The Multi-site 2009 EVALITA Spoken Dialog System Evaluation

Paolo Baggia[1], Francesco Cutugno[2], Morena Danieli[1], Roberto Pieraccini[3], Silvia Quarteroni[4], Giuseppe Riccardi[4] and Pierluigi Roberti[4]

[1]Loquendo SpA Torino
[2] Department of Physics – NLP Group - University of Napoli Federico II
[3]SpeechCycle Labs, New York, USA
[4]Department of Information Engineering & Computer Science (DISI) - University of Trento
{morena.danieli,paolo.baggia}loquendo.com; cutugno@na.infn.it;
roberto@speechcycle.com; silvia.quarteroni@gmail.com; riccardi@disi.unitn.it;
pierluigi.roberti@unitn.it

Abstract. This document presents the coordination and the evaluation procedures for the Spoken Dialogue System Task in EVALITA 2009. Three institutions participated into the competition, University of Trento, University of Naples and Loquendo. EVALITA participants were asked to develop a SDS application operating in the sales force domain, they were provided with a preliminary list of scenarios indicating system accounting modalities and a possible list of subtasks that should made possible. The three systems were hosted on a server at Trento, 19 volunteers called all of them. The calls have been recorded, transcribed and annotated. The evaluation work, based on scripting run on the annotations, has been mainly focused on assessing performance at the dialogue, task, and concept levels. Detailed results indicating the systems performances are reported in the paper. This document presents the coordination and the evaluation procedures for the Spoken Dialogue System Task in EVALITA 2009. Three institutions participated into the competition, University of Trento, University of Naples and Loquendo SpA.

Keywords: Spoken Dialogue Systems, Evaluation, Mixed Initiative.

1 Introduction

Research and development in the ICT area of Interactive Voice Responders (IVR), in Italy, both in academic and industrial environments, sometimes present a fragmented scenario that is typical of other emerging research fields. Notwithstanding this situation, Spoken Dialogue Systems (SDSs) are encountering a greater and greater interest and, on the industrial point of view, they are growing in number and quality thanks to an increasing number of small private companies acting in the tertiary industrial sector. These companies base their products on third party software as far as (but not only) ASR and TTS layers are concerned, building applications for their customers and avoiding, to some extent, to invest in research. With the exception of

Loquendo, and of one or two really smaller companies resulting from a spin-off from the ex IRST (presently FBK), speech engines and VXML interpreters are bought abroad, from companies that invested on Italian (and on other non-English languages) during the '90s.

Concerning academic efforts, with the exception of the Trento University, where a significant number of scientists and advanced students are involved in research on SDS, an extremely limited number of other groups contribute, with a minimal amount of participants, to the research in this field.

To promote and engage the NLP and speech community in the Spoken Dialog research field, we decided to launch the Spoken Dialog System track at EVALITA 2009. This is happening in parallel with similar initiatives in English and other languages such as the "Let's Go" challenge promoted in the United States [1]. The organisers also promised a wide dissemination of the obtained results. Participation to the campaign has not been encouraging as only three institutions took part in the competition, but with the help of the Italian Speech Sciences Association (AISV) the dissemination promise will be maintained.

2 Mixed Initiative SDS

As it is well known, SDS design can be approached in many different ways: in an ideal continuum ranging from the 'directed-dialogues' (also known as "system initiative dialogues"), that probably guarantees higher performances while limiting the designer's fantasy and dialogue naturalness, until to the 'user-driven' approach, in which the system responds to any request without imposing any constrain to the user, almost far from being feasible. The so-called 'mixed-initiative' dialogue systems are situated just in the middle, and constitute a good compromise between the rich potentialities of a user-initiative system and the system-initiative which is low-profile but less error-prone.

For these reasons the organisers of the SDS task choose the implementation of a mixed-initiative application as a challenge platform for EVALITA.
Mixed-initiative dialogs are becoming more and more frequent in many recently developed ICT products.

3 Preliminary guidelines

EVALITA participants were asked to develop a SDS application operating in the sales force domain. The system was thought to serve salespersons calling their company and (1) asking for data about customers or reviewing open orders/invoices, (2) requiring the opening of a new order of one or more positions, each of them including a product and a quantity; (3) searching the company catalog to find products and pricing and optionally discounts. Each call made to the SDS has been targeted to the completion of one or more of the three specific task listed above. A task has been considered successful if the input procedure ended with the correct recording of the

provided information into the database and if the output process delivered the expected responses.

