Multi-Site Data Collection for a Spoken Language Corpus

MADCOW *

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

ABSTRACT

This paper describes a recently collected spoken language corpus for the ATIS (Air Travel Information System) domain. This data collection effort has been co-ordinated by MADCOW (Multi-site ATIS Data COllection Working group). We summarize the motivation for this effort, the goals, the implementation of a multi-site data collection paradigm, and the accomplishments of MADCOW in monitoring the collection and distribution of 12,000 utterances of spontaneous speech from five sites for use in a multi-site common evaluation of speech, natural language and spoken language.

1. Introduction

Following the February 1991 DARPA Speech and Natural Language Workshop, the DARPA Spoken Language contractors decided to institute a multi-site data collection paradigm in order to:

- support a common evaluation on speech, natural language and spoken language;
- maximize the amount of data collected;
- provide some diversity in data collection paradigms;
- reduce cost to any one site by sharing the data collection activity across multiple participating sites.

To co-ordinate this effort, MADCOW was formed in May 1991 with a representative from each of the participating sites. This included the six sites planning to collect and evaluate on the data: AT&T, BBN, CMU, MIT, SRI and Paramax (formerly Unisys), it included NIST, which was responsible for data validation, distribution and selection and scoring of test material, and it included the Annotation group at SRI, responsible under a separate contract for annotating the data with database reference answers.

The charter of MADCOW was to implement the multi-site data collection paradigm, to monitor the distribution of the data, and to agree on a test paradigm for the multi-site data. The original goals for the data collection activity were to collect 10,000 training utterances and 1,000 test utterances, plus material for a dry-run test to be held in October 1991. Between May 1991 and February 1992, the following data have been collected under the MADCOW effort:

- 10,400 training utterances collected from 280 speakers at 5 sites, with speech and transcriptions disseminated to testing sites;
- 5,600 utterances annotated with database reference answers, distributed to testing sites;
- 300 annotated utterances used for October dry-run test;
- 1,000 annotated utterances used for the February 1992 test;
- 1,000 utterances set aside for a later test.

Significant data collection and evaluation infrastructure were already in place prior to the formation of MADCOW. This included the definition of the air travel planning task [11], the database (a relational version of an eleven city subset of the Official Airline Guide, containing airline, flight and ground transportation information, initially set up by C. Hemphill at TI and revised and extended by R. Moore and others at SRI), a comparator-based evaluation methodology for comparing database tuples to reference answers [1, 12], and several earlier ATIS corpora collected at TI [2] and SRI.

To implement the multi-site data collection effort, each site agreed to collect a corpus of 2200 utterances and to provide this corpus to NIST in a standard format, including speech data, transcriptions, and a logfile recording the subject's interaction with the data collection system.

David Pallett's group at NIST was responsible for validation and distribution of the training data as well as for

*This paper was written under the auspices of the Multi-Site ATIS Data Collection Working group (MADCOW) by L. Hirschman, M. Bates, D. Dahl, W. Fisher, J. Garofolo, K. Hunicke-Smith, D. Pallett, C. Pao, P. Price, and A. Rudnicky. In addition, many other people made important contributions to this work and are listed in the Acknowledgements section.

running the common evaluation. As data were submitted, NIST checked the data for conformity to the standard formats and randomly set aside 20% of each site's incoming data for test sets. For the common evaluation, NIST was responsible for the release of the test data, and the collection, scoring and analysis of the results, as well as for adjudication of questions about reference answers on the Spoken Language and Natural Language tests.

The Annotation group under Jared Bernstein at SRI was responsible for providing the database reference answers and for categorization of the data into context-independent (class A), context-dependent (class D) and unanswerable (class X) utterances. To facilitate timely agreement on specific issues, a special subgroup, chaired by Deborah Dahl, was formed under MADCOW, with responsibility for the Principles of Interpretation[1].

2. Collecting the Data

Data collection procedures were not standardized across sites. We know that variation in these procedures can lead to vast differences in the resulting data. Though standardizing is often important and has played a crucial role in other areas of the DARPA SLS program, it is also difficult and costly. Spoken language understanding as a human-computer interface tool is a new technology, and the space of potential variations is enormous and largely unexplored. We therefore chose to sample various points in this space, and to document the differences. This decision may be revised as we learn from this experiment. We outline in this section those aspects of data collection shared by all systems, and provide a separate section for each data collection site to highlight the unique aspects of that site.

