Speaker Sensitive Response Evaluation Model

JinYeong Bak, Alice Oh
jy.bak@kaist.ac.kr, alice.oh@kaist.edu
School of Computing, KAIST

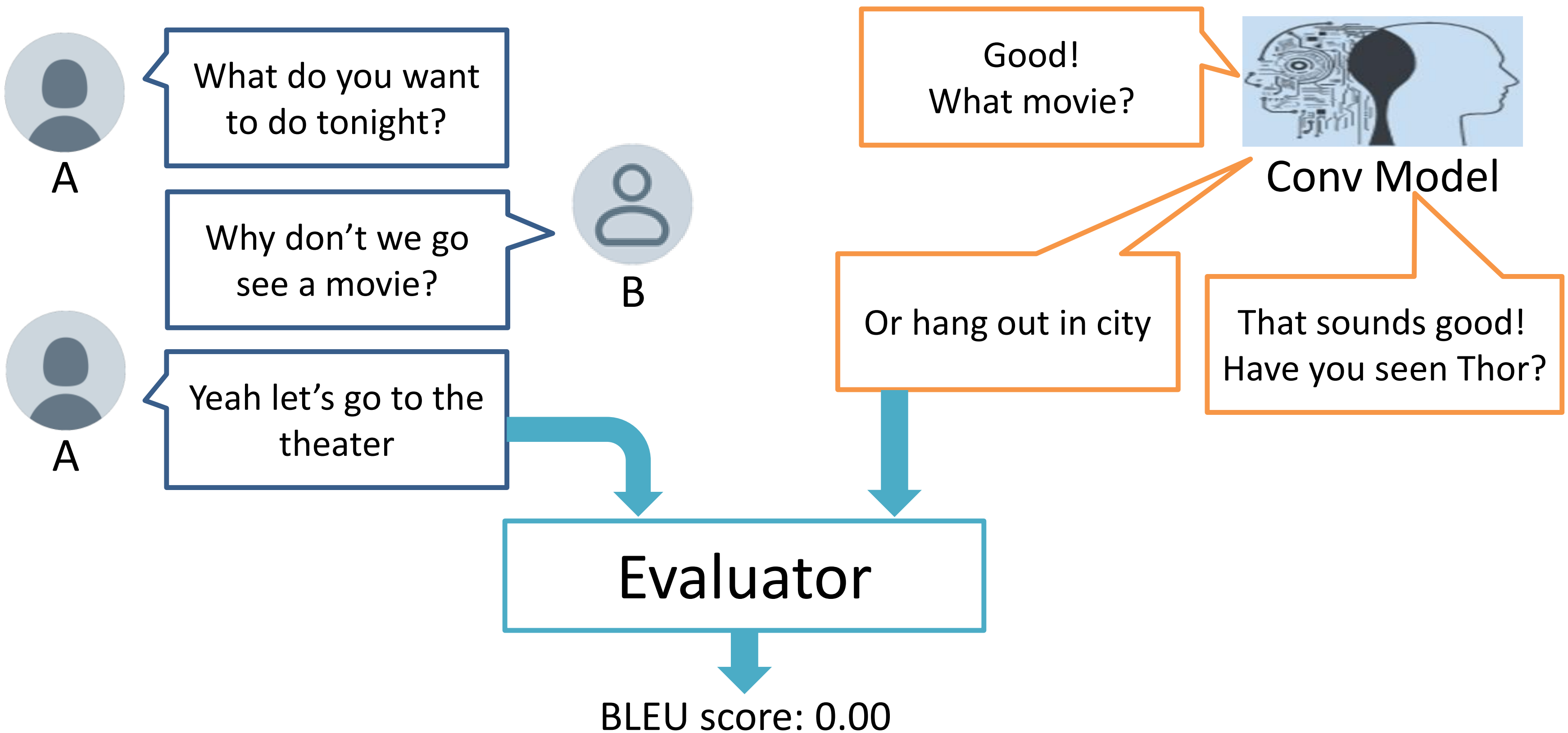
Motivation

1. Human annotation is resources-consuming
   - Requires money and evaluation time
   - Low scalability
   - i.e. Evaluating 450 responses in Amazon Mturk
     - Time: 5 hours
     - Cost: $300
     - Untrustworthy answers rate: 15%

Motivation

2. Responses of conversation can be various

Evaluation metrics take generated response and ground truth



Motivation

2. Responses of conversation can be various

Evaluation metrics take generated response and ground truth

Type	Utterance	Human	BLEU	Emb Avg
Context	A: What do you want to do tonight? B: Why don’t we go see a movie?			
Ground Truth	A: Yeah Let’s go to the theater			
Candidate 1	That sounds good! Have you seen Thor?	5.00		
Candidate 2	Good! What movie?	5.00		
Candidate 3	Or hang out in city	3.80		
Candidate 4	The weather is no good for walking	2.60		
Candidate 5	The sight is extra beautiful here	1.00		
Candidate 6	Enjoy your concert	1.00		

Motivation

3. Existing metrics that consider the given conversation
   – Automatic Dialogue Evaluation Model (ADEM) [Lowe et al., ACL 2017]
   – Referenced metric and Unreferenced metric Blended Evaluation Routine (RUBER) [Tao et al., AAAI 2018]

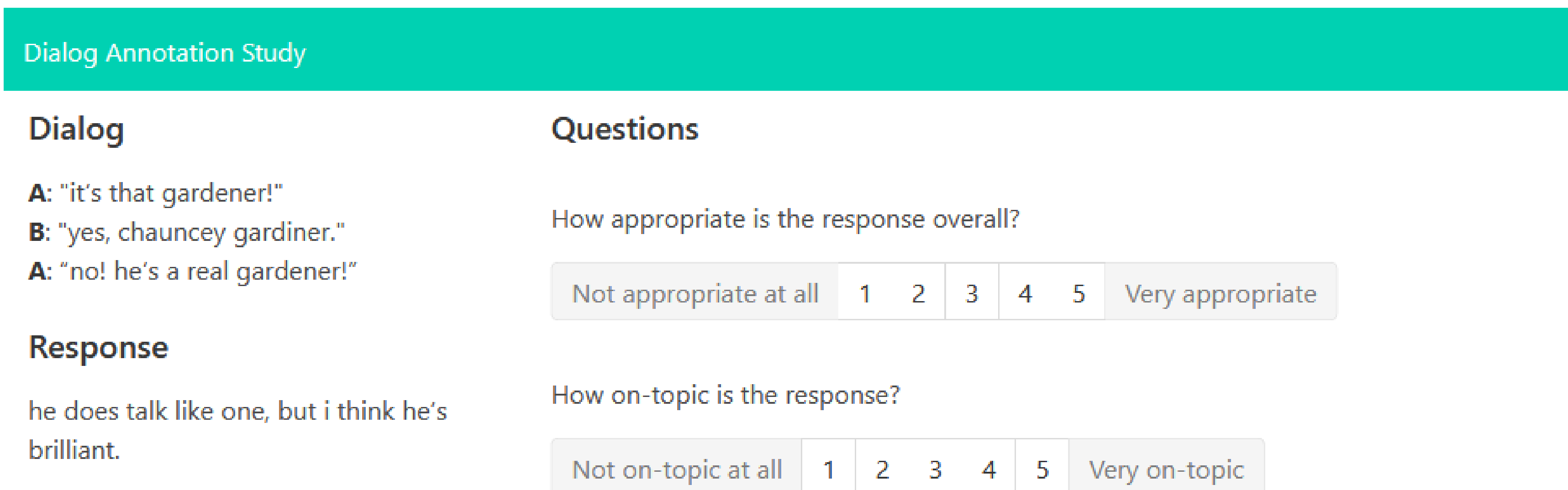
Motivation

3. Existing metrics that consider the given conversation
   - High scores to non-appropriate responses

Type	Utterance	Human	ADEM	RUBER
Context	A: What do you want to do tonight? B: Why don’t we go see a movie?			
Ground Truth	A: Yeah Let’s go to the theater			
Candidate 1	That sounds good! Have you seen Thor?	5.00	2.04	0.59
Candidate 2	Good! What movie?	5.00	2.06	0.55
Candidate 3	Or hang out in city	3.80	1.82	0.48
Candidate 4	The weather is no good for walking	2.60	**2.17**	0.47
Candidate 5	The sight is extra beautiful here	1.00	2.13	**0.64**
Candidate 6	Enjoy your concert	1.00	1.94	0.57

