# Learning to Control the Specificity in Neural Response Generation

**Ruqing Zhang**, Jiafeng Guo, Yixing Fan, Yanyan Lan, Jun Xu, Xueqi Cheng

1. CAS Key Lab of Network Data Science and Technology
Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China
2. University of Chinese Academy of Sciences, Beijing, China

中国科学院网络数据科学与技术重点实验室
Key Laboratory of Network Data Science & Technology, CAS

中国科学院计算技术研究所
INSTITUTE OF COMPUTING TECHNOLOGY, CHINESE ACADEMY OF SCIENCES

# Background - Dialog

## Task-Oriented Dialog

- Personal assistant, helps people complete specific tasks
- Combination of rules and statistical components


## Chit-Chat Dialog

- No specific goal, attempts to produce natural responses
- Using variants of seq2seq model

# Background – Neural Model

- utterance-response: *n-to-1* relationship
- e.g., the response “Must support! Cheer!” is used for 1216 different input utterances

**Rank-frequency distribution**

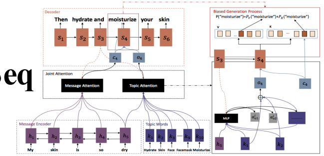

**Seq2Seq framework**

- treat all the utterance-response pairs uniformly
- employ a single model to learn the mapping between utterance and response

