Session 8: Statistical Language Modeling

Mitchell Marcus, Chair

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

1. Introduction

Over the past several years, the successful application of statistical techniques in natural language processing has penetrated further and further into written language technology, proceding with time from the periphery of written language processing into deeper and deeper aspects of language processing. At the periphery of natural language understanding, Hidden Markov Models were first applied over ten years ago to the problem of determining part of speech (POS). HMM POS taggers have yielded quite good results for many tasks (96%+ correct, on a per word basis), and have been widely used in written language systems for the last several years. A little closer in from the periphery, extensions to probabilistic context free parsing (PCFG) methods have greatly increased the accuracy of probabilistic parsing methods within the last several years; these methods condition the probabilities of standard CFG rules on aspects of extended lingustic context. Just within the last year or two, we have begun to see the first applications of statistical methods to the problem of word sense determination and lexical semantics. It is worthy of note that the first presentation of a majority of these techniques has been within this series of Workshops sponsored by ARPA.

It is a measure of how fast this field is progressing that a majority of papers in this session, six, are on lexical semantics, an area where the effective application of statistical techniques would have been unthinkable only a few years ago. One other paper addresses the question of how a POS tagger can be built using very limited amounts of training data, another presents a method for finding word associations and two others address various aspects of statistical parsing.

2. Part of Speech Tagging

The first paper in this session, by Matsukawa, Miller and Weischedel, describes a cascade of several components, sandwiching a novel algorithm between the output of an existing black-box segmentation and POS labelling system for Japanese, JUMAN, and the POST HMM POS tagger. The middle algorithm uses what the authors call example-based correction to change some of JUMAN's initial word segmentation and to add alternative POS tags from which POST can then make a final selection. (Japanese text is printed without spaces; determining where one word stops and another starts is a crucial problem in Japanese text processing.) The example-based correction method, closely related to a method presented by Brill at this workshop last year, uses a very small amount of training data to learn a set of symbolic transformation rules which augment or change the output of JUMAN in particular deterministic contexts.

3. Grammar Induction and Probabilistic Parsing

Most current methods for probabilistic parsing either estimate grammar rule probabilities directly from an annotated corpus or else use Baker's Inside/Outside algorithm (often in combination with some annotation) to estimate the parameters from an unannotated corpus. The I/O algorithm, however, maximizes the wrong objective function for purposes of recovering the expected grammatical structure for a given sentence; the I/O algorithm finds the model that maximizes the likelihood of the observed sentence *strings* without reference to the grammatical structure assigned to that string by the estimated grammar. Often, however, probabilistic parsing is used to derive a tree structure for use with a semantic analysis component based upon syntax directed translation; for this translation to work effectively, the details of the parse tree must be appropriate for tree-based semantic composition techniques. Current techniques are also inapplicable to the recently developed class of chunk parsers, parsers which use finite-state techniques to parse the non-recursive structures of the language, and then use another technique, usually related to dependency parsing, to connect these chunks together. Two papers in this session can be viewed as addressing one or both of these issues. The paper by Abney presents a new measure for evaluating parser performance tied directly to

grammatical structure, and suggests ways in which such a measure can be used for chunk parsing. Brill presents a new technique for parsing which extends the symbolic POS tagger he presented last year. Surprisingly, this simple technique performs as well as the best recent results using the I/O algorithm, using a very simple technique to learn less than two hundred purely symbolic rules which deterministically parse new input.

4. Lexical semantics: Sense class determination

The remaining papers in this session address three separate areas of lexical semantics. The first is sense class determination, determining, for example, whether a particular use of the word "newspaper" refers to the physical entity that sits by your front door in the morning, or the corporate entity that publishes it; whether a particular use of "line" means a product line, a queue, a line of text, a fishing line, etc. Several papers in this session address the question of how well automatic statistical techniques can discriminate between alternative word senses, and how much information such techniques must use. The paper by Leacock, Miller and Voorhees tests three different techniques for sense class determination: Bayesian decision theory, neural networks, and content vectors. These experiments show that the three techniques are statistically indistinguishable, each resolving between three different uses of "line" with an accuracy of about 76%, and between six different uses with an accuracy of about 73%. These techniques use an extended context of about 100 words around the target word; Yarowsky's paper presents a new technique which uses only five words on either side of the target word, but can provide roughly comparable results by itself. This new method might well be combined with one of these earlier techniques to provide improved performance over either technique individually.

5. Lexical semantics: adjectival scales

A second area of lexical semantics focuses on the semantics of adjectives that determine linguistic scales. For example, one set of adjectives lie on the linguistic scale from *hot* through *warm* and *cool* to *cold*, while another set lies on the scale that goes from *huge* through *big* to *little* to *tiny*. Many adjectives can be characterizing as picking out a point or range on some such scale. These scales play a role in human language understanding because of a phenomenon called *scalar implicature*, which underlies the fact that if someone asks if Tokyo is a big city, much better than replying "yes" is to say, "Well, no; it's actually quite huge". By the law of scalar implicature, one cannot felicitously assent to an assertion about a midpoint on a scale even if it is logically true, if an assertion about an extremum is also logically true. McKeown and Hatzivassiloglou take a first step toward using statistical techniques to automatically determine where adjectives fall along such scales by presenting a method which automatically clusters adjectives into groups which are closely related to such scales.

6. Lexical semantics: Selectional Restrictions

Another key aspect of lexical semantics is the determination of the selectional constraints of verbs; determining for each sense of any given verb what kinds of entities can serve as the subject for a given verb sense, and what kinds of entities can serve as objects. For example, for one meaning of *open*, the thing opened is most likely to be an entrance; for another meaning, a mouth; for another, a container; for another, a discourse. One key barrier to determining such selectional constraints automatically is a serious problem with sparse data; in a large corpus, a given verb is likely to occur with any particular noun as object in only a handful of instances. Two papers in this session automatically derive selectional restrictions, each with a different solution to this particular form of the sparse data problem. The paper by Resnik utilizes an information theoretic technique to automatically determine such selectional restrictions; this information is then used to resolve a number of syntactic ambiguities that any parser must deal with. Resnik uses the noun *is-a* network within Miller's WordNet to provide sufficiently large classes to obtain reliable results. Grishman and Sterling attack the problem of sparse data by using co-occurance smoothing on a set of fully automatically generated selectional constraints.

In one last paper in lexical semantics, Matsukawa presents a new method of determining word associations in Japanese text. Such word associations are useful in dealing with parsing ambiguities and should also prove useful for Japanese word segmentation.