Multi-Site Data Collection and Evaluation in Spoken Language Understanding

*L. Hirschman, M. Bates, D. Dahl, W. Fisher, J. Garofolo, D. Pallett, K. Hunicke-Smith, P. Price, A. Rudnicky, and E. Tzoukermann**

Contact: Lynette Hirschman, NE43-643 SLS Group
MIT Laboratory for Computer Science, Cambridge, MA 02139
e-mail: lynette@goldilocks.lcs.mit.edu

ABSTRACT

The Air Travel Information System (ATIS) domain serves as the common task for DARPA spoken language system research and development. The approaches and results possible in this rapidly growing area are structured by available corpora, annotations of that data, and evaluation methods. Coordination of this crucial infrastructure is the charter of the Multi-Site ATIS Data COllection Working group (MADCOW). We focus here on selection of training and test data, evaluation of language understanding, and the continuing search for evaluation methods that will correlate well with expected performance of the technology in applications.

1. Introduction

Data availability and evaluation procedures structure research possibilities: the type and amount of training data affects the performance of existing algorithms and limits the development of new algorithms; and evaluation procedures document progress, and force research choices in a world of limited resources. The recent rapid progress in spoken language understanding owes much to our success in collecting and distributing a large corpus of speech, transcriptions, and associated materials based on human-machine interactions in the air travel domain. A tight feedback loop between evaluation methodology and evaluation results has encouraged incremental extension to the evaluation methodology, to keep pace with the technology development. The paper reports on the data collection and evaluation efforts co-ordinated by MADCOW over the past year.

The multi-site data collection paradigm [3, 4] distributes the burden of data collection, provides data rapidly, educates multiple sites about data collection issues, and results in a more diverse pool of data than could be obtained with a single collection site. The resulting data represents a wide range of variability in speaker characteristics, speech style, language style and interaction style. It has allowed individual sites to experiment with data collection methods: by replacing various system components with a human, we collect the kind of data we can aim for in the future, while completely automated systems help us to focus on the major current issues in system accuracy and speed. Sites have also experimented with interface strategies: spoken output only, tabular output only, response summaries, spoken and written paraphrase, and system initiative may be more or less appropriate for different users and different tasks and all can dramatically affect the resulting data.

MADCOW's recent accomplishments include:

- Release of 14,000 utterances for training and test, including speech and transcriptions;
- Release of annotations for almost 10,000 of these utterances (7500 training utterances and three test sets of 2300 utterances total), balanced by site;
- A bug reporting and bug fix mechanism, to maintain the quality and consistency of the training data;
- An evaluation schedule that delivered training data and froze changes in the principles of interpretation[1] several months before the evaluation;
- An experiment with "end-to-end" evaluation that permits evaluation of aspects of the system not previously evaluable.

Table 1 shows the breakdown of all training data and Table 2 shows the breakdown for just the annotated data.[2]

2. Current Evaluation Methodology

When the ATIS task was developed in 1990 [9], little work had been done on formal evaluation of understanding for natural language interfaces.[3] In the absence of a generally accepted semantic representation,

*This paper was written the auspices of the Multi-Site ATIS Data Collection Working group (MADCOW). In addition to the authors, many other people, listed under the Acknowledgements section, made important contributions to this work.

[1] These are the principles that define how various vague or difficult phrases are to be interpreted; see section 2.1 below.

[2] A class A utterance can be interpreted by itself, with no additional context; a class D utterance requires an earlier "context-setting" utterance for its interpretation; and a class X utterance cannot be evaluated in terms of a reference database answer.