favor such general responses with high frequency

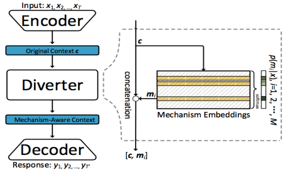
Performance

**TA-Seq2Seq**




- pre-defined a set of topics from an external corpus
- rely on external corpus

**MARM**



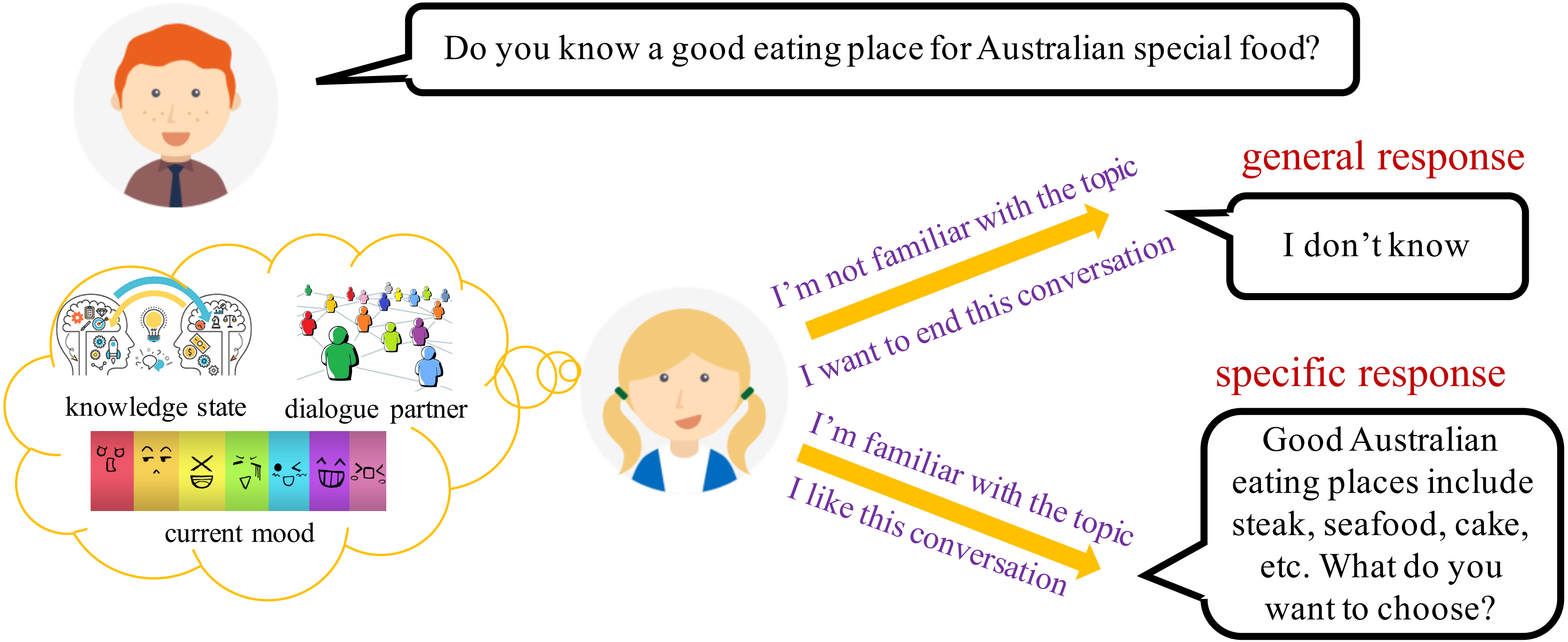

- introduce latent responding factors to model multiple responding mechanisms
- lack of interpretation

How to capture different utterance-response relationships ?

Conversation context
Topic information
Keyword
Coherence
Scenarios heuristics
● ● ●

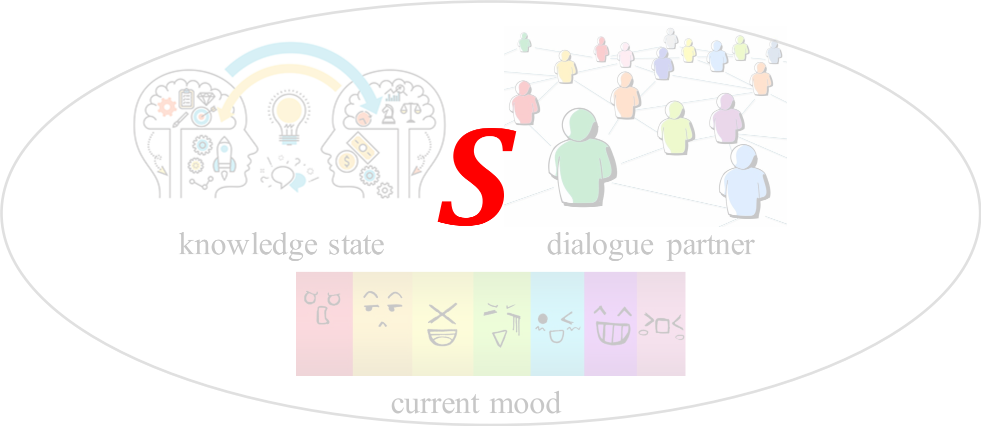
Our motivation comes from
**Human Conversation Process**

# Human Conversation Process



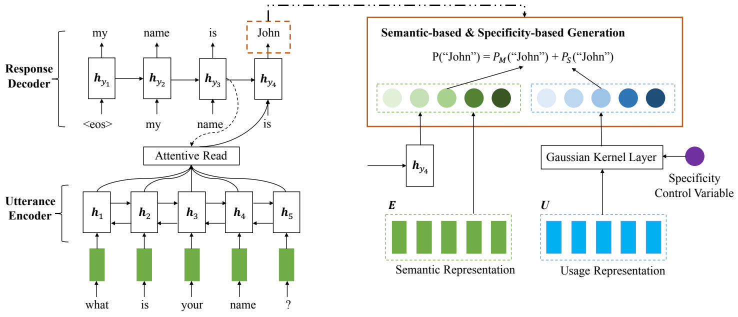
# Key Idea

- introduce an explicit specificity control variable ***s*** to represent the response purpose
  - *s* summarizes many latent factors into one variable
  - *s* has explicit meaning on specificity
  - *s* actively controls the generation of the response



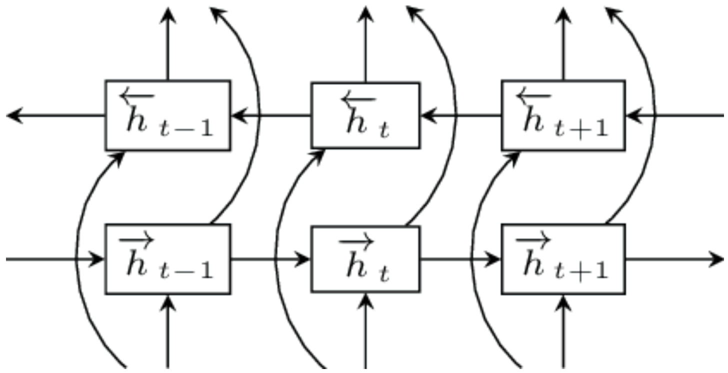
# Model Architecture

- the specificity control variable *s* is introduced into the Seq2Seq model
- single model -> multiple model
  - different <utterance, response>, different *s*, different models
- word representation
  - semantic representation: relates to the semantic meaning
  - usage representation: relates to the usage preference



