SIGDIAL 2010, 24th Sept 2010

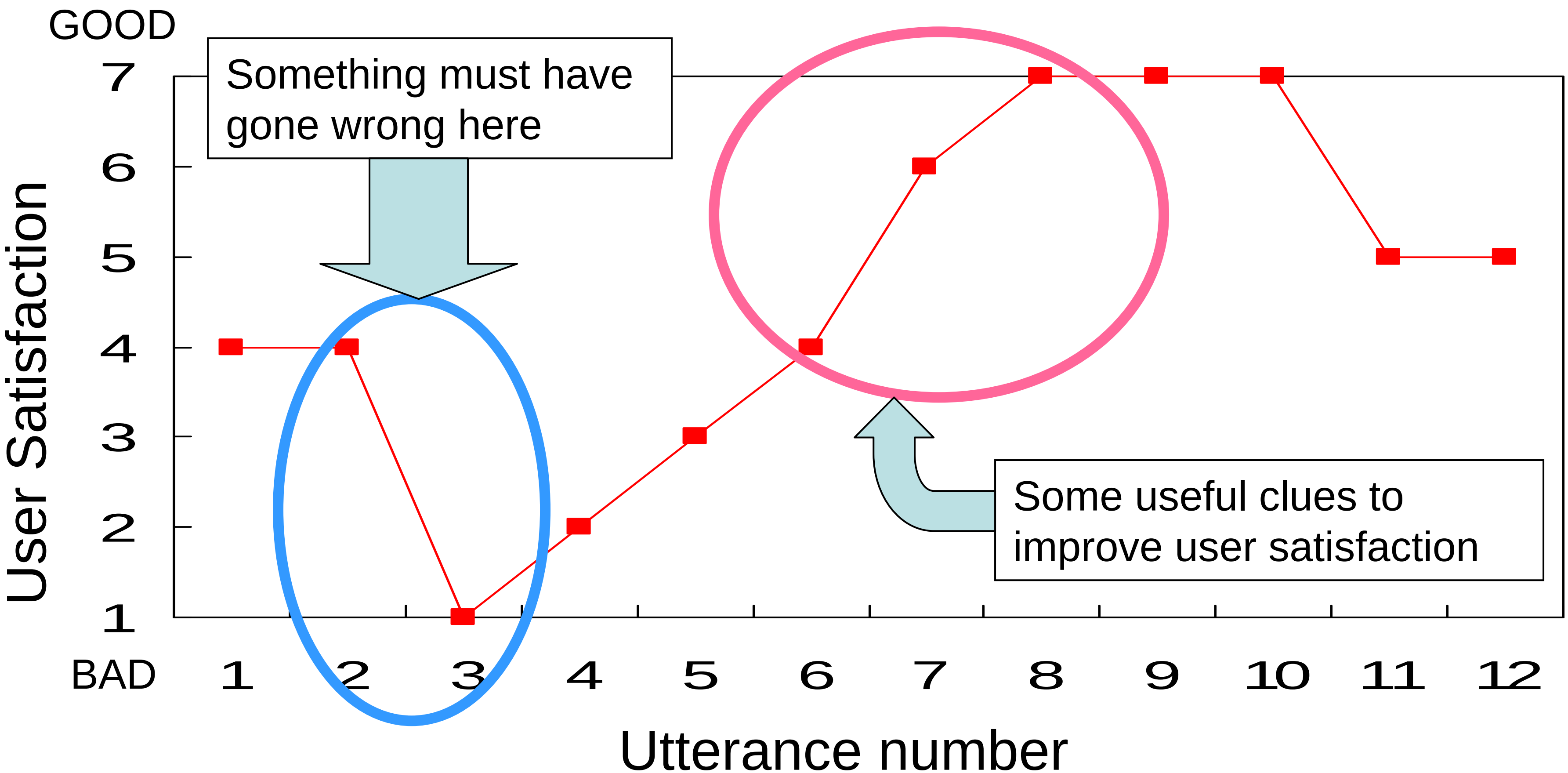
Modeling User Satisfaction Transitions in Dialogues from Overall Ratings

Ryuichiro Higashinaka, Yasuhiro Minami, Kohji Dohsaka, Toyomi Meguro

NTT Corporation

Objective

- Predict user satisfaction transitions during a dialogue
  - Useful for an analysis to improve dialogue systems