[3] This coincides with the beginnings of formal evaluation for written text, via the Message Understanding Conferences (MUCs)

Site	Speakers	Scenarios	Utterances
AT&T	57	200	2100
BBN	62	307	2277
CMU	47	214	2219
MIT	182	625	4910
SRI	90	148	2478
TOTAL	438	1494	13984

Table 1: Multi-site ATIS Data Summary

Site	Class A	Class D	Class X	Total
ATT	457 36.6%	497 39.8%	295 23.6%	1249 16.6%
BBN	858 56.2%	357 23.4%	312 20.4%	1527 20.3%
CMU	562 37.6%	340 22.7%	594 39.7%	1496 19.9%
MIT	663 37.7%	680 38.7%	414 23.6%	1757 23.4%
SRI	676 45.7%	618 41.8%	184 12.4%	1478 19.7%
Total	3216 42.8%	2492 33.2%	1799 24.0%	7507 100.0%

Table 2: Annotated Training Data Summary

the DARPA SLS community focused instead on "the right answer," as defined in terms of a database query task (air travel planning). This permitted evaluation by comparing "canonical" database answers to the system answers using a comparator program [1]. This approach was felt to be far easier, given proper definition of terms, than to agree on a standard semantic representation.

The original evaluation methodology was defined only for context-independent (*class A*) utterances. However, this left approximately half the data as unevaluable (see Table 2). Over the next two years, the evaluation method was extended to cover context-dependent queries (*class D* utterances), it was tightened by requiring that a correct answer lie within a minimal answer and a maximal answer (see section 2.1), and it was made more realistic by presenting utterances in scenario order, as spoken during the data collection phase, with no information about the class of an utterance. Thus, we now can evaluate on approximately 75% of the data (all but *class X* data – see Tables 2 and 4). We also introduced a *Weighted Error* metric because we believed, at least in some applications, wrong answers might be worse than "no answer":[4]

$$WeightedError = \\ \#(No\_Answer) + 2 * \#(Wrong\_Answer).$$

2.1. The Evaluation Mechanism

The comparator-based evaluation method compares human annotator-generated canonical ("reference") database answers to system generated answers. The annotators first classify utterances into context-independent (A), context-dependent (D) and unevaluable (X) classes. Each evaluable utterance (class A or D) is then given minimal and maximal reference answers. The minimal reference answer is generated using NLParse[5] and the maximal answer is generated algorithmically from the minimal answer. A correct answer must include all of the tuples contained in the minimal answer and no more tuples than contained in the maximal answer.

The Principles of Interpretation document provides an explicit interpretation for vague natural language expressions, e.g., "red-eye flight" or "mid-afternoon," and specifies other factors necessary to define reference answers, e.g., how context can override ambiguity in certain cases, or how utterances should be classified if they depend on previous unevaluable utterances. This document serves as a point of common reference for the annotators and the system developers, and permits evaluation of sentences that otherwise would be too vague to have a well-defined database reference answer.

The initial Principles of Interpretation was implemented in 1990. The document is now about 10 pages long, and includes interpretation decisions based on some 10,000 ATIS utterances. The document continues to grow but at a significantly slower rate, as fewer new issues arise. It is remarkable that such a small document has sufficed to provide well-defined interpretations for a corpus of this size. This demonstrates that rules for the interpretation of natural language utterances, at least in the ATIS domain, can be codified well enough to support an automatic evaluation process. Because this procedure was explicit and well-documented, two new sites were able to participate in the most recent evaluation (November 1992).

2.2. Testing on the MADCOW Data

The test data selection procedure was designed to ensure a balanced test set. Test data for the November 1992 evaluation were chosen using procedures similar to those for the November 1991 test [3]. As sites submitted data to NIST, NIST set aside approximately 20% of the utterances to create a pool of potential test data; some 1200 utterances constituted the November 1991

[8]. The MUC evaluation uses a domain-specific template as the basis for evaluation. To date, the goal of a domain-independent semantic representation, perhaps analogous to the minimal bracketing of the Penn Treebank database [2] for parsing, remains elusive.

[4] A recent experiment [5] showed that for one system, subjects were able to detect a system error before making their next query in 90% of the cases. In the remaining 10%, a system error caused the subject to lose several turns before recovering, leading to a reduced estimated weighting factor of 1.25 for errors in that system.

[5] NLParse is a database access product of Texas Instruments.

test pool; 1300 utterances constituted the November 1992 test pool.

