An End-to-End Evaluation of Two Situated Dialog Systems.

Lina Maria Rojas Barahona, Alejandra Lorenzo, Claire Gardent

An End-to-End Evaluation of Two Situated Dialog Systems

Lina M. Rojas-Barahona
Inria, LORIA, UMR 7503
Villers-lès-Nancy
F-54600, France
lina.rojas@loria.fr

Alejandra Lorenzo
Université de Lorraine
LORIA, UMR 7503
Vandoeuvre-lès-Nancy
F-54500, France
alejandra.lorenzo@loria.fr

Claire Gardent
CNRS, LORIA, UMR 7503
Vandoeuvre-lès-Nancy
F-54500, France
claire.gardent@loria.fr

Abstract

We present and evaluate two state-of-the art dialogue systems developed to support dialog with French speaking virtual characters in the context of a serious game: one hybrid statistical/symbolic and one purely statistical. We conducted a quantitative evaluation where we compare the accuracy of the interpreter and of the dialog manager used by each system; a user based evaluation based on 22 subjects using both the statistical and the hybrid system; and a corpus based evaluation where we examine such criteria as dialog coherence, dialog success, interpretation and generation errors in the corpus of Human-System interactions collected during the user-based evaluation. We show that although the statistical approach is slightly more robust, the hybrid strategy seems to be better at guiding the player through the game.

1 Introduction

In recent years, there has been much research on creating situated conversational characters i.e., virtual characters (VCs) that look and act like humans but inhabit a virtual environment (Gratch et al., 2002; Hofs et al., 2010; Traum et al., 2007; Johnson et al., 2005; Traum et al., 2008; DeVault et al., 2011).

In this paper, we focus on French speaking, situated conversational agents who interact with virtual characters in the context of a serious game designed to promote careers in the plastic industry (The Mission Plastechnologie game or MP). We present and compare two state-of-the art dialogue systems. The first system (H) is a hybrid approach that combines an information-state dialogue manager (Larsson and Traum, 2000) with a classifier for interpreting the players' phrases. The second system (QA) is a question/answering character model which predicts the system dialog move given a player's utterance (Leuski and Traum, 2008). Both systems use a generation-by-selection strategy (Leuski et al., 2006; Gandhe and Traum, 2007) where the system's utterances are selected from a corpus of possible outputs based on the dialog manager output. While previous work focuses on relatively short dialogs in a static setting, in our systems we consider long interactions in which dialogs occur in a setting that dynamically evolves as the game unfolds.

We evaluate the two dialog systems in the context of the 3D game they were developed for and seek to determine the degree to which a dialog system is operational. To answer this question, we analyse both systems with respect not only to quantitative metrics such as accuracy but also to user- and corpus-based metrics. User-based metrics are computed based on a questionnaire the users filled in; while corpus-based metrics are manually extracted from the corpus of Player-VC interactions collected during the user-based evaluation. As suggested by evaluation frameworks such as PARADISE (Walker et al., 1997) and SASSI (Hone and Graham, 2000), we show that a multiview evaluation permits a better assessment of how well the dialog system functions "in the real world". The metrics proposed assess dialog success and coherence, as well the costs of dialog components.

The paper is organized as follows. In Section 2,

we present the MP game, the dialogue strategies used in the different dialogs and the dialog data used for training. Section 3 presents the two dialog systems we compare. Section 4 presents the evaluation schemes used to compare these two systems and discusses the results obtained. Section 5 concludes with directions for further research.

2 Dialogues in the MP Game

We begin by describing the MP game, the dialogs in the MP game, the strategies used to guide the hybrid dialog manager and the data used for training.

2.1 The MP Game and Dialogs

The MP game is a multi-player quest where 3 teenagers seek to build a joystick in order to free their uncle trapped in a video game [1]. To build this joystick, the player (who alternatively represents anyone of these three teenagers) must explore the plastic factory and achieve 17 mandatory goals (*find the plans, get the appropriate mould, retrieve some plastic from the storing shed*, etc), as well as 11 optional goals which, when reached, provide them with extra information about the plastic industry (and therefore increases their knowledge of it).

In total, the player can achieve up to 28 game goals by conducting 12 separate dialogs in various parts of the virtual world. Each of the 12 dialogs in the MP game helps players to achieve the game goals. The player interacts with the virtual characters to obtain information that helps her to achieve these goals and, as a consequence, to increase her score in the game. Table 1 summarises the game goals and the contextual parameters (player's role, location in the virtual world, VCs present) associated with each dialog.

