Interaction Quality Estimation Using Long Short-Term Memories

Niklas Rach[1], **Wolfgang Minker**[1], **and Stefan Ultes**[2]
[1]Institute of Communication Engineering, Ulm University
{vorname.nachname}@uni-ulm.de
[2]Department of Engineering, University of Cambridge, UK
su259@cam.ac.uk

Abstract

For estimating the Interaction Quality (IQ) in Spoken Dialogue Systems (SDS), the dialogue history is of significant importance. Previous works included this information manually in the form of precomputed temporal features into the classification process. Here, we employ a deep learning architecture based on Long Short-Term Memories (LSTM) to extract this information automatically from the data, thus estimating IQ solely by using current exchange features. We show that it is thereby possible to achieve competitive results as in a scenario where manually optimized temporal features have been included.

1 Introduction

The increasing complexity of Spoken Dialogue Systems (SDS) and the requirements that come with this progress made automatized recognition and modeling of user states crucial to ensure natural and user adaptive interaction. User Satisfaction (US) is one important part of such a state. On the dialogue level (i.e. after the interaction is complete), it provides a measure for the interaction and allows to compare different SDS (Walker et al., 1997) or to learn appropriate dialogue strategies (Walker, 2000; Ultes et al., 2017a). However, if US is available in each turn, it can also be used for user adaptation (Ultes et al., 2011, 2012, 2016, 2014a).

In the scope of this work we focus on the Interaction Quality (IQ) as a turn-wise approach to US and propose a deep learning architecture to estimate it solely using exchange parameters[1]. In doing so, we show that with the proposed approach, manually optimized, pre-computed temporal information (as employed in previous work) is no longer required.

Diverse approaches for estimating the US were already proposed, including n-gram models (Hara et al., 2010) and Hidden Markov Models (Higashinaka et al., 2010a; Engelbrech et al., 2009) in different scenarios. Although the results were above the random baseline, the respective improvement was only minor. As it was discussed by Higashinaka et al. (2010b), one difficulty of this task lies in the subjective nature of US since it depends on the appreciation of the user.

IQ is a more objective approach to US that relies on the rating of experts instead of users (Schmitt and Ultes, 2015) and thus closes the gap between subjective valuation and objective criteria. The respective rating is given on a scale between 1 (extremely unsatisfied) and five (satisfied) after listening to audio records of the dialogue in question. A detailed study on the correlation between the IQ and a measure of the real US was provided by Ultes et al. (2013) and various approaches including Hidden Markov Models (Ultes et al., 2014b; Ultes and Minker, 2014), Support Vector Machines (Schmitt et al., 2011; Ultes and Minker, 2013), Ordinal Regression (El Asri et al., 2014) and Recurrent Neural Networks (Pragst et al., 2017) have been employed to estimate the IQ from exchange parameters. Although the results show a significant improvement to alternative approaches, the classification relies in each case on precomputed features modeling the dialogue history (so called *temporal features*).

Despite the good results, using *temporal features* requires insight into the correlations between the dialogue history and the IQ score as the timespan covered by the temporal information significantly influences the outcome (Ultes et al., 2017b). The required knowledge about this correlation is

[1] An exchange is a system turn followed by a user turn.

Proceedings of the SIGDIAL 2017 Conference, pages 164–169,
Saarbrücken, Germany, 15-17 August 2017. ©2017 Association for Computational Linguistics

usually not accessible and likely to be domain dependent thus rendering the respective approaches inflexible. In contrast, we employ a deep learning classifier to extract the required temporal information automatically and show that in doing so it is possible to achieve competitive results by only using exchange level parameters. In addition, we show that findings of previous works regarding the optimal amount of temporal information to be included may be retrieved in our approach by slightly varying the input sequences. Finally, the usability of our proposed architecture in real-life scenarios is discussed by looking at the percentage of usable IQ guesses.

The remainder of this paper is as follows: In Section 2 we discuss the LSTM based neural network architecture followed by a discussion of the employed data in Section 3. Section 4 presents the experiments and results and we close with a brief conclusion and outlook in Section 5.

2 LSTM-based Interaction Quality Estimation

Recurrent Neural Networks (RNN) include temporal correlations in the data into the classification process and are thus suitable for sequential tasks such as the one at hand. However, common approaches have shown to be inefficient in learning long-term dependencies (Bengio et al., 1994) due to a vanishing (or exploding) gradient. To tackle this problem, Hochreiter et al. (1997) introduced an architecture, called Long Short-Term Memory (LSTM) that allows to preserve temporal information, even if the correlated events are separated by a longer time. Since previous works showed that long time correlations are of importance for estimating the IQ, we consider LSTM a suitable approach for the reviewed scenario.