Motivation

3. Existing metrics that consider the given conversation
   – High scores to non-appropriate responses
   – Need human labeled score for responses to train the model



Dialog Annotation Study

Dialog

A: "it's that gardener!"
B: "yes, chauncey gardiner."
A: "no! he's a real gardener!"

Response

he does talk like one, but i think he's brilliant.

Questions

How appropriate is the response overall?

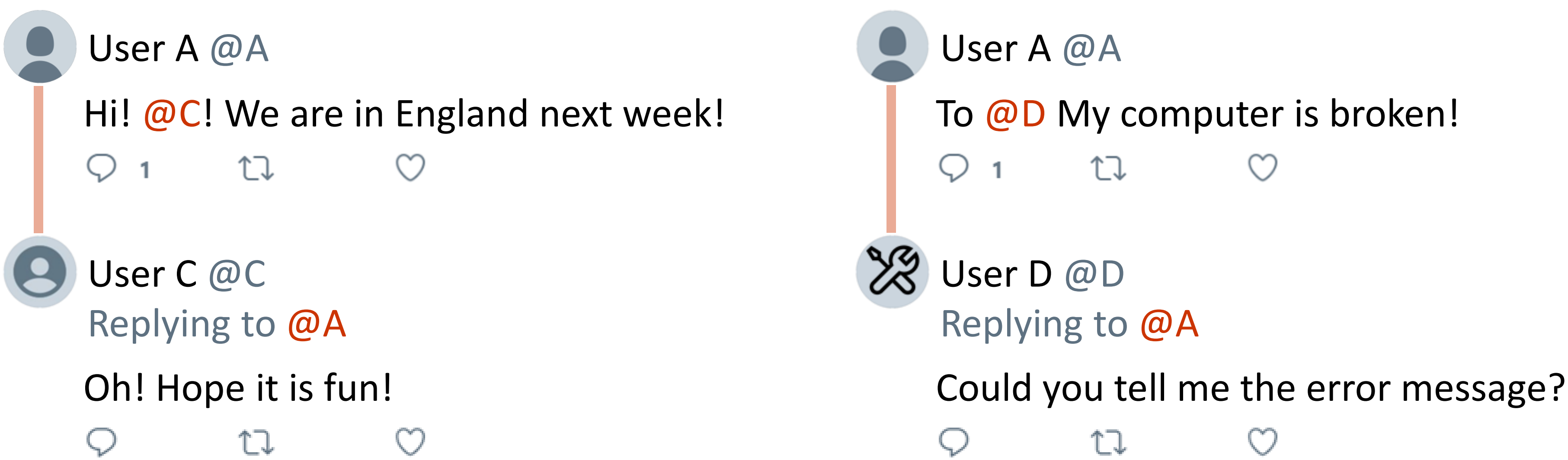
Not appropriate at all | 1 | 2 | 3 | 4 | 5 | Very appropriate

How on-topic is the response?

Not on-topic at all | 1 | 2 | 3 | 4 | 5 | Very on-topic

Motivation

4. Difference of utterances depends on the speakers
   - A speaker is likely to have similar utterances with same conversational partner
   - A speaker may say different utterances with different conversational partners



Preliminary Study – Experiment Setup

Difference of utterances depends on the speakers

- Categorize utterances into four sets
  - Same Conversation (SC_A): Speaker *A*'s utterances in a conversation
  - Same Partner (SP_A): *A*'s utterances in conversations with the same partner
  - Same Speaker (SS_A): *A*'s utterances
  - Random ($Rand_A$): Random utterances from speakers who are not *A*

Preliminary Study – Experiment Setup

Difference of utterances depends on the speakers

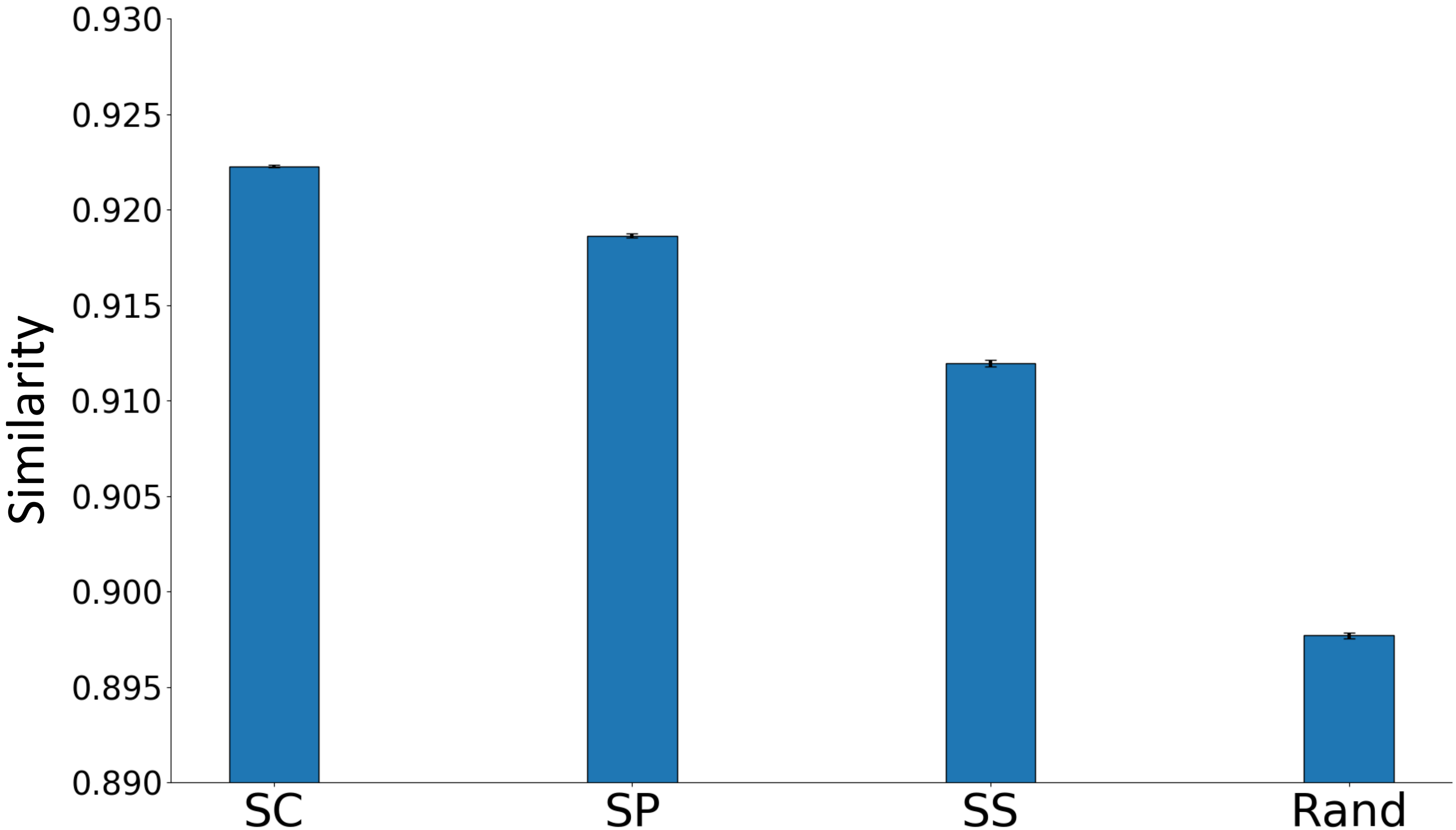
- Categorize utterances into four sets
  - Same Conversation (SC_A): Speaker *A*’s utterances in a conversation
  - Same Partner (SP_A): *A*’s utterances in conversations with the same partner
  - Same Speaker (SS_A): *A*’s utterances
  - Random ($Rand_A$): Random utterances from speakers who are not *A*
- Create utterance vectors by GloVe [Pennington et al., EMNLP 2014]
- Compute the similarity of vectors by Frobenius norm

