

Short Exploration in Probability
Michael Ellsworth

In this project, I hope to practice applying some ideas discussed in the paper, Probability for Linguists, and the PowerPoint, N-gram Models and the Sparsity Problem, both of which I downloaded from your web-site. I will use as a corpus the text of Thoreau's *Civil Disobedience*, accessible through ClassicReader.com.

Before commencing with various calculations of probability, it is important to know the properties of the text at hand. Thoreau's essay has 9,334 words (not counting punctuation marks as words). Excluding the punctuation marks, Linguistica counts 2,019 unique words in the essay and there are 290 sentences in the essay. The first sentence of the essay follows:

(1) I heartily accept the motto, "That government is best which governs least"; and I should like to see it acted up to more rapidly and systematically.

The goal for this exercise is to find rules which allow this sentence to be more probable than its actual probability in the English language, which, due to the recursive nature of phrase structure, is infinitesimally small. [That's not right: it depends on the mode that you set up, and most reasonable models assign a finite probability to finite strings (they wouldn't be reasonable if they didn't!).] Theoretically, if there are 290 unique sentences in the work, the probability of each one is $1/290$ or about .345%. [Hmmm. That's one model, but an odd one.] However, this is an unrealistic goal: the only way to achieve this number is if our "rules" simply wrote out each sentence in the essay as indivisible entities and assigned an equal probability to each one!

Let's begin with no knowledge of English. It could be (wrongly) assumed that every word has an equal chance of being chosen at any time. With this assumption, and restricting ourselves to the language in the text, each of the 26 words in (1) could be one of 2,019 unique words. In other words, they would be considered as independent events with a probability of $1 / 2,019$ that they are the words they are. Thus, the probability of (1) under this scheme is $(1 / 2,019)^{26} = 1.1654e-86$. The positive log probability is 285.7. This is an astoundingly low probability. In other words, this is a lousy model of English. Right –and it would be better if you divide this by the number of words, because every model is going to assign a lower probability to longer sentences, all other things being equal.

The next step would be to acknowledge that words are not uniformly distributed throughout the corpus. This is still treating the words as independent events, however, it shows an understanding that the word "the" is more common in the corpus than "government," which is more common than the word "heartily." The probability of each individual word is given by (number of occurrences in text) / (length of text). The frequencies of the words in (1) in the text are as follows. The numbers in parentheses are their frequency in (1) itself, if more than one.

I(2) - 195 the - 499 that - 136 to(2) - 361 and(2) - 351
 it - 199 is - 150 best - 6 government - 51 governs - 2
 least - 7 which - 74 should - 26 like - 11 see - 18
 more - 20 up - 15 rapidly, systemically, motto, acted, heartily, accept - 1

There are 9,334 words in the text, so the probability of 'T' is $195/9,334 = .02089$
 The probability of sentence (1) is then:

$$(195^2)(499)(136)(361)(351^2)(199)(150)(6)(51)(2)(7)(74)(26)(11)(18)(20)(15)(1^6) / 9,334^{26}$$

| = $1.007e-70$, positive log probability = 232.5. This is an enormous improvement. Good!

This is still a very, very low number. Our grammar is still shockingly rudimentary, denying even that words affect each other. A few simple rules can start to help increase the probability of (1).

(a) An article can not end an English sentence. The text has 240 a's, 37 an's and 499 the's. Therefore the 26th word in (1), "systematically," has a probability of $(1 / 9,334 - 499 - 240 - 37) = 1 / 8558 = 1.168e-4$. The probability of sentence (1) becomes $1.097e-70$.

(b) Only in trivial circumstances can a word follow itself. Let $f(i)$ = the frequency of the word in position i . The probability of (1) becomes

$$p = (195 / 9,334) (1 / (8558 - 351)) \sum_{i=2}^{25} (f(i) / (9,334 - f(i-1)))$$

first 'I' 'systematically' middle words

$$p = .02089 (.0001219) (8.602e30 / 1.431e95) = 1.531e-70.$$

- (c) An article can not follow an article.
- (d) A conjunction can not follow an article or a conjunction or begin a sentence.
- (e) A preposition can not typically follow an article. A verb can not follow an article.
- (f) A pronoun can not follow an article or (typically) a pronoun.
- (g) A negative word can not follow an article.

