Questions for Flipped Classroom Session of COMS 4705 Week 8, Fall 2014. (Michael Collins)

Question 1 This question concerns training IBM Model 2 for statistical machine translation. Assume that we have a bilingual corpus of English sentences $\mathbf{e}$ paired with sentences in a foreign language $\mathbf{f}$. Each English sentence is a string formed from the words $\{the, dog, ate, cat, a, banana\}$, while each foreign sentence is a string formed from the words $\{athe, adog, aate, acat, aa, abanana\}$. The set of training examples is as follows:

Training Example	English sentence $\mathbf{e}$	Foreign Sentence $\mathbf{f}$
1	*the dog ate*	*aate adog athe*
2	*the cat ate the banana*	*abanana athe aate acat athe*
3	*a dog ate a cat*	*acat aa aate adog aa*
4	*a cat ate*	*aate acat aa*

Recall that in IBM model 2 we have translation parameters of the form $t(f|e)$ where f is a foreign word, and e is an English word; and in addition we have alignment parameters of the form $q(j|il, m)$ where l is the length of the English sentence, m is the length of the foreign sentence, and this parameter denotes the probability of word i in the foreign string being aligned to word j in the English string.

Question 1a: Specify parameter values for IBM model 2 that result in $p(\mathbf{f}|\mathbf{e}) = 1$ for all training examples shown in the table above.

Question 1b: Now assume that the training set is as follows:

Training Example	English sentence $\mathbf{e}$	Foreign Sentence $\mathbf{f}$
1	*the dog ate*	*athe adog aate*
2	*the cat ate the banana*	*athe acat athe abanana aate*
3	*a dog ate a cat*	*aa adog aa acat aate*
4	*a cat ate*	*aa acat aate*

Can you define IBM Model 2 parameters such that $p(\mathbf{f}|\mathbf{e}) = 1$ for all examples in this training set? If your answer is yes, define the parameters of the model. If your answer is no, give a justification for your answer in 200 words or less.

Question 2 Consider an instance of IBM Model 2 with the following parameters:

- $\mathcal{E} = \{\text{the, dog}\}$
- $\mathcal{F} = \{\text{le, chien}\}$
- $q(1|1,2,2) = 0.7$, $q(2|1,2,2) = 0.3$, $q(1|2,2,2) = 0.4$, and $q(2|2,2,2) = 0.6$. (Recall that each q parameter is of the form $q(j|i,l,m)$ where j is the English position, i is the French position, l is the English sentence length, and m is the French sentence length.)
- $t(\text{le}|\text{the}) = 0.9$, $t(\text{chien}|\text{the}) = 0.1$, $t(\text{le}|\text{dog}) = 0.2$, $t(\text{chien}|\text{dog}) = 0.8$.

Question 2a: Now consider an instance where the English sentence $e =$ the dog, and the French sentence has length $m = 2$. Write down the probability for $p(f|e, m = 2)$ for the possible French sentences *le le*, *le chien*, *chien le*, *chien chien*.

Hint: Note that we have the identity

$$\begin{aligned} & \sum_{a_1=1}^{2} \sum_{a_2=1}^{2} \prod_{j=1}^{2} t(f_j|e_{a_j}) q(a_j|j,l,m) \\ = & \left(\sum_{a_1=1}^{2} t(f_1|e_{a_1}) q(a_1|1,l,m) \right) \times \left(\sum_{a_2=1}^{2} t(f_2|e_{a_2}) q(a_2|2,l,m) \right) \end{aligned}$$

Question 2b: Now say we have the English sentence $e =$ *the dog* paired with the French sentence $f =$ *le chien* in the training data. What is the value for

$$p(A_1 = 1|e, f, m = 2)$$

in this case?

Question 2c: Prove that the following identity holds:

$$\begin{aligned} & \sum_{a_1=0}^{l} \sum_{a_2=0}^{l} \ldots \sum_{a_m=0}^{l} \prod_{j=1}^{m} t(f_j|e_{a_j}) q(a_j|j,l,m) \\ = & \left(\sum_{a_1=0}^{l} t(f_1|e_{a_1}) q(a_1|1,l,m) \right) \\ & \times \left(\sum_{a_2=0}^{l} t(f_2|e_{a_2}) q(a_2|2,l,m) \right) \\ & \ldots \\ & \times \left(\sum_{a_m=0}^{l} t(f_m|e_{a_m}) q(a_m|2,l,m) \right) \end{aligned}$$

Why is this identity computationally useful?

Question 3 We will define a new model for statistical machine translation as follows:

$$p(f, a|e, m) = \prod_{j=1}^{m} t(f_j|e_{a_j}) d(a_j|a_{j-1}, l, m)$$

Here f is a French sequence of words $f_1, f_2, \ldots f_m$, a is a sequence of alignment variables $a_1, a_2, \ldots a_m$, and e is an English sequence of words $e_1, e_2, \ldots e_l$. Note that the probability $p(f, a|e, m)$ is conditioned on the identity of the English sentence, e, as well as the length of the French sentence, m. The parameters of the model are translation parameters of the form $t(f|e)$ and alignment parameters of the form $d(a_j|a_{j-1}, l, m)$. We assume that a_0 is defined to be 0. Note that in contrast to IBM Model 2, the alignment parameters are now modified to be conditioned upon the previous alignment variable.

Assume we have $m = 6$, and $e =$ *the dog ate the cat*. The model then defines a distribution $P_{M3}(f, a|$*the dog ate the cat*$, 6)$, as well as a distribution over French strings,

$$p(f|e, m) = \sum_{a} p(f, a|e, m)$$

Define an HMM which defines a distribution $p_{\text{HMM}}(f)$ over French strings which is identical to the distribution $p(f|e, m)$. Your HMM should have N states, a set of output symbols Σ that is the set of all possible French words, and parameters of the following types:

- $q(j|*)$ for $j = 1 \ldots N$ is the probability of choosing state j as the initial state.
- $q(k|j)$ for $j = 1, 2, \ldots, N$ and $k = 1, 2, \ldots, N$ is the probability of transitioning from state j to state k.
- $e(o|j)$ for $j = 1, 2, \ldots, N$ and $o \in \Sigma$ is the probability of emitting symbol o from state j.

This HMM will be slightly different from the HMMs seen in class, in that there is *no stop symbol*, and the HMM is used to define a distribution over state/symbol sequences of length 6 only. For example, the state sequence $\langle 1, 2, 1, 3, 4, 2 \rangle$ paired with the output symbol sequence *le chat aime le chien bleu* would have probability

$$q(1|*)q(2|1)q(1|2)q(3|1)q(4|3)q(2|4)$$

$$\times e(le|1)e(chat|2)e(aime|1)e(le|3)e(chien|4)e(bleu|2)$$

You should describe how all of the parameters in the HMM model can be written in terms of the translation model parameters $t(f|e)$ and $d(a_j|a_{j-1}, l, m)$. Hint: your HMM should have 5 states, where the states correspond to the 5 positions in the English string *the dog ate the cat*.