LI 569 — Intro to Computational Linguistics
Assignment 1: Probability Exercises

Prof. J. Eisner — Summer 2013
Due date: Sunday, June 30

General homework policies for this class:

- For all assignments, unless otherwise stated, you may work in groups of 2 or 3. Indeed, you'll probably learn better that way! When handing in a solution, please clearly indicate all of its authors.
- We expect you to do all the homework, but ultimately, you are here at LI for your own sake. Homework is how you will explore, extend, and internalize the ideas from lecture. Our grading may be somewhat holistic: that is, we will be looking for thoughtful engagement with the questions, but we won't be obsessing over points for each problem. So you are free to spend more time on the questions that are more meaningful to you.
- For more detailed feedback and discussion of your answers, try the TA's office hours. She may also post some general comments or clarifications after each assignment.
- We'll post instructions soon about how to submit your homework. Probably we will ask you to write up your answers using your favorite text editor or word processor, convert to PDF, and upload the PDF file to your CTools Dropbox.
- You may need some way of typing mathematical symbols. There are fancy solutions like TEX or the Insert Equation feature in MS Word. But it's also okay to use a more casual plain-text notation as suggested in Figure 1.

Objective of this assignment: Become comfortable working with conditional probability distributions. First you will do a hands-on activity to learn a new technique for modeling these distributions. Then you'll do some pencil-and-paper exercises on topics from the first lecture, like the probability axioms and how to use the chain rule to multiply probabilities together.

---

1. In our first class, we talked about estimating probabilities like

$$p(\text{Revere wins} \mid \text{weather's clear}, \text{ground is dry}, \text{jockey getting over sprain}, \text{Epitaph also in race}, \text{Epitaph was recently bought by Gonzalez}, \text{race is on May 17}, \ldots)$$

   We discussed the difficulty of taking all these conditions into account. We can't simply measure Revere's win rate among just those past races *that have all these conditions*. Such an estimate of the probability would be unreliable because the number of such past races is so small (perhaps 0).

So we considered "backing off" by throwing away some conditions. Yet as you may have imagined, a good handicapper doesn't do that! He or she *does* pay attention to all these conditions when estimating a horse's odds. In Richard Russo's excellent novel *The Risk Pool* (1988), the narrator remarks in passing:

> Courses in handicapping should be required, like composition and Western civilization, in our universities. For sheer complexity, there's nothing like a horse race, excepting life itself ... To weigh and evaluate a vast grid of information, much of it meaningless, and to arrive at sensible, if erroneous, conclusions, is a skill not to be sneezed at.

So what beats computing simple ratios of backed-off counts? Answer: Conditional log-linear models are hugely useful and we will draw on them throughout the class. Prepare to reconfigure your brain! Go play with the online toy at

http://cs.jhu.edu/~jason/465/hw-prob/loglin/

It should give you almost *physical* intutions about how conditional log-linear models behave.

This interactive visualization takes you through a series of lessons, including a supplementary reading. We expect this may take a few hours. Again, we recommend doing it with a friend or two. Feel free to post questions on Piazza. Here is what you should hand in:

(a) The lessons are peppered with guidance in the form of questions. Write down your answers to the most interesting questions (you get to decide which ones). For grading purposes we just want to see thoughtful engagement with the material. Probably writing out answers to 1–2 questions per lesson is about right.

(b) Now, how would you use this stuff in linguistics? Think over your own interests in linguistics or NLP. Give an example of a conditional distribution $p(y \mid x)$ that would be interesting to model. Your goal here might be to predict y from x, or to understand the properties of x that are predictive of y.

We plan to post a summary of the class's answers. So make sure to be clear about your problem and the model you'd build. What do x and y represent? What features $\vec{f}(x, y)$ would you use in your model?

(c) To satisfy our curiosity, tell us how long you spent on each of the following:

i. reading the handout on log-linear modeling (linked to from lesson 1)
ii. working through the lessons
iii. writing down your answers
iv. describing a problem where you can apply this technique

Frank Ferraro and I would also welcome any other feedback on the visualization. We just built it this fall and we're giving a talk about it in August. It has been used once by CS students, but not previously by linguistics students.

