UDC 519.234

S. Ultes, A. Schmitt, R. ElChab, W. Minker

STATISTICAL MODELING OF INTERACTION QUALITY IN SPOKEN DIALOGUE SYSTEMS: A COMPARISON OF (CONDITIONED) HIDDEN MARKOV MODEL-BASED CLASSIFIERS VS. SUPPORT VECTOR MACHINES

The Interaction Quality (IQ) metric has recently been introduced for measuring the quality of an interaction with a Spoken Dialogue System (SDS). The metric allows for an estimation of a quality score at arbitrary points in a spoken human-machine interaction. While previous work relied on Support Vector Machines (SVMs) for classifying the score based on a static feature vector representing the entire previous interaction, we evaluate a Conditioned Hidden Markov Model (CHMM) which accounts for the sequential character of the data and, in contrast to a regular Hidden Markov Model (HMM), provides class probabilities. The results show that a CHMM achieves an Unweighted Average Recall (UAR) of 0.39. Thereby it is outperformed by a regular HMM with an UAR of 0.44 and an SVM with an UAR of 0.49, both trained and evaluated under the same conditions.

Keywords: interaction quality, support vector machines.

To evaluate the quality of Spoken Dialogue Systems (SDSs), different measures exist. Unfortunately, objective metrics like, e. g., task completion or dialogue duration are not human-centered. Subjective measures compensate for this by modeling the user's subjective experience.

However, in human-machine dialogues, there is no easy way of deriving the user's satisfaction level. Furthermore, a regular user does not want to spend time answering questions about the performance of the system. Human-machine dialogues usually have no conversational character but are task oriented.

Therefore, approaches for determining the satisfaction level automatically have been under investigation for several years, most prominently the PARADISE framework by Walker et al. [1]. Assuming a linear dependency between objective measures and User Satisfaction (US), a linear regression model is applied to determine US on the dialogue level. This is not only very costly, as dialogues must be performed with real users, but also inadequate if quality on a finer level is of interest, e.g., on the exchange level. To overcome this issue, work by Schmitt et al. introduced a new metric for measuring the performance of a SDS on the exchange level called Interaction Quality (IQ) [2].

Human-machine dialogues may be regarded as a process evolving over time. A well-known statistical method for modeling such processes is the Hidden Markov Model (HMM). Since HMMs do not provide class probabilities, we present an approach for determining IQ using Conditioned Hidden Markov Models (CHMMs). They were originally introduced by Glodek et al. [3] who applied the model to laughter detection on audio-visual data.

In Section 2, we discuss other work on determining qualitative performance of SDSs and in Section 3 we present details about the definition of IQ and the data we use. Further, Section 4 presents a formal description of the CHMM. Evaluation is described in Section 5 and, finally, Section 6 concludes this work.

Related Work. Work on determining User Satisfaction using HMMs was performed by Engelbrecht et al. [4]. They predicted US at any point within the dialogue on a five-point scale. Evaluation was performed based on labels the users applied themselves during a Wizard-of-Oz experiment. The dialogue course paused during labeling. They achieved a Mean Squared Error of 0.086.

Further work which incorporates HMMs was presented by Higashinaka et al. [5]. The HMM was trained on US ratings at each exchange which were derived from ratings for the whole dialogue. The authors compare their approach with HMMs trained on manually annotated exchanges achieving a better performance for the latter.

Higashinaka et al. also present work on the prediction of turn-wise ratings for human-human (transcribed conversation) and human-machine (text dialogue from chat system) dialogues [6].

Ratings ranging from 1-7 were applied by two expert raters labeling smoothness, closeness, and willingness.

Dealing with true User Satisfaction, Schmitt et al. presented their work about statistical classification methods for automatic recognition of US [7]. The data was collected in a lab study where the users themselves had to rate the conversation during the ongoing dialogue. Labels were applied on a scale from 1 to 5. By applying a Support Vector Machine (SVM), they achieved an Unweighted Average Recall (UAR) of 49.2.

