Towards Quality-Adaptive Spoken Dialogue Management

Stefan Ultes, Alexander Schmitt, Wolfgang Minker
Dialogue Systems - Ulm University
Albert-Einstein-Allee 43
89081 Ulm, Germany
{stefan.ultes,alexander.schmitt,wolfgang.minker}@uni-ulm.de

Abstract

Information about the quality of a Spoken Dialogue System (SDS) is usually used only for comparing SDSs with each other or manually improving the dialogue strategy. This information, however, provides a means for inherently improving the dialogue performance by adapting the Dialogue Manager during the interaction accordingly. For a quality metric to be suitable, it must suffice certain conditions. Therefore, we address requirements for the quality metric and, additionally, present approaches for quality-adaptive dialogue management.

1 Introduction

For years, research has been focused on enabling Spoken Dialogue Systems (SDSs) to behave more adaptively to the user's expectations and needs. Möller et al. (2009) presented a taxonomy for quality of human-machine interaction, i.e., Quality of Service (QoS) and Quality of Experience (QoE). For QoE, several aspects are identified. They contribute to good user experience, e.g., interaction quality, usability and acceptability. These aspects can be combined to the term User Satisfaction (US), describing the degree by which the user is satisfied with the system's performance. The dialogue community has been investigating this aspect for years. Most prominently is the PARADISE framework by Walker et al. (2000) which maps objective performance metrics of an SDS to subjective user ratings.

Recent work mostly discusses how to evaluate Spoken Dialogue Systems. However, the issue of how this information can be useful for improving dialogue performance remains hardly addressed. Hence, we focus on exploring techniques for incorporating dialogue quality information into the Dialogue Manager (DM). This is accompanied by the problem of defining characteristics of a suitable dialogue quality metric.

In Section 2, we present related work both on measuring dialogue quality and on approaches for incorporating user state information into the DM. In Section 3, requirements for a quality metric are presented along with a suitable example. Section 4 presents our ongoing and future work on incorporating quality measures into dialogue strategies. Finally, Section 5 concludes this work.

2 Related Work

In recent years, several studies have been published on determining the qualitative performance of a SDS. Engelbrecht et al. (2009) predicted User Satisfaction on a five-point scale at any point within the dialogue using Hidden Markov Models (HMMs). Evaluation was based on labels the users applied themselves during a Wizard-of-Oz experiment. To guarantee for comparable conditions, the dialogue flow was controlled by predefined scenarios creating transcripts with equal length for each scenario.

Further work based on HMMs was presented by Higashinaka et al. (2010). The HMM was trained on US rated at each exchange. These exchange ratings were derived from ratings for the whole dialogue. The authors compare their approach with HMMs trained on manually annotated exchanges achieving a better performance for the latter.

NAACL-HLT 2012 Workshop on Future directions and needs in the Spoken Dialog Community: Tools and Data, pages 49–52,
Montréal, Canada, June 7, 2012.

In order to predict US, Hara et al. (2010) created n-gram models from dialogue acts (DA). Based on dialogues from real users interacting with a music retrieval system, overall ratings for the whole dialogue have been labeled on a five point scale after the interaction. An accuracy (i.e., rate of correctly predicted ratings) of 34% by a 3-gram model was the best performance which could be achieved.

Dealing with true User Satisfaction, Schmitt et al. presented their work about statistical classification methods for automatic recognition of US (Schmitt et al., 2011b). The data was collected in a lab study where the users themselves had to rate the conversation during the ongoing dialogue. Labels were applied on a scale from 1 to 5. Performing automatic classification using a Support Vector Machine (SVM), they achieved an Unweighted Average Recall (UAR) of 49.2 (i.e., average rate of correctly predicted ratings, compensated for unbalanced data).

An approach for affective dialogue modeling based on Partially Observable Markov Decision Processes (POMDPs) was presented by Bui et al. (2007). Adding stress to the dialogue state enables the dialogue manager to adapt to the user. To make belief-update tractable, the authors introduced Dynamic Decision Networks as means for reducing complexity.

Pittermann et al. (2007) presented another approach for adaptive dialogue management. The authors incorporated emotions by modeling the dialogue in a semi-stochastic way. Thus, an emotional dialogue model was created as a combination of a probabilistic emotional model and probabilistic dialogue model defining the current dialogue state.

