The ATIS Spoken Language Systems Pilot Corpus

Charles T. Hemphill, John J. Godfrey, George R. Doddington

Texas Instruments Incorporated
PO Box 655474, MS 238
Dallas, Texas 75265

Abstract

Speech research has made tremendous progress in the past using the following paradigm:

- define the research problem,
- collect a corpus to objectively measure progress, and
- solve the research problem.

Natural language research, on the other hand, has typically progressed without the benefit of any corpus of data with which to test research hypotheses. We describe the Air Travel Information System (ATIS) pilot corpus, a corpus designed to measure progress in Spoken Language Systems that include both a speech and natural language component. This pilot marks the first full-scale attempt to collect such a corpus and provides guidelines for future efforts.

Introduction

The ATIS corpus provides an opportunity to develop and evaluate speech systems that understand spontaneous speech. This corpus differs from its predecessor, the Resource Management corpus (Price *et al*, 1988), in at least four significant ways.

1. Instead of being read, the speech has many of the characteristics of spontaneous spoken language (*e.g.*, dysfluencies, false starts, and colloquial pronunciations).
2. The speech collection occurs in an office environment rather than a sound booth.
3. The grammar becomes part of the system under evaluation rather than a given part of the experiment.
4. The reference answer consists of the actual reply for the utterance rather than an orthographic transcription of the speech.

The evaluation methodology supported by ATIS depends on having a comparable representation of the answer for each utterance. This is accomplished by limiting the utterances to database queries, and the answers to a ground set of tuples from a fixed relational database. The ATIS corpus comprises the acoustic speech data for a query, transcriptions of that query, a set of tuples that constitute the answer, and the SQL expression for the query that produced the answer tuples.

The ATIS database consists of data obtained from the Official Airline Guide (OAG, 1990), organized under a relational schema. The database remained fixed throughout the pilot phase. It contains information about flights, fares, airlines, cities, airports, and ground services, and includes twenty-five supporting tables. The large majority of the questions posed by subjects can be answered from the database with a single relational query.

To collect the kind of English expected in a real working system, we simulate one. The subject, or "travel planner," is in one room, with those running the simulation in another. The subject speaks requests over a microphone and receives both a transcription of the speech and the answer on a computer screen. A session lasts approximately one hour, including detailed preliminary instructions and an exit questionnaire.

Two "wizards" carry out the simulation: one transcribes the query while the other produces the answer. The transcriber interprets any verbal editing by the subject and removes dysfluencies in order to produce an orthographic transcription of what the subject intended to say. At the same time, the answerer uses a natural language-oriented command language to produce an SQL expression that elicits the correct answer for the subject. On-line utilities maintain a complete log of the session, including time stamps.

At the conclusion of the session, the utterances are sorted into categories to determine those utterances suitable for objective evaluation. Finally, each utterance receives three different transcriptions. First, a checked version of the transcription produced during the session provides an appropriate input string for evaluating text-based natural language systems. Second, a slightly expanded version of this serves as a prompt in collecting a read version of the spontaneously spoken sentences. Finally, a more detailed orthographic transcription represents the speech actually uttered by the subject, appropriate for use in acoustic modeling.

Corpus Collection

About one session a day was conducted, using subjects recruited from within Texas Instruments. A typical session included approximately 20 minutes of introduction, 40 minutes of query time and 10 minutes for follow-up. Each session resulted in two speech files for each query and a complete log of the session. Figure 1 depicts the session procedure.

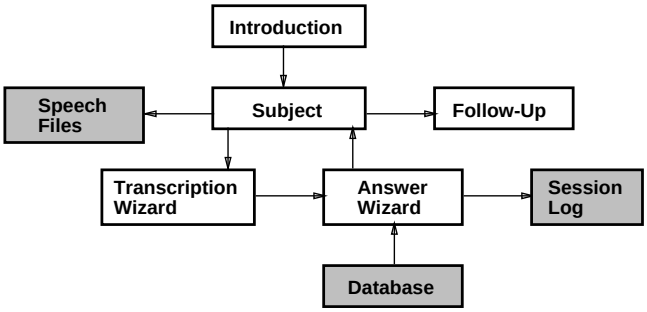

Figure 1: Subject Session Procedure

Session Introduction

The subjects were given the following instructions, both orally and in writing:

> The Air Travel Information System (ATIS) is a prototype of a voice-input information retrieval system. It has the same information that is contained in the Official Airline Guide (OAG) to help you make air travel plans. We would like you to participate in a trial use of this experimental system.