The herein employed architecture is thus built of a LSTM unit, consisting of two stacked LSTM cells, followed by a two-layer perceptron unit with sigmoid activation functions. The latter one is given as

$$F_{MLP} : y_t \rightarrow (g_2 \circ g_1)(y_t) \tag{1}$$

$$g_i(y_t) = sigm(W_i^T y_t + b_i) \tag{2}$$

where W_i denotes the weight matrix, b_i a bias vector and $sigm$ the element-wise sigmoid function. A LSTM cell on the other hand can be seen as function

$$f : x_t, c_{t-1}, h_{t-1} \rightarrow h_t, c_t \tag{3}$$

with h_t the output state, c_t the internal cell state and x_t the input of the LSTM at time step t. In a multilayer scenario, the input of a layer is the output of the previous one. A deeper discussion of the LSTM architecture including the respective formulas is provided for example in (Zaremba et al., 2014). The complete LSTM unit can thus be written as a function F_{LSTM} that processes a given input through two LSTM layers and maps it to an output state y_t. Combining this description with equation 1 yields

$$z_t = (F_{MLP} \circ \sigma \circ F_{LSTM})(x_t) \tag{4}$$

for the whole net with z_t the final IQ mapping of the input and σ the softmax normalization function. In the reviewed scenario, each LSTM layer consisted of 48 nodes whereas the perceptron unit had 48 nodes in the hidden layer and five nodes in the output layer. Therefore, the two LSTM layers are employed to extract the temporal information whereas the following perceptron layers serve as classifier that maps the output of the LSTM unit to the respective IQ scale. The whole net is depicted in Figure 1 and was implemented using Google's Tensorflow library (Abadi et al., 2016). Optimization was done by use of the Adaptive Gradient Algorithm (Duchi et al., 2011).

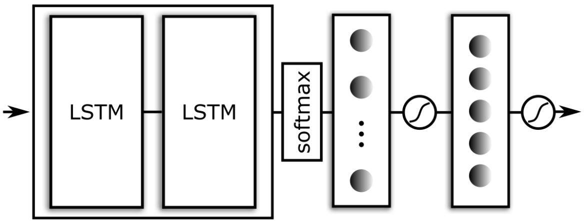
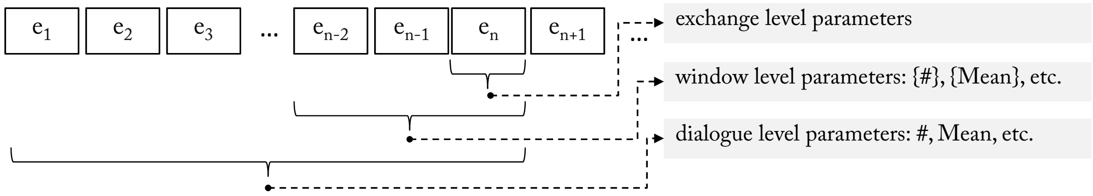


Figure 1: Sketch of the deep learning architecture in use. The left part contains the two stacked LSTM cells followed by a softmax normalization unit. The output is fed into a two layer perceptron with sigmoid activation functions.

3 The LEGO Corpus

To appropriately compare our results, we employ the LEGO coprus (Schmitt et al., 2012)—the same corpus as the authors of previous work. It is based on the "Let's Go Bus Information System" of the Carnegie Mellon university in Pittsburg (Raux et al., 2006) and consists of 200 dialogues including 4884 system-user exchanges. Each exchange was assigned with features from three instances of



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

Figure 2: The three parameter levels including the temporal features of the window and the dialogue level (Schmitt et al., 2012).

the SDS, namely the Automatic Speech Recognition (ASR), Natural Language Understanding (NLU) and the Dialogue Manager (DM). Furthermore, the corpus was annotated with an IQ rating by three experts following specific guidelines to achieve an objective measure (Schmitt et al., 2011). In doing so, an inter-annotator agreement of $\kappa = 0.54$ was achieved. For the final IQ score, the median of all three ratings was taken. To include temporal features into the corpus, three different interaction levels that are depicted in Figure 2 were considered:

- The exchange level contains all features regarding the current system-user exchange.
- The window level includes counts and means of numerical exchange level features from the previous n exchanges, where n is referred to as window size.
- The dialogue level contains counts and means of numerical exchange level features from all previous exchanges.