2.2 Dialog Data and Annotation

To train both classifiers, the one used by the hybrid and the one used by the QA system, we collected Human-Machine dialog data using a Wizard-of-Oz setting and manually annotated each turn with a dialog move. The resulting corpus (called Emospeech Corpus) and the annotation scheme (as well as the inter-annotator agreement) used are described in detail (Rojas-Barahona et al., 2012). Briefly, the Emospeech Corpus comprises 1249 dialogs, 10454 utterances and 168509 words. It contains 3609 player utterances consisting of 31613 word tokens and 2969 word types, with approximately 100 conversations for each dialog in the game. Turns were annotated with dialog moves (Traum and Larsson, 2003) capturing both domain knowledge (e.g., about the goals set by the game) and the set of core communicative acts.

[1]The MP game was created by Artefacto, `http://www.artefacto.fr/index_ok.htm`

2.3 Dialog Strategies

We identified four main dialog strategies underlying the 12 MP dialogs and used these to define the plans guiding the rule-based discourse management in the hybrid system. These strategies can be seen as transactions made up of conversational games (Carletta et al., 1997).

Strategy 1. This strategy is used in the first dialog only and consists of a single *Address Request* move by the VC followed by the player's answer: Lucas requests Ben to find the address of the Plastic Enterprise that must be hidden somewhere in the lab. Ben can accept, reject or ask for help. Lucas answers accordingly and ends the conversation.

Strategy 2. Nine dialogues follow this strategy. They include several (up to 5) requests for information and the corresponding system/player's exchange. Appendix A shows an example dialog following this strategy.

Strategy 3: This is a confirmation strategy where the VC first checks that the player has already achieved a given task, before informing her about the next step (e.g. dialogs with Melissa in Table 1).

Strategy 4. This strategy, exemplified in Appendix B, is similar to strategy 2 but additionally includes a negotiation step where the VC asks the player for help.

3 Dialogue Systems

The game and the two dialog systems built were integrated as agents within the Open Agent Architecture as shown in Figure 1. Both systems access a database for starting the appropriate dialogs at the appropriate place in the virtual world while simultaneously storing all interactions in the database.

Id	VC	Player	Goals	Location
1	Lucas	Ben	Find the address of the enterprise.	Uncle's place.
2	M.Jasper	Lucas	The manufacturing first step	Enterprise reception
3	Samir	Julie	Find the plans of the joystick *Optional: job, staff, studies, security policies*	Designing Office
4	Samir	Julie	Find out what to do next *Optional: jobs in the enterprise, staff in the enterprise*	Designing Office
5	Melissa	Lucas	Find the mould, optional where are the moulds	Plant
6	Melissa	Lucas	Find the right machine	Plant
7	Melissa	Lucas	Confirm you have found the right mould and machine and find out what to do next	Plant
8	Operator	Julie	Knowing about the material space and about the job *Optional: find out what to do in the case of failure* *helping to feed a machine with the right material*	Material Space
9	Serge	Ben	Perform quality tests. *Optional: VC's job*	Laboratory Tests
10	Serge	Ben	Find out what to do next. *Optional: know what happens with broken items*	Laboratory Tests
11	Sophia	Julie	Find the electronic components, knowing about VC's job	Finishing
12	Sophia	Lucas	Finishing process *Optional: know about conditioning the product*	Finishing

Table 1: Description of the 12 dialogs in the MP Game.

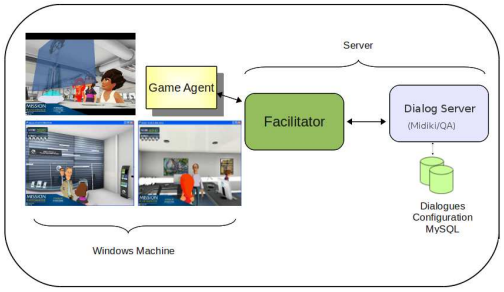

Figure 1: General Architecture for the dialog system: modules are implemented as agents within the Open Agent Architecture.

3.1 The Hybrid Dialogue System

The hybrid system combines an interpreter; a rule based, Information State Approach dialog manager; a generator; and the game/dialog communication components i.e., the OAA interface.

