The acquisition and dialog-act labeling of the EDECAN-SPORTS corpus

Lluís-F. Hurtado, Fernando García, Emilio Sanchis, Encarna Segarra

Departament de Sistemes Informàtics i Computació
Universitat Politècnica de València, Spain
{lhurtado, fgarcia, esanchis, esegarra}@dsic.upv.es

Abstract

In this paper, we present the acquisition and labeling processes of the EDECAN-SPORTS corpus, which is a corpus that is oriented to the development of multimodal dialog systems acquired in Spanish and Catalan. Two Wizards of Oz were used in order to better simulate the behavior of an actual system in terms of both the information used by the different modules and the communication mechanisms between these modules. User and system dialog-act labeling, as well as other information, have been obtained automatically using this acquisition method Some preliminary experimental results with the acquired corpus show the appropriateness of the proposed acquisition method for the development of dialog systems.

Keywords: Corpus Acquisition, Wizard of Oz, Dialog System

1. Introduction

A first step in the development of many spoken dialog systems is the acquisition of a training corpus that can be used for learning the models of the different modules of the system: automatic speech recognition, language understanding, and dialog management.

The quality of this corpus is directly related to its capability to represent the expected behavior of future interactions with real users. This is the reason why the Wizard of Oz technique (Fraser and Gilbert, 1991) is widely used during the acquisition process to simulate system behavior (Salber and Coutaz, 1993) (Serrano and Nigay, 2010). In this acquisition process, it is also important to use as many real implemented modules (speech recognizer, understanding, or TTS synthesizer) as possible in order to better simulate real system behavior.

In this paper, we present the acquisition and labeling of the EDECAN-SPORTS corpus. This corpus was acquired within the EDECAN Spanish project (Lleida et al., 2006). The EDECAN-SPORTS task consists of providing information about court availability as well as booking and cancellation of sports facilities at our University. The service is provided via a multimodal information kiosk placed in a public hall of an education center of the University. Some of the main characteristics of this task are the following: it is restricted at the semantic levels (i.e., related to a specific domain); and it is unrestricted at the syntactic and lexical levels (i.e., open vocabulary and spontaneous speech). Also, the system must access the information system of the University, fulfilling the constraints, formats, and protocols of this information system. The output is multimodal, combining speech with textual and graphic information. Multilingualism must be allowed given the language characteristics in our country (Spanish and Catalan).

Figure 1 shows the screen of the kiosk at a specific moment of interaction with a real user. The transcription of the TTS synthesizer is shown; in this example, the system output is *We indicate in the screen the available courts, do you want anything else?*. The screen also contains a table showing the available courts. This type of interaction with the user is more friendly than the telephone interaction since the user can see the different courts and timetables in the table, 6 different choices in this example. This amount of information cannot be provided in a natural manner via telephone.

We have designed the acquisition process using as many preliminary implemented modules as possible. To do this, we analyzed human-human dialogs provided by the sports department of our University, which had the same domain as the EDECAN-SPORTS task. From these dialogs, we defined the semantics of the task in terms of dialog-acts for both the user utterances and the system prompts, and we labeled the turns of the dialogs. Thus, we had a small initial corpus for the EDECAN-SPORTS task. From this small corpus, we learned a preliminary version of a statistical dialog manager that was developed in our laboratories (Griol et al., 2008). This dialog manager was used as a prototype in the supervised process of acquiring a larger corpus by means of the Wizard of Oz technique. Since the initial corpus is not large enough to train suitable models for automatic speech recognition nor language understanding modules, we did not have a preliminary version of this module for the acquisition process. Our proposal is based on using a specific Wizard of Oz to play the role of the natural language understanding module and a second Wizard of Oz to supervise the dialog manager prototype.

Both the use of two Wizard of Oz and a preliminary version of the dialog manager allow us to obtain an adequate corpus that simulates system behavior. In addition, the manual labeling effort is minimized because user and system dialog-acts are obtained at the end of the acquisition process.

The paper is structured as follows. Section 2. presents the semantic restrictions of the task, the set of user and system dialog-acts, and an example of a dialog. Section 3. describes a scheme of the platform developed for the acquisition of the EDECAN-SPORTS corpus. Section 4. describes the corpus acquired through this platform. Section 5. presents some preliminary experiment results of language understanding. And finally, some conclusions are presented in Section 6..

