Modeling User Satisfaction Transitions in Dialogues from Overall Ratings

Ryuichiro Higashinaka†, **Yasuhiro Minami**‡, **Kohji Dohsaka**‡, and **Toyomi Meguro**‡
† NTT Cyber Space Laboratories, NTT Corporation
‡ NTT Communication Science Laboratories, NTT Corporation
higashinaka.ryuichiro@lab.ntt.co.jp
{minami,dohsaka,meguro}@cslab.kecl.ntt.co.jp

Abstract

This paper proposes a novel approach for predicting user satisfaction transitions during a dialogue only from the ratings given to entire dialogues, with the aim of reducing the cost of creating reference ratings for utterances/dialogue-acts that have been necessary in conventional approaches. In our approach, we first train hidden Markov models (HMMs) of dialogue-act sequences associated with each overall rating. Then, we combine such rating-related HMMs into a single HMM to decode a sequence of dialogue-acts into state sequences representing to which overall rating each dialogue-act is most related, which leads to our rating predictions. Experimental results in two dialogue domains show that our approach can make reasonable predictions; it significantly outperforms a baseline and nears the upper bound of a supervised approach in some evaluation criteria. We also show that introducing states that represent dialogue-act sequences that occur commonly in all ratings into an HMM significantly improves prediction accuracy.

1 Introduction

In recent years, there has been intensive work on the automatic evaluation of dialogues (Walker et al., 1997; Möller et al., 2008). Automatic evaluation makes it possible to predict the performance of dialogue systems without the costly process of performing surveys with human subjects, leading to a rapid improvement cycle for dialogue systems. It is also useful for detecting problematic situations in an ongoing dialogue (Walker et al., 2002; Herm et al., 2008; Kim, 2007). In these studies, the typical approach is to train a prediction model, such as a regression or classification model, using features representing the whole or a part of a dialogue together with human reference labels (e.g., reference ratings). However, creating such reference labels by hand can be extremely costly when we want to predict user satisfaction transitions during a dialogue because we need to create reference labels after each utterance/dialogue-act in the training data (Engelbrecht et al., 2009).

This paper proposes a novel approach for predicting user satisfaction transitions during a dialogue only from the dialogues with overall ratings. The approach makes it possible to avoid creating reference labels for utterances/dialogue-acts and only requires a single reference label for each dialogue. More specifically, we predict the user satisfaction rating after each dialogue-act in a dialogue only by using dialogues with dialogue-level (overall) user satisfaction ratings as training data. Our basic approach is to train hidden Markov models (HMMs) of dialogue-act sequences associated with each overall rating and combine such rating-related HMMs into a single HMM. We use this combined HMM to decode a sequence of dialogue-acts by the Viterbi algorithm (Rabiner, 1990) into state sequences that indicate from which rating-related HMM each dialogue-act is most likely to have been generated, leading to our rating predictions for the dialogue-acts. This paper experimentally examines the validity of our approach and explores several model topologies for possible improvement.

In Section 2, we review related work on automatic evaluation of dialogues. In Section 3, we describe our approach in detail. In Section 4, we describe the experiment we performed to verify our approach and present the results. In Section 5, we summarize and mention future work.

2 Related Work

Regression models are typically utilized for evaluating the quality of an entire dialogue. Most famously, the PARADISE framework (Walker et al., 1997) learns from data a linear regression model that predicts dialogue-level user satisfaction from various objective characteristics of a dialogue that concern task success and dialogue costs. This framework is widely used today and a number of extensions have been proposed to improve the prediction performance (Möller et al., 2008); how-

Proceedings of SIGDIAL 2010: the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 18–27,
The University of Tokyo, September 24-25, 2010. ©2010 Association for Computational Linguistics

ever, it is not aimed at predicting user satisfaction transitions.

Classification models are widely employed to detect problematic situations in an ongoing dialogue. Walker et al. (2002) developed the Problematic Dialogue Predictor for the “How May I Help You” system (Gorin et al., 1997) to robustly transfer problematic calls to human operators in call routing tasks. They derive speech recognition, language understanding, and dialogue management features from the first few turns of a dialogue and apply a decision tree classifier to detect problematic calls. For a similar task, Hirschberg et al. (2004) and Herm et al. (2008) used prosodic and emotional features. Kim (2007) recently proposed an approach for online call quality monitoring so that problematic calls can be transferred to human operators as quickly as possible rather than waiting for the first few turns.

N-grams and HMM-based approaches have also been actively studied. Hara et al. (2010) proposed predicting the most likely user satisfaction level of a dialogue by using N-grams of dialogues for each satisfaction level in the music navigation domain. Isomura et al. (2009) used HMMs to evaluate the naturalness of a dialogue in their interview system. They trained HMMs that model dialogue-act sequences between human subjects and used them to evaluate human-machine dialogues by the output probabilities of the HMMs. Recently, there have been approaches to predict user satisfaction transitions by evaluating the quality of individual utterances in a dialogue. For example, Engelbrecht et al. (2009) predicted user satisfaction ratings after each user utterance by HMMs trained from utterance-level features and utterance-level reference ratings.

The problem with these approaches is that they require a lot of training data, especially when we want to predict the quality of smaller units such as utterances. Our aim is to reduce such cost. Our work is similar to Engelbrecht’s work (Engelbrecht et al., 2009) in that we use HMMs to predict user satisfaction transitions during a dialogue but different in that we only use dialogue-level ratings to model dialogue-act-level user satisfaction transitions.

3 Approach

We aim to predict user satisfaction transitions only from dialogues with overall ratings. More formally, given a dialogue d_i of a set of dialogues D $(= \{d_1 \dots d_N\})$, we want to predict the user satisfaction rating after each dialogue-act in d_i, namely, $r'(da(d_i, 1)) \dots r'(da(d_i, m_i))$, using D with their dialogue-level ratings $r(d_1) \dots r(d_N)$.

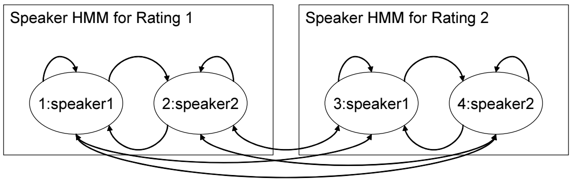


Figure 1: SHMMs connected ergodically. In the figure, an oval marked with speaker1/speaker2 indicates a state that emits speaker1/speaker2’s dialogue-acts. Arrows denote transitions and numbers before speaker1/speaker2 are state IDs. Boxes group together the states related to a particular overall rating.

Here, $da(d_i, l)$ denotes the l-th dialogue-act in d_i, N the total number of dialogues, and m_i the total number of dialogue-acts in d_i.

