Session 9: Automatic Acquisition of Linguistic Structure

Mitchell Marcus, Session Organizer
University of Pennsylvania
Philadelphia, PA 19104

This special session was devoted to a rapidly expanding focus of research in natural language processing. Five years ago, only two or three pioneering researchers were attempting to automatically extract and utilize higher level linguistic structure from large corpora; all of this work was within the stochastic modeling tradition. The seven papers presented within this session provide evidence that a wide range of research is now underway in the automatic acquisition of linguistic structure utilizing both symbolic and probabilistic techniques as well as combinations of the two.

Rayner and Samuelsson

The key idea behind this work, presented by Rayner, is that multiple steps in the derivation of the syntactic and semantic analysis of a natural language input can be composed into single analysis rules which can then be applied much more efficiently to succeeding inputs. The resulting rules are modular and fairly general in that the composition algorithm stops whenever the syntactic category NP is encountered. Applied to a prototype NL query system, a speedup by a factor of 30 on following inputs has been observed on a small test corpus of queries. It was clarified during the question session that the system uses either these special compiled rules on a fragment of text input or its original grammar; it never attempts to use both simultaneously.

Hindle and Rooth

This work demonstrated a method by which a conventional broad coverage parser could be used to bootstrap an automatic statistical procedure for deciding prepositional phrase attachment, a key and central problem in natural language understanding. This procedure correctly decided PP attachment with an accuracy of 78% in a set of test cases where a PP might either modify the immediately preceding NP or the previous verb. Surprisingly, human judges succeeded in determining the correct attachment at an accuracy of only 85% given just the lexical information that the procedure used (i.e. the verb, the head noun of the following object and the the preposition). Limiting decisions to cases where the procedure's confidence is greater than 95% gives the same accuracy as these human judges, as does using only information extracted from the Cobuild dictionary (for the subset of cases where it contained information).

Question session: Bob Moore pointed out during the discussion that many preposition choices in real discourse are nonstandard; four of the preposition choices in the 6/90 ATIS corpus were nonstandard including *Flights leaving from Boston* to *NY*. Hindle, who presented the paper, suggested that a sufficiently large corpus of materials would see such examples. In response to another question, he also suggested that one could build a more complex algorithm which used more classical semantic information if the confidence of the algorithm was low.

Chitrao and Grishman

Chitrao presented this paper, which demonstrated an improved technique for assigning probabilities to the productions of a context free grammar and using the resulting probabilistic context free grammar to select among the alternate parses of an input sentence. Rather than assigning probabilities to each context free production rule (e.g. $S \rightarrow NP\ VP$)) in isolation, the context of each production is taken into account, and the priority assigned to each rule is dependent on the context in which it is used. On a corpus of test sentences from the MUCK II training data, the parser derives the correct parse first about 6% more often using statistical techniques than without. Using the new context sensitive techniques gives an additional 7% increase in accuracy (to 37% correct, with another 37% in error only due to PP attachment errors). In response to a query, it was revealed that while the techniques discussed here work much better than unconstrained parsing, that the extensions of preference semantics presented at the last DARPA workshop work marginally better by some measures.

Sharman, Jelinek and Mercer

Accurately estimating the probabilities of each context free production in a probabilistic grammar intended for unrestricted text may well require a prohibitive amount of training material, if done straightforwardly. This paper, presented by Jelinek, suggests using the so-called ID/LP (immediate dominance/linear precedence) formalism to factor a set of context free productions into a set of *dominance* relations, stating which non-terminals can dominate which other symbols in the grammar, and a set of *precedence* relations, stating which symbols will precede other symbols in a derivation. An experiment was performed using the IBM-funded Lancaster treebank of one million words of hand-parsed text taken from the AP newswire to use a probabilistic ID/LP grammar to parse English sentences. Tests show that the parser yields either a correct or close-to-correct parse about 60% of the time (exactly correct 19%).

During the question session, Ken Church argued that parameterization on purely structural relations, such as

used in this and the previous paper would be strikingly less successful than parameterization on words, parameterizing perhaps (as in the Hindle and Rooth paper) on pairs of words in certain structural relations. Much discussion resulted from a side comment of Jelinek's: A group from IBM informally surveyed a number of of purportedly "broad-coverage" parsers within the US on a test set of short sentences less than 14 words in length, and discovered that the best parser correctly identifies the best parse for these short sentences with an accuracy of only 60%. While these results struck many of those attending as extremely atypical, those who had worked on such parsers felt the results were a fair and accurate representation of the state of the art. It is my belief that the development of techniques similar to those presented in this session will lead to a major improvement in the accuracy of parsers in the very near future.

Brill, Magerman, Marcus and Santorini

A report on several related pieces of research, this paper was presented by the current writer. This work investigates the possibility that the grammar of a language can be inferred automatically by a distributional analysis of a large corpus of text.This work presents a new algorithm which uses an information theoretic measure to derive a unlabeled bracketings of novel sentences using primarily an analysis of the distribution of part-of-speech n-grams in an appropriately annotated corpus. On sentences of length less than 15, this algorithm misplaces on average 2-3 brackets per sentence. Initial experiments indicate that deriving appropriate part-of-speech tags from raw texts using similar distributional might well be possible. In the discussion, it was suggested that asymmetrical information measures might yield better results. The presenter responded that this might well be the case; while mutual information has been widely investigated recently, other measures might well perform somewhat better on this task.

Gale and Church

After showing that the task of spelling correction is in many ways closely analogous to the task of speech recognition, Church demonstrated that many standard estimators fail to yield correct results when used as part of a stochastic spelling corrector. Because of anomalies that result from sparse data problems, both the maximum likelihood estimator and the expected likelihood estimator fail to yield good results in choosing the correct word to replace a misspelled word. The Good-Turing method makes the use of contextual information useful, even in the case of very sparse data. An algorithm that uses the G-T estimator to do spelling correction was presented.

Much of the discussion focussed on the fact that spelling correction can often be separated into the correction of true typos and misspellings due to ignorance. Church noted that the performance of spelling correctors for the later case could be improved by utilizing stress information and expecting unstressed vowels to be incorrect far more often than stressed vowels.

Lewis

Lewis's contribution marks the first paper presented at a DARPA Speech and Natural Language Workshop by a researcher in the area of information retrieval (IR); it combines work in classical IR with work in natural language processing. The work presented here derives from the view that automatic classification within classical IR systems can usefully be viewed as machine learning. The paper itself investigates the use of structural relations as determined by a conventional parser to create indexing phrases. It presents the results of preliminary experiments which use clustering techniques to aggregate together related syntactic phrases into single concept classes (i.e. single dimensions in a real valued, multi-dimensional concept space) used within the context of an experimental IR system.

During the discussion, it was clarified that the clustering technique used was nearest neighbor clustering using cosine correlation between vectors. To a comment that Young and Hayes achieved 100%recall and 90% precision in work done for the Carnegie Group, Lewis pointed out that the task of *categorization* which they were doing is a far different (and simpler) task than that of doing retrieval with respect to *arbitrary* queries.

I would like to thank Julia Hirschberg for taking on the duties of chairing the session itself at the workshop.