Adaptive Information Presentation for Spoken Dialogue Systems: Evaluation with human subjects

Verena Rieser[1], **Simon Keizer**[1,2], **Xingkun Liu**[1], **Oliver Lemon**[1]

[1] Heriot-Watt University
Edinburgh, United Kingdom

[2] University of Cambridge
Cambridge, United Kingdom
{v.rieser,s.keizer,x.liu,o.lemon}@hw.ac.uk

Abstract

We present evaluation results with human subjects for a novel data-driven approach to Natural Language Generation in spoken dialogue systems. We evaluate a trained Information Presentation (IP) strategy in a deployed tourist-information spoken dialogue system. The IP problem is formulated as statistical decision making under uncertainty using Reinforcement Learning, where both content planning and attribute selection are jointly optimised based on data collected in a Wizard-of-Oz study. After earlier work testing and training this model in simulation, we now present results from an extensive online user study, involving 131 users and more than 800 test dialogues, which explores its contribution to overall 'global' task success. We find that the trained Information Presentation strategy significantly improves dialogue task completion, with up to a 9.7% increase (30% relative) compared to the deployed dialogue system which uses conventional, hand-coded presentation prompts. We also present subjective evaluation results and discuss the implications of these results for future work in dialogue management and NLG.

1 Introduction

Natural Language Generation (NLG) for Spoken Dialogue Systems serves two goals. On the one hand the "local" NLG task is to present "enough" information to the user (for example helping them to feel confident that they have a good overview of the search results) while keeping the utterances short and understandable. On the other hand, better Information Presentation should also contribute to the "global/ overall" dialogue task, so as to maximise task completion.

We have developed a novel framework for adaptive Natural Language Generation (NLG) where the problem is formulated as incremental decision making under uncertainty, which can be approached using Reinforcement Learning (Lemon, 2008; Rieser and Lemon, 2009; Rieser et al., 2010).This model is also being explored by other researchers (Dethlefs et al., 2011; Dethlefs and Cuayáhuitl, 2011) and (Janarthanam and Lemon, 2010; Janarthanam et al., 2011). We have applied the theory to a variety of NLG problems, such as referring expression generation, and here we focus on adaptive Information Presentation (IP) in spoken dialogue. The IP model is adaptive to noisy feedback from the current generation context (e.g. a user, a surface realiser, and a TTS engine), and it incrementally adapts the IP policy at the turn level. Reinforcement Learning is used to automatically optimise the IP policy with respect to a data-driven objective function.

In previous simulation-based work, we demonstrated that this IP model "locally" outperforms other IP strategies as used by conventional dialogue systems (Rieser and Lemon, 2009), as well as a more elaborate IP baseline strategy mimicking human "wizard" IP behaviour (Rieser et al., 2010). We have now integrated this policy into a full online dialogue system using Voice Over IP (VoIP), and evaluated its performance with real users. In particular, we test its ability to contribute to overall dialogue task success.

In Section 2 we briefly review the NLG framework as planning under uncertainty and how it was tested and trained in simulation. Section 3 explains

Proceedings of the 13th European Workshop on Natural Language Generation (ENLG), pages 102–109,
Nancy, France, September 2011.

how this trained policy was integrated into a fully working spoken dialogue system. Section 4 describes the experimental setup. In Section 5 we present the results, and in Section 6 we conclude with a discussion.

2 NLG as planning under uncertainty

We follow the overall framework of NLG as planning under uncertainty (Lemon, 2008; Rieser and Lemon, 2009; Rieser et al., 2010), where each NLG action is a sequential decision point, based on the current dialogue context and the expected long-term utility or "reward" of the chosen NLG action. Other recent approaches describe this task as planning, e.g. (Koller and Petrick, 2008), or as utility-based decision making (Deemter, 2009), but not as a statistical planning problem, where uncertainty in the stochastic environment is explicitly modelled. Below, we apply this framework to Information Presentation strategies in SDS using Reinforcement Learning (RL) (Sutton and Barto, 1998), where the example task is to present a set of search results (e.g. restaurants) to users. In particular, we consider 7 possible policies for structuring the content (see Figure 1): Recommending one single item, comparing two items, summarising all items, or ordered combinations of those actions, e.g. first summarise all the retrieved items and then recommend one of them. The IP module has to decide which action to take next, how many attributes to mention, and when to stop generating. We use a sentence generator based on the stochastic sentence planner SPaRKy (Stent et al., 2004) for surface generation.