3 Interaction Quality Metric

In order to enable the Dialogue Manager to be quality-adaptive, the quality metric must suffice certain criteria. In this Section, we identify the important issues and render the requirements for a suitable quality metric.

3.1 General Aspects

For adapting the dialogue strategy to the quality of the dialogue, the quality metric is required to implement certain characteristics. We identify the following items:

- exchange-level quality measurement,
- automatically derivable features,
- domain-independent features,
- consistent labeling process,
- reproducible labels and
- unbiased labels.

The performance of a Spoken Dialogue System may be evaluated either on the dialogue level or on the exchange level. As dialogue management is performed after each system-user exchange, dynamic adaption of the dialogue strategy to the dialogue performance requires exchange-level performance measures. Therefor, Dialogue-level approaches are of no use. Furthermore, previous presented methods for exchange-level quality measuring could not achieve satisfying accuracy in predicting dialogue quality (Engelbrecht et al., 2009; Higashinaka et al., 2010).

Features serving as input variables for a classification algorithm must be automatically derivable from the dialogue system modules. This is important because other features, e.g., manually annotated dialogue acts (Higashinaka et al., 2010; Hara et al., 2010), produce high costs and are also not available immediately during run-time in order to use them as additional input to the Dialogue Manager. Furthermore, for creating a *general* quality metric, features have to be domain-independent, i.e., not depending on the task domain of the dialogue system.

Another important issue is the consistency of the labels. Labels applied by the users themselves are subject to large fluctuations among the different users (Lindgaard and Dudek, 2003). As this results in inconsistent labels, which do not suffice for creating a generally valid quality model, ratings applied by expert raters yield more consistent labels. The experts are asked to estimate the user's satisfaction following previously established rating guidelines. Furthermore, expert labelers are also not prone to be influenced by certain aspects of the SDS, which are not of interest in this context, e.g., the character of the synthesized voice. Therefore, they create less biased labels.

3.2 Interaction Quality

As metric, which fulfills all previously addressed requirements, we present the Interaction Quality (IQ) metric, see also (2011a). Based on dialogues from the "Let's Go Bus Information System" of the Carnegie Mellon University in Pittsburgh (Raux et al., 2006), IQ is labeled on a five point scale. The labels are (from best (5) to worst (1)) "satisfied", "slightly unsatisfied", "unsatisfied", "very unsatisfied" and "extremely unsatisfied". They are applied by expert raters following rating guidelines, which have been established to allow consistent and reproducible ratings.

Additionally, domain-independent features used for IQ recognition have been derived from the dialogue system modules automatically for each exchange grouped on three levels: the *exchange level*, the *dialogue level*, and the *window level*. As parameters like ASRCONFIDENCE or UTTERANCE can directly be acquired from the dialogue modules they constitute the *exchange level*. Based on this, counts, sums, means, and frequencies of *exchange level* parameters from multiple exchanges are computed to constitute the *dialogue level* (all exchanges up to the current one) and the *window level* (the three previous exchanges).

A corpus containing the labeled data has been published recently (Schmitt et al., in press) containing 200 calls annotated by three expert labelers, resulting in a total of 4,885 labeled exchanges. Using statistical classification of IQ based on SVMs achieves an Unweighted Average Recall of 0.58 (Schmitt et al., 2011a).

4 Quality-Adaptive Spoken Dialogue Management

The goal of our work is to enable Dialogue Managers to directly adapt to information about the quality of the ongoing dialogue. We present two different approaches that outline our ongoing and future work.

4.1 Dialogue Design-Patterns for Quality Adaption

Rule-based Dialogue Managers are still state-of-the-art for commercial SDSs. It is hardly arguable that making the rules quality-dependent is a promising way for dialogue improvement. However, the number of possibilities for adapting the dialogue strategy to the dialogue quality is high. Based on the SpeechCycle RPA Dialogue Manager, we are planning on identifying common dialogue situations in order to create design-patterns. These patterns can be applied as a general means of dealing with situations that arise by introducing quality-adaptiveness to the dialogue.

4.2 Statistical Quality-Adaptive Dialogue Management

For the incorporation of Interaction Quality into a statistical DM, two approaches have been found.

First, based on work on factored Partially Observable Markov Decision Processes by Williams and Young (2007) and similar to Bui et al. (2006), we presented our own approach for incorporating additional user state information (Ultes et al., 2011).

