A Spoken Dialogue System Based on a Dialogue Corpus Analysis

R. López-Cózar, A.J. Rubio, P. García, J.C. Segura

Dpto. Electrónica y Tecnología de Computadores
Facultad de Ciencias. Campus Universitario de Fuentenueva, s/n
Universidad de Granada. 18071 GRANADA (SPAIN)
Tel.: +34 958 243193, FAX: +34 958 243230, E-mail: {ramon,rubio,pedro,segura}@hal.ugr.es

Abstract

We are developing a spoken dialogue system for Spanish, called SAPLEN, which aims to deal with product orders and clients' queries in fast-food restaurants (López-Cózar *et al.*, 1997; López-Cózar & Rubio, 1997). In this paper we show the kind of information we have obtained from the analysis of a dialogue corpus that we recorded in a fast-food restaurant. We also describe how we have used this information to build the dialogue module of our system. Finally we comment upon some features of the system and show some of our experimental results.

1. Dialogue Corpus

We initially obtained a corpus of more than 500 dialogues from recordings taken in a fast-food restaurant. The quality of the voice is quite poor since the restaurant is a very noisy environment. From this corpus we have labelled manually a sub-corpus of 100 dialogues (called *A-corpus* in this paper) to obtain the following information:

- The set of words our system must recognise and understand, in order to analyse the sentences of the clients.
- The semantic structures which clients usually use, in order to prepare our system for the analysis of the sentences.
- The goals that the clients and the restaurant assistants try to achieve, in order to set up the strategy the system must follow.
- The set and structure of the sentences our system must generate, in order to carry out the system's natural language generation.

1.1 Analysis

In the A-corpus we have found 809 interactions. An interaction means a turn in the dialogue in which either the client or the assistant speaks. Within any turn the speaker may utter one or more sentences. The clients took 374 interactions from the total (46,22%) and the assistants took 435 interactions (53,77%). The number of interactions from the assistants is slightly higher than that of the clients because the assistants usually start and finish the dialogue, and they generally take the initiative in the conversation. We have also timed every dialogue. The timing starts when the client or the assistant begins to speak (generally uttering one greeting) and ends when the assistant gives the client the total price to pay. The time required for preparing the ordered products is also included. The average duration of a dialogue is 1 min 28 sec.

1.2 Labelling

We have used 55 different labels for the A-corpus. Each label is a compound of one speech-act type and the information provided by the user. These labels do not take grammatical considerations into account, merely attending to pragmatic aspects. Thus, the sentences *"Can I have a sandwich?"* and *"A sandwich, please"* are both labelled as an order for a sandwich. The 10 most frequently used labels are:

Label	%
order_product_amount	17,13
order_food_name	9,58
inform_confirmation	8,64
order_drink_name	8,13
inform_price	6,89
greeting	4,72
question_confirmation	4,50
inform_available_food_type	3,92
question_sell_more	3,55
inform_other	3,55

Table 1. 10 most frequently used labels in the A-corpus

Note that the speakers use many understanding confirmations (8,64%). The percentage of *inform_other* is also considerable (3,55%). In this latter label we have included the sentences related to the taking of the ordered products (*"here you are"*, *"you can take napkins here"*, etc.).

1.3 Lexical Classes

We have grouped the keywords in the A-corpus into lexical classes. The table below sets out some of the most important lexical classes our dialogue system uses, as well as some word examples (translated from Spanish into English).

Lexical class	Examples (translated)
amount	*one, two, three, ...*
food_type	*sandwich, salad, ...*
food_name	*cantábrico, ...*
food_complement	*ketchup, mayonnaise, ...*
temperature	*cold, warm, ...*
drink_name	*beer, wine, milkshake, ...*
size	*small, medium, large, ...*
taste	*orange, lemon, ...*
drink_complement	*ice, alcohol, ...*
interrogative	*what, how, ...*
affirmative	*yes, OK, perfect, ...*
negative	*no, not, ...*

Table 2. Some lexical classes the dialogue system uses

1.4 Semantic Structures

By analysing the A-corpus we can find out the semantic structures the clients will probably use to interact with the system. For example, for ordering food, the clients usually use one of the following semantic structures:

(1) <amount> <food_name>
(2) <amount> <food_name> <food_type>
(3) <food_name>
(4) <food_name> <food_type>

Semantic structure (1) appeared on 61,32% of the occasions when clients ordered food (for example, *"one cantábrico"*); type (2) appeared on 20,75% of occasions (for example, *"one cantábrico sandwich");* type (3) appeared on 16,98% of occasions (for example, *"cantábrico"*); and the type (4) appeared on 0,94% of occasions (for example, *"cantábrico sandwich"*). As can be seen, the clients most usually say just the amount and the food name. They normally omit the food type when it can be deduced from the food name (for example, *"cantábrico"* implies *"sandwich"*).