Prior work has presented a variety of IP strategies for structuring information (see examples in Table 1). For example, the SUMMARY strategy is used to guide the user's "focus of attention". It draws the user's attention to relevant attributes by grouping the current results from the database into clusters, e.g. (Polifroni and Walker, 2008; Demberg and Moore, 2006). Other studies investigate a COMPARE strategy, where the attributes of individual items from the database result are compared, e.g. (Walker et al., 2007; Nakatsu, 2008). Most work in SDS however uses a RECOMMEND strategy, where only the top ranking item from the database result is presented, e.g. (Young et al., 2007).

We jointly optimise these 7 content structuring strategies together with attribute selection, i.e. how many attributes to mention in each strategy (e.g. SUMMARY(3)+RECOMMEND(2) with number of attributes in brackets). Attribute types are ranked according to a pre-defined user model (i.e. cuisine, price range, location, food quality, and service quality). We formulate the problem as a Markov Decision Process (MDP), where states are dialogue system contexts and actions are NLG decisions. Each state-action pair is associated with a transition probability, which is the probability of moving from state s at time t to state s' at time $t+1$ after having performed action a when in state s. This transition probability is computed by the environment model (i.e. the user simulation and realiser), and explicitly captures the uncertainty in the generation environment. This is a major difference to other non-statistical planning approaches. Each transition is also associated with a reinforcement signal (or "reward") r_{t+1} describing how good the result of action a was when performed in state s. The aim of the MDP is to maximise the long-term expected reward of its decisions, resulting in a *policy* which maps each possible state to an 'optimal' action in that state (i.e. the action with the highest expected long-term reward) (Rieser and Lemon, 2011).

$$\begin{bmatrix} \text{ACTION:} \begin{bmatrix} \textit{IP:} \begin{Bmatrix} \text{SUMMARY} \\ \text{COMPARE} \\ \text{RECOMMEND} \end{Bmatrix} \begin{Bmatrix} \text{attr: } 1\text{-}5 \end{Bmatrix} \end{bmatrix} \\ \text{STATE:} \begin{bmatrix} \texttt{attributes:} \{1\text{-}15\} \\ \texttt{sentence:} \{2\text{-}18\} \\ \texttt{dbHitsFocus:} \{1\text{-}100\} \\ \texttt{userSelect:} \{0,1\} \\ \texttt{userAddInfo:} \{0,1\} \\ \texttt{userElse:} \{0,1\} \end{bmatrix} \end{bmatrix}$$

Figure 2: State-Action space for the IP problem

We treat IP as a hierarchical joint optimisation problem, where first one of the IP structures (1-3) is chosen and then the number of attributes is decided, as shown in Figure 2. At each generation step, the MDP can choose 1-5 attributes. This results in 215 possible strategies, given the ordering constraints displayed in Figure 1. Generation stops as soon as the user is predicted to select a presented item, i.e. the "local" IP task is successful.

