Session 7: Natural Language II

Salim Roukos

IBM Research Division, Thomas J. Watson Research Center,
Yorktown Heights, NY 10598

Context-free grammars (CFG) describe the possible derivations of a Non-Terminal (NT), denoted by A, independently of the context (or tree) in which this NT (also referred to as constituent) occurs. Typically there are several rewrite rules for a particular NT A. One straight-forward method for using a probabilistic model is to define a probabilistic CFG (PCFG) which assigns probabilities to the rewrite rules of a NT A that sum to one. This simple model therefore assumes a probabilistic independence assumption that the choice of the rule to expand A is independent of the context (parse tree) in which A occurs. As has transpired in the discussion period, this strong assumption was found to be objectionable by many. Several in the audience suggested using other grammatical formalisms to put probabilities on, where the independence assumption may be more acceptable. While this may be an approach, I suspect that there is a gold mine in using PCFG in a manner that is slightly more sophisticated than the above approach. For example, as **Magerman and Marcus** (and reference 3 in their paper) suggest, one may assign that the rewrite rule probability depend on the parent rule and the trigram of part-of-speech categories centered on the word that is the left corner of the rewrite rule. Other conditionings are possible and care should be taken that the parameters of the resulting probabilistic model can be estimated reliably from the training corpus. I think whether PCFGs or more sophisticated probabilistic grammars are needed for language processing would be best answered by using empirical experiments using standardized tests to allow for meaningful comparisons. To foster such a research program several ingredients are needed:

- Grammars (whether context free or not) that have a large enough coverage with a reasonable number of parses (say that the correct parse is with probability 99% in the top N parses where N is some agreed upon number.)
- Standard blind test sets annotated with the correct parse. See the paper by Black et al in these proceedings where a common marking is proposed.
- Standard training sets to allow parameter estimation and grammar development.

In trying to assess the value of PCFGs, one cannot forget the analogy to speech research where the HMM model has a strong independence assumption that has not yet shown to be a drastic barrier to improved performance.

Four of the five papers in this session address issues with PCFGs. Paper #2 by **DeMori and Kuhn** extends the algorithm to compute the probability that a sequence of words is an initial (prefix) substring of a sentence (see reference 8 in DeMori and Kuhn) to handle the case of an island: the probability that a string of words is an island (i.e., occur somewhere in a sentence.) They point out that the resulting computation is impractical. But they identify the special case where the gap length to the left of the island is known. Extending the gap length by one is only cubic. (In Paper #2, a dynamic cache language model using a tri-part-of-speech model is also described.) Paper #4 by **Kochman and Kupin** presents an algorithm that computes the probability that an LR parser for a PCFG will complete (accept a sentence.) Thereby, deriving the prefix probability (involves a matrix inversion) and then deriving update rules to compute the joint probability of the parser stack and input substring. Paper #5 by **Kupiec** presents a new organization on how to carry the probability computations (for parameter estimation of PCFGs) based on an extension that uses several trellises (*a la* forward-backward algorithm). The algorithm does not require the grammar to be in Chomsky Normal Form as required by the Inside-Outside algorithm but rather uses a recursive transition network representation of the grammar. As discussed earlier, Paper #3 by **Magerman and Marcus** makes the case that rule probabilities should depend on more context. However, they quickly abandon the probabilistic approach in favor of a heuristic score citing shortcomings of the independence assumption and unreliable probability estimates. I suspect that we will hear more on this subject as more empirical work is done to determine how best to deal with these issues. Finally, Paper #1 by **Bobrow** proposes a search strategy that uses several agenda to fill a chart in order to get an "acceptable" parse (a parse that leads to executable database access commands.) He introduces the use of the rule success probability which is estimated by frequency counts that the rule introduces terms that belong to the "acceptable" parse. Using heuristic weigthing of rule success probabilities yields a speedup by 1.8 compared to a full search CYK algorithm.