УДК 004.5

INTERACTION QUALITY: A REVIEW

S. Ultes, W. Minker

Ulm University
43, Albert-Einstein-Allee, Ulm, 89081, Germany
E-mail: stefan.ultes@uni-ulm.de, wolfgang.minker@uni-ulm.de

Automatically deriving the quality of a Spoken Dialogue System is an important task for both assessing dialogue systems and improving them. Work on automatic quality estimation for each system-user-exchange further holds the opportunity of using this quality information for online-adaption of the dialogues. The Interaction Quality paradigm is the first metric holding those features. Hence, this contribution gives an overview over the Interaction Quality paradigm and reviews recent estimation approaches. Furthermore, it renders drawbacks of the current approaches and proposes further directions in order to improve the estimation accuracy.

Keywords: spoken dialogue systems, dialogue assessment, machine learning.

КАЧЕСТВО ВЗАИМОДЕЙСТВИЯ: ОБЗОР

Ш. Ультес, В. Минкер

Университет города Ульма
Германия, 89081, Ульм, Аллея Альберта Эйнштейна, 43
E-mail: stefan.ultes@uni-ulm.de, wolfgang.minker@uni-ulm.de

Автоматическое извлечение качества речевой диалоговой система является важной задачей как для оценки диалоговых систем, так и для их улучшения. Работа автоматической оценки качества для каждого обмена между системой и пользователем даёт возможность использования информации о качестве для адаптации диалога в режиме реального времени. Парадигма качества взаимодействия – это первая метрика, которая может содержать такие свойства. Дается обзор парадигмы качества взаимодействия и существующих подходов к оценке качества. Далее рассматриваются недостатки существующих подходов и предлагаются направления для дальнейшего улучшения процесса оценки качества взаимодействия.

Ключевые слова: речевая диалоговая система, оценка диалога, машинное обучение.

Spoken Dialogue Systems (SDSs) play a key role in designing a human-machine interface to be natural as speech is one of the major channels of human communication. Assessing the quality of such SDSs has been discussed controversially in recent years. First work on deriving subjective metrics automatically has been performed by Walker et al. [1] resulting in the PARADISE framework, which is the current quasi-standard in this field. Briefly explained, a linear dependency is assumed between dialogue parameters and user satisfaction to estimate qualitative performance on the dialogue level.

As PARDIES does not allow for exchange-level quality measures, Schmitt et al. introduced a new paradigm called Interaction Quality (IQ) [2]. Based on interaction parameters, a statistical model is created to derive IQ automatically for each exchange. By that, it is possible to use the quality information for automatic adaption in spoken human-machine dialogues [3]. In this contribution, we will present an overview of work using IQ recognition. Furthermore, we will discuss the different approaches and propopse futer directions.

In Section 2, the initial work by Schmitt et al. is presented defining Interaction Quality. Following that, Section 3 presents further work on automatic recognition of Interaction Quality including a controverse discussion. Finally, Section 4 concludes by outlining important futur directions.

Interaction Quality. Information about the quality of the interaction between humans and an SDS may be used for several purposes. Besides using it to compare different systems, it may also be used for improving the dialogue design itself. PARADISE provides quality values on the dialogue level which allows for general optimization of the dialogue in an offline fashion. Unfortunately, this paradigm is not usable for online dialogue optimization where the dialogue system adapts to the current quality of the dialogue. Ultes et al. [3] identified several requirements for a quality metric to be suitable for the task of automatically deriving the quality of an ongoing dialogue using statistical classification approaches. Among those requirements are, e.g., exchange level quality measurement and automatically derivable features. The Interaction Quality (IQ) paradigm introduced by Schmitt et al. [2] offers both. It is based on features which are derived from the three dialogue system modules Automatic Speech Recognition (ASR), Natural Language Understanding (NLU), and Dialogue Management (DM).