The term temporal features thus refers to features of the window and dialogue level. The influence of these two additional levels as well as the choice of n on the automatized estimation of the IQ were studied (Ultes et al., 2017b) and serve as a baseline for this work.

4 Experiments and Results

In this section we discuss the results of the employed classifier in estimating the IQ for the annotated LEGO corpus. To distinguish the contribution of the parameters derived from different SDS instances to the IQ, three feature sets were employed that consisted of features assigned to the ASR, the DM and both:

ASR: *ASRRecognitionStatus* (string, status of the ASR), *Modality* (string, input modality of the user, either *speech* or *dtmf*), *ExMo* (string, expected modality of the user input, either *speech*, *dtmf*, *both* or *none*), *ASRConfidence* (float, confidence score of the ASR), *Barged-In?* (boolean, true if system was interrupted by the user), *UnExMo?* (boolean, true if the actual input modality did not match the expected one), *WPUT* (integer, words per user turn), *UTD* (float, utterance turn duration)

DM: *ActivityType* (string, type of activity), *RoleName* (string, function of the system turn), *RePromt?* (boolean, true if the current turn is a repromt), *WPST* (integer, words per system turn), *DD* (float, dialogue duration), *RoleIndex* (integer, tries necessary to get a desired response from the user)

Parameters that are either constant or task-related were discarded, including the two features from the NLU. To represent all parameters as a numerical input vector, non-numerical features were encoded in a one-hot vector. As in previous work, we used 10-fold cross validation to evaluate the outcomes. The results are compared in terms of Unweighted Average Recall[2] (UAR), Cohen's (linearly weighted) Kappa (Cohen, 1968) and Spearman's Rho (Spearman, 1904) to the ones achieved by Ultes et al. (2017b) with the best window size $n = 9$, the full feature set and a Support Vector Machine (SVM). Our results as well as the baseline value are shown in Table 1. For all three measures, the results with the full feature set are competitive to the baseline. Whereas the UAR is slightly below the reference value, κ and ρ show a small improvement. The results for the two subsets are visibly below the baseline for both UAR and κ whereas the DM value of ρ equals the respective reference value. Moreover, the DM features yield better results than the ASR features and thus contribute more to the overall IQ value,

[2] The arithmetic average of all class-wise recalls.

	features	#TF	UAR	κ	ρ
LSTM	ASR+DM	0	**0.548**	**0.684**	**0.832**
LSTM	ASR	0	0.502	0.636	0.796
LSTM	DM	0	0.516	0.654	0.812
SVM	ASR+DM	25	**0.549**	0.679	0.812

Table 1: The results of the LSTM approach in comparison to the SVM baseline (Ultes et al., 2017b), including the number of handcrafted temporal features in use (#TF) for each scenario.

which is in line with the outcomes of previous work (Ultes et al., 2015). It is stressed that none of the feature sets employed for the LSTM uses handcrafted temporal features nor needs them. Thus, we conclude that our approach is indeed capable of extracting the required temporal information automatically.

In addition, we investigate the temporal information extracted by the trained classifier by measuring the impact of one system-user exchange on following estimates. This allows a comparison of the extracted information in the herein discussed scenario with the manually set window size in previous work. To this end, we replaced the input vector of the second system-user exchange e_2 in each dialogue $D_i = (e_1^i, e_2^i, .., e_L^i)$ of the corpus $\{D_1, ..., D_M\}$ by the input associated with one out of 20 randomly picked exchanges e_r^j ($j \in \{1, \dots, M\}$) with assigned IQ value of 1. The modified dialogues

$$\tilde{D}_i = (e_1^i, e_r^j, ..., e_L^i) \qquad (5)$$

were then fed through a trained model of the 10-fold cross validation and the results were compared to the ones achieved with the original data by computing the sum of the absolute errors of each class. This was repeated for all 20 random picks and all 10 models (we employed different random picks for each model). The mean of this error over all dialogues, all trained models and all random picks for the replaced exchange was determined and is shown as a function of the system-user exchange number in Figure 3. This error indicates the impact one exchange has on the IQ estimate of following exchanges. We see that from exchange number 9 to exchange number 12 the error clearly decreases. A comparison with the referenced work shows that this drop is in the same range as the optimal window size $n = 9$ (that would correspond to exchange number 11).

