Data Collection And Evaluation

David S. Pallett
National Institute of Standards and Technology

ABSTRACT

This session focussed on two inter-related issues: (1) performance assessment for spoken language systems and (2) experience to date in speech corpora collection for these systems. The session included formal presentations from representatives of SRI International, MIT's Laboratory for Computer Science, BBN Systems and Technologies Corporation, and Carnegie Mellon University's School of Computer Science.

SESSION OVERVIEW

Material presented by Patti Price et al. of SRI International described collection of more than 12 hours of human-human interactive problem solving in the air travel planning domain. SRI has made use of this data to define an initial vocabulary and to define an interface for this domain. Recent efforts to conduct "Wizard" simulations of human-system interactions in this domain were described. Price noted that in the naturally occurring dialogues it is rare that a database query occurs. Rather, the user will state a plan (e.g., "I need to make a reservation") and then provide, in small steps, the pieces of information necessary for the "agent" to help accomplish the plan. Breaking the dialogue into small pieces of information, asking for frequent confirmation, and having the agent sometimes take an active role all seem to play a role in making the dialogue efficient. These findings suggest to Price that both arguments of naturalness (a strong motivating factor for the use of natural language in the first place) and efficiency argue for human-system interactions that will yield large numbers of sentences that are not database queries. In view of this, it was argued that performance assessment procedures must go beyond consideration of database query-answer pairs and include other mechanisms for assessment, such as the template-based method of the MUCK-2 approach.

In the first of the two papers by the group at MIT's Laboratory for Computer Science, Zue et al. describe the collection and preliminary analysis of a spontaneous speech corpus using a simulated human-system dialogue with the VOYAGER spoken language system. The Voyager system is made up of three components: (1) the SUMMIT speech recognition system which converts the speech into a set of word hypotheses, (2) the TINA natural language component, which provides a linguistic interpretation and a parse tree that is translated into a query language form, and (3) a modified version of the direction assistance program (developed by Jim Davis of

MIT's Media Laboratory). Spontaneous speech data were recorded from 100 subjects, and each subject was also recorded reading orthographic transcriptions of their spontaneous speech (minus false starts, hesitations and filled pauses). In collecting the spontaneous speech, a simulation was used that replaced the SUMMIT speech recognition component with a human typing the orthographic transcription of the spontaneous speech into the remainder of the system. The speech data, consisting of nearly 10,000 utterances, were subsequently digitized and transcribed. Comparison of corresponding spontaneous and read speech data show that the spontaneous utterances [i.e., sentences] are longer than their read counterparts, and that there is much more variability in the spontaneous speech. Pauses that are more frequent and longer account for much of the longer duration characterizing spontaneous speech. Non-speech vocalizations found in the spontaneous speech include mouth clicks, breath noise and filled pauses such as "um", "uh", or "ah". False starts occurred in almost 4% of the spontaneous sentences. The words following false starts in the spontaneous speech included "back ups" (to the same as the [apparent] intended word), a different word in the same [syntactic] category, a word from a new category, or a back up to repeat [several] words already uttered. This study concludes that the process of data collection in this simulation was relatively straightforward, and that incremental data collection can "be quite effective for development of spoken language systems".

In the second paper from Zue's group, a number of performance assessment issues are raised. It is suggested that spoken language systems should be evaluated along several dimensions. The dimensions include (1) accuracy of the system and its various modules (e.g., phonetic, word and sentence accuracy as well as linguistic and task completion accuracy), (2) coverage and habitability, (3) flexibility, and (4) efficiency (e.g. task completion time). Zue et al. note that evaluations of accuracy inevitably involve the use of references involving varying degrees of subjectivity. At higher levels, system outputs may involve more abstract information, complicating the process of automatic comparison with a reference output. The preliminary evaluation of the MIT Voyager system includes evaluation of the Summit speech recognition component. Using a 570 word lexicon and a word-pair grammar with a test set perplexity of 22 to constrain the search space, word accuracies of approximately 86% and sentence accuracies of 49% for sentences of about 8 words per sentence are reported. Analyses of natural language performance focussed on coverage in terms of percentage of sentences that could be parsed and perplexity. Overall system performance has been evaluated by several means, including a panel of naive users to judge the appropriateness of the responses of the system as well as the queries made by the subjects. Although data were available only for a small number of subjects, it appeared that "appropriate" responses together with "verbose, but otherwise correct" responses

accounted for approximately 85% of the responses. About 87% of the user queries were judged reasonable. The issue of efficiency was not addressed, since the system under discussion operates in about 12 times real time, precluding real-time interactive dialogues.