2. A **language model** is a probability function p that assigns probabilities to word sequences such as $\vec{w} = (\texttt{i}, \texttt{love}, \texttt{new}, \texttt{york})$.

Suppose $\vec{w} = w_1 w_2 \cdots w_n$ (a sequence of n words). As discussed in class, a **trigram language model** defines

$$p(\vec{w}) \stackrel{\text{def}}{=} p(w_1) \cdot p(w_2 \mid w_1) \cdot p(w_3 \mid w_1, w_2) \cdot p(w_4 \mid w_2, w_3) \cdots p(w_n \mid w_{n-2}, w_{n-1}) \quad (1)$$

	Text	Picts	LaTeX
$p(x \mid y)$	`p(x | y)`	`p(x | y)`	`p(x \mid y)`
$\neg x$	`NOT x`	`~x`	`\neg x`
$\bar{x}$ (set complement)	`COMPL(x)`	`\x`	`\bar{x}`
$x \subseteq y$	`x SUBSET y`	`x {= y`	`x \subseteq y`
$x \supseteq y$	`x SUPERSET y`	`x }= y`	`x \supseteq y`
$x \cup y$	`x UNION y`	`x U y`	`x \cup y`
$x \cap y$	`x INTERSECT y`	`x ^ y`	`x \cap y`
$x \geq y$	`x GREATEREQ y`	`x >= y`	`x \geq y`
$x \leq y$	`x LESSEQ y`	`x <= y`	`x \leq y`
$\emptyset$ (empty set)	`NULL`	`0`	`\emptyset`
$\mathcal{E}$ (event space)	`E`	`E`	`E`

Figure 1: Each column shows a different style of writing mathematical symbols.

on the assumption that the sequence was generated in the order $w_1, w_2, w_3 \ldots$ ("from left to right") with each word chosen in a way dependent only on the previous two words.[1]

(a) Expand the above definition of $p(\vec{w})$ using naive estimates of the parameters, such as

$$p(w_4 \mid w_2, w_3) \stackrel{\text{def}}{=} \frac{c(w_2 w_3 w_4)}{c(w_2 w_3)}$$

where $c(w_2 w_3 w_4)$ denotes the count of times the trigram $w_2 w_3 w_4$ was observed in a training corpus.

Remark: Naive parameter estimates of this sort are called **maximum-likelihood estimates** (MLE). They have the advantage that they maximize the probability (equivalently, minimize the perplexity) of the training data. But they will generally perform badly on test data, unless the training data were so abundant as to include all possible trigrams many times.

Hint: You will have to think about $p(w_1)$. It says that the first word w_1 was simply generated from a unigram model, conditioned on 0 words of context. Similarly, $p(w_2 \mid w_1)$ indicates that w_2 was generated from a bigram model, conditioned on only 1 word of context (namely w_1).

Remark: As a result, this setup is slightly different from the trigram model we discussed in class. Equation (1) doesn't model w_1 as the first word of a sentence. w_1 is just the first word you heard when you turned on the radio—the sentence might have started earlier, so w_1 isn't necessarily a word of the sort that starts sentences. Equation (1) also doesn't model w_n as the last word of a sentence. If $n = 10$ (chosen in advance), then w_n is just the tenth word you heard—the sentence might continue after that, so w_n isn't necessarily a word of the sort that ends sentences.

(b) One could also define a kind of reversed trigram language model $p_{reversed}$ that instead assumed the words were generated in reverse order ("from right to left"):

$$\begin{aligned} p_{reversed}(\vec{w}) \stackrel{\text{def}}{=} & \; p(w_n) \cdot p(w_{n-1} \mid w_n) \cdot p(w_{n-2} \mid w_{n-1}, w_n) \cdot p(w_{n-3} \mid w_{n-2}, w_{n-1}) \\ & \cdots p(w_2 \mid w_3, w_4) \cdot p(w_1 \mid w_2, w_3) \end{aligned} \tag{2}$$

[1] This is the "second-order Markov" assumption. It is an example of a conditional independence assumption: w_i is conditionally independent of $w_{i-3}, w_{i-4}, \ldots$ given w_{i-1}, w_{i-2}.