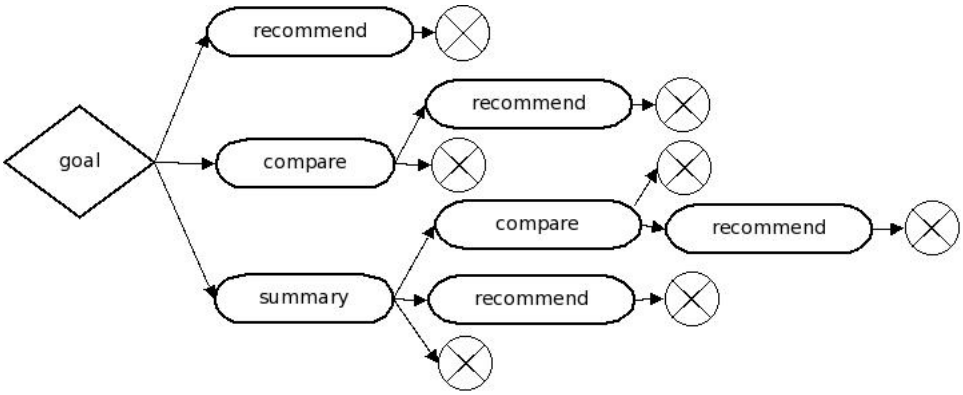

Figure 1: Possible Information Presentation structures (X=stop generation)

Strategy	Example utterance
SUMMARY	26 restaurants meet your query. There are 10 restaurants which serve Indian food and are in the cheap price range. There are also 16 others which are more expensive.
COMPARE	The restaurant called Maharajah and the restaurant called The Gandhi are both Indian restaurants. However, The Gandhi is in the cheap price range while Maharajah is moderately priced.
RECOMMEND	The restaurant called The Gandhi has the best overall quality amongst the matching restaurants. It is an Indian restaurant, and it is in the cheap price range.

Table 1: Example realisations, generated when the user provided `cuisine=Indian`, and where the NLG component has also selected the additional attribute `price` for presentation to the user.

States are represented as sets of dialogue system context features. The state space comprises "lower-level" features about the realiser behaviour (two discrete features representing the number of attributes and sentences generated so far) and three binary features representing the user's predicted next action, as well as "high-level" features provided by the Dialogue Manager (DM) (e.g. current database hits in the user's focus (`dbHitsFocus`)).

We train the policy in a simulated environment which is constructed from Wizard-of-Oz data (Liu et al., 2009). Simulated users for testing and training, as well as a data-driven reward function have been trained and evaluated using this data (Rieser et al., 2010). The data-driven reward function is formulated as a linear regression in equation (1) ($R^2 = .26$), which indicates that users like to be focused on a small set of database hits, which will enable them to choose an item ($valueUserReaction$), while keeping the IP utterances short (where $\#sentence$ is in the range [2-18]):

$$\begin{aligned} Reward &= 0.121 \times valueUserReaction \\ &\quad -1.2 \times \#DBhits \\ &\quad -1.43 \times \#sentence \end{aligned} \tag{1}$$

The policy was trained using the SHARSHA algorithm (a hierarchical version of SARSA) (Shapiro and Langley, 2002) with linear function approximation (Sutton and Barto, 1998).

3 System Integration

In order to evaluate our NLG strategy with real users, it was integrated into the 'CamInfo' system (Young et al., 2010), a spoken dialogue system providing tourist information for real locations in Cambridge. This baseline system has been made accessible by phone using VoIP technology, enabling out-of-lab evaluation with large numbers of users. Apart from practical advantages in managing evaluation campaigns, this development effort was also intended as a step towards evaluating spoken dialogue systems under more realistic conditions. Please note, however, that the users in this evaluation were still recruited and asked to complete predefined tasks (see Section 4), and therefore the evaluation might not be as realistic as an evaluation of a final deployed application with real users having real goals (Black et al., 2011).

The speech recogniser (ASR), semantic parser (SLU) and dialogue manager (DM) have all been

developed at Cambridge University. For speech synthesis (TTS), the Baratinoo synthesiser, developed at France Telecom, was used.

The DM uses a POMDP (Partially Observable Markov Decision Process) framework, allowing it to process N-Best lists of ASR hypotheses and keep track of multiple dialogue state hypotheses. The DM policy is trained to select system dialogue acts given a probability distribution over possible dialogue states. It has been shown that such dialogue managers can exploit the information in the N-Best lists (as opposed to only using the top ASR hypothesis) and are therefore particularly effective in noisy conditions (Young et al., 2010).