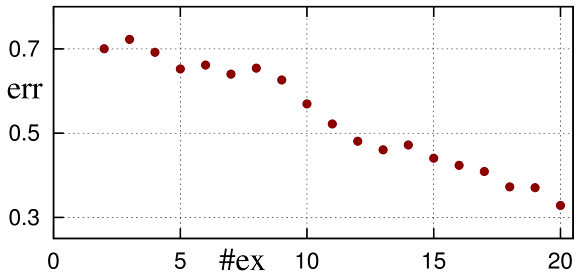


Figure 3: Mean error caused by the replacement of the second system-user exchange by a random picked exchange as a function of the exchange number.

Therefore the impact of the exchange in question is decreased in the same range as in a scenario were this impact is controlled manually. This indicates that similar temporal information that was employed therein is automatically extracted by our architecture.

In many classification scenarios, the classes are not ordered which means that in the case of a wrong guess it is irrelevant which class was chosen. However, as the IQ is an ordered scale, the distance of the wrong guess to the real class is of interest, especially in view of the application. We therefore compute the amount of guesses in which the classification was wrong only by one point (e.g. an instant of IQ 1 classified as IQ 2 or vice versa). This percentage δ can be derived directly from the confusion matrix C as

$$\delta = \frac{1}{N}\left(\sum_{k=1}^{K-1} C_{k,k+1} + \sum_{k=2}^{K} C_{k,k-1}\right) \qquad (6)$$

with N the number of total entries of C and K the number of classes, i.e. the dimension of C. Adding this value to the Accuracy (ACC) gives a percentage of usable guesses of the classifier. The results for the architecture used in this work and the best feature set (ASR + DM) are ACC=0.57 and δ=0.37, resulting in a sum of 0.94. In other words, considering a real-life scenario, 94% of the classifiers guesses could be used, for example for user adaptation. Again, these results are compared to the ones achieved with a SVM and the setup of (Ultes et al., 2017b) with a sum of 0.91. Evidently, the deep learning classifier outperforms the SVM approach in this metric.

5 Conclusion and Outlook

In this work, we investigated the estimation of the IQ with a deep learning classifier by only using exchange level parameters. It was shown that by use of the presented architecture, precomputed temporal features are no longer required and the IQ can be estimated with an UAR of 0.548. The results are competitive to the ones achieved with a SVM classifier and the whole feature set in earlier work. In addition, we compared the temporal information extracted by the classifier with the optimal window size from previous work and showed that our results match previous findings. Finally, the usability of the employed classifier in applications was discussed by computing the percentage of usable guesses in such a case. The result of 94% is below the outcome of the 0.91 achieved with the SVM and a complete feature set. Moreover, since our approach does not require any domain dependent information, it is much more flexible.

It is reasonable to assume that the difficulty of estimating the interaction quality and the amount of temporal information that is required rely on the complexity of the system and the interaction. Although the herein presented slot filling dialogue is comparatively basic, the IQ is influenced not only by technical aspects (e.g., the quality of the speech recognition) but also by the ability of the system to react appropriately. This influence is even stronger in more advanced tasks, where the user satisfaction (and thus the IQ as well) may also depend on the ability of the system to appropriately react on the users state including for example emotions and culture. Although this task differs from the one addressed here, we assume the presented architecture to be a good starting point for these scenarios as well due to its above discussed flexibility.

Thus, for future work the performance of this architecture in different scenarios and systems will be of interest, especially in systems were the IQ depends on additional aspects. Moreover, applying the presented architecture to estimate other user states or features used for user adaptation is also in the focus of future work.

Acknowledgments

This work is part of a project that has received funding from the *European Unions Horizon 2020 research and innovation programme* under grant agreement No 645012.

References

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al. 2016. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467* .

Yoshua Bengio, Patrice Simard, and Paolo Frasconi. 1994. Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks* 5(2):157–166.

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

John Duchi, Elad Hazan, and Yoram Singer. 2011. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research* 12(Jul):2121–2159.

Layla El Asri, Hatim Khouzaimi, Romain Laroche, and Olivier Pietquin. 2014. Ordinal regression for interaction quality prediction. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pages 3245–3249.

Klaus-Peter Engelbrech, Florian Gödde, Felix Hartard, Hamed Ketabdar, and Sebastian Möller. 2009. Modeling user satisfaction with Hidden Markov Model. *Proceedings of the SIGDIAL 2009 Conference: The 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue* (September):170–177.

