EXPANDING THE SCOPE OF THE ATIS TASK: THE ATIS-3 CORPUS

Deborah A. Dahl, Madeleine Bates, Michael Brown, William Fisher, Kate Hunicke-Smith, David Pallett, Christine Pao, Alexander Rudnicky, and Elizabeth Shriberg

Contact: Deborah Dahl
Unisys Corporation
P.O. Box 517
Paoli, PA 19301
email:dahl@vfl.paramax.com

ABSTRACT

The Air Travel Information System (ATIS) domain serves as the common evaluation task for ARPA spoken language system developers.[1] To support this task, the Multi-Site ATIS Data COllection Working group (MADCOW) coordinates data collection activities. This paper describes recent MADCOW activities. In particular, this paper describes the migration of the ATIS task to a richer relational database and development corpus (ATIS-3) and describes the ATIS-3 corpus. The expanded database, which includes information on 46 US and Canadian cities and 23,457 flights, was released in the fall of 1992, and data collection for the ATIS-3 corpus began shortly thereafter. The ATIS-3 corpus now consists of a total of 8297 released training utterances and 3211 utterances reserved for testing, collected at BBN, CMU, MIT, NIST and SRI. 2906 of the training utterances have been annotated with the correct information from the database. This paper describes the ATIS-3 corpus in detail, including breakdowns of data by type (e.g. context-independent, context-dependent, and unevaluable)and variations in the data collected at different sites. This paper also includes a description of the ATIS-3 database. Finally, we discuss future data collection and evaluation plans.

1. BACKGROUND

The ATIS task was first used as a common ARPA spoken language evaluation task in 1990 [1,2]. In the ATIS task, subjects obtain air travel information such as flight schedules, fares, and ground transportation from a relational database using spoken natural language, and use it to solve air travel planning scenarios. Although the core air travel planning task has remained the same since the beginning, its use in evaluation has gradually evolved over the years with the general objectives of increasing the match between the evaluation and a real task as well as increasing the accuracy of the metric.

The first official evaluation took place in February of 1991, following a dry run in June of 1990. In the 1991 evaluation, context-independent (Class A) queries as well as dialog pairs (D1) were evaluated. The score for a system was the weighted-error metric which included a penalty for incorrect answers as opposed to “No Answer”. Further refinements took place in the November 1992 evaluation, where Class D (utterances with context dependencies throughout the dialog) queries were evaluated. Another variation introduced in 1992 was the min-max criterion, in which the information provided by systems in the answer was required to fall between a minimum and a maximum amount. In the most recent evaluation, December 1993, the main change has been to drop the weighted error metric and report results based on the unweighted error, or 100-%T.

The 1993 ATIS spoken language understanding evaluation is the first evaluation based on the ATIS-3 corpus ([5]). The ATIS-3 corpus will also supply test data for the December 1994 ATIS evaluation. In addition, test data has also been reserved for a dry run of a semantic evaluation [6].

2. THE EXPANDED ATIS RELATIONAL DATABASE

The initial ATIS task was based on a relational database containing air travel information for 11 cities. Three corpora of spontaneous spoken language utterances (ATIS-0, ATIS-1 and ATIS-2) were collected with this database using a variety of paradigms, as described in [3,4]. As ATIS technology developed, it was felt that the initial ATIS task was unrealistically limited because of the small size of the database. Consequently, the database was expanded to include air travel information for 46 cities. The expanded database was released in the fall of 1992, and data collection began shortly thereafter.

The new database is based on air travel data obtained from the Official Airline Guide (OAG) in June 1992 and current at that time. The database includes information for 46 cities

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

and 52 airports in the US and Canada. The largest table in the expanded database, the flight table, includes information on 23,457 flights. This compares to 11 cities, 9 airports, and 765 flights in the earlier ATIS databases and clearly represents a significant scaling up of the ATIS task. Despite the fact that the number of flights in the database has been increased by over a factor of thirty, the conversion to the larger database has not caused any serious difficulties for the sites doing data collection, the sites doing evaluation, or the annotators.This result is encouraging, since it indicates that the SLS technology developed on a small database can scale up to a significantly bigger task.