# Model - Encoder

- Bi-RNN: modeling the utterance from both forward and backward directions
  - $\{\boldsymbol{h}_1^{\rightarrow}, \ldots, \boldsymbol{h}_T^{\rightarrow}\}\ \{\boldsymbol{h}_T^{\leftarrow}, \ldots, \boldsymbol{h}_1^{\leftarrow}\}$
  - $\mathbf{h}_t = [\boldsymbol{h}_t^{\rightarrow}, \boldsymbol{h}_{T-t+1}^{\leftarrow}]$



# Model - Decoder

- predict target word based on a mixture of two probabilities: the semantic-based and specificity-based generation probability

$$p(y_t) = \beta p_M(y_t) + \gamma p_S(y_t)$$

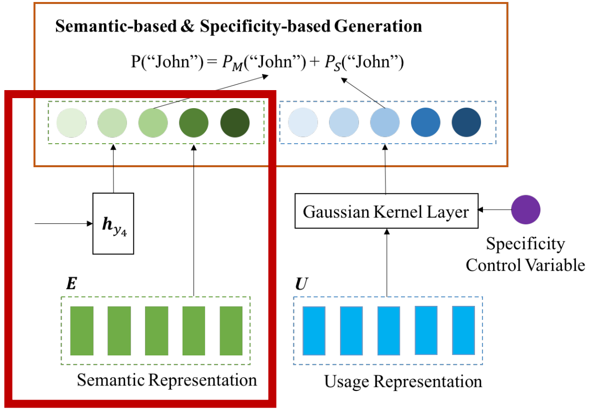
  - semantic-based probability

    \- decides what to say next given the input

$$p_M(y_t = w) = \boldsymbol{w}^T\left(\boldsymbol{W}_M^h \cdot \boldsymbol{h}_{y_t} + \boldsymbol{W}_{\boldsymbol{M}}^{\boldsymbol{e}} \cdot \boldsymbol{e}_{t-1} + \boldsymbol{b}_M\right)$$

hidden state

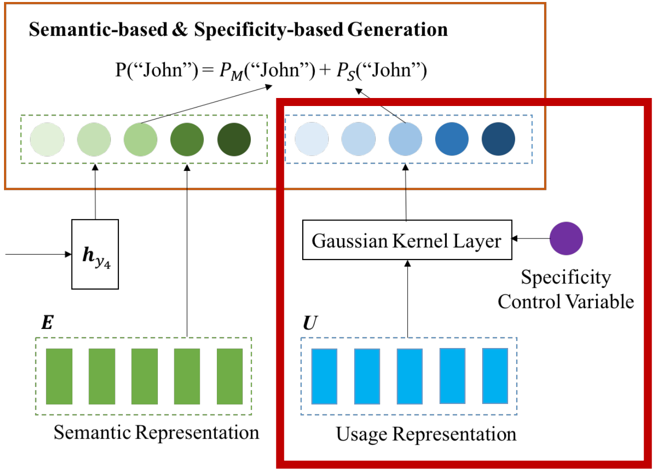
semantic representation



# Model - Decoder

➢ specificity-based probability

- decides how specific we should reply




- Gaussian Kernel layer
  - ✓ the specificity control variable interacts with the usage representation of words through the layer
  - ✓ let the word usage representation regress to the variable $s$ through certain mapping function (sigmoid)
- specificity control variable $s \in [0,1]$
  - ✓ 0 denotes the most general response
  - ✓ 1 denotes the most specific response

$$p_S(y_t = w) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(\Psi_S(\boldsymbol{U},\boldsymbol{w}) - s)^2}{2\sigma^2}\right)$$

$$\Psi_S(\boldsymbol{U},\boldsymbol{w}) = \sigma(\boldsymbol{w}^T(\boldsymbol{U} \cdot \boldsymbol{W}_U + \boldsymbol{b}_U))$$

usage representation

variance

# Model Training

- Objective function – log likelihood

$$\mathcal{L} = \sum_{(\boldsymbol{X},\boldsymbol{Y})\epsilon\mathcal{D}} \log P(\boldsymbol{Y}|\boldsymbol{X}, s; \theta)$$

- Training data: triples $(\boldsymbol{X}, \boldsymbol{Y}, s)$
- $s$ is not directly available in the raw conversation corpus

How to obtain $s$ to learn our model?

