Neural Response Generation with Dynamic Vocabularies

Yu Wu
SKLSDE, Beihang University
wuyu@buaa.edu.cn

Wei Wu
Microsoft Research
wuwei@microsoft.com

Dejian Yang
SKLSDE, Beihang University
dejianyang@buaa.edu.cn

Can Xu
Microsoft Research
Can.xu@microsoft.com

Zhoujun Li
SKLSDE, Beihang University
lizj@buaa.edu.cn

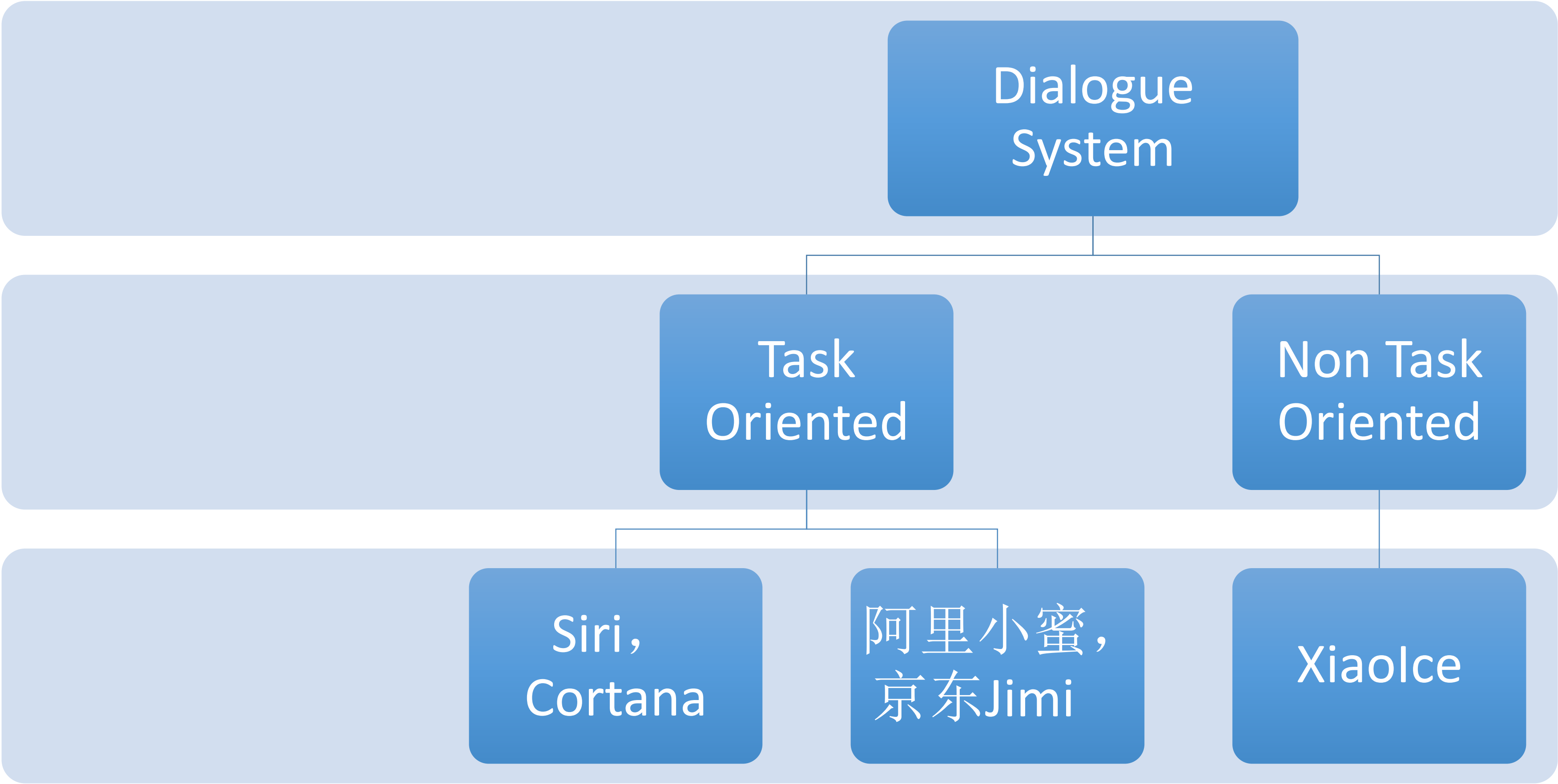