Our basic idea is to train HMMs representing dialogue-act sequences of dialogues for each overall rating and combine these rating-related HMMs into a single HMM that can assign ratings for dialogue-acts by estimating from which HMM each dialogue-act has most likely to have been generated by the Viterbi decoding. We use HMMs because they can deal with sequences that evolve over time and have been successfully utilized to model and evaluate dialogue-act sequences (Shirai, 1996; Isomura et al., 2009; Engelbrecht et al., 2009). The generative feature of an HMM is also useful when we want to build a probabilistic dialogue manager that produces the most likely dialogue-act sequences (Hori et al., 2008) or that aims to maximize a reward function in partially observable Markov decision processes (Williams and Young, 2007; Minami et al., 2009).

When there are K levels of user satisfaction as overall ratings, we create K HMMs each of which is trained using the dialogue-act sequences in dialogues $D_k \subset D$, where $D_k = \{\forall d_i, |r(d_i) = k\}$. We use the EM-algorithm to train HMMs. Here, we assume that each HMM has two states, each of which emits dialogue-acts of one of the conversational participants. This type of HMM is called a speaker HMM (**SHMM**) and has been successfully utilized to model two-party conversation (Meguro et al., 2009).

As an illustrative example, Fig. 1 shows two SHMMs for ratings 1 and 2 that are connected ergodically. We can simply use these connected SHMMs (namely, states 1, 2, 3, and 4) to decode a sequence of dialogue-acts into state sequences and thereby obtain rating predictions. For example, if the optimal state sequence obtained by the Viterbi decoding is {4, 2, 1, 3, 2}, we can convert it into ratings <2, 1, 1, 2, 1> using the ratings associated with the states.

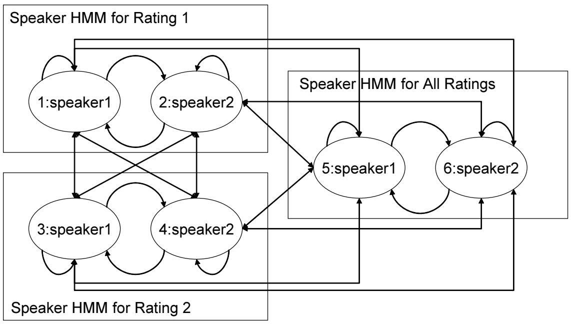


Figure 2: SHMMs with an additional SHMM trained from all dialogues.

Introducing Common States: The simple ergodic model may not be sufficient for appropriately assigning ratings to input dialogue-act sequences because it is often the case that there are dialogue-act sequences, such as greetings and question-answer pairs, that commonly occur in every dialogue. If we forcefully assign a rating for such dialogue-act sequences, it may result in degrading the prediction accuracy. Therefore, in addition to the simple ergodic model, we introduce another SHMM that represents dialogue-act sequences of dialogues for all ratings (see Fig. 2). This additional SHMM models dialogue-act sequences that occur commonly in all dialogues and it can simply be trained using all dialogues. Hence, we call the states in this SHMM **common states**. When this SHMM is added to the ergodic model, it may be possible to reduce the possibility of our having to forcefully assign inappropriate scores to common dialogue-act sequences. In this model, when the optimal state sequence is {1, 4, 5, 6, 2}, the predicted ratings become <1, 2, 0, 0, 1>. Here, we assume that the SHMM for all ratings corresponds to rating 0, which is reasonable because common dialogue-acts should not affect ratings. The obtained ratings can also be interpreted as <1, 2, 2, 2, 1> when we assume that the rating of a dialogue-act is taken over from the previous turn.

Using Concatenated Training: We have so far presented two model topologies, one with K SHMMs connected ergodically and the other with $K + 1$ SHMMs having an additional SHMM representing all ratings. However, we still have a problem; that is, we need to find optimal transition probabilities between the SHMMs of different ratings. Our solution is to use concatenated training (Lee, 1989). The procedure for concatenated training is illustrated in Fig. 3 and has the following three steps.

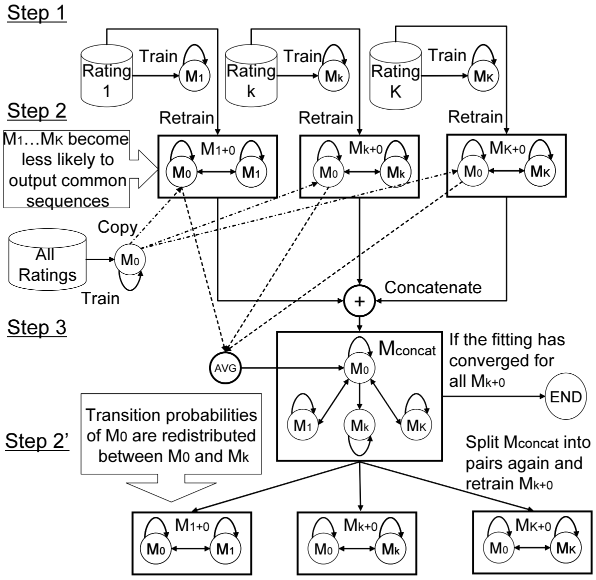

Step 1
Train
Rating 1
M1
Train
Rating k
Mk
Train
Rating K
MK
Step 2
Retrain
Retrain
Retrain
M1…MK become less likely to output common sequences
M1+0
M0
M1
Mk+0
M0
Mk
MK+0
M0
MK
Copy
All Ratings
M0
Train
Concatenate
Step 3
AVG
Mconcat
M0
M1
Mk
MK
If the fitting has converged for all Mk+0
END
Transition probabilities of M0 are redistributed between M0 and Mk
Step 2'
Split Mconcat into pairs again and retrain Mk+0
M1+0
M0
M1
Mk+0
M0
Mk
MK+0
M0
MK

Figure 3: Three steps to combine SHMMs using concatenated training.

step 1 Train an SHMM M_k ($M_k \in M, 1 \leq k \leq K$) using dialogues D_k, where $D_k = \{\forall d_i | r(d_i) = k\}$, and an SHMM M_0 using all dialogues; i.e., D. Here, K means the maximum level of user satisfaction and $r(d_i)$ the rating assigned to d_i.

step 2 Connect each $M_k \in M$ with a copy of M_0 using equal initial and transition probabilities (we call this connected model M_{k+0}) and retrain M_{k+0} with $\forall d_i \in D_k$, where $r(d_i) = k$.

step 3 Merge all models M_{k+0} $(1 \leq k \leq K)$ to produce one concatenated HMM (M_{concat}). Here, the output probabilities of the copies of M_0 are averaged over K when all models are merged to create a combined model. If the fitting of all M_{k+0} models has converged against the training data, exit this procedure; otherwise, go to step 2 by connecting a copy of M_0 and M_k for all k. Here, the transition probabilities from M_0 to $M_l (l \neq k)$ are summed and equally distributed between the copied M_0's self-loop and transitions to the states in M_k.

