A prototype of a spoken dialog system based on statistical models

E.Segarra, L.Hurtado, J.A.Gómez, F.García, J.Planells, J.Pastor, L.Ortega, M.Calvo, E.Sanchis

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

Abstract

We present in this paper a prototype of a spoken dialog system. One of the characteristics of this system is that most of the modules (speech recognition, understanding and dialog manager) are based on statistical models. The system has the possibility of easily change the task or the language by means of interchanging the different modules. We present in this case a dialog task consisting of booking of sports facilities in the University.
Index Terms: spoken dialog system, statistical model

1. Introduction

The development of spoken dialog systems is a complex process that involves the design, implementation and evaluation of a set of modules that deal with different knowledge sources.

Currently, some of the most successful approaches are based on statistical models estimated using training corpus. Statistical models have been widely used in speech recognition, language modeling and speech understanding. Although in the case of Dialog Managers most of the approaches are based on the manual design of the system behavior, over the last few years, approaches that use statistical models to represent the dialog manager have also been developed ([1], [2], [3], [4]).

In this field, we have recently developed an approach to manage the dialog using statistical models that can be learnt from a data corpus. This work has been applied within the EDECAN-SPORT [5] domain that consist of a task for the booking of sports facilities in the University. In order to increase the reliability of the system, some features were included in the system:

- Unlike classical slot filling tasks, our Dialog Manager takes its decisions not only based on the previous user turns but also considering the information supplied by the Application Manager and it can perform both actions: to provide information and to modify the application data (i.e. after booking or cancelling a court).
- In some cases the system must give many data to the user that are difficult to provide by speech in a concise way. That is the case of offering some courts in some days of the week. In order to convert this in a more friendly communication process a multimodal input/output is available in the system. Therefore in some cases the system gives the information by using a graphical interface, or the user can use the voice or the touchscreen
- In order to take into account some information about the preferences of the user, previous dialogs are saved and its information is used as a priori knowledge.

The prototype has been installed in a kiosk to be accessible to the users in some areas of the University. At the moment it is installed in our laboratory to evaluate its behavior.

2. The spoken dialog system

In figure 1 a scheme of the dialog system developed is presented. The system has been implemented using the architecture defined in the SD-TEAM project [6]. This architecture allows the integration, substitution and collaboration of the modules even if they are located in different computers. The system contains the habitual modules of a Dialog system, ASR, Language Understanding, Dialog Manager, Answer Generator and TTS, besides the specific modules related to the Application Manager and the multimodal user interface. The Application Manager controls the access to the database, not only to provide information but also to modify it when booking or cancellation must be done. In the case of multimodality, we have include two possibilities: the user has a touchscreen to select an item, and the system can give some informations in term of tables. It must be noted that both modules speech understanding and dialog manager are based on statistical models learnt from training samples, as it is described in next sections.

3. The training corpus

In order to design the system, we firstly analyzed human-human dialogs provided by the sports area of our university, which have the same domain that the defined for the EDECAN-SPORT task. From these dialogs we defined the semantics of the task in terms of dialog acts for both the user utterances and system prompts and we subsequently labeled these dialogs. Thus, we had a very small initial corpus for the EDECAN-SPORT task. From this small corpus we learned a preliminary version of the dialog manager. Then we acquired a training corpus by means of a Wizard of Oz technique [7], as it is shown in figure 2. The special characteristic of this acquisition is that we used the preliminary dialog manager learnt from the human-human corpus, and two Wizard were used: one of them for the understanding process and the other to supervise the dialog manager. The reason of this approach is to better simulate a human-machine interaction.

Using this approach a set of 240 dialogs was acquired for our task(a total of 18 different speakers from different Spanish regions). The languages involved in the acquisition were Spanish, Catalan and Basque. A set of 15 types of scenarios were defined in order to cover all the possible use cases of the task. The information available for each dialog consisted of four audio channels, the transcription of the user utterances (with an average of 5.1 user turns per dialog and 6.7 words per user turn) and the semantic labeling of the user and system turns.

