Quality-adaptive Spoken Dialogue Initiative Selection And Implications On Reward Modelling

Stefan Ultes, Matthias Kraus, Alexander Schmitt, and Wolfgang Minker
Ulm University
Albert-Einstein-Allee 43
89081 Ulm, Germany
{fistname.lastname}@uni-ulm.de

Abstract

Adapting Spoken Dialogue Systems to the user is supposed to result in more efficient and successful dialogues. In this work, we present an evaluation of a quality-adaptive strategy with a user simulator adapting the dialogue initiative dynamically during the ongoing interaction and show that it outperforms conventional non-adaptive strategies and a random strategy. Furthermore, we indicate a correlation between Interaction Quality and dialogue completion rate, task success rate, and average dialogue length. Finally, we analyze the correlation between task success and interaction quality in more detail identifying the usefulness of interaction quality for modelling the reward of reinforcement learning strategy optimization.

1 Introduction

Maximizing task success in task-oriented dialogue systems has always been a central claim of Spoken Dialogue (SDS) research. Today, commercial systems are still inflexible and do not adapt to users or the dialogue flow. This usually results in bad performance and in frequently unsuccessful dialogues. In recent years, adaptation strategies have been investigated for rendering SDS more flexible and robust. The aim of those strategies is to adapt the dialogue flow based on observations that are made during an ongoing dialogue.

One approach to observe and score the interaction between the system and the user is the Interaction Quality (IQ) (Schmitt and Ultes, 2015) originally presented by Schmitt et al. (2011). Their Interaction Quality paradigm is one of the first metrics which can be used for this purpose. A pilot user study on adapting the dialogue to the Interaction Quality by Ultes et al. (2014b) in a limited domain has already shown encouraging results. There, similar dialogue performance was achieved for both the strategy adapting the grounding mechanism to Interaction Quality and the strategy of always applying implicit confirmation prompts previously known to achieve best user feedback.

While the previous experiment showed encouraging results for adapting the grounding strategy, it is unclear if other aspects of a dialogue strategy may also be positively affected. Hence, in this contribution, we investigate if applying rules for adapting the dialogue initiative to IQ may also result in an increase in IQ and if other metrics like task success rate or dialogue completion rate may correlate[1].

To investigate this, we have designed a basic experiment having an IQ-adaptive dialogue strategy adapting the dialogue initiative. Depending on the IQ score, the system chooses between user-initiative, system-initiative and mixed-initiative. Moreover, the performance of four additional strategies is analyzed regarding a correlation between IQ and other performance measures.

Besides the interest in the general performance of the quality-adaptive strategy, we are specifically interested whether implications may be drawn from the experiments about the usage of IQ in a reinforcement learning setting for modelling the reward function.

The outline of the paper is as follows: in Section 2, we present significant related work on adaptive dialogue and quality metrics including the Interaction Quality (IQ) paradigm, a more abstract form of user satisfaction. All five dialogue strategies are described in detail in Section 3. The experimental setup including the test system in the the "Let's Go" domain is presented in Section 4

[1] Automatic optimization aims at maximizing a reward function. If IQ was contributing positively to this reward function, optimisation would naturally result in an increase in IQ. As we do not perform optimization, this correlation does not automatically exist

Proceedings of the SIGDIAL 2015 Conference, pages 374–383,
Prague, Czech Republic, 2-4 September 2015.

followed by a thorough presentation of the experimental results based on dialogues with the user simulator. Based on the experiments' results, inferences are drawn on using IQ for reward modelling. Finally, we conclude and outline future work in Section 6.

2 Significant Related Work

The field of adaptive dialogue spans over many different types of adaptation. While some systems adapt to their environment (e.g., (Heinroth et al., 2010)), the focus of this work lies on systems that adapt to the user and the characteristics of the interaction. More specifically, an emphasis is placed on dynamic adaptation to the user during the ongoing dialogue.

2.1 User-Adaptive Dialogue

A very prominent work closely related to the topic of this contribution has been presented by Litman and Pan (2002). They identify problematic situations in dialogues by analyzing the performance of the speech recognizer (ASR) and use this information to adapt the dialogue strategy. Each dialogue starts off with a user initiated strategy without confirmations. Depending on the ASR performance, a system-directed strategy with explicit confirmations may eventually be employed. Applied to TOOT, a system for getting information about train schedules, the authors achieved significant improvement in task success compared to a non-adaptive system. While Litman and Pan adapt to the ASR performance as indicator for problematic dialogues (being a system property representing an objective adaptation criterion), the user is put into the focus of adaptation in this work by using an abstract form of user satisfaction hence applying a subjective criterion.

Further work on user-adaptive dialogue has been presented by Gnjatović and Rösner (2008) adapting to the user's emotional state and by Nothdurft et al. (2012) adapting to the user knowledge. For both, only simulated or predefined user states are used while this work uses a real estimation module deriving the user satisfaction.