Cities and airports included in the new database are listed in the Appendix.

3. ATIS-3 DATA COLLECTION AND ANNOTATION

The ATIS-3 data was collected at BBN, CMU, MIT, NIST, and SRI. NIST participated in ATIS data collection for the first time in this round of data collection, using data collection software from both BBN and SRI.

Since the beginning of the ATIS task data collection paradigms have moved toward increasingly automated approaches. The original ATIS-0 corpus was collected using human wizards to both transcribe the subjects' utterances as well as to interpret them (the so-called "wizard of OZ" paradigm). In the ATIS-3 corpus, nearly all transcription and interpretation of the subjects' speech was done by the sites' ATIS systems themselves. The only exception was MIT, which collected data using a transcription wizard instead of a speech recognizer, while using MIT's natural language system to interpret the utterances. Automatic data collection has the advantage of reduced cost. In addition, the data is more realistic in the sense that it is obtained from subjects who are really talking to a computer. The disadvantage of automatic data collection is that imperfect processing by the spoken language system sometimes leads to the presence of artifacts in data collection, such as utterances repeated over and over again.

The general process of data collection and annotation as described in [3] has not changed in the ATIS-3 data collection effort. We summarize this process here for convenience.

Collected data is transcribed at the collecting site and sent to NIST, where it is logged and potential test data is held out. The data is then released to sites participating in the ATIS evaluations as initial, i.e. unannotated, data, and is simultaneously sent to SRI for annotation. During annotation, the data is classified into three categories:

- Class A: not dependent on context for interpretation
- Class D: dependent on context for interpretation
- Class X: unevaluable

The Principles of Interpretation document is used to categorize utterances in these three classes and also specifies how to interpret vague expressions which occur in utterances in Class A and D.

Annotated data is returned from SRI to NIST and released by NIST. A full set of data for a subject session includes the following files:

- .wav: speech waveform
- .log: session log
- .sro: detailed transcription
- .cat: categorization of query (A, D, X)

Table 1: Total ATIS-3 Data

	Training Pool (including 1993 Test Data)			Test			SemEval Dry Run			Total		
Site	#Spkr	# Sess	#Utts	#Spkr	#Sess	#Utts	#Spkr	#Sess	#Utts	#Spkr	#Sess	#Utts
BBN	14	55	1101	9	37	389				23	92	1490
CMU	15	177	1462	8	70	387				23	247	1849
MIT	30	146	954	25	120	418				55	266	1372
NIST	49	253	2510	22	179	201	12	67	500	71	432	3211
SRI	30	141	2326	6	27	418				36	168	2744
Total	125	693	8297			1813	12	67	500			10666

- .ref: minimal reference answer
- .rf2: maximal reference answer

3.1. The ATIS-3 Corpus

3.2. Initial Data

The total data collected for the ATIS-3 corpus consists of 12,047 utterances, of which 3,876 have been annotated. The data breaks down across sites as shown in Table 1. Approximately 20% of submitted data is allocated to the pool from which evaluation test data is drawn. In addition, 500 utterances from the NIST data collection activity have been reserved as test data for a possible dry run for semantic evaluation in ATIS [6]. This table does not include 1440 additional utterances collected at CMU which have not yet been released as initial data. Two subject-scnearios from the ATIS-3 corpus can be seen in Appendix 2. Note in particular the false starts typical of spontaneous speech.

3.3. Annotated Data

Slightly over 1/3, or 36%, of the released initial data has been annotated with the correct answers. Unannotated data includes data reserved for the December 1994 evaluation, which will be annotated just before the evaluation to insure that it is consistent with the Principles of Interpretation current at that time. Other unannotated data includes data from NIST and SRI which was received too late to be annotated.

The full corpus of annotated data also includes 667 sentences collected for the November 1992 logfile experiments [4]. Although these utterances were collected using the 11 city database, they were annotated using the expanded database. The rationale for this decision was that the annotators wished to get experience with the expanded database, and at the time, the logfile data was the only data available.

The annotated data breaks down into Classes A, D, and X by site as shown in Table 2.