NIST's goal was to select approximately 1000 test utterances from the test data pool, evenly balanced among the five collection sites (AT&T, BBN, CMU, MIT, and SRI). Utterances were selected by session, i.e., utterances occurring in one problem-solving scenario were selected as a group, avoiding sessions that seemed to be extreme outliers (e.g., in number of class X utterances, total number of utterances, or number of repeated utterances). Because the test pool contained only marginally more utterances than were needed for the test, it was not possible to simultaneously balance the test set for number of speakers, gender, or subject-scenarios. The test set contained 1002 utterances.[6] The breakdown of the data is shown in Table 3.

NIST verified and corrected the original transcriptions. However, some uncertainty about the transcriptions remained, due to inadequacies in the specifications for the transcription of difficult-to-understand speech, such as *sotto voce* speech. After the transcriptions were verified, the data were annotated by the SRI annotation group to produce categorizations and reference answers. A period for adjudication followed the test, where testing sites could request changes to the test data categorizations, reference answers, and transcriptions. The final post-adjudication classification of the test data set is shown in Table 4. Final evaluation results are reported in [6].

Collecting Site	Speakers	Scenarios	Utterances
ATT	7; 1M/ 6F	22	200
BBN	7; 3M/ 4F	28	201
CMU	4; 4M/ 0F	12	200
MIT	10; 3M/ 7F	37	201
SRI	9; 5M/ 4F	19	200
Total	37; 16M/21F	118	1002

Table 3: Multi-site ATIS Test Data November 1992

3. Limitations of the Current Evaluation

The current data collection and evaluation paradigm captures important dimensions of system behavior. However, we must constantly re-assess our evaluation procedures in terms of our goals, to ensure that our evaluation procedures help us to assess the suitability of a particular technology for a particular application, and to ensure that benchmark scores will correlate well with user satisfaction and efficiency when the technology is transferred to an application.

The advantage of using a pre-recorded corpus for evaluation is clear: the same data are used as input to all systems under evaluation, and each system's set of answers is used to automatically generate a benchmark score. This approach ensures a uniform input across all systems and removes human involvement from the benchmark testing process (except that human annotators define the reference answers). Any annotated set of data can be used repeatedly for iterative training. However, some of these same strengths also impose limitations on what we can evaluate.

First, there is the issue of the match between the reference answer and the user's need for useful information. The comparator method can count answers as correct despite system misunderstanding. For example, if a system misrecognizes "Tuesday" as "Wednesday" and the user realizes that the flight information shown is for Wednesday flights, the user may appropriately believe that the answer is wrong. However, if all flights have daily departures, the database answer will be *canonically* correct. On the other hand, useful (but not strictly correct) answers will be counted wrong, because there is no "partially correct" category for answers.

Second, mixed initiative in human-machine dialogue will be required for technology transfer in many spoken language understanding applications. But the evaluation paradigm actively discourages experimentation with mixed initiative. A query that is a response to a system-initiated query is classified as unevaluable if the user's response can only be understood in the context of the system's query. During evaluation, any system response that is a query will automatically be counted as incorrect (since only database answers can be correct).

The use of pre-recorded data also preserves artifacts of the data collection system. For example, much of the test data were collected using systems or components of systems to generate responses, rather than a human alone. As a result, the data include many instances of system errors that affect the user's next query. A user may have to repeat a query several times, or may correct some error that the data collection system (but not the system under evaluation) made. These are artificial phenomena that would disappear if the data collection and evaluation systems were identical.

Finally, the current paradigm does not take into account the speed of the response, which greatly affects the overall interaction. Demonstration systems at several sites

[6] The data recorded using the Sennheiser head-mounted noise-cancelling microphone were used as the test material for "official" speech recognition (SPREC) and spoken language system (SLS, NL) testing. For a subset of the utterances in the official test sets, recordings were also made using a desk-mounted Crown microphone.

