On Quality Ratings for Spoken Dialogue Systems – Experts vs. Users

Stefan Ultes, Alexander Schmitt, and Wolfgang Minker
Ulm University
Albert-Einstein-Allee 43
89073 Ulm, Germany
{stefan.ultes,alexander.schmitt,wolfgang.minker}@uni-ulm.de

Abstract

In the field of Intelligent User Interfaces, Spoken Dialogue Systems (SDSs) play a key role as speech represents a true intuitive means of human communication. Deriving information about its quality can help rendering SDSs more user-adaptive. Work on automatic estimation of subjective quality usually relies on statistical models. To create those, manual data annotation is required, which may be performed by actual users or by experts. Here, both variants have their advantages and drawbacks. In this paper, we analyze the relationship between user and expert ratings by investigating models which combine the advantages of both types of ratings. We explore two novel approaches using statistical classification methods and evaluate those with a pre-existing corpus providing user and expert ratings. After analyzing the results, we eventually recommend to use expert ratings instead of user ratings in general.

1 Introduction and Motivation

In human-machine interaction it is important that user interfaces can adapt to the specific requirements of its users. Handicapped persons or angry users, for example, have specific needs and should be treated differently than regular users.

Speech is a major component of modern user interfaces as it is the natural means of human communication. Therefore, it seems logical to use Spoken Dialogue Systems (SDS) as part of Intelligent User Interfaces enabling speech communication of different complexity reaching from simple spoken commands up to complex dialogues. Besides the spoken words, the speech signal also may be used to acquire information about the user state, e.g., about their emotional state (cf., e.g., (Polzehl et al., 2011))). By additional analysis of the human-computer-dialogues, even more abstract information may be derived, e.g., the quality of the system (cf., e.g., (Engelbrecht and Möller, 2010)). System quality information may be used to adapt the system's behavior online during the ongoing dialogue (cf. (Ultes et al., 2012)).

For determining the quality of Spoken Dialogue Systems, several aspects are of interest. Möller et al. (2009) presented a taxonomy of quality criteria. They describe quality as a bipartite issue consisting of Quality of Service (QoS) and Quality of Experience (QoE). Quality of Service describes objective criteria like dialogue duration or number of turns. While these are well-defined items that can be determined easily, Quality of Experience, which describes the user experience with subjective criteria, is more vague and without a sound definition, e.g., User Satisfaction (US).

Subjective aspects like US are either determined by using questionnaires like SASSI (Hone and Graham, 2000) or the ITU-standard augmented framework for questionnaires (Möller, 2003), or by using single-valued ratings, i.e., a rater only applies one single score. In general, two major categories of work on determining single-valued User Satisfaction exist. The satisfaction ratings are applied either

- by **users** during or right after the dialogue or
- by **experts** by listening to recorded dialogues.

Proceedings of NAACL-HLT 2013, pages 569–578,
Atlanta, Georgia, 9–14 June 2013. ©2013 Association for Computational Linguistics

In this work, users or *user raters* are people who actually perform a dialogue with the system and apply ratings while doing so. There is no constraint about their expertise in the field of Human Computer Interaction or Spoken Dialogue Systems: They may be novices or have a high expertise. With experts or *expert raters*, we refer to people who are not participating in the dialogue thus constituting a completely different set of people. Expert raters listen to recorded dialogues after the interactions and rate them by assuming the point of view of the actual person performing the dialogue. These experts are supposed to have some experience with dialogue systems. In this work, expert raters were "advanced students of computer science and engineering" (Schmitt et al., 2011a).

For *User* Satisfaction, ratings applied by the users seem to be clearly the better choice over ratings applied by third persons. However, determining true User Satisfaction is only possible by asking real users interacting with the system. Ideally, the ratings are applied by users talking to a system employed in the field, e.g., commercial systems, as these users have real concerns.