2. The Strategy for the System

We identified the goals of the clients and the assistants from the analysis of the A-corpus and from these we designed the strategy the dialogue system should follow (Denecke & Waibel, 1997). The general goals of the system are:

- Attempt to sell food
- Attempt to sell a drink
- Request for client's phone number
- Request for client's address
- Attempt to sell more restaurant products
- Confirm client's phone number
- Confirm client's address
- Confirm ordered products
- Confirm price to pay
- Confirm transportation time

More sub-goals can be generated from the interaction with the clients if they provide only partial information.

2.1 System's Natural Language Generation

From an analysis of the A-corpus we obtained the set of sentences our system must generate. The five most frequently used sentence types in the corpus are: *assistant's information about the price* (20,92%), *assistant's request for confirmation* (11,45%), *assistant's confirmation* (11,01%), *assistant's question about anything else to order* (10,79%), and *assistant's greeting* (8,37%). We use 40 sentence patterns for the generation of the sentences. The patterns generally consist of several concepts, expressions and vacant gaps. During the generation of the sentences, the system expands the concepts and expressions, and fills the gaps with the appropriate words.

2.2 Some other Features of the System

The initiative strategy for the dialogue is mixed (Giovanni & Zue, 1997; Larsen 1997). The system tends to drive the dialogue in order to achieve its goals and sub-goals. The users may, however, take the initiative whenever they choose and can say whatever they want at any time, so we have set up a mechanism to handle focus shifting. As yet our model system only processes written speech, but we are currently working on a module for speech recognition. The user can correct the system's misunderstandings or non-understandings explicitly, and the system can correct some (simulated) word recognition errors by means of implicit recovery (Danieli & Gerbino, 1995). The vocabulary size is about 500 words. The sentences are analysed by means of a robust bottom-up parsing. We use a semantic grammar for the analysis. The system deals with anaphors, ellipsis, ambiguity, and tautology. The interactions of the users are represented as frames.

2.3 Word Recognition Simulation

In our experiments with the model system we did not use a real word recogniser, using a simulator instead, which can include, change or remove words in the sentences uttered by the users, depending upon nine parameters which configure its reaction. A *noise level* (n_r) parameter represents the negative effect upon the user's voice signal of extraneous noise. We used the value n_r=0.25 as a maximum in our experiments, indicating that background noise was distorting 25% of the speaker's

signal. Four parameters decide how many words uttered by the user are made unrecognisable because of background noise. The system then processes sentences containing words that might have been inserted, changed or removed. Three parameters are used to calculate the confidence value associated to every word *w* in a sentence, *conf(w)*. The system uses expectations about what the user will probably say in his/her interaction. Finally, a *confidence threshold* (u_c) parameter decides whether every word *w* in a sentence is considered as having been correctly recognised. This is the case when $conf(w) \geq u_c$.

3. Experimental Results

We have evaluated our system by means of both objective and subjective methods (Billi *et al*., 1997; Churcher *et al*., 1997). To do this we used a corpus (called *B-corpus*) of 100 dialogues obtained from conversations between 100 test clients and our system. We designed 10 types of *scenario* to obtain the conversations. Each client selected one of them at random and tried to achieve the goals set for that scenario. None of the clients was familiar with the system beforehand. The B-corpus was divided into five groups of 20 dialogues each. We used a different value for the confidence threshold u_c for every group (0.0, 0.6, 0.7, 0.8 and 0.9), and a fixed value for the other parameters.

3.1 Subjective Evaluation Results

The subjective evaluation results were obtained from the test clients' opinions about the quality of the following features of the system: sentence understanding (SU), error recovery (ER), natural language generation (NLG), naturalness (NAT), transaction success (TS), task completion (TC), and speed (SP). In addition, we allotted a score for overall user satisfaction (SAT). The clients could rank the quality of every feature of the system as 1, 2, 3, 4 or 5 ("Very bad", "Bad", "OK", "Good", or "Very good", respectively). The following table sets out the most frequently expressed opinions for every feature of the system tested.

	u_c=0.0	u_c=0.6	u_c=0.7	u_c=0.8	u_c=0.9
SU	5	4	4	3	1
ER	5	4	3	4	1
NLG	5	4	5	5	4
NAT	4	5	4	4	4
TS	5	5	5	5	1
TC	5	5	5	5	1
SP	5	4	4	4	4
SAT	4	4	4	2	1
Total:	38	35	34	32	16

Table 3. Subjective evaluation results for the dialogue system

It can be seen in this table that the system's performance decreases as u_c increases. The system reaches its worst performance when u_c=0.9, because too many words are considered to have been incorrectly recognised.