Site	Class A	Class D	Class X	Total
ATT	48 (24.0%)	41 (20.5%)	111 (55.5%)	200 (20.0%)
BBN	97 (48.3%)	27 (13.4%)	77 (38.3%)	201 (20.1%)
CMU	76 (38.0%)	66 (33.0%)	58 (29.0%)	200 (20.0%)
MIT	100 (49.8%)	67 (33.3%)	34 (16.9%)	201 (20.1%)
SRI	106 (53.0%)	46 (23.0%)	48 (24.0%)	200 (20.0%)
Total:	427 (42.6%)	247 (24.7%)	328 (32.7%)	1002 (100.0%)

Table 4: Breakdown of Test Data by Class

have begun to diverge from those used in benchmark evaluations, in part, because the requirements of demonstrating or using the system are quite different from the requirements for generating reference database answers.

These limitations of the comparator-based evaluation preclude the evaluation of certain strategies that are likely to be crucial in technology transfer. In particular, we need to develop metrics that keep human subjects in the loop and support human-machine interaction. However, the use of human subjects introduces new issues in experimental design. Over the past year, MADCOW has begun to address these issues by designing a trial *end-to-end* evaluation.

4. End-to-End Evaluation Experiment

The end-to-end evaluation, designed to complement the comparator-based evaluation, included 1) objective measures such as timing information, and time to task completion, 2) human-derived judgements on correctness of system answers and user solutions (*logfile evaluation*), and 3) a user satisfaction questionnaire.

The unit of analysis for the new evaluation was a scenario, as completed by a single subject, using a particular system. This kept the user in the loop, permitting each system to be evaluated on its own inputs and outputs. The use of human evaluators allowed for assessing partial correctness, and provided the opportunity to score other system actions, such as mixed initiatives, error responses and diagnostic messages. The end-to-end evaluation included both task-level metrics (whether scenarios had been solved correctly and the time it took a subject to solve a scenario) and utterance-level metrics (query characteristics, system response characteristics, the durations of individual transactions).

An experimental evaluation took place in October 1992, to assess feasibility of the new evaluation method. We defined a common experimental design protocol and a common set of subject instructions (allowing some local variation). Each site submitted to NIST four travel planning scenarios that had a well-defined "solution set". ¿From these, NIST assembled two sets of four scenarios. Each site then ran eight subjects, each doing four scenarios, in a counter-balanced design. Five systems participated: the BBN, CMU, MIT and SRI spoken language systems, and the Paramax system using typed input.

A novel feature of the end-to-end experiment was the *logfile evaluation.* This technique, developed at MIT [7], is based on the logfile which records and timestamps all user/system interactions. A human evaluator, using an interactive program,[7] can review each user/system interaction and evaluate it by type of user request, type of system response, and correctness or appropriateness of response. For user requests, the following responses were distinguished: 1) *New Information,* 2) *Repeat,* 3) *Rephrase,* or 4) *Unevaluable.* For system responses, the evaluators categorized each response as follows:

Answer: further evaluated as *Correct, Incorrect Partially Correct or Can't Decide;*
System Initiated Directive: further evaluated as *Appropriate, Inappropriate, or Can't Decide;*
Diagnostic Message: further evaluated as *Appropriate, Inappropriate, or Can't Decide;*
Failure-to-Understand Message: no further evaluation.

The evaluator also assessed the scenario solution according to whether the subject finished and whether the answer belonged to the defined solution set.

To facilitate determination of the correctness of individual system responses, we agreed to follow the Principles of Interpretation, at least to the extent that an answer judged correct by these Principles would not be counted incorrect. For this experiment, logfile evaluation was performed independently by Bill Fisher (NIST) and Kate Hunicke-Smith (SRI Annotation), as well as by volunteers at MIT and BBN. This gave us experience in looking at the variability among evaluators of different levels of experience. We found that any two evaluators

[7]The program was developed by David Goodine at MIT; the evaluator instructions were written by Lynette Hirschman, with help from Lyn Bates, Christine Pao and the rest of MADCOW.

agreed about 90% of the time, and agreement among multiple evaluators decreased proportionally.

