IWSDS 2010, 2nd Oct 2010

Issues in Predicting User Satisfaction Transitions in Dialogues: Individual Differences, Evaluation Criteria, and Prediction Models

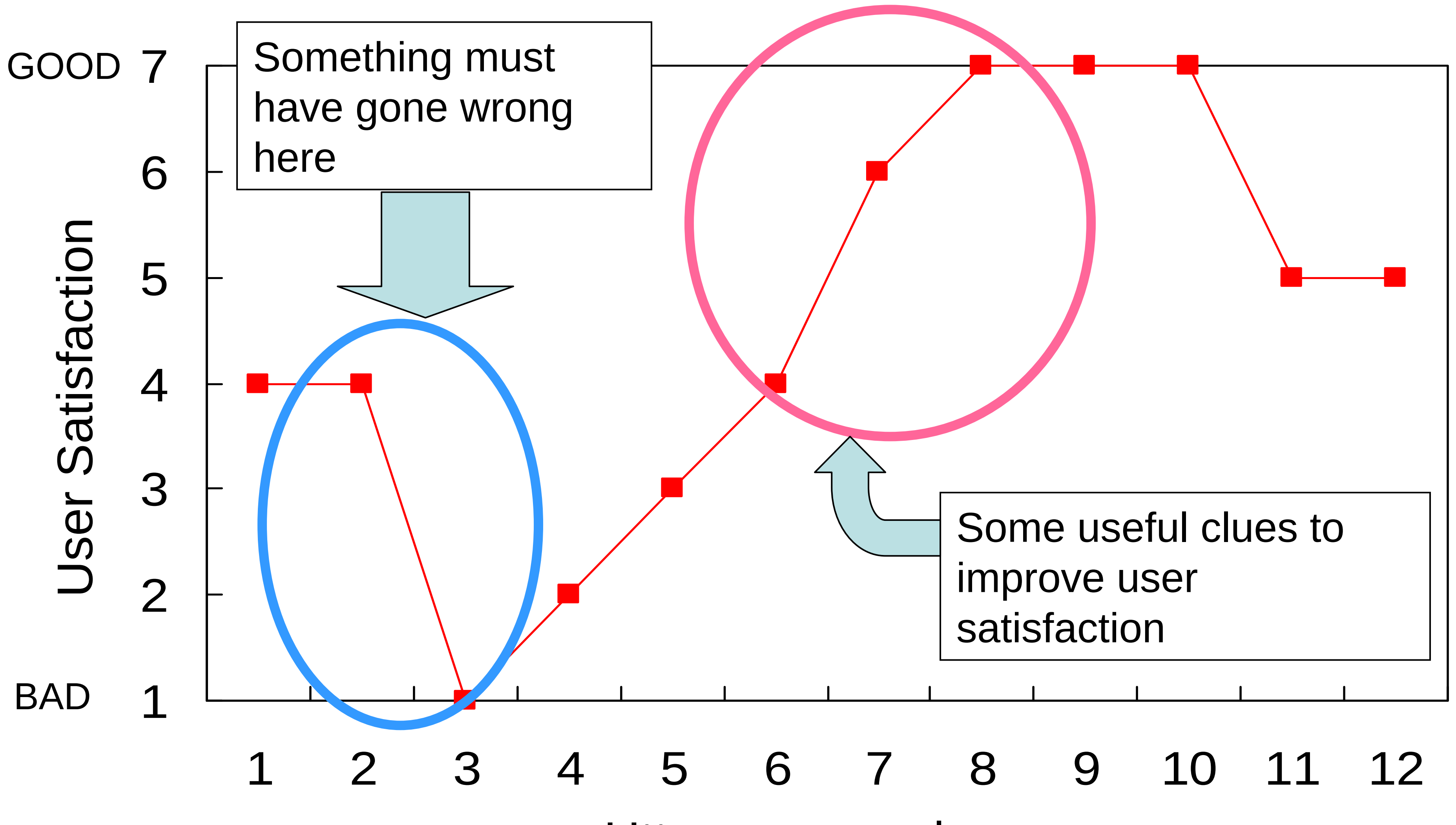
Ryuichiro Higashinaka, Yasuhiro Minami,
Kohji Dohsaka, Toyomi Meguro

NTT Corporation

Background

- Emerging work on predicting user satisfaction transitions during a dialogue
  - Useful for a turn-by-turn analysis of the performance of a dialogue system
  - Useful for pinpointing situations where the dialogue quality begins to degrade or improve
- Recent work
  - Modeling transitions by HMMs (Engelbrecht et al., 2009, Higashinaka et al., 2010)

User Satisfaction Transitions



Open Issues

- Individual differences
  - How user satisfaction transitions differ among raters?
- Evaluation criteria
  - What evaluation criteria to use for evaluating user satisfaction transitions?
- Prediction models
  - What model should we adopt for prediction?

