SESSION 10: CORPORA AND EVALUATION

Clifford J. Weinstein

MIT Lincoln Laboratory
244 Wood Street
Lexington, MA 02173

INTRODUCTION

This session on corpora and evaluation was composed of two distinct parts. Before the break, four papers dealing with a range of important aspects of evaluation of written language systems and spoken language systems were presented. A printed version of each of these papers is included in the conference proceedings. After the break, a series of informal reports (not included as proceedings papers) were given summarizing the work of the Corpora and Performance Evaluation Committee (CPEC) of the DARPA Spoken Language Systems (SLS) Program, with specific reports from several working groups which have been dealing with various aspects of corpora collection and performance evaluation in the SLS Program. A lively and extended discussion followed these working group reports, including presentations of a number of alternate viewpoints.

SUMMARY OF PAPERS PRESENTED

Third Message Understanding Conference (MUC-3): Phase 1 Status Report—Beth Sundheim

Beth Sundheim reported on the MUC-3 evaluation effort, for which a dry run had just been completed during the week prior to the Speech and Natural Language Workshop. The general sense of the presentation was that much progress had been made relative to prior MUC evaluation efforts. The number of sites reported had increased to twelve. The evaluation was broader in scope than previous ones in most respects, including text characteristics, task specifications, and performance measures. Specifically, the overgeneration and fallout measures were new in this round. A semi-automated scoring program had been developed and distributed to all sites, including computation of recall, precision, overgeneration, and fallout measures designed specifically for the MUC task. A selected set of dry run results was presented, most of which are summarized in the printed paper. It was emphasized that the dry run results should not necessarily be expected to be predictive of the results of official testing, which will take place in May 1991. As an item of interest, Sundheim noted tests which had been performed at NYU using a "system" which ignored all but the dateline of each message, and which actually outperformed some "serious" systems in some of the evaluation dimensions. In the discussion, Paul Bamberg suggested that the scoring could be modified so that a "trick" system could not perform well by simply relying on a priori probabilities and guessing the most likely message. The new scoring procedure would favor systems which produced the right information for unlikely input messages.

A Procedure for Quantitatively Comparing the Syntactic Coverage of English Grammars—Ezra Black, et al.

Ezra Black reported, on behalf of a group of fourteen co-authors/grammarians, on a new method of quantitatively comparing parses provided by different grammarians, when each grammarian provides the single parse which he/she would ideally want his/her grammar to produce. The comparison method is well-defined, relatively simple, and, more importantly, has been agreed upon by all the grammarians/co-authors. The procedure judges each parse only by its constituent boundaries (and not by the names assigned to these constituents), compares the parse to a standard parse, and produces two scores (the Recall Rate and the Crossing Parentheses Rate) for the candidate parse. Initially, the standard parse was taken as a majority parse derived from the grammarians' parses. Then it was determined that an independent standard parse (the "hand parse" from the University of Pennsylvania Treebank) worked just as well. The UPenn Treebank parse was then accepted as the standard parse. Based on the scores produced on a test sample consisting of 14 sentences from the Brown corpus, the hand parses produced by the grammarians were very consistent (average Recall 94% and average Crossing Rate <1%). Black stated that a key next step will be to apply this evaluation procedure to machine parses, and remarked that initial tests had indicated that the grammars did not do nearly as well as the grammarians (the Recall Rate on the best of four machine parsers tested so far was about 60%).

One questioner from the audience asked how the 14 sentences were selected ("pseudo-randomly"), and whether 14 was enough for a meaningful test. Black suggested that 14 was a reasonable start, and that it was intended that 50 sentences would be used for test in the next round, which would involve machine parsers. All those with suitable machine parsers were invited to participate in this next test. A second questioner asked whether the use of semantic knowledge by the grammarians made the test unfair to syntactic grammars. Black responded that the main issue was to develop a comparative measure for parses, independent of what knowledge sources were used.

Evaluating Text Categorization—David Lewis

David Lewis described and compared a variety of methods for evaluation of text categorization systems. He explained the relationship between text categorization, and the more-frequently-evaluated task of text retrieval. He defined the system effectiveness measures of recall, precision, and fallout, in terms of a contingency table for binary decisions, and discussed the difficult issue of defining a standard of correctness on test data sets. He contrasted "microaveraging" and

"macroaveraging" strategies in evaluation of retrieval systems, and commented that MUC uses microaveraging, where all decisions made on a document are considered as a single group when computing recall, precision, and fallout. He described indirect evaluation methods, wherein text categorization is evaluated based on the performance of a system (e.g., text retrieval or text extraction) which uses the results of the text categorization. The paper serves to place in perspective a number of important techniques, as well as issues which must be addressed, in evaluation of text categorization systems.