5. Lessons Learned

The experiment provided useful feedback on the risks and advantages of end-to-end evaluation, and will guide us in refining the procedure. For the initial trial, we made methodological compromises in several areas: a small number of subjects, no control over cross-site subject variability, few guidelines in developing or selecting scenarios. These compromises seemed reasonable to get the experiment started; however, the next iteration of end-to-end evaluation will need to introduce methodological changes to provide statistically valid data.

5.1. Sources of Variability

Valid comparisons of systems across sites require control over major sources of variability, so that the differences of interest can emerge. The use of human subjects in the evaluation creates a major source of variability, due to differences in the subjects pools available at various sites and the characteristics of individuals. We can minimize some of these differences, for example, by training all subjects to the same criterion across sites (to account for differences in background and familiarity with the domain), by using many subjects from each site (so that any one subject's idiosyncrasies have less of an effect on the results), and by ensuring that procedures for subject recruitment and data collection across sites are as similar as possible (we made a serious effort in this direction, but more could be done to reduce the cross-site variability that is otherwise confounded with the system under evaluation). An alternative would be to perform the evaluation at a common site. This would allow for greater uniformity in the data collection procedure, it could increase the uniformity of the subject pool, and would allow use of powerful experimental techniques (such as within-subject designs). Such a common-site evaluation, however, would pose other challenges, including the port of each system to a common site and platform, and the complex design needed to assess potential scenario order effects, system order effects, and their interaction.

Another source of variability is the set of travel planning scenarios the subjects were asked to solve. Certain scenarios posed serious problems for all systems; a few scenarios posed particular problems for specific systems. However, the data suggest that there was a subset that could perform a reasonable diagnostic function.

5.2. Logfile Evaluation

Somewhat unexpectedly, we found that logfile evaluation was a useful tool for system developers in identifying dialogue-related problems in their systems. The evaluator interface allowed for rapid evaluation (about 5-15 minutes per scenario). However, the evaluator instructions need refinement, the interface needs minor extensions, and most important, we need to design a procedure to produce a statistically reliable logfile evaluation score by combining assessments from multiple evaluators.

A remaining thorny problem is the definition of *correct*, *partially correct*, and *incorrect answers*. For this experiment, we used the Principles of Interpretation document to define a correct answer, so that we would not need to develop a new document for these purposes. For the next evaluation, we need definitions that reflect utility to the user, not just canonical correctness.[8]

Finally, we found that we could not rely on subjects to correctly complete the scenarios presented to them. In some cases, the subject was not able to find the answer, and in other cases, the subject did not follow directions regarding what information to provide in the answer. This made it difficult to compute accurate statistics for scenario-level metrics such as task completion and task completion time; this problem was exacerbated by the limited amount of data we collected.

6. Future Directions

We view evaluation as iterative; at each evaluation, we assess our procedures and try to improve them. The comparator-based evaluation is now stable and the November 1992 evaluation ran very smoothly. The availability of an expanded database will require a new data collection effort. Increasing emphasis on portability may have an impact on evaluation technology. In addition, we plan to continue our experiments with end-to-end evaluation, to work out some of the methodological problems described in the previous section.

The ATIS *relational database* has been expanded from 11 cities to 46 cities, to provide a more realistic task supporting more challenging scenarios. The database was constructed using data from the Official Airline Guide and now includes 23,457 flights (compared to 765 flights). The set of new cities was limited to 46 because it was felt that a larger set would result in an unwieldy database and would thus require the sites to devote too many resources to issues peripheral to their research, such as database management and query optimization. Data

[8] Originally, we had wanted to compare logfile scores to comparator-based scores. However, for several sites, the data collection and evaluation systems had diverged and it was not possible to simultaneously interact with the subjects and provide comparator-style answers. Therefore, we were not able to perform this experiment.

collection on this larger database is now beginning.

The *portability* of the technology (from application to application, and from language to language) becomes an increasing challenge as the technology improves, since more potential applications become possible. It still takes many hours of data collection and several person months of system development to port an application from one domain (e.g., air travel) to another similar domain (e.g., schedule management). Evaluating portability is still more challenging. Evaluation has a significant cost: the comparator-based method requires the definition of a training corpus and its collection, defining principles of interpretation, and (most expensively) the annotation of data. Therefore, if we believe that regular evaluations play an important role in guiding research, we need to find cost-effective ways of evaluating systems. End-to-end evaluation may provide some low-overhead techniques for quickly evaluating system performance in new domains.

