Ling 5801: Problem Set 4

Due via Carmen dropbox at 11:59 PM 11/4.

1. [10 pts.] Draw a recognizer chart (as in Section 10.3 of Lecture Notes 10 on context-free grammar recognition) for the sentence:

   'the clown intentionally fell down twice,'

   using the following grammar:

   S → NP VP
   NP → the clown
   VP → intentionally VP
   VP → fell down
   VP → VP twice

2. The probability of the sentence 'the cat hit the toy off the mat' is defined for a probabilistic grammar in Section 11.4 of Lecture Notes 11 on probabilistic parsing.

   (a) [5 pts.] Write an equation for the probability of the following tree in terms of specific grammar rule probabilities from this grammar (see Section 11.7 of Lecture Notes 11), e.g.: $P(\tau) = P(S \to NP\ VP \mid S) \cdot P(NP \to \text{the cat} \mid NP) \cdot \ldots$:
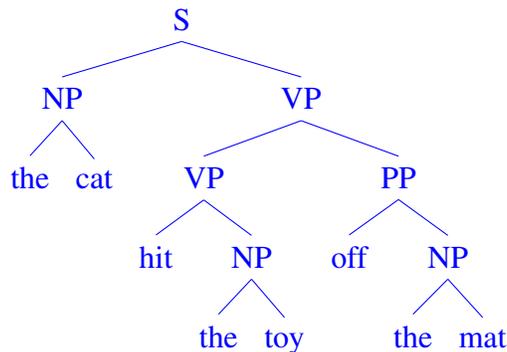
   (b) [5 pts.] Write an equation for the probability of the **sentence** 'the cat hit the toy off the mat' as a sum of products of specific grammar rule probabilities from this grammar, e.g.: $P(X_{1..T} = \text{the cat hit the toy off the mat}) = P(S \to NP\ VP \mid S) \cdot P(NP \to \text{the cat} \mid NP) \cdot \ldots + \ldots$. Remember that a sentence may have more than one possible phrase structure.

   (c) [5 pts.] Give an equation for the probability of the above sentence that uses the distributive axiom to reduce the number of multiplication and addition operations in the equation.

3. Write an equation defining each of the following:

   (a) [2 pts.] Write an equation defining a distribution $P(x \wedge y)$ for any values x and y of random variables X and Y in terms of a set of frequencies $F(x \wedge y)$ for all combinations of values of random variables X and Y. Your equation should be as short as possible.

(b) [2 pts.] Write an equation defining a distribution $P(x \mid y)$ for any values x and y of random variables X and Y in terms of a set of frequencies $F(x \wedge y)$ for all combinations of values of random variables X and Y. Your equation should be as short as possible.

(c) [3 pts.] Suppose you have:

   i. the probability in some data that people are or are not on the telephone,

   ii. the probability in some data that people speak slowly or quickly (above or below some threshold) given that they are on or not on the telephone, and

   iii. the probability in some data that people pronounce the /t/ sound (e.g. in latter) as [t] or as something like [d] given that they speak quickly or slowly, on or not on the telephone.

Write an equation that uses these probabilities to calculate the estimated probability that people will pronounce /t/ as [d] and be on the telephone, when the speech rate is unknown.

(d) [3 pts.] In the previous subproblem, how could you simplify this calculation using marginalization, and thus get probability estimates based on larger amounts of data, if you think that the pronunciation of /t/ just depends on the speech rate, but not on whether the speaker is or is not on the telephone?

4. [10 pts.] PROGRAMMING: Using the syntax described in the lecture notes, write a program called 'Q4.py' to read models for variables A and B from standard input in the following format, using the Model and CondModel classes from the model.py module described in the lecture notes on probability models (and posted individually on the web site):

```
A : on = .7
A : off = .3
B on : one = .1
B on : two = .9
B off : one = .3
B off : two = .3
B off : three = .4
```

and use these to calculate $P(b)$ for each modeled value of B using the following equation:

$$P(b) = \sum_{a \in A} P(a) \cdot P(b \mid a)$$

in the following format:

```
B : one = .16
B : two = .72
⋮
```

5. [10 pts.] PROGRAMMING: Using the syntax described in the lecture notes, write a program called 'Q5.py' to read in models for speaking on a phone $P(p)$, speech rate $P(r \mid p)$, and pronunciation of /t/ $P(t \mid p, r)$, from standard input in the following format:

```
P : on = .7
P : off = .3

⋮

R on : fast = .6
R on : slow = .4
R off : fast = .5
R off : slow = .5

⋮

T on fast : [t] = .9
T on fast : [d] = .1
T on slow : [t] = .7
T on slow : [d] = .3
T off fast : [t] = .3
T off slow : [t] = .8

⋮
```

and use them to calculate and print a probability model for $P(r \mid t)$ in the format:

```
R [t] : fast = .??
R [t] : slow = .??

⋮
```

Your program should be as short as possible. You may want to use the `Model` and `CondModel` classes from the `model.py` module described in the lecture notes on probability models (and posted individually on the web site), to help you read and write in this format. Hand in all inputs and outputs.