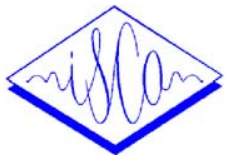

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

DATA COLLECTION AND PROCESSING IN THE CARNEGIE MELLON COMMUNICATOR

Maxine Eskenazi, Alexander Rudnicky, Karin Gregory, Paul Constantinides, Robert Brennan, Christina Bennett, and Jwan Allen
Carnegie Mellon University
5000 Forbes Ave., Pittsburgh, Pa. 15213 USA
max/air/pcc/cbennett/jwanr@cs.cmu.edu, gregory+@andrew.cmu.edu
http://www.speech.cs.cmu.edu

ABSTRACT

In order to create a useful, gracefully functioning system for travel arrangements, we have first observed the task as it is accomplished by a human. We then imitated the human while making the user believe he was dialoguing with an automatic system. As we gradually built our system, we devised ways to assess progress and to detect errors. The following described the manner in which the Carnegie Mellon Communicator was built, data collected, and assessment begun using these criteria.

Keywords: Corpora, Wizard of Oz, Travel Planning, Communicator

1. INTRODUCTION

We have collected over 3160 dialogs at present for the Carnegie Mellon Communicator. This system carries out a telephone dialog with the user in order to make plane, car, and hotel reservations for business travellers.

The automatic system was built gradually in a series of steps, proceeding from human-human speech to human-machine dialogs.

After examining 48 human-human dialogs collected with a travel agent, a Wizard-of-Oz (WOZ) system was created where the human operator typed responses to a user-spoken input. 107 dialogs were collected in this manner.

Thereafter a prototype version of the automatic system became available and was used to capture human-machine dialogs. The first prototype automatic system makes simple flight reservations of one or more legs and simple car and hotel reservations. The data obtained using the automatic system (as well as the WOZ data) has been transcribed. Word labels were used and we added several labels that describe nonspeech sounds related to the use of the telephone.

In view of extending the system to handle more complex travel planning phenomena (for example, where conflicts in meeting times and available transport might occur), a second Wizard of Oz experiment was carried out.

The dialogs have been assessed using a per-turn notation that reflects both system and user goals

2. DATA TYPES

2.1 Human-Human dialogues

As of the writing of this article, we have collected over 3160 dialogs for the Carnegie Mellon Communicator [1]. This system carries out a telephone dialog with the user in order to make plane, car, and hotel reservations for business travellers.

The automatic system was built in a series of steps, proceeding from human-human speech to human-machine dialogs. First we recorded an experienced travel agent making reservations for trips that people in the Carnegie Mellon Speech Group were taking in the upcoming months. These recordings took place in two stages.

In the first stage, the travel agent was aware that she was being recorded, but unaware of our goals. In the second stage, the travel agent was made aware of the most general project goals, i.e. creating the system described in the first paragraph above, and was instructed to be more terse, to implicitely encourage the caller to be terse as well, and to avoid out-of-domain conversation topics.

The dialogues were analysed as to their goals and subgoals, the timeline of the goals (is goal y dependant on goal x, and must it always follow x?), and the types of negociations that took place. The

10.21437/Eurospeech.1999-494

human-human dialogs are used as a standard reference in building and refining the automatic system (when new functions are added, when system response appears inappropriate, etc.). It is considered to be stable, and flexible, representing an expert's breadth of knowledge, ease of interaction, client profiles, and freedom to propose novel solutions where necessary.

2.2 First Wizard of Oz dialogues

After examining the 58 dialogs collected in this way, a Wizard-of-Oz (WOZ) system was created where the human operator typed in responses to user spoken input. The "Wizard" had observed the travel agent and was to imitate the latter's behavior and speech patterns as well as respect the limitations of the first totally automatic system, such as only dealing with US destinations and flights that are round trip. The Wizard used the USAIRWAYS Priority Travel Works webpage as a real source of information for the dialogs. The Wizard typed her responses and they were synthesized to the caller. The first trials of the Wizard imitated a "system initiative" type dialogue. This constraining interaction was soon abandoned in favor of "mixed initiative" dialogue. A total of 107 WOZ dialogs were collected in this manner.

2.3 Prototype dialogues

Thereafter a prototype version of the automatic system became available and was used to capture human-machine dialogs. The system outputs one wavefile per user turn. All system actions are recorded in a logfile. When the speech has been transcribed (see below), the transcriptions are merged with the logfile ("mergefile"), thus affording a complete, time-accurate, analysis of the dialogue.