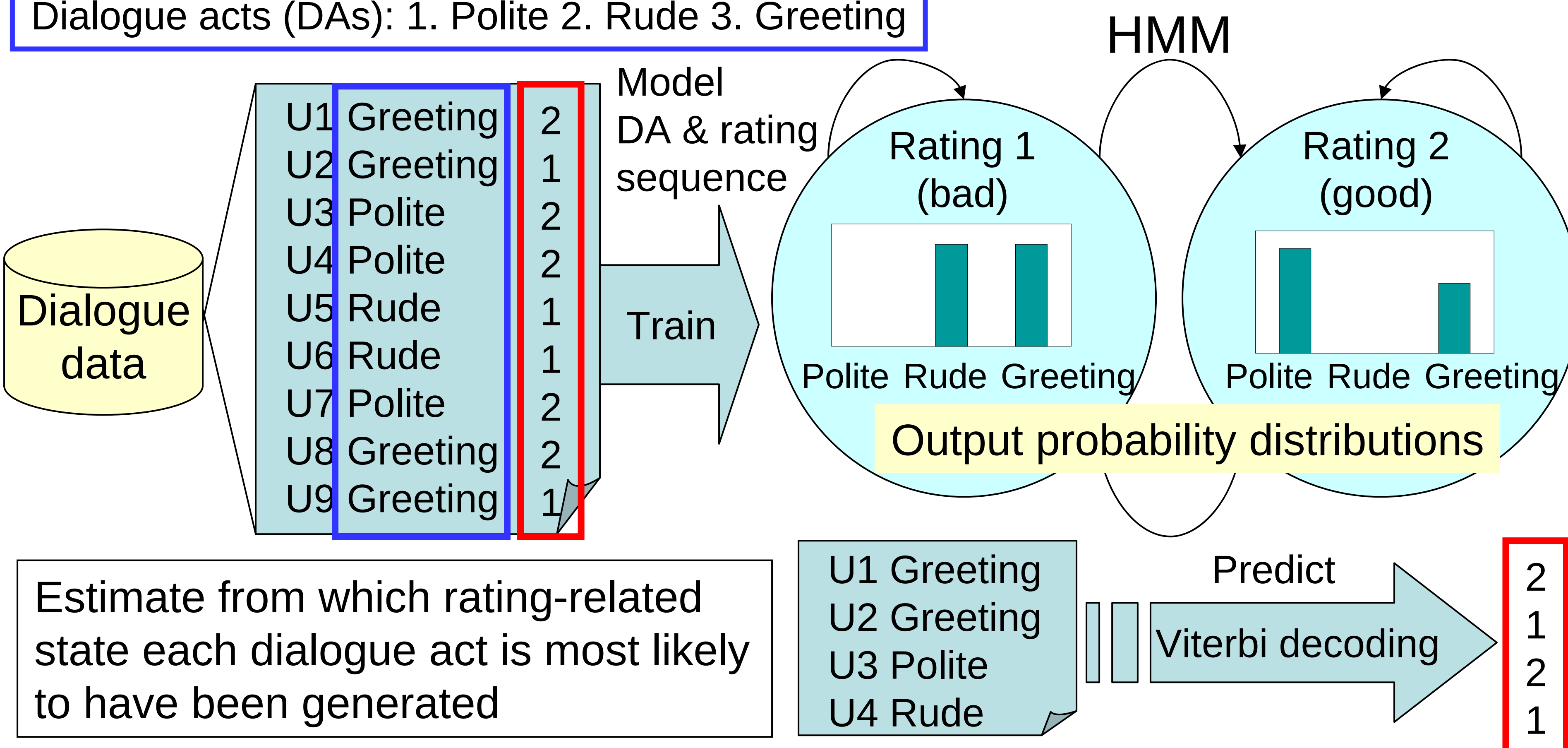
Related Work

- Plenty of work on predicting overall user satisfaction of a dialogue
  - PARADISE (Walker et al., Möller et al.)
  - N-gram based method (Hara et al.)
- Little work on predicting user satisfaction transitions during a dialogue
  - One exception: Engelbrecht et al., 2009
  - Models user satisfaction transitions using hidden Markov models (HMMs)