Outline

- Task, challenges, and ideas

- Our approach
  - Dynamic vocabulary for S2S learning

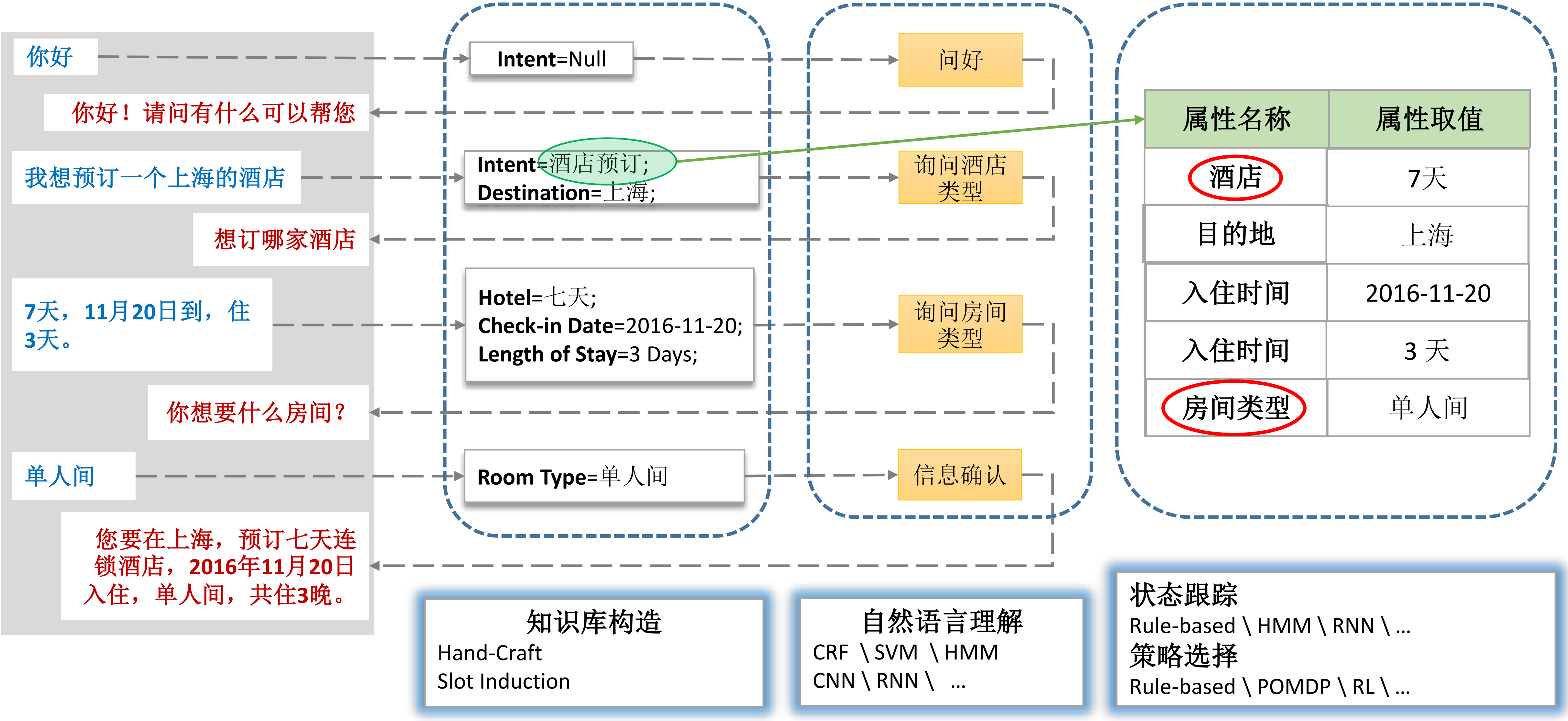
- Experiment
  - Experiment setup: data set and baseline methods
  - Evaluation and analysis

Taxonomy of dialogue systems



Chen, Hongshen, et al. "A Survey on Dialogue Systems: Recent Advances and New Frontiers."