We propose to acquire distant labels for $s$

# Distant Supervision

- Normalized Inverse Response Frequency (NIRF)
  - a response is more general if it corresponds to more input utterances
  - the Inverse Response Frequency (IRF) in a conversation corpus

$$\mathrm{IRF}_{\mathbf{Y}} = \log(1 + |\mathcal{R}|)/f_{\mathbf{Y}}$$

$$\mathrm{NIRF}_{\mathbf{Y}} = \frac{\mathrm{IRF}_{\mathbf{Y}} - \min_{\mathbf{Y}' \in \mathcal{R}}(\mathrm{IRF}_{\mathbf{Y}'})}{\max_{\mathbf{Y}' \in \mathcal{R}}(\mathrm{IRF}_{\mathbf{Y}'}) - \min_{\mathbf{Y}' \in \mathcal{R}}(\mathrm{IRF}_{\mathbf{Y}'})}$$

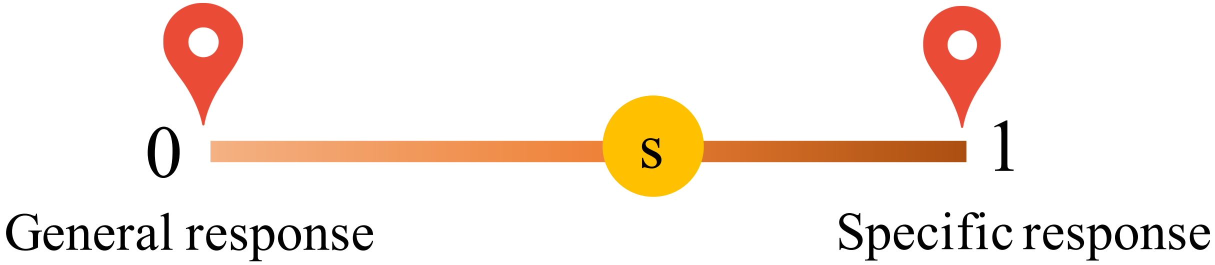
- Normalized Inverse Word Frequency (NIWF)
  - a response is more specific if it contains more specific words
  - the maximum of the Inverse Word Frequency (IWF) of all the words in a response

$$\mathrm{IWF}_{y} = \log(1 + |\mathcal{R}|)/f_{y}$$

$$\mathrm{IWF}_{\mathbf{Y}} = \max_{y \in \mathbf{Y}}(\mathrm{IWF}_{y})$$

# Specificity Controlled Response Generation

- Given a new input utterance, we can generate responses at different specificity levels by varying the control variable $s$
- Different s, different models, different responses
  - $s = 1$: the most informative response
  - $s \in [0,1]$: more dynamic , enrich the styles in the response
  - $s = 0$: the most general response



# Experiments - Dataset

- Short Text Conversation (STC) dataset
  - released in NTCIR-13
  - a large repository of post-comment pairs from the Sina Weibo
  - 3.8 million post-comment pairs
  - Jieba Chinese word segmenter

| | |
|---|---|
| Utterance-response pairs | 3,788,571 |
| Utterance vocabulary #w | 120,930 |
| Response vocabulary #w | 524,791 |
| Utterance max #w | 38 |
| Utterance avg #w | 13 |
| Response max #w | 74 |
| Response avg #w | 10 |