Subjects were not told whether that the "experimental system" was totally automated or involved human intervention. It was hoped that most subjects would believe that the system was real to elicit natural speech.

Subjects were informed about the contents of the relational database in a one page summary. The summary described the major database entities in fairly general terms to avoid influencing the vocabulary used during the session. To avoid some misconceptions in advance, subjects were told that the database did not contain information about hotels or rental cars.

The subject was next assigned a travel planning scenario, systematically chosen from a set of six scenarios designed to exercise various aspects of the database. For example, some scenarios focused on flight time constraints while others concentrated on fares. The scenarios did not specify particular times or cities in an effort to make the scenario more personal to the subject. The following example illustrates this:

> Plan the travel arrangements for a small family reunion. First pick a city where the get-together will be held. From 3 different cities (of your choice), find travel arrangements that are suitable for the family members who typify the "economy", "high class", and "adventurous" life styles.

After receiving the scenario, subjects were left with the instructions and given five minutes to plan the details of the scenarios. Subjects were given pen and paper on which to write the details and to take notes during the session.

Finally, subjects were given instructions regarding the operation of the system. The "system", from the subjects perspective, consisted of a 19 inch color monitor running the X Window System, and a head-mounted Sennheiser (HMD 410-6) microphone. A desk mounted Crown (PCC-160 phase coherent cardioid) microphone was also used to record the speech. The "office" contained a sparc-station cpu and disk to replicate office noise, and a wall map of the United States to help subjects solve their scenarios.

The monitor screen was divided into two regions: a large, scrollable window for system output and a smaller window for speech interaction. The system used a "push-to-talk" input mechanism, whereby speech collection occurred while a suitably marked mouse button was depressed. Subjects were given the opportunity to cancel an utterance for a period of time equal to the length of the utterance.

A single sentence was used for all subjects to illustrate the push-to-talk mechanism and interaction with the system:

> Show me all the nonstop flights between Atlanta and Philadelphia.

This sentence was processed as if the system actually responded to the utterance, including a transcription of the speech on the subject's display followed by the answer in table format.

Session Queries

After the introduction, subjects were given approximately 40 minutes to complete the task described in the scenario. If they finished early, subjects were instructed to select another scenario or to explore the capabilities of the system. After the 40 minutes, subjects were given the opportunity to continue, finally ending the session by saying "all done".

Once the actual session started, subjects cycled through thinking, querying, waiting, and writing. While the thinking portion of the session actually required the most time, the query portion required the most resources.

Several things happened at once as a given subject spoke a query. While speech from both the head-mounted and desk-mounted microphones was recorded, one wizard began to transcribe the speech and the other wizard began to answer the query. A playback capability could be used if needed by the transcription wizard. The answer wizard was constrained not to send the answer before the transcription wizard finished the transcription. Typically, the subject received the typed

transcription a few seconds after speaking and the answer approximately 20 seconds later.

Each wizard each had their own X Window terminal. The transcription wizard used a gnuemacs-based tool that checked the spelling of the transcription and sent the transcription to both the answer wizard and the subject. Despite the transcription wizard's best efforts, some transcription mistakes did reach the subject: occasionally words were omitted, inserted, or substituted (e.g., "fight" for "flight").

The answer wizard used a tool called NLParse (Hemphill *et al*, 1987) to form the answer to the subjects queries. This tool used a natural language-oriented command language to produce a set of tuples for the answer. NLParse provides a set of menus to help convey the limited coverage to the wizard. In practice, the answer wizard knew the coverage and used typing with escape completion to enter the appropriate NLParse command. NLParse provides several advantages as a wizard tool:

- every answerable query (with respect to the database) receives an answer,
- the NLParse query language avoids ambiguity,
- the wizard formulates the answer in terms of database entities, and
- the wizard can easily discern the correctness of the answer.