For such Spoken Dialogue Systems, though, it is not easy to get users to apply quality ratings to the dialogue – especially for each system-user-exchange. The users would have to rate either by pressing a button on the phone or by speech, which would significantly influence the performance of the dialogue. Longer dialogues imply longer call durations which cost money. Further, most callers only want to quickly get some information from the system. Therefore, it may be assumed that most users do not want to engage in dialogues which are artificially made longer. This also inhabits the risk that users who participated in long dialogues do not want to call again. Therefore, collecting ratings applied by users are considered to be expensive. One possible way of overcoming the problem of rating input would be to use some special installation which enables the users to provide ratings more easily (cf. (Schmitt et al., 2011b)). However, this is also expensive and the system's usability would be very restricted. Further, this setup could most likely only be used in a lab situation.

Expert raters, on the other hand, are able to simply listen to the recorded dialogues and to apply ratings, e.g., by using a specialized rating software. This process is much easier and does not require the same amount of effort needed for acquiring user ratings. Further, as already pointed out, we refer to experts as people who have some basic understanding of dialogue systems but are not required to be high-level experts in the field. That is why we believe that these people can be found easily.

As both categories of ratings have their advantages and disadvantages, this contribution aims at learning about the differences and similarities of user and expert ratings with the ultimate goal of either being able to predict user ratings more efficiently or of advocating for replacing the use of user ratings by using only expert ratings in general.

Therefore, this work analyzes the relation between quality ratings applied by user and expert raters by analyzing approaches which take advantage of both categories: Using the less expensive rating process with expert raters and still predicting *real* User Satisfaction ratings. Moreover, this works' goal is to shed light on the question whether information about one rating (in this case the less expensive expert ratings) may be used to predict the other rating (the more expensive user ratings). For this, we present two approaches applying two different statistical classification methods for a showcase corpus. Results of both methods are compared to a given baseline.

The remainder of this paper is organized as follows. First, we give a brief overview of work done in both categories (user ratings vs. expert ratings) in Section 2 and present our choice of data the analysis in this paper is based on in Section 3. Further, evaluation metrics are illustrated in Section 4 and approaches on facilitating prediction of user rater scores by expert rater information are presented in Section 5 followed by an evaluation and discussion of the results in Section 6.

2 Significant Related Work

Predicting User Satisfaction for SDSs has been in the focus of research for many years, most famously the PARADISE framework by Walker et al. (1997). The authors assume a linear dependency between quantitative parameters derived from the dialogue and US, modeling this dependency using linear regression. Unfortunately, for generating the regression model, weighting factors have to be computed for each system anew. This generates high costs as dialogues have to be performed with real users where each user further has to complete a questionnaire after completing the dialogue. Moreover, in the PARADISE framework, only quality measurement for the whole dialogue (or system) is allowed. However, this is not suitable for using quality information for online adaption of the dialogue (cf. (Ultes et al., 2012)). Furthermore, PARADISE relies on questionnaires while we focus on work using single-valued ratings.

Numerous work on predicting User Satisfaction as a single-valued rating task for each system-user-exchange has been performed in both categories. This work is briefly presented in the following.

2.1 Expert Ratings

Higashinaka et al. (2010a) proposed a model to predict turn-wise ratings for human-human dialogues (transcribed conversation) and human-machine dialogues (text from chat system). Ratings ranging from 1-7 were applied by two expert raters labeling “Smoothness”, “Closeness”, and “Willingness” not achieving a Match Rate per Rating (MR/R)[1] of more than 0.2-0.24. This results are only slightly above the random baseline of 0.14. Further work by Higashinaka et al. (2010b) uses ratings for overall dialogues to predict ratings for each system-user-exchange. Again, evaluating in three user satisfaction categories “Smoothness”, “Closeness”, and “Willingness” with ratings ranging from 1-7 achieved best performance of 0.19 MR/R.

Interaction Quality (IQ) has been introduced by Schmitt et al. (2011a) as an alternative performance measure to User Satisfaction. In their terminology, US ratings are only applied by users. As their presented measure uses ratings applied by expert raters, a different term is used. Each system-user exchange was annotated by three different raters using strict guidelines. The ratings ranging from 1-5 are used as target variable for statistical classifiers using a set of automatically derivable interaction parameters as input. They achieve a MR/R of 0.58.

