Improving Cross Database Prediction of Dialogue Quality Using Mixture of Experts

Klaus-Peter Engelbrecht, Hamed Ketabdar, Sebastian Möller

Deutsche Telekom Laboratories, Quality and Usability Lab, TU Berlin

{Klaus-Peter.Engelbrecht; Hamed.Ketabdar; Sebastian.Moeller}@telekom.de

Abstract

Models for the prediction of user judgments from interaction data can be used in different contexts such as system quality assessment, monitoring of deployed systems, or as a reward function in learned dialog managers. Such models still show a considerable lack with respect to their generalizability [6]. This paper specifically addresses this issue. We propose to use a Mixture of Experts approach for cross-database predictions. In Mixture of Experts, several classifiers are trained on subsets of the data showing specific characteristics. Predictions of each expert model are combined for the overall prediction result. We show that such an approach can improve the cross-database prediction accuracy.

Index Terms: user judgments, quality prediction, evaluation, spoken dialog systems

1. Introduction

Models for the prediction of user judgments about Spoken Dialog Systems from interaction data can be used in different contexts such as system quality assessment [1][2], monitoring of deployed systems [3], or as reward function in learned dialog managers [4]. Thus, there has been continuous interest in such models since the introduction of PARADISE [5]. The basic assumption of PARADISE is that user satisfaction is determined by the successful accomplishment of the desired task, as well as the costs associated with this. Both can be measured in terms of interaction parameters calculated from log files of the interaction or manual annotations of these. User satisfaction, in turn, can be measured with questionnaires. PARADISE now states that Linear Regression can be used to predict the user satisfaction judgments from the interaction parameters.

The accuracy of such models when applied to unseen data has been a main issue of concern. In particular, it has been shown that prediction accuracy decreases drastically if the predicted data stem from an independent database or system [6]. Walker et al. [7], in turn, showed that cross database predictions are possible without considerable loss of accuracy. However, the performance of the models cannot be granted, as there are contradicting results. In this paper, we specifically address this issue of model generalizability.

In a previous study [8], it was demonstrated how the focus of attention, guiding what aspects of the system determine the user judgment, changes depending on the severity of another aspect – in that case, the number of speech understanding errors. Thus the performance of judgment prediction models needs to take into account possible shifts in the users' attention.

One way to do this could be to create different models for different system types or contexts of usage, and select the model fitting the case to predict best. Ideally, this selection would be done automatically in a reliable way. As a first approach towards this, we propose in this paper to apply a concept from machine learning called Mixture of Experts. In the Mixture of Experts approach, different experts are specialized on different parts of a problem to be solved. Each expert provides a decision or estimation regarding the aspect of the problem for which it is designed. Outputs of different experts are then combined based on a proper fusion strategy to arrive at a final decision or estimation.

Mixtures of Experts have been applied in different fields in order to solve recognition, prediction or estimation related tasks. López-Cózar et al. [9] use a multi-classifier system for emotion detection in interaction data with Spoken Dialog Systems. Audio-visual speech recognition is another field of research in which the concept of Mixture of Experts is actively used [10]. In this case, one expert is specialized on speech recognition based on acoustic data, and a second expert is specialized on speech recognition based on visual data representing lip movements. The recognition results of the two experts are combined resulting in enhanced speech recognition performance. Multi-stream speech processing systems [11][12] can also be considered as Mixture of Experts structures in which different experts are used for extracting different types of acoustic features. In addition, different classifiers are used as experts modelling different parts of the feature space.

To date, only one approach to user judgment prediction using a multi-classifier system is known to us. Möller [13] divides the problem of dialog quality into different aspects, such as efficiency, or dialog symmetry. These aspects are taken from a taxonomy structuring the constituents of dialog quality. Relevant interaction parameters are identified as predictors for each aspect, and Linear Regression is utilized to make predictions of each aspect. Then, the overall satisfaction is predicted from the predictions for the single aspects using Linear Regression. However, this procedure did not improve the prediction accuracy. The approach proposed by us differs to that in [13], as in our work fusion is based on selection of individual results rather than their combination.

In the following sections, we explain in more depth the concept of Mixture of Experts and how it relates to the problem of cross-database predictions of user judgments. We then present and discuss results using 3 different databases to account for the models' ability to generalize across databases and systems.