Engelbrecht et al., 2009

- Uses manually-labeled reference transitions

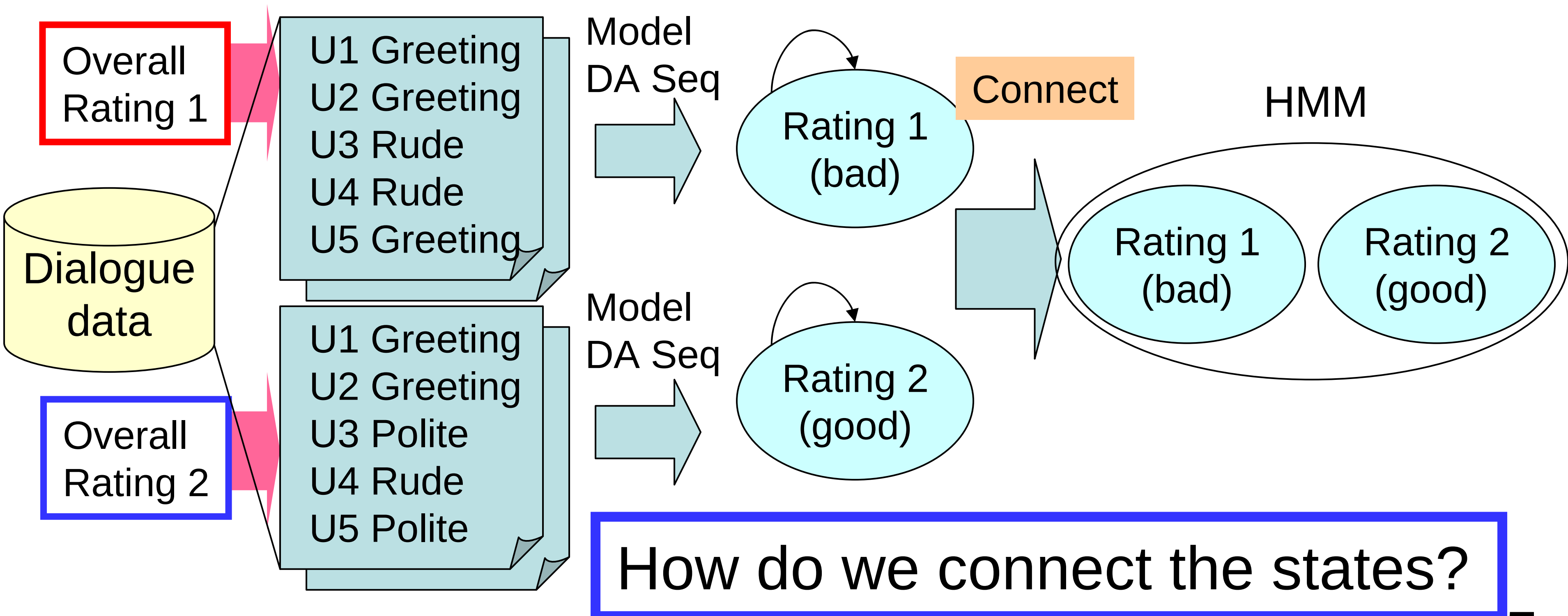
Dialogue acts (DAs): 1. Polite 2. Rude 3. Greeting



Problem: High cost for making reference data

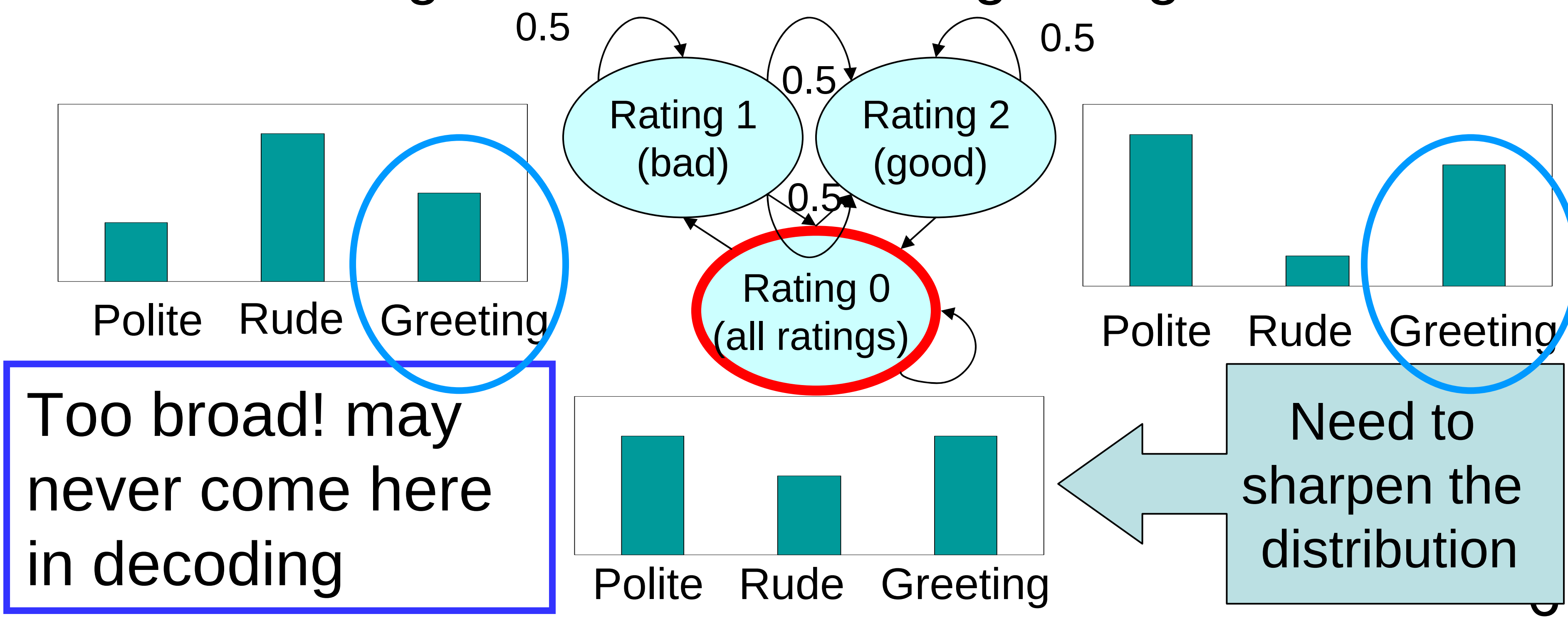
Approach

- Train an HMM from dialogues with overall ratings
  - No utterance-level references are necessary
  - Requires only a single overall rating per dialogue