Twitter Conversation Corpus

- A Twitter conversation
  - Five or more tweets
  - At least two replies by each user

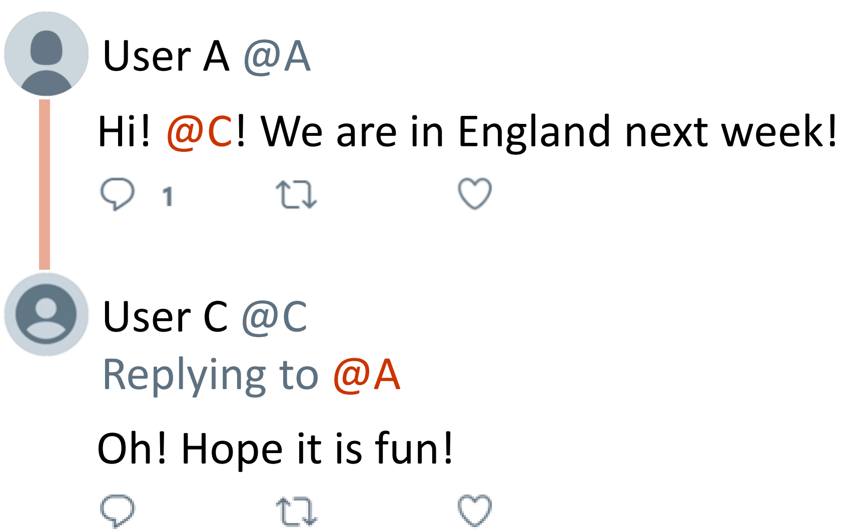
- Statistics
  - 27K users
  - 107K dyads
  - 770K conversations
  - 6M tweets
  - 7 years (2007-2013)

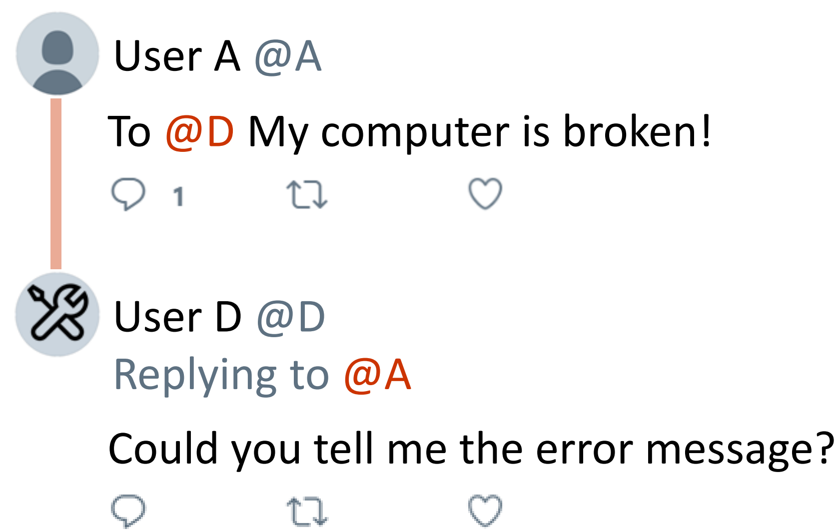
Preliminary Study – Result



Motivation

4. Difference of utterances depends on the speakers
   - A speaker is likely to have similar utterances with same conversational partner
   - A speaker may say different utterances with different conversational partners

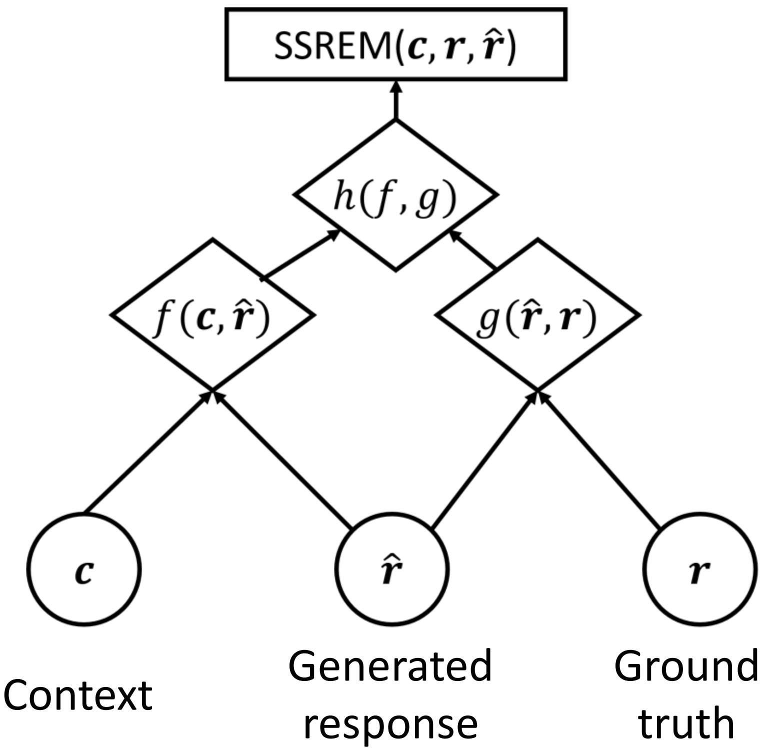




SSREM

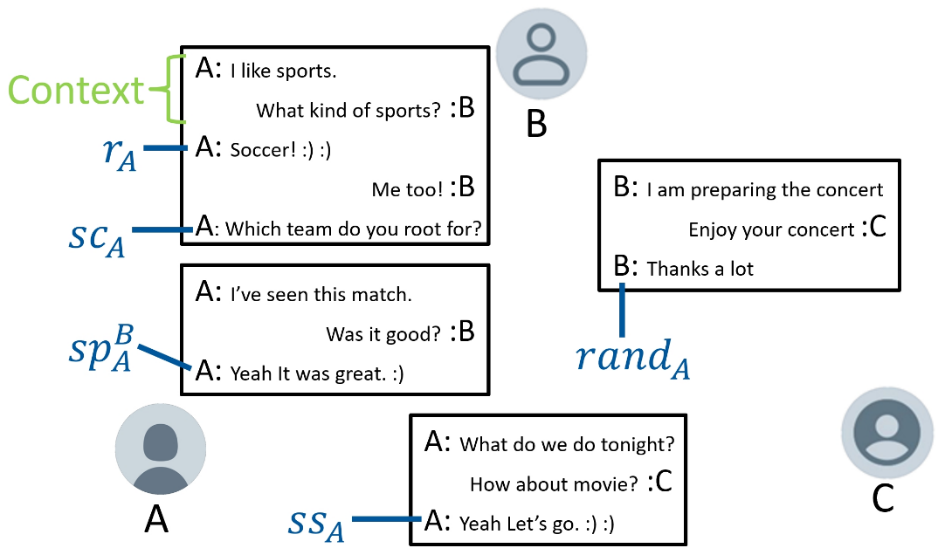
Speaker Sensitive Response Evaluation Model