However, the NLParse query language was not originally designed for rapid query entry, prompting several small grammar enhancements during the pilot.

The answer wizard's terminal also included a gnuemacs-based utility that created a session log. This included the transcription, the NLParse input, the resulting SQL expression, and the set of tuples constituting the answer. The answer wizard sent only the set of tuples to the subject.

The ATIS Database

The ATIS database was designed to model as much of a real-world resource as possible. In particular, we tried to model the printed OAG in a straightforward manner. With this approach, we could rely on travel data expertise from Official Airline Guides, Incorporated. We also used the data directly from the OAG and did not invent any data — something that is difficult to accomplish in a realistic manner. Additionally, the printed OAG was available to all sites and provided a form of documentation for the database.

The relational schema were designed to help answer queries in an intuitive manner, with no attempt to maximize the speech collected (*e.g.*, by supplying narrow tables as answers). Toward this end, entities were represented with simple sets or lists in the most direct way.

Session Follow-Up

After the query phase of the session, subjects were given a brief questionnaire to let us know what they thought of the system. This consisted of the following ten questions with possible answers of "yes" "maybe/sometimes", "no" or "no opinion":

1. Were you able to get the travel information you needed?
2. Were you satisfied with the way the information was presented?
3. Did the responses contain the kinds of information you were seeking?
4. Were the answers provided quickly enough?
5. Would you prefer this method to looking up the information in a book?
6. Did the system understand your requests the first time?
7. If the system did not understand you, could you easily find another way to get the information on a later try?
8. Was the travel planning scenario appropriate for a trial use of the system?
9. Do you think a person unfamiliar with computers could use the system easily?
10. Do you think a human was interpreting your questions?

After the questionnaire, the subjects were given a chance to ask questions, and were informed that the system was a simulation involving human intervention. Finally, we thanked our subjects with their choice of either a mug or a T-shirt.

Corpus Processing

After data collection, a rather elaborate series of processing steps was required before the subject's utterances actually became part of the corpus. A session resulted in a set of speech files and a session log that formed the raw materials for the corpus. Figure 2 illustrates the processing steps.

Transcriptions

To facilitate use of the corpus, three transcriptions were provided with each query. A more detailed transcription document specifies the details of these, with the rationale explained below.

- **NL_input**: This transcription is a corrected version of the on-the-fly session transcription, corrected while reviewing the subject's speech off-line. This transcription reflects the speech as the subject meant to say it, that is, dysfluencies corrected

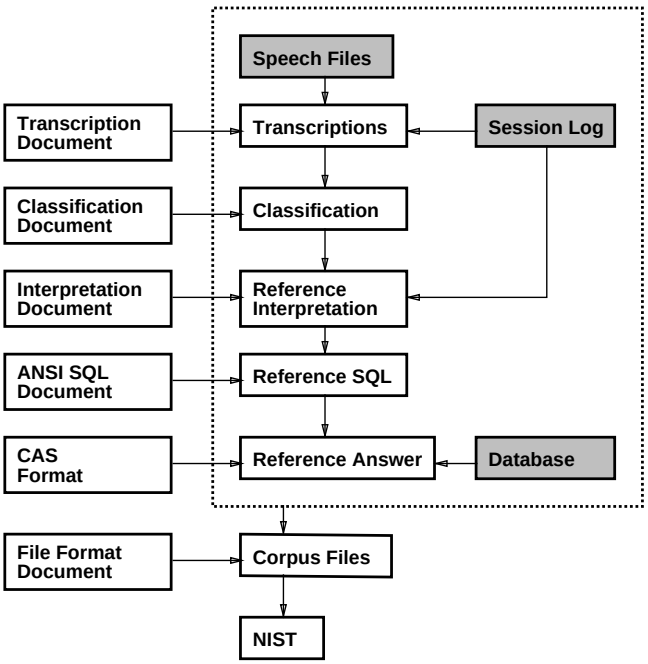

Figure 2: Corpus Processing Steps

verbally by the subject were corrected. The orthography of the transcription obeys common English conventions. It is suitable as input to a text-based natural language systems. Example:

> Where is the stop for USAir flight number 37 from Philadelphia to San Francisco?