In concatenated training, the transition and output probabilities can be optimized between M_0 and M_k, meaning that the output probabilities of dialogue-act sequences that are common and also found in M_k can be moved from M_k to M_0. This makes the distribution of M_k sharp (not broad/uniform), making it likely to output only the dialogue-acts specific to a rating k. As regards M_0, its distribution of output probabilities can also be sharpened for dialogue-acts that occur commonly in all ratings. This sharpening of distributions is likely to be helpful in assigning appropriate ratings to dialogue-act sequences. In the next section, we experimentally examine how these proposed HMMs perform in modeling and predicting user satisfaction transitions in dialogue.

4 Experiment

To verify our approach, we first prepared dialogue data. Then, we trained our HMMs and compared them with a random baseline and an upper bound that uses a supervised approach; that is, an HMM is trained using reference labels on the dialogue-act level.

4.1 Dialogue Data

We used dialogues in two domains; the animal discussion (**AD**) domain and the attentive listening (**AL**) domain. All dialogues are in Japanese. In both domains, the data we used were text dialogues. We did not use spoken dialogue data because we wanted to avoid particular problems of voice, such as filled pauses and overlaps, although we aim to deal with spoken dialogue in the future.

4.1.1 Animal Discussion

We used the dialogue data in the AD domain that we previously collected (Higashinaka et al., 2008). In this domain, the system and user talk about likes and dislikes about animals via a text chat interface. The data consist of 1000 dialogues between a dialogue system and 50 human users. Each user conversed with the system 20 times, including two example dialogues at the beginning. All user/system utterances have been annotated with dialogue-acts. There are 29 dialogue-act types including those related to self-disclosure, question, response, and greetings. For example, a dialogue-act DISC-P denotes one's self-disclosure about a proposition P. Here, P is either `like(X,A)` or `dislike(X,A)` where X is a conversational participant and A a certain animal. DISC-R denotes one's self-disclosure of a reason for a proposition. See (Higashinaka et al., 2008) for the details of the dialogue-acts.

For our experiment, we created two subsets of the data. We first extracted 180 dialogues by taking all 18 non-example dialogues for the initial ten users sorted by user ID (**AD-SUB1**; 4147 user dialogue-acts and 6628 system dialogue-acts). Then, from AD-SUB1, we randomly extracted nine dialogues per user to form another subset of 90 dialogues (**AD-SUB2**; 2050 user dialogue-acts and 3290 system dialogue-acts). An annotator, who was not one of the authors, labeled AD-SUB1 with dialogue-level user satisfaction ratings and AD-SUB2 with utterance-level ratings. More specifically, each dialogue/utterance was given three different user satisfaction ratings related to "Smoothness of the conversation", "Closeness perceived by the user towards the system", and "Willingness to continue the conversation". The ratings ranged from 1 to 7, where 1 is the worst and 7 the best (see Fig. 4 for examples of utterance-level and overall ratings given by the annotator for an excerpt of a dialogue). In a manner similar to (Evanini et al., 2008), we used a third-person's user satisfaction rating for the sake of consistency.

	Utterance (dialogue-acts)	Sm	Cl	Wi
SYS	Do you like rabbits? (DA: Q-DISC-P)	6	6	6
USR	I like rabbits. They are cute. (DA: DISC-P, DISC-R)			
SYS	Indeed they are cute. (DA: REPEAT)	6	6	6
SYS	Tell me why you like rabbits. (DA: Q-DISC-R-OTHER)	6	5	6
USR	I like them because they are small and warm. (DA: DISC-P-R)			
SYS	You like them because they are warm. (DA: REPEAT)	7	5	7
	Overall rating for the dialogue	7	5	6

Figure 4: Excerpt of a dialogue with utterance-level user satisfaction ratings for smoothness (Sm), closeness (Cl), and willingness (Wi) in the AD domain. SYS and USR denote system and user, respectively. The dialogue was translated by the authors.

For utterance-level ratings, the annotator carefully read each utterance and gave ratings after each system utterance according to how she would have felt after receiving each system utterance if she had been the user in the dialogue. To make the situation more realistic, she was not allowed to look down at the dialogue after the current utterance. At the beginning of a dialogue, the ratings always started from four (neutral). When the annotator gave dialogue-level ratings, she looked through the entire dialogue and rated its quality (smoothness, closeness, and willingness) according to how she would have felt after having had the dialogue in question.

4.1.2 Attentive Listening

We collected human-human listening-oriented dialogues in a manner similar to (Meguro et al., 2009). In this AL domain, a listener attentively listens to the other in order to satisfy the speaker's desire to speak and to make himself/herself heard. We collected such listening-oriented dialogues using a website where users taking the roles of listeners and speakers were matched up to have conversations. There were ten listeners who always stayed at the website and 37 speakers who could talk to them anytime the listeners were available. They were all paid for their participation. A conversation was done through a text-chat interface.

The use of facial and other non-linguistic expressions were not allowed for analysis purposes. The participants were instructed to end the conversation approximately after ten minutes. Within a three-week period, each speaker was instructed to have at least two conversations a day, resulting in our collecting 1260 listening-oriented dialogues.

Two independent annotators labeled each utterance with 40 dialogue-act types, including those related to self-disclosure, question, internal argument, sympathy, and information giving. The inter-annotator agreement was reasonable, with 0.57 in Cohen's κ. Although we cannot describe the full details of our dialogue-acts for lack of space, we have dialogue-acts DISC-EVAL-POS for one's self-disclosure of his/her positive evaluation towards a certain entity, DISC-EXP for one's self-disclosure of his/her experience, and SELF-Q-DESIRE for one's internal argument about his/her desire (e.g., "Have I ever wanted to go abroad?"). We used the dialogue-act annotation of one of the annotators in this work.

An annotator gave dialogue-level user satisfaction ratings to all 1260 dialogues (**AL-ALL**; 31779 speaker dialogue-acts and 28681 listener dialogue-acts). Then, we made a subset of the data by randomly selecting ten dialogues for each of the ten listeners to obtain 100 dialogues (**AL-SUB1**; 2453 speaker dialogue-acts and 2197 listener dialogue-acts). Finally, the annotator gave utterance-level ratings to AL-SUB1. The utterance-level ratings were given only after listeners' utterances. The annotator gave three ratings as in the AD domain; namely, smoothness, closeness, and good listening. Instead of willingness, we have a "good listener" criterion asking for how good the annotator thinks the listener is from the viewpoint of attentive listening; for example, how well the listener is making it easy for the speaker to speak. All ratings ranged from 1 to 7. See Fig. 5 for a sample dialogue in the AL domain with utterance-level and overall ratings given by the annotator.

4.2 Training HMMs

From the dialogue data and their dialogue-level ratings, we created our proposed HMMs. We had five topology variations:

ergodic0: The simple ergodic model with no additional SHMM for all ratings. See Fig. 1 for the topology. This HMM has 7 SHMMs connected ergodically with equal initial/transition probabilities.

ergodic1: The simple ergodic model with an additional SHMM for all ratings. See Fig. 2 for the topology. This HMM has 8 (7 + 1) SHMMs connected ergodically with equal initial/transition probabilities.