2.2 User Ratings

An approach presented by Engelbrecht et al. (2009) uses Hidden Markov Models (HMMs) to model the SDS as a process evolving over time. User Satisfaction was predicted at any point within the dialogue on a 5 point scale. Evaluation was performed based on labels the users applied themselves during the dialogue.

Hara et al. (2010) derived turn level ratings from an overall score applied by the users after the dialogue. Using n-gram models reflecting the dialogue history, the achieved results for recognizing User Satisfaction on a 5 point scale showed to be hardly above chance.

Work by Schmitt et al. (2011b) deals with determining User Satisfaction from ratings applied by the users themselves during the dialogues. A statistical classification model was trained using automatically derived interaction parameter to predict User Satisfaction for each system-user-exchange on a 5-point scale achieving an MR/R of 0.49.

3 Corpus

The corpus used by Schmitt et al. (2011b) not only contains user ratings but also expert ratings which makes it a perfect candidate for our research presented in this paper. Adopting the terminology by Schmitt et al., user ratings are described as User Satisfaction (US) whereas expert ratings are referred to with the term Interaction Quality (IQ) (cf. (Schmitt et al., 2011a)). The data used for all experiments of this work was collected by Schmitt et al. (2011b) during a lab user study with 38 users in the domain of the “Let’s Go Bus Information” system (Raux et al., 2006) of the Carnegie Mellon University in Pittsburgh. 128 calls were collected consisting of a total of 2,897 system-user exchanges. Both ratings, IQ and US, are at a scale from 1 to 5 where 1 stands for “extremely unsatisfied” and 5 for “satisfied”. Each dialogue starts with a rating of 5 as the user is expected to be satisfied in the beginning because nothing unsatisfying has happened yet.

Further, the corpus also provides interaction parameters which may be used as input variables for the IQ and US recognition models. These parameters have been derived automatically from three dialogue modules: Automatic Speech Recog-

[1] MR/R is equal to Unweighted Average Recall (UAR) which is explained in Section 4.

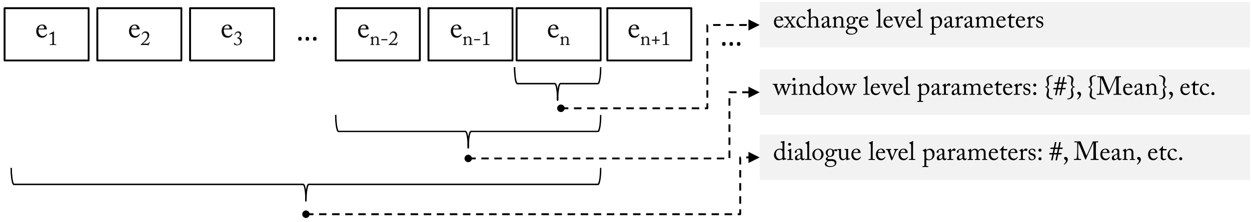
e1
e2
e3
...
en-2
en-1
en
en+1
...
exchange level parameters
window level parameters: {#}, {Mean}, etc.
dialogue level parameters: #, Mean, etc.

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 dialog steps (here $n = 3$); the dialog level, measuring overall performance values from the entire previous interaction.*

nition, Spoken Language Understanding, and Dialogue Management. Furthermore, the parameters are modeled on three different levels (see Figure 1):

- *Exchange level* parameters can be derived directly from the respective dialogue modules, e.g., `ASRConfidence`.
- *Dialogue level* parameters consist of counts (#), means (Mean), etc. of the exchange level parameters calculated from all exchanges of the whole dialogue up to the current exchange, e.g., `MeanASRConfidence`.
- *Window level* parameters consist of counts ({#}), means ({Mean}), etc. of the exchange level parameters calculated from the last three exchanges, e.g., {`Mean`}`ASRConfidence`.