The original data collected at TI and SRI used two human "wizards." As the subject spoke a sentence, one person provided a fast transcription, while the other used NLParse[2] to generate an SQL query to access the database. At all sites subjects were led to believe they are talking to a fully automated system. For data collected at SRI, this was true; all other sites used some automatic speech recognition and/or natural language understanding, with varying amounts of human transcription and error correction. AT&T used only audio outputs; all other sites used a computer screen to display tables of data and other information to the subject. The two standard microphones were the Sennheiser HMD-410 close talking microphone and the Crown PCC-160 table-top microphone. Table 1 shows the total data collected by site, including training and test data[3].

```
You have only three days for job hunting, and you have
arranged job interviews in two different cities! (The
interview times will depend on your flight schedule.)
Start from City-A and plan the flight and ground
transportation itinerary to City-B and City-C, and back
home to City-A.
```

Figure 1: Common ATIS scenario

All sites used a set of air travel planning "scenarios" (problems) for subjects to solve; BBN supplemented these with problems more like general database query tasks. The scenarios varied greatly in complexity and in the number of queries required for a solution. For sites using a wizard, the wizard was constrained in behavior, and did not represent human-like capabilities, though the wizard's role varied from site to site. By agreement, one "common scenario" was designated, shown in Figure 1, and sites agreed to collect 10% of their data using this common scenario.

All sites (except BBN) used a debriefing questionnaire which explained the nature of the experiment, unveiled the deception of the wizard, and elicited comments from the subject on the experience. All sites automatically generated logfiles documenting subject queries, system responses and time stamps for all key events. A sample log file is shown in Figure 2; the user input is marked as "Utterance", the SQL is marked as "Query", the wizard input as "Sentence", the system display as "Result", and there are timestamps to mark when speech recording begins, when the sentence is sent for processing, and when the answer is returned.

Site	Speakers	Scenarios	Utterances
AT&T	50	176	1887
BBN	62	307	2277
CMU	43	196	2480
MIT	75	250	2265
MIT: old DB	96	320	2940
SRI	72	134	2301
TOTAL	398	1383	14150

Table 1: Multi-site ATIS data summary

2.1. BBN Data Collection

The BBN data collection setup employed an interactive subject and wizard interface based on X-windows. The

[1] The Principles of Interpretation are the set of rules governing the interpretation of various types of queries, e.g, the meaning of "around 10 AM", or the definition of what constitutes a "red-eye flight"; see the *Principles of Interpretation* section below.

[2] NLParse is a Texas Instruments propriety system, made available to the DARPA research community for the ATIS application

[3] The numbers in this table reflect total data collected, which differs from the amount of data *released* by NIST, quoted in the previous section.

```
[UtteranceID:] 1
[Timestamp: Sent speech for utterance 1 at 10:38:36]
[Begin Utterance: 1]
do you have any flights from Pittsburgh to Boston on
Wednesday of next week . in the morning
[End Utterance: 1]

[Begin Sentence: 1]
pit to boston wednesday morning
[End Sentence: 1]

[Timestamp: Sent sentence for utterance 2 at 10:38:56]
[Begin Query: 1]
select distinct flight.airline_flight,flight.from_airport,
  flight.to_airport,flight.departure_time,flight.arrival_time,
  flight.time_elapsed,flight.stops from flight
where flight.from_airport in ('PIT')
  and flight.to_airport in ('BOS')
  and((flight_days like '%WE%'
  and flight_days not like 'NOT%')
    or (flight_days like 'DAILY'))
  and(flight.departure_time<=1200)
[End Query: 1]

[Begin Result: 1]
------------------------------------------------------------
AIRL#    FROM  TO    LEAVE ARRIVE  DURA  STOPS
------------------------------------------------------------
US674    PIT   BOS   1200   1328    88    0
US732    PIT   BOS    710    839    89    0
US736    PIT   BOS    840   1006    86    0
[End Result: 1]

[Timestamp: Sent answer for utterance 1 at 10:39:00]
```

Figure 2: Sample log file (excerpt)

subject's queries and answers were stacked on the color screen for later examination or other manipulation by the subject. The system also used BBN's real-time BYBLOS speech recognition system as the front-end; the wizard had the choice of using the speech recognition output or correcting it. This choice allowed the wizard to give feedback (in terms of errorful speech recognition) to the subject that may have encouraged the subject to speak more clearly. Certainly there would be such feedback in a real system.