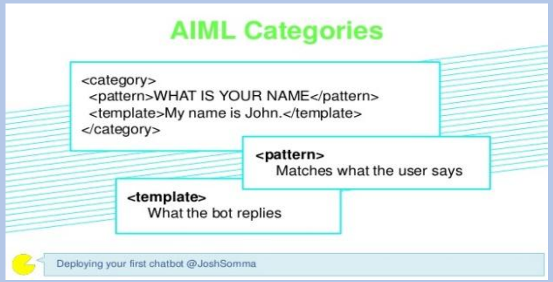
Task Oriented Chatbot



Genre of Chatbots

Templated based Chatbot

- Fill slots in a pre-defined sentence.



- Controllable, interpretable
- Low coverage

Retrieval based Chatbot

- Select proper responses from a pre-defined index.

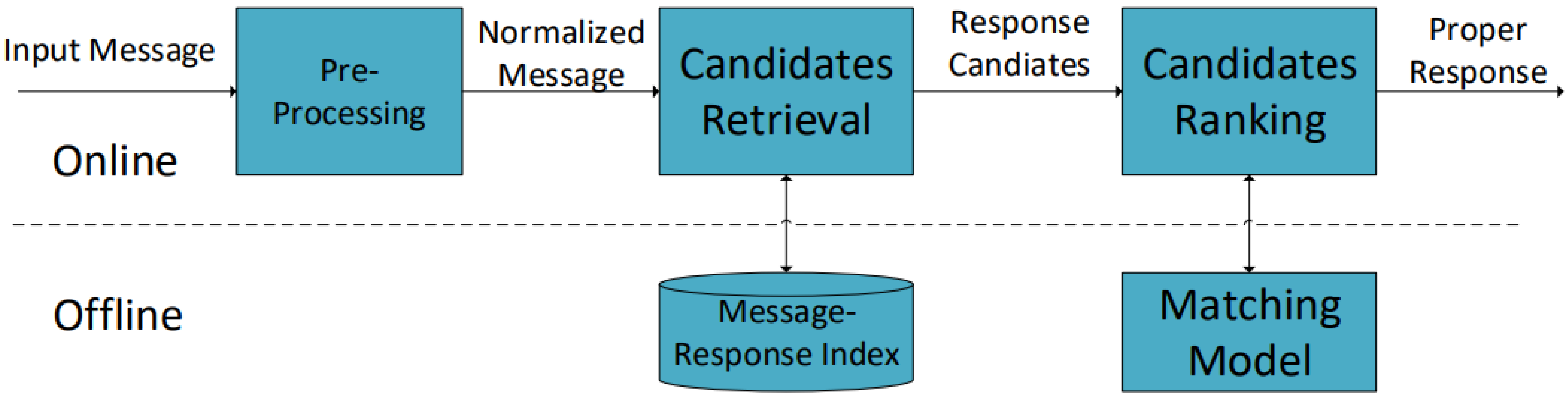
- Fluent, interesting and informative replies.
- Heavily rely on the index.

Generation based Chatbot

- Generate a relevant response to a history.

