

Learning to Control the Specificity in Neural Response Generation

ACL 2018

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

University of Chinese Academy of Sciences, Beijing, China

CAS Key Lab of Network Data Science and Technology, Institute of Computing Technology

Presenter: Yang Zhao

AI Lab NLP Centre

Outline

- Introduction
- Related works
- Specificity Controlled Seq2Seq Model
- Data and Baselines
- Results and Analysis
- Conclusion and Takeaways


Outline

- Introduction
- Related works
- Specificity Controlled Seq2Seq Model
- Data and Baselines
- Results and Analysis
- Conclusion

Introduction

- A widely used approach to chit-chat dialog is learning a generative conversational model (Seq2Seq and its variants) from social conversation data
- Seq2seq originally for machine translation (1-to-1, semantically equivalent)
- Dialog could be **many-to-1**

Examples of Many-to-1 in Dialog

- X_1 : How is the weather?
 - X_2 : How is the dress?
 - X_3 : We are going to have a trip tomorrow!
- 
- Y: It's good.

*These responses are safe but **boring***

Related Works to Generic Response

- As an early work, (Li+, 2016a) used Maximum Mutual Information (MMI) to penalize general responses
- A persona-based neural conversation model later introduced by (Li+, 2016b)
- (Yao+, 2016) improve the specificity with RL by using the averaged IDF score of the words in the response as a reward
- (Shen+, 2017) presented a CVAE framework for generating specific responses based on specific attributes

*Different from above work, the presented work introduce an **explicit** specificity control variable into Seq2Seq model*

Outline

- Introduction
- Related works
- **Specificity Controlled Seq2Seq Model**
- Data and Baselines
- Results and Analysis
- Conclusion

Problem Statement

$$\mathbf{X} = (x_1, x_2, \dots, x_T)$$

$$\mathbf{Y} = (y_1, y_2, \dots, y_{T'})$$



$p(\mathbf{Y}|\mathbf{X}, s)$ over the corpus D

- introduce an explicit variable s
 - s would have explicit meaning on specificity
 - s could control the generation of the response \mathbf{Y} given the input utterance \mathbf{X}