	Utterance (dialogue-acts)	Sm	Cl	GL
LIS	You know, in spring, Japanese food tastes delicious. (DA: DISC-EVAL-POS)	5	5	5
SPK	This time every year, I make a plan to go on a healthy diet. But . . . (DA: DISC-HABIT)			
LIS	Uh-huh (DA: ACK)	6	5	6
SPK	The temperature goes up suddenly! (DA: INFO)			
SPK	It's always too late! (DA: DISC-EVAL-NEG)			
LIS	Clothing worn gets less and less while not being able to lose weight. (DA: DISC-FACT)	6	6	6
SPK	Well, people around me soon get used to my body shape though. (DA: DISC-FACT)			
	Overall rating for the dialogue	7	7	7

Figure 5: Excerpt of a dialogue with utterance-level user satisfaction ratings for smoothness (Sm), closeness (Cl), and good listener (GL) in the AL domain. SPK and LIS denote speaker and listener, respectively. Both the speaker and listener are human.

ergodic2: Same as ergodic1 except that the number of common states is doubled so that common dialogue-act sequences can be more accurately modeled. Note that without concatenated training, SHMMs for each rating may also have sharp distributions for common sequences. One possible solution to avoid this is to sharpen the distributions of common states by increasing its number of states.

concat1: 8 (7 + 1) SHMMs combined using concatenated training. See Fig. 3 for the topology.

concat2: Same as concat1 except that the number of common states is doubled.

[See Appendices A and B for the actual examples of the obtained models]

4.2.1 Baseline and Upper Bound

We created the following baseline (random) and upper bound (supervised) models for comparison:

random: This outputs ratings 1–7 at random.

supervised: This is an HMM trained in a manner similar to (Engelbrecht et al., 2009). This model is the same as ergodic0 in topology but different in that the initial, transition, and output probabilities are trained in a supervised manner using the dialogue-acts and dialogue-act-level reference ratings in AD-SUB2 and AL-SUB1. Since we only have ratings for system/listener utterances in the corpora, in order to make training data, we assumed that the ratings for dialogue-acts corresponding to user/speaker utterances were the same as those after the previous system/listener utterances. This model simulates the ideal situation where we possess user satisfaction ratings for all dialogue-acts in the data.

4.3 Evaluation Procedure

We performed a ten-fold cross validation. We first separated utterance-level labeled data (i.e., AD-SUB2 or AL-SUB1) into 10 disjoint sets. Then, for each set S, we used dialogue-level labeled data (i.e., AD-SUB1 or AL-ALL) excluding S for training HMMs. Here, 'supervised' only used the utterance-level labeled data excluding S for training. Then, we made the models (i.e., ergodic0, ergodic1, ergodic2, concat1, concat2, random and supervised) output rating sequences for the dialogue-acts in S and evaluated them with the reference ratings in S. We repeated this process ten times to evaluate the overall performance.

Since utterance-level ratings are provided only after system/listener utterances, we only evaluated ratings after dialogue-acts corresponding to system/listener utterances. When a system/listener utterance contained multiple dialogue-acts, the dialogue-acts were assumed to have the same rating as that utterance. When the output rating sequences contain 0, which can be the case for ergodic1–2 and concat1–2, the 0 is replaced by the most previous non-zero rating. When 0 is found at the beginning of a dialogue, it remained 0. Although our reference ratings always started with four (cf. Section 4.1.1), we did not use this information to fill initial zeros because we wanted to evaluate the prediction accuracy when we do not have any prior knowledge. Since some models may benefit from avoiding evaluating dialogue-acts at the beginning because of these zeros, we simply compared the rating sequences where all models produced non-zero values. For example, when we have three output rating sequences <0,5,6,0,4>, <0,0,1,2,0>, and <1,2,3,4,5> for a given dialogue-act sequence, the zeros that follow non-zero values are first filled with their preceeding values, and thereby we obtain <0,5,6,6,4>, <0,0,1,2,2>, and <1,2,3,4,5>. Then, by cropping the common non-zero span, we obtain <6,6,4>, <1,2,2>, and <3,4,5>, and use these rating sequences for evaluation.

4.3.1 Evaluation Criteria

We used two kinds of evaluation criteria: one for evaluating individual matches and the other for evaluating distributions.

Evaluating Individual Matches: We used the match rate and mean absolute error to evaluate the matching of reference and hypothesis rating sequences. They are derived by the equations shown below. In the equations, R $(= \{R_1 \dots R_L\})$ and H $(= \{H_1 \dots H_L\})$ denote reference and hypothesis rating sequences for a dialogue, respectively. L is the length of R and H (Note that they have the same length).

- **Match Rate (MR)**

$$\mathrm{MR}(R,H) = \frac{1}{L}\sum_{i=1}^{L}\mathrm{match}(R_i,H_i), \qquad (1)$$

where 'match' returns 1 or 0 depending on whether a rating in R matches that in H.

- **Mean Absolute Error (MAE)**

$$\mathrm{MAE}(R,H) = \frac{1}{L}\sum_{i=1}^{L}|R_i - H_i|. \qquad (2)$$

Evaluating Distributions: In generative models, it is important that the output distribution matches that of the reference. Therefore, we additionally use Kullback-Leibler divergence, match rate per rating, and mean absolute error per rating. The Kullback-Leibler divergence evaluates the shape of output distributions. The match rate per rating and mean absolute error per rating evaluate how accurately each individual rating can be predicted; namely, the accuracy for predicting dialogue-acts with one rating is equally valued with those for other ratings irrespective of the distribution of ratings in the reference. It is important to use these metrics in the practical as well as information theoretic sense because it is no use predicting only easy-to-guess ratings; we should be able to correctly predict rare but still important cases. For example, rating 1 in human-human dialogue is quite rare; however, predicting it is very important for detecting problematic situations in a dialogue.

- **Kullback-Leibler Divergence (KL)**

$$\mathrm{KL}(\mathbf{R},\mathbf{H}) = \sum_{r=1}^{K}\mathrm{P}(\mathbf{H},r)\cdot\log(\frac{\mathrm{P}(\mathbf{H},r)}{\mathrm{P}(\mathbf{R},r)}), \qquad (3)$$

where K is the maximum user satisfaction rating (i.e. 7 in this experiment), $\mathbf{R}$ and $\mathbf{H}$ denote the sequentially concatenated reference/hypothesis rating sequences of the entire dialogues, and $\mathrm{P}(*,r)$ denotes the occurrence probability that a rating r is found in an arbitrary rating sequence.