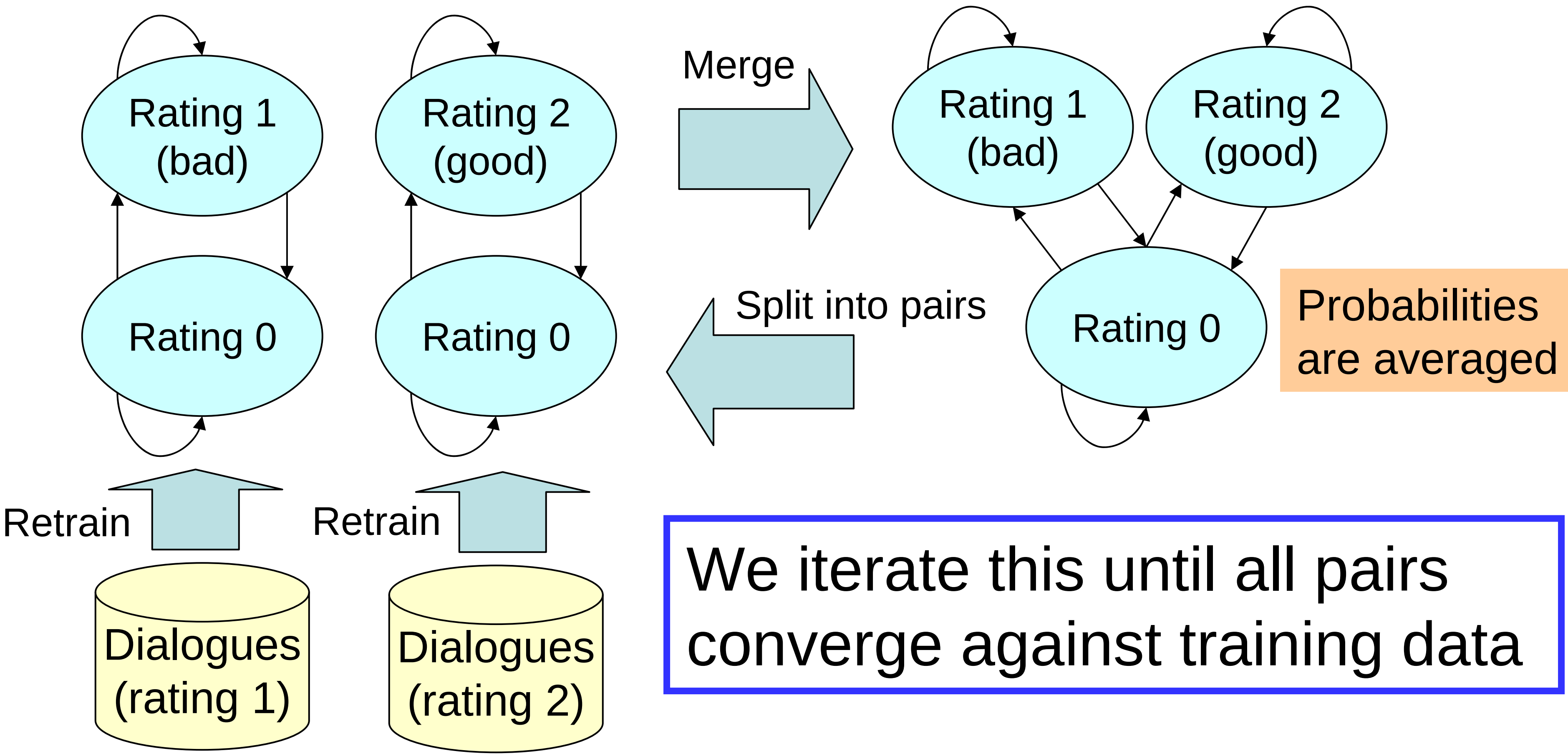
How to connect the states

- Connect with equal transition probabilities
  ➔ Common dialogue acts (e.g. Greeting) are always predicted as the same rating
- Add a state (common state) trained from the data of all ratings to avoid evaluating rating-wide DAs