Participant have been provided with a preliminary list of scenarios indicating system accounting modalities and a possible list of subtasks that should made possible in all competitor applications. It was decided that no application would have modified the database by dynamically adding new words (i.e. new clients or new salesmen) to the application vocabulary (even if it is in principle possible in some cases) and, consequently to the VXML grammars.

4 Participants, hosting and calling sessions

Three applications took part into the competition:

1. Loquendo (by Paolo Baggia and Enrico Giraudo)
2. UniTn (by Stefan Rigo, Evgeny A. Stepanov, Pierluigi Roberti, Silvia Quarteroni, and Giuseppe Riccardi)
3. UniNa (by Gianluca Mignini and Francesco Cutugno)

All the three systems were based on the VoxNauta ™ platform. It was then decided to host them on the same server at University of Trento, and to organise all the calling sessions using analogical telephone lines and normal phones. Five volunteers from each participant site made 4 calls to the two other systems (i.e. nobody called its own system), the server randomly selected an even number of calls to each system.

5 Evaluation

The EVALITA SDS campaign was characterized by two main features. First, the participant systems shared three main components:

a) The data, i.e. the SALES database;

b) The domain model, i.e. concepts relating to the sales domain. Concepts are further specified by attributes (e.g. product is specified by its price), as illustrated in Table 1.

c) The task model, i.e. the tasks the systems needed to address, such as listing customers and placing product orders. The full list of tasks is reported in Table 2.

Table 1. EVALITA SDS Concepts and attributes

EVALITA SDS Concepts	
Concept	Attributes
Customer	Id, name, surname, address, shop_name, city
Product	Id, description, amount, brand, category, price, discount, discount_amount
Order	Id
Salesperson	Id, name, surname

The second key feature of the EVALITA SDS track was the presence of a central, Web-based repository maintained by a team of the Department of Computer Science of the University of Trento (UniTN, henceforth). All the dialogues were stored, transcribed, annotated, and evaluated thanks to a set of tools running on the UniTN repository.

The above listed two aspects made it possible to conduct a head-to-head evaluation of the three participant systems on the grounds of a common set of scenarios and evaluation metrics.

Table 2. Tasks defined in the EVALITA SDS domain

EVALITA SDS Tasks	
Name	Description
Identify representative	Verify representative ID
Ask customer detail	Obtain a given customer's address, shop name, etc.
List orders	List orders currently active by representative
Show last order	Show last order placed by representative
List customers	List customers assigned to representative
New order	Place a new order for a product
List products by category	
List products by brand	
List products – other	List products in general or according to other criteria than category or brand, e.g. discounts
Search single product	Obtain information about a specific product
Ask for help	
Exit application	

The evaluation setup is discussed in Section 6, while Section 7 reports the results obtained by the three participants, and Section 8 draws a conclusion on the present year task.

6 Evaluation Setup

Three groups took part in the EVALITA SDS competition, i.e. Loquendo, UniTN, and University of Naples (UniNA, henceforth). The EVALITA SDS evaluation was organized as follows.

Each participant site recruited a set of five subjects who were available to place four telephone calls each, the latter being randomly assigned to one of the participants' systems other than the one developed by the callers' site.

Each call aimed at performing one out of ten application scenarios designed specifically for the SDS task (Section 7). Each call was stored in the central repository.

Data stored in the repository was the source used by the UniTN Web tool to visualize dialogs. Moreover, each participant group had an account on the Web tool that allowed them to transcribe and annotate their own dialogs, with the aim of minimizing transcription errors.

Based on transcription, annotation, and on a number of events registered by the repository platform (call start/end, hang-ups, etc.), a number of metrics were applied to evaluate each participant's dialogs with respect to several objective metrics.

6.1 Transcription and Annotation

Each dialogue turn was transcribed by one of three Italian mother tongue scribes, one per participant institution. Scribes followed the guidelines at:
http://cicerone.dit.unitn.it/DialogStatistics/Transcription/indexTra.php.

Once transcribed, each dialogue turn was annotated in terms of:

- Tasks REQUESTED and COMPLETED during the interaction,
- CONCEPTS and VALUES mentioned in the utterance.

An initial annotation was carried out separately by the three participants; however, a number of inconsistencies in task and concept annotation was detected. This made a second annotation necessary in order to comparatively evaluate the three systems.

6.2 Evaluation Metrics

A set of objective metrics was designed following previous work on SDS evaluation [2], [3] [4]. The evaluation work has been mainly focused on assessing performance at the dialogue, task, and concept levels.