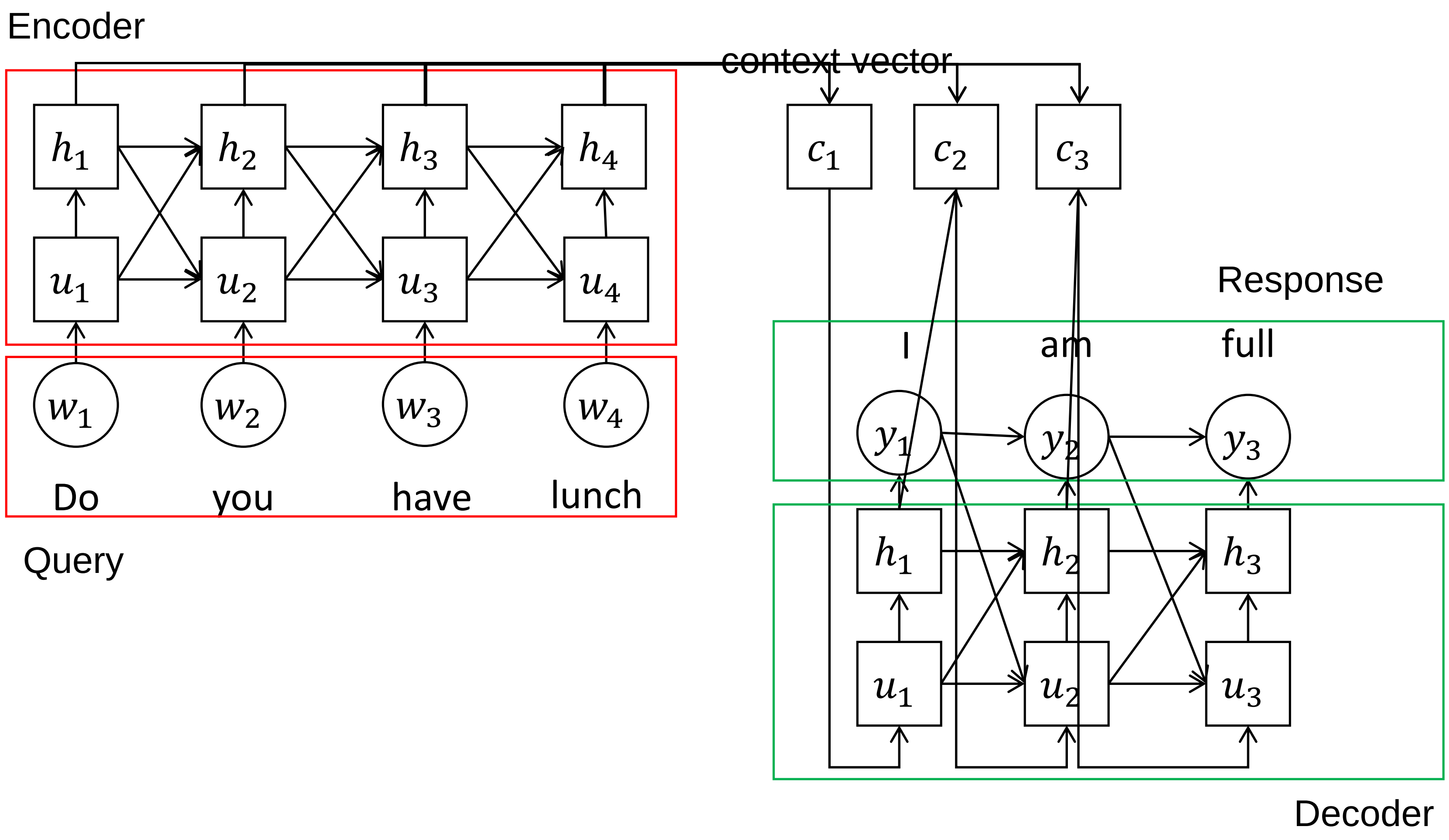
- Flexible, less human efforts.
- Ungrammatical, non-sense and general replied.

Pipeline of a retrieval based chatbot



Ji et al. An Information Retrieval Approach to Short Text Conversation

Sequence to Sequence Model for Chatbots



O Vinyals et al. *A Neural Conversation Model*

Retrieval v.s. Generation

Retrieval

- Pros
  - Diverse and fluent responses
  - Fluent responses
  - Flexible system
  - Easy to evaluate (L2R)
- Cons
  - Random responses
  - Bundled with query-response pairs
  - Difficult to be context-aware