The scenarios included not only trip planning scenarios, but also problem solving scenarios involving more general kinds of database access, e.g., finding the hub city for an airline X. This was done to try to elicit a richer range of language use.

2.2. CMU Data Collection

The Carnegie Mellon University (CMU) data collection system incorporated a working ATIS system [15] and a wizard. The subject sat at a computer that displayed a window containing system output, and another window that acted as an interface to the "recognition" system which used a push-and-hold protocol to record speech. Two channels of data were recorded, using both the Sennheiser and the Crown microphones. An Ariel DM-N digitizer and a Symetrix 202 microphone pre-amplifier completed the equipment. The wizard, sitting two cubicles away in an open-plan lab, listened to the subject directly through headphones. A modified version of the CMU ATIS system was used to assist the wizard in database access. The wizard could paraphrase the subject's query or correct recognition errors before database access. Retrieved information was previewed by the wizard before being sent to the subject's display. The wizard also had available a set of standard "error" replies to be sent to the subject when appropriate (e.g., when the subject asked questions outside the domain).

Subjects were recruited from the university environment; they ranged in age from 18 to 38, with a mean of 24 years. The subjects were introduced to the system by an experimenter who explained the procedure and sat with the subject during the first scenario. Standard air travel planning scenarios were used. The experimenter then left the enclosure, but was available if problems arose. Subjects completed as many scenarios as fit into an hour-long session. A maximum of 6 scenarios were available; an average of 4.6 were completed in the data collected to date.

2.3. MIT Data Collection

The MIT data collection paradigm emphasized interactive data collection and dialogue, using the MIT ATIS system [10, 13]. Data were collected by asking subjects to solve scenarios using the system; the experimenter sat in another room and transcribed a "clean" version of the subject's speech input. The transcriber eliminated hesitations, "ums" and false starts, but otherwise simply transmitted a transcription of what the subject said. The natural language component then translated the transcribed input into a database query and returned the display to the user. The MIT system produced several forms of output for the subject, including a summary of the question being answered (in both written and spoken form) and a reformatted tabular display without cryptic abbreviations. The system also supported a capability for system-initiated clarification dialogue to handle cases where the user underspecified a query. For example, if the user specified only a destination, the system would ask where the subject was departing from.

Subjects were recruited mainly from MIT and consisted of undergraduates, graduate students and employees. Each subject was given a $10 gift certificate to a local store. A data collection session lasted approximately 45 minutes; it included an introduction by the experimenter (who also acted as transcriber); practice with the push-and-hold-to-talk mechanism; the solution of three or four scenarios (moving from simple scenarios to more complex ones involving booking a flight); and completion of a debriefing questionnaire. The data were collected in an office-noise environment using an Ariel Pro-Port A/D system connected to a Sun Sparcstation.

2.4. AT&T Data Collection

The AT&T ATIS data were collected using a partially simulated, speech-in/speech-out spoken language system [9]. The natural language and database access components of the AT&T system were essentially identical to those of the MIT ATIS system [10]. The interface with the the subject was designed to simulate an actual telephone-based dialogue: the system provided all information in the form of synthesized speech, as opposed to displaying information on a computer terminal. Speech data were captured simultaneously using (1) the Sennheiser microphone amplified by a Shure FP11 microphone-to-line amplifier, and (2) a standard carbon button-based telephone handset (over local telephone lines). Digitization was performed by an Ariel Pro-Port A/D system.