In particular, our overall dialogue-level metrics were the mean and standard deviation of the dialogue length, expressed both in time-to-completion (measured in seconds), and in number of turns, where one "turn" is a couple consisting in a system utterance and a user one.

At the task level, we measured the average and standard deviation in the number of turns required to complete a task, as well as task success rate.

The success rate of a task t_i in a collection of dialogues C, or $tsr_C(t_i)$, is defined as the ratio between the number of correct completions of t_i, named $corr_C(t_i)$, and the number of requests of t_i found in C, named $req_C(t_i)$: $tsr_C(t_i) = corr_C(t_i)/req_C(t_i)$. Note that "correct completion" means that not only there was no misrecognition in the type of task to perform, but also that such a task was performed with the correct parameters (e.g. amount, brand and customer of a given product in the *New order* task).

At the concept level, we measured precision and recall[1].

6.3 Evaluation Scenarios

Ten scenarios were developed to evaluate each participant SDS. Each scenario represents a typical interaction with the system, as illustrated in Table 3.

[1] Given A, the set of concepts annotated by the annotator and B, the set of concepts understood by the SDS, precision is defined as: $P = (A \cap B)/B$, while recall is defined as: $(A \cap B)/A$.

Table 3. One of the 10 evaluation scenarios for EVALITA SDS

Scenario 1
1. Identificarsi come Fabrizio Villa (n. id. 1)
2. Richiedere la lista degli ordini aperti di Mario Bianchi.
3. Sapere l'eventuale sconto per un Product della categoria pasta.
4. Inserire l'Order per Mario Bianchi di 50 carote della marca Bio.

The recruited subjects could choose 4 out of the 10 scenarios and perform one call per scenario by dialing a dedicated telephone number.

Each call from each group subject was randomly routed with equal probability to one of the other two groups' SDS. This allowed the fairest possible setting for the evaluation.

In addition, to the "cross-evaluation" performed by the three groups, external subjects submitted calls that they were routed with equal probability to one of the three participant SDSs.

The evaluation took place in October 2009 and lasted five days: one day per participant group followed by two days dedicated to external callers.

Section 4 summarizes and compares the results obtained by the different participants.

7 Results

Out of the 134 calls collected in total, we selected a working subset of 20 calls per system by initially discarding extremely short dialogs (i.e. lasting less than 30 seconds), and then by randomly discarding part of the remaining ones. This step was carried out to remove trial calls from our analysis and to even out the fact that the random routing of dialogues assigned a different number of calls to each application.

Table 4 reports the general figures relating to the different participant systems. We note that, while Loquendo and UniTN recorded a similar number of turns, interactions with the UniNA application were sensibly shorter. A closer look into the dialogues shows that calls routed to the UniNA system generally concerned tasks requiring a lower number of turns, e.g. *Ask customer detail* instead of *New order* (see also Table 5).

Moreover, the UniTN application tended to ask for explicit confirmation from the user more frequently (particularly in the first task, *Identify representative*); this results in an average of 24.4 turns in UniTN dialogues against 18.9 in Loquendo.

Finally, different caller behaviors could be observed: in some cases, when tasks were not successfully performed, some callers tended to re-try them, while others moved on to the following task.

Table 4. Dialog level statistics

Participant	**Duration (sec)**	**Duration (# Turns)**
UniNA	145.8±72.7	11.0±5.7
Loquendo	182.2±84.7	18.9±8.9
UniTN	206.4±81.7	24.4±10.1

Table 5 reports the success rates for the different tasks defined in the EVALITA SDS domain, as well as the time taken to complete them.