- Input
  - Context $\boldsymbol{c}$
  - Generated response $\hat{\boldsymbol{r}}$
  - Ground truth $\boldsymbol{r}$
- Algorithm
  - Measures correlation $\hat{\boldsymbol{r}}$ and $\boldsymbol{c}$
    Learnable function $f(\boldsymbol{c}, \hat{\boldsymbol{r}})$
  - Measures correlation $\hat{\boldsymbol{r}}$ and $\boldsymbol{r}$
    Pre-defined function $g(\hat{\boldsymbol{r}}, \boldsymbol{r})$
  - Blends the correlations
    Pre-defined function $h(f, g)$



SSREM - Training f function

- Define $R_{cand} = \{r_A, sc_A, sp_A, ss_A, rand_A\}$ for a context
  - r_A: ground truth response
  - sc_A: one utterance in the same conversation set SC_A
  - sp_A: one utterance in the same partner set SP_A
  - ss_A: one utterance in the same speaker set SS_A
  - $rand_A$: one utterance in the random set $Rand_A$



SSREM - Training f function

- Define $R_{cand} = \{r_A, sc_A, sp_A, ss_A, rand_A\}$ for a context
- Build a classification problem

  Identifies ground truth response r_A from R_{cand}
- Make a classifier that uses the function f

  Probability of r_A given context $\boldsymbol{c}$ and R_{cand}

$$p(r_A|\boldsymbol{c}, R_{cand}) = \frac{\exp(f(\boldsymbol{c}, r_A))}{\sum_{r' \in R_{cand}} \exp(f(\boldsymbol{c}, r'))}$$

- Use Twitter conversation corpus to train f

  27K users, 107K dyads, 770K conversations, 6M tweets

Result

Type	Utterance	Human	RUBER	SSREM
Context	A: What do you want to do tonight? B: Why don't we go see a movie?			
Ground Truth	A: Yeah Let's go to the theater			
Candidate 1	That sounds good! Have you seen Thor?	5.00	0.59	**0.64**
Candidate 2	Good! What movie?	5.00	0.55	0.62
Candidate 3	Or hang out in city	3.80	0.48	0.49
Candidate 4	The weather is no good for walking	2.60	0.47	0.44
Candidate 5	The sight is extra beautiful here	1.00	**0.64**	0.38
Candidate 6	Enjoy your concert	1.00	0.57	0.33

Experiment 1

- Goal: Correlation with human scores
- Human scores
  - Annotate the appropriateness of 1,200 responses
    - Twitter conversations
    - Movie scripts
  - Use Amazon MTurk

Human Score	1	2	3	4	5
Twitter	211	258	342	278	71
Movie	279	267	311	217	126

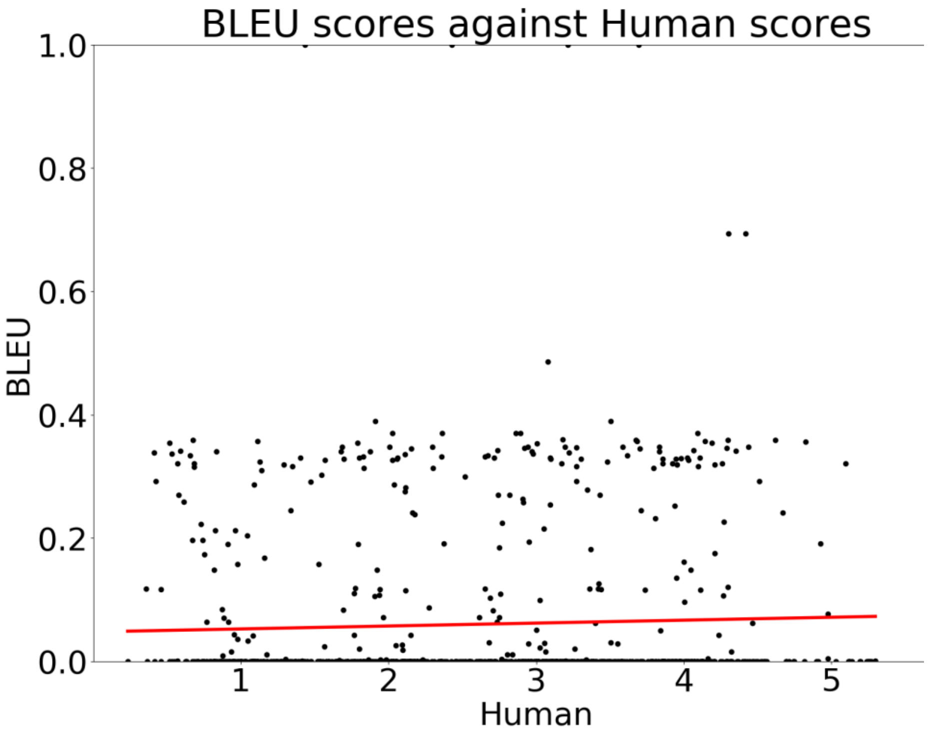
Experiment 1

- Comparison metrics
  - BLEU [Papineni et al., ACL 2002]
  - ROUGE-L [Lin, TSBO 2004]
  - EMB [Liu et al., EMNLP 2016]
  - RUBER [Tao et al., AAAI 2018]
  - RSREM ($R_{cand} = \{r_A, rand_A^{(1)}, rand_A^{(2)}, rand_A^{(3)}, rand_A^{(4)}\}$)
- Correlation
  - Spearman
  - Pearson