(1) Individual Differences

- Subjective nature of user satisfaction
- Prediction model made from one rater's transitions may not generalize
- Need to investigate how raters agree in rating user satisfaction transitions

•We check correlations and distributions of ratings between different raters
•We discuss the feasibility of creating a general prediction model

(2) Evaluation Criteria

- In any engineering work, it is necessary to establish an evaluation measure
- No established measure
- Mean squared error of rating probabilities
  - Used in Engelbrecht et al. 2009
  - Limitation: dialogue has to follow a predefined scenario

    ➔ too restrictive for common use

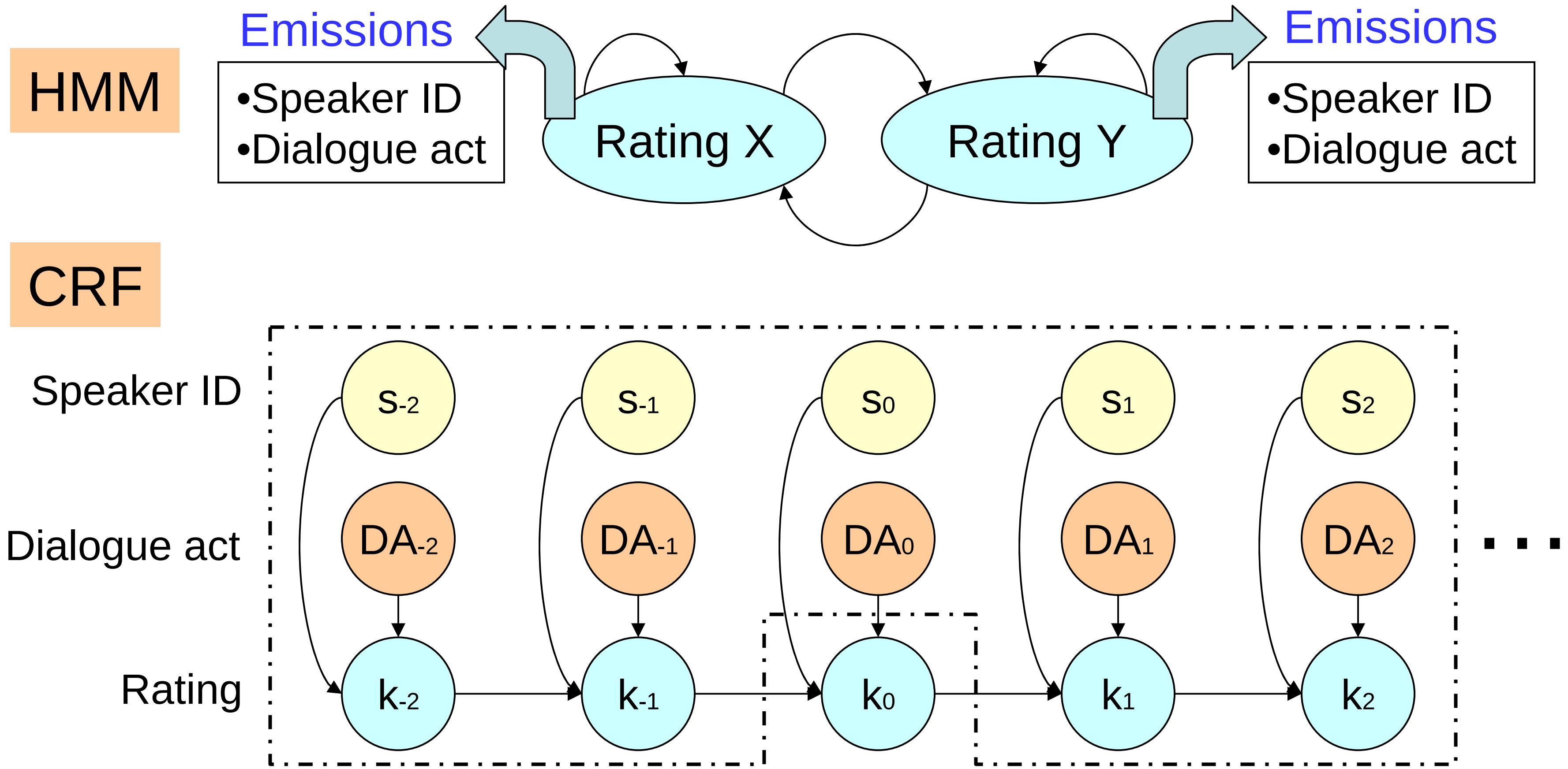
•We propose several candidates for evaluation metrics and experimentally decide the best one