- **Match Rate per rating (MR/r)**

$$\mathrm{MR/r}(\mathbf{R},\mathbf{H}) = \frac{1}{K}\sum_{\mathrm{r}=1}^{K}\frac{\sum_{i\in\{i|\mathbf{R}_i=r\}}\mathrm{match}(\mathbf{R}_i,\mathbf{H}_i)}{\sum_{i\in\{i|\mathbf{R}_i=r\}}1}, \qquad (4)$$

	Criterion	random	ergodic0	ergodic1	ergodic2	concat1	concat2	supervised
Smoothness	MR	0.142_{e0e1}	0.111	0.111	0.157_{e0e1}	0.153	$\mathbf{0.199}_{e0e1r}$	$0.275_{c1e0e1e2r}$
	MAE	1.988_{e0e1}	2.212	2.212	1.980	1.936_{e0e1}	$\mathbf{1.870}_{e0e1}$	$1.420_{c1c2e0e1e2r}$
	KL	0.287	0.699	0.699	0.562	**0.280**	0.369	0.162
	MR/r	0.143	0.137	0.137	0.176	0.136	**0.177**	0.217
	MAE/r	2.286	2.414	2.414	**2.152**	2.301	2.206	1.782
Closeness	MR	0.143	0.129	0.129	0.171_{e0e1}	0.174	$\mathbf{0.189}_{e0e1}$	$0.279_{c1c2e0e1e2r}$
	MAE	2.028	2.066	2.066	1.964	$\mathbf{1.798}_{e0e1r}$	1.886	$1.431_{c1c2e0e1e2r}$
	KL	0.195	0.449	0.449	0.261	**0.138**	0.263	0.092
	MR/r	0.143	0.156	0.156	**0.170**	0.155	0.164	0.231
	MAE/r	2.283	2.236	2.236	2.221	2.079	**2.067**	1.702
Willingness	MR	0.143_{e0e1}	0.112	0.112	0.180_{e0e1}	0.152	$\mathbf{0.183}_{e0e1}$	$0.283_{c1c2e0e1e2r}$
	MAE	2.005	2.133	2.133	1.962	$\mathbf{1.801}_{e0e1r}$	1.882	$1.403_{c1c2e0e1e2r}$
	KL	**0.225**	0.568	0.568	0.507	0.238	0.255	0.125
	MR/r	0.143	0.152	0.152	**0.192**	0.181	0.167	0.224
	MAE/r	2.286	2.258	2.258	2.107	**1.958**	2.164	1.705

Table 1: The match rate (MR), mean absolute error (MAE), Kullback-Leibler divergence (KL), match rate per rating (MR/r) and mean absolute error per rating (MAE/r) for our proposed HMMs, the random baseline, and the upper bound (supervised) for the AD domain. 'e0–e2', 'c1–c2', and 'r' indicate the statistical significance ($p<0.01$) over ergodic0–2, concat1–2, and random, respectively. **Bold font** indicates the best value within each row (except for 'supervised').

where $\mathbf{R}_i$ and $\mathbf{H}_i$ denote ratings at i-th positions.

- **Mean Absolute Error per rating (MAE/r)**

$$\mathrm{MAE/r}(\mathbf{R},\mathbf{H}) = \frac{1}{K}\sum_{\mathrm{r}=1}^{K}\frac{\sum_{i\in\{i|\mathbf{R}_i=r\}}|\mathbf{R}_i-\mathbf{H}_i|}{\sum_{i\in\{i|\mathbf{R}_i=r\}}1}. \tag{5}$$

4.4 Evaluation Results

Tables 1 and 2 show the evaluation results for the AD and AL domains, respectively. The MR and MAE values are averaged over all dialogues. To compare the means of the MR and MAE, we performed a non-parametric multiple comparison test [Steel-Dwass test (Dwass, 1960)]. We did not perform a statistical test for other criteria because it was difficult to perform sample-wise comparison for distributions. Naturally, 'supervised' is the best performing model for all criteria in both domains. Therefore, we focus on how much our proposed models differ from the baseline (random) and the upper bound (supervised).

In the AD domain, we find that ergodic0 and ergodic1 performed rather poorly and concat1 and concat2 performed fairly well, significantly outperforming the random baseline. However, it is also clear that we still need a great deal of improvement for our models to reach the level of 'supervised'. A promising sign is that concat2 is not significantly different from 'supervised' in smoothness. Here, ergodic0 and ergodic1 returned the exact same results. This means that the state transition paths did not go through the common states at all in ergodic1, suggesting that the common states in ergodic1 have very broad output distributions and the optimal path could not go through the common states, instead preferring other states having sharper distributions. However, this phenomenon was rightly avoided by introducing more common states as seen in the results for ergodic2; nonetheless, as the results for concat1 and concat2 indicate, the transition probabilities have to be trained appropriately to obtain better results.

In the AL domain, although the tendency of the evaluation results is the same as that for the AD domain, concat2 is clearly the best performing model. It outperformed other models in almost all cases except for "Good Listener" for which concat1 performed better. In fact, the MR/r and MAE/r of concat1 are quite close to those of 'supervised', suggesting the potential of our approach.

Overall, although we still need further improvement in order for our models to be closer to the upper bound, we showed that we can, to some extent, predict user satisfaction transitions in a dialogue only from overall ratings of dialogues using our proposed HMMs. We also showed that model topologies and learning methods can make significant differences. Especially, we found the introduction of common states to be crucial in making appropriate models for prediction. Since our models, especially concat2, significantly outperformed the baseline, we believe that our approach can be one of the viable options for automatically predicting user satisfaction transitions when there exist only overall rating data.

5 Summary and Future Work

We presented a novel approach for modeling user satisfaction transitions only from dialogues with overall ratings. The experimental results show that it is possible to predict user satisfaction transi-

	Criterion	random	ergodic0	ergodic1	ergodic2	concat1	concat2	supervised
Smoothness	MR	0.143_{e0e1e2}	0.069	0.069	0.131_{e0e1}	0.173_{e0e1}	$\mathbf{0.243}_{c1e0e1e2r}$	$0.439_{c1c2e0e1e2r}$
	MAE	1.868_{e0e1e2}	2.519	2.519	2.433	$1.687_{e0e1e2r}$	$\mathbf{1.594}_{e0e1e2r}$	$0.802_{c1c2e0e1e2r}$
	KL	0.989	2.253	2.253	2.319	0.851	**0.753**	0.087
	MR/r	0.141	0.118	0.118	0.156	0.161	**0.167**	0.231
	MAE/r	2.289	2.500	2.500	2.492	2.093	**2.077**	1.868
Closeness	MR	0.143_{e0e1}	0.050	0.050	0.175_{e0e1}	0.158_{e0e1}	$\mathbf{0.263}_{c1e0e1e2r}$	$0.425_{c1c2e0e1e2r}$
	MAE	1.849_{e0e1e2}	2.357	2.357	2.316	1.778_{e0e1e2}	$\mathbf{1.562}_{e0e1e2r}$	$0.890_{c1c2e0e1e2r}$
	KL	1.022	2.137	2.137	2.220	1.155	**0.909**	0.109
	MR/r	0.143	0.090	0.090	0.122	0.117	**0.159**	0.237
	MAE/r	2.281	2.577	2.577	2.811	2.260	**2.039**	1.972
Good Listener	MR	0.143_{e0e1}	0.075	0.075	0.145_{e0e1}	0.199_{e0e1}	$\mathbf{0.206}_{e0e1e2}$	$0.422_{c1c2e0e1e2r}$
	MAE	1.890_{e0e1e2}	2.237	2.237	2.150	$\mathbf{1.634}_{e0e1e2r}$	$\mathbf{1.634}_{e0e1e2r}$	$0.852_{c1c2e0e1e2r}$
	KL	0.945	1.738	1.738	1.782	0.924	**0.824**	0.087
	MR/r	0.143	0.121	0.121	0.184	**0.224**	0.200	0.227
	MAE/r	2.284	2.358	2.358	2.236	**1.911**	2.083	1.769