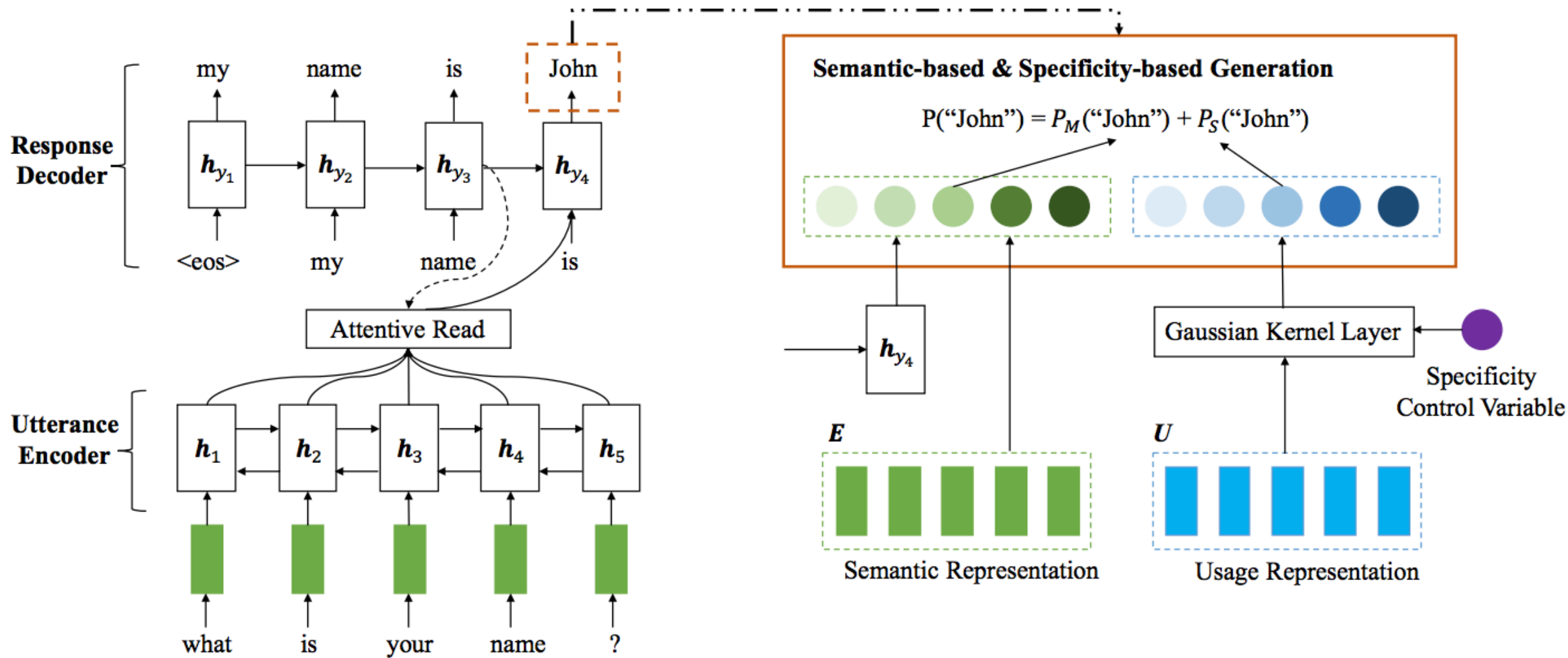
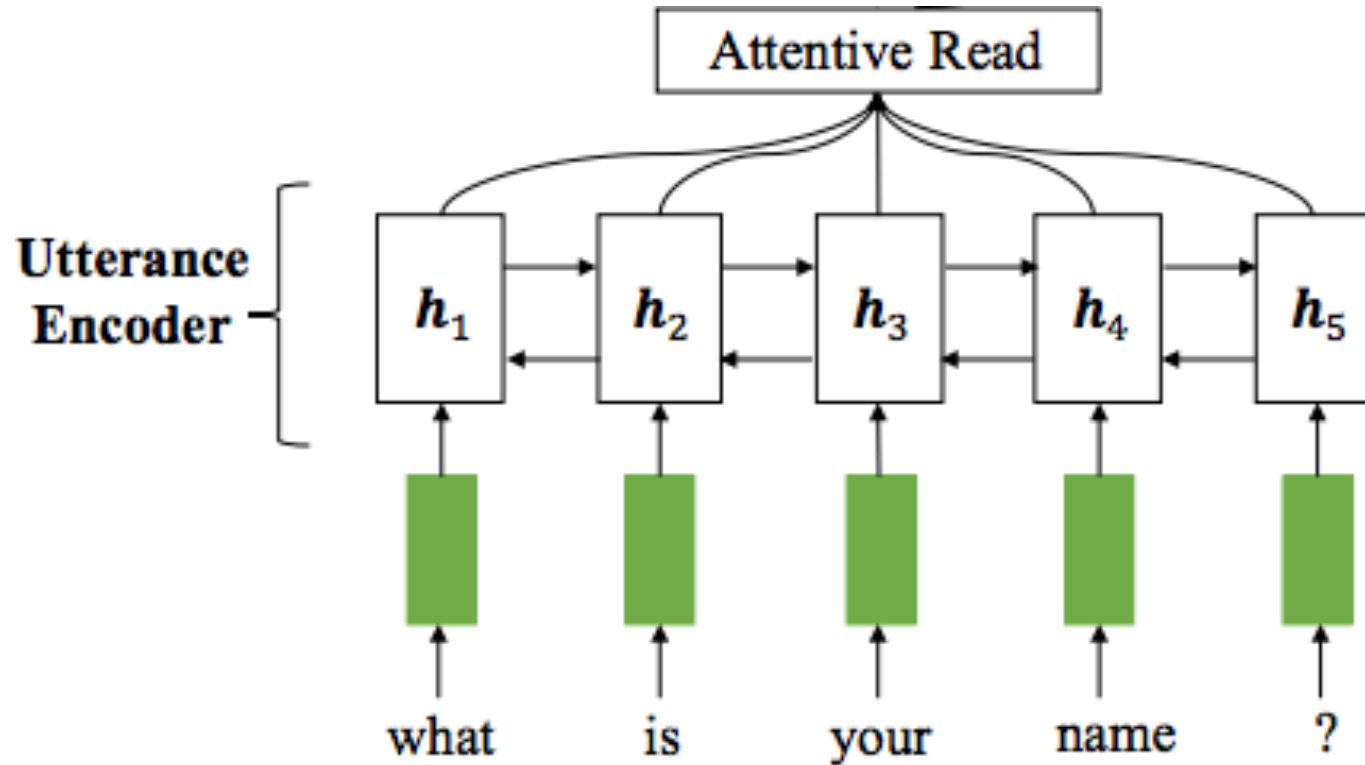


