

# Finding compounds

John Goldsmith

September 23, 2007

Download our version of the Brown Corpus, and read in all the words. Alphabetize the words. Pick 100 consecutive words in this ranking to constitute your test set. Decide which of them are compounds, and which are not. If you are uncertain, do not guess, and replace the uncertain word with the next word on your list. Feel free to combine your test set with other people's test sets in the class (so that you can end up with multiples of hundreds of words—more is always better).

Write an algorithm to find compound words. Give a high-level description, and in pseudo-code. This should not take more than 20 or 30 lines. Your goal is just to explain how you are trying to solve the problem.

Run it on all of the words of the Brown Corpus. Assume that the function you have written gives a yes/no answer to the question, Is this word a compound? Calculate precision and recall on its ability to correctly identify compounds on your test set.

Here is a link to a list of some 2200 compounds that someone posted on the internet: <http://www.rickwalton.com/curricul/compound.htm>. Calculate how many of these words are identified as compounds by your algorithm, when these words are added to the Brown Corpus.

In addition, look at other words in the corpus, and explain in your own words what your algorithm does right and what it does wrong.

P.S. This problem is probably a lot harder than it looks. You might have to worry about not identifying words like “meat” and “mean” as compounds.