Joining Words into Meaningful Phrases through Virtual Text Compression

Eitan Pechenick
Dr. Peter Dodds
Department of Mathematics and Statistics
University of Vermont

April 19, 2012

Abstract

When perusing any manner of written expression, knowing the vocabulary does not suffice. A human reader employs a wealth of experience and intuition to determine, for instance, which words in a given sentence belong together as a single phrase. Clearly, “New York” is meaningful, as are some larger phrases building upon it—e.g. “New York City” and “New York State.” In particular, an accomplished reader would not attempt to interpret any of these phrases one word at a time—or, for that matter, try to interpret a word like ”new” one letter at a time—because there is nothing to be gained in the attempt. Similarly, a reader might treat “I am” the same way, since “am” is highly correlated with “I.”

There are a variety of measures for computing the correlation between two words in a given text, which are effective at ranking pairs of words by intuitive meaningfulness. However, these tend to have two drawbacks. First, a human has to choose a correlation threshold. If the threshold is too high, the computer rejects too many meaningful phrases. If it is too low, the computer accepts too many junk phrases. Second, it remains difficult to rank longer phrases.

The method explored here aims to address both issues. To capture long phrases, new phrases are built iteratively. At any step, a new phrase may be built from two phrases of arbitrary length. To address the issue of reasonable thresholds, new phrases are chosen so that if a particular text is compressed efficiently based on the current dictionary, adding the new phrase to the dictionary will improve the compression rate. The aim, then, is to compute a dictionary of phrases that optimizes compression of a text under an efficient scheme and to compare this with an intuitive sense of meaning.