Table 2: Evaluation results for the AL domain. See Table 1 for the notations in the table.

tions to some extent by our approach and that introducing common states and concatenated training can significantly improve prediction accuracy. For improvement, we plan to explore new dialogic features for emissions, different topologies, and other optimization functions, such as discriminative ones. We also need to validate our approach using dialogue-act recognition results instead of hand-labeled dialogue-acts. We also want to apply our approach to sequence mining in dialogues where we have categories instead of ratings for dialogues. It is also necessary to test whether our HMMs can be generalized over different raters, since user satisfaction ratings may differ greatly among individuals. Although there remain such issues, we believe we have presented a new direction in automatic evaluation of dialogues and the experimental results show that our approach is promising.

References

Meyer Dwass. 1960. Some k-sample rank-order tests. In Ingram Olkin et al., editor, *Contributions to Probability and Statistics*, pages 198–202. Stanford University Press.

Klaus-Peter Engelbrecht, Florian Gödde, Felix Hartard, Hamed Ketabdar, and Sebastian Möller. 2009. Modeling user satisfaction with hidden Markov models. In *Proc. SIGDIAL*, pages 170–177.

Keelan Evanini, Phillip Hunter, Jackson Liscombe, David Suendermann, Krishna Dayanidhi, and Roberto Pieraccini. 2008. Caller experience: A method for evaluating dialog systems and its automatic prediction. In *Proc. SLT*, pages 129–132.

Allen L. Gorin, Giuseppe Riccardi, and Jerry H. Wright. 1997. How may I help you? *Speech Communication*, 23(1-2):113–127.

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 *Proc. LREC*, pages 78–83.

Ota Herm, Alexander Schmitt, and Jackson Liscombe. 2008. When calls go wrong: How to detect problematic calls based on log-files and emotions? In *Proc. INTERSPEECH*, pages 463–466.

Ryuichiro Higashinaka, Kohji Dohsaka, and Hideki Isozaki. 2008. Effects of self-disclosure and empathy in human-computer dialogue. In *Proc. SLT*, pages 109–112.

Julia Hirschberg, Diane Litman, and Marc Swerts. 2004. Prosodic and other cues to speech recognition failures. *Speech Communication*, 43:155–175.

Chiori Hori, Kiyonori Ohtake, Teruhisa Misu, Hideki Kashioka, and Satoshi Nakamura. 2008. Dialog management using weighted finite-state transducers. In *Proc. INTERSPEECH*, pages 211–214.

Naoki Isomura, Fujio Toriumi, and Kenichiro Ishii. 2009. Evaluation method of non-task-oriented dialogue system using HMM. *IEICE Transactions on Information and Systems*, J92-D(4):542–551.

Woosung Kim. 2007. Online call quality monitoring for automating agent-based call centers. In *Proc. INTERSPEECH*, pages 130–133.

Kai-Fu Lee. 1989. *Automatic speech recognition: the development of the SPHINX system*. Kluwer Academic Publishers.

Toyomi Meguro, Ryuichiro Higashinaka, Kohji Dohsaka, Yasuhiro Minami, and Hideki Isozaki. 2009. Analysis of listening-oriented dialogue for building listening agents. In *Proc. SIGDIAL*, pages 124–127.

Yasuhiro Minami, Akira Mori, Toyomi Meguro, Ryuichiro Higashinaka, Kohji Dohsaka, and Eisaku Maeda. 2009. Dialogue control algorithm for ambient intelligence based on partially observable Markov decision processes. In *Proc. IWSDS*, pages 254–263.

Sebastian Möller, Klaus-Peter Engelbrecht, and Robert Schleicher. 2008. Predicting the quality and usability of spoken dialogue services. *Speech Communication*, 50(8-9):730–744.

Lawrence R. Rabiner. 1990. A tutorial on hidden Markov models and selected applications in speech recognition. *Readings in speech recognition*, 53(3):267–296.

Katsuhiko Shirai. 1996. Modeling of spoken dialogue with and without visual information. In *Proc. ICSLP*, volume 1, pages 188–191.