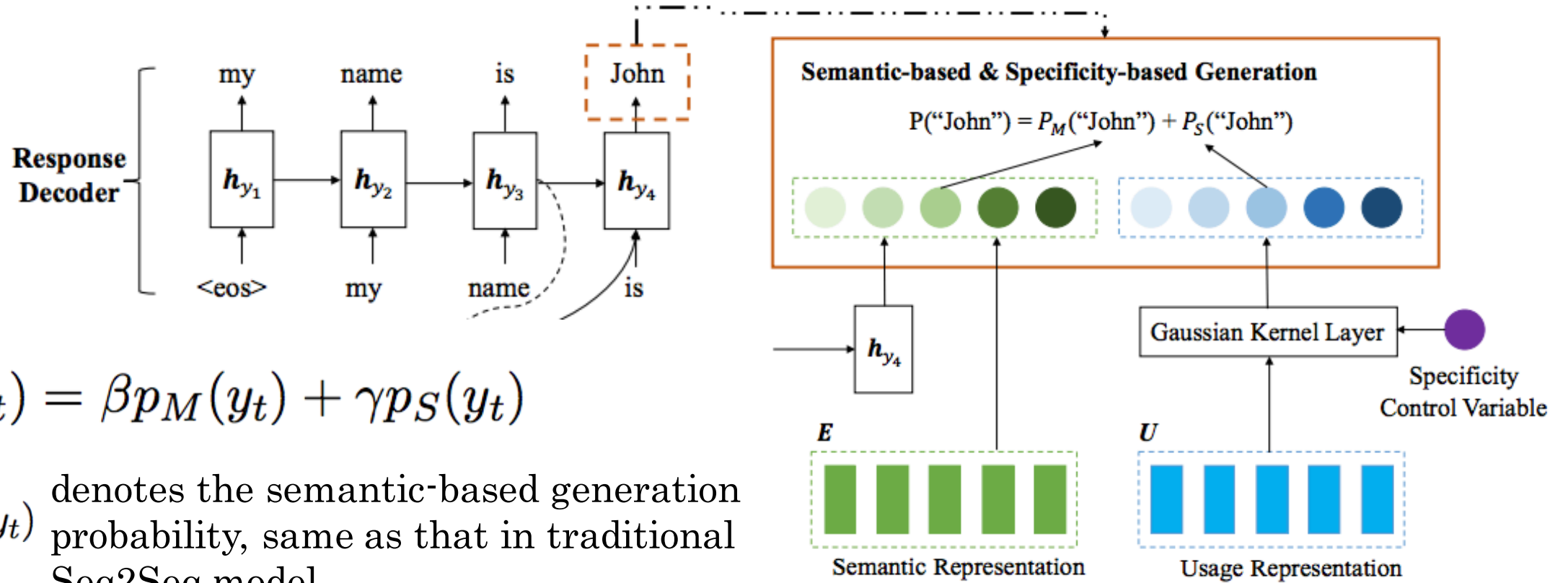
Figure 2: The overall architecture of SC-Seq2Seq model.