Before each recording session, the experimenter provided the subject with a brief verbal explanation of the task, a page of written instructions, a summary of the ATIS database domain, and a list of travel planning scenarios. The system initiated the dialogue at the beginning of the recording session, and responded after every utterance with information or with an error message. The experimenter controlled recording from the keyboard, starting recording as soon as the system response ended, and stopping recording when the subject appeared to have completed a sentence. The experimenter then transcribed what the subject said, excluding false starts, and sent the transcription to the system, which automatically generated the synthesized response. A complete session lasted about an hour, including initial instruction, a two-part recording session with a five minute break, and a debriefing questionnaire.

Subjects for data collection were recruited from local civic organizations, and collection took place during working hours. As a result, 82 percent of the subjects were female, and subjects ranged in age from 29 to 77, with a median age of 55. In return for each subject's participation, a donation was made to the civic organization through which he or she was recruited.

2.5. SRI Data Collection

The SRI data collection system used SRI's SLS system; there was no wizard in the loop. The basic characteristics of the DECIPHER speech recognition component are described in [4], [6], and the basic characteristics of the natural language understanding component are described in [3]. Two channels of data were recorded, using both the Sennheiser and the Crown microphones. Subjects clicked a mouse button to talk, and the system decided when the utterance was complete. The data were collected in an office-noise environment, using a Sonitech Spirit-30 DSP board for A/D connected to a Sun Sparcstation.

Subjects were recruited from other groups at SRI, from a nearby university, and from a volunteer organization. They were given a brief overview of the system and its capabilities, and were then asked to solve one or several air travel planning scenarios. The interface allowed the user to move to the context of a previous question. Some subjects used the real-time hardware version of the DECIPHER system [5], [16]; others used the software version of the system. Other parameters that were varied included: instructions to subjects regarding what they should do when the system made errors, the interface to the context-setting mechanism, and the number of scenarios and sessions. See [14] for details on the interface and the conditions that were varied from subject to subject.

3. Distributing the Data

During the MADCOW collection effort, NIST was primarily responsible for two steps in the data pipeline: (1) quality control and distribution of "initial" unannotated data received from the collection sites; and (2) quality control and distribution of annotated data from the SRI annotators.

3.1. Distribution of Initial Data

Initial (unannotated) data were received on 8mm tar-formatted tapes from the collection sites, logged into the file "madcow-tapes.log", and placed in queue for distribution. The initial data consisted of a .log file for each subject-scenario, and .wav (NIST-headered speech waveform with standard header fields) and .sro (speech recognition detailed transcription) files for each utterance. The 8mm tapes were downloaded and the initial data and filename/directory structure were verified for format compliance using a suite of shell program verification programs. Non-compliant data were either fixed at NIST or returned to the site for correction, depending on the degree and number of problems. Twenty percent of the utterances from each collection site was then set aside as potential test data. The remaining data for training were assigned an initial release ID (date) and the textual non-waveform data were then made available to the collection and annotation sites via anonymous ftp. The tape log file, "madcow-tapes.log" was updated with the release date. A cumulative lexicon in the file "lexicon.doc.<DATE>" was also updated with each new release. During the peak of data collection activity, these releases occurred at weekly intervals. When enough waveforms (.wav) had accumulated to fill a CD-ROM (630 Mb), the waveforms were premastered on an ISO-9660 8mm tape which was then sent to MIT for "one-off" (recordable) CD-ROM production. Upon receipt of each CD-ROM from MIT, the initial release ID(s) of the data on the CD-ROM were recorded in the file "madcow-waves.log", and the CD-ROMs were shipped overnight to the MADCOW sites.

3.2. Distribution of Annotated Data

Annotated data from SRI were downloaded at NIST via ftp. The data were organized by initial release date in the standard ATIS file and directory structure and contained files for the query categorization (.cat), wizard input to NLParse (.win)[4], the SQL for the minimal answer (.sql), the SQL for the maximal answer (.sq2, generated from the minimal SQL) and the corresponding minimal and maximal reference answers (.ref, .rf2).

The .cat, .ref, and .rf2 files in the release were verified for format compliance using a suite of verification programs. A classification summary was then generated for the release and the data made available to the MADCOW sites via anonymous ftp. The "madcow-answers.log" file was updated with the release date.

3.3. Data Distribution Summary

Table 2 shows a summary by site and class of the annotated MADCOW data distributed by NIST as of December 20, 1991.