If the annotated data is broken down by site as well as by class, it can be noted that there is a wide range of variation across sites in the relative proportion of A, D, and X queries, which can be seen in Figure 1. We believe this is largely attributable to the effects of different data collection scenarios used by the different sites. The practical consequences of this effect are that an understanding of how scenarios achieve this effect might lead to the development of techniques for improving system performance for particular applications.

4. ATIS PLANS

4.1. Development Test Data

Despite the fact that nearly 25,000 ATIS utterances have been collected since 1990, no standard development test data exists for ATIS. Sites have individually constructed development test sets from the training data and evaluation test sets, but this method makes inter-site comparisons difficult. While inter-site comparisons are a major goal of the official evaluation, variations in the test set from year to year make comparisons across years problematic. In addition, if evaluation test data is used after the evaluation as development test data, it is contaminated by system developers looking at it in detail for the purposes of adjudication. The existence of a development test corpus will also extend the usefulness of the ATIS-3 corpus after all training and evaluation test data is released by providing a source of unseen data. For these reasons MADCOW has decided to collect an additional 300-500 utterances from BBN, CMU, MIT, NIST, and SRI, to be designated development test data. This data is to be collected in the spring of 1994 and will have a high priority for early annotation.

4.2. Alternative Evaluations

MADCOW is also interested in exploring evaluation paradigms other than the standard CAS evaluation. These include the end-to-end/logfile approach described in [4], as well as the semantic evaluation paradigm described in [6].

Table 1: Number of A, D, and X utterances in ATIS-3 data

Training Data					December 93 Evaluation Test Data					Total
Site	Class A	Class D	Class X	Total	Site	Class A	Class D	Class X	Total	
BBN	282	182	193	657	BBN	89	57	53	199	856
CMU	417	197	103	717	CMU	113	50	36	199	916
MIT	391	239	121	751	MIT	82	50	32	164	915
SRI	329	351	106	786	SRI	80	86	34	200	986
NIST	0	0	0	0	NIST	84	82	82	203	203
Total	1419	969	523	2911		448	325	192	965	3876

End to End: In 1992 MADCOW defined and carried out a dry run evaluation of approaches in which a human judge rules on the correctness or appropriateness of each system response and, in which task-level metrics, such as time-to-complete task and correctness of solution are measured [4]. On the basis of an analysis of the experiment discussed in [4] performed by Alex Rudnicky, we have determined that in order to obtain statistically reliable results it will be necessary to reduce extraneous sources of variation as much as possible; consequently, a within-subjects design is highly desirable.[1] Although we have not continued to actively develop this approach, we believe that it may be useful in the future as we move to increasingly realistic tasks.

Semantic Evaluation: The goal of semantic evaluation is to define a level of representation which focuses specifically on language understanding, as opposed to task performance, in a maximally task-independent way. This approach has the advantage of minimizing the number of extraneous tasks required of system developers participating in evaluations. In addition, it is anticipated that much of the work done in developing the semantic evaluation will carry over to new tasks.

Aside from the specific representation used, which is discussed in detail for ATIS in [6], the infrastructure for carrying out a semantic evaluation is remarkably parallel to that required by the current CAS evaluations. That is, data needs to be collected, annotated according to a set of well-defined rules, and distributed to sites. In addition ancillary software is required for scoring and to assist in annotation.

[1] We would like to acknowledge David Pisoni of Indiana University and Astrid Schmidt-Nielsen of the Naval Research Lab for their helpful comments on the end-to-end evaluation procedure.

4.3. Beyond ATIS-3

MADCOW has also begun to explore follow-on tasks to ATIS to be implemented for the 1995 evaluation cycle. Although the details of future tasks remain to be specified, telephone tasks are of high interest, since they stimulate both research on telephone speech as well as interactive dialog. In addition, telephone tasks are useful because the subjects do not have to be physically located near the data collecting system, thus making it possible for subjects in different geographic areas to interact with a range of data collection systems simply by using another telephone number.