By manipulating the notation, show that the two models are identical (i.e., $p(\vec{w}) = p_{reversed}(\vec{w})$ for any $\vec{w}$) provided that both models use MLE parameters estimated from the same training data (see problem 2a).

(c) Suppose that our data include sentence delimiters: sentences are delimited by the special word BOS at the start and the special word EOS at the end. For example, in the following sequence $\vec{w}$, we have $w_1 =$ BOS and $w_{16} =$ EOS and $\vec{w}$ conists of 3 complete sentences.

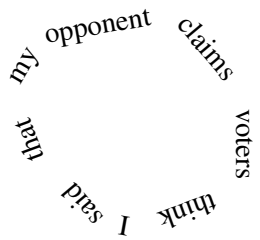
BOS `do you think so` EOS BOS `yes` EOS BOS `at least i thought so` EOS

The probability of

BOS `do you think the` EOS

should be extremely low under any good language model. Why? In the case of the trigram model, which parameter or parameters are responsible for making this probability low?

(d) Some politicians seem to speak in circles. Instead of uttering strings with a start and an end, they blow round linguistic objects out of their mouths, like smoke rings:



I wondered, how would we build a language model to describe the probability of such objects? Well, problems like this show up in computer vision, since images don't have a natural start and end either. Consider the similar problem of modeling the colors of these 4 pixels, using bigrams of adjacent pixel colors:

A	B
D	C

A well-known computer vision paper by J. Besag (1974) proposed approximating $p(A, B, C, D)$ by a product $p(A \mid B) \cdot p(B \mid C) \cdot p(C \mid D) \cdot p(D \mid A)$.

So here's the question. Can Besag's approximation be justified by using the chain rule plus backoff? If so, show how. If not, fix it as best you can. Discuss.

3. The following should strengthen your understanding of the basic properties of probability. But if this type of math is unfamiliar to you, don't freak out or waste a lot of time struggling (unless you want to). Just skip this problem, or do what you can with help from friends.

Let $\mathcal{E} \neq \emptyset$ denote the event space (it's just a set, also known as the outcome space or sample space). Let p be a function that assigns a real number in $[0, 1]$ to any subset of $\mathcal{E}$. This number is called the probability of the subset.

You are told that p satisfies the following two axioms:

- $p(\mathcal{E}) = 1$.
- $p(X \cup Y) = p(X) + p(Y)$ provided that $X \cap Y = \emptyset$.[2]

[2] In fact, probability functions p are also required to satisfy a generalization of this second axiom: if $X_1, X_2, X_3, \ldots$ is an infinite sequence of disjoint sets, then $p(\bigcup_{i=1}^{\infty} X_i) = \sum_{i=1}^{\infty} p(X_i)$. But you don't need this for this assignment.

As a matter of notation, remember that the **conditional probability** $p(X \mid Z) \stackrel{\text{def}}{=} \frac{p(X \cap Z)}{p(Z)}$. For example, singing in the rain is one of my favorite rainy-day activities: so my ratio $p(\text{singing} \mid \text{rainy}) = \frac{p(\text{singing AND rainy})}{p(\text{rainy})}$ is high. Here the predicate "singing" picks out the set of singing events in $\mathcal{E}$, "rainy" picks out the set of rainy events, and the conjoined predicate "singing AND rainy" picks out the intersection of these two sets—that is, all events that are both singing AND rainy.

(a) In class, we said that if $Y \subseteq Z$, then $p(Y) \leq p(Z)$.
Prove that this actually follows from the axioms above. You may use any and all set manipulations you like. Remember that $p(A) = 0$ does not imply that $A = \emptyset$ (why not?), and similarly, that $p(B) = p(C)$ does not imply that $B = C$ (even if $B \subseteq C$).