4 Evaluation metrics

For measuring the performance of the classification algorithms, we rely on *Unweighted Average Recall (UAR)*, *Cohen's Kappa* and *Spearman's Rho*. The latter two also represent a measure for similarity of paired data. All measures will be briefly described in the following:

Unweighted Average Recall The Unweighted Average Recall (UAR) is defined as the sum of all class-wise recalls r_c divided by the number of classes $|C|$:

$$UAR = \frac{1}{|C|} \sum_{c \in C} r_c \; . \qquad (1)$$

Recall r_c for class c is defined as

$$r_c = \frac{1}{|R_c|} \sum_{i=1}^{|R_c|} \delta_{h_i r_i} \; , \qquad (2)$$

where δ is the Kronecker-delta, h_i and r_i represent the corresponding hypothesis-reference-pair of rating i, and $|R_c|$ the total number of all ratings of class c. In other words, UAR for multi-class classification problems is the accuracy corrected by the effects of unbalanced data.

Cohen's Kappa To measure the relative agreement between two corresponding sets of ratings, the number of label agreements corrected by the chance level of agreement divided by the maximum proportion of times the labelers could agree is computed. κ is defined as

$$\kappa = \frac{p_0 - p_c}{1 - p_c} \; , \qquad (3)$$

where p_0 is the rate of agreement and p_c is the chance agreement (Cohen, 1960). As US and IQ are on an ordinal scale, a weighting factor w is introduced reducing the discount of disagreements the smaller the difference is between two ratings (Cohen, 1968):

$$w = \frac{|r_1 - r_2|}{|r_{max} - r_{min}|} \; . \qquad (4)$$

Here, r_1 and r_2 denote the rating pair and r_{max} and r_{min} the maximal and minimal rating. This results in $w = 0$ for agreement and $w = 1$ if the ratings have maximal difference.

Spearman's Rho The correlation of two variables describes the degree by that one variable can be expressed by the other. *Spearman's Rank Correlation Coefficient* is a non-parametric method assuming a monotonic function between the two variables (Spearman, 1904). It is defined by

$$\rho = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}} , \qquad (5)$$

where x_i and y_i are corresponding ranked ratings and $\bar{x}$ and $\bar{y}$ the mean ranks. Thus, two sets of ratings can have total correlation even if they never agree. This would happen if all ratings are shifted by the same value, for example.

5 Recognition of US Using IQ Information

As discussed in Section 1, automatic recognition of ratings applied by users as performed by Schmitt et al. (2011b) for User Satisfaction is time-consuming and expensive. Therefore, approaches are presented which facilitate expert ratings, i.e., Interaction Quality, with the hope of making US recognition more feasible. IQ an US are strongly related as both metrics represent the same quantity applied by different rater groups. Results of the Mann-Whitney U test, which is used to test for significant difference between Interaction Quality and User Satisfaction, show their difference ($p < 0.05$) but values for Cohen's Kappa (Cohen, 1960) and Spearman's Rank Correlation Coefficient (Spearman, 1904) emphasize the that IQ and US are quite similar. Achieving $\kappa = 0.5$ can be considered as a *moderate* agreement according to Landis and Koch's Kappa Benchmark Scale (Landis and Koch, 1977). Furthermore, a correlation of $\rho = 0.66$ ($p < 0.01$) indicates a *strong* relationship between IQ and US (Cohen, 1988).

While it has been shown that user and expert ratings are similar, it is desirable nonetheless to being able to predict real user ratings. These ratings are the desired kind of ratings when it comes to subjective dialogue system assessment. Only users can give a rating about their satisfaction level, i.e., how they like the system and the interaction with the system. However, user ratings are expensive as elaborated in Section 1. Therefore, we investigate approaches to recognize US which rely on means of IQ recognition.

