Evolving new lexical association measures using genetic programming

A collocation is an expression consisting of two or more words that correspond to some conventional way of saying things. Various definitions of a collocation range from identifying it with an idiom to saying that a collocation is just a set of words occurring together more often than by chance. There are many possible applications of collocations: automatic language generation, word sense disambiguation, improving text categorization, information retrieval, etc. As different applications require different types of collocations that are often not found in dictionaries, automatic extraction of collocations from large textual corpora has been the focus of much research in the last decade. Automatic extraction of collocations is usually performed by employing lexical association measures (AMs) to indicate how strongly the words comprising an n-gram are associated. Most approaches concentrate on improving and combining known lexical association measures such as log-likelihood, pointwise mutual information, chi square or Dice coefficient. In this presentation, a new approach based on genetic programming is presented. Our preliminary experimental results show that the evolved measures outperform three known association measures.