Name: Solutions

Email: hollingk@gmail.com

CMSC/LING 723 - Computational Linguistics I

Midterm Exam 18 Oct 2011

Question	Points
1	7
2	5
3a	10
3b	3
4	5
5	5
6	5
7a	10
7b	15
7c	5
7d	20
7e	10
Total	100

For Question 1, consider the following set of sentences as a corpus:

<s> I am Sam </s>
<s> Sam I am </s>
<s> I do not like green eggs and ham </s>
<s> I do not like them , Sam I am </s>

Question 1a. (1 point) How many bigrams are there in this corpus (total)?

27 total; 18 unique.

Question 1b. (1 point) What is the most frequent bigram in this corpus?

"I am" and "<s> I" are tied, at 3 occurrences each.

Question 1b. (2 points) What is the probability of the bigram "*I am*" under an unsmoothed model (i.e., $P_{MLE}(am \mid I)$)?

$\frac{\text{c(I am)}}{\text{c(I)}} = \frac{3}{5} = 0.6$

Question 1b. (2 points) What is the probability of the bigram "*I do*" under an unsmoothed model (i.e., $P_{MLE}(do \mid I)$)?

$\frac{\text{c(I do)}}{\text{c(I)}} = \frac{2}{5} = 0.4$

Question 1c. (1 point) What is the most frequent co-occurrence bigram in this corpus?

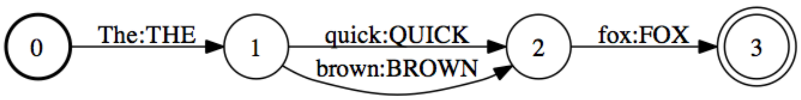
"<s> I" and "I </s>" are tied at 5 occurrences each.

Question 2. (5 points) What is the Soundex representation for the name "*Engekrethson*"?

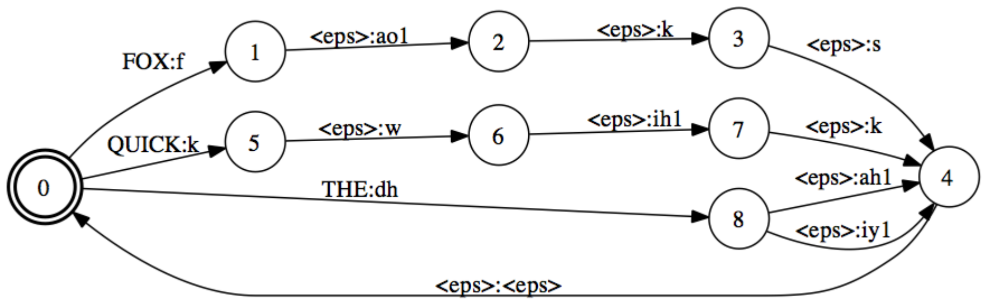
E522 (E526 is incorrect, and indicates improper treatment of the "gek" subsequence.)

For Question 3, consider the following two FSTs.

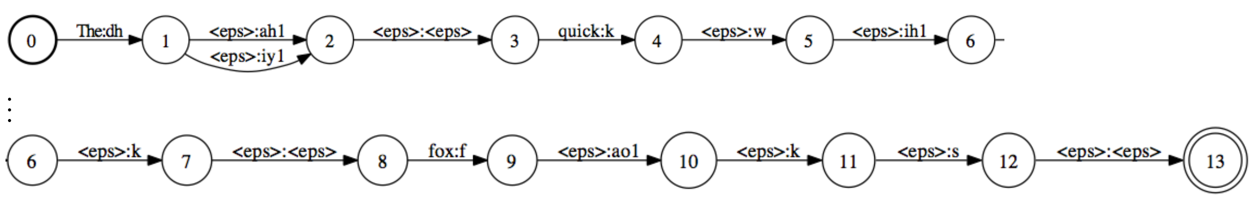
FST1:



FST2:



Question 3a. (10 points) Draw the composition of these two transducers (i.e., FST2 ∘ FST1).



Question 3b. (3 points) Is FST1 a deterministic transducer? Is FST2? Is your composed FST for 3a deterministic?

Yes, no, no.

Question 4. (5 points) Explain the Markov Assumption and why we use it. You may use equations in your answer if necessary, and/or give an example of how we might use the Markov Assumption.

The Markov Assumption is that events are conditionally independents of other events that occurred $n+1$ time steps prior (for a Markov order-n model). We use the Markov assumption because it is computationally intractable to condition on every previous event in our history; the assumption can also resolve some sparsity issues.

Question 5. (5 points) Explain the difference between interpolation models and backoff models of smoothing.

Interpolation models use information from all of the models; backoff models use information from only one model, "backing off" to a lower-order model only if there is not enough evidence to use information from a higher-order model.

Question 6. (5 points) Consider the following scenario: we have a text corpus of 10,000 words, and are provided with a vocabulary list of 100 words for the corpus. There are 5 words in our vocabulary that were never observed in the corpus. If we have observed the word "*the*" 900 times in our corpus, what is the expected frequency estimate of this word according to Laplace's Law?