The Interpreter Module In the hybrid system, the interpreter is a classifier trained on the annotated data (cf. section 2.2), which maps the player's utterance to a dialog move. To build the classifier we experimented with both SVM (Support Vector Machine) and LR (Logistic Regression) [2] using different sets of features: utterances were preprocessed by removing stop words and converting content words to unaccented lowercase. Afterwards, we experiment with and without TF*IDF (term frequency*Inverse Document Frequency) filtering and context moves (from 0 to 4 previous dialog moves).

Since the game consist of a number of different dialogs, two options naturally arise: we could either train a single classifier on the whole dataset or train one classifier per dialog. Hence the data sets (and the number of categories to be learned) differ depending on whether we trained one classifier on the whole game data or a classifier for each dialog.

In addition, hand-written rules are used to correct misclassification errors as follows. The best prediction given by the classifier is matched against the expected move determined by the last move stored in the information-state (see below the description of the dialog manager). In case of a mismatch, the interpreter selects a move that is valid in the current context and updates the information state with this move. For instance, if the last move is a yes/no question, *confirm* and *disconfirm* dialog acts are valid moves; for propositional questions, $\texttt{Goal}_i$ is a valid

[2] We used LIBSVM (Chang and Lin, 2001) and MALLET (McCallum, 2002) for the LR classifier with L1 Regularisation.

dialog move [3]; etc. Since the process is non deterministic, this re-interpretation process may improve the system response but it may also be incorrect. For instance, in the following example re-interpretation correctly overrides the dialog move output by the classifier to a move asking the VC (Samir) for the next goal to be achieved.

Samir: Vous avez d'autres questions pour moi?

Do you have other questions ?

Julie: comment cela fonctionne ?

How does it work?

Samir: Eh bien , il va falloir la fabriquer maintenant cette manette . Allez voir Mélissa.

Well, you have to build the joystick now. Go to see Melissa.

In contrast, the dialog below shows a case where re-interpretation fails. Instead of re-interpreting the player's (Julie) input as a request for the next goal, it outputs a request for information about the staff thereby yielding an incoherent exchange.

Samir: D' autres questions ?

Other questions?

Julie: oui qu'est-ce que je peux faire avec ces bouts de papier ?

yes, what can I do with these pieces of paper ?

Samir: Et bien sachez qu'il y a de plus en plus des femmes dans cette industrie ...

you should know there are more and more women in this industry ...

The Dialog Manager We designed a plan for each dialog strategy and extended Midiki (Burke et al., 2003) to support the OAA architecture and access the relational database specifying the configurations of the different dialogs in the game. Each time a new dialog is launched, the information state is loaded with the corresponding dialog-context (e.g., speakers, list of goals to be discussed) and the plan modeling the corresponding dialog strategy. To support dialog management, we implemented a set of update and selection rules for integrating players' moves, handling the information-state and for preparing the agenda according to the plan. More specifically, the following rules are executed at runtime: *Integration*: integrates dialog moves (e.g., questions, answers, acknowledgments) in the information state (questions are listed in the *Question Under Discussion*, answers change the *Commond Ground*, player answers are integrated in response to VCs questions). *Manage Plan*: searches the next action in the plan. *Refill Agenda*: updates the agenda with the next action and *Selection*: selects the next dialog move according to the plan. Once the system move has been selected, the Generator searches an appropriate verbalisation.

The Generator As mentioned above, the generator implements a generation-by-selection strategy. Given the dialog move output by the dialog manager, the generator selects any utterance in this corpus that is labeled with this dialog move and with the identifier of the current dialog.

In addition, two types of dialog moves are given special treatment. The first two moves of each dialog are systematically constrained to be a welcome greeting followed by either a request to pursue a goal (`ask(Goal`$_i$`)`) or a proposal to help (`ask(task(X))`). Furthermore, propositional questions (i.e., proposals by the system to discuss additional topics) were annotated separately with their respective dialog goals. For example, Samir's sentence: *Are you interested in hearing about my job, the people that work here or the security policies?*, was annotated with the goals: *job*, *staff* and *security_policies*. For these dialog acts, the generator checks the list of current missing goals so as to retrieve an appropriate propositional question. In this way, the system can coherently direct the player by suggesting possible topics without using vague and repetitive sentences such as *Would you like to know more?*.