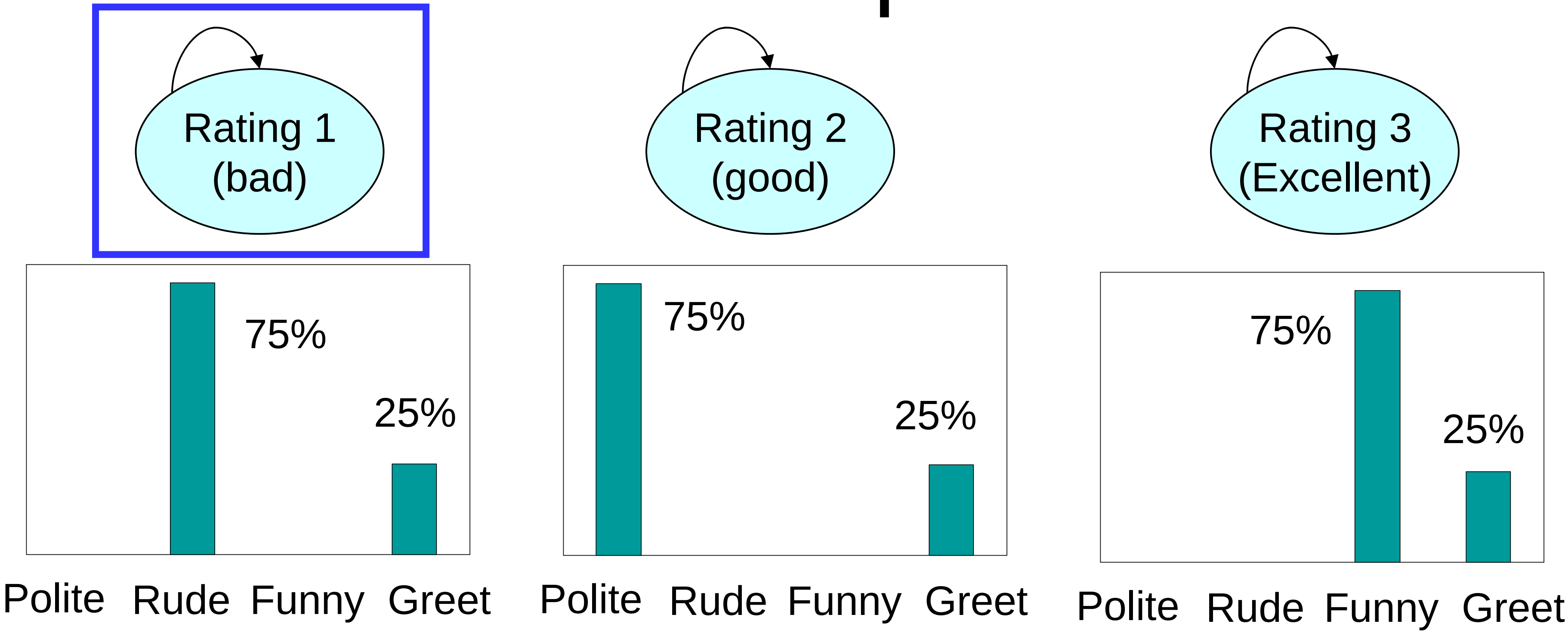
Solution: Concatenated Training

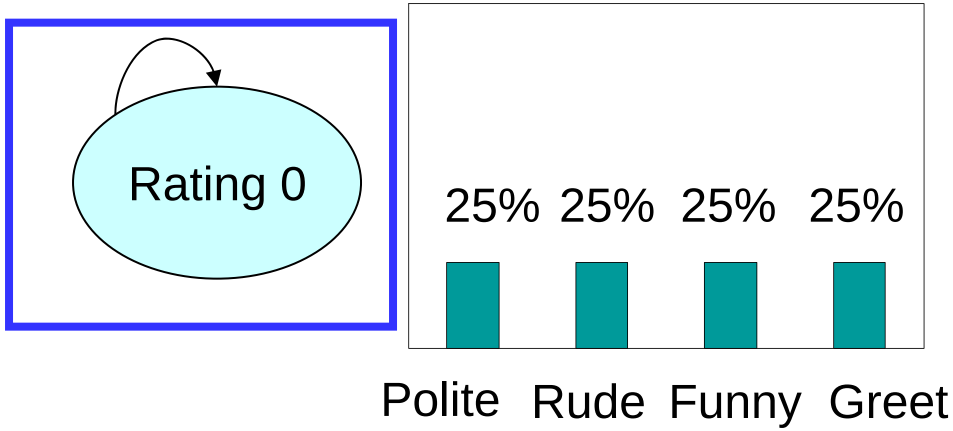
- We can sharpen the output distributions