Encoder



Bi-directional GRU-RNN (Choi, 2014)

Decoder

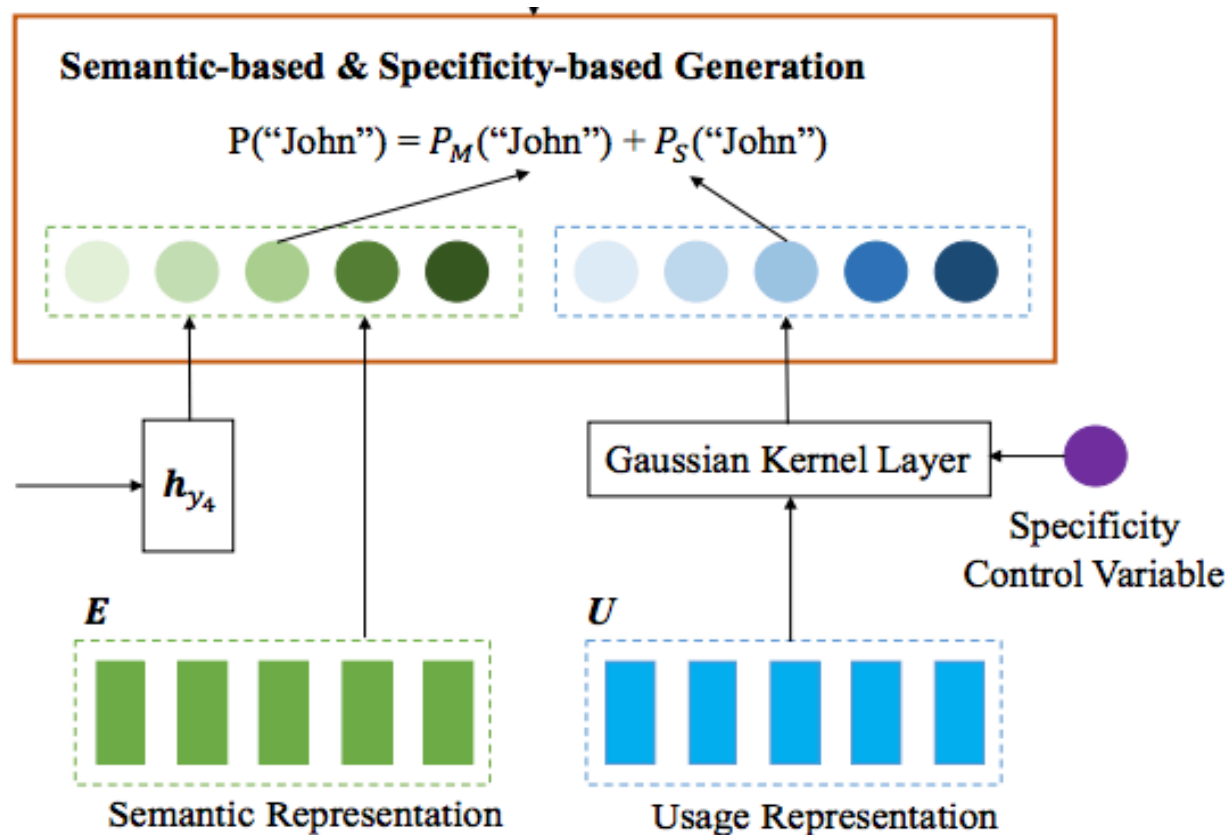


$$p(y_t) = \beta p_M(y_t) + \gamma p_S(y_t)$$

$p_M(y_t)$ denotes the semantic-based generation probability, same as that in traditional Seq2Seq model

$p_S(y_t)$ denotes the generation probability of the target word given the specificity control variable s