The system makes simple flight reservations of one or more legs, simple car reservations and simple hotel reservations. It is generally available and may be accessed toll free at 1-877-CMU-PLAN in the United States, and, elsewhere at 1-412-CMU-1084.

2.4 Second Wizard of Oz dialogues

In view of extending the system to handle more complex travel planning phenomena (for example, conflicts in meeting times and available transport), we carried out a second Wizard of Oz experiment. This time the Wizard had flight information for many smaller cities as well as information about which airport, for multi-airport cities, had the most convenient flights. She also had charts of driving times and distances between pairs of cities, and between landmarks and the five closest airports.

The user had one of four scenarios to enact. Each scenario involved three to four meetings, small cities, and tight schedules. The user had to either change schedules or transport (a car instead of flying) on his own, or in verbal negociation with the Wizard ("you might leave the night before instead"). Although schedules were tight, we could tighten them even more in the future to force the user to change appointments, implying the use of a multi-agent system including a scheduler.

The four scenarios' content can be characterised as follows, where the italicised place name is where connection times were tight:

1) Washington DC -> San Diego -> Point Loma -> *Miramar* -> *Monterey* -> Washington DC
2) Detroit -> Pittsburgh -> *Seven Springs* -> Lansdowne VA -> Detroit
3) Cleveland -> Philadelphia -> *Hawthorne NY* -> Washington DC -> Cleveland
4) Boston -> *Pittsburgh* -> *Youngstown* -> Boston

The goals addressed in all the dialogues were classified as follows:

- in every dialog (7 goals): user id; out leg; return leg; car; hotel; pay with credit card; end session.
- optional (7 goals): intermediate legs; summary of itinerary; cost; driving info; obtaining a map; sending email; communication management.

There were 22 to 44 turns per dialogue. This number seems to be speaker-related (two of the speakers did two dialogs each), but data is too sparse to confirm this. There were 8 to 14 goals per scenario. This broke down as follows:

Scenario	**number of goals**
1	12-13
2	10-11
3	10-14
4	8-13

Table 1. Number of goals per scenario

Table 1 shows that scenarios 3 and 4 have a more variable number of goals. We need further evidence to determine if this is due to the way the scenarios were defined, which users got these scenarios, or some other source.

The user took the initiative in the dialogue for from 3% and 48% of the turns. From questions asked after the sessions, we attribute this to speaker differences; it does not seem related to a specific scenario. The users tend to fall into a continuum going from those who used the system for information only to those who relied on the system to help define options (fly or drive, for example), and make decisions. Users took the initiative in the dialogue for the following reasons: 1) change to another goal; 2) request information outside present goal; 3) communication management (meta-goal).

2.5 MOVIELINE dialogues

A smaller application was constructed with the goal of obtaining large amounts of speech data under conditions not too different from the CMU Communicator. Where the CMU Communicator is still a prototype and can give real flight times, for example, as of this writing, it still requires the final reservation to be made by a travel agent. Movieline offers information immediately - the caller thus derives an immediate benefit and uses the system more voluntarily. This system gives movie listings and times for cinemas around Pittsburgh. Speech and logfiles are collected in the same way as for the Communicator.

Type	No. dialogs	No. Utts	Duration
Human-Human	58	1800	~1hr
WOZ 1	107	1992	1.3hrs
Automatic	2861	44783	11.4hrs
Movieline	122	8057	2.8hrs
WOZ 2	16	487	~.3hrs
TOTALS	3164	51119	16.8hrs

Table 2. Quantity of different types of data

3. TRANSCRIPTION

This large amount of data (see Table 2) has been labelled as we collect it. The labelling is carried out in-house and tends to follow collection at a constant one-day-behind (if today is Tuesday, data collected through yesterday, Monday, has been transcribed).

There are three transcribers. Each wavefile is seen by at least two of them, one to annotate and the other to correct. They use the Carnegie Mellon Scribe annotation tool, which assists them by automating the process of loading and playing audio files from a dialog session, and writing the transcription to the appropriate file. The tool allows a transcriber to traverse the logged data, and listen to the logged audio from a given session. The transcriber's i.d. information, along with the date and time of the transcription, is stored in the directory of each session, and is displayed when the session is loaded. Scribe also presents the list of accepted tags for annotating the transcription, that can be automatically inserted using hot keys. Additionally Scribe uses the dictionary from the respective system to check the quality of the transcription, and alert the transcriber to possible misspellings before he commits to them.