Using user ratings to improve the dialogue performance in a reinforcement learning (RL) approach has been presented by Walker (2000), Rieser and Lemon (2008), Janarthanam and Lemon (2008), and Gašić et al. (2013). Walker applied RL to a MDP-based dialogue system ELVIS for accessing emails over the phone. They modeled the reward function using the PARADISE framework (Walker et al., 1997) showing that the resulting policy improved the system performance in terms of user satisfaction significantly. The resulting best policy showed, among other aspects, that the system-initiative strategy was found to work best. The group of Lemon also employed PARADISE for modelling the reward function. Using reinforcement learning, they found an optimal dialogue strategy for result presentation (Rieser and Lemon, 2008) or referring expressions (Janarthanam and Lemon, 2008) for natural language generation.

For a POMDP-based dialogue manager, Gašić et al. use a reward function based on user ratings to train the optimized policy. The user ratings are acquired using Amazon Mechanical Turk. They show that their approach converges much faster than conventional approaches using a user simulator. However, their approach does not allow for adapting the course of the dialogue online but relies on a pre-optimized dialogue strategy.

Finally, not directly providing user adaptivity but allowing for reacting to specific dialogue situations in a rule-based manner is VoiceXML (Oshry et al., 2007). By counting the number of "reprompts" or "nomatches", a suitable strategy may be selected. While these parameters are also part of the Interaction Quality used for adaptation within this work, the Interaction Quality captures more complex effects than the simple rules of VoiceXML. These effects may not be modeled easily using rules (Ultes and Minker, 2013).

2.2 Interaction Quality

While there is numerous work on investigating turn-wise quality ratings for SDSs, e.g., Engelbrecht et al. (2009), Higashinaka et al. (2010) and Hara et al. (2010), the Interaction Quality paradigm by Schmitt et al. (2011) seems to be the only metric fulfilling the requirements for adapting the dialogue online (Ultes et al., 2012).

For rendering an SDS adaptive to the user's satisfaction level, a module is needed to automatically derive the satisfaction from the ongoing interaction. For creating this module, usually, dialogues have to be annotated with ratings describing the user's satisfaction level. Schmitt et al. (2015) proposed a measure called "Interaction Quality" (IQ) which fulfills the requirements of a

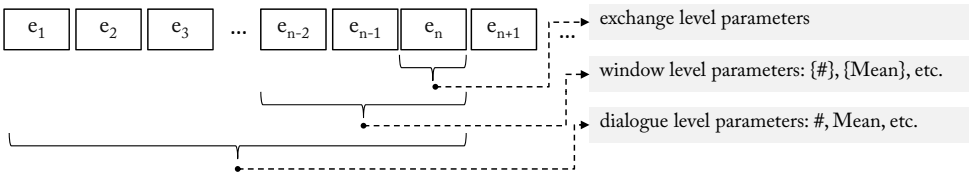

Figure 1: The three different modeling levels representing the interaction at exchange e_n: The most detailed exchange level, comprising parameters of the current exchange; the window level, capturing important parameters from the previous n dialogue steps (here $n = 3$); the dialogue level, measuring overall performance values from the entire previous interaction.

quality metric for adaptive dialogue identified by Ultes et al. (2012). For Schmitt et al., the main aspect of user satisfaction is that it is assigned by real users. However, this seems to be impractical in many real world scenarios. Hence, the usage of expert raters is proposed. Further studies have also shown a high correlation between quality ratings applied by experts and users (Ultes et al., 2013).

The IQ paradigm is based on automatically deriving interaction parameters from the SDS and feed these parameters into a statistical classification module which predicts the IQ level of the ongoing interaction at the current system-user-exchange [2]. The interaction parameters are rendered on three levels (see Figure 1): the exchange level, the window level, and the dialogue level. The exchange level comprises parameters derived from SDS modules Automatic Speech Recognizer, Spoken Language Understanding, and Dialogue Management directly. Parameters on the window and the dialogue level are sums, means, frequencies or counts of exchange level parameters. While dialogue level parameters are computed out of all exchanges up to the current exchange, window level parameters are only computed out of the last three exchanges.

These interaction parameters are used as input variables to a statistical classification module. The statistical model is trained based on annotated dialogues of the Lets Go Bus Information System in Pittsburgh, USA (Raux et al., 2006). Each of the 4,885 exchanges (200 calls) has been annotated by three different raters resulting in a rating agreement of $\kappa = 0.54$. The final IQ value of the three raters is derived using the median. Furthermore, the raters had to follow labeling guidelines to enable a consistent labeling process (Schmitt et al., 2012). Schmitt et al. (2011) estimated IQ with a Support Vector Machine using only automatically derivable parameters achieving an unweighted average recall of 0.59.