# Experiments – Model Analysis

| | Models | distinct-1 | distinct-2 | BLEU-1 | BLEU-2 | Average | Extrema |
|---|---|---|---|---|---|---|---|
| SC-Seq2Seq$_{NIRF}$ | $s = 1$ | 5258/0.064 | 16195/0.269 | 15.109 | 7.023 | 0.578 | 0.380 |
| | $s = 0.8$ | 5337/0.065 | 16105/0.271 | 15.112 | 7.003 | 0.578 | 0.381 |
| | $s = 0.5$ | 5318/0.065 | 16183/0.269 | 15.054 | 7.001 | 0.578 | 0.380 |
| | $s = 0.2$ | 5323/0.065 | 16087/0.270 | 15.168 | 7.032 | 0.580 | 0.380 |
| | $s = 0$ | 5397/0.066 | 16319/0.271 | 15.093 | 7.011 | 0.577 | 0.380 |
| SC-Seq2Seq$_{NIWF}$ | $s = 1$ | **11588/0.116** | **27144/0.347** | 12.392 | 5.869 | 0.554 | 0.353 |
| | $s = 0.8$ | 6006/0.051 | 17843/0.257 | 11.492 | 5.703 | 0.553 | 0.350 |
| | $s = 0.5$ | 2835/0.050 | 9537/0.235 | **16.122** | **7.674** | **0.609** | **0.399** |
| | $s = 0.2$ | 1534/0.048 | 5117/0.218 | 8.313 | 4.058 | 0.542 | 0.335 |
| | $s = 0$ | 1038/0.046 | 3154/0.211 | 4.417 | 3.283 | 0.549 | 0.334 |

Table 2: Model analysis of our SC-Seq2Seq under the automatic evaluation.

1. We vary the control variable s by setting it to five different values (i.e., 0, 0.2, 0.5, 0.8, 1)
2. NIWF (word-based) is a good distant label for the response specificity

# Experiments – Model Analysis

| | Models | distinct-1 | distinct-2 | BLEU-1 | BLEU-2 | Average | Extrema |
|---|---|---|---|---|---|---|---|
| $\text{SC-Seq2Seq}_{\text{NIRF}}$ | $s = 1$ | 5258/0.064 | 16195/0.269 | 15.109 | 7.023 | 0.578 | 0.380 |
| | $s = 0.8$ | 5337/0.065 | 16105/0.271 | 15.112 | 7.003 | 0.578 | 0.381 |
| | $s = 0.5$ | 5318/0.065 | 16183/0.269 | 15.054 | 7.001 | 0.578 | 0.380 |
| | $s = 0.2$ | 5323/0.065 | 16087/0.270 | 15.168 | 7.032 | 0.580 | 0.380 |
| | $s = 0$ | 5397/0.066 | 16319/0.271 | 15.093 | 7.011 | 0.577 | 0.380 |
| $\text{SC-Seq2Seq}_{\text{NIWF}}$ | $s = 1$ | **11588/0.116** | **27144/0.347** | 12.392 | 5.869 | 0.554 | 0.353 |
| | $s = 0.8$ | 6006/0.051 | 17843/0.257 | 11.492 | 5.703 | 0.553 | 0.350 |
| | $s = 0.5$ | 2835/0.050 | 9537/0.235 | **16.122** | **7.674** | **0.609** | **0.399** |
| | $s = 0.2$ | 1534/0.048 | 5117/0.218 | 8.313 | 4.058 | 0.542 | 0.335 |
| | $s = 0$ | 1038/0.046 | 3154/0.211 | 4.417 | 3.283 | 0.549 | 0.334 |

specific

general

Table 2: Model analysis of our SC-Seq2Seq under the automatic evaluation.

1. Varying the variable s from 0 to 1, the generated responses turn from general to specific
2. Different s -> different models -> different focus

# Experiments – Comparisons

| Models | distinct-1 | distinct-2 | BLEU-1 | BLEU-2 | Average | Extrema |
|---|---|---|---|---|---|---|
| Seq2Seq-att | 5048/0.060 | 15976/0.168 | 15.062 | 6.964 | 0.575 | 0.376 |
| MMI-bidi | 5074/0.082 | 12162/0.287 | 15.772 | 7.215 | 0.586 | 0.381 |
| MARM | 2566/0.096 | 3294/0.312 | 7.321 | 3.774 | 0.512 | 0.336 |
| Seq2Seq+IDF | 4722/0.052 | 15384/0.229 | 14.423 | 6.743 | 0.572 | 0.369 |
| SC-Seq2Seq$_{NIWF,s=1}$ | **11588/0.116** | **27144/0.347** | 12.392 | 5.869 | 0.554 | 0.353 |
| SC-Seq2Seq$_{NIWF,s=0.5}$ | 2835/0.050 | 9537/0.235 | **16.122** | **7.674** | **0.609** | **0.399** |

Table 3: Comparisons between our SC-Seq2Seq and the baselines under the automatic evaluation.

