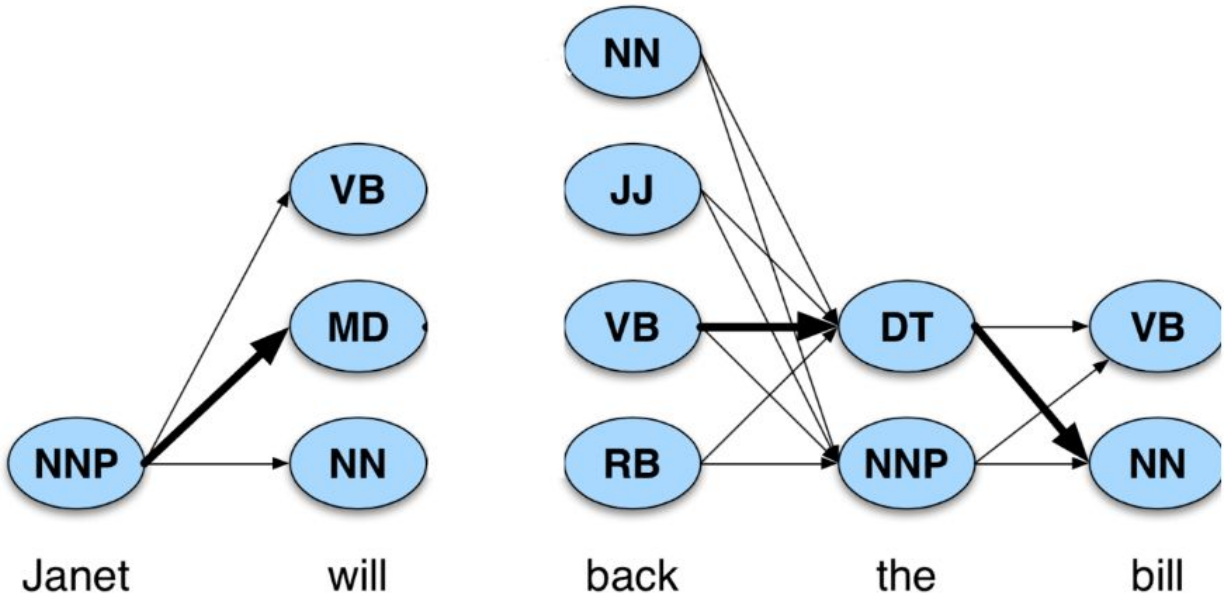
Since the max of the probabilities is 1.6E-11, This means that P(MD)*P(VB|MD)*P(back|VB) is the likelihood that is kept as the likelihood of 'back' being a verb classification.

Back given RB:

(3E-8)*P(RB|MD)*P(back|RB) = (3E-8)*(.1698)*(.0104)=**5.29e-11**
(2.3E-13)*P(RB|VB)*P(back|RB)=(2.3E-13)*(.0514)*(.0104)=1.2E-16
(1.1e-10)*P(RB|NN)*(P(back|RB)=(1.1e-10)*(.0177)*(.0104)=2.0E-13
Since the max of the probabilities is 5.3E-11, this means that P(MD)*P(RB|MD)*P(back|RB) is kept as the likelihood of 'back' being an RB classification

	NNP	MD	VB	JJ	NN	RB	DT
<s>	0.2767	0.0006	0.0031	0.0453	0.0449	0.0510	0.2026
NNP	0.3777	0.0110	0.0009	0.0084	0.0584	0.0090	0.0025
MD	0.0008	0.0002	0.7968	0.0005	0.0008	0.1698	0.0041
VB	0.0322	0.0005	0.0050	0.0837	0.0615	0.0514	0.2231
JJ	0.0366	0.0004	0.0001	0.0733	0.4509	0.0036	0.0036
NN	0.0096	0.0176	0.0014	0.0086	0.1216	0.0177	0.0068
RB	0.0068	0.0102	0.1011	0.1012	0.0120	0.0728	0.0479
DT	0.1147	0.0021	0.0002	0.2157	0.4744	0.0102	0.0017

3)



The first step is estimating how to classify the word will. To do this, it is first taking the word ‘Janet’, as this is the preceding word. ‘Janet’ has an unsmoothed probability of only being an “NNP’, so there is only one preceding, non-zero state. Next, the probability of a verb, given an NNP, the probability of an MD, given an NNP, and the probability of an NN, given an NNP, are calculated, as are the emission probabilities of will being a VB, MD, or NN. Since there is only one preceding possibility, there is only one possible entry per state, and the max probability for each state is the only possible possibility. From the graph, ‘will’ has the highest likelihood of being an MD.

The second image shows three consecutive words. The first word, ‘back’ has four possible non-zero states. The second word, ‘the’ has two possible states. For each state in ‘back’ the probability associated with that state, times the probability of DT or NNP, given the four preceding tags, times the probability of ‘the’, given the DT or NNP are calculated. For each of those two states, the highest of the probabilities calculated are chosen. Once this is accomplished, the same process is repeated for the word ‘bill’. From the above graph, it can be seen that the VB -> DT -> NN is the most likely path.