The natural language generation component of this baseline system is a standard rule-based surface realiser covering the full range of system dialogue acts that the dialogue manager can produce. It has only one IP strategy, i.e., the system only provides information about database entries in the form of single venue recommendations (the RECOMMEND strategy, see Table 1). The attributes of the venue to be presented are selected heuristically. In the extended version of the system, the IP strategy is replaced by our trained NLG component, which is optimised to decide between different IP strategies.

We follow a hybrid between statistical and rule-based approaches in order to integrate the trained policy: higher-level hand-coded rules impose a set of constraints on the statistical policy. Note that the possibility of constraining statistical policies with hard-coded rules is increasingly required for developing commercial dialogue systems (Williams, 2008). We follow a modular approach for integration, where the NLG and Dialogue Management strategies were trained separately (we discuss this issue further below).

We impose the following rule-based constraints on our policy in order to make it compatible with the (separately trained) DM policy:

- The chosen IP strategy must end with in a RECOMMEND action, since the DM expects (exactly one) named entity to be mentioned.
- COMPARE actions are excluded in order to not introduce new named entities that the user may refer to later (since the DM was not optimised under this condition).
- The attribute selection is forced to present at least the attributes chosen by the DM.

The remaining decision points are: choosing between RECOMMEND and SUMMARY+RECOMMEND, as well as selecting additional attributes to present to the user. Although this is a somewhat limited version of the fully optimised IP strategy, it is still interesting to discover whether even a limited amount of NLG optimisation (in terms of more elaborate IP strategies and attribute selection) has an effect on overall global system performance.

Hence, in this real user evaluation, we compared the baseline system, incorporating a single recommendation IP strategy only, with the extended system, incorporating our trained NLG IP policy. In a previous proof-of-concept study (Rieser and Lemon, 2009) a similar rule-based baseline NLG strategy (RECOMMEND only) was shown to be outperformed in simulation. We now test whether these results transfer to real user settings. In the remainder of this paper we will refer to the baseline system as *BASE* system and to the system with the integrated trained IP strategy as *TIP*.

4 Experimental Setup

For the evaluation of the two systems, two approaches to managing subjects were taken. In the first approach, subjects were recruited using mail-shots and web-based advertising amongst people from Cambridge and Edinburgh, mostly students. From the resulting pool of subjects, people were gradually invited to start the tasks, in their own time, and within a given trial period of around two weeks. After the trial period, they were paid (using PayPal) per completed task, with a required minimum of 15 tasks, and a maximum of 40 tasks. For the two systems, this resulted in a corpus of 304 dialogues. In the second approach, an alternative method of managing subjects was used, using Amazon Mechanical Turk (Jurcicek et al., 2011). In this setup, tasks are published as so-called HITs (Human Intelligence Tasks) on a web-server and registered workers can complete them. This setup resulted in 532 collected dialogues for the two systems compared[1]. In the remainder of this paper, we will refer to the corpus

[1]This evaluation was part of a bigger evaluation campaign, in which 2046 dialogues were collected in total.

obtained with 'locally' managed subjects as *Feb11-LOC* and to the corpus obtained using Amazon Mechanical Turk as *Feb11-AMT*.

In both of the above-mentioned approaches, the subjects were directed to a webpage with detailed instructions and for each task, a phone number to call and the scenario to follow. The subjects were randomly assigned to interact with one of the systems (BASE or TIP). A scenario describes a place to eat in town, with some constraints, for example: *"You want to find a moderately priced restaurant and it should be in the Riverside area. You want to know the address, phone number, and type of food."*. After the dialogue, the subjects were asked to fill in a short questionnaire, assessing the impact of IP strategies on the users' perception of various system components:

Q1. Did you find all the information you were looking for? [Yes / No]

Please state your attitude towards the following statements:

Q2. The system understood me well. [1 – 6]

Q3. The phrasing of the system's responses was good. [1 – 6]

Q4. The system's voice was of good quality. [1 – 6]

1: strongly disagree	4: slightly agree
2: disagree	5: agree
3: slightly disagree	6: strongly agree

Table 2 summarises the two corpora of collected data. For the Feb11-AMT corpus, considerably more subjects were used, although many of them did only a small number of tasks. For the Feb11-LOC corpus, it was more difficult to recruit many subjects, but in this setup, the subjects could be asked to complete a minimum number of tasks, hence the higher average number of dialogues per user.