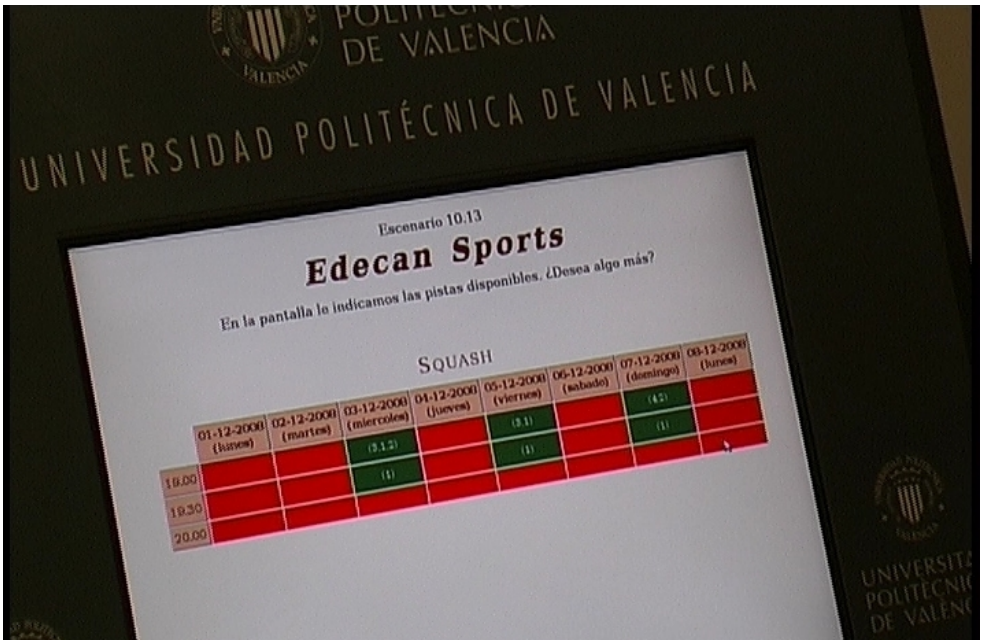

Figure 1: The kiosk screen at a specific moment of interaction with a real user

2. Acquisition design

To ensure that the acquired corpus would be useful to learn models for the main modules of the spoken dialog system, the acquisition was thoroughly designed. Although the user was allowed to use open vocabulary and spontaneous speech, the acquired dialogs had to be restricted at the semantic level, that is, the semantics of the dialogs had to match the semantics of the task.
To ensure that the dialogs were semantically restricted, 11 types of scenarios were defined for the acquisition to cover all the possible use cases of the task. Each scenario type contained up to four different goals that the user should achieve:

- AVAILABILITY the user ask for the court availability, the user can also give some additional information and restrictions, such as SPORT, DATE, HOUR, etc.
- BOOKING the user books a sport court; the system needs to know at least SPORT, DATE and HOUR.
- CANCELLATION the user request to cancel a reservation made previously.
- BOOKED the user ask for information about a sport court that was previously booked.

They were 4 basic scenarios, one for each one of the 4 goals. We designed 7 additional scenarios that include a combination of two basic scenarios: BOOKING + BOOKING, CANCELLATION + CANCELLATION, BOOKED + CANCELLATION, AVAILABILITY + BOOKED, AVAILABILITY + CANCELLATION, BOOKING + BOOKED, and BOOKING + CANCELLATION. Some goals were specifically designed to be impossible to achieve. Hence, by including negative situations, the acquired corpus should have more variability.

2.1. User and system turn representation

Based on the analysis of the human-human corpus provided by the sports department, the semantics of the task (the set of user and system dialog-acts) was defined using the concept of frame.
A set of 4 task-dependent concepts, 3 task-independent concepts, and 6 attributes were defined to represent the semantics of the user turns.