5.1 Belief-Based Sequential Recognition

Methods used for IQ and US recognition by Schmitt et al. (2011b; 2011a) suffer from the fact that the sequential character of the data is modeled inadequately as they assume statistical independence between the single exchanges (recognition of IQ and US does *not* depend on the respective value of the previous exchange). Hence, we present a Markovian approach overcoming these issues. A probability distribution over all US states, called *belief state*, is updated after each system-user-exchange taking also into account the *belief state* of the previous exchange. This belief update[2] is equivalent to the Forward Algorithm known from Hidden Markov Models (cf. (Rabiner, 1989)). In doing so, the new US probabilities also depend on the US values of the previous exchange. Moreover, a latent variable is introduced in order to decouple the target variable US with the variable the observation probability depends on IQ. This results in an indirect approach for recognizing User Satisfaction that is based on the more affordable recognition of Interaction Quality assuming that a universal mapping between IQ and US exists.

Thus, to determine the probability $b(US)$ of having the true User Satisfaction label US after the current system-user-exchange, we rely on Interaction Quality recognition, whose observation probability is depicted as $P(o|IQ)$. Furthermore, for coupling both quantities, we introduce a coherence probability $P(IQ|US)$. Belief update for estimating the new values for $b'(US')$ is as follows:

$$b'(US') = \alpha \cdot \sum_{IQ'} P(o'|IQ') \cdot P(IQ'|US') \cdot \sum_{US} P(US'|US) b(US) \qquad (6)$$

The observation probability $P(o'|IQ')$ is modeled using confidence scores of classifiers applied for IQ recognition. Further, we compute the sum over all previous US beliefs $b(US)$ weighted by the transition probability $P(US'|US)$. Both, transition and coherence probability have been computed by taking the frequency of their occurrences in the training data. The α factor is used for normalization only.

Since we are aiming at generating an estimate $\hat{US}$

[2] Terminology is taken from Partially Observable Markov Decision Processes, cf. (Kaelbling et al., 1998)

at each exchange, it is calculated by

$$\hat{US} = \arg\max_{US'} b'(US') \tag{7}$$

generating a sequence of estimates for each dialogue.

As the action of the system a can be expected to influence the satisfaction level of the user, action-dependency is added to Equation 6 resulting in

$$b'(US') = \alpha \cdot \sum_{IQ'} P(o'|IQ') \cdot P(IQ'|US', a) \cdot \sum_{US} P(US'|US, a) b(US). \tag{8}$$

Hence, each system action a influences coherence and transition probabilities. It should be noted that action-dependency can only be introduced as in a SDS each turn a system action is selected and executed by the dialogue manager.

5.2 Model Exchange

While in *Belief-Based Sequential Recognition*, probability models are used for coupling expert and user ratings explicitly, a simpler approach has also been examined. A statistical classifier trained on the target variable IQ is used to evaluate classification of the target variable US. This seems to be reasonable as the set of scores and meaning of the scores of both metrics are equivalent. Furthermore, necessary prerequisites are fulfilled: the sample corpus contains both labels, the labels for US and IQ correspond, and both recognition approaches are based on the same feature set.

6 Experiments and Results

For evaluating *Belief-Based Sequential Recognition*, not only the absolute performance is of interest but also how this performance is influenced by the characteristics of the observation probability, i.e., the performance of the applied statistical classification approach and the variance of their confidence scores. In order to obtain different confidence characteristics, multiple classification algorithms, or algorithm variants respectively, are needed. Hence, five statistical classifiers have been chosen arbitrarily to produce the observation probabilities for *Belief-Based Sequential Recognition*:

- SVM[3] with cubic kernel
- SVM with RBF-kernel
- Naive Bayes
- Naive Bayes with kernel
- Rule Induction

In contrast to Schmitt et al. (2011b; 2011a), a reduced feature set was used consisting of 43 parameters as some textual parameters were removed which are very specific and take many different values, e.g., UTTERANCE (the system utterance) or INTERPRETATION (the interpretation of the speech input).