Site	Class A	Class X	Class D	Total
ATT	164 34.8%	118 25.1%	189 40.1%	471 8.4%
BBN	850 55.6%	334 21.8%	345 22.6%	1529 27.3%
CMU	430 38.3%	403 35.9%	289 25.8%	1122 20.1%
MIT	671 38.2%	406 23.1%	680 38.7%	1757 31.4%
SRI	335 46.8%	82 11.5%	299 41.8%	716 12.8%
Total	2450 43.8%	1343 24.0%	1802 32.2%	5595 100.0%

Table 2: Distribution of the annotated training data summary

3.4. Common Documentation

To facilitate common data exchange, MADCOW developed a set of documents which specify the formats for each common file type, listed below:

- .wav - NIST-headered speech waveform with standard header fields
- .log - Session log
- .sro - SR-output detailed transcription
- .cat - Query categorization
- .ref - Minimal reference answer
- .rf2 - Maximal reference answer

In addition, documentation was developed to specify directory and filename structures, as well as file contents. To insure conformity, NIST created and distributed format verification software for each file type and for directory/filename structures. The specifications documents and verification software are maintained for public distribution in NIST's anonymous ftp directory. NIST also maintains documentation for the transcription conventions, logfile formats, categorization principles and principles of database interpretation, also available in NIST's anonymous ftp directory.

To track the flow of data through the distribution "pipeline" during data collection, NIST maintained and published the data flow logs and documentation modifications in weekly electronic mail reports to MADCOW.

4. The Evaluation Paradigm

The diversity of data collection paradigms was a concern for MADCOW. To control for potential effects introduced by this diversity, it was agreed that test sets would consist of comparable amounts of data from each site (regardless of the amount of training material available from that site). In addition, benchmark test results would be displayed in an $N * M$ matrix form (for the N systems under test from the M data collecting sites). For the February 1992 tests, the number of collecting sites (M) was 5. This format was intended to indicate if data from one collecting site were "outliers" and whether a site performed particularly well on locally collected data.

The February 1992 Evaluation required sites to generate answers for data presented in units consisting of a "subject-scenario". The utterances from the scenario

[4] This was used to produce the minimal SQL query; see section on Annotation

were presented in sequence, with no annotation as to the class of the utterances. For scoring purposes, as in previous ATIS Benchmark tests [7], test queries were grouped into several classes on the basis of annotations. Results for the context-independent sentences (Class A) and context-dependent sentences (class D) were computed and tabulated separately, along with an overall score (A + D). Class X queries ("unanswerable" queries) were not included in the NL or SLS tests, but were included in the SPREC tests (since valid .lsn transcriptions existed for these utterances). The matrix tabulations reported on % correct, % incorrect and % weighted error[5] defined as $[2 * (\%False) + (\%No\_Answer)]$.

The February 1992 results also reflected a new method of computing answer "correctness" using both a minimal and a maximal database reference answer. The objective was to guard against "overgeneration": getting answers correct by including all possible facts about a given flight or fare, rather than by understanding what specific information was requested. This method (proposed by R. Moore and implemented by Moore and E. Jackson of SRI) specified the maximum relevant information for any query, and required that the correct answer contain at least the minimal correct information, and no more than the maximum. This method was first used during the October 1991 "dry run" and was adopted as the standard scoring procedure by the DARPA Spoken Language Coordinating Committee.

Three types of performance assessment tests were computed on the ATIS MADCOW Benchmark Test Data: SPeech RECognition (SPREC), Natural Language (NL), and Spoken Language Systems (SLS) tests. Details of these tests, and a summary of "official" reported results, are to be found elsewhere in these Proceedings [8].

5. Annotation

The goal of annotation was to classify utterances and provide database reference answers for the subjects' queries in the ATIS domain. These reference answers were used by the system developers and by NIST to evaluate the responses of the MADCOW natural language and spoken language systems.