Once the corpus was acquired a semi-automatic annotation process was performed. For the user turns (the set of user dialog acts) we defined four task-dependent concepts (*Availability*, *Booking*, *Booked*, *Cancellation*), three task-independent concepts (*Affirmation*, *Negation*, and *Not-Understood*) and six attributes (*Sport*, *Hour*, *Date*, *Court-Type*, *Court-Number*, and

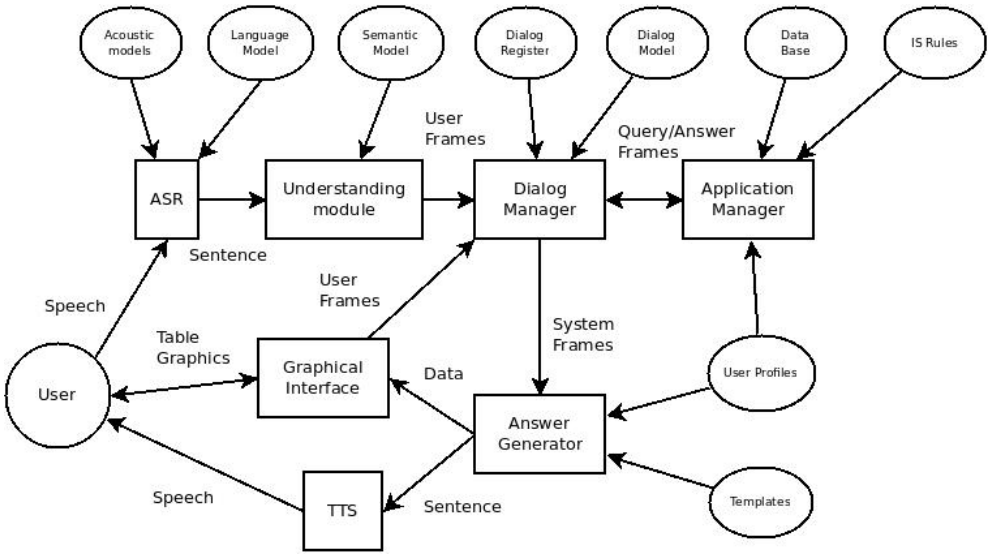

Figure 1: *The EDECAN architecture.*

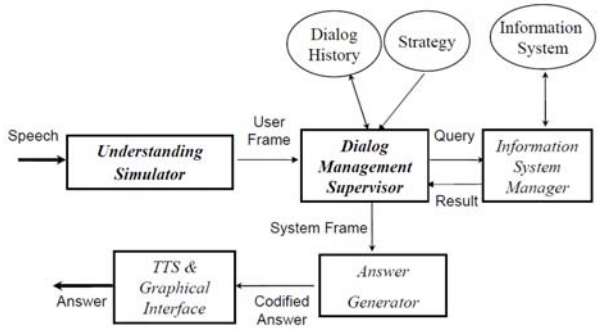

Figure 2: *Acquisition schema using a Wizard of Oz technique.*

Order-Number).

An example of the semantic interpretation of an input sentence is shown below:

I want to book a basketball court for tomorrow.

Semantic Representation:

(Booking)

Sport: basketball

Date: tomorrow

The labeling of the system turns in terms of system dialog acts is similar to the one defined for the user turns. A total of 21 concepts were defined: Task-independent concepts (*Opening* and *Closing*); concepts used to inform the user about the result of a specific query (*Availability*, *Booking*, *Booked*, and *Cancellation*), concepts defined to ask for the attributes that are necessary for a specific query (*Sport*, *Date*, *Hour*, *Court-Number*, and *Court-Type*), concepts used for the confirmation of concepts (*Confirmation-Availability*, *Confirmation-Booking*, *Confirmation-Booked* and *Confirmation-Cancellation*), and attributes (*Confirmation-Sport*, *Confirmation-Date*, *Confirmation-Hour* and *Confirmation-CourtType*).

An example of the labeling of a system turn is shown below:

To play basketball tomorrow, there are two courts: court number 3 at 10:00 and court number 1 at 16:00. Please choose one.

Semantic Representation:

(Booking-Choice)

Sport: basketball

Date: tomorrow

Hour: 10:00 16:00

Court-Number: 3 1

This annotated corpus was used to learn the language models, semantic models and Dialog Manager models.