3.2 The QA System

The QA system combines a classifier that matches players' turns to system dialog moves with the same generation-by-selection algorithm used in the hybrid system. Like the hybrid system, the QA dialog system maintains a constant interaction with the game to allow for the game coherence and continuity. In particular, at the end of each dialog, the game goals that have been reached through the dialog are returned to the game server so that the game can proceed and be updated with the appropriate score.

For the QA system, we trained a Logistic Regression (LR) classifier for each dialog in the game.

[3] The system asks the player for the goal to be discussed: $ask(task(X))$ and the player answers one goal in the situated dialog: $Goal_i$.

We applied the same preprocessing and extracted the same features used for training the interpreter in the hybrid system.

3.3 Discussion and Interaction Examples

As shown above, while the QA system simply matches the player's input to a system response, the hybrid system has a much more elaborate dialog management policy including re-interpretation and the use of game and dialog history information, to guide the dialog. As a result, the QA dialogs are generally more spontaneous, giving the player more initiative whereas the hybrid system enforces a more System-Driven dialog strategy thereby guiding the player through the game. The following two examples illustrate this point. Sophia is the system and Lucas the player. The game context provides two goals which can be discussed: one mandatory (To find out that finishing the manufacturing process involves conditioning the product), and one optional (To be told what conditioning the product means). The examples show how the hybrid system proposes both goals to the user while the QA engine provides the same information fluently.

Hybrid system

Sophia: Vous avez fait du bon travail les enfants!

You did a great job kids!

Sophia: Bien. Que puis-je faire pour vous ?

What can I do for you?

Lucas: J'ai fini la manette

I've finished the joystick

Sophia: Normalement j'ai encore quelques étapes à réaliser avant l'envoi au client. Vous voulez en savoir plus ?

Normally I still have to condition the product. Do you want to know more about conditioning a product?

Lucas: Oui

Yes, ok

Sophia: Conditionner le produit veut dire l'emballer soigneusement pour pouvoir l'expédier au client. Aldo, le logisticien assure qu'on ne manque pas de stock et de gérer les livraisons et les expéditions.

Conditioning the product means packing it carefully in order to ship it to the customer. Aldo,the logistics coordinator ensures there is no shortage of stock and he is also in charge of the deliveries and shipments.

Sophia: Je vous laisse tester votre manette . Au revoir

I leave you test out your joystick. Goodbye

Question Answer

Sophia: hé bien c'est du bon travail tout ça.

Good! That's a great job kids!

Sophia: Alors , où en sommes nous?

So where are we?

Lucas:voilà nous avons utilisé les composants. Quelle est l'étape suivante?

We used the components, what is the next step?

Sophia: On finis la fabrication par conditionner le produit.

We finish the fabrication by conditioning the product.

Lucas:Ok

Sophia: Conditioner le produit veut dire l'emballer soigneusement pour pouvoir l'expédier au client . Le logisticien assure qu'on ne manque pas de stock et de gérer les livraisons et les expéditions.

Conditioning the product means packing it carefully in order to ship it to the customer. Aldo,the logistics coordinator ensures there is no shortage of stock and he is also in charge of the deliveries and shipments.

Sophia: Et voila, je vous dis au revoir

Ok, I say good bye!

4 Evaluation

In evaluating the two systems, we seek to compare their usability: Which system is best suited for use by real users in the context of the MP serious game? We also seek to better understand which module causes which errors and why. To address these questions we conducted a quantitative evaluation where we compare the accuracy of the interpreter and the dialog manager integrated in each system; a user based evaluation involving 22 subjects using both the QA and the hybrid system; and a corpus based evaluation where we examine such criteria as dialog coherence, dialog success, interpretation and generation errors in the corpus of Human-System interactions collected during the user-based evaluation.

4.1 Quantitative Evaluation

We begin by evaluating the accuracy of the interpreter and the dialog manager used by the hybrid and the QA system respectively. These two classifiers were trained on the Emospeech corpus mentioned above and evaluated with 30-fold cross-validation.