We iterate this until all pairs converge against training data

Example

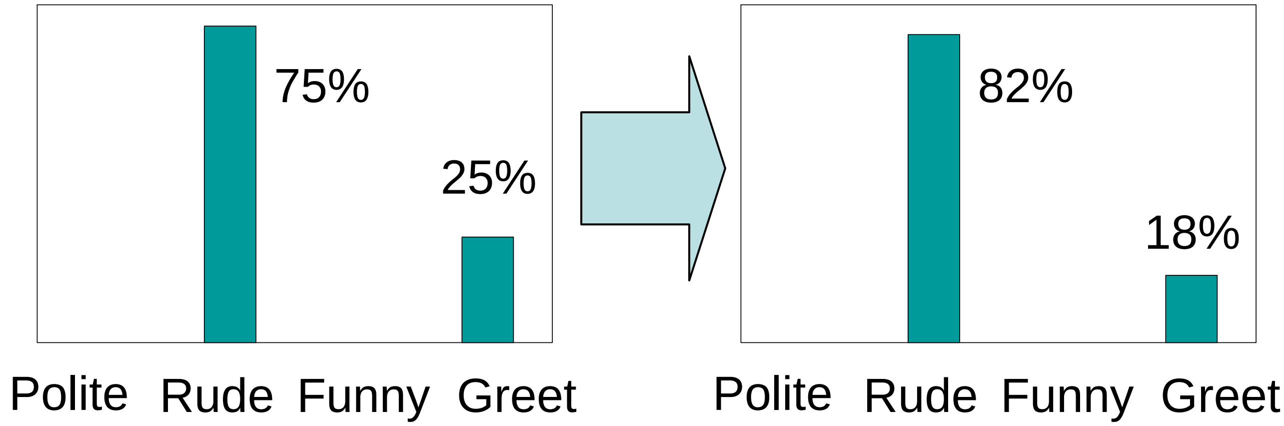


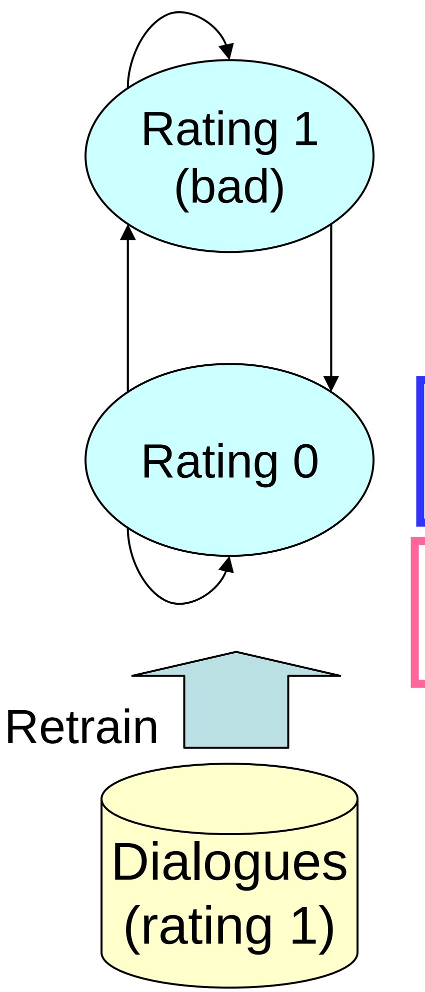


NOTE:
We ignore transition probabilities.
We assume the same initial probabilities