- Task-dependent concepts represent the user's intentions: (AVAILABILITY), (BOOKING), (CANCELLATION), (BOOKED).
- Task-independent concepts represent generic concepts: (ACCEPTANCE), (REJECTION), (NOT_UNDERSTOOD).
- Attributes represent constraints on the user's intentions: SPORT, HOUR, DATE, COURT_TYPE, COURT_NUMBER, ORDER_NUMBER.

A set of 19 concepts were defined to represent the semantics of the system turns:

- To ask for specific information: (SPORT), (HOUR), (DATE), (COURT_TYPE), (COURT_NUMBER), (ORDER_NUMBER), (BOOKING_CHOICE), (CANCELLATION_CHOICE).
- To show information to the user: (AVAILABILITY), (BOOKED).
- To ask for explicit confirmation: (CONFIRMATION), (CONFIRMATION_BOOKING), (CONFIRMATION_CANCELLATION).
- To inform the user of the result of an action: (BOOKING), (CANCELLATION), (ERROR_INFO).

- To start and finish the dialog: (OPENING), (NEW_QUERY), (CLOSING).

Figure 2 shows an example of a labeled dialog that was translated to English from the original Catalan.

S1: Welcome to the sport service. How can I help you?
(OPENING)

U1: I want to know the availability of tennis courts on May the 26th in the evening.
(AVAILABILITY)
SPORT:tennis
DATE: 26-05-2011
HOUR: evening

S2: There are three available hours on May the 26th in the evening: from four to five, from six to seven, and from seven to eight. Which do you want to book?
(AVAILABILITY)
AV-NUM:3
DATE: 26-05-2011
HOUR: [16:00-17:00][18:00-19:00][19:00-20:00]
(BOOKING_CHOICE)

U2: I would like to book the one from six to seven.
(BOOKING)
HOUR:[18:00-19:00]

S3: Do you want to book a tennis court on May the 26th from six to seven?
(CONFIRMATION_BOOKING)
SPORT:tennis
DATE: 26-05-2011
HOUR: [18:00-19:00]

U3: Yes, I do.
(ACCEPTANCE)

S4: I have just booked you the tennis court number 4 on May the 26th from six to seven in the evening. Do you want anything else?
(BOOKING)
SPORT:tennis
COURT_NUM:4
DATE: 26-05-2011
HOUR: [18:00-19:00]
(NEW QUERY)

U4: No, thank you.
(REJECTION)

S5: Thank you for using the sport service. Goodbye.
(CLOSING)

Figure 2: A labeled dialog from the EDECAN-SPORTS corpus

3. The acquisition platform

Figure 3 shows a scheme of the platform developed for the acquisition of the EDECAN-SPORTS corpus. The first wizard listens to the user utterances and provides a semantic representation of these utterances in terms of frames. This way the semantic labeling of the user turns is performed at the same time as the acquisition process.

In order to have more realistic samples to learn the dialog manager model even when recognition/understanding errors are present, the correct semantic representation is automatically modified to introduce some errors. This error-simulation (Garcia et al., 2007) is based on the analysis of the errors in the recognition and understanding processes generated when our models were used with another corpus from a task with similar characteristics. During the acquisition, both semantic labellings (the correct one and the modified one) are stored.

We have developed an approach to dialog management using a statistical model that is estimated from a dialog corpus (Griol et al., 2008). This model was automatically learned from a dialog corpus labeled in terms of dialog-acts. From the human-human corpus, a prototype of the dialog manager module was implemented to be included in our acquisition system. The dialog manager module receives the modified semantic representation of the user turns and generates two outputs in terms of dialog-acts: a query to the University sports department information system manager, and the answer to the user.

The second wizard supervises the behavior of the dialog manager prototype. Sometimes the results (specifically their cardinality) can influence the response of the dialog manager. Therefore, the supervision of the dialog manager is carried out by means of two applications. The first application is used to supervise the results of the queries to the information system. The second one is used to supervise the answer to the user that is automatically generated by the dialog manager. The Wizard of Oz corrects the answer when she/he considers that it is inadequate according to the dialog state and the result of the query to the information system. In the acquisition of the EDECAN-SPORTS corpus, less than 15% of the system answers had to be corrected by the Wizard of Oz.

All the information that had been considered by the dialog manager in order to choose its actions was stored for each system turn.