Interaction Quality. For Interaction Quality recognition, we use the LEGO corpus published by Schmitt et al. [8]. It is based on 347 calls to the "Let's Go Bus Information System" of the Carnegie Mellon University in Pittsburgh [9] recorded in 2006. Labels for IQ have been assigned by three expert raters to 200 calls consisting of 4,885 exchanges in total. IQ was labeled on a scale from 1 (extremely unsatisfied) to 5 (satisfied). As the users are expected to be satisfied at the beginning, each dialogue's initial rating is 5.

Parameters used as input variables for the IQ model have been derived from the dialogue system modules automatically for each exchange. Further, parameters on three levels have been created: the exchange level, the dialogue level, and the window level.

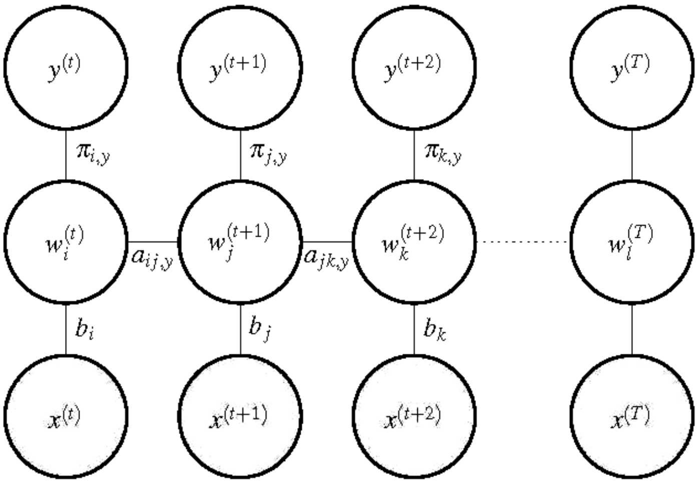

Fig. 1. General graphical representation of the CHMM model in the discrete time domain. For each time step t , $y^{(t)}$ represents the most likely label and $w_i^{(t)}$ the most likely hidden state given observation $x^{(t)}$. b_i represents the probability for the observation and $\pi_{i,y}$ the label probability. $a_{ij,y}$ defines the probability of transitioning from state $w_i^{(t)}$ to state $w_i^{(t+1)}$

As parameters like ASRCONFIDENCE or UTTERANCE can directly be acquired from the dialogue modules they constitute the exchange level. Based on this, counts, sums, means, and frequencies of exchange level parameters from multiple exchanges are computed to constitute the dialogue level (all exchanges up to the current one) and the window level (the three previous exchanges).

Schmitt et al. [2] performed IQ recognition on this data using SVMs. They achieved an Unweighted Average Recall (UAR) of 0.58.

CHMM. Conditioned Hidden Markov Models [3] are an extension of regular HMMs. They provide probabilities for multiple classes. A sequence diagram illustrating the principle operation method of the CHMM in the time domain is shown in Figure 1.

Model Description. Like the continuous HMM, the CHMM also consists of a discrete set of hidden states $w_i \in W$ and a vector space of observations $X \subseteq R^n$. A separate emission probability $b_i(x^{(t)})$ is linked to each state defining the likelihood of observation $x^{(t)} \in X$ at time t while being in state w_i. Further, $a_{ij,y} = p(w^{(t)} = w_j \mid w^{(t-1)} = w_i, y^{(t)} = y)$ defines the transition probability of transitioning from state w_i to w_j. In contrast to the regular HMM, the transition probability distribution also depends on the class label $y \in Y$. This results in the transition matrix $A \in R^{|W|\times|W|\times|Y|}$.

Furthermore, the meaning of the initial probability $\pi_{i,y} = p(w^{(1)} = w_i \mid y^{(1)} = y)$ for state w_i is altered. It additionally represents the label probability for label y at any time with the corresponding matrix $\pi \in R^{|W|\times|y|}$. An schematic example of a CHMM with two labels and three hidden states is illustrated in Figure 2.