3.2 Objective Evaluation Results

The objective evaluation was made in the laboratory by analysing the B-corpus manually. This corpus consists of 1194 user turns, and 1288 system turns. We selected 100 user turns from it at random in order to get the scores on simulated word recognition (WR), simulated keyword recognition (KWR), and implicit recovery (IR). By analysing the whole B-corpus we were able to measure objectively the behaviour of the system in terms of the following features: sentence recognition (SR), sentence understanding (SU), turn correction ratio (TCR), contextual appropriateness (CA) and transaction success (TS). Each client was told that he could abandon the conversation if the behaviour of the system became unacceptable to him/her. The table below shows the results thus obtained.

	u_c=0.0	u_c=0.6	u_c=0.7	u_c=0.8	u_c=0.9
WR	100	90.47	90.24	79.59	33.33
KWR	100	90.99	91.63	77.17	41.15
IR	-	46.66	46.87	37.5	18.98
SR	100	70.0	68.0	52.0	21.0
SU	85.71	82.25	69.08	55.97	24.75
TCR	4.27	10.46	15.79	26.45	56.5
AC	85	79.77	73.72	54.8	37.87
TS	84.1	56.41	76.92	31.42	13.63
Aband.	0	0	0	0.25	0.75

Table 4. Objective evaluation results for the dialogue system (%)

It can be seen in this table that the system's performance decreases significantly when $u_c \geq 0.8$ and performs worst when u_c=0.9. It is noteworthy that no user abandoned the conversation with the system when $u_c < 0.8$, and that 75% of them abandoned the dialogue when u_c=0.9. From the results of our subjective and objective evaluations we can conclude that 0.7 is the highest value for u_c if the system is to be judged as performing acceptably.

4. Future Work

We are currently designing a module for speech recognition. A voice activity detector trained in a discriminative manner distinguishes between voice and background noise. The speech signal is pre-emphasised and segmented into frames. Every frame is analysed and represented by a vector which includes 14 Mel frequency Cepstral coefficients, the energy, and the first and second derivatives. We are using context-independent phone-like units modeled by SCHMM (Semi Continuous Hidden Markov Models). The language is modeled by a bigram.

At the moment, the system stores little knowledge concerning what it has said before. We intend to enhance natural language generation by taking into account the previous system's outputs to avoid some sentence repetitions. The system uses an implicit strategy for the confirmations, which is appropriate when few recognition errors occur. As a future development to the system we are considering incorporating a mixed confirmation strategy, thus rendering the confirmation strategy implicit by default, whilst being automatically changeable to explicit whenever the system detects that many recognition errors are occurring.

5. Conclusions

We have described here the kind of information we have obtained from the analysis of a dialogue corpus. This information allowed us to identify the set of words our system must recognise and understand, the semantic structures the restaurant clients usually use, the goals that the clients and the assistants try to achieve, and the set of sentences our system must generate. We have mentioned the strategies used to handle the initiative and the confirmations, together with other features of the system. We have also described how we have gone about simulating word recognition and how we are currently setting up a module for real word recognition. Finally, we have set out our experimental results, obtained by both objective and subjective methods.

6. References

Billi *et al*. (1997). Roberto Billi, Giuseppe Castagneri, Morena Danielli. "Field trial evaluations of two different information inquiry systems". Speech Communication 97.

Churcher *et al*. (1997). Gavin E. Churcher, Eric S. Atwell, Clive Souter. "Generic template to evaluate integrated components in spoken dialogue systems". EUROSPEECH '97.

Danieli & Gerbino (1995). Morena Danieli, Elisabetta Gerbino. "Metrics for evaluating dialogue strategies in a spoken language system". Proceedings of the AAAI Spring Symposium on Empirical Methods in Discourse Interpretation and Generation.

Denecke & Waibel (1997). Matthias Denecke, Alex Waibel. "Dialogue strategies guiding users to their communicative goals". EUROSPEECH '97.

Giovanni & Zue (1997). Giovanni Flammia and Victor Zue. "Learning the structure of mixed initiative dialogues using a corpus of annotated conversations". EUROSPEECH '97.

Larsen (1997). Lars Bo Larsen. "A strategy for mixed-initiative dialogue control". EUROSPEECH '97.

López-Cózar *et al.* (1997). "A Voice Activated Dialog System for Fast-Food Restaurant Applications". Ramón López-Cózar, Pedro García, J. Díaz, Antonio J. Rubio. EUROSPEECH '97.

López-Cózar *et al.* (1997). "A Knowledge Representation Model for a Voiced Dialogue System". R. López-Cózar, A. J. Rubio, P. García, J. Díaz. Speech and Computer (SPECOM '97).

López-Cózar & Rubio (1997). "Una Introducción al Mecanismo de Generación de Lenguaje Natural utilizado por el Sistema SAPLEN". Ramón López-Cózar Delgado, Antonio J. Rubio Ayuso. Magazine no. 21. Spanish Society for Natural Language Processing.

López-Cózar & Rubio (1997). "SAPLEN: Un Sistema de Diálogo en Lenguaje Natural para una Aplicación Comercial". Ramón López-Cózar Delgado, Antonio J. Rubio Ayuso. Proc. III Jornadas de Informática '97. Spanish Association for Informatics and Automatics (AEIA).