These two WOz allow us to simulate the real system behavior since they take their decisions by considering the same information that will be supplied to the future modules.

4. The EDECAN-SPORTS corpus

We acquired a set of 165 dialogs for the EDECAN-SPORTS task using the platform described above. A total of 16 different speakers from different geographic origins (the headquarters of the research teams of the EDECAN consortium participated in this endeavor). The languages involved in the acquisition were Spanish (15 speakers) and Catalan (3 speakers). Two of the speakers acquired dialogs in both Spanish and Catalan.

The information available for each dialog consists of four audio channels, the transcription of the user utterances, and the semantic labeling of the user and system turns. Table 1 shows the main characteristics of the acquired corpora in both Spanish and Catalan.

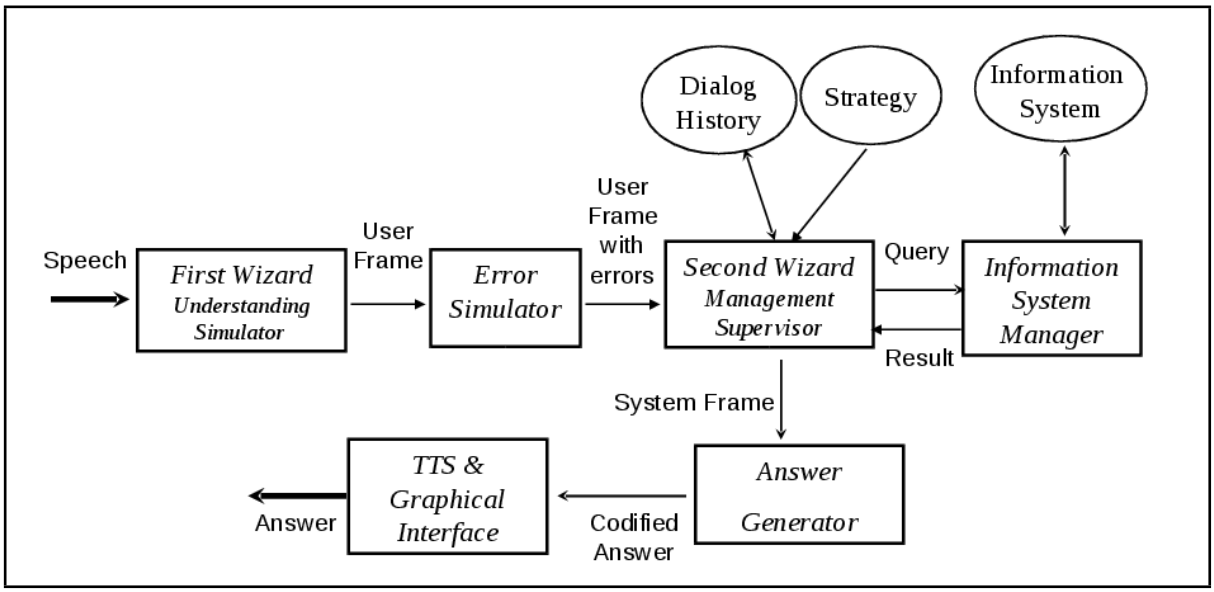

Figure 3: The platform for the acquisition of the EDECAN-SPORTS corpus

	Spanish	Catalan
Number of speakers	15	3
Number of dialogs	137	28
Number of user turns	687	144
Average user turns per dialog	5.01	5.14
Number of words	4,740	995
Vocabulary size	335	179
Average words per user turn	6.90	6.91
Number of user concepts/attributes	1,283	285
Number of system concepts	954	212

Table 1: Main characteristics of the EDECAN-SPORTS corpus

5. Corpus evaluation

In order to evaluate the acquired corpus, a preliminary language understanding experiment was performed. Two different understanding models were learned using the labeled corpora, one for Spanish and another one for Catalan. Each corpus was split into one training set (with 70% of the user turns) and one test set (with the remaining 30% of the turns). The training sets contained 477 turns for Spanish and 109 for Catalan, while the test sets included 210 turns for Spanish and 35 for Catalan.