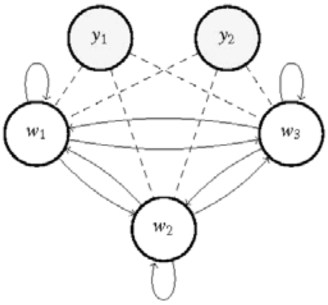

Fig. 2. This is an example of a CHMM with 2 labels and 3 hidden states. The dashed lines representthe label dependence of the hidden states, while the full lines illustrate state transitions. Please note that state transitions also depend on the labels which is not shown here

According to Glodek et al. [3], the likelihood of an observation sequence $x^{(n)}$ with corresponding label sequence $y^{(n)}$ is given by

$$\begin{aligned} p(x^{(n)}, w^{(n)} \mid y^{(n)}, \lambda) &= \sum_{w \in W} p(w^{(1)} = w \mid y^{(1)}, \pi) \\ &\cdot \prod_{t=2}^{T} p(w^{(t)} = w_j \mid w^{(t-1)} = w_i, y^{(t)}, A) \\ &\cdot \prod_{t=1}^{T} p(x^{(t)} \mid w^{(t)} = w_j, \theta), \end{aligned} \quad (1)$$

where $w^{(n)}$ denotes the sequence of the hidden states. Further, in this work, the emission probability $b_j(x^{(t)}) = p(x^{(t)} \mid w^{(t)} = w_j, \theta)$ is modeled as a Gaussian Mixture Model (GMM) with the parameter set $\theta = \{\{\varphi_{j,k}\}_k^K, \{\mu_{j,k}\}_k^K, \{\Sigma_{j,k}\}_k^K\}$. The parameter set λ

describing the complete CHMM is defined by $\lambda = \{\pi, \mathrm{A}, \theta\}$.

Learning. The learning phase consists of two parts: *initialization* and *training*.

For initialization, the *k*-means algorithm [10] is used and the number of clusters k corresponds to the number of hidden states. After clustering initial observation sequences with their corresponding label sequences, the transition probabilities are updated according to the transitions between the clusters, given the labels. The initial probabilities are updated according to the cluster and the corresponding label that each element belongs to.

Training is performed using the Baum-Welch algorithm, which is heavily dependent on the initialization. When comparing the HMM explained by Rabiner et al. [11] to the CHMM, several changes (Changes in Eq.: 19, 20, 24, 25, 27, 37, 40a, 40b and 40c from [11]) must be applied to the Baum-Welch algorithm.

The αs and βs of the Forward-Backward algorithm as given by Glodek et al. [3] are

$$a_{t,y}(j) = b_j(x^{(t)}) \cdot \sum_{i \in W} a_{ij,y} \cdot a_{t-1,y}(i) \tag{2a}$$

$$a_{1,y}(j) = b_j(x^{(1)}) \cdot \pi_{j,y} \tag{2b}$$

$$\beta_{t,y}(i) = \sum_{j \in W} a_{ij,y} \cdot b_j(x^{(t+1)}) \cdot \beta_{t+1,y}(j) \tag{3a}$$

$$\beta_{T,y}(i) = 1 \tag{3b}$$

$$\beta_{0,y}(i) = \sum_{j \in W} \pi_{j,y} \cdot b_j(x^{(1)}) \cdot \beta_{1,y}(j) \tag{3c}$$

The state beliefs $\gamma_{t,y}(j)$ and the transition beliefs $\xi_{t-1,t,y}(i,j)$ are then computed by using

$$\gamma_{t,y}(j) = \frac{\alpha_{t,y}(j) \cdot \beta_{t,y}(j)}{p(X)} \tag{4}$$

$$\xi_{t-1,t,y}(i,j) = \frac{\alpha_{t-1,y}(i) \cdot b_j(x^{(t)}) \cdot a_{ij,y} \cdot \beta_{t,y}(j)}{p(X)} \tag{5}$$

where $\sum_{t=1}^{T-1} \gamma_{t,y}(i)$ is the expected number of transitions from w_i given y and $\sum_{t=1}^{T-1} \xi_{t-1,t,y}(i,j)$ is the expected number of transitions from w_i to w_j given y.