[2] A system turn followed by an user turn

3 Quality-Adaptive Dialogue

Within this section, we describe one part of the main contribution of rendering the dialogue initiative adaptive to Interaction Quality and compare the resulting strategy to several non-adaptive strategies. Conventional dialogue initiative categories are *user initiative*, *system initiative*, and *mixed initiative* (McTear, 2004). As there are different interpretations of what these initiative categories mean, we stick to the understanding of initiative as used by Litman and Pan (2002): the initiative influences the openness of the system question and the set of allowed user responses. The latter is realized by defining which slot values provided by the user are processed by the system and which are discarded. Hence, for *user initiative*, the system asks an open question allowing the user to respond with information for any slot. For *mixed initiative*, the system poses a question directly addressing a slot. However, the user may still provide information for any slot. This is in contrast to the *system initiative*, where the user may only respond with the slot addressed by the system. For instance, if the system asks for the arrival place and the user responds with a destination place, this information may either be used (*mixed initiative*) or discarded (*system initiative*).

In this work, five different strategies are created. Three basic non-adaptive strategies are compared against one adaptive and one random adaptive strategy. All of these strategies can be generated from the flow diagram in Figure 2 by varying the IQ value. The non-adaptive *user*, *system*, and *mixed initiative strategy* are well known concepts and will not further be described. In order to keep the strategies comparable, all have a similar structure: in each strategy, the system starts with

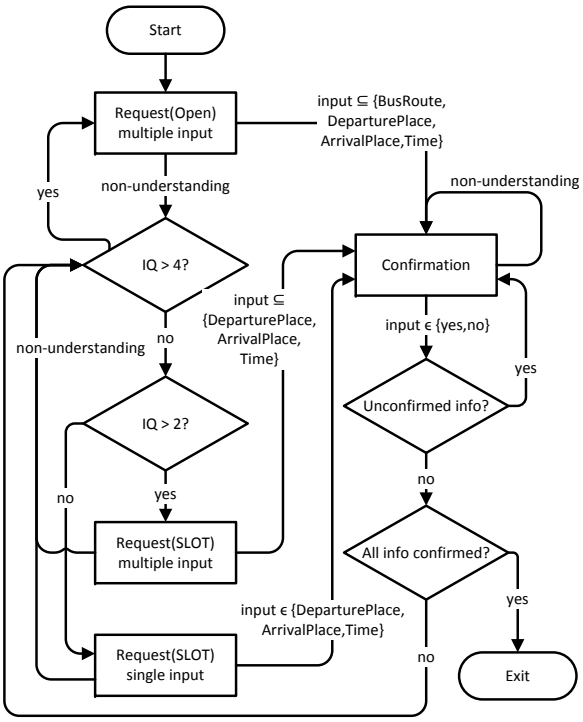

Figure 2: The flow chart describing the adaptive and non-adaptive strategies. For the adaptive strategy, the course of the dialogue as well as the allowed user input are influenced by the IQ value. For the random strategy, the IQ values are generated randomly. The non-adaptive strategies are realized by fixed IQ values: $IQ = 5$ for the user initiative strategy always posing open requests, $IQ = 3$ and $IQ = 1$ for mixed and system initiative explicitly requesting slot information. Provision of the bus route was not mandatory.

an open request allowing the user to respond with information for all slots. The system first continues with confirming provided information before continuing strategy-specific.

For adapting the initiative based on IQ, the *adaptive strategy* utilizes the basic concepts of the non-adaptive strategies, i.e., the pairs of system question and its restriction on the user input. Hence, the way missing information is requested depends on the Interaction Quality. For an IQ value of five, an open request is placed. For an IQ value greater than two, information for all missing slots is allowed as user input (same behavior as in the mixed initiative strategy) while only the requested information is allowed otherwise. If unconfirmed slot information is present, the strategy decides to first initialize grounding before other missing information is requested. If the user provides information for an already confirmed slot, this information is discarded. The same behavior is implemented into the *user* and *mixed* initiative strategies. Note that the thresholds between the different adaptation levels have been defined arbitrarily based on human judgement. An example dialogue is depicted in Figure 3.

System:	*Request(Open)*	
User:	Non-understanding	$IQ = 5$
System:	*Request(Open)*	
User:	Inform(Travel Time: 8pm)	$IQ = 5$
System:	*Confirm(Travel Time: 8pm)*	
User:	Deny	$IQ = 3$
System:	*Request(Departure place)*	
User:	Inform(Travel Time: now)	$IQ = 3$
System:	*Confirm(Travel Time: now)*	
User:	...	

Figure 3: Example dialogue of the adaptive strategy. As the IQ value is 5 in the beginning, the system requests openly for information. After the IQ value has dropped to 3, the mixed initiative is active. Hence, the system asks for specific information directly still allowing input for other slots.

The *random strategy* uses the same dialogue definition as the adaptive strategy. However, the initiative is selected randomly.