The resulting feature set consists of the following parameters (parameter names are in accordance with the parameter names of the LEGO corpus (Schmitt et al., 2012)):

Exchange Level ACTIVITY, ACTIVITYTYPE, UTD, BARGED-IN?, ASRCONFIDENCE, MEANASRCONFIDENCE, TURNNUMBER, MODALITY, LOOPNAME, ASRRECOGNITIONSTATUS, ROLEINDEX, ROLENAME, NOISE?, HELPREQUEST?, REPROMPT?, WPST, WPUT

Dialogue Level #BARGEINS #ASRSUCCESS, #HELPREQUESTS, #TIMEOUTS, #TIMEOUTS_ASRREJECTIONS, #ASRREJECTIONS, #REPROMPTS, #SYSTEMQUESTIONS, #SYSTEMTURNS, #USERTURNS, %BARGEINS, %ASRSUCCESS, %HELPREQUESTS, %TIMEOUTS, %TIMEOUTS_ASRREJECTIONS, %ASRREJECTIONS, %REPROMPTS

Window Level {#}TIMEOUTS_ASRREJCTIONS, {#}HELPREQUESTS, {#}ASRREJECTIONS, {MEAN}ASRCONFIDENCE, {#}TIMEOUTS, {#}REPROMPTS, {#}SYSTEMQUESTIONS, {#}ASRSUCCESS, {#}BARGEINS

All results are evaluated with respect to the reference experiment of direct US recognition (US recognition using models trained on US). This is performed in accordance to Schmitt et al. (2011b) using the statistical classification algorithms stated

[3] Support Vector Machine, cf. (Vapnik, 1995)

Table 1: *Results (UAR, Cohen's Kappa, and Spearman's Rho) of 10-fold cross-validation for US recognition of US recognition using models trained on US*

Classifier	UAR	κ	ρ
SVM (cubic Kernel)	0.39	0.33	0.48
SVM (RBF-Kernel)	0.39	0.42	0.55
Naive Bayes	0.36	0.40	0.55
Naive Bayes (Kernel)	0.42	0.44	0.59
Rule Induction	0.50	0.51	0.61

Table 2: *Results (UAR, Cohen's Kappa, and Spearman's Rho) of 10-fold cross-validation for US recognition of the Model Exchange approach (trained on IQ, evaluated on US)*

Classifier	UAR	κ	ρ
SVM (cubic Kernel)	0.34	0.42	0.55
SVM (RBF-Kernel)	0.34	0.42	0.58
Naive Bayes	0.35	0.40	0.57
Naive Bayes (Kernel)	0.34	0.37	0.60
Rule Induction	0.34	0.42	0.59

Table 3: *Results (UAR, Cohen's Kappa, and Spearman's Rho) of 10-fold cross-validation for US recognition of action-independent Belief-Based Sequential Recognition*

Classifier	UAR	κ	ρ
SVM (cubic Kernel)	0.28	0.36	0.48
SVM (RBF-Kernel)	0.30	0.40	0.54
Naive Bayes	0.32	0.39	0.54
Naive Bayes (Kernel)	0.33	0.45	0.61
Rule Induction	0.33	0.47	0.63

Table 4: *Results (UAR, Cohen's Kappa, and Spearman's Rho) of 10-fold cross-validation for US recognition of action-dependent Belief-Based Sequential Recognition*

Classifier	UAR	κ	ρ
SVM (cubic Kernel)	0.28	0.35	0.48
SVM (RBF-Kernel)	0.29	0.40	0.54
Naive Bayes	0.32	0.40	0.55
Naive Bayes (Kernel)	0.34	0.44	0.60
Rule Induction	0.35	0.47	0.62

above. The performance of the reference experiment is shown in Table 1.

Using the same feature set, these classification algorithms are also applied for the evaluation of the *Model Exchange* approach using 10-fold cross validation. Note that the parameters of the classifiers also remained the same. The data was partitioned randomly on exchange level, i.e., without regarding their belonging to a specific dialogue. The measured results of the *Model Exchange* approach for the five classification methods can be seen in Table 2.