Parameter learning is performed after evaluation of N sequences, updating the initial probabilities using the following formula

$$\pi_{i,y} = \frac{\begin{array}{c}\textit{expected number of times of being} \\ \textit{in } w_i \textit{ at time } t = 1 \textit{ given } y\end{array}}{\begin{array}{c}\textit{expepcted number of times of being} \\ \textit{in all } w \textit{ at } t = 1 \textit{ given } y\end{array}} = \tag{6}$$

$$= \frac{\sum_{l=1}^{N} \delta_{y^{(1)(l)}=y} \gamma_{1,y^{(n)(l)}}(i)}{\sum_{l=1}^{N} \sum_{j \in w_1} \delta_{y^{(1)(l)}=y} \gamma_{1,y^{(n)(l)}}(j)}$$

where $\sum_{i=1}^{n} \pi_{i,y} = 1$ and δ is the Kronecker delta.

The update for the transition probabilities after evaluating N sequences is

$$a_{ij,y} = \frac{\textit{expected number of transitions from } w_i \textit{ to } w_j \textit{ given } y}{\textit{expected number of transitions from } w_i \textit{ given } y} =$$

$$= \frac{\sum_{l=1}^{N} \sum_{t=0}^{T-1} \xi_{t-1,t,y^{(n)(l)}}(i,j) \delta_{y^{(t)(l)}=y}}{\sum_{l=1}^{N} \sum_{t=0}^{T-1} \gamma_{t,y^{(n)(l)}}(j) \delta_{y^{(t)(l)}=y}} \tag{7}$$

where

$$\forall_{y \in Y} \sum_{j=1}^{n} a_{ij,y} = 1.$$

The emission probabilities can be computed in accordance with the methods presented by Rabiner et al. [11]. As the state beliefs depend on y, a sum over all labels has to be applied in order to create label independent emission probabilities.

Evaluation. The Viterbi algorithm generates a sequence of expected labels which are evaluated with the metrics, which described in the following.

Results for CHMM experiments according to the number of hidden states along with results for regular HMM and SVM classification. The '*' indicates the best result

Class	# states	UAR	Kappa	Rho
SMO	-	0.49	0.61	0.77
HMM	5	0.44	0.56	0.72
CHMM	5	0.38	0.40	0.56
	6	0.38	0.39	0.57
	7	0.35	0.40	0.59
	8	0.37	0.41	0.59
	9*	0.39	0.43	0.60
	10	0.37	0.39	0.55
	11	0.36	0.41	0.58

Metrics. The **Unweighted Average Recall** (UAR) for multi-class classification problems is the accuracy corrected by the effects of unbalanced data.

To measure the relative agreement between two corresponding sets of ratings we apply **Cohen's Kappa** [12]. It is defined by 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. In order to take account for ordinal scores, a weighting factor w is introduced reducing the discount of disagreements the closer the ratings are together [13]:

$$w = \frac{|r_1 - r_2|}{range} \tag{8}$$

Here, r_1 and r_2 denote the rating pair and *range* the maximum distance which may occur between two ratings. This results in $w = 0$ for agreement and $w = 1$ if the ratings differ the most.

For measuring the correlation between two variables, *Spearman's Rank Correlation Coefficient* is used [14]. It is a non-parametric method assuming a monotonic function between the two variables, defined by

$$\rho = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}, \tag{9}$$

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

Setup and Results. For the experiments, we used the LEGO corpus presented in Section 3. Since the values of multiple parameters are constant for most exchanges, they are excluded. Otherwise, this would have resulted in rows of zeros during computation of the covariance matrices of the feature vectors. A row of zeros in the covariance matrix will make it irreversible, which will cause errors during the computation of the emission probabilities.

The model operated with a vector of 29 dimensions. Results of the experiments are presented in Table 1. The data is ranked according to the number of hidden states used for the model. The accuracy decreased remarkably after passing the threshold of 9 states, where the highest values for UAR, κ, and ρ could be achieved.

