Abstract

This thesis investigates different statistical methods for the automatic extraction of lexical chunks. While no standard definition of lexical chunks exists, it is generally agreed that they are groups of two or more words which are often used together and whose meaning is often non-compositional, i.e., the meaning of the chunk is not fully determinable from the meaning of the individual words. Examples of lexical chunks include things like collocations, phrasal verbs, compound nouns, idioms, and phrases of greeting. The automatic extraction of lexical chunks is useful for many NLP applications, such as Machine Translation, Information Retrieval, and parsing; it is also important for foreign language teaching, as chunks have been implicated as important in language learning and processing.

In this thesis, I compare five different statistical association measures for the extraction of lexical chunks consisting of up to five words. These association measures are: Raw Frequency, Mutual Information, Log Likelihood, Mutual Expectation, and Symmetric Conditional Probabilities. All measures have been used with some success in automatic extraction tasks, although much previous work has focused only on specific types of chunks, such as collocations. My work differs from much of the previous work in that it allows for the extraction of any type of chunk, and it allows for the extraction of chunks with gaps in them (e.g, the chunk “as X as”, where the X could be filled by a number of different words).

Each measure was used in a training program, in which lexical chunks were extracted from a corpus of roughly 5.8 million words. The program then examined an article from the New York Times and extracted all the chunks it found using the database formed in training. These chunks were then compared to a gold standard list which had been compiled from a combination of idiom/collocation dictionaries and human judgments. It was found that Raw Frequency performed the best, but in further tests, a new method of training was employed which involved splitting the training corpus into 365 subcorpora, formed from the individual texts making up the corpus. This method was found to vastly improve performance, particularly recall, for most measures, especially Mutual Information and Log Likelihood.

These improvements were investigated and found to be mainly due to increased extraction of chunks consisting of a combination of content words (verbs, nouns, etc.) and function words (prepositions, conjunctions, etc.). The full corpus training had resulted in those measures finding mainly chunks consisting of only content words, which, though useful for the extraction of collocations, make up only a part of lexical chunks. The split corpus method is thus useful in the particular area of lexical chunk extraction, which is a unique area as lexical chunks form such a broad and difficult-to-define category.