The dialogues of all strategies continue until all mandatory slots contain a confirmed value or the user terminates the interaction. If the user responds with information about a slot which is not in the set of allowed slot information, these values are discarded. This may lead to a 'Non-Understanding' (or 'out-of-grammar' user input) even though the user has provided information.

4 Experiments and Results

Evaluation of the dialogue strategies presented in Section 3 is performed using an adaptive dialogue system interacting with a user simulator. A user simulator offers an easy and cost-effective way for getting a basic impression about the performance of the designed dialogue strategies. Furthermore, we describe the setup of the experiments followed by a discussion of the results.

4.1 Let's Go Domain

For evaluating the adaptive strategies, we use the Let's Go Domain as it represents a domain of suitable complexity. The Let's Go Bus Information System (Raux et al., 2006) is a live system in Pittsburgh, USA providing bus schedule information to

the user. The Let's Go User Simulator (LGUS) by Lee and Eskenazi (2012) is used for evaluation to replace the need for human evaluators.

The dialogue goal of Let's Go consists of four slots: bus number, departure place, arrival place, and travel time. However, the bus number is not mandatory. The original system contains more than 300,000 arrival or departure places, respectively. To acquire information about the specific goal of the user, the system may use one out of nine system actions to which the user responds with a subset of six user actions. In LGUS, the user actions are accompanied with a confidence score simulating automatic speech recognition performance. The system action is either requesting for information or explicitly confirming previously shared information. Hence, the user may either provide information about a certain slot or affirm or deny a slot value.

Any combination of the user actions is possible—even having contradicting information present, e.g., informing about two different values of the same slot or affirming and denying a value at the same time. As problems with the speech recognition and language understanding modules are also modeled by LGUS, these effects are reflected by the user action 'Non-Understanding'.

4.2 Experimental Setup

In order to evaluate the dialogue strategies, we use the adaptive dialogue manager OwlSpeak (Ultes and Minker, 2014), originally created by Heinroth et al. (2010), extended for including quality-adaptivity (Ultes et al., 2014a). OwlSpeak is based on the Model-View-Presenter paradigm separating the dialogue description and dialogue state in the model from the dialogue control logic in the presenter. Originally, the interface to a voice browser using VoiceXML (Oshry et al., 2007) is embedded in the view. For this work, the view has been replaced in order to provide an interface to LGUS which is instantiated as a server application communicating to other modules using JSON (Crockford, 2006). Furthermore, the system has been extended to handle multi-slot user input.

For rendering the system adaptive, Ultes et al. (2014a) included an interaction estimation module into the system. It is based on the Support Vector Machine (SVM (Vapnik, 1995)) implementation LibSVM (Chang and Lin, 2011) using a linear kernel. Interaction with real users requires a more complex system than an interaction with a simulated user. Thus, some SDS modules are missing and not all parameters of the IQ paradigm are available. This results in a feature set of only 16 parameters[3]. The trained model achieves an unweighted average recall[4] of 0.56[5] on the training data using 10-fold cross-validation which is a considerably good performance. All exchanges of the LEGO corpus (Schmitt et al., 2012) have been used for training.

Evaluation of the dialogue strategies is performed by creating 5,000 simulated dialogues for each strategy. Like Raux et al. (2006), short dialogues ($\leq$ 5 exchanges[6]) which are considered "not [to] be genuine attempts at using the system" are excluded from all statistics in this paper.

Three objective metrics are used to evaluate the dialogue performance: the average dialogue length (ADL), the dialogue completion rate (DCR) and task success rate (TSR). The ADL is modeled by the average number of exchanges per completed dialogue. A dialogue is regarded as being completed if the system provides a result—whether correct or not—to the user. Hence, DCR represents the ratio of dialogues for which the system was able to provide a result, i.e., provide schedule information:

$$DCR = \frac{\#completed}{\#all} .$$

TSR is the ratio of completed dialogues where the user goal matches the information the system acquired during the interaction:

$$TSR = \frac{\#correctResult}{\#completed} .$$

Here, only destination place, arrival place, and travel time are considered as the bus number is not a mandatory slot and hence not necessary for providing information to the user.

As a correlation between objective measures and IQ is investigated, the average IQ value (AIQ) is calculated for each strategy based on the IQ

[3]The parameters applied are ASRRecognitionStatus, ASRConfidence, RePrompt?, #Exchanges, ActivityType, Confirmation?, MeanASRConfidence, #ASRSuccess, %ASRSuccess, #ASRRejections, %ASRRejections, {Mean}ASRConfidence, {#}ASRSuccess, {#}ASRRejections, {#}RePrompts, {#}SystemQuestions.

[4]The arithemtic average over all class-wise recalls.

[5]Comparable to the best-know approaches.

[6]The minimum number of exchanges to successfully complete the dialogue is 5.