EM algorithm updates probabilities based on the expected counts

75*0.75=56.25
25*0.5=12.5
➔ 4.5:1

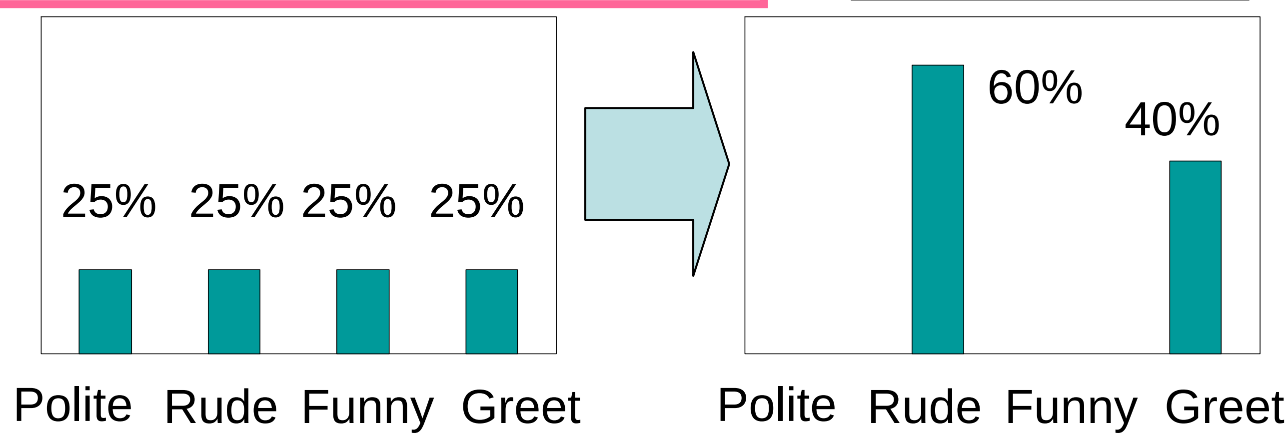




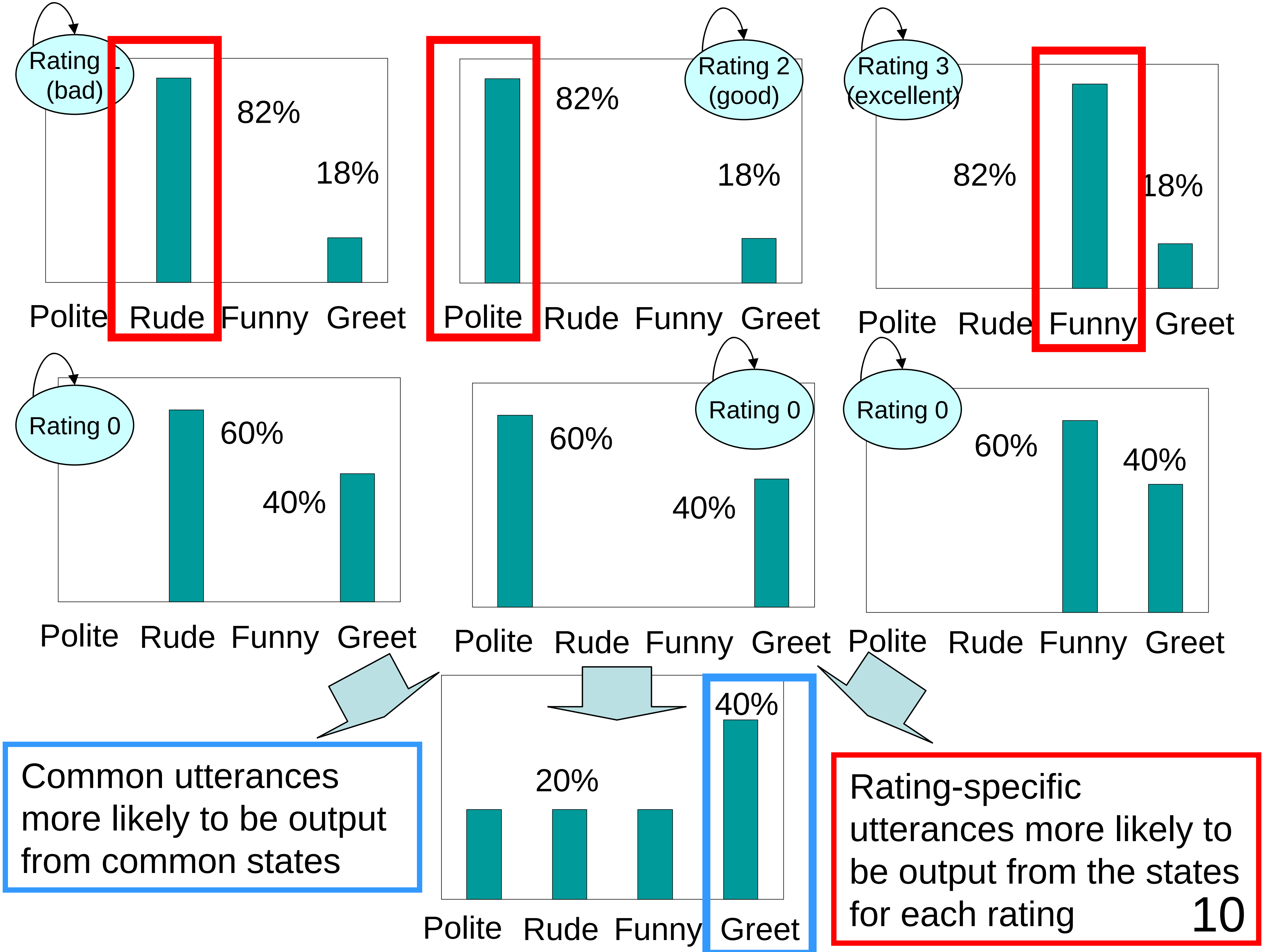
75% of “Rude” output from Rating 1
25% of “Rude” output from Rating 0