When $s = 1$, our SC-Seq2Seq$_{\mathrm{NIWF}}$ model can achieve the best specificity performance

# Experiments – Comparisons

| Models | distinct-1 | distinct-2 | BLEU-1 | BLEU-2 | Average | Extrema |
|---|---|---|---|---|---|---|
| Seq2Seq-att | 5048/0.060 | 15976/0.168 | 15.062 | 6.964 | 0.575 | 0.376 |
| MMI-bidi | 5074/0.082 | 12162/0.287 | 15.772 | 7.215 | 0.586 | 0.381 |
| MARM | 2566/0.096 | 3294/0.312 | 7.321 | 3.774 | 0.512 | 0.336 |
| Seq2Seq+IDF | 4722/0.052 | 15384/0.229 | 14.423 | 6.743 | 0.572 | 0.369 |
| SC-Seq2Seq$_{NIWF,s=1}$ | **11588/0.116** | **27144/0.347** | 12.392 | 5.869 | 0.554 | 0.353 |
| SC-Seq2Seq$_{NIWF,s=0.5}$ | 2835/0.050 | 9537/0.235 | **16.122** | **7.674** | **0.609** | **0.399** |

Table 3: Comparisons between our SC-Seq2Seq and the baselines under the automatic evaluation.

1. our SC-Seq2Seq$_{NIWF}$ model can best fit the ground truth data
2. there are diverse responses in real data in terms of specificity

# Experiments – Comparisons

| | +2 | +1 | +0 | kappa |
|---|---|---|---|---|
| Seq2Seq-att | 29.32% | 25.27% | 45.41% | 0.448 |
| MMI-bidi | 30.40% | 24.85% | 44.75% | 0.471 |
| MARM | 20.11% | 27.96% | 51.93% | 0.404 |
| Seq2Seq+IDF | 28.81% | 23.87% | 47.33% | 0.418 |
| SC-Seq2Seq$_{\text{NIWF},s=1}$ | 42.47% | 14.29% | 43.24% | 0.507 |
| SC-Seq2Seq$_{\text{NIWF},s=0.5}$ | 20.62% | 40.16% | 39.22% | 0.451 |
| SC-Seq2Seq$_{\text{NIWF},s=0}$ | 14.34% | 46.38% | 39.28% | 0.526 |

Table 4: Results on the human evaluation.

1. SC-Seq2Seq$_{\text{NIWF},s=1}$ generates the most informative responses and interesting and the least general responses than all the baseline models
2. The largest kappa value is achieved by SC-Seq2SeqNIWF,s=0