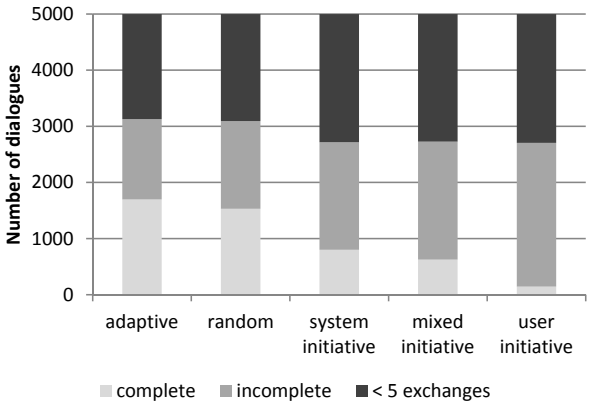

Figure 4: The ratio of omitted dialogues due to their length (< 5 exchanges), the completed dialogues (complete), and the dialogues which have been aborted by the user (incomplete) with respect to the dialogue strategy. While the amount of short dialogues is similar for each strategy, the number of completed dialogues varies strongly.

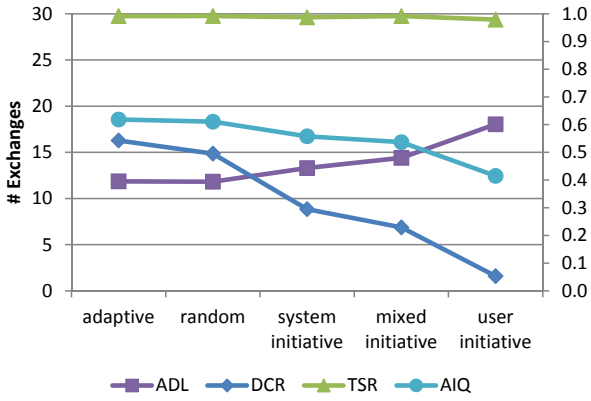

Figure 5: The average dialogue length (ADL), task success rate (TSR), the dialogue completion rate (DCR), and the average Interaction Quality (AIQ) for all for dialogue strategies. With decreasing DCR, also AIQ decreases and ADL increases. (AIQ values are normalized to the interval [0–1].)

value of the last exchange of each dialogue. Furthermore, this measure is also used to investigate if adapting the course of the dialogue to IQ also results in higher IQ values.

4.3 Experimental Results

Figure 4 shows the ratio of complete, incomplete, and omitted dialogues for each strategy with respect to the total 5,000 dialogues. As can be seen, about the same ratio of dialogues is omitted due to being too short. The DCR clearly varies more strongly for the five strategies.

The results for DCR, TSR, ADL, and AIQ are presented in Table 1 and Figure 5. TSR is almost the same for all strategies, meaning that, if a dialogue completes, the system almost always found the correct user goal. DCR, ADL and AIQ on the other hand vary strongly. They strongly correlate with a Pearson's correlation of $\rho = -0.953$ ($\alpha < 0.05$) for DCR and ADL, $\rho = 0.960$ ($\alpha < 0.01$) for DCR and AIQ, and $\rho = -.997$ ($\alpha < 0.01$) for ADL and AIQ. This shows that by improving IQ, being a subjective measure, an increase in objective measures may be expected.

Comparing the performance of the adaptive strategy to the three non-adaptive strategy clearly shows that the adaptive strategy performs significantly best for all metrics. With a DCR of of 54.27%, the performance is comparable to the rate achieved on the training data of LGUS (cf. (Lee and Eskenazi, 2012)). The non-adaptive strategies achieve a much lower DCR having the system initiative strategy as second best with only 29.48%. This performance goes together with shorter dialogues shown by the ADL. Furthermore, the results for DCR clearly show that the user initiative strategy is unusable. Thus, this strategy will not be analyzed any further.

Furthermore, it is of interest if better objective performance also results in better IQ values for the complete dialogue. This is especially important since it is imperative for the relevance of the Interaction Quality. Adapting to IQ to improve the dialogue must also result in an increase of the IQ value. This effect has been validated by these experiments. The adaptive strategy has a significant higher average IQ (AIQ) value calculated from the IQ value for the whole dialogues, i.e., the IQ value of the last system-user-exchange, than all other non-adaptive strategies.

The question remains if adapting to IQ is the actual reason for the improvement. Maybe, the user simulated with LGUS only "likes" diversified initiative prompts better which is represented by the random strategy. While this statement is true to some extent (see ADL), reasonably adapting to IQ further improves the system performance significantly as shown by DCR and AIQ.