The annotators began with the transcribed .sro files, and determined the possible interpretations of each utterance, classifying them as one of the following:

- A: context-independent
- D: context-dependent (classification includes tag(s) pointing to the context-setting query or queries).
- X: unevaluable (explanation provided by a tag in the classification).

```
.sro #1: show me flights from Pittsburgh to Boston on september
fourth in the morning
.cat #1: A:
.win #1: List morning flights from Pittsburgh and to Boston and
flying on 9/4/91
.ref #1: (138860 138861 138862)
.rf2 #1: Very long; may contain the following information:
flight ID, flight#, airline, times, date, day name, frequency, city
names, city codes, airport codes.
.sro #2:  what classes of service are there on flight U S seven
thirty ˜
.cat #2:  X: trunc-utt
.sro #3:  are there meals on that flight
.cat #3:  X: context-dependent:  Q002
.sro #4:  are there meals on U S seven thirty two
.cat #4:  D: testably-ambiguous
interp#1: yes/no context-dep:Q1
interp#2: wh-ques context-dep:Q1
.win #4:  List food services served on flights from Pittsburgh and
to Boston and flying on 9/4/91 and whose airline code is US and
whose flight number is 732
.ref #4:  (YES OR
(("B" 1 "COACH")
( "B" 1 "FIRST")))
.rf2 #4:
(YES OR
(("B"  1  "COACH"  "PIT"  "BOS"  "WED"  "US"  732  "PPIT
"BBOS" 9/4/91 138860 "PITTSBURGH" "BOSTON" "DAILY")
("B"  1  "FIRST"  "PIT"  "BOS"  "WED"  "US"  732  "PPIT"
"BBOS" 9/4/91 138860 "PITTSBURGH" "BOSTON""DAILY")
))
```

Figure 3: Annotation files from a sample ATIS session

Those utterances which were evaluable (class A or D) were translated into an English-like form (.win for wizard input) that could be interpreted by NLParse, a menu-driven program that converts English-like sentences into database queries expressed in SQL. Annotation decisions about how to translate the .sros were guided by the "Principles of Interpretation" (see the next section). After the .sro form of an utterance was classified and translated, the work was checked thoroughly by another annotator and by various checking programs. NLParse was then run to generate an SQL form in a .sql file. Finally a series of batch programs was run on each .sql file to produce the minimal and maximal reference answers (.ref and .rf2 files) for the corresponding utterance.

Figure 3 shows the annotation files created for a sample ATIS dialogue. Each line in italics identifies the file; the .sro file is the input; the .cat, .win, .ref and .rf2 files are created during the annotation procedure. Sentence #1 is class A, and has as its minimal reference answer the set of flight IDs for flights meeting the constraints. The maximal answer contains all of the columns used in the .sql query to constrain the answer; the answer is too

[5] This single number performance measure was first introduced for the February 1991 tests.

large to be displayed here. The .sro for sentence #2 ends with a truncation (marked by a tilde ~), which causes it to be classified as X (unevaluable). Thus no .win, .ref or .rf2 files are generated. Sentence #3 is a context-dependent utterance, due to the anaphoric expression *that flight.* It depends on #2, but since #2 is class X, #3 is also classified as X, following the principle that anything that depends on a class X (unevaluable) sentence must itself be unevaluable. Finally, sentence #4 is a yes-no question, which may have two answers: either *YES* or the set of entities satisfying the constraints. This sentence is also context-dependent, since it refers to flight US 732 between Pittsburgh and Boston. (Flight 732 may go to other cities, thus context is needed to establish the segment of interest). The minimal reference answer to the question about meals is defined to be the triple (meal,number,class). The maximal answer can include any information used in the .sql to generate the minimal answer.

6. Principles of Interpretation

In order to carry out an objective evaluation, it was necessary to be able to say whether an answer was right or wrong. In turn, deciding on the right answer often depended on how particular words and constructions in the query were interpreted. Thus, it was recognized early on in the development of the ATIS common task that it would be necessary to agree on specific definitions for certain vague expressions. In addition, given the current database, there was often more than one reasonable way of relating particular queries to the database. To insure objectivity in the evaluation, decisions about how to interpret queries had to be documented in such a way that all participants in the evaluation had access to them. The Principles of Interpretation document describes the interpretation of queries with respect to the ATIS database. This document was used both by system developers to train their systems and by the annotators for developing reference answers.

Examples of decisions in the Principles of Interpretation include: the meaning of terms like *early morning*, classification of a snack as a meal for the purposes of answering questions about meals, and the meaning of constructions such as *between X and Y*, defined for ATIS to mean "from X to Y".