A Proposal for Incremental Dialogue Evaluation—Madeleine Bates and Damaris Ayuso

Lyn Bates presented two proposed techniques for evaluation of SLS systems that deal with dialogue. The examples in this paper dealt directly with the Air Travel Information System (ATIS) task being used in the SLS program, hence this paper served as a good bridge to the ATIS-oriented presentations after the break. The first proposal suggested a methodology for comparing systems based on their ability to handle diverse utterances in context. For example, each system would be tested with a specific "context-setting" query pair, followed by a set of (say, 10) alternative "next-utterances," each of which would directly follow the context-setting query pair. The system would be judged based on its answers for each of the alternative "next-utterances." The second proposal suggested a modification of the performance metric to encourage partial understanding. This proposal was aimed at permitting systems some leeway in responding to partially-understood queries, as well as providing credit for reasonable responses. It was noted that the judgment of whether an answer was reasonable would almost certainly have to be made by human arbiters. Both proposals in this paper were offered for further consideration by the SLS evaluation committees.

CORPORA AND PERFORMANCE EVALUATION COMMITTEE (CPEC) REPORTS

The collection and distribution of common speech corpora (including text transcriptions), and the definition and execution of procedures and standards for the performance evaluation of SLS systems processing these corpora, have been central activities in the SLS program. These efforts have been carried out under the aegis of a Corpora and Performance Evaluation Committee (CPEC), with the tasks divided among several Working Groups. This part of the session was devoted to reports by the CPEC chairperson (David Pallett) and the chairpersons of each of the Working Groups. Each report included a brief summary of (1) the charter and goals of the group; (2) activities, issues addressed, decisions, and accomplishments; and (3) open issues and work remaining to be done. The ensuing discussion included presentation of a number of alternative viewpoints on some of the issues addressed.

The order of presentations was as follows:

CPEC Overview—Dave Pallett
ATIS Corpora Working Group Report—Dave Pallett
Performance Evaluation Working Group Report—Bill Fisher
ATIS Relational Database Working Group Report—Bob Moore
ATIS Speech Recognition Working Group Report—Victor Zue
Informal ATIS Speech Recognition Baseline Definition—Doug Paul
Speech Corpora Working Group Report—Janet Baker

CPEC Overview—Dave Pallett

CPEC Members. Dave Pallett (NIST, Chair); Janet Baker (Dragon); Lyn Bates; (BBN); Bob Moore (SRI); Alex Rudnicky (CMU); Victor Zue (MIT-LCS).

CPEC Issues and Role. Pallett identified some of the many issues and details which CPEC has had to address to achieve the goals of a useful common corpus and meaningful evaluation procedures. For ATIS, these issues included: subject scenarios; subject instructions; wizard instructions; display/feedback; data collection protocols (e.g., push-to-talk); transcription conventions; data classification; class definition (A, D1, etc.); comparator revisions; canonical answer formats. The specifics of these issues were delegated to the Working Groups, with CPEC responsible for integration.

Pallett also described the relationship between the CPEC and working groups, the contracting agent (NIST), and the database contractors (TI and SRI). Starting around July 91, NIST became DARPA's agent for data collection contracts, and hence became the official interface with the contractors. NIST, in turn, had the challenge of integrating the decisions and advice of the Working Groups into its direction of the contractors' efforts.

CPEC Challenges, February 91. Pallett noted the following challenges for CPEC and the Working Groups in the months to come: (1) more ATIS data, collected at a faster rate; (2) better, more consistent, ATIS data; (3) improved documentation and communication; (4) refinements in evaluation procedures; and (5) development of new evaluation procedures.

ATIS Corpora Working Group Report—Dave Pallett

This report was integrated with the CPEC report, since Pallett chairs both groups.

ATIS Corpora Working Group Members. Dave Pallett (NIST, Chair); Charles Hemphill (TI); Lew Norton (UNISYS); Patti Price (SRI); Alex Rudnicky (CMU); Stephanie Seneff (MIT); Jay Wilpon (ATT-BL).

ATIS Corpora Work Group Primary Task. Agree on data collection paradigm for "contracted" ATIS Corpora (additional, ad hoc ATIS data was collected by some of the SLS groups).

ATIS Corpora Status. Training data for the February 91 evaluation was provided by both contractors (SRI and TI). The test data for the February 91 evaluation was provided by TI.

Performance Evaluation Working Group Report—Bill Fisher

Performance Evaluation Working Group Members. Bill Fisher (NIST, Chair); Janet Baker (Dragon); Debbie Dahl (UNISYS); Lyn Bates (BBN); Lynette Hirschman (MIT-LCS); Doug Appelt (SRI); Wayne Ward (CMU).

Performance Evaluation Issues and Accomplishments. Issues addressed by this Working Group included: principles of interpretation, common answer specification, query classification, and strategies for dialog evaluation. A major accomplishment was the addition of a test involving some dialog dependency (the "D1" class of query), in addition to the purely class A tests conducted in June 90.

Areas for Further Work. Updates are needed for the Principles of Interpretation Document and for the Classification Document. The suite of tests needs to be carefully extended to include more dialog phenomena.