Hybrid System As we mentioned in section 3.1, since the game includes different dialogs, a natural question arise: whether to implement the inter-

preter with a single classifier for the whole dataset, or using a different classifier for each dialog in the game. To answer this question, we compared the accuracy reached in each case. The details of these experiments are described in (Rojas-Barahona et al., 2012). The highest accuracy is reported when using a single classifier for the *whole game*, reaching an accuracy of 90.26%, as opposed to 88.22% in average *for each dialog*. In both cases, the classifier used is LR, with L1 regularisation and applying the tf*idf filtering. However, although the classifier trained on the whole dialog data has better accuracy (learning a model per dialog often run into the sparse data issue), we observed that, in practice, it often predicted interpretations that were unrelated to the current dialog thereby introducing incoherent responses in dialogs. For instance, in the dialog below, the player wants to know how waste is managed in the factory. The best prediction given by the interpreter is a goal related to another dialog thereby creating a mismatch with the DM expectations. Re-interpretation then fails producing a system response that informs the player of the next goal to be pursued in the game instead of answering the player's request.

Ben: Comment on gère les dechets ici?

How is the waste managed here ?

Serge: Allez voir Sophia pour qu'elle vous fournisse les composants électroniques nécessaires à votre manette.

Go and see Sophia, she'll give you the electronic components you need for your joystick.

For the user based experiment, we therefore use the LR models with one classifier per dialog.

QA System For evaluating the QA classifier, we also compared results with or without tf*idf filtering. The best results were obtained by the LR classifier *for each dialog* with tf*idf filtering yielding an accuracy of 88.27% as shown in Table 2.

Id	w/o Tf*Idf	w Tf*Idf
1	83.33	82.93
2	93.55	91.8
3	72	80.95
4	80	82.47
5	95.24	93.98
6	97.56	97.5
7	97.5	97.44
8	70.59	76
9	92.77	91.14
10	85.53	86.49
11	83.51	87.5
12	94.12	91.04
Avg.	87.14	88.27

Table 2: Results of the LR classifier for mapping players' utterances to system moves, with content-words and a context of four previous system moves, with and without tf*idf filtering.

4.2 Preliminary User-Based Evaluation

The accuracy of the interpreter and the dialog manager used by the hybrid and the QA system only gives partial information on the usability of the dialog engine in a situated setting. We therefore conducted a user-based evaluation which aims to assess the following points: interpretation quality, overall system quality, dialog clarity, game clarity and timing. We invited 22 subjects to play the game twice, once with one system and once with the other. The experiment is biased however in that the players always used the hybrid system first. This is because in practice, the QA system often fail to provide novice players with enough guidance to play the game. This can be fixed by having the player first use the hybrid system. Interestingly, the game guidance made possible by the Information State approach is effective in guiding players through the game e.g., by proposing new goals to be discussed at an appropriate point in the dialog; and by taking dialog history into account.

After playing, each user completed the questionnaire shown in Table 3. For those criteria such as dialog and game clarity, we do not report the scores since these are clearly impacted by how many times the player has played the game. Table 4 shows the mean of the quantitative scores given by the 22 subjects for interpretation, overall system quality and timing. We computed a significance test between the scores given by the subjects, using the Wilcoxon signed-rank test[4]. As shown in the Table, for all criteria, except Q.4, the QA performs significantly ($p < 0.01$) better than the Hybrid system.

[4]The Wilcoxon signed-rank test is the non-parametric alternative to the paired t-test for correlated samples, applicable, e.g. when dealing with measures which cannot be assumed to have equal-interval scales, as is usual with user questionnaires.

	Interpretation
Q.1	Did you have the feeling the virtual characters understood you? (very bad 1 ... 100 very good)
	Overall System Quality
Q.2	Did you find the conversations coherent? (very bad 1 . . . 100 very good)
Q.3	Did you enjoy talking with the virtual characters? (very annoying 1 ... 100 very enjoyable)
Q.4	Would you prefer playing the game without conversations with virtual characters? (yes/no)
Q.5	What is your overall evaluation of the quality of the conversations? (very bad 1 . . . 100 very good)
	Dialogue clarity
Q.6	How easy was it to understand what you were supposed to ask? (very difficult 1 ... 100 very easy)
Q.7	How clear was the information given by the virtual characters? (totally unclear 1 ... 100 very clear)
Q.8	How effective were the instructions at helping you complete the game? (not effective 1 ... 100 very effective)
	Game clarity
Q.9	How easy was it to understand the game? (totally unclear 1 ... 100 very clear)
	Timing
Q.10	Were the system responses too slow (1) / just at the right time (2) / too fast (3)

Table 3: Questionnaire filled by the subjects that played with both dialog systems.