2. Mixture of Experts for Dialogue Quality Prediction

In the Mixture of Experts approach, the initial problem is presented to different decision-taking components (experts). The experts should be designed in a way that allows them to handle different parts of the problem space. To achieve a prediction, each expert first makes a decision based on the presented problem at input. The decisions made by different

experts are then collected and combined according to a proper combination strategy, resulting in a final decision. This approach can also be viewed as a multi-stream information processing approach, where the output of each expert can be assumed as one information stream. Multi-stream systems take the advantage of obtaining information from multiple complementary sources of information (in our case, outputs of different experts) to arrive at a decision. The redundancy which exists in multi-stream systems makes them more robust against failures of some streams in the system. Moreover, they may result in improved performance when all the experts specialized on different tasks work reliably and their outputs are optimally combined.

Knowing the concept of Mixture of Experts, the next question is how to create such experts. In principle, the experts should be able to provide complementary information. In this work, our main aim is to increase generalizability of the dialog quality prediction model. One way to create suitable experts for this task is to train different experts with different databases. This increases the bias towards certain dialogue categories. When a new dialogue is presented to such a Mixture of Experts system, there is a chance that one of the experts is more specialized on this type of dialogue, resulting in a better prediction of the corresponding quality judgment as compared to using a single model trained on all data. This immediately leads to increased generalizability of the whole system in dealing with new dialogues of different sources.

Another way of creating proper experts is to use different machine learning approaches. Each machine learning model comes with different assumptions about the distribution of data (features), and hence performs better when exposed to data with a similar distribution. As different dialogue types may also come with different distributions of data, combining different modelling approaches as different experts increases the chance that a new dialogue matches one of the experts' feature modeling spaces.

A second key issue in the Mixture of Experts approach is the strategy for combination of the experts' outputs. The combination strategy should take advantage of complementary information in different experts' outputs efficiently. For this purpose, the combination strategy should be able to measure how reliable the output of each expert is for predicting the quality of dialogue. In Section 4, we present the fusion strategies used in this work.

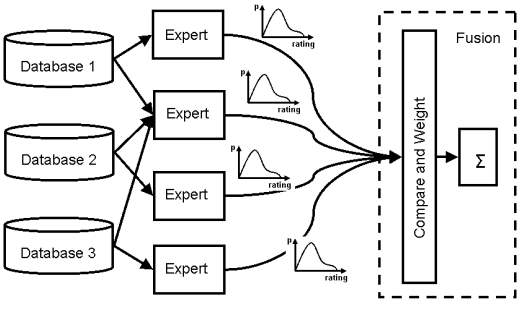

Figure 1. *Mixture of Experts as applied in this work.*

3. Databases

In order to see variability in the expert models, we employ data from 3 different databases. The most distinguishing features of each database will be described in this section.

Database 1 (D1) was acquired with 3 commercial spoken dialog systems from the public transport information domain. The systems differ in the deployed dialog strategy and system voice, as well as in how the desired information is provided. Fifteen users came to the lab and called each system through a regular telephone line, resulting in 45 calls.

Database 2 (D2) was collected with the Boris restaurant information system [13]. The system allows users to find a restaurant's name and address according to different criteria. The latter are collected in a mixed-initiative dialog, using keyword spotting for semantic analysis. Forty users spoke to the system in a Wizard-of-Oz experiment where the speech recognition was simulated by a human transcriber. Each user performed 5 tasks, leading to 200 rated dialogs.

Database 3 (D3) stems from an experiment with the INSPIRE smart home system [14]. The system allows speech control over lamps, TV, an electronic program guide, a fan, and an answering machine, resulting in a variety of simple and complex tasks. For the experiment, the system was set up in a living room environment inside a usability lab. Thus, users speak "into the room". Similar to Boris, the system features mixed-initiative and keyword spotting for semantic analysis, and the speech recognizer was replaced by a human transcriber. Fourteen users performed 14 tasks with the system, resulting in 196 dialogs.