(3) Prediction Models

- Hidden Markov models (HMMs)
  - Used in previous work
  - Generative model
- Conditional random fields (CRFs)
  - Recent trend in sequential labeling
  - Best performance in many NLP tasks
  - Discriminative model

•We compare HMMs and CRFs to investigate which model is more suitable for the task of predicting user satisfaction transitions

HMMs and CRFs



Prediction is done by finding the most likely rating sequence for the sequence of speaker IDs and dialogue acts

Data Collection

- Dialogue data (text chat) in two domains
  - Animal Discussion (AD)
    - Discuss likes and dislikes about animals
    - Human-system dialogue
    - Useful for obtaining preferences of users
  - Attentive Listening (AL)
    - Listener attentively listens to the speaker to satisfy the speaker's desire to be heard
    - Human-human dialogue
    - Useful for counseling purposes

Data Statistics

AD Domain: 90 dialogues				
	# Utterances	# Dialogue-acts	Avg	SD
All	5180	5340	59.33	17.54
User	1890	2050	22.78	6.60
System	3290	3290	36.56	11.81

AL Domain: 100 dialogues				
	# Utterances	# Dialogue-acts	Avg	SD
All	3951	4650	46.50	8.99
Speaker	2103	2453	24.53	5.69
Listener	1848	2197	21.97	5.25

Data Annotation

- User satisfaction ratings by two raters
  - Raters rated each system (listener) utterance as if they were the user (speaker)
  - 7-levels (1: bad ⇔ 7: good)
  - Third-party ratings for consistency
  - User satisfaction ratings from three aspects
    - Smoothness of a dialogue
    - Closeness perceived by the user
    - Willingness to talk or Good Listener
- Dialogue acts for all utterances

Example: Animal Discussion

	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

29 dialogue act types

Example: Attentive Listening

	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 when 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)			

Listener self-discloses a lot to propel the speaker to speak

40 dialogue act types

Individual Differences

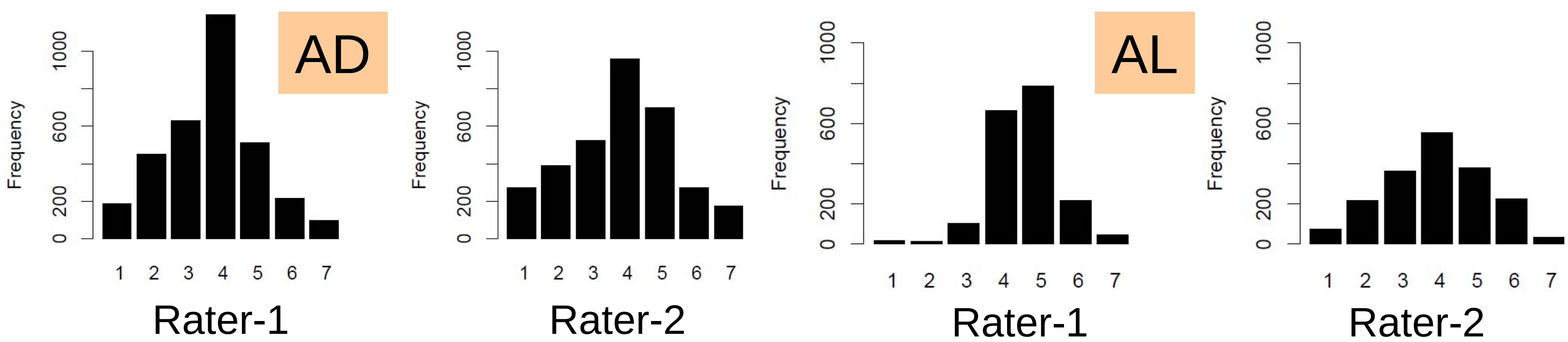
- Correlations between the two raters

Granularity	AD Domain			AL Domain		
	Smoothness	Closeness	Willingness	Smoothness	Closeness	Good Listener
(a) 7 ratings	0.18	0.15	0.27	0.18	0.10	0.11
(b) 3 ratings	0.17	0.13	0.18	0.04	0.05	0.11
(c) 3 ratings	0.13	0.11	0.21	0.14	0.08	0.08
(d) 2 ratings	0.20	0.17	0.31	0.18	0.13	0.14
(e) 2 ratings	0.30	0.30	0.32	0.18	0.11	0.04