The results are computed using 6-fold cross validation. When evaluating the performances for each fold, best performance was achieved for 9 states with an UAR of 0.45, Cohen's κ of 0.58, and Spearman's ρ of 0.74.

To define a baseline, we rely on the approach by Schmitt et al. [2]. Using the same features, we trained a Support Vector Machine (SVM) with a linear kernel. The results are shown in Table. Unfortunately, the CHMM approach was not able to outperform the baseline. This is most likely caused by the fact that only little training data was available.

Furthermore, we conducted an experiment using regular HMMs. Using 5 hidden states, we assigned a label to each hidden state. As depicted in Table 1, the CHMM performed worse than the HMM approach. Though performance of both models is dramatically influenced by the lack of data, the CHMM is rather prone to this, as (*number of labels*) · (*number of hidden states*) additional probability models have to be estimated.

As dialogues have a sequential structure, an approach for estimating Interaction Quality on the exchange level has been evaluated using Conditioned Hidden Markov Models. Experiments were conducted for measuring its performance. The best result with 9 hidden states is outperformed vastly by previously presented methods based on SVM classification. We identified the lack of training data as the cause for this.

References

1. Walker M., Litman D., Kamm C. A., Abella A., Paradise: a framework for evaluating spoken dialogue agents // Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics. Morristown, NJ, USA: Association for Computational Linguistics, 1997. P. 271–280.

2. Schmitt A., Schatz B., Minker W. Modeling and predicting quality in spoken humancomputer interaction // Proceedings of the SIGDIAL 2011 Conf. Portland, Oregon : Association for Computational Linguistics, Jun. 2011.

3. Glodek M., Scherer S., Schwenker F. Conditioned hidden markov model fusion for multimodal classification // Proceedings of the 12th Annual Conf. of the Intern. Speech Communication Association (INTERSPEECH 2011). International Speech Communication Association, Aug. 2011. P. 2269–2272.

4. Engelbrecht K.-P., G¨odde F., Hartard F., Ketabdar H., M¨oller S. Modeling user satisfaction with hidden markov model // SIGDIAL '09: Proc. of the SIGDIAL 2009 Conf. Morristown, NJ, USA: Association for Computational Linguistics, 2009. P. 170–177.

5. Higashinaka R., Minami Y., Dohsaka K., Meguro T. Modeling user satisfaction transitions in dialogues from overall ratings // Proceedings of the SIGDIAL 2010 Conf. Tokyo : Association for Computational Linguistics, Sep. 2010. P. 18–27.

6. Higashinaka R., Minami Y., Dohsaka K. Meguro T. Issues in predicting user satisfaction transitions in dialogues: Individual differences, evaluation criteria, and prediction models // Spoken Dialogue Systems for Ambient Environments, Lecture Notes in Computer Science, in G. Lee, J. Mariani, W. Minker, and S. Nakamura, Eds. Springer Berlin : Heidelberg. 2010. Vol. 6392. P. 48–60.

7. Schmitt A., Schatz B., Minker W. A statistical approach for estimating user satisfaction in spoken human-machine interaction // Proceedings of the IEEE Jordan Conf. on Applied Electrical Engineering and Computing Technologies (AEECT). Amman : IEEE, Dec. 2011.

8. Schmitt A., Ultes S., Minker W. A parameterized and annotated corpus of the cmu let's go bus information system // Intern. Conf. on Language Resources and Evaluation (LREC), in-press.

9. Raux A., Bohus D., Langner B., Black A. W., Eskenazi M. Doing research on a deployed spoken dialogue system: One year of lets go! experience // Proc. of the Intern. Conf. on Speech and Language Processing (ICSLP), Sep. 2006.

10. Faber V. Clustering and the continuous k-means algorithm. Los Alamos Science. 1994. № 22. P. 138–144.

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

12. Cohen J. A coefficient of agreement for nominal scales // Educational and Psychological Measurement. 1960. Vol. 20, № 1. Apr. P. 37–46.

13. Cohen J. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit // Psychological bulletin. 1968. Vol. 70, № 4. P. 213.

