SESSION 4: STATISTICAL LANGUAGE MODELING

Aravind K. Joshi, Chair

Department of Computer and Information Science
University of Pennsylvania
Philadelphia, PA 19104-6389

1. Introduction

Corpus based Natural Language Processing (NLP) is now a well established paradigm in NLP. The availability of large corpora, often annotated in various way has led to the development of a variety of approaches to statistical language modeling. The papers in this session represent many of these important approaches. I will try to classify these papers along different dimensions, thus providing the reader an overview as well as some understanding of the future directions of the work in this area.

There are two major motivations for research in statistical NLP, which are not necessarily independent of each other.

1. Robust Parsing: For processing free texts, hand crafted grammars are neither practical nor reliable. Statistical techniques are necessary both for robustness and efficiency. The use of statistical techniques for part of speech tagging and parsing is clearly motivated by these considerations.

2. Automatic Acquisition of Linguistic Structure: Here the goal is to use statistical techniques to discover linguistic structure by processing large corpus. The two motivations are clearly not independent, however the latter is more concerned with the extent to which the structure can be discovered statistically and the extent to which it has to be provided from outside.

2. Adaptive Stochastic Modeling

Improvements in stochastic language modeling can be obtained by using adaptive techniques. Della Pietra et al. describe an algorithm to adapt a $n-gram$ language model to a document as it is dictated. Rosenfeld and Huang describe an adaptive technique which uses information about within-document word sequence correlations, where one word sequence triggers another, causing its estimated probability to be raised. Such adaptive techniques are essential as the vocabulary size increases.

3. Part of Speech Tagging

Statistical techniques have been very successful in the task of part of speech tagging. There are two papers in this session dealing with part of speech tagging, representing two different perspectives. Black et al. describe the use of decision trees to estimate probabilities of words appearing with various parts of speech, given the context in which the word appears. Decision trees are used to take care of some of the problems of modeling long-distance dependencies.

Statistical techniques were introduced for part of speech have been more successful than the rule based techniques for the task of part of speech tagging. These rules are, of course, hand crafted. Brill presents a rule based tagger which automatically acquires its rules and tags from a corpus based analysis. Its accuracy is comparable to stochastic taggers. Brill's paper is an example of how statistical techniques can be used to acquire structure, thus opening possibilities for overcoming the limitations of usual rule based approaches to language processing.

4. Grammar Inference and Probabilistic Parsing

Grammar inference is a challenging problem for statistical approaches to natural language processing because the standard techniques based on finite-state models are incapable to represent hierarchical structure of natural language.

The parameter estimation methods have already been extended by Baker to stochastic context-free grammars. Pereira and Schabes describe some of the difficulties with the inside-outside algorithm, in particular the growth of local maxima as the number of nonterminals increases and the possible divergence between the structure inferred and the qualitative linguistic judgments. They propose an extension of the inside-outside algorithm using a partially parsed corpus in order to provide a tighter connection between the hierarchical structure and the inferred grammar.

Stochastic approaches to grammar inference and parsing

are significantly enhanced by combining lexical, structural, and contextual information. Several papers in this session describe different techniques for achieving this combination. Magerman and Weir describe a probabilistic agenda based chart parsing algorithm which uses a probabilistic technique for modeling where edges in the chart are likely to occur in the agenda-based chart-parsing process, enabling the parser to avoid the worst case behavior.

Mark et al. use a stochastic context-free grammar (CFG) combined with the $n - gram$ statistics, which provide some 'local' contextual information. They then describe techniques for parameter estimation.

Black et al. describe a history based approach for combining some lexical, syntactic, semantic, and structural information. They use the leftmost derivation of the parse tree to specify the context. Although they describe their approach using a CFG, it appears that the approach is more general and not necessarily limited to CFGs.

Schabes describes stochastic Lexicalized Tree-adjoining Grammars(LTAG). He shows how the inside-outside reestimation algorithm for stochastic CFGs can be extended to stochastic LTAGs. The LTAGs provide a framework for integrating hierarchical, syntactic, and lexical information in the grammar formalism itself, thereby allowing the specification of co-occurrence relationships directly.

Hindle also presents a parser that combines lexical and grammatical constraints into a uniform grammatical representation. In this sense, the papers by Schabes and Hindle are closely related. A new aspect of Hindle's parser is that it uses analogy to guess the likelihood of constructions outside the grammar.

The paper by Brill and Marcus, although I have classified it in the general category of grammar inference and probabilistic parsing, has a somewhat different flavor. Brill and Marcus present an algorithm for the acquisition of phrase structure grammar in an unsupervised manner. Their approach is based on the well-known distributional analysis techniques proposed by Zellig Harris in the early 50's. These techniques were not actively pursued earlier as it was not possible to work with large corpora in those days. Now it is possible to do so. Brill and Marcus use the entropy measure to evaluate the distributional similarity of items, something that can be carried out with the help of large corpora. The techniques as proposed by Harris were meant to be used by linguists doing the field work, judging the distributional similarity by questioning informants in the field.

5. Summary

I have identified the current trends in statistical language modeling by classifying the papers in the categories described above. The trend of combining statistical and grammatical information in some uniform manner will definitely continue and we should expect both theoretical and experimental results in the near future.

There is no reason to suppose that these statistical techniques are applicable only at the sentence level. It is very likely that these techniques will be applicable to certain aspects of discourse also. Again, it is important here to combine statistical information with information about discourse structure in a uniform fashion. Unlike grammatical structure, we still know little about discourse structure. Hence, research on discourse structure is crucial if successful application of statistical techniques is to be achieved.