- **prompting_text**: This transcription expands any acoustically ambiguous lexical tokens found in the **NL_input** transcription while listening to the subject's speech. This transcription serves as a prompt in a later read-speech session, allowing a comparisons of read and spontaneous speech. Example:

  > Where is the stop for USAir flight number thirty-seven from Philadelphia to San Francisco?

- **SR_output**: This transcription includes a detailed description of the major acoustic events in the query. It is created from the **prompting_text** while listening to the subject's speech and includes all the dysfluencies previously ignored. Abbreviations and numbers are expanded to eliminate open-class lexical items. This transcription serves as a point of comparison for speech recognition systems that output an orthographic transcription. Example:

  > Where is the stop [uh] for U S, Air flight number thirty seven, from Philadelphia to San Francisco

For interim testing purposes, a Standard, Normalized Orthographic Representation (SNOR) was created algorithmically from the **SR_output** transcription. Punctuation and dysfluencies were removed, resulting in something resembling the **NL_input** transcription, but with abbreviations and numbers expanded. Example:

> WHERE IS THE STOP FOR U S AIR FLIGHT NUMBER THIRTY SEVEN FROM PHILADELPHIA TO SAN FRANCISCO

Classification

Not all queries were equally suited for evaluating spoken language systems. Accordingly, each query received a classification to help circumscribe the types of queries desired for training and testing. The classifications themselves were determined through a committee and defined several dimensions:

- context-dependent/context-removable/context-independent
- ambiguous (vague)/clear
- unanswerable/answerable
- ill-formed (grossly)/well-formed
- noncooperative/cooperative

The committee defined evaluable queries (for June, 1990) as those not classified by the first term in each set. In addition to these, the following simple classifications were supplied to help sites analyze results:

- ungrammatical/grammatical
- multi-sentence/single-sentence

Reference Interpretation

An interpretation document was defined, which specifies the details of how to interpret an ATIS query, both for the answer wizard and for the SLS sites. For example, for consistency it was ruled that a flight serving a snack would be considered as a flight with a meal. The document provides a mapping of concepts expressed in English to concepts encoded in the relational database. The NLParse commands reflect these conventions and were included in the corpus to facilitate maintenance since it was usually easier to determine the correctness of the reference answer by looking at the NLParse command rather than the resulting SQL expression. In the event of an erroneous answer, correction occurs by simply amending the NLParse command.

Reference SQL

The pilot corpus includes the ANSI-standard SQL expression that produced the reference answer, which is the "final word" on the interpretation of a subject's query. It also provides some degree of database independence. For example, as long as the relational schema remain fixed, we can add new cities to the database, rerun the SQL against the database, and produce a new corpus that includes the new cities. This works as long as the evaluation criteria excludes context-dependent queries.

Reference Answer

The reference answer consists of the set of tuples resulting from the evaluation of the reference SQL with respect to the official ATIS database. This is actually redundant, but makes scoring easier for most sites. The tuples are formatted according the Common Answer Specification (CAS) format (Boisen *et al*, 1989). This format amounts to representing the answer in Lisp syntax to aid in automatic scoring.

Corpus Files

All of the items mentioned above were formatted into files and shipped to the National Institute of Standards and Technology (NIST). NIST then distributed the corpus to interested sites. A file format document exists to help sites install the data.

Results

Forty-one sessions containing 1041 utterances were collected over 8 weeks, nine of which were designated as training material by NIST. Each session consisted of 25.4 queries per session on average. Table 1 describes the utterance statistics for each Pilot Distribution (PD).

PD	Weeks	Sessions	Utt	Utt/Sess
1	2	9	234	26.0
2	2	10	245	24.5
3	2	10	236	23.6
4	1	7	197	28.1
5	1	5	129	25.8
total	8	41	1041	25.4

Table 1: Session Utterance Statistics

Table 2 describes the time statistics for each PD. Each session consisted of approximately 40 minutes of query time with an average rate of 39.1 queries per hour. The average time between queries of 1.5 minutes included subject thinking time, and about 22 seconds for the wizard to send the answer to the subject after the transcription.