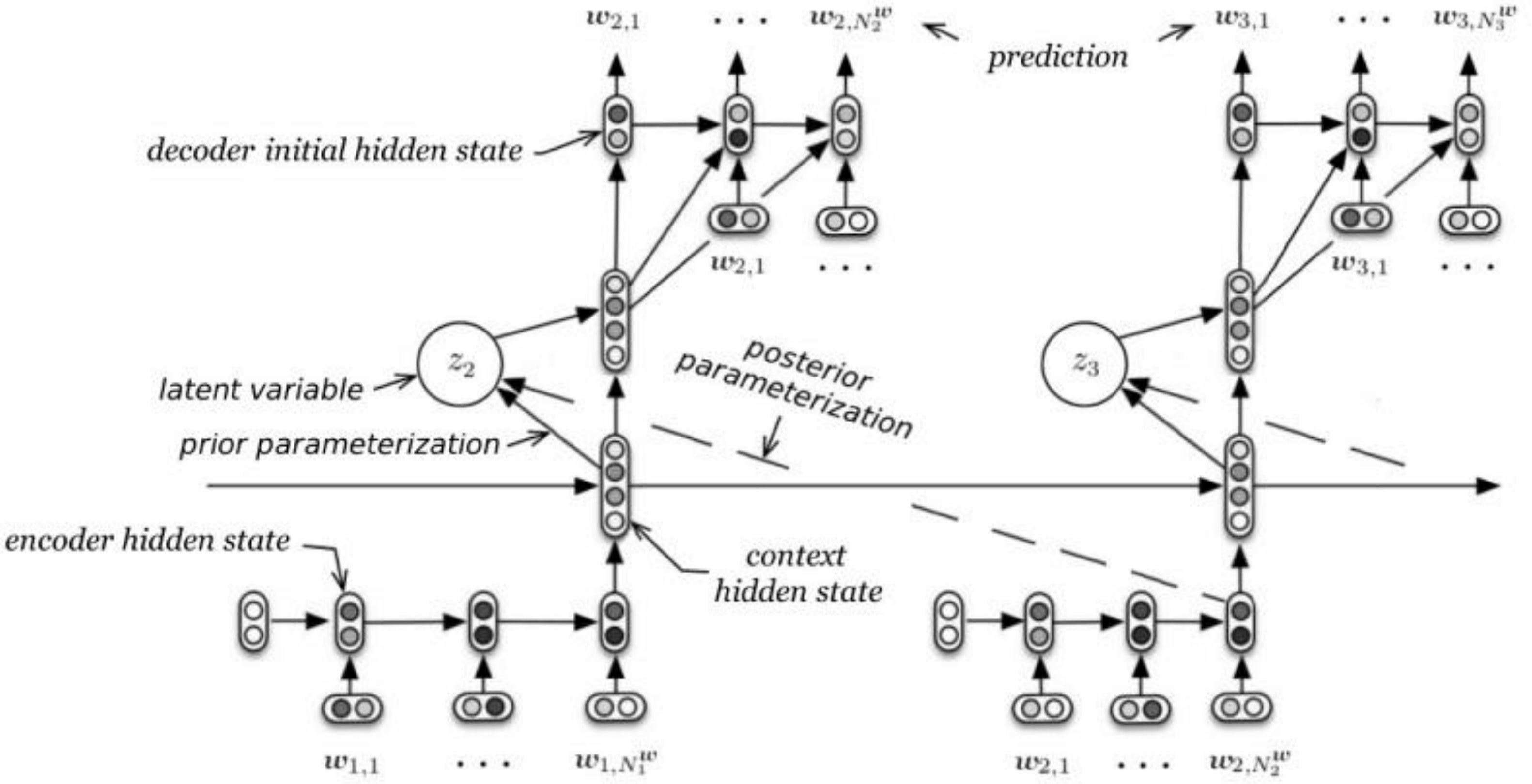
Generation

- Pros
  - End-to-end learning
  - Safe responses
  - Easy to be context-aware, emotional and controllable.
- Cons
  - Hard to evaluate
  - Boring and disfluent responses
  - Require experienced developers
  - UNK