In the factored POMDP by Williams and Young (2007), the state of the underlying process is defined as $s = (u, g, h)$. To incorporate IQ, it is extended by adding the IQ-state s_{iq}, resulting in $s = (u, g, h, s_{iq})$.

Following the concept of user acts, we further introduce IQ-acts iq that describe the current quality predicted by the classification algorithm for the current exchange. Incorporating IQ acts into observation o results in the two-dimensional observation space

$$O = U \times IQ,$$

where U denotes the set of all user actions and IQ the set of all possible Interaction Quality values.

Second, for training an optimal policy for action selection in POMDPs, a reward function has to be defined. Common reward functions are task-oriented and based on task success and dialogue length. As an example, a considerable positive reward is given for reaching the task goal, a considerable negative reward for aborting the dialogue, and a small negative reward for each exchange in order to keep the dialogue short. Interaction Quality scores offer an interesting and promising way of defining a reward function, e.g., by rewarding improvements in IQ. By that, strategies that try to keep the quality at an overall high can be trained allowing for a better user experience.

5 Conclusion

For incorporating information about the dialogue quality into the Dialogue Manager, we identified characteristics of a quality metric defining necessary prerequisites for being used during dialogue management. Further, the Interaction Quality metric has been proposed as measure, which suffices all requirements. In addition, we presented concrete approaches of incorporating IQ into the DM outlining our ongoing and future work.

Acknowledgements

We would like to thank Maxine Eskenazi, Alan Black, Lori Levin, Rita Singh, Antoine Raux and Brian Langner from the Lets Go Lab at Carnegie Mellon University, Pittsburgh, for providing the Lets Go Sample Corpus. We would further like to thank Roberto Pieraccini and David Suendermann from SpeechCycle, Inc., New York, for providing the SpeechCycle RPA Dialogue Manager.

References

T. H. Bui, J. Zwiers, M. Poel, and A. Nijholt. 2006. Toward affective dialogue modeling using partially observable markov decision processes. In *Proceedings of workshop emotion and computing, 29th annual German conference on artificial intelligence*.

T. H. Bui, M. Poel, A. Nijholt, and J. Zwiers. 2007. A tractable ddn-pomdp approach to affective dialogue modeling for general probabilistic frame-based dialogue systems. In *Proceedings of the 5th IJCAI Workshop on Knowledge and Reasoning in Practical Dialogue Systems*, pages 34–37.

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. ACL.

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 *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may. ELRA.

Ryuichiro Higashinaka, Yasuhiro Minami, Kohji Dohsaka, and Toyomi Meguro. 2010. Modeling user satisfaction transitions in dialogues from overall ratings. In *Proceedings of the SIGDIAL 2010 Conference*, pages 18–27, Tokyo, Japan, September. Association for Computational Linguistics.

Gitte Lindgaard and Cathy Dudek. 2003. What is this evasive beast we call user satisfaction? *Interacting with Computers*, 15(3):429–452.

Sebastian Möller, Klaus-Peter Engelbrecht, C. Kühnel, I. Wechsung, and B. Weiss. 2009. A taxonomy of quality of service and quality of experience of multimodal human-machine interaction. In *Quality of Multimedia Experience, 2009. QoMEx 2009. International Workshop on*, pages 7–12, July.

Johannes Pittermann, A. Pittermann, Hong Meng, and W. Minker. 2007. Towards an emotion-sensitive spoken dialogue system - classification and dialogue modeling. In *Intelligent Environments, 2007. IE 07. 3rd IET International Conference on*, pages 239 –246, September.

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

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

Alexander Schmitt, Benjamin Schatz, and Wolfgang Minker. 2011b. A statistical approach for estimating user satisfaction in spoken human-machine interaction. In *Proceedings of the IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT)*, Amman, Jordan, December. IEEE.

Alexander Schmitt, Stefan Ultes, and Wolfgang Minker. in-press. A parameterized and annotated corpus of the cmu let's go bus information system. In *International Conference on Language Resources and Evaluation (LREC)*.

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.

Marilyn Walker, Candace Kamm, and Diane Litman. 2000. Towards developing general models of usability with paradise. *Nat. Lang. Eng.*, 6(3-4):363–377.

Jason D. Williams and Steve J. Young. 2007. Partially observable markov decision processes for spoken dialog systems. *Computer Speech and Language*, (21):393–422.