5 Reward Modelling with Interaction Quality

The presented results clearly show that AIQ and DCR are correlated. As almost all completed di-

Strategy	DCR	TSR	ADL	AIQ
adaptive	54.27%	99.18%	11.86	3.47**
random	49.53%	99.22%	11.82**	3.44**
system initiative	29.48%	98.75%	13.30*	3.23
mixed initiative	22.91%	99.20%	14.40*	3.15**
user initiative	5.32%	97.92%	18.04	2.66

Table 1: The results of the experiments for the five strategies given by dialogue completion rate (DCR), task success rate (TSR), average dialogue length (ADL) and average Interaction Quality (AIQ) rating the complete interaction for all completed dialogues. All results for DCR and TSR are significantly different (chi-squared test). Significant differences in ADL (unpaired t-test) and AIQ (Mann-Whitney U test) with the respective column below are marked with ** for the level of $\alpha < 0.01$ and with * for $\alpha < 0.05$. All other comparisons between non-neighbors are significant with $\alpha < 0.01$

alogues were also successful, a correlation between AIQ and task success may also be assumed. In this section, we investigate if this correlation may be exploited for modelling the reward function for reinforcement learning approaches to dialogue management. This would be very beneficial, as for state-of-the-art reinforcement learning approaches to dialogue management, e.g., (Lemon and Pietquin, 2012; Young et al., 2013), a positive or negative reward is added at the end of each dialogue depending on the successful achievement of the task. However, to do this, usually, the true user goal has to be know. This is either possible by asking the user or by using a user simulator for training. Here, Gašić et al. have shown that optimizing the strategy with real user dialogues yields better strategies than using a user simulator. However, asking the user to provide whether they consider the dialogue to be successful is time consuming and interruptive thus only possible in artificial lab settings. If there was a metric which allowed to automatically detect successful, or, more generally, good dialogues, this metric would be very useful for the described situation yielding the opportunity to optimize on real dialogues without disrupting the users.

Therefore, the correlation of the final IQ value and task success is analyzed. Based on all strategies, the dialogues are evaluated regarding the success rate with respect to the final IQ value and the dialogue length. An example for dialogue lengths

DL	IQ	success	failure	# dialogues
9	1	0.0%	100.0%	487
	2	0.0%	100.0%	40
	3	37.2%	62.9%	253
	4	93.8%	6.3%	512
	5	0.0%	100.0%	2
10	1	0.0%	100.0%	452
	2	0.0%	100.0%	38
	3	42.4%	57.6%	172
	4	96.6%	3.5%	406
	5	0.0%	100.0%	3
11	1	0.0%	100.0%	405
	2	2.9%	97.1%	35
	3	47.8%	52.3%	178
	4	84.0%	16.0%	100
	5	-	-	0
12	1	0.3%	99.7%	329
	2	23.1%	76.9%	52
	3	78.5%	21.6%	297
	4	96.3%	3.7%	270
	5	0.0%	100.0%	1

Table 2: Example of the task success rate with respect to IQ and the dialogue length (DL). Disregarding rows with less then 15 dialogues, there is clearly a trend for higher task success rates if the IQ value increases as well.

of 9–12 is depicted in Table 2. To compute those, again, dialogues with less than five exchanges are excluded. Clearly, a trend can be identified for higher task success rates when having a high final IQ for all dialogue lengths[7].

Based on this finding, an IQ threshold may be defined which separates dialogues regarded as being successful and dialogues regarded as being not successful. For a threshold of four, for example, all dialogues with a final IQ of five and four may be regarded as successful while all other dialogues are regarded as failure. However, not all dialogues above the threshold are necessarily actually successful and not all dialogues below the threshold are necessarily actually unsuccessful. Hence, to find a good threshold, the precision—representing this relationship—is calculated for both success and failure dialogues for different thresholds. The results are depicted in Table 3.

The best overall threshold indicated by a maximum unweighted average precision[8] (UAP) is four achieving a precision of 0.863 for success and of 0.826 for failure. While a threshold of four is also

[7]Rows with less than 15 dialogues are disregarded as sufficient data is needed to compute reasonable task success rates.

[8]The arithmetic average over all class-wise precisions.

Success IQ ≥	Precision Success	Failure	UAP
5	0.448	0.669	0.559
4	0.863	0.826	0.845
3	0.652	0.888	0.770
2	0.646	0.995	0.820
1	0.331	-	0.166

Table 3: The precision of success and failure dialogues (along with the unweighted average precision (UAP)) when setting all dialogue with final IQ greater or equal a given IQ value to be successful and the remainder to be a failure.

Success IQ ≥	Recall Success	Failure	UAR
5	0.008	0.995	0.502
4	0.595	0.953	0.774
3	0.798	0.789	0.794
2	0.992	0.730	0.861
1	1.000	-	0.500

Table 4: The recall of success and failure dialogues (along with the unweighted average recall (UAR)) when setting all dialogue with final IQ greater or equal a given IQ value to be successful and the remainder to be a failure.

the best threshold for success, the highest precision for failure is a threshold of two, i.e., regarding all dialogues as being a failure with a final IQ of one. Hence, to further maximize UAP, two thresholds may be defined: four for success and two for failure. This results in an UAP of 0.929 not regarding all dialogues with a final IQ of two or three.