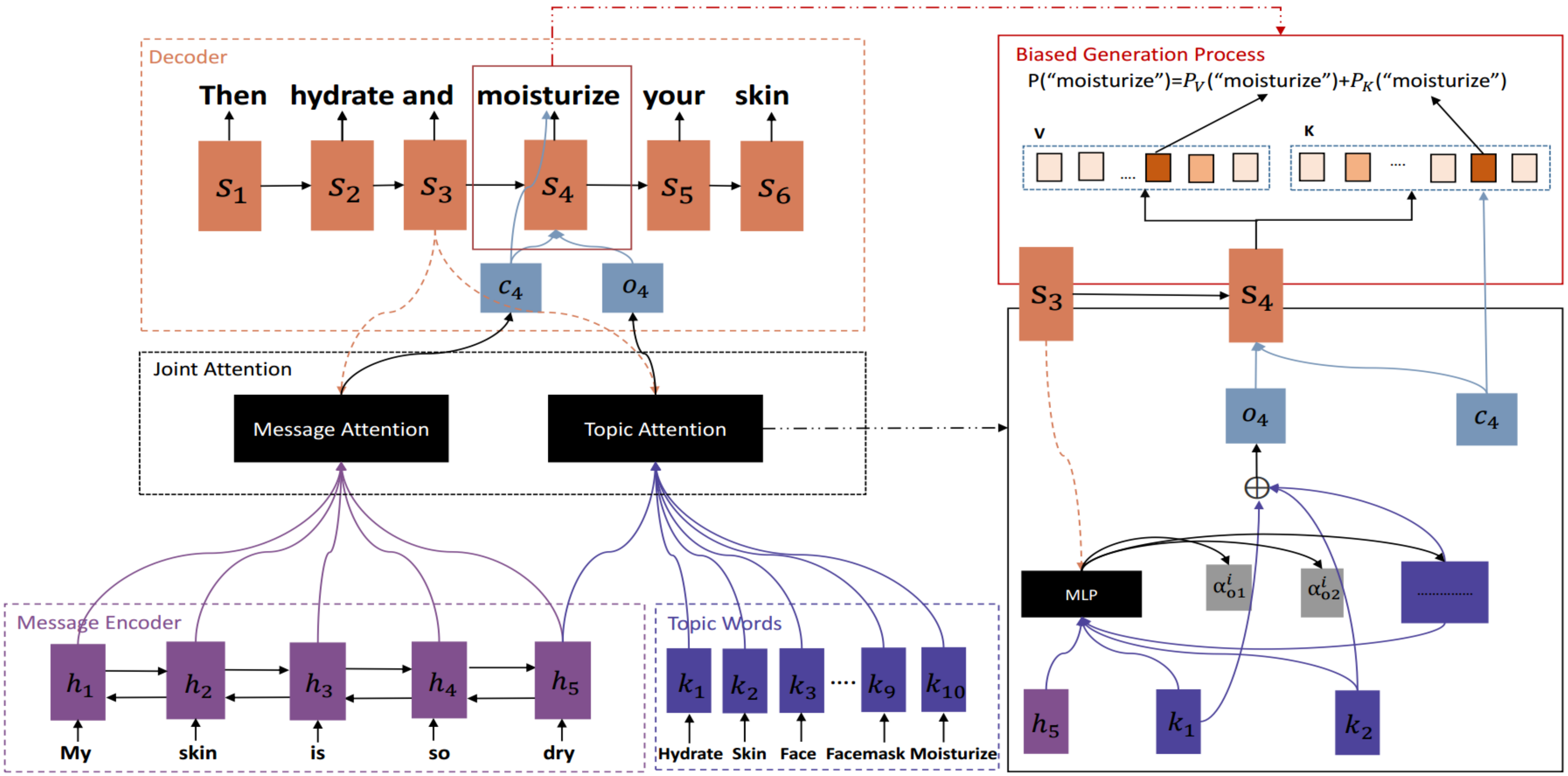
Challenges of Generative Chatbots

- The fluency problem
  - 你有多无聊-> 无聊的无聊 (how bored you are -> bored's bored)
- The "UNK" problem
  - Specific entities, low frequency words cannot be generated
- Boring responses / diversity
  - Easy to generate responses like "I do not know" "why", "haha" etc.
  - Especially bad on long queries

Existing Methods: CVAE (Serban et al. AAAI 2017, Zhao et al. ACL2017)



Existing Methods: complex models （Xing et al. AAAI 2017）



Existing Methods: Heavy rerank algorithms

Li et al. NAACL 2016:

$$\begin{aligned}\hat{T} &= \arg\max_T \big\{(1-\lambda)\log p(T|S) \\ &\qquad + \lambda \log p(S|T) - \lambda \log p(S)\big\} \\ &= \arg\max_T \big\{(1-\lambda)\log p(T|S) + \lambda \log p(S|T)\big\}\end{aligned}$$

Mei et al. AAAI 2017: LDA based reranking algorithm.

Existing Methods: Reinforcement algorithm

- Reinforcement Learning
  - Policy Gradient： Li et al. EMNLP 2016: Use $P(S|T) + \lambda P(T|S)$ as a reward
  - Value based network: 宋皓宇，张伟男，刘挺 基于DQN的开放域多轮对话策略学习

- GAN
  - SeqGAN ： Li et al. EMNLP 2017: GAN for response generation
  - Gan with an approximate embedding layer. Xu et al. EMNLP 2017

Intuition

- Only a small part of words are useful in the decoding.
  - Function words should be included.
    - Function words guarantee grammatical correctness and fluency of responses.
    - 的，了，我，你….
  - Words that are relevant to the context should be included.
    - Content words, on the other hand, express semantics of responses.
  - How to select content words?
    - Alignment model does not work for dialogue.
    - We need to train a model that capable of allocating a dynamic vocabulary for each input.