(b) Use the above fact to prove that conditional probabilities $p(X \mid Z)$, just like ordinary probabilities, always fall in the range $[0, 1]$.

(c) In class, we said that $p(\emptyset) = 0$. Again, prove that this actually follows from the axioms above.

(d) Let $\bar{X}$ denote $\mathcal{E} - X$. Prove from the axioms that $p(X) = 1 - p(\bar{X})$. For example, $p(\text{singing}) = 1 - p(\text{NOT singing})$.

(e) Prove from the axioms that $p(\text{singing AND rainy} \mid \text{rainy}) = p(\text{singing} \mid \text{rainy})$.

(f) Prove from the axioms that $p(X \mid Y) = 1 - p(\bar{X} \mid Y)$. For example, $p(\text{singing} \mid \text{rainy}) = 1 - p(\text{NOT singing} \mid \text{rainy})$. This is a generalization of 3d.

(g) Simplify: $\big(p(X \mid Y) \cdot p(Y) + p(X \mid \bar{Y}) \cdot p(\bar{Y})\big) \cdot p(\bar{Z} \mid X)/p(\bar{Z})$

(h) Under what conditions is it true that $p(\text{singing OR rainy}) = p(\text{singing}) + p(\text{rainy})$?

(i) Under what conditions is it true that $p(\text{singing AND rainy}) = p(\text{singing}) \cdot p(\text{rainy})$?

(j) Suppose you know that $p(X \mid Y) = 0$. Prove that $p(X \mid Y, Z) = 0$.[3]

(k) Suppose you know that $p(W \mid Y) = 1$. Prove that $p(W \mid Y, Z) = 1$.

4.

(a)	$p(\neg\text{shoe} \mid \neg\text{nail}) = 1$	*For want of a nail the shoe was lost,*
(b)	$p(\neg\text{horse} \mid \neg\text{shoe}) = 1$	*For want of a shoe the horse was lost,*
(c)	$p(\neg\text{race} \mid \neg\text{horse}) = 1$	*For want of a horse the race was lost,*
(d)	$p(\neg\text{fortune} \mid \neg\text{race}) = 1$	*For want of a race the fortune was lost,*
(e)	$p(\neg\text{fortune} \mid \neg\text{nail}) = 1$	*And all for the want of a horseshoe nail.*

Show carefully that (e) follows from (a)–(d). *Hint:* Consider

$$p(\neg\text{fortune}, \neg\text{race}, \neg\text{horse}, \neg\text{shoe} \mid \neg\text{nail}),$$

as well as the "chain rule" from class and the facts proved in problems 3a, 3b, and 3k.

Note: The $\neg$ symbol denotes the boolean operator NOT.

Note: This problem is supposed to convince you that logic is just a special case of probability theory. An excellent and wide-ranging book developing this theme, by the influential statistician E. T. Jaynes, is *Probability Theory: The Logic of Science.* See http://bayes.wustl.edu/ for more readings.

Note: Be glad I didn't ask you to prove the correct operation of Figure 2!

[3] More precisely, $p(X \mid Y, Z)$ could be either 0 or undefined, namely 0/0. (There do exist advanced ways to redefine conditional probability to avoid this 0/0 problem. Even then, though, one may want a probability measure p to leave some probabilities or conditional probabilities undefined. This turns out to be important for reasons beyond the scope of this course: e.g. http://en.wikipedia.org/wiki/Vitali_set.)