Frequencies in the text of select words:

their - 51 of - 258 not - 143 in - 119 as - 108
 they - 96 but - 93 for - 93 be - 88 my - 54
 are - 77 he - 72 or - 67 with - 68 was - 52
 this - 60 have - 67 if - 58 his - 56

The probability of "motto" following "the" changes from $1 / 8835$ to $1 / 7458$.
 The probability of "heartily" following "I" changes from $1 / 9139$ to $1 / 8750$
 The probability of "should" following "I" changes from $26 / 9139$ to $26 / 8750$

The probability of "I" following "and" changes from 195 / 8983 to 195 / 8823
The probability of "systematically" following "and" changes from 1 / 8983 to 195 / 8823
The probability of "I" beginning the sentence changes from 195 / 9334 to 195 / 8823

The probability of the sentence has increased marginally, to 2.170e-70.
The positive log of this probability is about 231.4.

Unfortunately, rules (a) - (g) only change the probability, in total, by a factor of about 2. Why is this? They seem to be barking up the wrong tree! Actually, they are barking up the correct tree - that of conditional probability. The bark is pretty feeble however, because I am using a somewhat roundabout way to get to conditional probability: somewhat contrived grammar rules. A proper approach to conditional probability would find the probability of an item after the preceding (or following) item is known. Well, you're absolutely right about all of that, but it's certainly instructive to see how much (or in this case, how little) the formulation of such generalizations give us in quantitative terms.

Consider the following somewhat dramatic sentence from Civil Disobedience:

(2) I see this blood flowing now.

Frequencies: I - 195, see - 18, this - 60, blood - 5, flowing - 1, now - 28.

Treating each word as an independent event, the probability of sentence (2) is:

$195(18)(60)(5)(28) / 9,334^6 = 4.458e-17$. The positive log probability is 54.3.

One thing to note is that the probability is MUCH higher than the probability of sentence (1). Why is this so? Simply, because it's much shorter! [right, which is why it is very reasonable to focus on the average log probability for a sentence or string of sentences, not the raw probability or log prob] Logically, this makes sense, since a 6 word sentence has many less room for differentiation. But a rigorous mathematical treatment of this would realize that the two sentences of difference lengths are simply irrelevant to each other unless the STOP element (the period) is considered. Rolling a dice enough times will result in very differentiated outcomes, even if the number of faces on the dice is lowered. Still, this sentence can be made much more probable if one uses conditional probability.

Conditional probability $\rightarrow p(A | B) = p(A \text{ and } B) / p(B)$

There are two ways to begin finding the conditional probability of sentence (2). One is to find the likelihood of "I" beginning a sentence - in Civil Disobedience, Thoreau must be a bit self-centered : "I" begins 44 / 290 sentences. The other option is to ignore the punctuation, and just see that I follows the word "death." The probability of "I," given "death," is 1 / 2. Since we are treating sentences as autonomous, the first treatment seems better. Yes By using the biology concept of START and STOP codons, the probability for the initial "I" can be formalized.

290 instances of "START" and 44 of "START I" - $p(I | \text{START } _) = 44/290 = .1517$
The rest of the words follow the same formula:

195 instances of "I" and 3 of "I see," - $p(\text{see} | I _) = 3/195 = .0154$
18 instances of "see" and 1 of "see this" - $p(\text{this} | \text{see } _) = 1/18 = .0556$
60 instances of "this" and 1 of "this blood," - $p(\text{blood} | \text{this } _) = 1/60 = .1667$
5 instances of "blood" and 1 of "blood flowing" - $p(\text{flowing} | \text{blood } _) = 1/5 = .2$
1 instances of "flowing" and 1 of "flowing now" - $p(\text{now} | \text{flowing } _) = 1.0$

As intimated above, a proper treatment of this sentence would include the following:
18 instances of "now" and 4 of "now STOP" - $p(\text{STOP} | \text{now } _) = 4/18 = .2222$

The probability of the sentence is the product of all these conditional probabilities, which is equal to $9.606e-8$. The positive log probability is 23.3. This is an enormous improvement! The same process can be used with sentence (1), but it would be extremely tedious without automation! Now, an aside.

Sentences (3) and (4) have probabilities of 0.
(3) The constructivist educators saw themselves as radical.
(4) I see this blood flowing.