Also note, that the Word Error Rate (WER) is relatively high in both corpora. This is partly due to the fact that the ASR module had not been trained properly for this particular domain due to lack of training data. Furthermore, some of the subjects were non-native speakers and some subjects used Skype to call the systems, which causes distortion of the audio signal. These conditions are the same for both BASE and TIP systems. Despite the high ASR error rates, overall task completion rates were high, due to the robustness of the POMDP dialogue manager.

Corpus	nDials	AvgTurns	nUsers	nDsUsr	WER
Feb11-LOC	304	11.48	19	16.00	56.5
Feb11-AMT	532	10.09	113	4.71	53.6

Table 2: Overview of collected data, with for each corpus the number of dialogues (nDials), the average number of user turns per dialogue (AvgTurns), the number of unique users (nUsers), the average number of dialogues per user (nDsUsr), and the word error rate (WER).

The overall most frequently employed IP strategy is SUMMARY(2)+RECOMMEND(2), see Table 3. Also, note that the trained policy never employed more than 3 attributes, and always chose to use the same number of attributes for its combined IP strategies.

Frequ.	Strategy(Attributes)
1	RECOMMEND(1)
123	RECOMMEND(2)
163	RECOMMEND(3)
254	SUMMARY(1)+RECOMMEND(1)
778	SUMMARY(2)+RECOMMEND(2)
270	SUMMARY(3)+RECOMMEND(3)

Table 3: Frequency of occurrences of each IP strategy observed in the evaluation with number of attributes in brackets.

5 Results

After processing the log files and completed user questionnaires, both objective and subjective performance measures were computed in order to compare the systems.

5.1 Objective evaluation

For the objective evaluation of the two dialogue systems we focused on measuring goal completion rates, which can be done in different ways. First, we can take the goal specification assigned to the user for each dialogue and then analyse the system dialogue acts. *Partial completion* (ObjSucc-PC) is achieved when the system has offered a venue that matches the constraints as specified in the assigned goal, for example it has provided the name of a cheap chinese restaurant in the riverside area. *Full completion* (ObSucc-FC) is achieved when the system has also provided the required additional information about that venue, for example the phone number and address.

In Table 4, all success rates obtained from the February 2011 evaluation are given, for the corpus

Corpus	System	nDials	nTurns	SubjSucc	ObjSucc-PC	ObjSucc-FC
Feb11-LOC	BASE	199	11.69	65.33 (6.61)	73.37 (6.14)	46.73 (6.93)
	TIP	105	11.02	60.00 (9.37)	77.23 (8.02)	49.50 (9.56)
Feb11-AMT	BASE	402	9.86	64.18 (4.69)	51.00 (4.89)	28.86 (4.43)
	TIP	130	10.83	56.15 (8.53)	60.77 (8.39)	37.69 (8.33)
Feb11-TOT	BASE	601	10.46	64.56 (3.82)	58.40 (3.94)	34.78 (3.81)
	TIP	235	10.91	57.87 (6.31)	**68.09 (5.96)**$*$	**42.98 (6.33)**$*$

Table 4: Overview of all success rates (%) obtained for the two corpora, including subjective success obtained from Q1 of the user questionnaire(SubjSucc), objective success based on assigned goals (ObjSucc-PC for partial completion and ObjSucc-FC for full completion). 95% confidence intervals for all success rates are indicated in brackets; statistically significant improvements ($p < 0.05$ using a z-test) are indicated with an asterisk (*). Also given are the number of dialogues (nDials) and dialogue length in terms of the average number of user turns per dialogue(nTurns).