4. The speech recognition module

The SD-TEAM architecture allows the integration of multiple ASR modules. We are using, in an interchangeable way, both Loquendo ASR and the one developed in our laboratory.

The latter is HMM based, uses a standard speech preprocessor, a n-gram language model and a Viterbi-based search. The speech signal is pre-emphasized by means of a high-pass FIR filter $H(z) = 1 - 0.95z^{-1}$ and then pre-processed to obtain a sequence of frames or acoustic vectors. A 20 ms Hamming window is applied every 10 ms to obtain each frame, which contains 39 parameters: Energy, the first 12 MFCC, and their first and second derivatives. The HMM were trained by means of HTK from the Albayzin Spanish corpus. Albayzin is a phonetically balanced corpus consisting of six hours of speech [8].

5. The understanding module

We propose an understanding process [9] that works in two phases (see figure 3).

The first phase consists of a transduction of the input sentence in terms of an intermediate semantic language. In the second phase, a set of rules transduces this intermediate representation in terms of frames. As the intermediate language is close to the frame representation, this phase only requires a small set of rules to construct the frame. This second phase consists of the following: the deletion of irrelevant segments of the input sentence, the reordering of the relevant concepts and attributes that appeared in the user sentence following an order which has been defined a priori, the automatic instantiation of certain task-dependent values, etc. This last action consists of

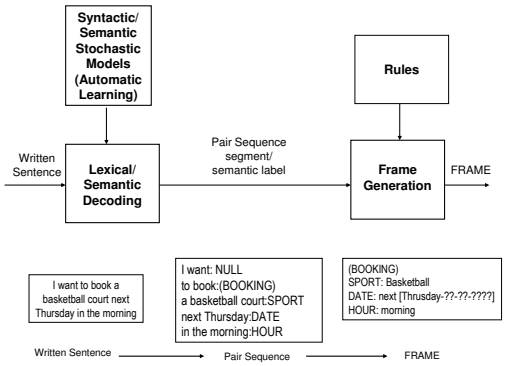

Figure 3: *Understanding module diagram.*

the conversion of dates and hours into their canonical values. For example, "on September the 15th" into "[2010-15-09]".

For the intermediate language 14 labels were defined, corresponding to the so-called semantic units: *NULL*, *(AVAILABILITY)*, *SPORT*, *(REJECTION)*, *(BOOKING)*, *(ACCEPTANCE)*, *DATE*, *HOUR*, *COURT-NUM*, *(BOOKED)*, *(CANCELLATION)*, *ORDER-NUM*, *COURT-TYPE*, and *NOT*. The goal of the first phase is to find the best sequence of semantic units given the input sentence and a two-level statistical semantic modelization is used in this phase. Figure 4 shows the statistical semantic model. The meaning of each sentence is represented as a sequence of semantic units, and it is associated a segmentation of the sentence in terms of the corresponding semantic units. From an annotated training corpus we learn two kind of models: one of them represents the concatenations of semantic units, and the other represents the lexical realization of each semantic unit (that is, the model of segments of words associated to each semantic units). In both cases the models used for this task are bigrams, that is, both bigrams of semantic units and of words into each semantic unit. The decoding process consists of a Viterbi search over the integrated network, that supplies not only the best sequence of semantic units but also the segmentation of the input sentence associated to it. This segmentation is used in the second phase of the semantic module to associate the values to the attributes (after a normalization process, if necessary).

In other words, given the input sentence $w = w_1 w_2 \cdots w_n \in W$, the process consists of finding the sequence of semantic units $v = v_1 v_2 \cdots v_k \in V$ which maximizes the probability:

$$\hat{v} = \underset{v}{\operatorname{argmax}}\, P(w|v)P(v)$$

The term $P(w|v)$ is the probability of the sequence of word w given the sequence of semantic units v. We estimate this probability (following the Viterbi algorithm) as the maximum for all segmentations of w in $|v|$ segments.

$$P(w|v) = \max_{\forall l_1, l_2, \ldots, l_{t-1}} \Big\{ P(w_1, \ldots, w_{l_1}|v_1)P(w_{l_1}) \cdot P(w_{l_1+1}, \cdots, w_{l_2}|v_2) \cdot \cdots \cdot P(w_{l_{k_1}+1}, \ldots, w_n|v_k) \Big\}$$