In each experiment, we acquired user judgments of the "Overall Quality" using a scale recommended by the ITU-T Rec. P.851 [15]. This is a 5-point scale with extended margins. However, in order to avoid data sparsity problems when training the models, we decided to reduce the number of scale levels to 3:

- 1 = worse than "fair" (31%)
- 2 = "fair" (27%)
- 3 = better than "fair" (42%)

The annotations made to the corpora are given in Table 1. Some of them required manual annotation, while others could be calculated from the log files, if these were available.

4. Experiments and Results

4.1. Method

In order to get reliable estimations of the predictive power of the models, we evaluate them on unseen data using cross validation. As the relations between interaction parameters and judgments may well be user specific [16], we decided to leave out all data of one user at each iteration of the cross validation. Thus, for each result presented predictions were made for all users in the database.

To obtain a prediction, we train prediction models on each database, as well as on the data from all three databases together (Fig. 1). We explicitly include data from the database from which the test cases were taken in the training. The model trained on the same database may be more accurate than other models, which may improve the prediction result. However, our basic assumption when applying Mixture of Experts is that in most cases one model will fit the given test case better than any other model, and that we can make use of this by an appropriate model fusion algorithm.

As a performance baseline, we use the results obtained from the experts trained on a single database, as well as the result achieved with the model trained on all data except the test cases.

The results achieved with the Mixture of Experts method are compared to the baseline results using the Pearson correlation (r) and the mean absolute prediction error (MAE). In cases where MAE is smaller than the baseline, we can test for significance of this difference using a t-test.

In order to enable advanced fusion methods, we use classifiers which allow predicting a probability distribution of the possible ratings. Probabilistic models are appropriate for this case. We report results using two different models: Naïve Bayes and Markov Chains. Naïve Bayes takes as input interaction parameters as defined in Table 1 on the right hand side. While the model defines an expected distribution of the features for each class to be predicted, a prediction of the probability for each class C given the feature values F can be obtained by application of Bayes' rule:

$$P(C|F) = P(F|C)*P(C) / P(F).$$

Markov Chains, on the other hand, take as input the raw interaction features as listed on the left hand side of Table 1. Thus, each dialog is represented as a sequence of events. By evaluating the transition probabilities between different states for each class to predict, it is possible to determine the likelihood of a sequence to "belong to" this class [17].

$$P(C|\{f_1,\ldots f_n\}) = P(f_1 \rightarrow f_2|C) * \ldots * P(f_{n-1} \rightarrow f_n|C).$$

Table 1. *Interaction features and corresponding interaction parameters.*

Annotated Feature	Tags	Interaction Parameters
PARSING	▪ correct ▪ partially correct ▪ failed ▪ incorrect	▪ *#PA:CO* ▪ *#PA:PA* ▪ *#PA:FA* ▪ *#PA:IC*
#TURN	▪ current turn number	▪ *#TURNS*
#WORDS	▪ number of words in the system prompt	▪ *WPST*
CONFIRM	▪ explicit ▪ implicit ▪ none ▪ n.a.	▪ *#CONF:EX* ▪ *#CONF:IM* ▪ *#CONF:NO* ▪ *#CONF:NA*
SSA	▪ complex (ask for several constraints) ▪ simple (ask for 1 constraint) ▪ extra simple (ask for selection) ▪ provide info	▪ *#SSA:COM* ▪ *#SSA:SI* ▪ *#SSA:ES* ▪ *#SSA:INFO*
VOICE	▪ TTS ▪ CANNED	▪ *%CANNED*
TS	▪ success ▪ failure	▪ *TS*

4.2. Results

The results for both modeling approaches are presented in Table 2 and Table 3 respectively. The first four rows of each table list the baseline results as defined above. Next are listed the result achieved by simply combining all predicted distributions with equal weights. The predicted value is obtained from the summed probability distributions by calculating the mean value μ:

$$\mu = (P(1|F)*1 + P(2|F)*2 + P(3|F)*3) / 3.$$

We also tested more complex fusion rules based on the entropy of the predicted distributions or the agreement of the predictions. Both measures allow weighting the contribution of each expert to the overall prediction result. However, while entropy measures the decisiveness of the prediction, the agreement of the predictions implements a majority voting approach. As the entropy-based fusion results were clearly below the baseline, the tables include only results from the agreement-based fusion.