N=10,000
V=100
C(the)=900
$C_{LAP} = P_{LAP} * N = \frac{C+1}{N+V} * N = \frac{901}{10{,}100} * 10,000 = 892.08$

For Question 7, consider the following string: *bats can fly*

An HMM POS-tag model is described by the following two tables:

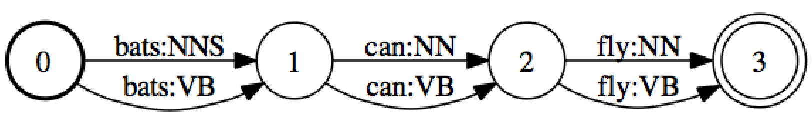
$b_j(v_k) = \mathbf{P}(v_k|\tau_j)$

	k:	1	2	3
j		bats	can	fly
1	NN	0	$\frac{1}{2}$	$\frac{1}{2}$
2	NNS	1	0	0
3	VB	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$

$a_{ij} = \mathbf{P}(\tau_j|\tau_i)$

	j:	0	1	2	3
i		</s>	NN	NNS	VB
0	<s>	0	0	$\frac{1}{2}$	$\frac{1}{2}$
1	NN	$\frac{1}{4}$	$\frac{1}{4}$	0	$\frac{1}{2}$
2	NNS	0	$\frac{1}{2}$	0	$\frac{1}{2}$
3	VB	$\frac{1}{2}$	$\frac{1}{4}$	0	$\frac{1}{4}$

Question 7a. (10 points) Based on the model given above, draw an (unweighted) FST of word:tag sequences for the string "*bats can fly*", using as few states as possible. Show only the word:tag sequences which have probability greater than zero.



Question 7b. (15 points) Use the following table (trellis) to fill in the Viterbi forward probability $\alpha_i^{\max}(t)$, as well as the corresponding back pointers, given the model from the previous page.

(see 1st table on next page; v=Viterbi, **red** $v^{\max}$=Viterbi-best values, ψ=back pointers)

Question 7c. (5 points) What is the Viterbi-best tag sequence according to the above chart?

NNS NN VB

Question 7d. (20 points). What is the probability of the string "*bats can fly*" according to the aforementioned model? You may re-use the table (trellis) on the previous page, or re-create another below.

$\frac{105+224}{13{,}824} = \frac{329}{13{,}824} = 0.023799$

(see 2nd table on next page for work)

Question 7e. (10 points) What is the perplexity of the aforementioned model on the string "*bats can fly*"?

$0.023799^{-1/N}, N = 4$
$= 2.5460$

	t:	0	1	2	3	4
j	τ_k:	<s>	bats	can	fly	</s>
0		1				$v_{\text{<s>}}(4) = v_{\text{NN}}(3) * a_{\text{NN},\text{<s>}} * b$ $= \frac{1}{64} * \frac{1}{4} * 1 = \frac{1}{256}$ $v^{\max}_{\text{<s>}}(4) = v_{\text{VB}}(3) * a_{\text{VB},\text{<s>}} * b$ $= \frac{1}{48} * \frac{1}{2} * 1 = \frac{1}{96}$ $\psi_{\text{<s>}}(4) = \text{VB}$
1	NN	0	$b_{\text{NN,bats}} = 0$	$v^{\max}_{\text{NN}}(2) = v_{\text{NNS}}(1) * a_{\text{NNS,NN}} * b_{\text{NN,can}}$ $= \frac{1}{2} * \frac{1}{2} * \frac{1}{2} = \frac{1}{8}$ $v_{\text{NN}}(2) = v_{\text{VB}}(1) * a_{\text{VB,NN}} * b_{\text{NN,can}}$ $= \frac{1}{6} * \frac{1}{4} * \frac{1}{2} = \frac{1}{48}$ $\psi_{\text{NN}}(2) = \text{NNS}$	$v^{\max}_{\text{NN}}(3) = v_{\text{NN}}(2) * a_{\text{NN,NN}} * b_{\text{NN,fly}}$ $= \frac{1}{8} * \frac{1}{4} * \frac{1}{2} = \frac{1}{64}$ $v_{\text{NN}}(3) = v_{\text{VB}}(2) * a_{\text{VB,NN}} * b_{\text{NN,fly}}$ $= \frac{1}{12} * \frac{1}{4} * \frac{1}{2} = \frac{1}{96}$ $\psi_{\text{NN}}(3) = \text{NN}$	
2	NNS	0	$v_{\text{NNS}}(1) = \alpha_{\text{<s>}}(0) * a_{\text{<s>},\text{NNS}} * b_{\text{NNS,bats}}$ $= 1 * \frac{1}{2} * 1 = \frac{1}{2}$ $\psi_{\text{NNS}}(1) = \text{<s>}$	$b_{\text{NNS,can}} = 0$	$b_{\text{NNS,fly}} = 0$	
3	VB	0	$v_{\text{VB}}(1) = \alpha_{\text{<s>}}(0) * a_{\text{<s>},\text{VB}} * b_{\text{VB,bats}}$ $= 1 * \frac{1}{3} * \frac{1}{2} = \frac{1}{6}$ $\psi_{\text{VB}}(1) = \text{<s>}$	$v^{\max}_{\text{VB}}(2) = v_{\text{NNS}}(1) a_{\text{NNS,VB}} * b_{\text{VB,can}}$ $= \frac{1}{2} * \frac{1}{2} * \frac{1}{3} = \frac{1}{12}$ $v_{\text{VB}}(2) = v_{\text{VB}}(1) a_{\text{VB,VB}} * b_{\text{VB,can}}$ $= \frac{1}{6} * \frac{1}{4} * \frac{1}{3} = \frac{1}{72}$ $\psi_{\text{VB}}(2) = \text{NNS}$	$v^{\max}_{\text{VB}}(3) = v_{\text{NN}}(2) a_{\text{NN,VB}} * b_{\text{VB,fly}}$ $= \frac{1}{8} * \frac{1}{2} * \frac{1}{3} = \frac{1}{48}$ $v_{\text{VB}}(3) = v_{\text{VB}}(2) a_{\text{VB,VB}} * b_{\text{VB,fly}}$ $= \frac{1}{12} * \frac{1}{4} * \frac{1}{3} = \frac{1}{144}$ $\psi_{\text{VB}}(3) = \text{NN}$	