PD	Min	Ave	Min/Utt	Sec/Ans	Utt/Hr
1	355	39.4	1.5	23.5	39.6
2	354	35.4	1.4	21.2	41.5
3	391	39.1	1.7	24.2	36.2
4	302	43.1	1.5	19.6	39.1
5	196	39.1	1.5	21.6	39.5
total	1598	39.0	1.5	22.1	39.1

Table 2: Session Time Statistics

The average utterance length (in words) varied according to the transcription: 10.2 for **NL_input**, 11.7 for **SR_output** (expanded lexical items and dysfluencies), and 11.3 for **NL_SNOR** (expanded lexical items). Eighteen percent of the utterances contained some form of dysfluency.

Of the 1041 utterances collected, 740 were judged evaluable according to the June 1990 criteria: not classified as context-dependent, ambiguous, ill-formed, unanswerable, or noncooperative. These results are shown in Table 3, broken down according to PD. The table also shows that if we relax these criteria to exclude only ambiguous and unanswerable utterances, the yield would increase from 71% to 80%.

PD	Utt	J-unevl	%J-evl	relax	%evl
1	234	88	62	73	68
2	245	73	70	52	79
3	236	47	80	32	86
4	197	58	70	27	86
5	129	35	73	19	85
total	1041	301	71	203	80

Table 3: Session Yield of Evaluable Utterances

Subjects generally enjoyed the sessions, as reflected in Table 4 (the tally includes two subjects not included in the corpus). The answers to questions were typically not provided quickly enough, as might be expected in a simulation. Some subjects defined an acceptable response time as under 5 seconds. Of the subjects that thought a human was interpreting the questions, some knew in advance, some misinterpreted the question ("Did the system *behave* as if a human was interpreting your questions?"), and some were tipped-off by the amazing ability of the system to recognize speech in the face of gross dysfluencies.

Q	Yes	Maybe/Sometimes	No	No Opinion
1	27	16	0	0
2	32	10	1	0
3	31	9	2	0
4	2	19	22	0
5	29	10	4	0
6	26	15	1	0
7	24	4	4	7
8	40	1	1	1
9	26	13	3	1
10	8	7	22	5

Table 4: Answers to the Questionnaire

Subjects also supplied general comments. Some subjects felt uncomfortable with computers or the system:

> "The topic was not my thing, but the voice activation was fascinating."

while other subjects were more enthusiastic:

"The system looks like a winner. It needs some fine-tuning and perhaps faster response, but otherwise it's a very promising tool."

Conclusions

The ATIS SLS pilot corpus has proved that objective evaluation of spoken language systems is both possible and beneficial. The pilot corpus has also served to clarify many points in the data collection procedure. In this effort, we have learned that a spontaneous speech corpus is more expensive to collect than a read speech one, but provides an opportunity to evaluate spoken language systems under realistic conditions.

Acknowledgments

This work was supported by the Defense Advanced Research Projects Agency and monitored by the Naval Space and Warfare Systems Command under Contract No. N00039-85-C-0338. The views and conclusions expressed here do not represent the official policies, either expressed or implied, of the Defense Advanced Research Projects Agency or the United States Government.

We gratefully acknowledge the publishers of the Official Airline Guide for travel data and consulting help. We thank the subjects for their participation, Jane McDaniel for her invaluable assistance in all phases of the corpus collection, and the many members of the various committees for their expert advice.

References

[1] Boisen, Sean, Lance A. Ramshaw, Damaris Ayuso, and Madeleine Bates, "A Proposal for SLS evaluation," in *Proceedings of the DARPA Speech and Natural Language Workshop*, October 1989.

[2] Hemphill, Charles T., Inderjeet Mani, and Steven L. Bossie, "A Combined Free-Form and Menu-Mode Natural Language Interface", *Abridged Proceedings of the Second International Conference on Human-Computer Interaction*, Honolulu, Hawaii, 1987.

[3] Official Airline Guides, *Official Airline Guide, North American Edition with Fares*, Oakbrook, Illinois, Volume 16, No. 7, January 1, 1990.

[4] Price, P.J., W.M. Fisher, J. Bernstein, D.S. Pallett, "The DARPA 1000-Word Resource Management Database for Continuous Speech Recognition", *Proceedings of ICASSP*, 1988.