To learn the statistical model from the training corpora and evaluate the test corpora, we used Conditional Random Fields, which is a framework for building probabilistic models to segment and label sequence data (Lafferty et al., 2001). Specifically, we used the CRF++ open source toolkit (*http://code.google.com/p/crfpp/*). The default parameters were used for the experimentation.

In order to evaluate the understanding results, two measures were used:

- The Correct Turn Rate (CTR). This measure represents the percentage of user turns for which the output of the language understanding process was exactly as in the reference.
- The Concept Accuracy (CA). This measure is similar to the Word Accuracy measure widely used in Automatic Speech Recognition but applied to understanding tasks (i.e., the evaluation units are the concepts and attributes).

It must be noted that, in order to increase the coverage of the understanding model, a categorization process was performed throughout all the experimentation. The used categories included months, day of the week, numbers, and sport names.

The results for both Spanish and Catalan are presented in Table 2. While the difference between the experimental result for Spanish and for Catalan in the CA measure is small, this difference increases a lot in the CTR measure. The lack of training data for the Catalan corpus results in a greater number of user turns not being decoded as in the reference, in comparison with the corpus in Spanish. This difference is barely noticeable when all the units of the semantic interpretation (i.e., the concepts and the attributes) are taken into account.

	Spanish	Catalan
CTR	84.8	68.6
CA	90.8	86.1

Table 2: Results for the language understanding evaluation of the EDECAN-SPORTS corpus

In addition, we have used the EDECAN-SPORTS corpus described in this paper to learn dialog management models based on stochastic finite-state transducers developed in our laboratories. More details may be found in (Hurtado et al., 2010).

6. Conclusions

In this paper, we have presented the acquisition and labeling of the EDECAN-SPORTS corpus. In order to obtain an adequate corpus and minimize the manual labeling effort, we used preliminary implemented modules.

We also used two Wizards of Oz that had the same limited information as what will be used by the future automatic system. Due to the method of acquisition, the labeling was performed at the same time as the acquisition process.

The experimental results presented in this work show the adequacy of the corpus and the labeling for the development of dialog systems.

Acknowledgements

Work partially supported by the Spanish MICINN under contract TIN2011-28169-C05-01 and by the Vicerrectorat d'Investigació, Desenvolupament i Innovació of the Universitat Politècnica de València under contract 20100982.

7. References

M. Fraser and G. Gilbert. 1991. Simulating speech systems. In *Computer Speech and Language*, volume 5, pages 81–99.

F. Garcia, L.F. Hurtado, D. Griol, M. Castro, E. Segarra, and E. Sanchis. 2007. Recognition and Understanding Simulation for a Spoken Dialog Corpus Acquisition. In *TSD 2007*, volume 4629 of *LNAI*, pages 574–581. Springer.

D. Griol, L. F. Hurtado, E. Segarra, and E. Sanchis. 2008. A statistical approach to spoken dialog systems design and evaluation. *Speech Communication*, 50(7-9):666–682.

Lluís-F. Hurtado, Joaquin Planells, Encarna Segarra, Emilio Sanchis, and David Griol. 2010. A stochastic finite-state transducer approach to spoken dialog management. In *Proc. of InterSpeech 2010*, pages 3002–3005, Makuhari, Chiba, Japan.

John Lafferty, Andrew McCallum, and Fernando Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. 18th International Conf. on Machine Learning*, pages 282–289. Morgan Kaufmann, San Francisco, CA.

E. Lleida, E. Segarra, M. I. Torres, and J. Macías-Guarasa. 2006. EDECÁN: sistEma de Diálogo multidominio con adaptación al contExto aCústico y de AplicaciónN. In *IV Jornadas en Tecnologia del Habla*, pages 291–296, Zaragoza, Spain.

Daniel Salber and Jolle Coutaz. 1993. A wizard of oz platform for the study of multimodal systems. In *Proc. of Join IFIP/ACM Conference on Human Factors in Computer Systems, INTERCHI'93*, Amsterdam, The Netherlands.

M. Serrano and L. Nigay. 2010. A wizard of oz component-based approach for rapidly prototyping and testing input multimodal interfaces. *Journal on Multimodal User Interfaces*, 3(3):215–225.