14. Spearman C. The proof and measurement of association between two things // American Journal of Psychology. 1904. Vol. 15. P. 88–103.

С. Ультес, А. Шмитт, Р. Эль Хаб, В. Минкер

СТАТИСТИЧЕСКОЕ МОДЕЛИРОВАНИЕ КАЧЕСТВА ВЗАИМОДЕЙСТВИЯ В РЕЧЕВЫХ ДИАЛОГОВЫХ СИСТЕМАХ: СРАВНЕНИЕ КЛАССИФИКАТОРОВ, ОСНОВАННЫХ НА (УСЛОВНЫХ) СКРЫТЫХ МАРКОВСКИХ МОДЕЛЯХ, И МАШИН ОПОРНЫХ ВЕКТОРОВ

В последнее время была представлена метрика качества диалога для оценки качества взаимодействия с языковой диалоговой системой. Эта метрика позволяет получить оценку качества в произвольный момент взаимодействия человека с машиной. В то время как предыдущая работа базировалась на методе опорных векторов (SVM) для классификации качества взаимодействия на основе статического характеристического вектора, представляющего всю предысторию взаимодействия, здесь мы исследуем условные скрытые марковские модели (CHMM), которые принимают во внимание последовательный характер данных и, в отличие от стандартных скрытых марковских моделей (HMM), высчитывают вероятности классов. Экспериментальные результаты показали, что CHMM достигла значения невзвешенного среднего вызова (UAR) равного 0.39. Таким образом, алгоритм уступает HMM с UAR равным 0.44 и SVM с UAR равным 0.49. Все алгоритмы тренировались и исследовались в равных условиях.

Ключевые слова: качество взаимодействия, машины опорных векторов.

UDC 519.234

K. V. Zablotskaya, S. G. Zablotskiy, F. Fernández-Martínez, W. Minker

LANGUAGE STYLE MATCHING AND VERBAL INTELLIGENCE

In this paper language style matching of speakers yielding different verbal intelligence was analyzed. The work is based on a corpus consisting of 100 descriptions of a short film (monologues), 56 discussions about the same topic (dialogues) and verbal intelligence scores of the test persons. According to the results, higher verbal intelligent speakers showed a greater degree of language style matching when describing the film and were able to better adapt to their dialogue partners compared to lower verbal intelligent participants.

Keywords: spoken dialogue system, linguistic analysis, ANOVA .

Statistical approaches are often applied to text analysis and information retrieval problems. For example, TF-IDF measures may be used for classification of documents into a fixed number of predefined categories. Comparing a document with special dictionaries may be helpful for its content and semantic analysis. Linguistic analysis of texts allows researchers to determine additional information about authors: age, social status, emotions, psychological state, etc. In this research we applied a relatively new statistical method, tokens n-gram distributions, to the analysis of language style matching (LSM).

When two speakers are talking to each other, they try to adapt to their dialogue partner and to somehow synchronize their verbal behaviors. This phenomenon was investigated in [1]. In [2] linguistic style matching in human-human conversations was analyzed. For the linguistic analysis all the utterances were compared with a special dictionary which contained words sorted by a number of categories. The usage of each category was analyzed on conversation and turn-by-turn levels and showed that speakers synchronized their words when talking to each other. In [3] college students were asked to write answers to several questions formulated in different styles. It was shown that students followed the language styles of written questions.

In this paper we analyzed LSM of speakers with different verbal intelligence. This investigation may be helpful for improvement of user-friendliness of spoken language dialogue systems (SLDSs). SLDSs which automatically adapt to users' language styles and change their dialogue strategies may help users to feel more comfortable when interacting with them. However, it is necessary to know how different speakers change their own language styles in order to adapt to their dialogue partners. In [3] it was shown that students with higher grades matched the linguistic styles of asked questions closer than other students participated in the experiment. In this research we analyzed spoken utterances of people with different verbal intelligence. In the first part of this research we analyzed similarity between language styles of verbal descriptions of a short film (monologues) made by test persons who participated in our experiments (German