Defining a threshold based on precision yields the downside that some actually successful dialogues are regarded as failure and vice versa. In fact, defining a threshold of four results in a recall—representing the percentage of dialogues being regarded as successful out of all truly successful dialogues—of 0.595 as shown in Table 4. This means that more than 40% of all truely successful dialogues are regarded as failure which is not ideal. Additionally, a recall of 0.953 for failure means that less than 5% of all truely failing dialogues are regarded as success. However, using the two thresholds defined above results in better rates. Still, 4.7% of all failing dialogues are regared as success. However, only 0.8% of all successful dialogues are regarded as failure which is much better. Having two thresholds, though, results in the need for more training dialogues as all dialogues between the two thresholds are obmitted: only 64% of all dialogues are used for training resulting in the need for 56% more dialogues for training.

6 Conclusion and Future Work

The contribution of this work is two-fold: first, we analyzed the performance of an adaptive dialogue strategy adapting the dialogue initiative to Interaction Quality and answered the question if IQ and objective measures correlate in such a setting. By comparing five different strategies, we could show that the dialogue completion rate, the average dialogue length, and the average interaction quality strongly correlate. In addition, we could show that the adaptive strategy clearly outperforms all non-adaptive strategies as well as the random strategy. Hence, not only the grounding strategy but also the dialogue initiative is suitable for rule-based quality-adaptive dialogue.

Second, we performed a more detailed analysis of the correlation of task success and Interaction Quality showing that by defining IQ thresholds separating dialogues regarded as success and failure is a reasonable approach achieving an unweighted average precision of 0.929. This is of special interest for reinforcement learning where this could be used to automatically detect task success. However, not all dialogues could be used for training the dialogue strategy resulting in the need for 56% more dialogues. Moreover, the effects on the resulting strategy of regarding dialogues which are truly failing as successful (in the sense of keeping the user satisfied) is unclear and must be anlyzed in a further study performing reinforcement learning with the proposed method.

For future work on quality-adaptive dialogue, the same adaptation techniques should be tested with real users. While user simulators offer a good means of evaluating dialogues easily, real users usually give new insight by showing unseen behavior. Furthermore, other adaptation mechanisms may be applied, e.g., in a statistical dialogue management setting (Ultes et al., 2011).

Acknowledgments

This work was supported by the Transregional Collaborative Research Centre SFB/TRR 62 “Companion-Technology for Cognitive Technical Systems” which is funded by the German Research Foundation (DFG). Additionally, we would like to thank Sungjin Lee and Maxine Eskenazi for providing access to the Let’s Go User Simulator.

References

Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

Douglas Crockford. 2006. RFC 4627 - The application/json Media Type for JavaScript Object Notation (JSON). Technical report, IETF, July.

Klaus-Peter Engelbrecht, Florian Gödde, Felix Hartard, Hamed Ketabdar, and Sebastian Möller. 2009. Modeling user satisfaction with hidden markov model. In *SIGDIAL '09: Proceedings of the SIGDIAL 2009 Conference*, pages 170–177, Morristown, NJ, USA. Association for Computational Linguistics.

Milica Gačić, Catherine Breslin, Matthew Henderson, Dongho Kim, Martin Szummer, Blaise Thomson, Pirros Tsiakoulis, and Steve J. Young. 2013. Online policy optimisation of bayesian spoken dialogue systems via human interaction. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8367–8371. IEEE.

Milan Gnjatović and Dietmar Rösner. 2008. Adaptive dialogue management in the nimitek prototype system. In *PIT '08: Proceedings of the 4th IEEE tutorial and research workshop on Perception and Interactive Technologies for Speech-Based Systems*, pages 14–25, Berlin, Heidelberg. Springer-Verlag.

Sunao Hara, Norihide Kitaoka, and Kazuya Takeda. 2010. Estimation method of user satisfaction using n-gram-based dialog history model for spoken dialog system. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May. European Language Resources Association (ELRA).

Tobias Heinroth, Dan Denich, and Alexander Schmitt. 2010. Owlspeak - adaptive spoken dialogue within intelligent environments. In *IEEE PerCom Workshop Proceedings*, March. presented as part of SmartE Workshop.

Ryuichiro Higashinaka, Yasuhiro Minami, Kohji Dohsaka, and Toyomi Meguro. 2010. Issues in predicting user satisfaction transitions in dialogues: Individual differences, evaluation criteria, and prediction models. In Gary Lee, Joseph Mariani, Wolfgang Minker, and Satoshi Nakamura, editors, *Spoken Dialogue Systems for Ambient Environments*, volume 6392 of *Lecture Notes in Computer Science*, pages 48–60. Springer Berlin / Heidelberg. 10.1007/978-3-642-16202-2_5.

Srinivasan Janarthanam and Oliver Lemon. 2008. User simulations for online adaptation and knowledge-alignment in troubleshooting dialogue systems. *Semantics and Pragmatics of Dialogue (LONDIAL)*, page 45.