When 7 ratings are converted into 2 ratings

Spearman's rank correlation coefficients

- Distributions of the ratings



Individual Differences (cont’d)

- Very low correlation between raters
  - Even decisions about good/bad do not match
- Distributions may vary greatly
  - Especially for human-human dialogues

Currently, it would be difficult to create a general prediction model

➔ We aim to create a rater-dependent prediction model in this work

Evaluation Criteria

- Six possible metrics to calculate the similarity between reference transitions and hypothesis transitions

Ref:	4 4 3 2 2 1 2 3
Hyp:	4 5 6 5 2 1 4 5

1. Match Rate (MR)
2. Mean Absolute Error (MAE)
3. Spearman's rank correlation coefficient (rho)
4. Kullback-Leibler Divergence (KL)
5. Match Rate per Rating (MR/r)
6. Mean Absolute Error per Rating (MAE/r)

•Equally treat difficult and easy-to guess ratings

- R: reference transitions for a dialogue
- H: hypothesis transitions
- L: length of a dialogue (# utterances)

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

How exactly two ratings match

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

Distance between the two transitions

$$\rho(R, H) = \frac{\sum_{i=1}^{L} (R_i - \bar{R})(H_i - \bar{H})}{\sqrt{\sum_{i=1}^{L} (R_i - \bar{R})^2 \sum_{i=1}^{L} (H_i - \bar{H})^2}}$$

Similarity of rating orders

- $\mathbf{R}$: reference transitions for all dialogues
- $\mathbf{H}$: hypothesis transitions
- K: maximum user satisfaction level (=7)

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

Similarity of rating distributions

$$\mathrm{MR/r}(\mathbf{R},\mathbf{H}) = \boxed{\frac{1}{K}\sum_{\mathrm{r}=1}^{K}} \overbrace{\boxed{\frac{\sum_{i\in\{i|\mathbf{R}_i=r\}} \mathrm{match}(\mathbf{R}_i,\mathbf{H}_i)}{\sum_{i\in\{i|\mathbf{R}_i=r\}} 1}}}^{\text{Match Rate}}$$

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

MR and MAE per each rating

- Equally treats difficult and easy-to guess ratings
- Important to predict rare but important cases

Assumptions for choosing the best metric

- The suitable metric
  - should show the lowest performance for “random choice” and “no choice” (e.g., majority baseline)
    ← they do not perform any prediction
  - should show similar performance values for the data of different raters
    ← the difficulty of prediction should be independent of the raters

Experiment

- Trained HMMs and CRFs using the reference user satisfaction transitions of each rater for each domain
- Random and majority baselines
- Procedure
  - Choose the best metric according to our assumptions
  - Analyze the performance of HMMs and CRFs using the best metric

The best metric

- Random and majority baselines beat HMMs and CRFs in MR, MAE, and MAE/r
- Spearman's rank correlation (rho) and KL greatly differ depending on the rater
- MR/r beats random and majority baselines and have similar values for different raters

MR/r becomes our recommended evaluation metric

Results (MR/r)

AD domain

	Smoothness		Closeness		Willingness	
	HMM	CRF	HMM	CRF	HMM	CRF
Rater-1	0.217	0.172	0.231	0.162	0.224	0.208
Rater-2	0.210	0.177	0.232	0.176	0.234	0.238

AL domain

	Smoothness		Closeness		Good Listener	
	HMM	CRF	HMM	CRF	HMM	CRF
Rater-1	0.228	0.193	0.231	0.190	0.222	0.202
Rater-2	0.210	0.185	0.195	0.168	0.208	0.185

HMMs outperform CRFs in most cases

Summary and future work

- Three issues in predicting transitions
  – Individual differences
    - Large differences between raters
    - It is better to aim for rater-dependent model
  – Evaluation criteria
    - Match Rate per rating (MR/r)
  – Prediction models
    - HMMs outperform CRFs
    - CRFs overtuned to output likely ratings
- Future work
  – other metrics, improving prediction performance with other features