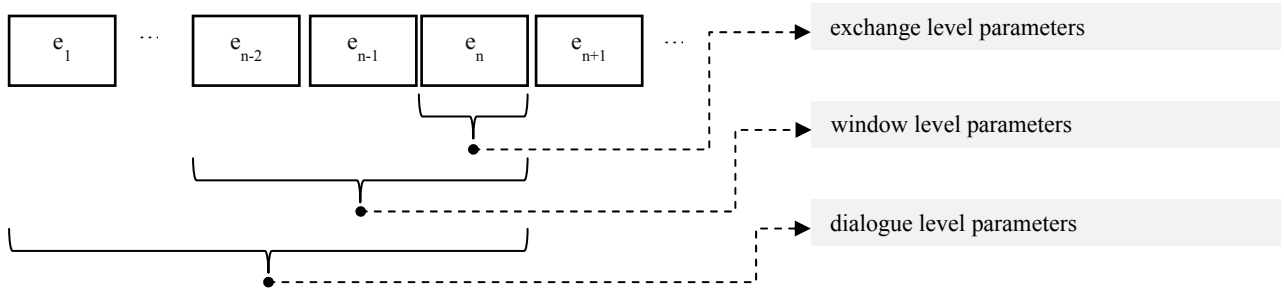

Fig. 1. The three different levels of interaction parameters

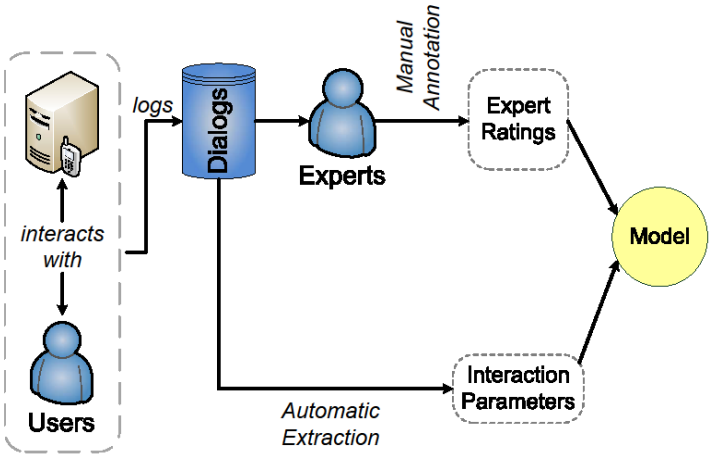

Fig. 2. Process for Interaction Quality estimation according to Schmitt et al. [2]

Parameters from the ASR module are, e.g., the confidence value, from the NLU module the semantic interpretation of the user input, and from the DM module information about the system action being a reprompts or an attempt to elicit common ground.

Moreover, these interaction parameters are designed on three levels: the exchange level, comprising information about the current system-user-exchange, the dialogue level, comprising information about the complete dialogue up to the current exchange, and the window level, comprising information about the n last exchanges. This is illustrated in Figure 1. A complete list of features can be found in [4].

Schmitt et al. [2] tackled the problem of automatic estimation of the IQ using a Support Vector Machine (SVM) [5] as statistical classification algorithm. Hence, they regard the problem as estimating five independent classes. The general process applied by Schmitt et al. for IQ classification is illustrated in fig. 2.

Dialogues between users and a dialogue system are recorded and logged. These logs are then analyzed by experts who apply quality ratings manually for each system user exchange on a scale from five (satisfied) to one (extremely unsatisfied). Furthermore, each exchange is annotated by three different raters. By following labeling guidelines, a certain degree of consistency between the raters is achieved still allowing enough freedom for individual ratings. For each exchange, a final rating is calculated by taking the median of the three expert ratings. The recordings, the logs of the dialogues and the corresponding labels have been published under the title LEGO corpus [4].

Using the final quality rating as target variable for the classifier and using the previously presented interaction parameter as features, Schmitt et al. achieved an unweighted average recall (UAR) of 0.59. The UAR is defined as the arithmetic average of all class-wise recalls thus eliminating the effects of unbalanced data.

While it can be argued that user ratings should be preferred over expert ratings as only real users of the system can truly give an opinion about its quality, asking user directly holds some drawbacks as they can be considered to be collected more expensively. Furthermore, expert and user ratings are quite similar so that expert ratings can easily function as a good replacement for user ratings, cf. Ultes et al. [6].

Approaches on Interaction Quality Recognition. Schmitt et al. are the first to achieve acceptable results for quality estimation of SDSs on the exchange-level. Using their ground work, further approaches have been investigated analyzing more aspects of IQ recognition and pursuing an improvement of interaction quality.

Markovian Approaches. The approach by Schmitt et al. has a major drawback: there, all exchanges are considered to be independent of each other. However, there is a temporal link between the exchanges of one dialogue. To overcome this, Ultes et al. [7] replaced the SVM with two classification models inherently taking into account temporal dependencies.