Some wavefiles undergo a third verification. After labelling and rechecking, we automatically detect any words in the transcripts that are not in the system lexicon or accepted human and non-human noise notation). The list of the files where oovs (words not in our lexicon) are detected (and specifics about what has been detected) are given to the transcribers. They then access the indicated files and make any necessary changes.

As mentioned above, word labels were used. We added a few conventions and several labels that describe nonspeech sounds related to the use of the telephone. First, we are transcribing in truecase, where all words are written in lower case except proper nouns, which have their first letter capitalised. Second, some proper names that are composed of several words which are often seen together, are written together, separated by an underscore rather than a space.

Sounds annotated include feedback from the synthesizer, hangup, knocking the handset, etc. The transcribers note the difference between a sudden sound and a continuous one during speech. Novel human speech notations include “mumble” where the caller is talking to someone else, and “far” where the transcriber perceives that the caller is holding the handset at a greater distance from his mouth than expected.

4. SYSTEM ASSESSMENT

We have been using logfiles, transcripts, and mergefiles to derive useful information for system assessment. We believe that using the following annotations on the above data will help us:

The mergefiles are hand annotated *per turn* in an Excel sheet with the following: dialog number, system or user initiative; type/cause of error (if present); system subgoal; user subgoal.

Label	Meaning
LxAC	Legx Arrival City
LxDD	Legx Departure Date
LxWH	Legx Need a Hotel
LxHL	Legx Hotel Location
LxCO	Legx Accept, Reserve

Table 3. Examples of subgoal (system and user) notation

and *per goal* with the following: number of turns; was-goal-completed; number of errors; mean duration per turn; system backoff; number of system subgoal repeats; number of user subgoal repeats; number of times system subgoal = previous user subgoal.

The subgoal notation shown in Table 3 is similar to the attribute values used by [3] and others. For system backoff, we simply note whether the system had asked for confirmation of a flight or hotel reservation ("Is that okay?", concluding a goal) and then was flexible enough, in case of a problem, to be able to go back within the goal and discuss one of the subgoals that the reservation was dependant on (say, "Do you want a hotel downtown?"). The notation we use should be useful in error detection [2] in our system.

In order to establish assessment criteria that we may use to compare different versions of our system over time and compare our system to others, we combine several of the measures above. One figure we are finding useful combines system and user repeats, as a function of the number of turns. We also use the percentage of turns in the within the goal where the system subgoal was the same as the user's previous subgoal ("I'd like to go to Pittsburgh"; "Where would you like to go?"). The latter figure gives us, where the number of turns for a given goal is above the expected number, a way to separate the "bothersome" parts of the dialog, such as the aforementioned Pittsburgh example, from extra turns in the dialog where the user is navigating possibilities such as later flights, and may be very happy with his interaction with the system. We now plan to compare our figures to other criteria, such as user satisfaction in order to validate their usefulness.

5. CONCLUSION

We have described data collection and transcription for the Carnegie Mellon Communicator. We have analysed some of the interaction in the second set of Wizard of Oz data (WOZ2). We are also starting to analyse the dialogues we obtain with the automatic system in view of overall assessment through subgoal notation.

This research was sponsored by the Space and Naval Warfare Systems Center, San Diego, under Grant No. N66001-99-1-8905. The content of the information in this publication does not necessarily reflect the position or the policy of the US Government, and no official endorsement should be inferred.

6. REFERENCES

[1] Rudnicky, A. , Thayer, E., Constantinides, P., Tchou., C.,Shern, R., Lenzo, K., Xu, W., Oh, A., (1999), Creating Natural Dialogs in the Carneige Mellon Communicator System, *Proceedings Eurospeech 99*.

[2] Constantinides, P., Rudnicky, A., (1999), Dialog Analysis in the Carnegie Mellon Communicator, *Proceedings Eurospeech 99*.

[3] Walker, M., Litman, D., Kamm, C., Abella, A., (1997), PARADISE: A framework for evaluating spoken dialogue agents, *Proceedings of the 35th annual meeting of the Association for Computational Linguistics (ACL-97)*, p. 271-280.