with data from locally recruited subjects (Feb11-LOC), and the corpus with data from Amazon Mechanical Turk workers, as well as both corpora pooled together (Feb11-TOT). The results show that the system with our NLG component (TIP) outperforms the baseline system (BASE) on all objective success rates in both corpora. Relative improvements of up to 30% for full completion on the Feb11-AMT corpus were obtained. After pooling the two corpora together, we have a sufficient number of dialogues to show that the improvement from our NLG strategy is statistically significant on both partial and full completion (using a 2-tailed z-test for two proportions).

It is also interesting to note that the average number of user turns per dialogue is not significantly different between systems in both corpora, suggesting that the contribution of the trained IP policy to system performance manifests itself primarily in terms of effectiveness rather than efficiency. By providing more useful information to the user, the system might help them to find an appropriate venue in fewer turns, but due to the lengthy system prompts, more turns might be needed to recover from speech recognition errors (see WER in Table 2).

5.2 Subjective evaluation

Table 5 summarises the subjective user scores from the questionnaire (see Section 4). In terms of subjective success rates (Q1), the baseline system (BASE) obtains slightly higher scores on both corpora, although no statistically significant differences were found. We will further discuss these results in section 6.

When comparing the other subjective scores (Q2–Q4) on a scale of [1–6], using a Mann-Whitney test, the only case where a statistically significant difference is found between the two systems is the score for *Q4:VoiceQuality* in the Feb11-LOC corpus, where the baseline system is significantly better. Since the the TTS voice is exactly the same for both systems, the difference in perceived voice quality might be influenced by the longer system prompts for the TIP system. However, we don't see this pattern in the Feb11-AMT corpus.

Corpus	System	Q1	Q2	Q3	Q4
Feb11-LOC	BASE	65.33	3.69	3.94	**4.23***
	TIP	60.00	3.44	3.70	3.91
Feb11-AMT	BASE	64.18	3.92	4.16	3.81
	TIP	56.15	3.87	4.30	3.85
Feb11-TOT	BASE	64.56	3.85	4.10	3.95
	TIP	57.87	3.68	4.03	3.88

Table 5: Subjective evaluation results, based on the questionnaire [Q1-Q4], where an asterisk (*) denotes a significant difference at $p < 0.05$ (using a z-test for Q1 and a Mann-Whitney test for Q2–Q4).

We also compared the Mechanical Turk setup to the setup where subjects where recruited locally (Feb11-AMT vs. Feb11-LOC for both systems). For the TIP system, *Q2:Understanding* and *Q3:Phrasing* are significantly higher in the Feb11-AMT corpus compared to the FEB11-LOC corpus. Similarly, the BASE system performs significantly better for *Q3:Phrasing* under the Mechanical Turk setting. However, when combining the results for all the subjective scores (similar to the objective scores), none of the differences are significant.

In sum, there is no difference in user ratings between the original BASE system and the TIP system with the integrated trained NLG strategy, except for *Q4:VoiceQuality*, which is better rated for

the BASE system in the Feb11-LOC corpus, even though the systems had identical TTS. The difference in ratings between the Feb11-LOC and Feb11-AMT corpora suggests that the way in which subjects are recruited, instructed and payed, as well as the user population targeted, has an impact on subjective ratings obtained.

6 Discussion

Following previous work on a novel NLG model in which content planning and attribute selection are formulated as statistical planning under uncertainty, this paper has presented results of the evaluation of this NLG model with real users, focussing on contribution to overall task success in spoken dialogue systems. The NLG model that was trained in a simulated environment was integrated in a deployed spoken dialogue system for tourist information and evaluated in an online experiment with 131 real users and over 800 dialogues. The results showed that the trained Information Presentation model significantly improves objective dialogue task completion, with up to a 30% relative increase (+9.7% raw improvement) compared to a state-of-the-art deployed dialogue system that generates conventional, hand-coded presentation prompts. This outcome confirms earlier results from a previous proof-of-concept study (Rieser and Lemon, 2009), where a similar baseline was shown to be outperformed in simulation.