While the results are significantly above chance[4], comparing them to the reference experiment reveals that in terms of UAR the reference experiment outperforms *Model Exchange* for all five classifiers. The achieved κ and ρ values show similar scores for both the reference experiment and the *Model Exchange* approach. However, in the data used for the experiments, the amount of occurrences of the ratings was not balanced (equal for all classes) which has been identified as the most likely reason for this effect.

Experiments for *Belief-Based Sequential Recognition* have also been performed using 10-fold cross validation. As complete dialogues and the order of exchanges within the dialogues are important for this approach, the data was partitioned randomly on the dialogue level. As previously explained, for the probability distributions of the observation probability model, classification results of IQ recognition with 10-fold cross validation has been used in order to get good estimates for the whole data set. Results for the action-independent version can be seen in Table 3.

For the action-dependent version, four different basic actions Announcement, Confirmation, Question, and Wait have been used, generating results presented in Table 4. The results illustrate that neither action-independent nor action-dependent *Belief-Based Sequential Recognition* can outperform the reference experiment (cf. Table 1). Still, both variants achieve results clearly above chance. Again, the unbalanced data causes κ and ρ to be similar to the reference experiment.

A comparison of the action-independent with the action-dependent approach shows almost no differences in their performances. Only a slight tendency towards better UARs for action-dependency can be spotted.

Figure 2 displays the performances of both variants of *Belief-Based Sequential Recognition* along with performance of IQ recognition and the variance σ^2 of the corresponding confidence distribu-

[4] UAR of 0.2 for five classes

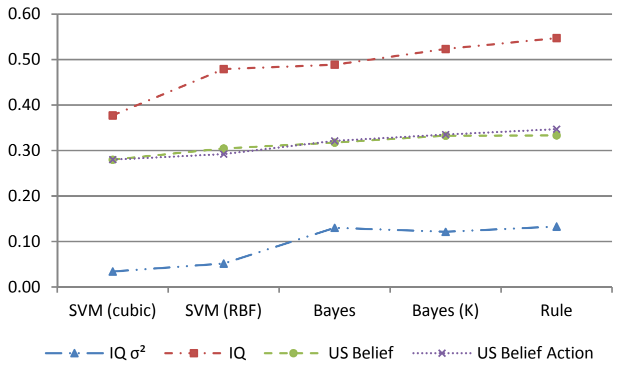

Figure 2: *UAR of IQ recognition and Belief-Based Sequential Recognition along with σ^2 of confidence distributions of IQ recognition*

Table 5: *Recognition performance and variance of confidence distributions for IQ recognition*

Classifier	σ^2	UAR	κ	ρ
SVM (cubic Kernel)	0.03	0.38	0.54	0.69
SVM (RBF-Kernel)	0.05	0.48	0.65	0.77
Naive Bayes	0.13	0.49	0.57	0.71
Naive Bayes (Kernel)	0.12	0.52	0.59	0.73
Rule Induction	0.13	0.55	0.68	0.79

tion (cf. Table 5). It can easily be seen that with rising UAR for IQ recognition, σ^2 also rises. This directly transfers to the performance of the *Belief-Based Sequential Recognition*. The more accurate the observation performance, the more accurate the belief prediction. Furthermore, when comparing the action-dependent to the action-independent variant of *Belief-Based Sequential Recognition*, better IQ performance and therefore a higher variance also causes slightly better results for the action-dependent variant. These differences, however, are only marginally. Therefore, they do not allow for drawing a conclusion.

7 Conclusions

For estimating User Satisfaction-like ratings, two categories exist: work relying on user ratings and work relying on expert ratings. To learn something about their differences and similarities, we explored the possibility of using the information encoded in the expert ratings to predict user ratings with the hope to get acceptable user rating prediction results. Therefore, we investigated if it is possible to determine the preferred true User Satisfaction value based on less expensive expert ratings. For this, a corpus containing both kinds of ratings was chosen, i.e., User Satisfaction (US) and Interaction Quality (IQ) ratings. Furthermore, interaction parameters were used to create statistical recognition models for predicting IQ and US, respectively. Two approaches have been investigated: *Belief-Based Sequential Recognition*, which is based on an HMM-like structure with IQ as an additional latent variable, and *Model Exchange*, which uses statistical models trained on IQ to recognize US. Unfortunately, neither *Belief-Based Sequential Recognition* nor *Model Exchange* achieved results with an acceptable UAR.

