SESSION 8 & 9: STATISTICAL AND LEARNING METHODS

Frederick Jelinek, Chair

Center for Speech Processing
The Johns Hopkins University
Baltimore, MD 21218

The statistical and learning methods of these two sessions are related to text processing. The presented techniques should facilitate achieving the desired goal of text understanding with all of its potential exploitation: speech understanding, gisting, information retrieval, indexing, machine translation, etc.

For many years, the conventional approach consisted of an attempt to design systems by experts who gained their knowledge through personal experience, introspection, and information exchange with other experts. As time went on, and claims of promise failed to be redeemed, the realization began to take hold that facts about language are too complex to be either listed or incorporated in rules specified by humans. Something always seemed to have been forgotten, and surprisingly frequent cases remained untreated.

Influenced in no small measure by the success of self-organizing methods in speech recognition, a growing circle of researchers have concluded that decisive progress in the field of language can only be achieved when the necessary knowledge will be extracted directly from data, lots and lots of diverse data. The proper work of experts is to secure data, create facilities for the reliable, plentiful, and appropriate annotation of the data, and help design systems that can extract the appropriate information from this data.

The results presented in the sessions on statistical and learning methods provide help in carrying out the last task.

The papers by Miller, Chodorow, Landes, Leacock, and Thomas (Using a Semantic Concordance for Sense Identification) and by Bruce and Wiebe (A New Approach to Sense Disambiguation) present methods that can be used to annotate in context the sense in which a word is used. E.g., when we say BANK, do we mean the institution for the deposit of money, or the shore of a river?

The training data for the work of the next four papers are treebanks.

The paper by Miller and Fox (Automatic Grammar Acquisition) attempts to learn statistical rules of a context-dependent grammar appropriate for shift-reduced parsers.

Assuming that text understanding will require parsing (constituent analysis), and realizing that grammar is only a tool that (for this application) is not needed for its own sake, three papers are dedicated to direct parsing without an explicit specification of any grammar. Two of these papers are statistical. They are related not just by authorship, but by use of techniques extracting the needed probabilities. The article by Ratnaparkhi and Roukos (A Maximum Entropy Model for Prepositional Phrase Attachment) deals with the decision whether in a sequence consisting of verb, noun, and prepositional phrases, the last two form a compound noun phrase or all three form a unit (e.g. does "ate the banana on the plate" specify a manner of eating or the location of the banana?). The article by Jelinek, Lafferty, Magerman, Mercer, Ratnaparkhi, and Roukos (Decision Tree Parsing Using a Hidden Derivation Model) presents a new method of constructing a statistical, direct, grammar-less parser attaching annotated trees to given word strings. Finally, Brill (A Report of Recent Progress in Transformation-Based Error-Driven Learning) discusses a deterministic method of direct parsing and tagging (part-of-speech annotation) in which an original standard parse or tagging is modified in a step by step fashion until the final form is attained.

Of course, parsing itself is only a means to an end. What one is really after is text understanding which the parse should facilitate. Miller, Bobrow, Schwartz, and Ingria (Statistical Language Processing Using Hidden Understanding Models) attempt to extract meaning directly and develop a corresponding methodology applicable to restricted domains such as the ATIS task. Given a sentence, their goal is to fill a template.

Meng, Seneff, and Zue (Phonological Parsing for Bi-directional Letter-to-Sound / Sound-to-Letter Generation) deal with a problem interesting to speech recognition and synthesis: given a spelled word, what should be its phonetic realization, and vice versa: given a phonetic string what is its likely spelling? The latter problem may arise when a person pronounces a word whose phonetic structure the system can recognize but which is new to its vocabulary.

The processing of Japanese text is complicated by the fact that there are no word delimiters. Japanese text consists of sequences of kanji followed by kana characters signaling inflection, politeness, and other information. The segmentation of such text is conventionally accomplished by deterministic rules. Papageorgiu (Japanese Word Segmentation by Hidden Markov Model) uses statistics.

Finally, Pereira, Riley, and Sproat (Weighted Rational Transductions and their Application to Human Language Processing) present a new algebraic uniform representation which they claim to be applicable to varied information sources, such as pronunciation dictionaries, language models, and lattices. If practically successful, this automata theory approach will surely be considerably elaborated.