LI 569 — Intro to Computational Linguistics
Assignment 2: Dyna, Parsing, Bayes

Prof. J. Eisner — Summer 2013
(Due date: Saturday, July 6: but please do problems 1–2 before Friday's 1:30 class)

General homework policies for this class: Same as on assignment 1. In particular, you may work in groups of 2 or 3. When handing in a solution, please clearly indicate all of its authors. Please upload your solution to CTools as a PDF file.

Objectives of this assignment: (1) Prepare for the upcoming parsing lecture on Friday, July 5. (2) Practice Bayes' Theorem.

1. As Darcey previously announced on Piazza, you should work through her new Dyna tutorial at `http://cs.jhu.edu/~darcey/dyna-tutorial.pdf`. This focuses on counting unigrams and bigrams in a corpus. She may still add one more relevant section, in which case she'll announce that on Piazza.

   You should hand in answers to at least 3 of the exercises in the tutorial (more if you can).

   Make sure to do the tutorial before Friday's lecture, where we'll use Dyna notation for something much more ambitious: parsing! And the next homework will ask you to use Dyna to play with parsing algorithms.

   Note that this implementation of Dyna is a *very* new and incomplete prototype—Wes and Tim are continuing to improve it even as you use it. Thanks for helping us test it!

2. Attempt the following puzzles from the North American Computational Linguistic Olympiad (a high school contest that does not require any prior background).

   - **One, Two, Tree**: `http://cs.jhu.edu/~jason/licl/hw2/onetwotree.pdf`
   - **Twodee**: `http://cs.jhu.edu/~jason/licl/hw2/twodee.pdf`

   These puzzles should warm you up for Friday's lecture on parsing—they'll give you a sense of what parsing is trying to do and how it might go about it.

   Since the puzzles come from a timed contest, don't spend too long mulling them over—do what you can and move on. Mostly they'are about seeing how a context-free grammar restricts the set of parses and yet leaves some ambiguity.

   You should be able to get most parts. However, parts R5, R8, R9, and H4 are tricky "math team" type puzzles that ask you to quantify the *amount* of ambiguity—few of the contestants solved these parts completely, so kudos to you if you can. (We'll post the answers later so you can learn from them.)

   For the Twodee puzzle, you may find it easier to print it out and write directly on the printout. You can hand in the printout in person at Friday's class.

3. Beavers can make three cries, which they use to communicate. `bwa` and `bwee` usually mean something like “come” and “go” respectively, and are used during dam maintenance. `kiki` means “watch out!” The following **conditional probability table** shows the probability of the various cries in different situations.[1]

$p(cry \mid situation)$	Predator!	Timber!	I need help!
`bwa`	0	0.1	0.8
`bwee`	0	0.6	0.1
`kiki`	1.0	0.3	0.1

(a) Notice that each column of the above table sums to 1. Write an equation stating this, in the form $\sum_{variable} p(\cdots) = 1$.

(b) A certain colony of beavers has already cut down all the trees around their dam. As there are no more to chew, $p(timber) = 0$. Getting rid of the trees has also reduced $p(predator)$ to 0.2. These facts are shown in the following **joint probability table**. Fill in the rest of the table, using the previous table and the laws of probability. (Note that the meaning of each table is given in its top left cell.)

$p(cry, situation)$	Predator!	Timber!	I need help!	TOTAL
`bwa`				
`bwee`				
`kiki`				
TOTAL	0.2	0		

(c) A beaver in this colony cries `kiki`. Given this cry, other beavers try to figure out the probability that there is a predator.

i. This probability is written as: $p(\underline{\qquad\qquad})$

ii. It can be rewritten without the $\mid$ symbol as: $\underline{\qquad\qquad}$

iii. Using the above tables, its value is: $\underline{\qquad\qquad}$

iv. Alternatively, Bayes’ Theorem allows you to express this probability as:

$$\frac{p(\underline{\qquad\quad})\cdot p(\underline{\qquad\quad})}{p(\underline{\qquad\quad})\cdot p(\underline{\qquad\quad}) + p(\underline{\qquad\quad})\cdot p(\underline{\qquad\quad}) + p(\underline{\qquad\quad})\cdot p(\underline{\qquad\quad})}$$

v. Using the above tables, the value of this is:

$$\frac{\underline{\qquad\quad}\cdot\underline{\qquad\quad}}{\underline{\qquad\quad}\cdot\underline{\qquad\quad}+\underline{\qquad\quad}\cdot\underline{\qquad\quad}+\underline{\qquad\quad}\cdot\underline{\qquad\quad}}$$

This should give the same result as in part iii., and it should be clear that they are really the same computation—by constructing table (b) and doing part iii., you were *implicitly* using Bayes’ Theorem. (I told you it was a trivial theorem!)

[1] I’m sorry to confess I made all of this up, but there are people who do real research on animal communication.

4. All cars are either red or blue. The witness claimed the car that hit the pedestrian was blue. Witnesses are believed to be about 80% reliable in reporting car color (regardless of the actual car color). But only 10% of all cars are blue.

   (a) Write an equation relating the following quantities and perhaps other quantities:

   $$p(\mathit{Actual} = \text{blue})$$
   $$p(\mathit{Actual} = \text{blue} \mid \mathit{Claimed} = \text{blue})$$
   $$p(\mathit{Claimed} = \text{blue} \mid \mathit{Actual} = \text{blue})$$

   Reminder: Here, *Claimed* and *Actual* are *random variables*, which means that they are functions over some outcome space. For example, the probability that $\mathit{Claimed} = \text{blue}$ really means the probability of getting an outcome x such that $\mathit{Claimed}(x) = \text{blue}$. We are implicitly assuming that the space of outcomes x is something like the set of witnessed car accidents.

   (b) Match the three probabilities above with the following terms: *prior probability*, *likelihood of the evidence*, *posterior probability*.

   (c) Give the values of all three probabilities. (Hint: Use Bayes' Theorem.) Which probability should the judge care about?

   (d) Let's suppose the numbers 80% and 10% are specific to Michigan. So in the previous problem, you were implicitly using the following more general version of Bayes' Theorem:

   $$p(A \mid B, Y) = \frac{p(B \mid A, Y) \cdot p(A \mid Y)}{p(B \mid Y)}$$

   where Y is $\mathit{city} = \text{Michigan}$. Just as 3f generalized 3d on Homework 1, by adding a "background" condition Y, this version generalizes Bayes' Theorem. Carefully prove it.

   (e) Now prove the more detailed version

   $$p(A \mid B, Y) = \frac{p(B \mid A, Y) \cdot p(A \mid Y)}{p(B \mid A, Y) \cdot p(A \mid Y) + p(B \mid \bar{A}, Y) \cdot p(\bar{A} \mid Y)}$$

   which gives a practical way of finding the denominator in the question 4d.

   (f) Write out the equation given in question 4e with A, B, and Y replaced by specific propositions from the red-and-blue car problem. For example, Y is "$\mathit{State} = \text{Michigan}$" (or just "Michigan" for short). Now replace the probabilities with actual numbers from the problem, such as 0.8.

   Writing out a real case of this important formula is straightforward, but may be good for you.