The paper by Boisen, Ramshaw and Bates from BBN describes "an automatic, essentially domain-independent" means of evaluating spoken language systems that provide answers to queries from a database. This proposal was developed out of an understanding that some consensus has been achieved on a number of issues including: (1) "Common evaluation involves working on a common domain (or domains). A common corpus of development queries (in both spoken and transcribed form), and answers to those queries in some canonical format, are therefore required.", (2) "One basis for system evaluation will be answers to queries from a common database, perhaps in addition to other measures." (3) "Automatic evaluation methods should be used whenever they are feasible". The proposal for evaluation on a DARPA common task has as a key component a program designated a "Comparator" that compares canonical answers to the answers supplied by a spoken language system. Answers are to be expressed in the form of a "Common Answer Specification (CAS)", as described in the proposal. The proposed comparator is a Common LISP program for comparing system output expressed in CAS format with canonical answers. Much as Zue et al. note, Boisen et al. note that "evaluation requires human judgement, and therefore the best we can expect from a program is comparison, not evaluation". BBN has prepared a small corpus of queries and their answers for the (proposed) "Common Personnel Database" to illustrate the use of the CAS format and as a check on the clarity and completeness of the CAS. Finally, Boisen et al. note that the collection of any corpus "for SLS development and testing will be more useful if it is easily sub-divided into easier and harder cases", and they propose candidate categorizations, starting from a default case in which no extra-sentential context is required, to the more difficult categories involving "local" extra-sentential reference, ellipsis cases, non-local references and [even] more complex cases. It is argued that these principles of categorization should be followed in implementing SLS evaluations.

In BBN's second paper in this session, Derr and Schwartz described the development of a new grammar that can be used in assessing the performance of speech recognition systems. It is a "statistical first-order class grammar" that has been developed for two different task domains (the DARPA Resource Management domain and a 2000 word personnel database domain). Derr and Schwartz argue that the existing two grammatical conditions (the "no grammar" or null grammar and the word-pair grammar cases) "suffer from several inadequacies". The null grammar provides "only a worst-case recognition test point", while the word-pair grammar not only excludes "many reasonable word sequences", but the use of the word-pair grammar yields such high recognition performance that reliable measurement of system improvements (i.e. statistically significant inferences of improvements) cannot be obtained without use of very

large development and evaluation test sets. Given a priori assignment of words to classes, the statistics of BBN's class grammar were counted directly from training data by counting the number of transitions from each class to each other class. Using 99 classes for the lexicon of the DARPA Resource Management task domain, the class grammar provides a perplexity of approximately 75 and recognition error rates that are in-between the results for the word-pair and null grammars. The increased error rate is noteworthy not for the fact that recognition performance is degraded per se, but for the fact that, using this grammar, incremental improvements may be shown to be statistically significant using smaller test sets and less test and machine time. [Following this presentation, it was observed that incremental improvements could also be shown to be statistically significant with the null grammar, but without the apparent benefit of higher word accuracies (i.e., the performance using the class grammar is shown to be somewhat better than the worst-case results for the null grammar). The critical issue is one of defining the desiderata for constraining grammars and the relationship of these grammars to those that are in some sense "natural" to the task sub-language.] Further advantages outlined for this approach include the fact that it is readily adapted to new task domains and that it is "tunable" (i.e., can be set to provide varying perplexity) by varying the number of classes in the grammar.

The paper by Rudnicky et al. from CMU presents results of a study of a spoken language interface involving a complex problem-solving task. A group of users was asked to perform 40 spreadsheet tasks, each successive task being carried out in a different modality (speech or typing). The voice spreadsheet consists of the UNIX-based spreadsheet program "SC" interfaced to a recognizer embodying the Sphinx speech recognition technology. The CMU study is noteworthy for the fact that, of the systems discussed in this session, it is the only system to provide near-real time interactions without the intervention of a "Wizard". For a task vocabulary of 271 words [and a constraining grammar with a perplexity of 52], word accuracies of 92.7% to 94.9% were achieved for spontaneous and read speech. Analyses conducted by Rudnicky et al. include discussions of semantic accuracy, grammaticality and language habitability. Spontaneous speech events are discussed in three categories: lexical, extra-lexical, and non-lexical. Detailed analyses also include "the time it takes to do things". Since the implementation used at CMU for these studies processed speech in about 2 times real time, it is perhaps not surprising that total task time was greater for speech input than keyboard. However, by accounting for processing "overhead" times and proposing a halving of the present error rate, Rudnicky et al. estimate that task completion times for speech and keyboard should be "equivalent". Current efforts are directed toward achieving a true real-time implementation and improving system accuracy.