The subjective scores however were quite similar between the two systems, and in terms of perceived success rate, the baseline system scored better, though not statistically significantly. One possible explanation is that the more elaborate TIP strategy might have somehow obscured the users' perceptions of task completion (even though the objective task completion was significantly higher).

An important factor that may have influenced the results, was that the word error rate was relatively high throughout the data. The more elaborate information presentation prompts from the integrated system (TIP) might have exacerbated the many speech recognition problems, where the DM might have falsely initiated a lengthy Information Presentation prompt after a mis-recognition error. This is also suggested by the analysis of dialogue length, which turned out to be very similar between the two systems. By providing more useful information to the user, the TIP system might help them to find an appropriate venue in fewer turns, but due to the lengthy system prompts, more turns might be needed to recover from speech recognition errors.

Although these evaluation results are very positive, a system setup which combines separately trained dialogue manager and NLG components is not ideal. In this case the dialogue manager was trained in a setup where only the single item recommendation strategy for IP is used. Therefore, for the dialogue manager state update, only dialogue acts for such IP prompts are expected. If the trained NLG model decides to use an alternative IP strategy, a mismatch is then potentially caused between what the dialogue manager planned and what is actually presented to the real user. Therefore, the NLG module might result in user behaviour that the dialogue manager is not optimised for. As a practical compromise it was therefore decided (as explained above) to require all IP prompts to end with a single item recommendation, and the COMPARE strategy was blocked during the evaluation. Therefore, neither DM nor NLG were trained for the final operating conditions that they would experience in this application, though the constraints on NLG mentioned above meant that the DM's chosen actions were maintained. In future work we therefore strive to jointly optimise the DM and NLG strategies (see also (Lemon, 2011)), and it is likely that full use of an optimised IP strategy would lead to an even greater performance boost in the overall system. We would expect that a joint optimisation of DM and NLG policies would prevent the DM from initiating long IP prompts after likely mis-recognitions. We predict that the results obtained in this study would be even stronger for a jointly-optimised DM+NLG strategy, and we pursue this in current work.

Finally, we note that the overall framework has also been used for optimising generation of referring expressions, including adaptive generation of temporal referring expressions, where similar results have been found in boosting overall task success of spoken dialogue systems (Janarthanam et al., 2011). This set of results shows that there are significant 'global' benefits to be gained by viewing NLG as statistical planning under uncertainty.

Acknowledgments

The research leading to these results has received funding from the EC's 7th Framework Programme (FP7/2007-2013) under grant agreement no. 216594 (CLASSiC project www.classic-project.org), and (FP7/2011-2014) under grant agreement no. 270019 (SpaceBook project).

References

Alan W Black, Susanne Burger, Alistair Conkie, Helen Hastie, Simon Keizer, Oliver Lemon, Nicolas Merigaud, Gabriel Parent, Gabriel Schubiner, Blaise Thomson, Jason D. Williams, Kai Yu, Steve Young, and Maxine Eskenazi. 2011. Spoken Dialog Challenge 2010: Comparison of Live and Control Test Results. In *Proceedings of SIGDIAL*.

Kees van Deemter. 2009. What game theory can do for NLG: the case of vague language. In *12th European Workshop on Natural Language Generation (ENLG)*.

Vera Demberg and Johanna Moore. 2006. Information presentation in spoken dialogue systems. In *Proc. of the Conference of the European Chapter of the ACL (EACL)*.

Nina Dethlefs and Heriberto Cuayáhuitl. 2011. Hierarchical Reinforcement Learning and Hidden Markov Models for Task-Oriented Natural Language Generation. In *Proc. of ACL*.