Interpretation. Question Q.1 aims to captures the user's assessment of the dialog system ability to correctly interpret the player's utterances. The QA system scores 0.7 points higher than the Hybrid system suggesting better question/answer coherence for this system. One possible reason is that while the hybrid system detects any incoherence and either tries to fix it using re-interpretation (which as we saw sometimes yields an incoherent dialog) or make it explicit (using a misunderstanding dialog act i.e., a request for rephrasing), the QA system systematically provides a direct answer to the player's input.

The relatively low scores assigned by the user to the interpretation capabilities of the two systems (57.36 and 64.55 respectively) show that the high accuracy of the interpreter and the dialog manager is not a sufficient criteria for assessing the usability of a dialog system.

Timing. One important factor for the usability of a system is of course real time runtimes. The evaluation shows that overall the speed of the QA system was judged more adequate. Interestingly though the difference between the two systems stems no so much from cases where the hybrid approach is too slow than from cases where it is too fast. These cases are due to the fact that while the QA system always issues one-turn answer, the rule based dialog based approach used in the hybrid system often produce two consecutive turns, one answering the player and the other attempting to guide her towards the following game goal.

In sum, although the QA system seems more robust and better at supporting coherent dialogs, the hybrid system seems to be more effective at guiding the player through the game.

	Question	Hybrid	QA
Interpr.	Q.1	57.36	64.55 (*)
Sys Qual.	Q.2	57.78	60.68 (*)
	Q.3	60.77	66.45 (*)
	Q.4/no	86.37	81.82
	Q.5	59.54	65.68 (*)
	Avg.	66.12	68.66 (*)
Timing	Q.10	2.25	2.05 (*)

Table 4: Mean of the quantitative scores given by 22 individuals. (*) denotes statistical significance at $p < 0.01$ (two-tailed significance level).

4.3 Corpus-Based Evaluation

The User-Based evaluation resulted in the collection of 298 dialogs (690 player and 1813 system turns) with the Hybrid system and 261 dialogs (773 player and 1411 system turns) with the QA system. To better understand the causes of the scores derived from the user-filled questionnaire, we performed manual error analysis on this data focusing on dialog incoherences, dialog success, dialog management and generation errors (reported in Table 5).

DM Errors The count of dialog management (DM) errors is the ratio $\frac{WR}{P}$ of wrong system responses on counts of player's input. In essence this metrics permits comparing the accuracy of the QA dialog manager with that of the hybrid system. On average there is no clear distinction between the two systems.

Generation Errors The system response selected by the generation component might be contextually inappropriate for at least two reasons. First, it may contain information which is unrelated to the current context. Second, it might have been imprecisely or incorrectly annotated. For instance, in the dialog below, the annotation of the turn *Yes, thanks. What do you want me to do?* did not indicate that the turn included a *Confirm* dialog move. Selecting this turn in the absence of a yes/no question resulted in a contextually inappropriate system response.

SYSTEM: Bonjour les petits jeunes je suis le préparateur matiére.

Hello kids, I am the raw material responsible

SYSTEM: Oui merci. Vous me voulez quoi en fait ?

Yes, thanks. What do you want me to do?

PLAYER: je veux en savoir plus sur cet endroit.

I would like to know more about this place

As shown in Table 5, for both systems, there were few generation errors.

Id	%DM H.	%DM. QA	%Gen H. & QA
1	0.0	4.55	0.57
2	10.81	12.00	1.02
3	10.38	12.04	1.49
4	16.22	14.86	0.32
5	10.34	2.13	1.46
6	0.0	0.0	0.94
7	9.52	4.0	0.0
8	11.68	7.08	2.06
9	2.13	26.47	0.76
10	15.63	16.13	6.08
11	11.94	8.33	3.19
12	14.29	8.16	3.17
Avg.	9.41	9.65	1.76

Table 5: DM and generation errors detected in the hybrid and the QA systems.

Unsuccessful Dialogs We counted as unsuccessful those dialogs that were closed before discussing the mandatory goals. The results are shown in Table 6. Overall the QA system is more robust leading to the mandatory goals being discussed in almost all dialogs. One exception was dialog 8, where the system went into a loop due to the player repeating the same sequence of dialog moves. We fixed this by integrating a loop detection step in the QA dialog manager. For the hybrid system, dialog 7, a dialog involving the confirmation strategy (cf. section 2) is the most problematic. In this case, the DM rules used to handle this strategy are inappropriate in that whenever the system fails to identify a contextually appropriate response, it simply says so and quits the dialog. The example illustrates the difficulty of developing a complete and coherent DM rule system.