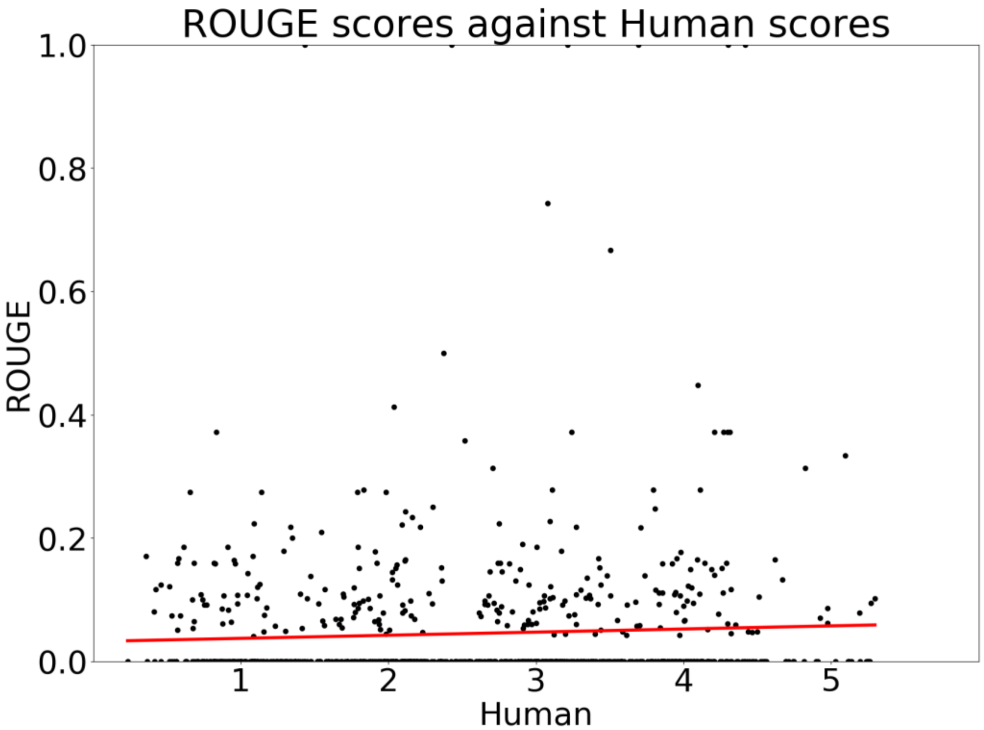
Experiment 1 - Result

Metric	Spearman	Pearson
BLEU	0.024 (0.472)	0.041 (0.227)
ROUGE	0.024 (0.471)	0.052 (0.124)
EMB	0.006 (0.861)	0.012 (0.720)
RUBER	0.044 (0.192)	0.046 (0.177)
RSREM	0.088 (< 0.01)	0.101 (< 0.01)
SSREM	**0.392 (< 0.001)**	**0.378 (< 0.001)**