The high correlation between expert and user ratings, depicted by high values for Cohen's κ and Spearman's ρ, already allow the conclusion that expert ratings can be used as a good replacement for user ratings. Moreover, the presented recognition results of the *Model Exchange* approach being clearly above chance underpin the strong similarity of IQ and US. Furthermore, IQ recognition is much more reliable and accurate than US recognition (shown by higher UAR, κ and ρ values).

While the experiments disproved the hope of getting acceptable user rating prediction results, the obtained results confirmed the similarity between both kinds of ratings. And as it is not necessary to use user ratings for most applications, e.g., for using the quality information to automatically improve the interaction (cf. (Ultes et al., 2012)), we believe that it suffices to use expert ratings as those can be acquired easier and less expensively and are similar enough to user ratings. Prompting the user to apply quality ratings in everyday situations with real-life systems will always be annoying to the user while recording of such interactions are always much easier to rate.

By providing a study for determining quality ratings of dialogues, we hope to encourage other researchers to look into this research for other parameters, e.g., emotion recognition.

References

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. In *Educational and Psychological Measurement*, volume 20, pages 37–46, April.

Jacob Cohen. 1968. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4):213.

Jacob Cohen. 1988. *Statistical power analysis for the behavioral sciences*. New Jersey: Lawrence Erlbaum Associates, July.

Klaus-Peter Engelbrecht and Sebastian Möller. 2010. A User Model to Predict User Satisfaction with Spoken Dialog Systems. In Gary Geunbae Lee, Joseph Mariani, Wolfgang Minker, and Satoshi Nakamura, editors, *Spoken Dialogue Systems for Ambient Environments. 2nd Int. Workshop on Spoken Dialogue Systems Technology*, Lecture Notes in Artificial Intelligence, pages 150–155. Springer, October.

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.

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 (Conference Chair), 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).

Ryuichiro Higashinaka, Yasuhiro Minami, Kohji Dohsaka, and Toyomi Meguro. 2010a. 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.

Ryuichiro Higashinaka, Yasuhiro Minami, Kohji Dohsaka, and Toyomi Meguro. 2010b. 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.

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

L. P. Kaelbling, M. L. Littman, and A. R. Cassandra. 1998. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101(1-2):99–134.

J. R. Landis and G. G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174, March.

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.

Sebastian Möller. 2003. Subjective Quality Evaluation of Telephone Services Based on Spoken Dialogue Systems. ITU-T Recommendation P.851, International Telecommunication Union, Geneva, Switzerland, November. Based on ITU-T Contr. COM 12-59 (2003).

Tim Polzehl, Alexander Schmitt, and Florian Metze. 2011. Salient features for anger recognition in german and english ivr portals. In Wolfgang Minker, Gary Geunbae Lee, Satoshi Nakamura, and Joseph Mariani, editors, *Spoken Dialogue Systems Technology and Design*, pages 83–105. Springer New York. 10.1007/978-1-4419-7934-6_4.

Lawrence R. Rabiner. 1989. *A tutorial on hidden Markov models and selected applications in speech recognition*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

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. 2012. A parameterized and annotated corpus of the cmu let's go bus information system. In *International Conference on Language Resources and Evaluation (LREC)*.

C. Spearman. 1904. The proof and measurement of association between two things. *American Journal of Psychology*, 15:88–103.

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 Commu-*

nity: Tools and Data (SDCTD 2012), pages 49–52, Montréal, Canada, June. Association for Computational Linguistics.

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

Marilyn Walker, Diane 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.