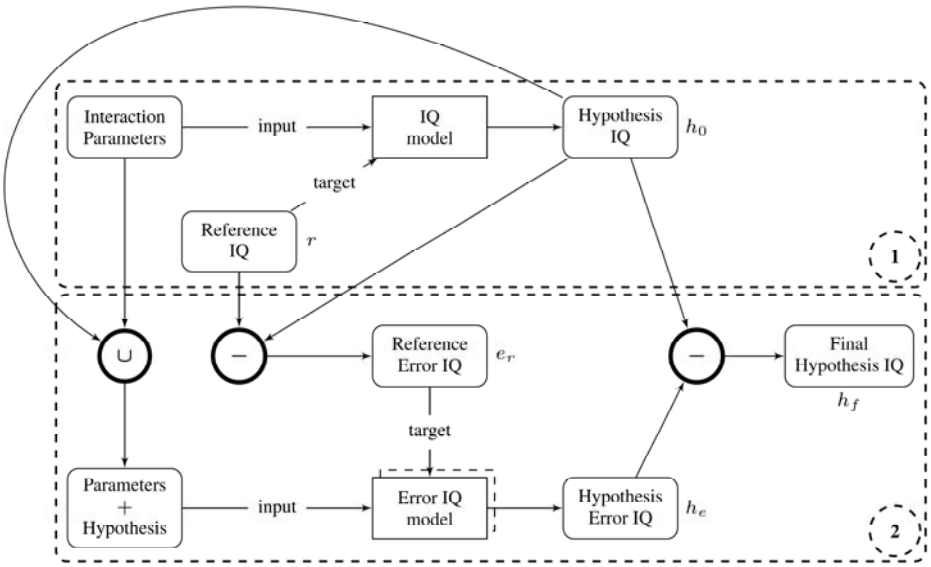

Fig. 3. Two-stage IQ classification using error correction

They applied a Hidden Markov Model (HMM) and a Conditioned Hidden Markov Model (CHMM) for estimating the target variable Interaction Quality based on interaction parameters. For applying a multi-class problem for the HMM, a model was instantiated consisting of five hidden states where each state was statically linked to one of the five quality classes. As the CHMM inherently provides class probabilities, no linking was necessary. Optimizing the state number resulted in a CHMM with nine hidden states. Both models used GMMs for modeling the observation probability. Unfortunately, both algorithms could not outperform the SVM baseline achieving only an UAR of 0.44 for the HMM and 0.39 for the CHMM.

A further approach by Ultes et al. exploiting the temporal character of the exchanges resulted in the Hybrid HMM approach. Here, they used static classifiers like an SVM or a Rule Learner trained in accordance to Schmitt et al. However, along with the classification results, also a confidence score for each quality class was computed. Adopting the HMM approach mentioned above, they used these confidence scores as observation probability achieving an improvement compared to plain static classification of up to 5% relative improvement.

Hierarchical approaches.

A total different approach to Interaction Quality Recognition has been presented by Ultes et al. rendering IQ recognition in two stages (figure 3).. In stage one, regular IQ recondition is performed. The resulting hypotheses are then compared with the reference IQ values and an error is calculated. A second classification model is then used in stage two targeting the error. There, the hypothesis of stage one is used as additional feature. They achieved relative improvement of SVM classification of 4.1 % in UAR and 0.5 % for Rule Induction. Again, the absolute performance for Rule Induction was much better than for SVM classification. ***Discussion.*** It is notable that for all tested approaches, Rule Induction performs best. Superficially, this might indicate that the problem is not very hard and therefore simple rule defining is sufficient. If we take a deeper look, the situation is different though. Rule Induction results in a big number of rules: In an example applying 6-flod cross-validation, 311 rules cover 5312 exchanges averaged per fold. This results in an average of 17 samples handled with one rule. Moreover, only 22 rules per fold cover more than or equal to 30 exchanges while 289 cover less than 30 exchanges. This shows that there are a high number of rules which cover outliers. One reason why this is not covered by the SVM might be that during SVM training, those are “pruned” for generalization reasons.

Analyzing the rules in order to get a better view on IQ itself does not reveal any new information. The only two conclusions which can be drawn are, first, that exchanges at the end of the dialogue (long dialogue duration) and exchanges with a lot of preceding reprompts have generally a low IQ value. Second, exchanges belonging to the beginning of a dialogue, which have little preceding reprompts, or which have a high ASR confidence have a good IQ value in general.