The basis of the agreement-based fusion is to determine which class was predicted by most experts with the highest probability. Then either this mode is chosen as the predicted value ("most frequent value"), or the distributions which agreed on this mode are summed and the mean value is calculated ("Mean of distributions with most frequent mode"). In addition, we can either include the model trained on all databases, or exclude it, as it is not strictly an "expert" in our sense.

For both classifiers, the prediction made by the Mixture of Experts is seconded by one of the plain models. However, in practice, the prediction of the Mixture of Experts is the best choice as it is more reliable and still very accurate. While training only with D3 yields a very good result for the Naïve Bayes model, for the Markov Chain model the result is relatively weak. The baseline model trained on all databases in conjunction performs well throughout and is thus reliable; however, in our trials, the Mixture of Experts combining all four baseline models was comparable or better (in case of Naïve Bayes, a test of the hypothesis that the prediction error is smaller than the baseline's error yielded $t(415)=4.8767$; $p<0.01$).

Table 2. *Results with the Naïve Bayes model. Given are Pearson's r and the mean absolute error for the baseline and several Mixture of Experts models.*

Model	*r*	*MAE*
D1	-0.24	1.11
D2	0.38	0.65
D3	**0.51**	**0.56**
D1+2+3	**0.42**	**0.70**
{D1, D2, D3, D1+2+3}	0.40	0.69
Most freq. mode {D1, D2, D3}	0.44	0.57
Most freq. mode {D1, D2, D3, D1+2+3}	0.41	0.60
Mean of distr. with most freq. mode {D1, D2, D3}	**0.47**	**0.58**
Mean of distr. with most freq. mode {D1, D2, D3, D1+2+3}	0.45	0.61

Table 3. *Results with the Markov Chain model. Given are Pearson's r and the mean absolute error for the baseline and several Mixture of Experts models.*

Model	*r*	*MAE*
D1	0.32	0.70
D2	0.23	0.76
D3	0.30	0.70
D1+2+3	**0.46**	**0.60**
{D1, D2, D3, D1+2+3}	0.44	0.66
Most freq. mode {D1, D2, D3}	0.27	0.73
Most freq. mode {D1, D2, D3, D1+2+3}	0.43	0.57
Mean of distr. with most freq. mode {D1, D2, D3}	0.33	0.70
Mean of distr. with most freq. mode {D1, D2, D3, D1+2+3}	**0.46**	**0.61**

Table 4. *Results achieved with mixtures of all models. Given are Pearson's r and the mean absolute error for the baseline and the best fusion methods.*

Model	*r*	*MAE*
D1+2+3	0.46	0.60
Most freq. mode {D1MC, D2MC, D3MC, D1+2+3MC, D1NB, D2NB, D3NB, D1+2+3NB}	0.46	0.57
Mean of distr. with most freq. mode {D1MC, D2MC, D3MC, D1+2+3MC, D1NB, D2NB, D3NB, D1+2+3NB}	0.47	0.61
{D1+2+3MC, D1+2+3NB}	0.47	0.65

Lastly, we tested if a mixture of the different modelling approaches trained on each of the databases can further improve the prediction. The results are somewhat controversial, as we need to decide for either higher correlation or a lower prediction error (Table 4). However, the results still show that combining different model algorithms does not harm, but has the potential to improve the prediction.

Overall, the best fusion method seems to be the mean of the sum of those distributions agreeing on the most frequent mode. The models trained on single databases as well as those trained on their combination should be included in the mixture.

5. Conclusion and Future Work

In this paper, we introduced a new paradigm for the improvement of prediction models for user judgments on Spoken Dialog Systems. In particular, we aimed at an improved performance on dialogs from unseen databases or systems. In applying the concept of Mixture of Experts with relatively straightforward fusion strategies, we could show that this approach is useful for increasing the reliability – and to some degree the accuracy – of such prediction models.

As a by-product to our work, we found that a relatively small amount of additional training data from different databases has a positive effect on the model quality. While intuitively it seems trivial that the model becomes better with more training data, previous results (e.g. [6]) suggested that far more data are needed to improve cross database predictions, as the relationships between interaction parameters and user judgments seemed too diverse.