Experiment 1 - Result

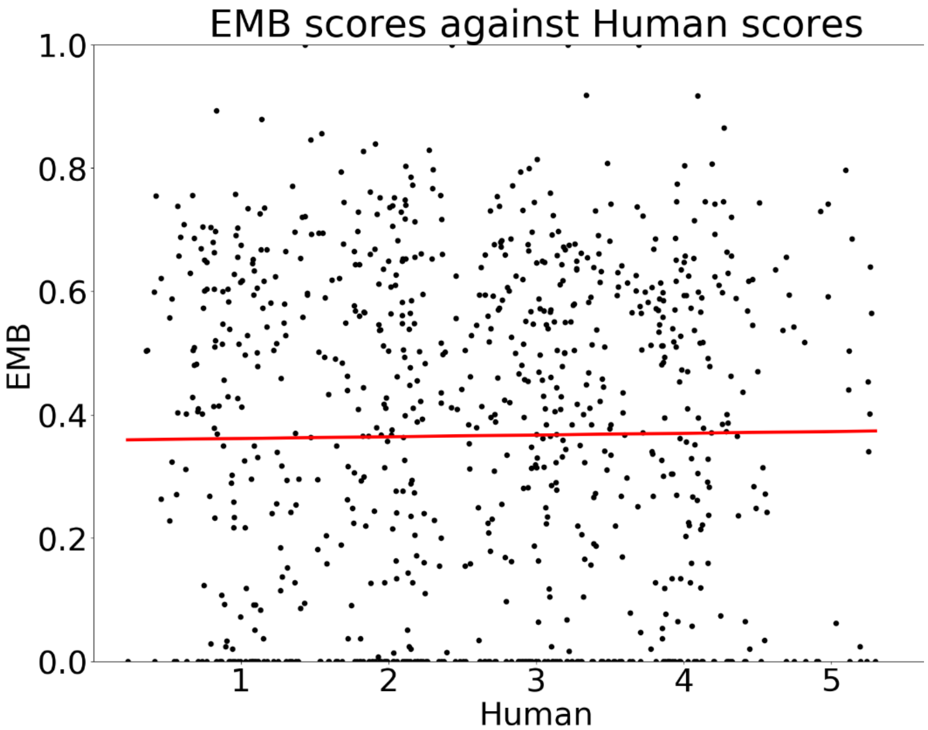


coeff: 0.005

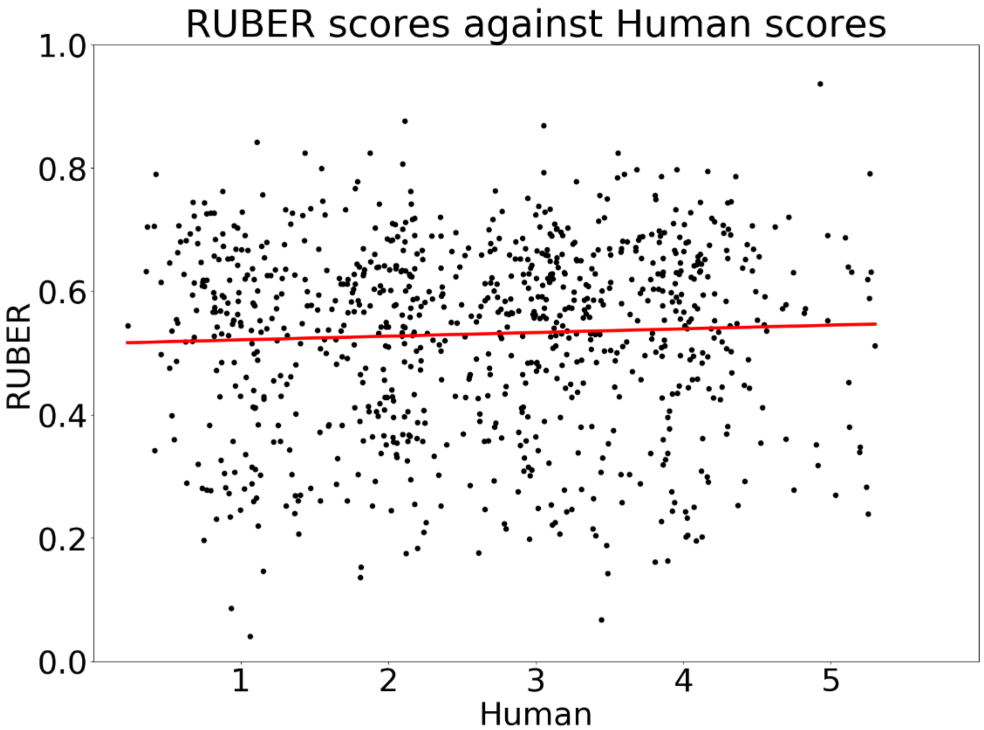


coeff: 0.005

Experiment 1 - Result

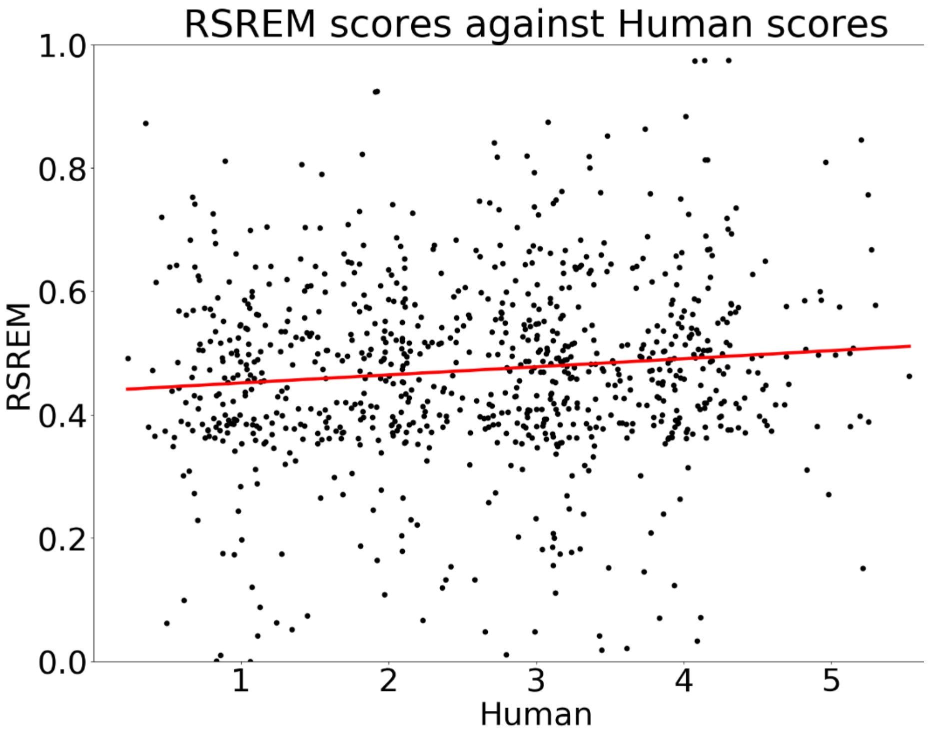


coeff: 0.003

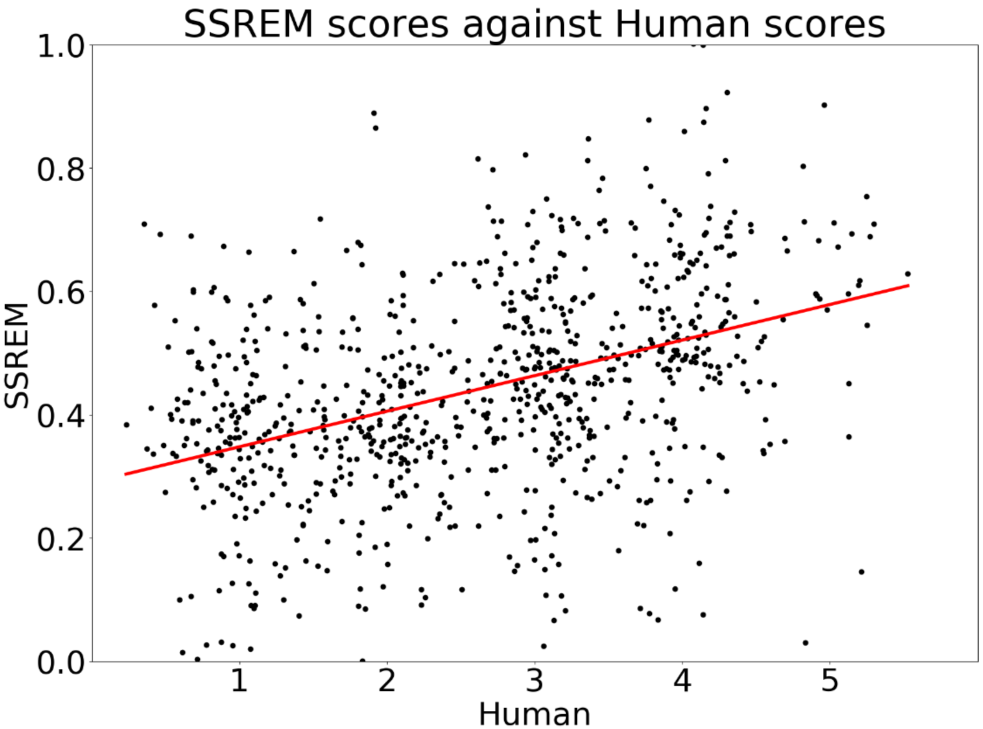


coeff: 0.006

Experiment 1 - Result



coeff: 0.013

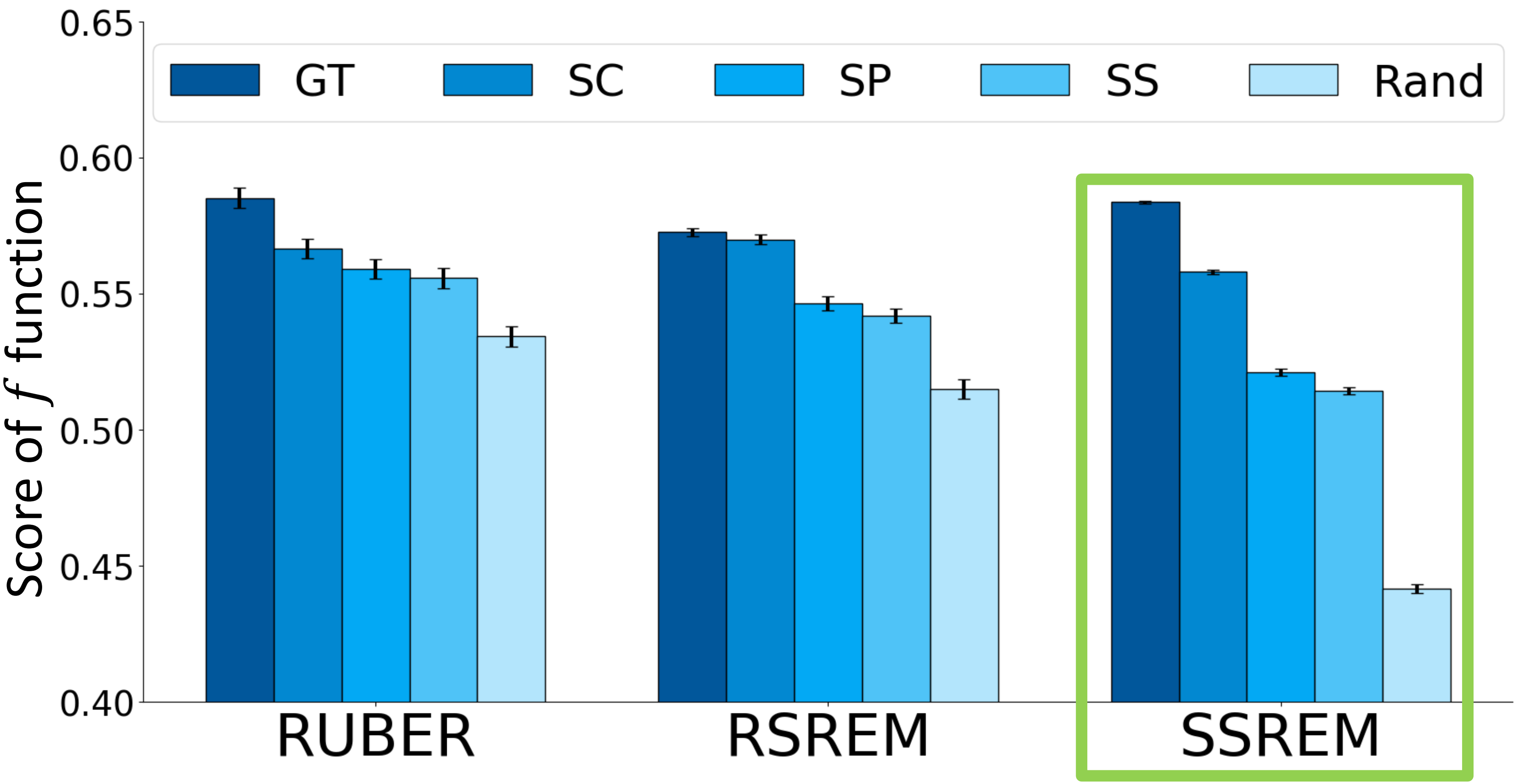


coeff: 0.058

Experiment 2

- Goal: Identifying true/false responses
- Responses
  - True: ground truth (GT)
  - False: SC, SP, SS, $Rand$
- Comparison metrics
  - RUBER [Tao et al., AAAI 2018]
  - RSREM ($R_{cand} = \{r_A, rand_A^{(1)}, rand_A^{(2)}, rand_A^{(3)}, rand_A^{(4)}\}$)