	t:	0	1	2	3	4
j	τ_k:	<s>	bats	can	fly	</s>
0		1				$\alpha_{\text{<s>}}(4) = \alpha_{\text{NN}}(3) * a_{\text{NN},\text{<s>}} * b$ $+ \alpha_{\text{VB}}(3) * a_{\text{VB},\text{<s>}} * b$ $= \frac{35}{1152} * \frac{1}{4} * 1 + \frac{7}{216} * \frac{1}{2} * 1$ $= \frac{35}{4608} + \frac{7}{432} = \frac{105+224}{13{,}824}$ $= 0.023799$
1	NN	0	$b_{\text{NN,bats}} = 0$	$\alpha_{\text{NN}}(2) = \alpha_{\text{NNS}}(1) * a_{\text{NNS,NN}} * b_{\text{NN,can}}$ $+ \alpha_{\text{VB}}(1) * a_{\text{VB,NN}} * b_{\text{NN,can}}$ $= \frac{1}{2} * \frac{1}{2} * \frac{1}{2} + \frac{1}{6} * \frac{1}{4} * \frac{1}{2}$ $= \frac{1}{8} + \frac{1}{48} = \frac{7}{48}$	$\alpha_{\text{NN}}(3) = \alpha_{\text{NN}}(2) * a_{\text{NN,NN}} * b_{\text{NN,fly}}$ $+ \alpha_{\text{VB}}(2) * a_{\text{VB,NN}} * b_{\text{NN,fly}}$ $= \frac{7}{48} * \frac{1}{4} * \frac{1}{2} + \frac{7}{72} * \frac{1}{4} * \frac{1}{2}$ $= \frac{7}{384} + \frac{7}{576} = \frac{35}{1152}$	
2	NNS	0	$\alpha_{\text{NNS}}(1) = \alpha_{\text{<s>}}(0) * a_{\text{<s>},\text{NNS}} * b_{\text{NNS,bats}}$ $= 1 * \frac{1}{2} * 1 = \frac{1}{2}$	$b_{\text{NNS,can}} = 0$	$b_{\text{NNS,fly}} = 0$	
3	VB	0	$\alpha_{\text{VB}}(1) = \alpha_{\text{<s>}}(0) * a_{\text{<s>},\text{VB}} * b_{\text{VB,bats}}$ $= \frac{1}{3} * \frac{1}{2} = \frac{1}{6}$	$\alpha_{\text{VB}}(2) = \alpha_{\text{NNS}}(1) a_{\text{NNS,VB}} * b_{\text{VB,can}}$ $+ \alpha_{\text{VB}}(1) a_{\text{VB,VB}} * b_{\text{VB,can}}$ $= \frac{1}{2} * \frac{1}{2} * \frac{1}{3} + \frac{1}{6} * \frac{1}{4} * \frac{1}{3}$ $= \frac{1}{12} + \frac{1}{72} = \frac{7}{72}$	$\alpha_{\text{VB}}(3) = \alpha_{\text{NN}}(2) a_{\text{NN,VB}} * b_{\text{VB,fly}}$ $+ \alpha_{\text{VB}}(2) a_{\text{VB,VB}} * b_{\text{VB,fly}}$ $= \frac{7}{48} * \frac{1}{2} * \frac{1}{3} + \frac{7}{72} * \frac{1}{4} * \frac{1}{3}$ $= \frac{7}{288} + \frac{7}{864} = \frac{7}{216}$	

Extra Credit. (5 points) Design and write up (including solutions) your own test question for a future Computational Linguistics I course.