Decoder - $p_M(y_t)$

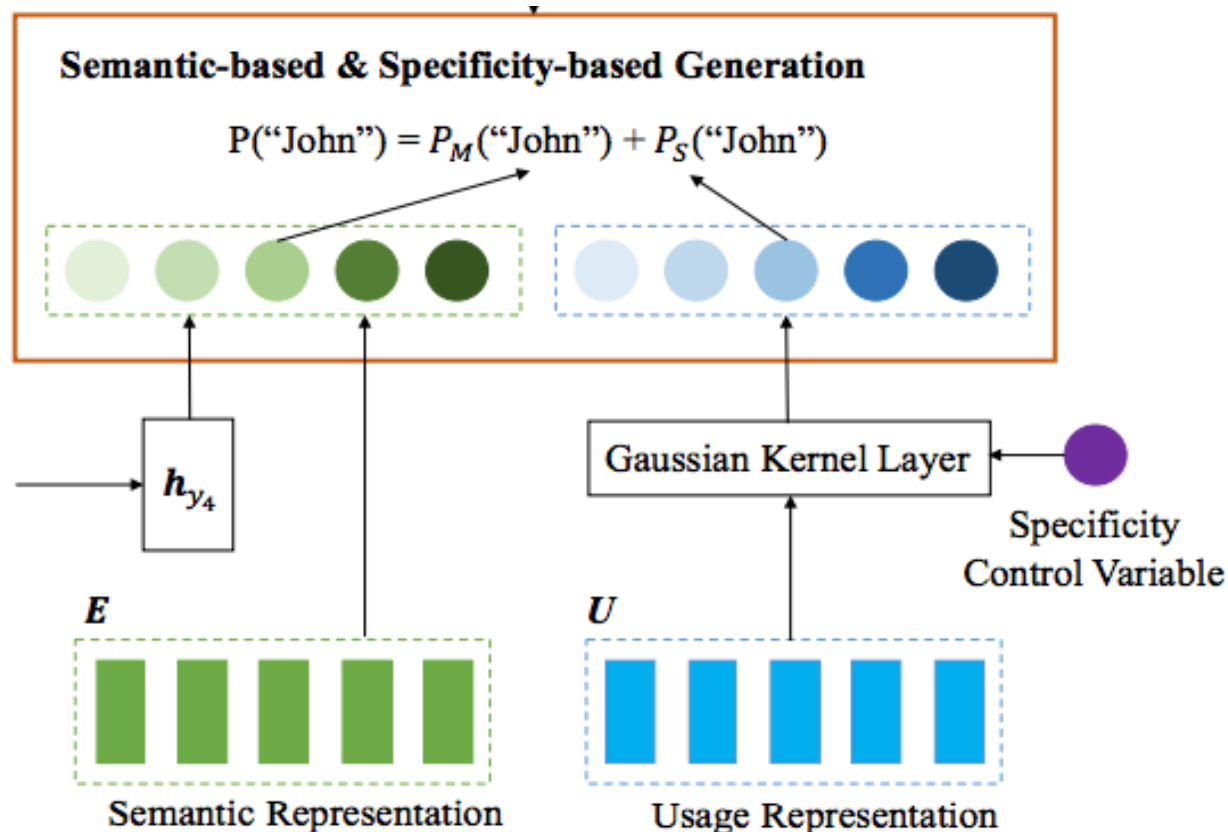


$$p_M(y_t = w) = \mathbf{w}^T (\mathbf{W}_M^h \cdot \mathbf{h}_{y_t} + \mathbf{W}_M^e \cdot \mathbf{e}_{t-1} + \mathbf{b}_M)$$

where \mathbf{w} is a one-hot indicator vector of the word w and \mathbf{e}_{t-1} is the semantic representation of the $t - 1$ -th generated word in decoder. \mathbf{W}_M^h , \mathbf{W}_M^e and \mathbf{b}_M are parameters. \mathbf{h}_{y_t} is the t -th hidden state in the decoder which is computed by:

$$\mathbf{h}_{y_t} = f(y_{t-1}, \mathbf{h}_{y_{t-1}}, \mathbf{c}_t)$$

Decoder - $p_S(y_t)$



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

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

where σ^2 is the variance, and $\Psi_S(\cdot)$ maps the word usage representation into a real value with the specificity control variable s as the mean of the Gaussian distribution. \mathbf{W}_U and \mathbf{b}_U are parameters to be learned.

Distant Supervision for specificity control variable s

1. Normalized Inverse Word Frequency
2. Normalized Inverse Response Frequency

Normalized Inverse Word Frequency

- Assumption - *the sentence is more specific if it contains more specific words.*
- for a word y in the response \mathbf{Y}

$$\text{IWF}_y = \log(1 + |\mathcal{R}|) / f_y$$

where f_y denotes the number of responses in \mathcal{R} containing the word y . where $|\mathcal{R}|$ denotes the size of the response collection \mathcal{R}

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

we find that the best performance can be achieved by using the maximum of the IWF of all the words in \mathbf{Y}

Normalized Inverse Response Frequency

Assumption - a response is more general if it corresponds to more input utterances in the corpus.

