Using compression to identify acronyms in text

Stuart Yeates, David Bainbridge and Ian H. Witten
Department of Computer Science, University of Waikato, Hamilton, New Zealand
{s.yeates, d.bainbridge, i.witten}@cs.waikato.ac.nz

Finding acronyms and their definitions in free text is useful for many purposes. Previous acronym definition finders relied heavily on heuristic methods. In contrast, we have developed a new method that uses several PPM models to encode the acronym in terms of its definition.

Four different attributes of each acronym are encoded using a PPMD order 5 model: (a) whether the acronym occurred before or after its definition (direction) (b) the distance between the acronym and the definition (first-word offset) (c) the pattern of words in the definition with letters in the acronym (subsequent-word offsets) and (d) the number of letters taken from each of those words. These models, taken together, give a complete encoding of the acronym in terms of its definition. The models were trained on 1080 acronyms extracted from 150 documents. A model of plain text was trained using 100 independent documents from the same collection.

Each word of the testing data was considered to be an acronym if it could be constructed from the initial letters of the 16 words on either side. If the word could be composed in more than one way, each possibility was considered. The entropy of the four attributes is calculated, the entropy of the word as normal text. The ratio of the two is compared to a threshold to determine whether to declare the word an acronym.

The new method is shown to outperform existing heuristic methods for acronym extraction. Figure 1 shows the results of separate experiments performed with acronyms of two or more letters and acronyms of three or more letters. The difference between the two recall-precision curves can be explained by the higher probability that a given non-acronym, two-letter, word can be formed from the words around it.