ACKNOWLEDGEMENTS

The MADCOW committee would like to acknowledge the contributions of the following people to the shared data collection efforts. At AT&T, Enrico Bocchieri and Bruce Buntschuh, At BBN, Beverly Schwartz, Sandra Peters, and Robert Ingria, at CMU, Robert Weide, Yuzong Chang, and Eric Thayer, at MIT, Lynette Hirschman and Joe Polifroni, at NIST, John Garofolo, Jon Fiscus, and Bruce Lund, at SRI Goh Kawai and Tom Kuhn (annotators) and at Unisys, Lew Norton.

REFERENCES

1. Price, P. "Evaluation of spoken language systems: the ATIS domain". In Proceedings of the speech and natural language workshop. Morgan-Kaufmann, 1990.

2. Hemphill, C. T., J. J. Godfrey, and G. R. Doddington. "The ATIS spoken language systems pilot corpus". In Proceedings of the speech and natural language workshop. Morgan Kaufmann, 1990.

Figure 1: Class A, D, and X in Training Data

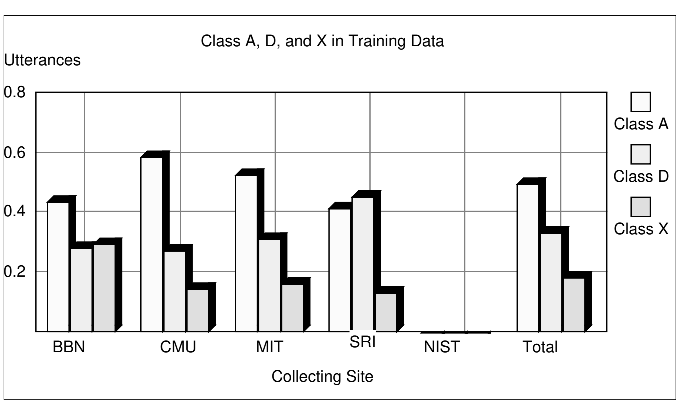
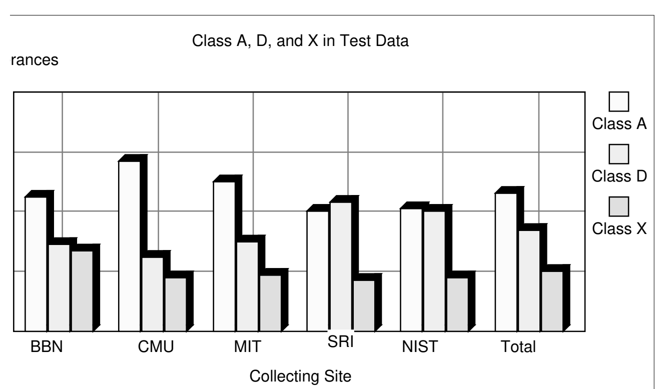


Figure 2: Class A, D, and X in Test Data



3. MADCOW. "Multi-site data collection for a spoken language corpus" Proceedings of the fifth DARPA speech and natural language workshop. Morgan Kaufmann, 1992.

4. Hirschman, L. M. Bates, D. Dahl, W. Fisher, D. Pallett, Kate Hunicke-Smith. P. Price, A. Rudnicky, and E. Tzoukermann."Multi-site data collection and evaluation in spoken language understanding". Proceedings of the Human Language Technology Workshop, March, 1993.

5. Pallett, David, Jonathan Fiscus, William Fisher, John Garofolo, Bruce Lund, Mark Pryzbocki, "1993 Benchmark Tests for the ARPA Spoken Language Program" (this volume).

6. Moore, R. "Semantic Evaluation for Spoken Language Systems" (this volume).

APPENDIX 1

Cities and airports included in the expanded ATIS relational database:

Table 1: Cities

Nashville, TN	Boston, MA	Burbank, CA
Baltimore, MD	Chicago, IL	Cleveland, OH
Charlotte, NC	Columbus, OH	Cincinnati, OH
Denver, CO	Dallas, TX	Detroit, MI
Fort Worth, TX	Houston, TX	Westchester County, NY
Indianapolis, IN	Newark, NJ	Las Vegas, NV
Los Angeles, CA	Long Beach, CA	Atlanta, GA
Memphis, TN	Miami, FL	Kansas City, MO
Milwaukee, WI	Minneapolis, MN	New York, NY
Oakland, CA	Ontario, CA	Orlando, FL
Philadelphia, PA	Phoenix, AZ	Pittsburgh, PA
St. Paul, MN	San Diego, CA	Seattle, WA
San Francisco, CA	San Jose, CA	Salt Lake City, UT
St. Louis, MO	St. Petersburg, FL	Tacoma, WA
Tampa, FL	Washington, DC	Montreal, PQ
Toronto, ON		