An important issue regarding the Interaction Quality paradigm lies in the data. Until now, only one dataset has been analyzed based on Let’s Go. Furthermore, it consists of many domain-specific parameters. While some of the more recent approaches removed parameters which strongly depend on the domain, the resulting set is still not completely domain independent. In order to be able to establish IQ as a general tool for evaluating and adapting Spoken Dialogue Systems, it is imperative to extend its analysis to other dialogue domains.

Conclusion. In this paper, we presented a review of the Interaction Quality metric for assessing Spoken Dialogue Systems on the exchange-level. While several approaches have already been tried, still some shortcomings exist. First of all, the selection of parameter is very domain-specific. While recent work removed most of those parameters, it still has to be shown that the IQ metric also works for other domains. Furthermore, while previous work regarded the problem as a classification task, the

suitability of applying regression methods for IQ estimation may be analyzed regarding it as estimating a continuous mathematical function. Finally, Conditioned Random Fields have shown to work well for sequence tagging. As this is related applying an IQ value to each exchange of a sequence of exchanges, those may also increase IQ recognition performance.

References

1. Walker M., Litman D., Kamm C. A., Abella A. PARADISE: a framework for evaluating spoken dialogue agents, in *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*, Morristown, NJ, USA, 1997.

2. Schmitt A., Schatz B., Minker W. MODELING AND PREDICTING QUALITY IN SPOKEN HUMAN-COMPUTER INTERACTION, in *Proceedings of the SIGDIAL 2011 Conference*, Portland, Oregon, USA, 2011.

3. Ultes S., Schmitt A., Minker W. 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)*, Montréal, Canada, 2012.

4. Schmitt A., Ultes S., Minker W. A Parameterized and Annotated Corpus of the CMU Let's Go Bus Information System, in *International Conference on Language Resources and Evaluation (LREC)*, 2012.

5. Vapnik V. N. The nature of statistical learning theory, New York, NY, USA: Springer-Verlag New York, Inc., 1995.

6. Ultes S., Schmitt A., Minker W. 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*, to appear.

7. Ultes S., ElChabb R., Minker W. Application and Evaluation of a Conditioned Hidden Markov Model for Estimating Interaction Quality of Spoken Dialogue Systems, in *Proceedings of the 4th International Workshop on Spoken Language Dialog System (IWSDS)*, 2012.

УДК 004.93

ON THE CONCEALMENT OF TRANSMISSION ERRORS FOR DISTRIBUTED SPEECH RECOGNITION SYSTEMS

D. Zaykovskiy

Ulm University
43, Albert-Einstein-Allee, Ulm, 89081, Germany. E-mail: dmitry@Zaykovskiy.de

The client-server speech recognition systems face the challenge to provide consistent performance over diverse channel conditions. It is therefore necessary to develop methods which could anticipate the effect of the transmission errors. In this paper we consider an error mitigation approach which does not modify the original data; instead it tries to reconstruct lost information at the receiver via interpolation of successfully transmitted features. Using the packet identification number the DSR server is able to decide unambiguously which packets were lost and which were closest packets received without error. With correctly received packets before and after the burst, error mitigation module can interpolate missing features.

Keywords: distributed speech recognition, transmission error mitigation, interpolation.

О СОГЛАСОВАНИИ ОШИБОК ПЕРЕДАЧИ В РАСПРЕДЕЛЕННЫХ СИСТЕМАХ РАСПОЗНАВАНИЯ РЕЧИ

Д. Зайковский

Университет города Ульма
Германия, 89081, Ульм, Аллея Альберта Эйнштейна 43. E-mail: dmitry@Zaykovskiy.de

В клиент-серверных системах распознавания речи стоит задача обеспечить последовательную работу в различных условиях канала передачи данных. Таким образом, необходимо разработать методы, которые позволят снизить эффект ошибок при передаче данных. В данной работе рассматривается подход для сглаживания ошибки, который не изменяет исходные данные, вместо этого он старается воссоздать потерянную информацию в приемнике с помощью интерполяции успешно переданных функций. Используя номер пакета идентификации сервера DSR, можно однозначно решить, какие пакеты были потеряны, и какие ближе к полу-