Id	%Uns. H.	%Inco. H.	%Uns. QA.	%Inc. QA.
1	0	0.0	0.0	0.0
2	0	0.0	0.0	0.0
3	6.67	3.33	7.41	0.0
4	7.14	0.0	0.0	4.0
5	3.85	0.0	0.0	0.0
6	0.0	0.0	0.0	0.0
7	21.21	0.0	0.0	0.0
8	3.70	0.0	15.63	3.13
9	0.0	0.0	0.0	4.35
10	0.0	6.67	0.0	16.67
11	3.45	6.90	0.0	3.70
12	4.17	4.17	4.55	4.55
Avg.	4.89	1.76	4.47	3.03

Table 6: Overall dialog errors, the percentage of unsuccessful dialogs

Incoherent Dialogs We counted as incoherent, dialogs where most system answers were unrelated to the player's input. As shown in Table 6, despite interpretation and generation imprecisions, most dialogs were globally coherent. They made sense according to the game context: they were related to the task to be solved by the player in the game, and the generated instructions were correctly understood. The hybrid system produces slightly less incoherent dialogs probably because of its re-interpretation mechanism which permits correcting contextually invalid dialog moves.

5 Conclusion

We have presented a multi-view evaluation of two system architectures for conversational agents situated in a serious game. Although the QA system seems more robust and is easier to deploy, the hybrid dialog engine seems to fare better in terms of game logic in that it guides the player more effec-

tively through the game. The evaluation shows the importance of assessing not only the dialog engine accuracy but also its usability in the setting it was designed for. In future work, we plan to compute a regression model of user satisfaction for applying reinforcement learning and find the optimal strategy. In addition, we plan to extend the comparison to other domains such as language learning and complex negociation dialogs.

6 Acknowledgments

The research presented in this paper was partially supported by the Eurostar EmoSpeech project and by the European Fund for Regional Development within the framework of the INTERREG IV A Allegro Project.

References

C. Burke, C. Doran, A. Gertner, A. Gregorowicz, L. Harper, J. Korb, and D. Loehr. 2003. Dialogue complexity with portability?: research directions for the information state approach. In *Proceedings of the HLT-NAACL 2003 workshop on Research directions in dialogue processing - Volume 7*.

Jean Carletta, Stephen Isard, Gwyneth Doherty-Sneddon, Amy Isard, Jacqueline C. Kowtko, and Anne H. Anderson. 1997. The reliability of a dialogue structure coding scheme. *Comput. Linguist.*, 23(1):13–31, March.

Chih C. Chang and Chih J. Lin, 2001. *LIBSVM: a library for support vector machines*.

David DeVault, Anton Leuski, and Kenji Sagae. 2011. An evaluation of alternative strategies for implementing dialogue policies using statistical classification and hand-authored rules. In *5th International Joint Conference on Natural Language Processing (IJCNLP 2011)*.

Sudeep Gandhe and David Traum. 2007. Creating spoken dialogue characters from corpora without annotations. In *Proceedings of 8th Conference in the Annual Series of Interspeech Events*, pages 2201–2204.

Jonathan Gratch, Jeff Rickel, Elisabeth André, Justine Cassell, Eric Petajan, and Norman Badler. 2002. Creating interactive virtual humans: Some assembly required. *IEEE Intelligent Systems*, 17:54–63, July.

Dennis Hofs, Mariët Theune, and Rieks Akker op den. 2010. Natural interaction with a virtual guide in a virtual environment: A multimodal dialogue system. *Journal on Multimodal User Interfaces*, 3(1-2):141–153, March. Open Access.

Kate S. Hone and Robert Graham. 2000. Towards a tool for the subjective assessment of speech system interfaces (sassi). *Nat. Lang. Eng.*, 6(3-4):287–303, September.

W. L. Johnson, H. H. Vilhjálmsson, and S. Marsella. 2005. Serious games for language learning: How much game, how much AI? In *Artificial Intelligence in Education*.

