Project 1

Due: 23 Jan 2004

The most nearly universal feature of modern NLP is constant interaction with large corpora of text; and the second most nearly universal is the gathering of all sorts of statistics for the purpose of ranking things. This project will get you used to dealing with corpora and statistics.

1 The corpus

The corpus we will use is derived from the Penn Treebank II Wall Street Journal corpus. Specifically, it is sections 2–21 (about 950K words), processed to remove empty nodes and “null” elements. These sections are usually used for the training portion of various algorithms, with the remaining sections reserved for other purposes (like testing).

As mentioned in class, the format of this corpus is parse trees in S-expression format. That is, each subtree is bracketed in parentheses, with the first element being the root node label, and subsequent elements representing children of the root (themselves represented as S-expressions). The top-level root node of each sentence is unlabelled, although in some treebank representations it is given the label “S1”.

Nonterminal labels can be broken up into a few parts, separated by hyphens and equal signs. The first part is always the actual phrase label. This is followed by function tags (of which there may be more than one) and then trace labels (which are numeric but can be any number). The phrase label is always present; the rest are optional.

Preterminal labels do not carry extra annotation. Note that some preterminals contain a hyphen—don’t confuse this with the hyphen that separates labels from function tags.

Terminals—words—can contain nearly anything. They are guaranteed not to contain whitespace or parentheses.

Figure 1 shows a sample S-expression sentence (the first in our corpus). The majority of the sentence is taken up by a prepositional phrase that is (metaphorically) locative, hence PP-LOC. The noun phrase that is the subject (SBJ) of the sentence contains several smaller noun phrases, variously
Figure 1: A sample treebank sentence

modified. The main verb phrase is headed by the auxiliary verb “was”, whose complement is another verb phrase headed by “attributed”.

2 The project

You’ll write a program that reads in the entire training and counts various words and labels to gauge their relative frequency. You have some flexibility in the exact formatting of your output, but you should not hard-code any filenames. Either accept the filename on the command line or read from standard input (or both), but document your interface decisions in a README file.

You’ll hand in your project by making a tar file (email me if you don’t know how) and copying it into the dropbox in the cs395C directory on leibniz. You’re welcome to do it in any programming language I can run on leibniz
(obviously including java and C++, and if you have other ideas let me know).

In addition to the electronic handin of code, hand in on paper the writeup and a printout of the required statistics. If it is not obvious, include in your README file instructions on how to replicate the printed-out statistics you hand in.

3 The statistics

Compile the following lists:

- The top twenty most frequent words; include both their total count and their probability of occurrence (i.e. if you picked a random word, what is the likelihood that it would be this word?)
- All nonterminals, ranked by frequency; include both count and probability.
- All preterminals, ranked by frequency; include both count and probability.
- For three preterminals (your choice), list the top ten most frequent words in that category, including count and (conditional) probability. Note that the counts on the words should only include those occurrences of the word that are tagged with this category.
- Pick some other interesting statistic on which to rank words, preterminals, nonterminals, or some combination thereof. List the top ten items according to this ranking.

4 The writeup

Write down any insights granted by the statistics rankings you compiled. Why did you pick the preterminals you did? What is interesting about the other distribution you selected? Did the results prove as interesting as you expected?

This writeup can be just a couple of paragraphs, certainly no longer than a page or two.