Table 5. Task durations (#turns: mean±std.dev.) and success rates

Task	UniNA		Loquendo		UniTN	
	Duration (turns)	**Tsr (corr/req)**	**Duration (turns)**	**Tsr (corr/req)**	**Duration (turns)**	**Tsr (corr/req)**
Identify representative	1.9 ± 0.4	100.0% (19/19)	2.4 ± 0.8	95.0% (19/20)	3.1 ± 0.5	90.5% (19/21)
Ask customer detail	2.0 ± 0.0	83.3% (5/6)	2.3 ± 0.5	88.9% (8/9)	3.4 ± 1.6	54.6% (12/22)
List orders	2.5 ± 1.5	0.0% (0/8)	2.0 ± 0.0	80.0% (4/5)	3.0 ± 0.0	75.0% (3/4)
Show last order	2.0 ±0.0	100% (1/1)	-	-	-	-
List customers	2.0 ± 0.0	50.0% (2/4)	2.0 ± 0.0	0.0% (0/8)	3.0 ± 0.0	66.7% (2/3)
New order	4.6 ± 1.5	36.4% (4/11)	4.3 ± 1.8	42.9% (9/21)	7.5 ± 2.8	63.2% (12/19)
List products by category	3.0 ± 1.0	14.3% (1/7)	-	-	3.0 ± 0.0	100.0% (3/3)
List products by brand	-	-	-	-	3.0 ± 0.0	50.0% (1/2)
List products – other	2.0 ± 0.0	0.0% (0/4)	3.0 ± 0.8	25.0% (2/8)	3.8 ± 1.6	44.4% (4/9)
Search single product	2.3 ± 0.4	55.6% (5/9)	2.8 ± 1.6	77.8% (14/18)	3.5 ± 2.5	78.6% (11/14)
Ask for help	2.0 ± 0.0	100% (3/3)	-	-	2.0 ± 0.0	100.0% (2/2)
Exit application	2.5 ± 0.5	100.0% (5/5)	-	0.0% (0/1)	2.4 ± 0.8	25.0% (4/16)
Overall (corr/req)	**-**	**58.4% (45/77)**	**-**	**62.2% (56/90)**	**-**	**63.5% (73/115)**

Our first observation is that for the tasks where a sufficient number of cases can be examined (i.e. the number of requests of such a task is sufficiently large), the three systems exhibit similar task success rates.

Moreover, the global task success rate, computed as the number of successful completions divided by the number of requests for all tasks, reveals very close values, i.e. around 60%. The UniTN system tends to take longer to perform tasks, however this results in a more successful task completion record. A possible factor penalizing the UniNA system (reaching the lowest success rate of 58.4%) is the fact that the latter did not support the *List product by brand* task[2], which is was requested 7 times and was annotated as correctly addressed only once. Moreover, it seemed to have troubles with the *List orders* task. In contrast, UniNA was the system better

[2] This task was not originally present in the delivered guidelines.

supporting *Identify representative* and *Exit application*, both of which reached 100% success.

In any case, due to the small amount of dialogs examined, the difference in task success across the three applications cannot be significantly judged.
Unfortunately, some discrepancies were found between each participant system's internal concept specification, and the concept specification in the annotation protocol. Although we are currently solving this issue, the latter made it impossible to disclose concept precision and recall in the current report due to time constraints.

7 Conclusions

This report illustrates the spoken dialogue system evaluation task carried on within the 2009 EVALITA campaign. Due to the novelty and well-known difficulties related to this kind of natural language evaluation task, a great deal of work has been devoted to design an effective and fair evaluation methodology. That work included both the identification of the different aspects of the experimental set up (common database, common tasks, scenarios, subjects' recruitment, etc…) and the identification of a set of dialogue metrics to be used in the evaluation of the results. While defining the evaluation setting was almost uncontroversial, as we partly expected, we faced difficulties related with the dialogue annotation task that was characterized by inconsistencies among the different annotators. This in turn required a re-annotation of all the collected dialogues. We believe that the inconsistencies are mainly due to the different backgrounds, and perhaps degree of direction received by the annotators.

Despite these difficulties, it was interesting to notice that, when developed on a common task model, the participant systems reported quite similar task success rate, where the differences are more related to different interaction styles performed by the dialogue agent. We suppose that these more subtle differences may have an impact on the quality of the interaction, as it is perceived by the users of the dialogue application, and that more subject-oriented evaluation could be taken into account for next evaluation dialogue campaigns.

References

1. Black, A., Eskenazi, M.: The Spoken Dialog Challenge. In: Proceedings of SigDial (2009)
2. Danieli, M., Gerbino, E.: Metrics for evaluating dialogue strategies in a spoken language system. In: Working Notes of the AAAI Spring Symposium on Empirical Methods in Discourse Interpretation and Generation (1995)
3. Levin, E., Narayanan, S., Pieraccini, R., Biatov, K., Bocchieri, E., Di Fabbrizio, G., Eckert, W., Lee, S., Pokrovsky, A., Rahim, M., Ruscitti, P., Walker, M.: The AT&T-DARPA Communicator Mixed-Initiative Spoken Dialog System. In: Proceedings of ICSLP (2000)
4. Varges, S., Riccardi, G., Quarteroni, S.: Persistent Information State in a Data-Centric Architecture. In: Proceedings of SigDial (2008)