Experiment 2 - Result



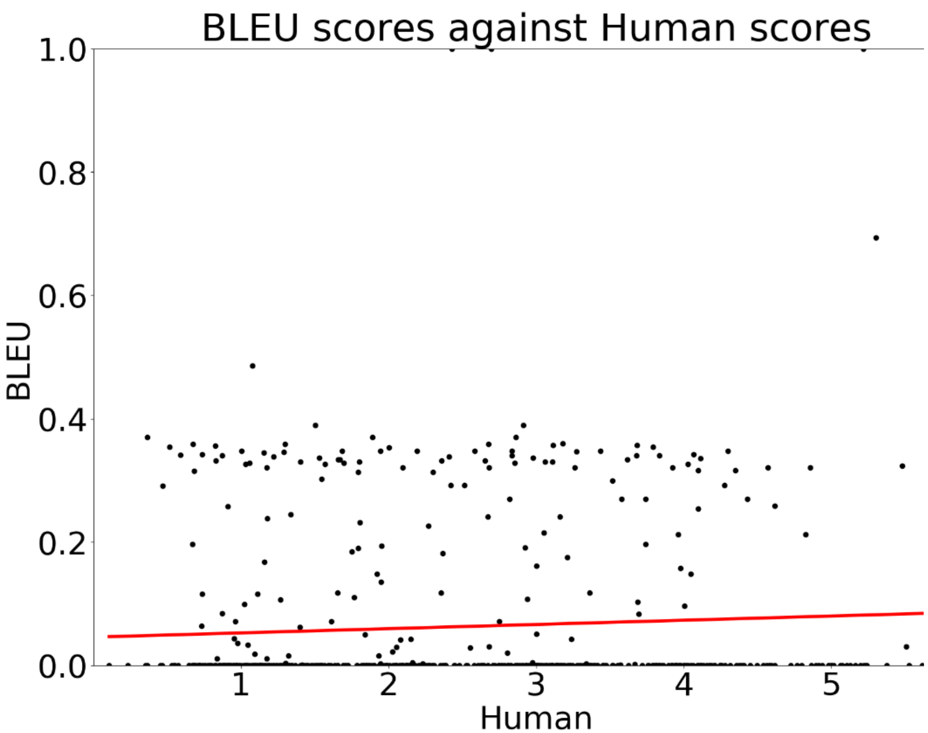
Experiment 3

- Goal: applicability of SSREM
- Data
  - Train: Twitter conversation corpus
  - Test: Movie script
- Method
  - Measuring correlation with human scores
  - Identifying true/false responses

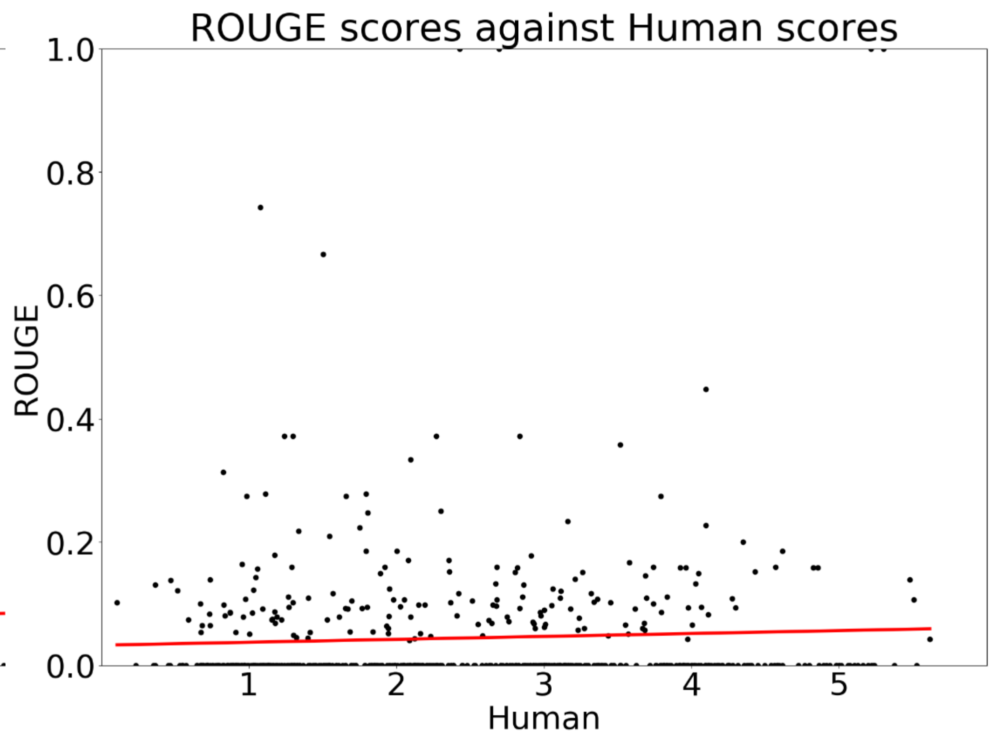
Experiment 3 - Result

Metric	Spearman	Pearson
BLEU	0.036 (0.378)	0.063 (0.124)
ROUGE	0.041 (0.322)	0.054 (0.191)
EMB	0.022 (0.586)	0.010 (0.815)
RUBER	0.004 (0.920)	-0.009 (0.817)
RSREM	0.009 (0.817)	0.024 (0.550)
SSREM	**0.132 (< 0.001)**	**0.119 (< 0.005)**

