Figure 12.3 The Probabilistic CYK algorithm for finding the maximum probability parse of a string of \textit{num\_words} words given a PCFG grammar with \textit{num\_rules} rules in Chomsky Normal Form (after Collins (1999) and Aho and Ullman (1972).) \textit{back} is an array of back-pointers used to recover the best parse. The \textit{build\_tree} function is left as an exercise to the reader.

When a treebank is unavailable, the counts needed for computing PCFG probabilities can be generated by first parsing a corpus. If sentences were unambiguous, it would be as simple as this: parse the corpus, increment a counter for every rule in the parse, and then normalize to get probabilities. However, since most sentences are ambiguous, in practice we need to keep a separate count for each parse of a sentence and weight each partial count by the probability of the parse it appears in. The standard algorithm for computing this is called the Inside-Outside algorithm, and was proposed by Baker (1979) as a generalization of the forward-backward algorithm of Chapter 7. See Manning and Schütze (1999) for a complete description of...