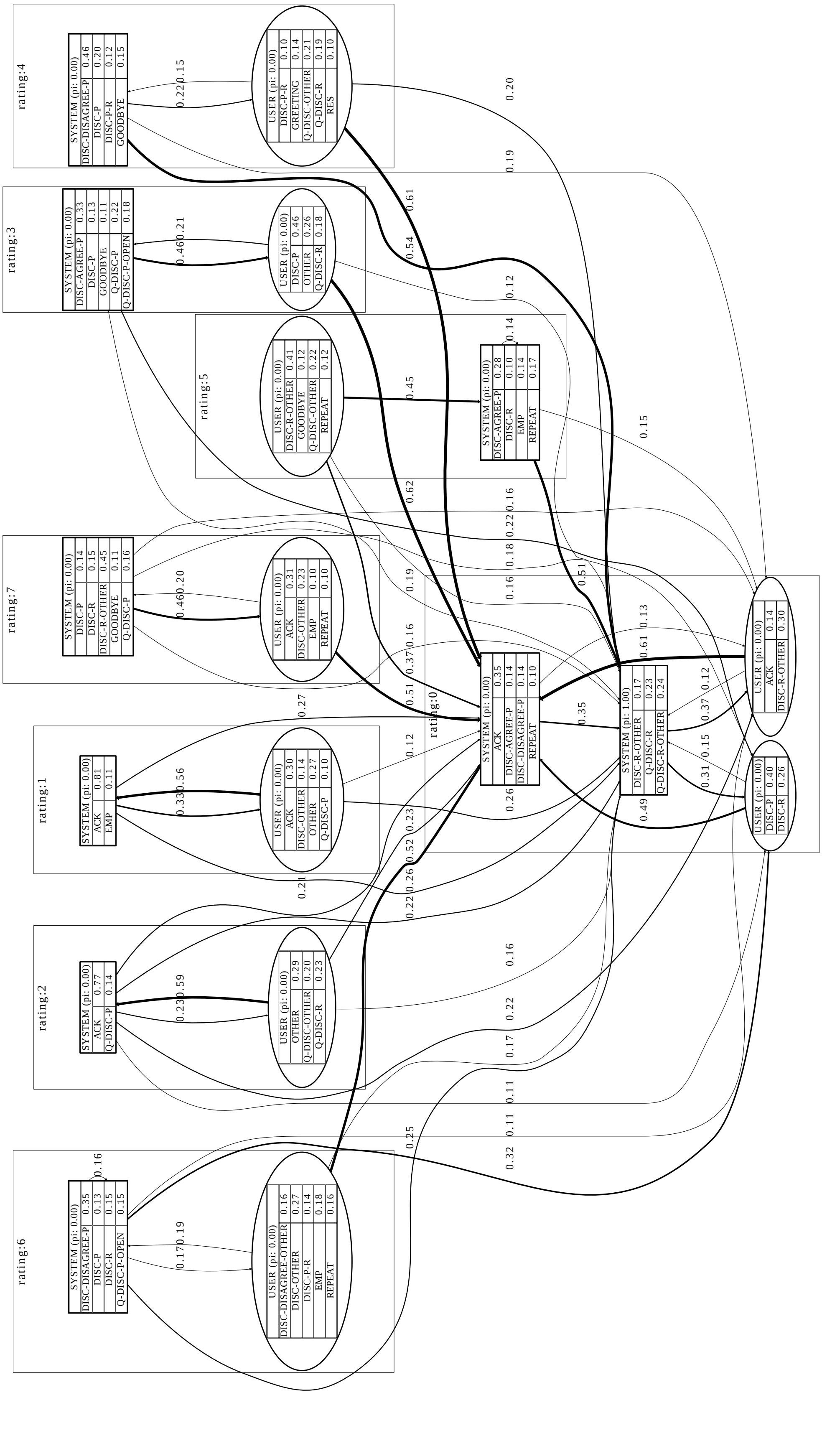
Marilyn A. Walker, Diane Litman, Candace A. Kamm, and Alicia Abella. 1997. PARADISE: A framework for evaluating spoken dialogue agents. In *Proc. EACL*, pages 271–280.

Marilyn A. Walker, Irene Langkilde-Geary, Helen Wright Hastie, Jerry Wright, and Allen Gorin. 2002. Automatically training a problematic dialogue predictor for a spoken dialogue system. *Journal of Artificial Intelligence Research*, 16(1):293–319.

Jason D. Williams and Steve Young. 2007. Partially observable Markov decision processes for spoken dialog systems. *Computer Speech & Language*, 21(2):393–422.

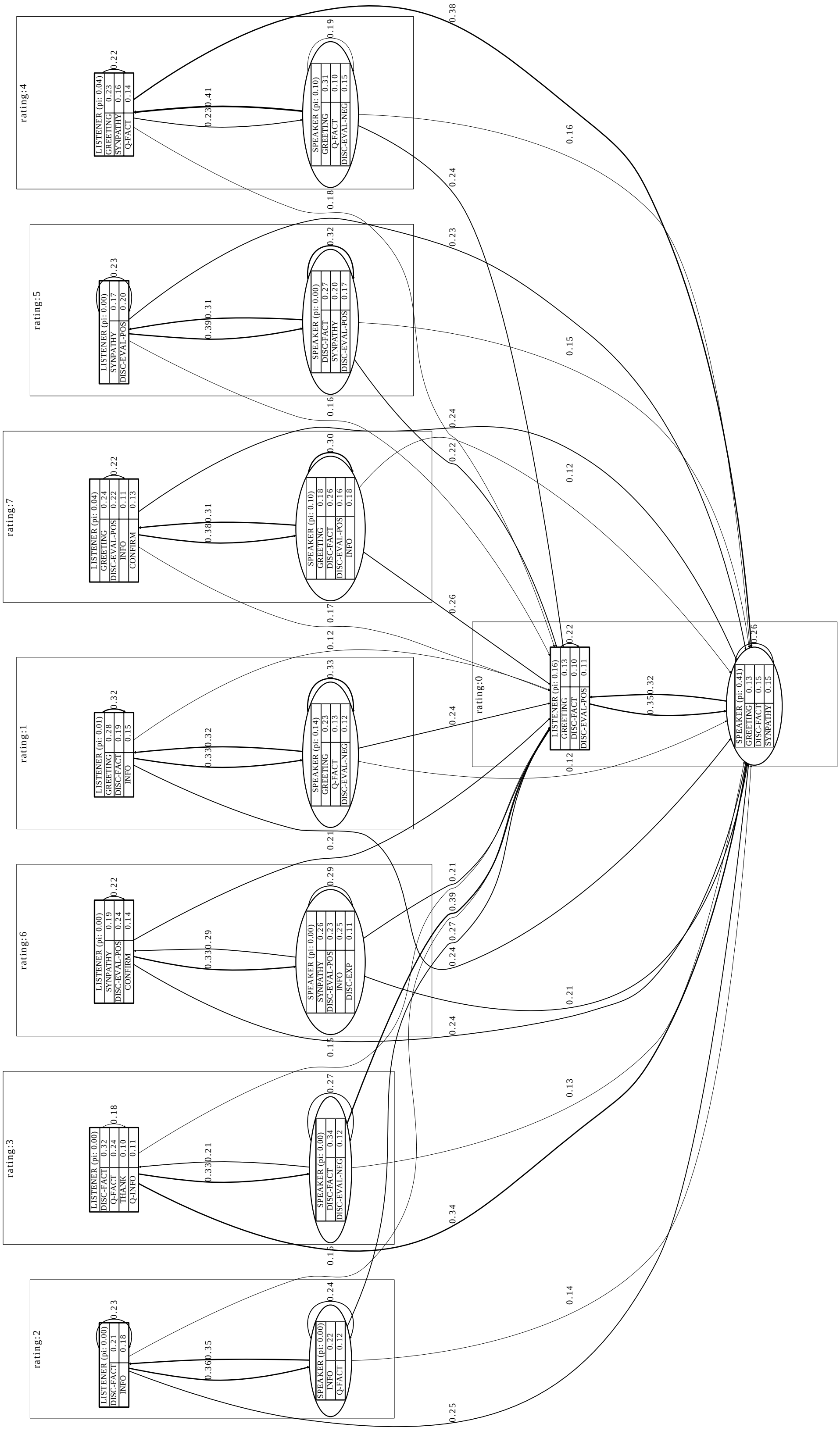
Appendix A. HMM obtained by concat2 for Willingness rating in the AD domain.