Table 2: Airports

William B. Hartsfield Atlanta Intl., Atlanta, Georgia
Nashville International, Nashville, Tennessee
Logan International, Boston, Massachusetts
Burbank, Burbank, California
Baltimore/Washington International, Baltimore, Maryland
Hopkins International, Cleveland, Ohio
Charlotte/Douglas International, Charlotte, North Carolina
Port Columbus International, Columbus, Ohio
Cincinnati/Northern Kentucky Intl., Cincinnati, Ohio
Love Field, Dallas/Ft. Worth, Texas
Washington National, Washington, D.C.
Stapleton International, Denver, Colorado
Detroit City, Detroit, Michigan
Dallas/Fort Worth International, Dallas/Ft. Worth, Texas
Metropolitan Wayne County, Detroit, Michigan
Newark International, Newark, New Jersey
Hobby, Houston, Texas
Westchester County, Westchester County, New York
Dulles International, Washington, D.C.
Houston Intercontinental, Houston, Texas
Indianapolis International, Indianapolis, Indiana
John F. Kennedy International, New York
Mccarran International, Las Vegas, Nevada
Los Angeles International, Los Angeles, California
La Guardia, New York NY
Long Beach Municipal, Long Beach, California
Kansas City International, Kansas City, Missouri
Orlando International, Orlando, Florida
Midway, Chicago, Illinois
Memphis International, Memphis, Tennessee
Miami International, Miami, Florida
General Mitchell International, Milwaukee, Wisconsin
Minneapolis/St. Paul International, Minneapolis/St. Paul, Mn
Metropolitan Oakland International, Oakland, California

Table 2: Airports

Ontario International, Ontario, California
O'Hare International, Chicago, Illinois
Philadelphia International, Philadelphia PA
Sky Harbor International, Phoenix, Arizona
St. Petersburg/Clearwater International, Tampa/St. Petersburg, Florida
Greater Pittsburgh International, Pittsburgh, Pennsylvania
Lindbergh Field/San Diego International, San Diego, California
Seattle/Tacoma International, Seattle/Tacoma, Washington
San Francisco International, San Francisco, California
San Jose International, San Jose, California
Salt Lake City International, Salt Lake City, Utah
Lambert St. Louis International, St. Louis, Missouri
Tampa International, Tampa/St. Petersburg, Florida
Buttonville, Toronto, Ontario
Mirabel, Montreal, Quebec
Toronto Island, Toronto, Ontario
Dorval International, Montreal, Quebec
Lester B. Pearson International, Toronto

APPENDIX 2

4.4. Sample Subject-Scenarios from the ATIS-3 Corpus (Data collected at NIST using SRI data collection system)

Example 1:

Scenario

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 Milwaukee and plan the flight and ground transportation itinerary to City-A and City-B, and back toMilwaukee.

x07016sx: i would like a morning flight from milwaukee to den- to denver colorado please with ground transportation

x07026sx: i would like a morning flight from milwaukee to denver colorado please

x07036sx: what type of ground transportation from the airport to denver

x07036sx: i would like an afternoon flight from denver colorado to dallas texas

x07046sx: what type of ground transportation from the airport to dallas

x07056sx: i want a evening flight from dallas to milwaukee

x07066sx: what type of ground transportation from the airport to milwaukee

Example 2:

Scenario:

Determine the type of aircraft used on a flight from Cleveland to Dallas that leaves before noon.

x02011sx: may i see all the flights from cleveland to , dallas

x02021sx.sro: can you show me the flights that leave before noon , only

x02031sx.sro: could you sh- please show me the types of aircraft used on these flights