- for a word y in the response Y

$\text{IRF}_Y = \log(1 + |\mathcal{R}|)/f_Y$ For a response $Y \in \mathcal{R}$, let f_Y denote its corpus frequency in \mathcal{R} where \mathcal{R} are all the responses in corpus

$\text{NIRF}_Y = \frac{\text{IRF}_Y - \min_{Y' \in \mathcal{R}}(\text{IRF}_{Y'})}{\max_{Y' \in \mathcal{R}}(\text{IRF}_{Y'}) - \min_{Y' \in \mathcal{R}}(\text{IRF}_{Y'})}$ where $\max(\text{IRFR})$ and $\min(\text{IRFR})$ denotes the maximal and minimum IRF value in \mathcal{R} respectively

Outline

- Introduction
- Related works
- Specificity Controlled Seq2Seq Model
- **Data and Baselines**
- Results and Analysis
- Conclusion

Data - Sina Weibo Short Text Conversation dataset

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

Table 1: Short Text Conversation (STC) data statistics: #w denotes the number of Chinese words.

- randomly selected two subsets as the development and test dataset, each containing 10k pairs. The left pairs are used for training.

Baseline

- the standard Seq2Seq model with the attention mechanism ([Bahdanau+, 2015](#))
- MMI-bidi: the Seq2Seq model using Maximum Mutual Information (MMI) as the objective function to reorder the generated responses ([Li+., 2016a](#))
- MARM: the Seq2Seq model with a probabilistic framework to model the latent responding mechanisms ([Zhou et al., 2017](#))
- Seq2Seq+IDF: an extension of Seq2Seq-att by optimizing specificity under the reinforcement learning framework, where the reward is calculated as the sentence level IDF score of the generated response ([Yao et al., 2016](#))

Evaluation

- To evaluate the specificity/diversity of the responses, distinct-1 & distinct-2 (Li et al., 2016a) that counts the numbers of distinct unigrams and bi-grams in the generated responses, and divide the numbers by total number of generated unigrams and bigrams
- BLEU (Papineni et al., 2002)
- **Average & Extrema** (Serban et al., 2017): Average and Extrema projects the generated response and the ground truth response into two separate vectors by taking the mean over the word embeddings or taking the extremum of each dimension respectively, and then computes the cosine similarity between them
- Human Evaluation - Annotators evaluate 300 random sampled test utterance (all inter-agreement scores are larger than 0.4)

Outline

- Introduction
- Related works
- Specificity Controlled Seq2Seq Model
- Data and Baselines
- **Results and Analysis**
- Conclusion

Results upon Automatic Evaluation

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

- (1) MMI-bidi can improve the specificity (in terms of distinct ratios) over the traditional Seq2Seq-att model
- (2) MARM can achieve the best distinct ratios among the baseline methods, but the worst in terms of the distinct numbers. The results indicate that MARM tends to generate specific but very short responses
- (3) Seq2seq+IDF (RL enhanced) does not show much advantages

Human Evaluation

	+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 _{NIWF,s=1}	42.47%	14.29%	43.24%	0.507
SC-Seq2Seq _{NIWF,s=0.5}	20.62%	40.16%	39.22%	0.451
SC-Seq2Seq _{NIWF,s=0}	14.34%	46.38%	39.28%	0.526

- (1) +2: *the response is not only semantically relevant and grammatical, but also informative and interesting*
- (2) +1: *the response is grammatical and can be used as a response to the utterance, but is too trivial (e.g., “I don’t know”)*
- (3) +0: *the response is semantically irrelevant or ungrammatical (e.g., grammatical errors or UNK)*

Cases

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!

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!