# Experiments – Case study

**Utterance**: 哎哟! 被发现了! 汤唯有必要那么美吗? Oh! Got caught! Why is Tang Wei so beautiful?

**Seq2Seq-att**: 我超喜欢她！ I really like her!

**MMI-bidi**: 好美。我超喜欢她！ So beautiful. I really like her!

**MARM**: 我不喜欢汤唯。I don't like Tang Wei.

**Seq2Seq+IDF**: 我超喜欢她的，我喜欢。I really like her, I like.

**s=1**: 我超喜欢她的眼妆，她的气质很美！ I really like her eye make-up, and her temperament is so beautiful!

**s=0.8**: 我超喜欢她的，她的眼睛很雪亮！ I really like her, and her eyes are so bright!

**s=0.5**: 我觉得她很美，很喜欢她！ I think she is very beautiful, and I really like her!

**s=0.2**: 我觉得她很美！ I think she is very beautiful.

**s=0**: 是的。Yes.

**Utterance**: 将来也要养一只这么乖的的萨摩，让它陪着孩子一起长大。I will keep such a sweet samoyed, and let it accompany my children to grow up.

**Seq2Seq-att**: 这不是萨摩吗? Is it a samoyed?

**MMI-bidi**: 好可爱的萨摩。What a lovely samoyed.

**MARM**: 好可爱啊。So lovely.

**Seq2Seq+IDF**: 好可爱的狗狗。What a lovely dog.

**s=1**: 萨摩耶好可爱，真想养一只! The samoyed is so lovely, and I really want to keep one!

**s=0.8**: 萨摩好可爱，好想掐掐。The samoyed is so lovely, and I really want to pinch it.

**s=0.5**: 好可爱的狗狗，好可爱的狗狗。What a lovely dog, what a lovely dog.

**s=0.2**: 好可爱！ 好可爱！ So lovely, so lovely!

**s=0**: 好可爱！ So lovely!

The responses generated by the four baselines are often quite general and short

# Experiments – Case study

**Utterance**: 哎哟! 被发现了! 汤唯有必要那么美吗? Oh! Got caught! Why is Tang Wei so beautiful?

**Seq2Seq-att**: 我超喜欢她！ I really like her!

**MMI-bidi**: 好美。我超喜欢她！ So beautiful. I really like her!

**MARM**: 我不喜欢汤唯。I don't like Tang Wei.

**Seq2Seq+IDF**: 我超喜欢她的，我喜欢。I really like her, I like.

**s=1**: 我超喜欢她的眼妆，她的气质很美！ I really like her eye make-up, and her temperament is so beautiful!

**s=0.8**: 我超喜欢她的，她的眼睛很雪亮！ I really like her, and her eyes are so bright!

**s=0.5**: 我觉得她很美，很喜欢她！ I think she is very beautiful, and I really like her!

**s=0.2**: 我觉得她很美！ I think she is very beautiful.

**s=0**: 是的。Yes.

**Utterance**: 将来也要养一只这么乖的的萨摩，让它陪着孩子一起长大。I will keep such a sweet samoyed, and let it accompany my children to grow up.

**Seq2Seq-att**: 这不是萨摩吗? Is it a samoyed?

**MMI-bidi**: 好可爱的萨摩。What a lovely samoyed.

**MARM**: 好可爱啊。So lovely.

**Seq2Seq+IDF**: 好可爱的狗狗。What a lovely dog.

**s=1**: 萨摩耶好可爱，真想养一只! The samoyed is so lovely, and I really want to keep one!

**s=0.8**: 萨摩好可爱，好想掐掐。The samoyed is so lovely, and I really want to pinch it.

**s=0.5**: 好可爱的狗狗，好可爱的狗狗。What a lovely dog, what a lovely dog.

**s=0.2**: 好可爱！ 好可爱！ So lovely, so lovely!

**s=0**: 好可爱！ So lovely!

With s from 1 to 0, SC-Seq2Seq$_{NIWF}$ can generate very long and specific responses, to more general and shorter responses.

# Experiments – Analysis

| 爸爸(dad) | | 水果(fruits) | | 脂肪肝(fatty liver) | | 单反相机(DSLR) | |
|---|---|---|---|---|---|---|---|
| Usage | Semantic | Usage | Semantic | Usage | Semantic | Usage | Semantic |
| 更好(better) | 妈妈(mother) | 尝试(attempt) | 蔬菜(vegetables) | 坐久(outsit) | 胖(fat) | 亚洲杯(Asian Cup) | 照相机(camera) |
| 睡觉(sleep) | 哥哥(brother) | 诱惑(tempt) | 牛奶(milk) | 素食主义(vegetarian) | 减肥(diet) | 读取(read) | 摄影(photography) |
| 快乐(happy) | 老公(husband) | 表现(express) | 西瓜(watermelon) | 散步(walk) | 高血压(hypertension) | 半球(hemispherical) | 镜头(shot) |
| 无聊(boring) | 爷爷(grandfather) | 拥有(own) | 米饭(rice) | 因果关系(causality) | 亚健康(sub-health) | 防辐射(anti-radiation) | 影楼(studio) |
| 电影(movie) | 姑娘(girl) | 梦想(dream) | 巧克力(chocolate) | 哑铃(dumbbell) | 呕吐(emesis) | 无人机(UAV) | 写真(image) |

Table 6: Target words and their top-5 similar words under usage and semantic representations respectively.


(a) usage (b) semantic

Figure 3: t-SNE embeddings of usage and semantic vectors.

1. Neighbors based on semantic representations are semantically related
2. Neighbors based on usage representations are not so related but with similar specificity levels

# Conclusion

- We argue
  - employing a single model to learn the mapping between the utterance and response will inevitably favor general responses
- We propose
  - an explicit specificity control variable is introduced into the Seq2Seq model handle different utterance-response relationships in terms of specificity
- Future work
  - employ some reinforcement learning technique to learn to adjust the control variable depending on users' feedbacks
  - apply to other tasks, like summarization, QA, etc

# Thanks Q & A

- Name: Ruqing Zhang | Email: zhangruqing@software.ict.ac.cn