ATIS Relational Database Working Group Report—Bob Moore

Working Group Members. Bob Moore (SRI, Chair); Rusty Bobrow (BBN); John Garofolo (NIST); Charles Hemphill (TI); Don McKay (UNISYS); Joe Polifroni (MIT/LCS).

Goals, Approach, Issues. Moore emphasized that the database obtained from OAG was not in relational database form. He commented that credit was due to TI, particularly Charles Hemphill, for doing a good job, under heavy time pressure, in organizing the ten-city database. He noted several issues which the Working Group had addressed, including displays, schema, etc. He noted controversies about certain issues, such as the display presented to the user. He suggested that it would be very desirable to add connecting flights to the database, as this would make ATIS a richer task; and noted that significant work, including serious changes to the schema, would be needed to include connecting flights. In the discussion, the Stephanie Seneff suggested that it would be important to include not only connecting flights, but also their fare structure.

ATIS Speech Recognition Ad Hoc Working Group Report—Victor Zue

Working Group Members. Victor Zue (MIT-LCS, Chair); Janet Baker (Dragon); Kai-Fu Lee (Apple); Hy Murveit (SRI); Doug Paul (MIT-LL); Rich Schwartz (BBN); Rich Stern (CMU); Jay Wilpon (ATT-BL).

Goal, Issues, and Outcome. The goal of this Working Group was to propose, in the ATIS domain: (1) a speech recognition test protocol including definition of training and test sets, a language model, and a vocabulary; (2) a scoring procedure. Issues addressed by the Group included:

(1) Modularity of the speech recognition component and appropriateness of the evaluation--the key issues here were how to evaluate speech recognition in an environment where speech recognition is only an intermediate result, and speech understanding is the goal.

(2) Need to converge quickly on a solution.

After considerable discussion, the Working Group reached agreement on the test set and on a modified scoring algorithm.

Training set, language model, and vocabulary were left open for the official February 91 evaluation. Further consideration of ATIS speech recognition evaluation was deferred until after the February 91 tests. However, an informal baseline specification of training set, language model, and vocabulary were proposed by Doug Paul and Rich Schwartz.

Informal ATIS CSR Baseline Test Definition--Doug Paul

Doug Paul described the informal ATIS CSR baseline test specification developed by himself and Rich Schwartz. This specification included:

(1) a designated set of acoustic training and development test data;

(2) a vocabulary of 1065 words, consisting of 921 observed words and additional words added to close classes (e.g., months, days); and

(3) a perplexity 17.8 bigram backoff language model, including the extended vocabulary and an "unknown word" class.

All data and information for this baseline was made available to all sites, with a note encouraging sites to test both under the baseline conditions and under other conditions of their choosing. BBN and Lincoln conducted tests using these baseline conditions, and other sites made use of parts of the baseline data, such as the vocabulary or the grammar.

Speech Corpora Working Group Report—Janet Baker

Speech Corpora Working Group Members. Janet Baker (Dragon, Chair); Francis Kubala (BBN); Doug Paul (Lincoln); Bob Weide (CMU); Mitch Weintraub (SRI).

Goal and Approach. The goals of this Working Group are to identify key research areas and provide resources not currently available, to stimulate and advance research and evaluation in continuous speech recognition (CSR) in the DARPA SLS Program. The Group's approach was to define a set of desired attributes for CSR corpora, and identify those attributes most important to the research groups in the SLS program.

Proposal for CSR Corpora. The Working Group is proposing collection of speech from two different application domains: (1) the HANSARD domain of Canadian parliamentary hearings; and (2) the CALS domain of logistical material in maintenance repair manuals for planes, tanks, and other military equipment. Both domains are supported by large text corpora on CD-ROM. HANSARD has government/political flavor, general English, large vocabulary, high perplexity. CALS has military flavor, specialized, modest vocabulary and perplexity. Read speech would be collected for both domains,

to support a variety of training and test conditions for CSR research.

Discussion Period

Francis Kubala (a member of the Speech Corpora Working Group) presented a dissenting point-of-view with respect to CSR corpora collection. He argued that to achieve the goal of operational human/machine interfaces, the most valuable data would be spontaneous evaluation test data from several domains, and suggested that the SLS demonstration domains should be used. He suggested several problems with Hansard (not a human/machine domain; read speech; single unfamiliar domain).

John Makhoul described a range of data collection scenarios for collecting goal-directed spontaneous speech. These scenarios include: (1) using a Wizard (as at SRI and TI); (2) using a typist, with a machine to interpret queries (as at MIT); and (3) using a real-time speech recognition (new suggestion) system in conjunction with a Wizard or typist. The third scenario was now becoming possible at sites (including BBN) which have real-time recognizers, and was proposed to have advantages both in realism and efficiency.

A good deal of discussion followed, giving various views on both the CSR corpus proposal, and on collection of speech corpora in general.