Nina Dethlefs, Heriberto Cuayáhuitl, and Jette Viethen. 2011. Optimising Natural Language Generation Decision Making For Situated Dialogue. In *Proc. of SIGDIAL*.

Srinivasan Janarthanam and Oliver Lemon. 2010. Adaptive Referring Expression Generation in Spoken Dialogue Systems: Evaluation with Real Users. In *Proceedings of SIGDIAL*.

Srinivasan Janarthanam, Helen Hastie, Oliver Lemon, and Xingkun Liu. 2011. "The day after the day after tomorrow?" a machine learning approach to adaptive temporal expression generation: training and evaluation with real users. In *Proc. of SIGDIAL*.

F. Jurcicek, S. Keizer, M. Gasic, F. Mairesse, B. Thomson, K. Yu, and S. Young. 2011. Real user evaluation of spoken dialogue systems using amazon mechanical turk. In *Proc. Interspeech*, Florence, Italy, August.

Alexander Koller and Ronald Petrick. 2008. Experiences with Planning for Natural Language Generation. In *ICAPS*.

Oliver Lemon. 2008. Adaptive natural language generation in dialogue using Reinforcement Learning. In *Proc. of the 12th SEMdial Workshop on on the Semantics and Pragmatics of Dialogues*, London, UK, June.

Oliver Lemon. 2011. Learning what to say and how to say it: joint optimization of spoken dialogue management and Natural Language Generation. *Computer Speech and Language*, 25(2):210–221.

Xingkun Liu, Verena Rieser, and Oliver Lemon. 2009. A Wizard-of-Oz interface to study information presentation strategies for spoken dialogue systems. In *Proc. of the 1st International Workshop on Spoken Dialogue Systems*.

Crystal Nakatsu. 2008. Learning contrastive connectives in sentence realization ranking. In *Proc. of SIGdial Workshop on Discourse and Dialogue*.

Joseph Polifroni and Marilyn Walker. 2008. Intensional Summaries as Cooperative Responses in Dialogue Automation and Evaluation. In *Proceedings of ACL*.

Verena Rieser and Oliver Lemon. 2009. Natural Language Generation as Planning Under Uncertainty for Spoken Dialogue Systems. In *Proc. of EACL*.

Verena Rieser and Oliver Lemon. 2011. Learning and Evaluation of Dialogue Strategies for new Applications: Empirical Methods for Optimization from Small Data Sets. *Computational Linguistics*, 37(1).

Verena Rieser, Oliver Lemon, and Xingkun Liu. 2010. Optimising Information Presentation for Spoken Dialogue Systems. In *Proceedings of ACL*, pages 1009–1018, Uppsala, Sweden, July.

Dan Shapiro and P. Langley. 2002. Separating skills from preference: Using learning to program by reward. In *Proc. of the 19th International Conference on Machine Learning (ICML)*, pages 570–577, Sydney, Australia, July.

Amanda Stent, Rashmi Prasad, and Marilyn Walker. 2004. Trainable sentence planning for complex information presentation in spoken dialog systems. In *Proc. ACL*.

R. Sutton and A. Barto. 1998. *Reinforcement Learning*. MIT Press.

Marilyn Walker, Amanda Stent, François Mairesse, and Rashmi Prasad. 2007. Individual and domain adaptation in sentence planning for dialogue. *Journal of Artificial Intelligence Research (JAIR)*, 30:413–456.

Jason D. Williams. 2008. The best of both worlds: Unifying conventional dialog systems and POMDP. In *Proceedings of Interspeech*.

SJ Young, J Schatzmann, K Weilhammer, and H Ye. 2007. The Hidden Information State Approach to Dialog Management. In *ICASSP 2007*.

S. Young, M. Gašić, S. Keizer, F. Mairesse, B. Thomson, and K. Yu. 2010. The Hidden Information State model: a practical framework for POMDP based spoken dialogue management. *Computer Speech and Language*, 24(2):150–174.