If bigram models are used, the probability of each segment given the associated semantic unit is:

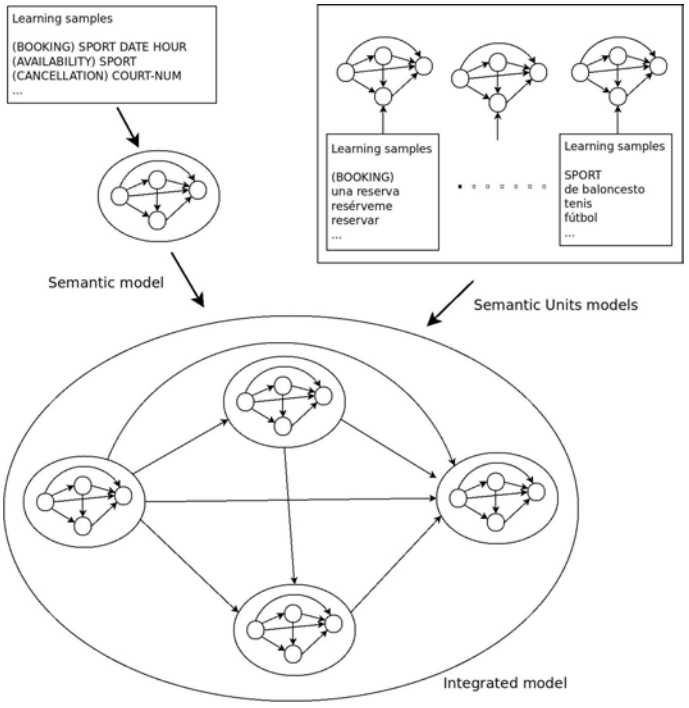

Figure 4: *Understanding model.*

$$P(w_i, \ldots, w_j|v_s) = \prod_{k=i}^{j} P(w_k|w_{k-1}, v_s)$$

and term $P(v)$ is the bigram probability of the sequence v.

$$P(v) = \prod_{i-1}^{k} P(v_i|v_{i-1})$$

6. Dialog Manager

The dialog model proposed in [10] is based on the transduction concept and on the use of Stochastic Finite-State Transducers. In other words, given a state of the system and a user turn, a system turn is generated and a transition to a new state is done. Therefore, dialog management is based on the modelization of the sequences of system and user dialog turns pairs. Thus, a dialog describes a path in the transducer model from its initial state to a final one.

In a dialog system, the Dialog Manager (DM) is the module devoted to choose the best system answer according to its dialog model during the dialog sequence. We consider a dialog as a sequence of pairs $(u_i, a_i), i = 1 \ldots n$, where u_i is the user utterance at time i and a_i is the answer of the system to this utterance. The system answer a_i is selected taking into account not only u_i but also all the information provided by the user throughout all the dialog sequence. All this information (in terms of concepts and attributes) is stored in a data structure we call Dialog Register (DR).

We have developed a statistical DM based on the use of a Stochastic Finite-State Transducer (SFST). A SFST is defined formally by a 6-tuple $(Q, \Sigma, \Delta, q_0, p, f)$. In our approach the input alphabet represents all the allowed user utterances and the output alphabet includes all the system answers defined for