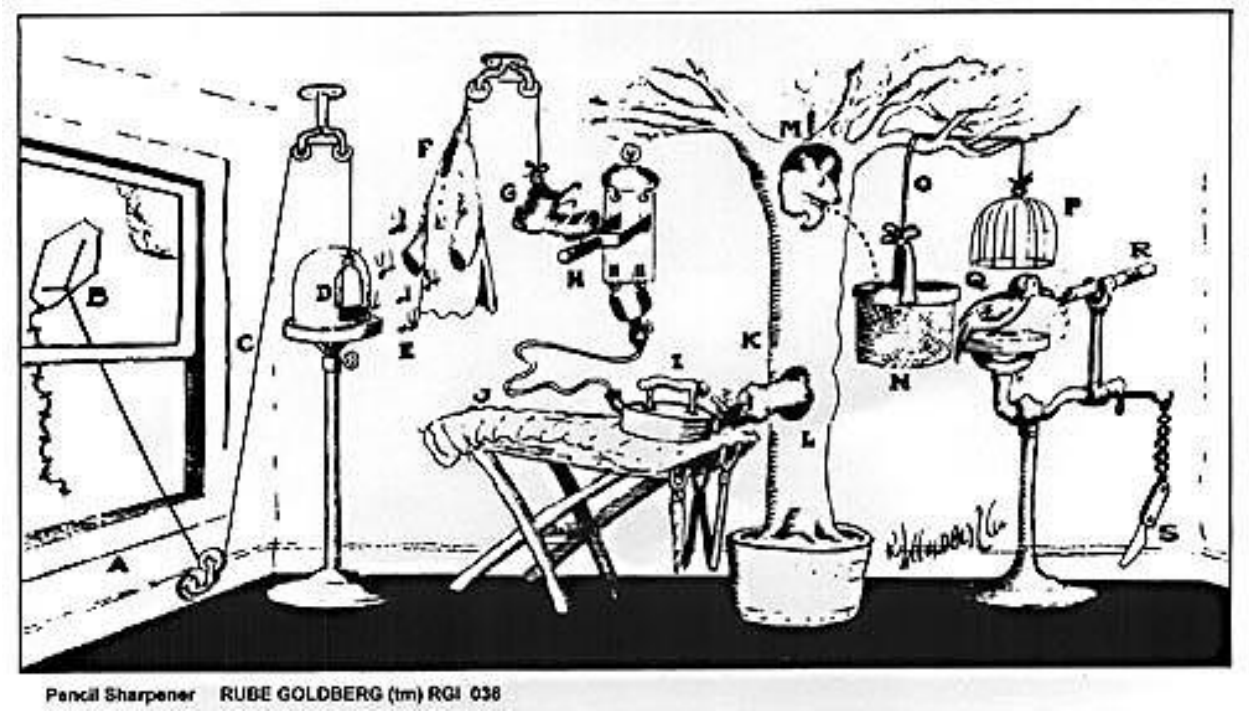


Figure 2: Rube Goldberg gets his think-tank working and evolves the simplified pencil-sharpener. Open window (A) and fly kite (B). String (C) lifts small door (D) allowing moths (E) to escape and eat red flannel shirt (F). As weight of shirt becomes less, shoe (G) steps on switch (H) which heats electric iron (I) and burns hole in pants (J). Smoke (K) enters hole in tree (L), smoking out opossum (M) which jumps into basket (N), pulling rope (O) and lifting cage (P), allowing woodpecker (Q) to chew wood from pencil (R), exposing lead. Emergency knife (S) is always handy in case opossum or the woodpecker gets sick and can't work.

5. [Added later via Piazza] Friday's grammar writing lab takes the form of a friendly competition among small teams (who are attempting to do the impossible, with diabolical rules of the game). I've run this game several times before and it's always a hit. It will teach you a lot about context-free grammars, hidden Markov models, probabilities, cross-entropy, grammaticality judgments, and more!

   To prepare:

   (a) Make sure that you bring your laptop, and that you are able to log into your Linux account and do basic stuff like editing files. Note that you may have to install a program on your laptop. *Please get all this working in advance—you won't want to waste time doing it during the competition.*

   (b) "Orwellspeak" is an entertaining little puzzle[4] that you can find at `http://cs.jhu.edu/~jason/licl/hw1/orwellspeak.pdf`. Part 1 serves as good warmup for our grammar writing competition, and Part 2 relates that to the n-gram models from today's lecture. Doing at least Part 1 before the lab is a good idea. If you want to write up your answers to Part 1 and/or Part 2, you can hand them in for extra credit.

[4]This was one of the puzzles that I wrote for the North American Computational Linguistics Olympiad—a high school contest that does not require any prior background.