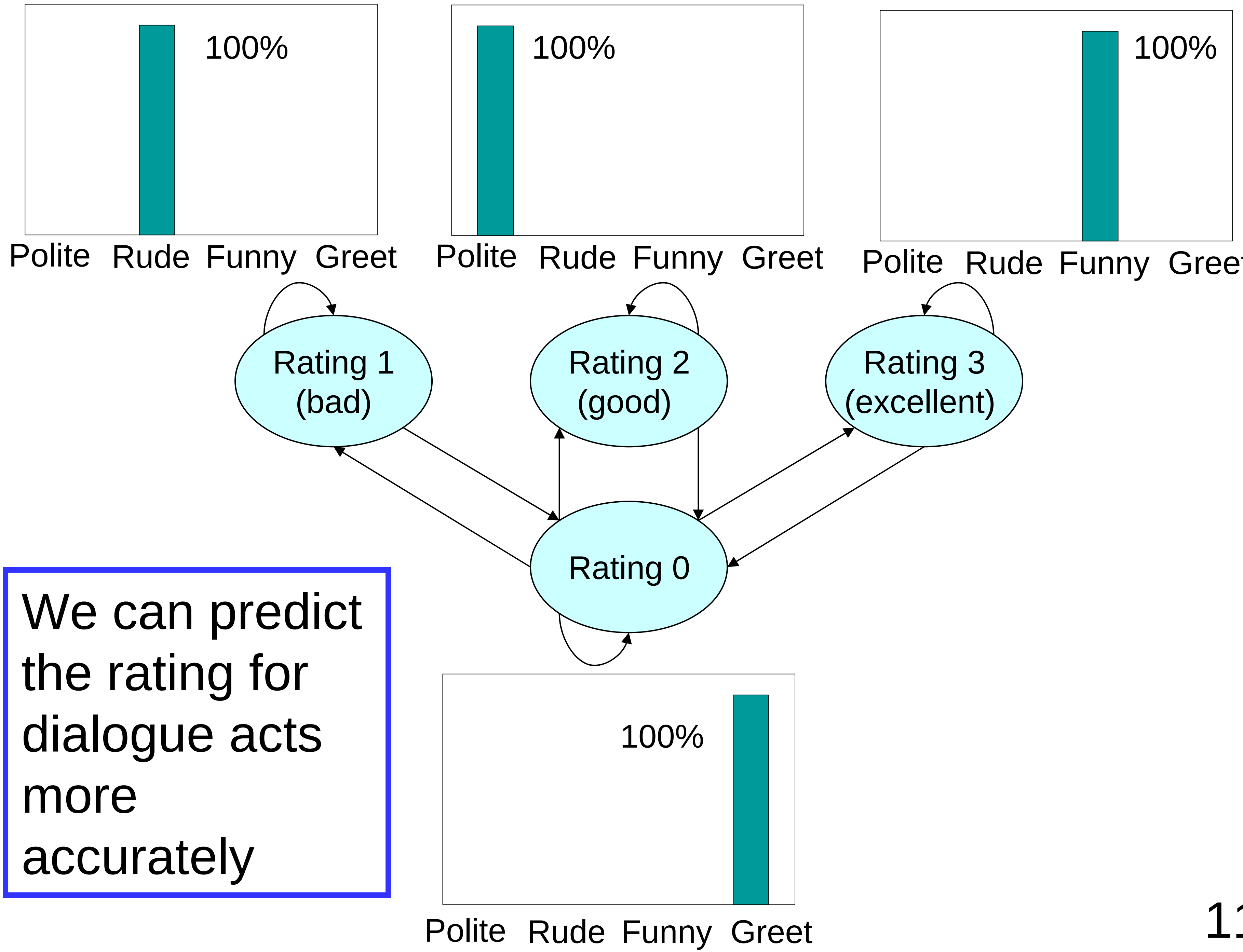
50% of “Greet” output from Rating 1
50% of “Greet” output from Rating 0

75*0.25=18.75
25*0.5=12.5
➔ 1.5:1



“Rude” 75 counts
“Greet” 25 counts


Rating 1 (bad)
82%
18%
Polite Rude Funny Greet
82%
Rating 2 (good)
18%
Polite Rude Funny Greet
Rating 3 (excellent)
82%
18%
Polite Rude Funny Greet
Rating 0
60%
40%
Polite Rude Funny Greet
60%
Rating 0
40%
Polite Rude Funny Greet
Rating 0
60%
40%
Polite Rude Funny Greet
40%
20%
Polite Rude Funny Greet
Common utterances more likely to be output from common states
Rating-specific utterances more likely to be output from the states for each rating



We can predict the rating for dialogue acts more accurately

Experiment

- Dialogue data (text chat) in two domains
  - Animal Discussion (human-system)
    - User and system discuss about animals
  - Attentive Listening (human-human)
    - Listener attentively listens to a speaker
  - Annotated with dialogue acts and US ratings (7-levels)

	Animal Discussion	Attentive Listening
# dialogues with overall ratings	180	1260
# dialogues with utterance-level ratings	90 (subset)	100 (subset)
Dialogue act types	29	40

- Comparison with a random baseline and an upper bound (Engelbrecht et al.) trained from reference transitions

Evaluation Criterion

- Match rate per rating (MR/r)
  - Equally treats easy and difficult-to-guess ratings
    - We want to predict rare but important ratings
  - Reliable than other metrics

(Higashinaka et al., 2010, IWSDS)

	U1	U2	U3	U4	U5	U6	U7	U8	U9	U10	U11	U12
Ref	4	4	5	5	6	6	7	4	3	2	1	1
Hyp	4	0	0	6	6	7	0	5	0	3	1	0
Hyp’	4	4	4	6	6	7	7	5	5	3	1	1

Rating	1	2	3	4	5	6	7
Match rate	2/2	0/1	0/1	2/3	0/2	1/2	1/1

Avg → MR/r 0.45

Results

Simple connection of rating-related states with equal transition probabilities

Animal Discussion domain

	random	simple	concat	Upper B
Smoothness	0.143	0.137	0.177	0.217
Closeness	0.143	0.156	0.164	0.231
Willingness	0.143	0.152	0.181	0.224

Attentive Listening domain

	random	simple	concat	Upper B
Smoothness	0.141	0.118	0.167	0.231
Closeness	0.143	0.090	0.159	0.237
Good Listener	0.143	0.121	0.224	0.227

Common states with concatenated training improve prediction performance

Summary and Future work

- A novel approach for predicting user satisfaction transitions using HMMs
  - Uses only the dialogues with overall user satisfaction ratings
  - Reduces the cost for training prediction models
  - Can be a viable option for evaluation
    - Overall ratings can be obtained easily
- Future work
  - New emissions to improve prediction accuracy
  - Apply our HMMs to other tasks