S. Larsson and D. Traum. 2000. Information state and dialogue management in the TRINDI dialogue move engine toolkit. *Natural Language Engineering*, 6:323–340.

Anton Leuski and David Traum. 2008. A statistical approach for text processing in virtual humans. In *Proceedings of the 26th Army Science Conference*.

Anton Leuski, Ronakkumar Patel, David Traum, and Brandon Kennedy. 2006. Building effective question answering characters. In *Proceedings of the 7th SIGDIAL Workshop on Discourse and Dialogue*, pages 18–27.

Andrew Kachites McCallum. 2002. Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu.

Lina M. Rojas-Barahona, Alejandra Lorenzo, and Claire Gardent. 2012. Building and exploiting a corpus of dialog interactions between french speaking virtual and human agents. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (to be published)*.

David R. Traum and Staffan Larsson. 2003. The information state approach to dialogue management. In Jan Kuppevelt, Ronnie W. Smith, and Nancy Ide, editors, *Current and New Directions in Discourse and Dialogue*, volume 22 of *Text, Speech and Language Technology*, pages 325–353. Springer Netherlands. 10.1007/978-94-010-0019-2_15.

David R. Traum, Antonio Roque, Anton Leuski, Panayiotis Georgiou, Jillian Gerten, Bilyana Martinovski, Shrikanth Narayanan, Susan Robinson, and Ashish Vaswani. 2007. Hassan: A virtual human for tactical questioning. In *Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue*, page 71–74, September.

David Traum, Stacy C. Marsella, Jonathan Gratch, Jina Lee, and Arno Hartholt. 2008. Multi-party, multi-issue, multi-strategy negotiation for multi-modal virtual agents. In *Proceedings of the 8th international conference on Intelligent Virtual Agents*, IVA '08, pages 117–130, Berlin, Heidelberg. Springer-Verlag.

Marilyn A. Walker, Donald Hindle, Jeanne Fromer, Giuseppe Di Fabbrizio, and Craig Mestel. 1997. Evaluating competing agent strategies for a voice email agent. *CoRR*, cmp-lg/9706019.

A. Example of a dialog with the hybrid system. Translation into English of a dialog between Samir (the system) and Julie (the player) (Strategy 2), it corresponds to the third dialog in Table 1

Samir: Hello my name is Samir, the product designer

Samir: What are you doing here young people?

Julie: we come to build the joystick of Professor Geekman

Samir: You are in the right place. The plans are in the closet

Samir: Before would you like to hear about my job, the studies I did or my colleagues

Julie: Tell me about your job.

Samir: Well, for instance, is me who devise and design the items we construct in this enterprise

Samir: Do you want to know more about my colleagues or about the security measures applied here?

Julie: What do your colleagues do?

Samir: My job is universal! In fact it can please both girls that boys...

Samir: In your opinion, what studies should be done to get there

Julie: Designing studies?

Samir: This is a creative job. Several training can lead to the job arts, industrial design, computer-aided design

Samir: And does the safety standards interest you?

Julie: That is important the safety yes, tell me more

Samir: You know, here you are in a strategic place.
The novelties, the products of tomorrow, is here that they are born. They must be protected. For that reason we control the entrance here.

Samir: Well, Good Luck. Bye!

B. Translation to English of a dialog with the QA system. Dialog between the raw material operator (the system) and Julie (the player), it corresponds to the 8th dialog in Table 1(Strategy 4).

Preparateur1: Hello I am the raw material responsible. I can provide you a lot of information about this place.

Preparateur1: What can I do for you ?

Julie: What do you do here?

Preparateur1: My role is to ensure machines are supplied with the right materials,in good colors. I turn on the drying ovens and the whole supply system.

Julie: How does it work the supply system?

Preparateur1: Well when we get the pellet material, they are wet ...

Julie: What does it happen when there is a problem?

Preparateur1: In that case, I call the technician, he is the best for fixing everything

Julie:Thanks for the information, what should I do now?

Preparateur1: It is not complicated and you will learn things. I have a new machine to supply but I can not find the right material. Would you give me a hand?

Julie: yes

Preparateur1: You must supply this machine with a new generation of plastic. This machine allows the manufacture of plastic bags

Julie: ok

Preparateur1: Let's go!

Julie: Ok, Let's start!

Preparateur1: Great, Thanks!

Preparateur1: You are very kind, thank you.