A subgroup on the Principles of Interpretation was formed to discuss and make decisions on new issues of interpretation as they arose data collection and annotation. A representative from each site served on this subgroup. This insured that all sites were notified when changes or additions occurred in the Principles, and allowed each site to have input into the decision process. It was important to make careful decisions, because any revision could cause previously annotated data to become inconsistent with the revised Principles of Interpretation. On the other hand, in many cases there was no one "correct" way of interpreting something, for example, the classification of a snack as a meal. In cases like this, the main goal was to make sure that all participants understood the chosen interpretation.

It was agreed that reference answers should emphasize literal understanding of an utterance, rather than a cooperative answer which might cause more information to be included than what was actually requested. However, to support systems used for demonstrations and for data collection as well as for evaluation, answers needed to be minimally cooperative, since otherwise demonstration systems would have to answer differently from evaluation systems. Thus the main criterion was how well the proposed interpretation reflected understanding of the query, with some consideration for providing a cooperative answer.

7. Conclusion

The MADCOW experiment in multi-site data collection and evaluation has been successful. The participating sites have collected a rich corpus of training data, have put in place methods for distributing the data, and have devised test procedures to evaluate speech, natural language, and spoken language results on a test corpus. The resources made available by the multi-site paradigm have allowed us to collect more data and to learn more about data collection than would have been possible with only one or two sites collecting data under a special contract.

Some difficult problems still remain. Our shared goal is to build interactive spoken language systems; however, our evaluation methods rely on a canned corpus and evaluate a system's recognition performance under static conditions that are not representative of the interactive environments in which these systems will eventually be used. In addition, the ATIS task has been limited so far to a small, static subset of the air travel domain. These difficulties will increase as research sites develop different approaches to actively managing interaction with the user: different processing strategies will generate divergent behaviors on the part of users, but this divergence will lessen the validity of tests that assume comparable responses to a sequence of queries.

The MADCOW collection and evaluation procedures have provided effective tools for assessing the current capabilities of interactive spoken language systems. However, we must continue to improve our methods of data collection and evaluation. For example, we have only just begun to explore the use of real-time spoken language systems for data collection and evaluation. We also need to more towards a larger, more realistic, database. As our spoken language systems evolve, data collection and evaluation methods must evolve with them.

8. Acknowledgements

The Multi-site ATIS Data Collection Working group includes a representative from each site involved in the ATIS data collection effort: Lynette Hirschman, Chair (MIT), Jay Wilpon (AT&T), Madeleine Bates (BBN), Alexander Rudnicky (CMU), David Pallett (NIST), Deborah Dahl (Paramax), Patti Price (SRI) and Jared Bernstein (SRI-Annotation). Many more people have made key contributions and have been an integral part of the multi-site effort. This includes the staff at NIST (Nancy Dahlgren, William Fisher, Jon Fiscus, Kathy Gallo, John Garofolo, Brett Tjaden) and the Annotation Group at SRI (Kate Hunicke-Smith, Harry Bratt, and Beth Bryson). In addition, Robert Moore and Eric Jackson (SRI) provided critical help in designing and debugging the code to support the minimal/maximal answer criterion. Individuals responsible for the data collection at the various sites include: Christine Pao (AT&T), Robert Bobrow, Robert Ingria, David Stallard, and Varda Shaked (BBN), Robert Weide, Cynthia Neelan and Sondra Ahlen (CMU), Joseph Polifroni, Stephanie Seneff, and Christie Clark Winterton (MIT), and Elizabeth Shriberg, Elizabeth Wade, and Steven Tepper (SRI). Deborah Dahl (Paramax), together with William Fisher (NIST), took on responsibility for maintaining the Principles of Interpretation in conjunction with the SRI Annotators. The Principles of Interpretation Working Group included Evelyne Tzoukermann (AT&T), David Stallard (BBN), Wayne Ward (CMU), Stephanie Seneff (MIT), and Robert Moore (SRI). Michael Phillips and Christie Clark Winterton (MIT) were responsible for creation of the CD-ROMs at MIT. Finally, many individuals at the various sites played an important role in checking and debugging the data and their annotations, including Robert Bobrow, Robert Ingria, and David Stallard (BBN), Joseph Polifroni and Stephanie Seneff (MIT), William Fisher, Nancy Dahlgren, Jon Fiscus, John Garofolo, Brett Tjaden (NIST), Lewis Norton (Paramax), and Eric Jackson (SRI).