Experiment 3 - Result

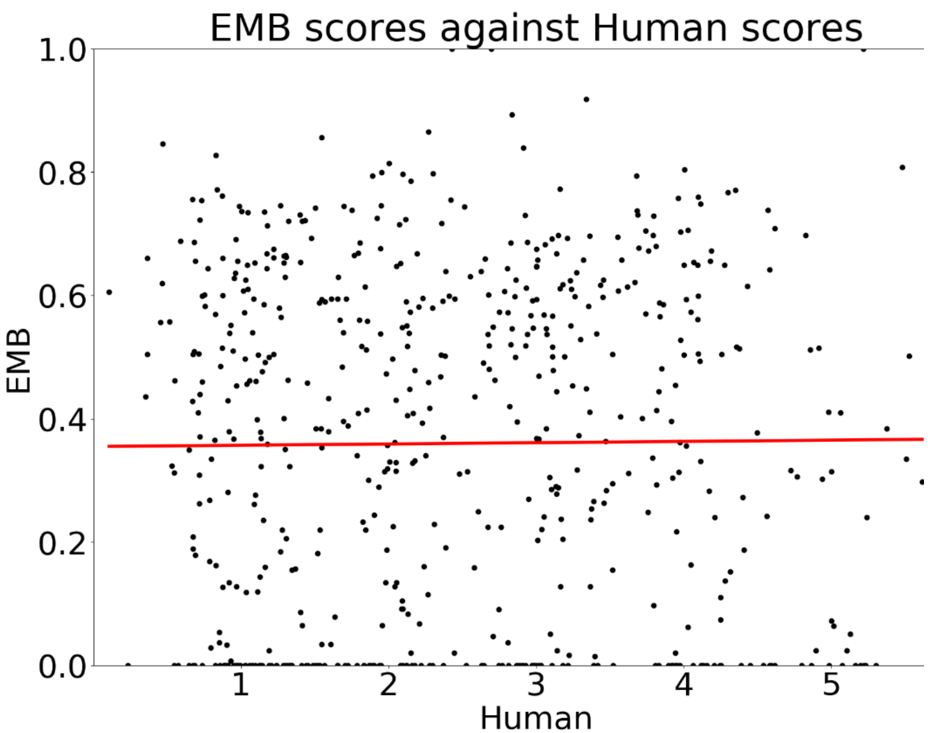


coeff: 0.007

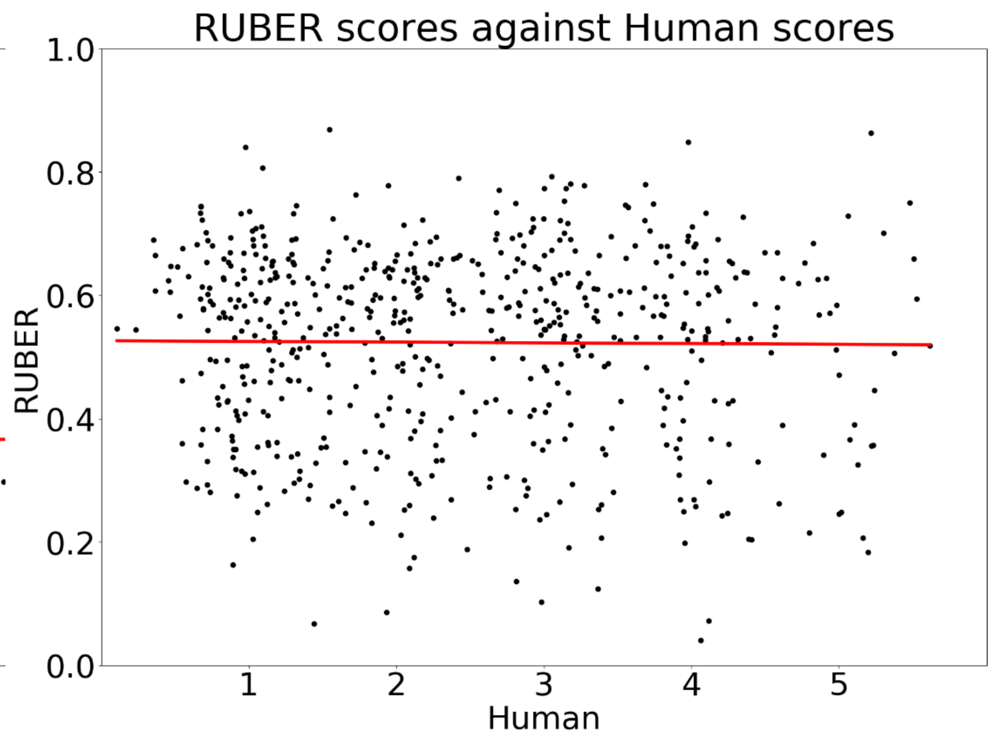


coeff: 0.005

Experiment 3 - Result



coeff: 0.002



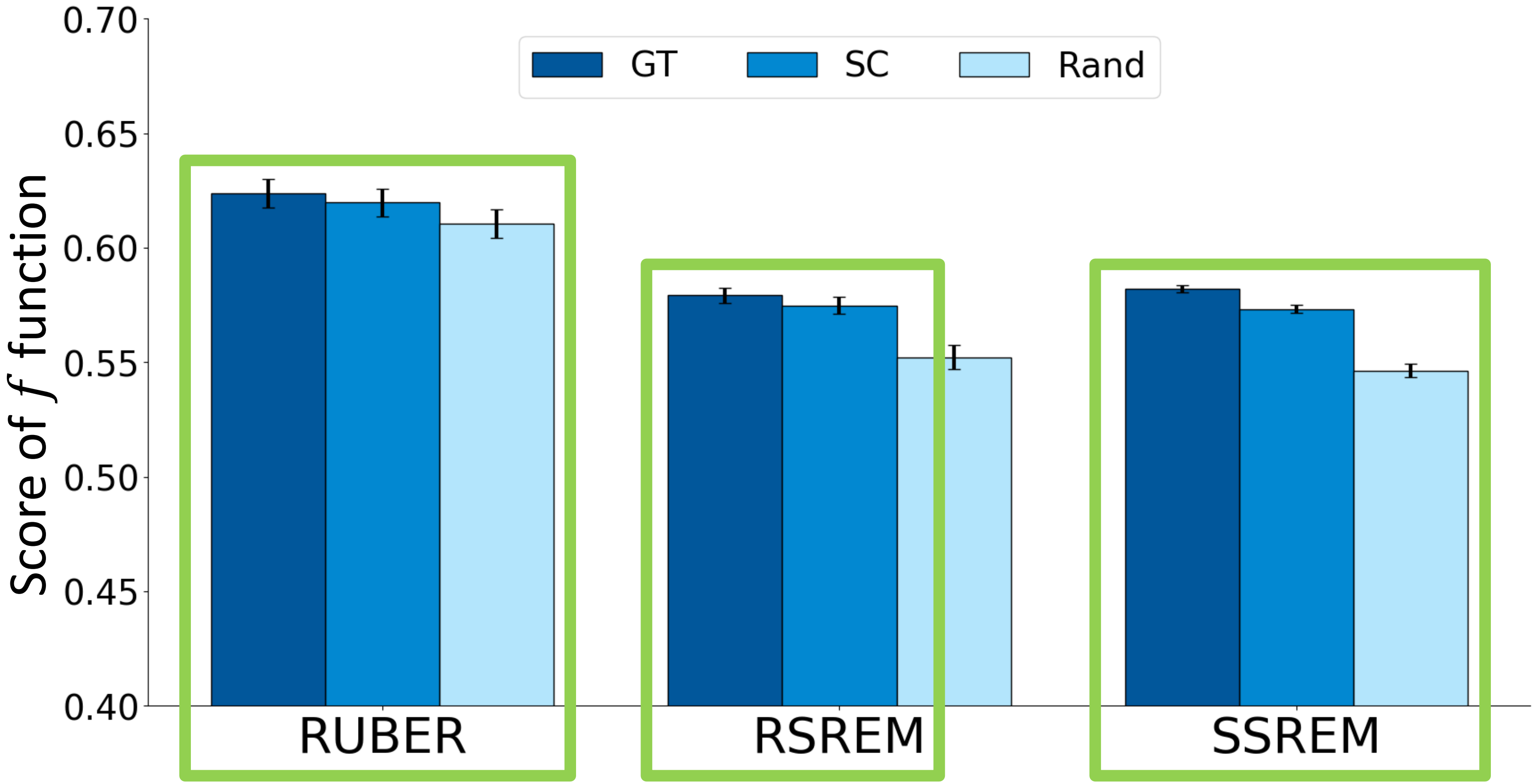
coeff: 0.001

Experiment 3 - Result

coeff: 0.008

coeff: 0.047

Experiment 3 - Result



Example

A

Morning Twitter xo

Morning!! x

B

A

How did u sleep?

good!! wbu??

B (GT)

SSREM: 0.830

ok but it wasn’t long enough

Conv Model

Conclusion

- Suggested new evaluation model for responses (SSREM)
  - Examines conversational context and ground truth response
  - Trains without human labeled data
- Showed experiment results
  - Measure correlations with human scores
  - Identify true / false responses
- Showed applicability of SSREM
  - Test SSREM on the same conversation corpus
  - Test SSREM on different conversation corpus

Thank you!

Any questions or comments?

JinYeong Bak

jy.bak@kaist.ac.kr, nosyu.github.io

U&I Lab, KAIST