Unfortunately, the results achieved even with Mixture of Experts are still too low for practical applications. Therefore, we plan to improve the models in some details. For example, the individual models could be optimized on unseen data to improve their contribution to the Mixture of Experts. In addition, there are a number of interaction parameters which we did not use in this work, but which could provide informative input for the prediction.

Also, it would be interesting to see how the performance of the models develops if more databases are used for the training of individual experts and a combined model. What the results in this paper definitely suggest is that by adding data of a few databases the prediction generalizability problem might be solved.

References

[1] H. Ai, F. Weng, "User Simulation as Testing for Spoken Dialog Systems," in *Proc. of SIGdial*, 2008, pp. 164-171.

[2] S. Möller, R. Englert, K.-P. Engelbrecht, V. Hafner, A. Jameson, A. Oulasvirta, et al., „MeMo: Towards Automatic Usability Evaluation of Spoken Dialogue Services by User Error Simulations," in *Proc. of Interspeech*, 2006, pp. 1786-1789.

[3] K. Evanini, P. Hunter, J. Liscombe, D. Suendermann, K. Dayanidhi, R. Pieraccini, "Caller Experience: A Method for Evaluating Dialog Systems and its Automatic Prediction," in *Proc. of SLT*, 2008, pp. 129-132.

[4] V. Rieser, O. Lemon, "Automatic Learning and Evaluation of User-Centered Objective Functions for Dialogue System Optimisation," in *Proc. of LREC*, 2008, pp. 2356-2361.

[5] M. Walker, D. Litman, C. Kamm, A. Abella, "PARADISE: A Framework for Evaluating Spoken Dialogue Agents," in *Proc. of ACL/EACL*, 1997, pp. 271–280.

[6] S. Möller, K.-P. Engelbrecht, R. Schleicher, "Predicting the Quality and Usability of Spoken Dialogue Services," *Speech Communication*, vol. 50, pp. 730-744, 2008.

[7] M. Walker, C. Kamm, D. Litman, „Towards Developing General Models of Usability with PARADISE," *Natural Language Engineering*, vol. 6, no. 3-4, pp. 363-377, 2000.

[8] K.-P. Engelbrecht, C. Kuehnel, S. Möller, "Weighting the Coefficients in PARADISE Models to Increase Their Generalizability," in *Proc. of PIT*, 2008, pp. 289-292.

[9] R. López-Cózar, Z. Callejas, M. Kroul, J. Nouza, and J. Silovský, "Two-Level Fusion to Improve Emotion Classification in Spoken Dialogue Systems," in: P. Sojka, et al. (Eds.): *Text Speech, and Dialogue*, Springer, Berlin, 2008.

[10] S. Dupont and J. Luettin, "Audio-visual Speech Modeling for Continuous Speech Recognition," IEEE Transactions on Multimedia, vol. 2, no. 3, pp. 141-151, 2000.

[11] H. Bourlard, and S. Dupont, "Sub-band-based Speech Recognition," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing*, 1997, pp. 1251-1254.

[12] K. Kirchhoff, "Combining Articulatory and Acoustic Information for Speech Recognition in Noisy and Reverberation Environments," in *Proc. Int. Conf. on Spoken Language*, 1998, pp. 891-894.

[13] S. Möller, *Quality of Telephone-based Spoken Dialog Systems*. New York: Springer, 2005.

[14] F. Goedde, S. Moeller, K.-P. Engelbrecht, C. Kuehnel, R. Schleicher, A. Naumann, and M. Wolters, „Study of a Speech-based Smart Home System with Older Users," in *Proc. Of IUI4AAL*, 2008, Frauenhofer IRB Verlag, Stuttgart.

[15] ITU-T Rec. P.851, *Subjective Quality Evaluation of Telephone Services Based on Spoken Dialogue Systems*, International Telecommunication Union, Geneva, 2003.

[16] K.-P. Engelbrecht, S. Möller, R. Schleicher, I. Wechsung, "Analysis of PARADISE Models for Individual Users of a Spoken Dialog System," in *Proc. of ESSV*, 2008, pp. 86-93.

[17] K.-P. Engelbrecht and S. Möller, "Sequential Classifiers for the Prediction of User Judgments about Spoken Dialog Systems," *Speech Communication*, submitted.