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 *LREC*.

Ryuichiro Higashinaka, Yasuhiro Minami, Kohji Dohsaka, and Toyomi Meguro. 2010a. Modeling user satisfaction transitions in dialogues from overall ratings. In *Proceedings of the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Association for Computational Linguistics, pages 18–27.

Ryuichiro Higashinaka, Yasuhiro Minami, Kohji Dohsaka, and Toyomi Meguro. 2010b. Modeling User Satisfaction Transitions in Dialogues from Overall Ratings. *Proceedings of the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue* pages 18–27.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9(8):1735–1780.

Louisa Pragst, Stefan Ultes, and Wolfgang Minker. 2017. Recurrent neural network interaction quality estimation. In Kristiina Jokinen and Graham Wilcock, editors, *Dialogues with Social Robots: Enablements, Analyses, and Evaluation*, Springer Singapore, Singapore, pages 381–393.

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 *INTERSPEECH*.

Alexander Schmitt, Benjamin Schatz, and Wolfgang Minker. 2011. Modeling and predicting quality in spoken human-computer interaction. *Proceedings of the SIGDIAL 2011 Conference* pages 173–184.

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* 74:12–36.

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.

Charles Spearman. 1904. The proof and measurement of association between two things. *The American journal of psychology* 15(1):72–101.

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

Stefan Ultes, Hüseyin Dikme, and Wolfgang Minker. 2016. Dialogue Management for User-Centered Adaptive Dialogue. In Alexander I. Rudnicky, Antoine Raux, Ian Lane, and Teruhisa Misu, editors, *Situated Dialog in Speech-Based Human-Computer Interaction*, Springer International Publishing, Cham, pages 51–61. https://doi.org/10.1007/978-3-319-21834-2_5.

Stefan Ultes, Robert ElChab, and Wolfgang Minker. 2014b. Application and evaluation of a conditioned hidden markov model for estimating interaction quality of spoken dialogue systems. In *Natural Interaction with Robots, Knowbots and Smartphones*, Springer, pages 303–312.

Stefan Ultes, Tobias Heinroth, Alexander Schmitt, and Wolfgang Minker. 2011. A theoretical framework for a user-centered spoken dialog manager. In Ramón López-Cózar and Tetsunori Kobayashi, editors, *Proceedings of the Paralinguistic Information and its Integration in Spoken Dialogue Systems Workshop*. Springer New York, New York, NY, pages 241–246. https://doi.org/10.1007/978-1-4614-1335-6_24.

Stefan Ultes, Juliana Miehle, and Wolfgang Minker. 2017a. On the applicability of a user satisfaction-based reward for dialogue policy learning. In *Proceedings of the 8th International Workshop On Spoken Dialogue Systems (IWSDS)*.

Stefan Ultes and Wolfgang Minker. 2013. Improving Interaction Quality Recognition Using Error Correction. *Proceedings of the SIGDIAL 2013 Conference* (August):122–126.

Stefan Ultes and Wolfgang Minker. 2014. Interaction quality estimation in spoken dialogue systems using hybrid-hmms. In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*. Association for Computational Linguistics, pages 208–217.

Stefan Ultes, María Jesús Platero Sánchez, Alexander Schmitt, and Wolfgang Minker. 2015. Analysis of an extended interaction quality corpus. In *Natural Language Dialog Systems and Intelligent Assistants*, Springer, pages 41–52.

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)*. Association for Computational Linguistics, Montréal, Canada, pages 49–52. http://www.aclweb.org/anthology/W12-1819.

Stefan Ultes, Alexander Schmitt, and Wolfgang Minker. 2013. On Quality Ratings for Spoken Dialogue Systems – Experts vs. Users. *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (June):569–578.

Stefan Ultes, Alexander Schmitt, and Wolfgang Minker. 2017b. Analysis of temporal features for interaction quality estimation. In *Dialogues with Social Robots*, Springer, pages 367–379.

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.

Marilyn A. Walker, Diane J. Litman, Candace A. Kamm, Alicia Abella, Marilyn A. Walker, Diane J. Litman, Candace A. Kamm, and Alicia Abella. 1997. Paradise. *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics* - pages 271–280. https://doi.org/10.3115/979617.979652.

Wojciech Zaremba, Ilya Sutskever, and Oriol Vinyals. 2014. Recurrent neural network regularization. *arXiv preprint arXiv:1409.2329* .