Figure 5: *The kiosk used for the testing.*

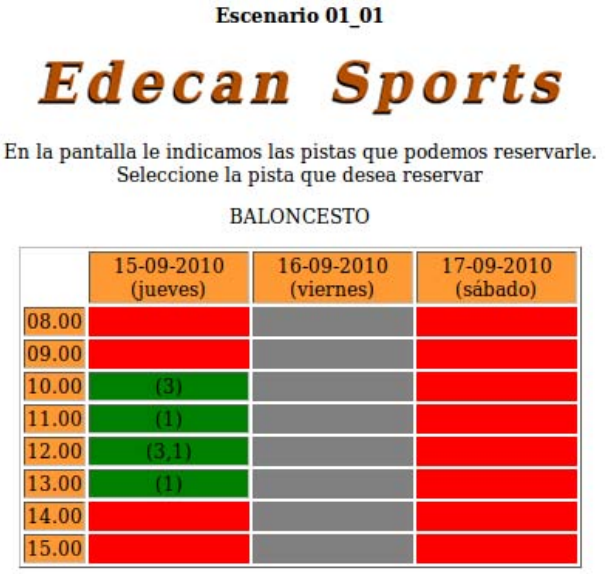

Figure 6: *Screenshot of the visual information for the user.*

the task, that is, the set of system dialog acts. $p(q, u, a, q_0) = Pr(u, a, q_0|q)$ is the transition probability from q to q_0 by observing u and emitting a. In our proposed approach, the selection of the best next system answer at time i (a_i) is made by means of the following local maximization:

$$\hat{a_i} = \underset{a_i \in \Delta}{\operatorname{argmax}}\, p(q_{i-1}, u_i, a_i, q_i) = \underset{a_i \in \Delta}{\operatorname{argmax}}\, Pr(u_i, a_i|q_{i-1})$$

The dialog ends when a final state q_f , is reached. From this point of view, a dialog can be seen as a path in the transducer from the initial state q_0 to the final state q_f .

7. Multimodal interface

The way to generate the system answers is template-based. That is, the Dialog Manager generates an answer frame that is supplied to the Answer Generator module. This module, using some predefined templates generates the sentence for the TTS. Also, as said before, some informations are presented in a table on the screen (see figures 5 and 6). This process is controlled by the Graphical Interface module that also manages the touch screen that can be used as used input for the Dialog Manager.

8. Conclusions

In this paper we have presented a dialog system based on statistical models that has been applied to a task of booking sport facilities in the University. The system has been implemented using a flexible architecture defined for the SD-TEAM project. This prototype shows that it is possible to build complete dialog systems based on statistical models at different levels. One advantage of this modelization is that the statistical models can be dinamically trained when real users interact with the system. We have also explored the possibility of combining multimodality in the input as well as in the output.

9. Acknowledgements

Work partially supported by the Spanish MEC and FEDER under contract TIN2008-06856-C05-02.

10. References

[1] J. Williams and S. Young. 2007. “Partially Observable Markov Decision Processes for Spoken Dialog Systems”. In Computer Speech and Language 21(2), pages 393-422.

[2] F. Torres, L. F. Hurtado, F. Garca, E. Sanchis, E. Segarra. 2005. “Error handling in a stochastic dialog system through confidence measures”. SPEECH COMMUNICATION (ISSN 0167-6393). Volume 45. Pages 211-229.

[3] D. Griol, L. F. Hurtado, E. Segarra, and E. Sanchis. 2008. A statistical approach to spoken dialog systems design and evaluation. In Speech Communication, volume 50, pages 666-682.

[4] David Griol, Llus F. Hurtado, Encarna Segarra, Emilio Sanchis. 2008. “A Statistical approach to spoken dialog systems design and evaluation”. SPEECH COMMUNICATION. Volume 22. Pages 666-682.

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

[6] R. Justo, M.I. Torres, E. Lleida, E. Sanchis, R. de Cordoba, J. Macias-Guarasa. 2010 “SD-TEAM: Tecnologas de aprendizaje interactivo, autoevaluacin y multimodalidad en sistemas de diálogo hablado multidominio”. Procesamiento de Lenguaje Natural n.45 pp. 331-332, Sept, 2010.

[7] 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.

[8] Moreno, A., Poch, D., Bonafonte, A., Lleida, E., Llisterri, J., Mariño, J. B. and Nadeu, C. 1993. “Albayzin Speech Database: Design of the Phonetic Corpus”, in Proceedings of Eurospeech, 1993, volume 1, pages 653–656. Berlin (Germany).

[9] L. F. Hurtado, E. Segarra, F. Garcia, E. Sanchis. “Language Understanding using n-multigram Models”. 4th Internacional Conference EsTAL. Lecture Notes in Computer Science LNAI 3230 (ISSN 0302-9743).

[10] L.F. Hurtado, J. Planells, E. Segarra, E. Sanchis, D. Griol. “A Stochastic Finite-State Transducer Approach to Spoken Dialog Management”. In Proc of Interspeech’2010, pag. 3002-3005, 2010.