Integrating Speech and Natural Language

PI: Salim Roukos

(617) 873-3452, sroukos@bbn.com
BBN Systems and Technologies Corporation
10 Moulton Street, Cambridge, MA 02238

The overall goal of this project is to integrate speech and natural language knowledge sources to build a speech understanding system for human-machine communication using spoken English. The speech knowledge sources use acoustic models based on hidden Markov modeling techniques. The natural language knowledge sources use a Unification grammar formalism for describing the syntax of English, a higher-order intensional logic language for representing the meaning of an utterance, and a 'Montague Grammar style' framework for interfacing syntax and semantics.

The objective of an integrated search strategy is to find the globally optimal (by acoustic likelihood) interpretation of the input given the constraints that are imposed by the syntactic and semantic components. Our approach in the BBN Spoken Language System (SLS) uses hidden Markov word models to determine a 'dense' word lattice (to minimize errors due to early decisions) of possible words present in the input speech. Then, a *lattice parser* is used to find all parses for all the syntactically possible word sequences present in the lattice (ordered by acoustic likelihood). Finally, a semantic interpreter determines from the ordered list of possible word sequences the highest scoring meaningful word sequence. That word sequence is the recognized sentence and its meaning is the interpretation of the input speech.

The lattice parser is an extension of our bottom-up word-synchronous parser for the Unification grammar to accept a word lattice as input. The algorithm determines all possible grammatical word sequences and ranks them by acoustic likelihood scores. Then, that list of grammatical sentences is processed to determine the highest scoring word sequence that is also meaningful using the application domain semantics.

The SLS system was evaluated using the DARPA Resource Management database. For the language modeling components (syntax and semantics), we used a training corpus of 791 sentences, which has been used for system development, and a test corpus of 200 sentences, which was unseen by the system developers. The grammar coverage was 92 % for the training corpus and 81 % for the test corpus. The coverage of the semantic interpreter was 75 % and 52 % for the two corpora, respectively.

We have also measured the speech understanding performance on three speakers from the 1000-word Resource Management database. The word accuracy on the 1000-word task improves from 71 % when no language model is used to 87 % when the syntax alone is used and to 92 % when both syntax and semantics are used.