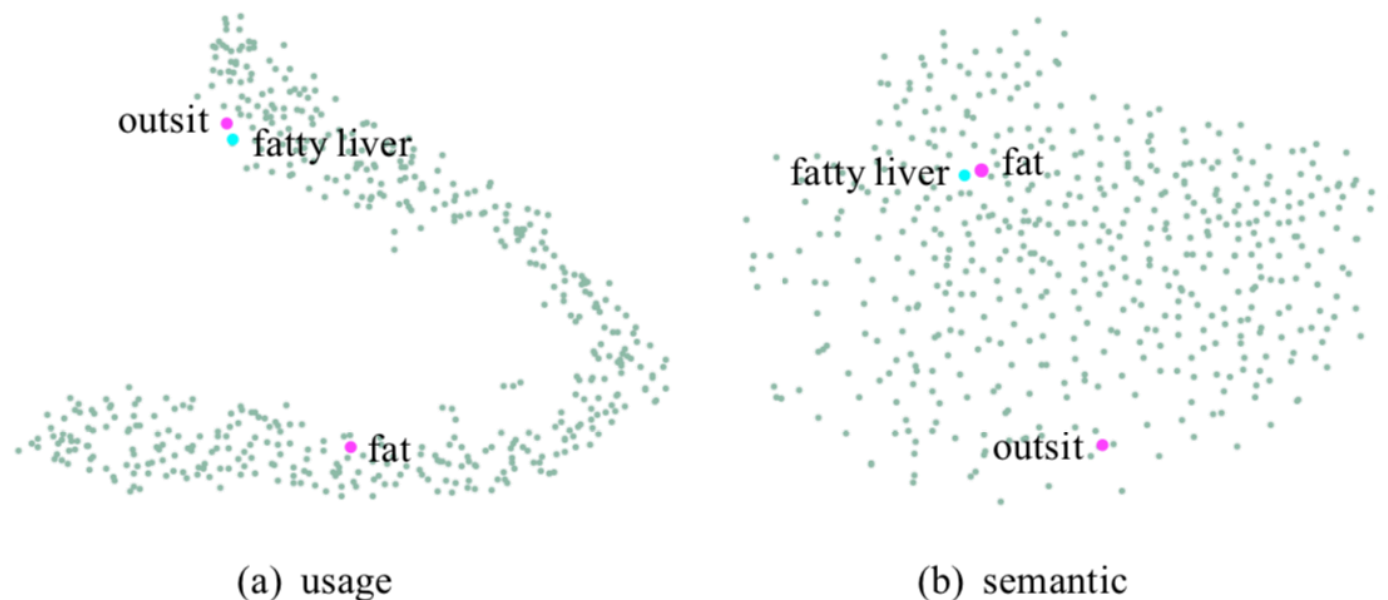
s=0.8: 萨摩好可爱，好想掐掐。 The samoyed is so lovely, and I really want to pinch it.

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!

Analysis on Usage Representations



- *Randomly sample 500 words from our SC-Seq2SeqNIWF and apply t -SNE to visualize both usage and semantic embeddings*

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

- In usage space, specific words “脂肪肝(fatty liver)” and “久坐(outsit)” closely, far from the general words like “胖(fat)”
- In semantic space, “脂肪肝(fatty liver)” is close to “胖(fat)” since they are semantically related

Analysis on Usage Representations


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)

- Neighbors based on semantic representations are semantically related, while neighbors based on usage representations are not so related but with similar specificity levels

Outline

- Introduction
- Related works
- Specificity Controlled Seq2Seq Model
- Data and Baselines
- Results and Analysis
- Conclusion

Conclusion and Takeaways from this work

- Seq2seq original for MT (1-to-1), could be inappropriate dialog system (Many-to-1 and 1-to-Many)
- X: balbalablablab Y: It's good.
- X: What are you doing? 
 - Y₁: I am having lunch.
 - Y₂: I am watching TV.
 - Y₃: I am hanging out with my friend.

Conclusion and Takeaways from this work

- Seq2seq original for MT (1-to-1), could be inappropriate dialog system (Many-to-1 and 1-to-Many)
- Learn specificity control variable S via distant supervision during training, and dynamic control S $[0, 1]$ to affect the response at different specificity levels
- Words could have multiple embeddings, reflecting different property (e.g. usage embedding)

Thanks