With the *end-to-end evaluation* experiment, we have made progress in creating a procedure that accurately assesses the usability of current spoken language technology and provides useful feedback for the improvement of this technology. The procedure needs to be further refined to reliably identify differences among systems and it must embody principles that can assess strengths and weaknesses of different systems for different purposes. In developing evaluation procedures that involve human interactions, we need to carefully assess the validity of the measures we use. For example a measure such as the number of utterances per scenario may seem relevant (e.g., the subject was frustrated with answers and had to repeat a question several times), but in fact may reflect irrelevant aspects of the process (the subject was intrigued by the system and wanted to push its limits in various ways). Meaningful evaluation will require metrics that have been systematically investigated and have been shown to measure relevant properties.

MADCOW has played a central role in developing and coordinating the multi-site data collection and evaluation paradigm. It will also play an active role in defining new methodologies, such as end-to-end evaluation, to support evaluation of interactive spoken language systems. We believe that end-to-end evaluation will allow us to assess the trade-offs among various component-level decisions (in speech recognition, natural language processing and interface design), bringing spoken language systems closer to eventual deployment.

7. ACKNOWLEDGEMENTS

We would particularly like to acknowledge the contribution of Nancy Dahlgren at NIST; prior to her accident, Nancy made important contributions to the annotation and debugging of the data. We greatly missed her participation during the final evaluation.

In addition to the authors, the following people made a valuable contribution to the process: at ATT: J. Wilpon; at BBN: R. Bobrow, R. Ingria, J. Makhoul, V. Shaked, and D. Stallard; at CMU: C. Neelan, E. Thayer, and R. Weide; at MIT: D. Goodine, J. Polifroni, C. Pao, M. Phillips, and S. Seneff; at NIST: N. Dahlgren, J. Fiscus, and B. Tjaden; at Paramax: L. Norton, and R. Nilson; and at SRI: H. Bratt, R. Moore, E. Shriberg, and E. Wade.

References

1. Bates, M., S. Boisen, and J. Makhoul, "Developing an Evaluation Methodology for Spoken Language Systems," *Proc. Third DARPA Speech and Language Workshop*, R. Stern (ed.), Morgan Kaufmann, June 1990.
2. Black, E., *et al.*, "A Procedure for Quantitatively Comparing the Syntactic Coverage of English Grammars," *Proc. Third DARPA Speech and Language Workshop*, P. Price (ed.), Morgan Kaufmann, June 1991.
3. Hirschman, L., *et al.*, "Multi-Site Data Collection for a Spoken Language Corpus", *Proc. Fifth Speech and Natural Language Workshop*, M. Marcus (ed.), Morgan Kaufmann, Arden House, NY, February 1992.
4. Hirschman, L., *et al.*, "Multi-Site Data Collection for a Spoken Language Corpus," *Proc. of the ICSLP*, Banff, Canada, October 1992.
5. Hirschman, L. and C. Pao, "The Cost of Errors in a Spoken Language Systems," submitted to Eurospeech-93, Berlin 1993.
6. Pallett, D., Fiscus, J., Fisher, W., and J. Garofolo, "Benchmark Tests for the Spoken Language Program," *Proc. DARPA Human Language Technology Workshop*, Princeton, March 1993.
7. Polifroni, J., Hirschman, L., Seneff, S. and V. Zue, "Experiments in Evaluating Interactive Spoken Language Systems" *Proc. DARPA Speech and Natural Language Workshop*, M. Marcus (ed.), Arden House, NY, February 1992.
8. *Proc. Fourth Message Understanding Conf.*, Morgan Kaufmann, McLean, June 1992.
9. Price P., "Evaluation of Spoken Language Systems: The ATIS Domain," *Proc. Third DARPA Speech and Language Workshop*, P. Price (ed.), Morgan Kaufmann, June 1990.