This HMM is the model obtained for one of the folds in the experiment. Square and oval states emit a system's dialogue-act and a user's dialogue-act, respectively. Emissions (dialogue-acts) are shown in each state as a table with their probabilities. Only the emissions and transitions over the probability of 0.1 are displayed for the sake of brevity. Here, 'pi' denotes initial probability.


rating:4
SYSTEM (pi: 0.00) | DISC-DISAGREE-P 0.46 | DISC-P 0.20 | DISC-P-R 0.12 | GOODBYE 0.15
USER (pi: 0.00) | DISC-P-R 0.10 | GREETING 0.14 | Q-DISC-OTHER 0.21 | Q-DISC-R 0.19 | RES 0.10
rating:3
SYSTEM (pi: 0.00) | DISC-AGREE-P 0.33 | DISC-P 0.13 | GOODBYE 0.11 | Q-DISC-P 0.22 | Q-DISC-P-OPEN 0.18
USER (pi: 0.00) | DISC-P 0.46 | OTHER 0.26 | Q-DISC-R 0.18
rating:5
USER (pi: 0.00) | DISC-R-OTHER 0.41 | GOODBYE 0.12 | Q-DISC-OTHER 0.22 | REPEAT 0.12
SYSTEM (pi: 0.00) | DISC-AGREE-P 0.28 | DISC-R 0.10 | EMP 0.14 | REPEAT 0.17
rating:7
SYSTEM (pi: 0.00) | DISC-P 0.14 | DISC-R 0.15 | DISC-R-OTHER 0.45 | GOODBYE 0.11 | Q-DISC-P 0.16
USER (pi: 0.00) | ACK 0.31 | DISC-OTHER 0.23 | EMP 0.10 | REPEAT 0.10
rating:1
SYSTEM (pi: 0.00) | ACK 0.81 | EMP 0.11
USER (pi: 0.00) | ACK 0.30 | DISC-OTHER 0.14 | OTHER 0.27 | Q-DISC-P 0.10
rating:2
SYSTEM (pi: 0.00) | ACK 0.77 | Q-DISC-P 0.14
USER (pi: 0.00) | OTHER 0.29 | Q-DISC-OTHER 0.20 | Q-DISC-R 0.23
rating:6
SYSTEM (pi: 0.00) | DISC-DISAGREE-P 0.35 | DISC-P 0.13 | DISC-R 0.15 | Q-DISC-P-OPEN 0.15
USER (pi: 0.00) | DISC-DISAGREE-OTHER 0.16 | DISC-OTHER 0.27 | DISC-P-R 0.14 | EMP 0.18 | REPEAT 0.16
rating:0
SYSTEM (pi: 0.00) | ACK 0.35 | DISC-AGREE-P 0.14 | DISC-DISAGREE-P 0.14 | REPEAT 0.10
SYSTEM (pi: 1.00) | DISC-R-OTHER 0.17 | Q-DISC-R 0.23 | Q-DISC-R-OTHER 0.24
USER (pi: 0.00) | ACK 0.14 | DISC-R-OTHER 0.30
USER (pi: 0.00) | DISC-P 0.40 | DISC-R 0.26
0.22 0.15
0.20
0.19
0.61
0.54
0.46 0.21
0.12
0.14
0.45
0.15
0.62
0.16
0.22
0.18
0.51
0.16
0.19
0.46 0.20
0.13
0.16
0.37
0.61
0.27
0.51
0.35
0.12
0.37
0.15
0.33 0.56
0.12
0.31
0.26
0.49
0.23
0.52
0.26
0.21
0.22
0.16
0.23 0.59
0.22
0.17
0.11
0.11
0.25
0.32
0.16
0.17 0.19

Appendix B. HMM obtained by concat1 for Good Listener rating in the AL domain.

This HMM is the model obtained for one of the folds in the experiment. Square and oval states emit a listener's dialogue-act and a speaker's dialogue-act, respectively. We find DICS-EVAL-NEG (self-disclosure of one's evaluation with a negative polarity) in the rating score 1 and DICS-EVAL-POS in the rating score 7, indicating that it may be better to make speakers talk about positive evaluations to be a good listener.


rating:4
LISTENER (pi: 0.04) | GREETING 0.23 | SYNPATHY 0.16 | Q-FACT 0.14
SPEAKER (pi: 0.10) | GREETING 0.31 | Q-FACT 0.10 | DISC-EVAL-NEG 0.15
0.22
0.19
0.230.41
0.18
0.38
0.16
0.24
rating:5
LISTENER (pi: 0.00) | SYNPATHY 0.17 | DISC-EVAL-POS 0.20
SPEAKER (pi: 0.00) | DISC-FACT 0.27 | SYNPATHY 0.20 | DISC-EVAL-POS 0.17
0.23
0.32
0.390.31
0.16
0.23
0.15
rating:7
LISTENER (pi: 0.04) | GREETING 0.24 | DISC-EVAL-POS 0.22 | INFO 0.11 | CONFIRM 0.13
SPEAKER (pi: 0.10) | GREETING 0.18 | DISC-FACT 0.26 | DISC-EVAL-POS 0.16 | INFO 0.18
0.22
0.30
0.380.31
0.17
0.24
0.22
0.12
0.26
rating:1
LISTENER (pi: 0.01) | GREETING 0.28 | DISC-FACT 0.19 | INFO 0.15
SPEAKER (pi: 0.14) | GREETING 0.23 | Q-FACT 0.13 | DISC-EVAL-NEG 0.12
0.32
0.33
0.330.32
0.12
0.21
0.24
rating:0
LISTENER (pi: 0.16) | GREETING 0.13 | DISC-FACT 0.10 | DISC-EVAL-POS 0.11
SPEAKER (pi: 0.41) | GREETING 0.13 | DISC-FACT 0.15 | SYNPATHY 0.15
0.22
0.26
0.350.32
0.12
rating:6
LISTENER (pi: 0.00) | SYNPATHY 0.19 | DISC-EVAL-POS 0.24 | CONFIRM 0.14
SPEAKER (pi: 0.00) | SYNPATHY 0.26 | DISC-EVAL-POS 0.23 | INFO 0.25 | DISC-EXP 0.11
0.22
0.29
0.330.29
0.15
0.21
0.39
0.27
0.24
0.21
0.24
rating:3
LISTENER (pi: 0.00) | DISC-FACT 0.32 | Q-FACT 0.24 | THANK 0.10 | Q-INFO 0.11
SPEAKER (pi: 0.00) | DISC-FACT 0.34 | DISC-EVAL-NEG 0.12
0.18
0.27
0.330.21
0.15
0.13
0.34
rating:2
LISTENER (pi: 0.00) | DISC-FACT 0.21 | INFO 0.18
SPEAKER (pi: 0.00) | INFO 0.22 | Q-FACT 0.12
0.23
0.24
0.360.35
0.14
0.25