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. Hemphill, C. T., J. J. Godfrey, and G. R. Doddington, "The ATIS Spoken Language System Pilot Corpus," *Proc. Third DARPA Speech and Language Workshop*, R. Stern (ed.), Morgan Kaufmann, June 1990.
3. Jackson, E., D. Appelt, J. Bear, R. Moore, A. Podlozny, "A Template Matcher for Robust NL Interpretation," *Proc. DARPA Speech and Natural Language Workshop*, P. Price (ed.), Morgan Kaufmann, 1991.
4. Murveit, H., J. Butzberger, and M. Weintraub,"Speech Recognition in SRI's Resource Management and ATIS Systems," *Proc. DARPA Speech and Natural Language Workshop*, P. Price (ed.), Morgan Kaufmann, 1991.
5. Murveit, H. and M. Weintraub, "Real-Time Speech Recognition System," *Proc. DARPA Speech and Natural Language Workshop*, P. Price (ed.), Morgan Kaufmann, 1991.
6. Murveit, H., J. Butzberger, and M. Weintraub, "Performance of SRI's Decipher Speech Recognition System on DARPA's ATIS Task," *Proc. DARPA Speech and Natural Language Workshop*, M. Marcus (ed.), Morgan Kaufmann, 1992.
7. Pallett, D., "Session 2: DARPA Resource Management and ATIS Benchmark Test Poster Session", *Proc. DARPA Speech and Natural Language Workshop Workshop*, P. Price (ed.), Morgan Kaufmann, 1991.
8. Pallett, D., et al. "February 1992 DARPA ATIS Benchmark Test Results Summary," *Proc. DARPA Speech and Natural Language Workshop*, M. Marcus (ed.), Morgan Kaufmann, 1992.
9. Pao, C. and J. Wilpon, "Spontaneous Speech Collection for the ATIS Domain with an Aural User-Feedback Paradigm," *Proc. DARPA Speech and Natural Language Workshop*, M. Marcus (ed.), Morgan Kaufmann, 1992.
10. Polifroni, J., S. Seneff, V. W. Zue, and L. Hirschman, , "ATIS Data Collection at MIT," DARPA SLS Note 8, Spoken Language Systems Group, MIT Laboratory for Computer Science, Cambridge, MA, November, 1990.
11. Price P., "Evaluation of Spoken Language Systems: The ATIS Domain," *Proc. Third DARPA Speech and Language Workshop*, P. Price (ed.), Morgan Kaufmann, June 1990.
12. Ramshaw, L. A. and S. Boisen, "An SLS Answer Comparator," SLS Note 7, BBN Systems and Technologies Corporation, Cambridge, MA, May 1990.
13. Seneff, S., L. Hirschman, and V. Zue, "Interactive Problem Solving and Dialogue in the ATIS Domain," *Proc. Fourth DARPA Speech and Language Workshop*, P. Price (ed.), Morgan Kaufmann, February 1991.
14. Shriberg, E., E. Wade, and P. Price, "Human-Machine Problem Solving Using Spoken Language Systems (SLS): Factors Affecting Performance and User Satisfaction" *Proc. DARPA Speech and Natural Language Workshop*, M. Marcus (ed.), Morgan Kaufmann, 1992.
15. Ward, W., "Evaluation of the CMU ATIS system" *Proc. Fourth DARPA Speech and Language Workshop*, P. Price (ed.), Morgan Kaufmann, February 1991.
16. Weintraub, M., G. Chen, P. Mankoski, H. Murveit, A. Stolzle, S. Narayanaswamy, R. Yu, B. Richards, M. Srivastava, J. Rabay, R. Broderson, "The SRI/UCB Real-Time Speech Recognition System," *Proc. DARPA Speech and Natural Language Workshop*, M. Marcus (ed.), Morgan Kaufmann, 1992.