What happened? Sentence (3) is easy to explain. It is not in the text, and it could not be. The constructivist education movement didn't happen for nearly a century after the writing of *Civil Disobedience*. Even if the sentence was not an anachronism, the sentence is not part of the language of political discourse which Thoreau used any more than the sentence "I bought bread at market." The probabilities of educators, constructivist, and radical are all zero.

What about sentence (4)? It easily could have been included in the text, as it is almost identical to sentence (2), which is. Let's see where it broke down:

$p(4) = (44/290)(3/195)(1/18)(1/60)(1/5)(0/1) = 0$, positive log probability is infinite.

The sentence broke down at the end. $p(\text{STOP} | \text{flowing } _) = 0$. This is because in the text, there is only one appearance of "flowing" and that sentence doesn't end. This probability, which seems unreasonable in some regard is actually quite accurate in other regards: the probability of finding sentence (4) in *Civil Disobedience* is precisely 0! This shows that the probability models discussed so far are NOT models of English, models of early 19th century English, models of Thoreau's speech community, or of political discourse. They are models of the text, which is much too small to represent anything greater than itself. Well put.

The problem can be somewhat abridged by acknowledging it and doing something about it. One option is the Laplace method. In this method, all possible combinations are said to be possible, as long as the words are found in the text (therefore, it won't help us deal

with sentence (3) and the constructivists). If they are not found in the corpus of data, they are given the frequency of 1. If they are found in the data, then their frequency is said to be one greater than the actual frequency. The probability of any event is therefore $f(\text{actual}) + 1 / (\# \text{ of unique words in text})^2 + (\# \text{ of words in text})$. Since we are now considering periods as words, there are now 9,624 words in the text and 2,020 unique words.

Let's find the probability of sentences (2) and (4) using this method.

START I see this blood flowing [now] STOP

$$(2) \quad (44+1) (3+1) (1+1) (1+1) (1+1) (1+1) (1+1) (1+1) = 3.008e-43 \\ / ((2,020)^2 + 9,624)^7$$

$$(4) \quad (44+1) (3+1) (1+1) (1+1) (1+1) (0+1) = 3.076e-37 \\ / ((2,020)^2 + 9,624)^6$$

Positive log probabilities: (2) = 141.2, (4) = 121.3

Sentence (4), which is not found in the text, now has a much greater probability than sentence (2), which is means likely in real language, since (2) makes a temporal claim that sentence (4) did not - thus limiting its usage. Of course, this semantic reasoning is not directly called forth above, but it helps validate the results. The probability model is no longer completely restricted to the text at hand.

That's not to say that the model isn't without problems. For example,

(5) Civil. Blood I you. Me.

$$p(5) = (0+1)(1+1)(1+1)(0+1)(0+1)(0+1)(0+1)(5+1) / (2,020^2 + 9,624)^8 = 3.065e-52$$

Positive log probability of (5) = 171.1

Clearly, (5) is garbage Oh I don't know. Sounds kind of poetic to me, but it is assigned a positive probability, and one that is not extremely low, either. Sentence (1), actually from the text, has, with it 26 words, almost certainly a lower probability than this garbage! The Laplace model in this case assigns $\sim 2,000^2 = 4,000,000$ slots to unseen data, and $\sim 9,000(2) = \sim 18,000$ slots to seen data. In other words, $\sim .45\%$ of the probability mass is assigned to seen data! This is unacceptable. One of the most frequent pairs in the text is "of the," which occurs 67 times, or .6962% of all the pairs in the text. In the Laplace model, this pair is given a probability of $1.6626e-5$, which is much too low. The Brown corpus, a fairly reasonable approximation of the English language, sets the probability of "of the" at .9492% of the pairs. The Laplace model is wrong by 3 orders of magnitude!!! This is simply because Laplace considers such pairs as "the the," "or nor," "have wrote," and ". ." to be legitimate English.

The Laplace model is not an acceptable way to organize linguistic data, and much more powerful and complicated methods are available. Unfortunately, they will not be discussed here!

--

Well, I'm glad that you undertook to think about statistical language models on your own. I thought briefly about trying to cover some of this in the Intro course, and decided it would be too hard. Very nice!

Chris' comments:

Very interesting. You obviously need some syntax to help probabilistic modelling of text sentences.