Sungjin Lee and Maxine Eskenazi. 2012. An unsupervised approach to user simulation: toward self-improving dialog systems. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 50–59. Association for Computational Linguistics, July.

Oliver Lemon and Olivier Pietquin. 2012. *Data-Driven Methods for Adaptive Spoken Dialogue Systems*. Springer New York.

Diane J. Litman and Shimei Pan. 2002. Designing and evaluating an adaptive spoken dialogue system. *User Modeling and User-Adapted Interaction*, 12(2-3):111–137.

Michael F. McTear. 2004. *Spoken Dialogue Technology: Towards the Conversational User Interface*. Springer, London.

Florian Nothdurft, Frank Honold, and Peter Kurzok. 2012. Using explanations for runtime dialogue adaptation. In *Proceedings of the 14th ACM International Conference on Multimodal Interaction*, pages 63–64. ACM, October.

M. Oshry, R. Auburn, P. Baggia, M. Bodell, D. Burke, D. Burnett, E. Candell, J. Carter, S. Mcglashan, A. Lee, B. Porter, and K. Rehor. 2007. Voice extensible markup language (voicexml) version 2.1. Technical report, W3C - Voice Browser Working Group, June.

Antoine Raux, Dan Bohus, Brian Langner, Alan W. Black, and Maxine Eskenazi. 2006. Doing research on a deployed spoken dialogue system: One year of let's go! experience. In *Proc. of the International Conference on Speech and Language Processing (ICSLP)*, September.

Verena Rieser and Oliver Lemon. 2008. Learning effective multimodal dialogue strategies from wizard-of-oz data: Bootstrapping and evaluation. In *ACL*, pages 638–646.

Alexander Schmitt and Stefan Ultes. 2015. Interaction quality: Assessing the quality of ongoing spoken dialog interaction by expertsand how it relates to user satisfaction. *Speech Communication*.

Alexander Schmitt, Benjamin Schatz, and Wolfgang Minker. 2011. Modeling and predicting quality in spoken human-computer interaction. In *Proceedings of the SIGDIAL 2011 Conference*, pages 173–184, Portland, Oregon, USA, June. Association for Computational Linguistics.

Alexander Schmitt, Stefan Ultes, and Wolfgang Minker. 2012. A parameterized and annotated spoken dialog corpus of the cmu let's go bus information system. In *International Conference on Language Resources and Evaluation (LREC)*, pages 3369–337, May.

Stefan Ultes and Wolfgang Minker. 2013. Interaction quality: A review. *Bulletin of Siberian State Aerospace University named after academician M.F. Reshetnev*, (4):153–156.

Stefan Ultes and Wolfgang Minker. 2014. Managing adaptive spoken dialogue for intelligent environments. *Journal of Ambient Intelligence and Smart Environments*, 6(5):523–539, August.

Stefan Ultes, Tobias Heinroth, Alexander Schmitt, and Wolfgang Minker. 2011. A theoretical framework for a user-centered spoken dialog manager. In *Proceedings of the Paralinguistic Information and its Integration in Spoken Dialogue Systems Workshop*, pages 241 – 246. Springer, September.

Stefan Ultes, Alexander Schmitt, and Wolfgang Minker. 2012. Towards quality-adaptive spoken dialogue management. In *NAACL-HLT Workshop on Future directions and needs in the Spoken Dialog Community: Tools and Data (SDCTD 2012)*, pages 49–52, Montréal, Canada, June. Association for Computational Linguistics.

Stefan Ultes, Alexander Schmitt, and Wolfgang Minker. 2013. On quality ratings for spoken dialogue systems – experts vs. users. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 569–578. Association for Computational Linguistics, June.

Stefan Ultes, Hüseyin Dikme, and Wolfgang Minker. 2014a. Dialogue management for user-centered adaptive dialogue. In *Proceedings of the 5th International Workshop On Spoken Dialogue Systems (IWSDS)*, January.

Stefan Ultes, Hüseyin Dikme, and Wolfgang Minker. 2014b. First insight into quality-adaptive dialogue. In *International Conference on Language Resources and Evaluation (LREC)*, pages 246–251, May.

Vladimir N. Vapnik. 1995. *The nature of statistical learning theory*. Springer-Verlag New York, Inc., New York, NY, USA.

Marilyn Walker, Diane J. Litman, Candace A. Kamm, and Alicia Abella. 1997. Paradise: a framework for evaluating spoken dialogue agents. In *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*, pages 271–280, Morristown, NJ, USA. Association for Computational Linguistics.

Marilyn Walker. 2000. An application of reinforcement learning to dialogue strategy selection in a spoken dialogue system for email. *Journal of Artificial Intelligence Research*, 12:387–416.

Steve J. Young, Milica Gačić, Blaise Thomson, and Jason D. Williams. 2013. Pomdp-based statistical spoken dialog systems: A review. *Proceedings of the IEEE*, 101(5):1160–1179.