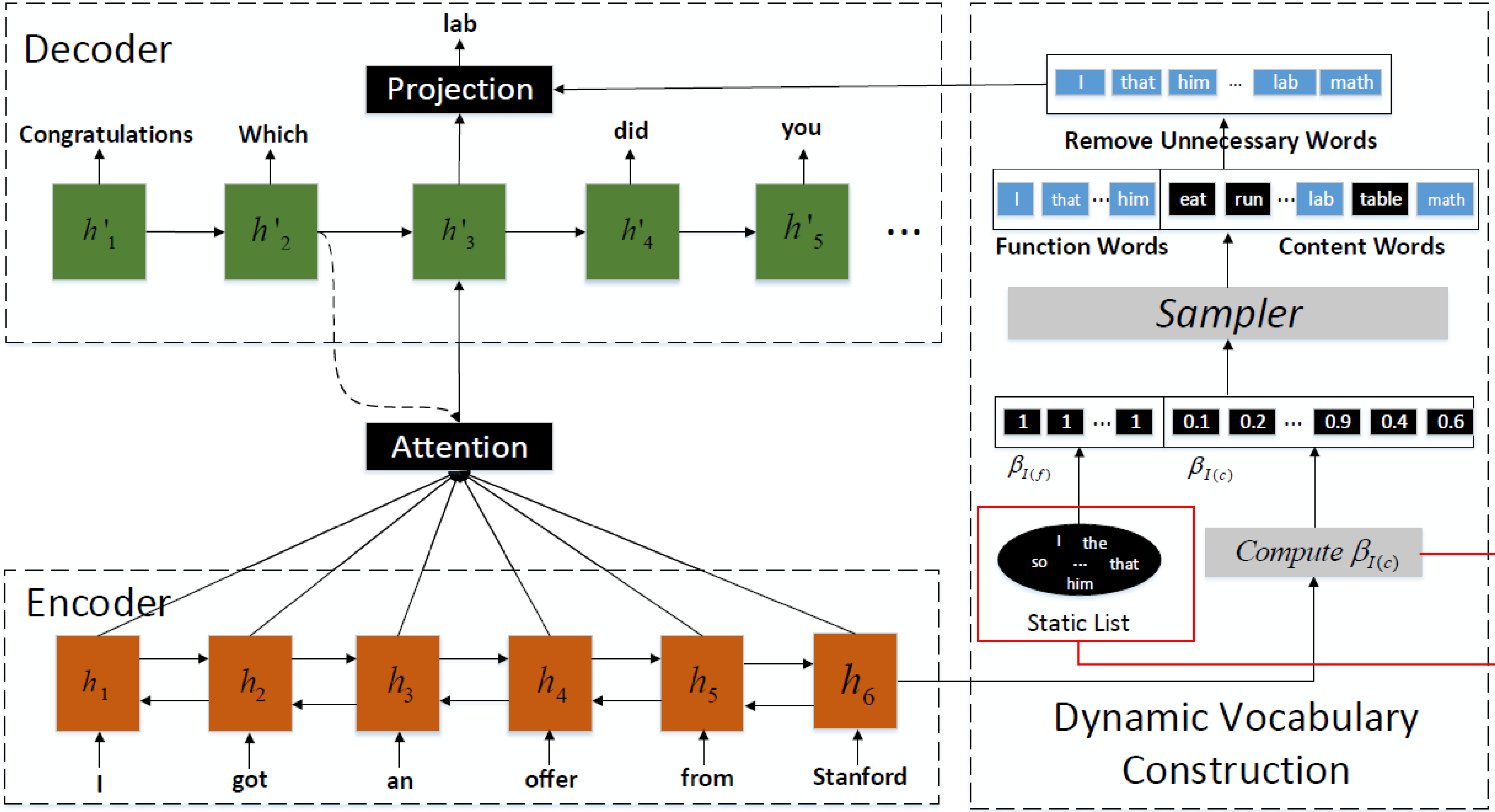
Key Ideas

- Construct a dynamic vocabulary for each input.
  - Save online decoding time
    - It is a time consuming operation to convert a hidden vector into a vocabulary distribution.
    - Matrix multiplication is sensitive to the matrix dimension.
  - Filter irrelevant words
    - Only a small part of words can be used in the decoding.
    - Filter out irrelevant words.

Outline

- Task, challenges, and ideas
- Our approach
  - Dynamic vocabulary for S2S learning
- Experiment
  - Experiment setup: data set and baseline methods
  - Evaluation and analysis

Dynamic Vocabulary Sequence to Sequence (DVS2S)

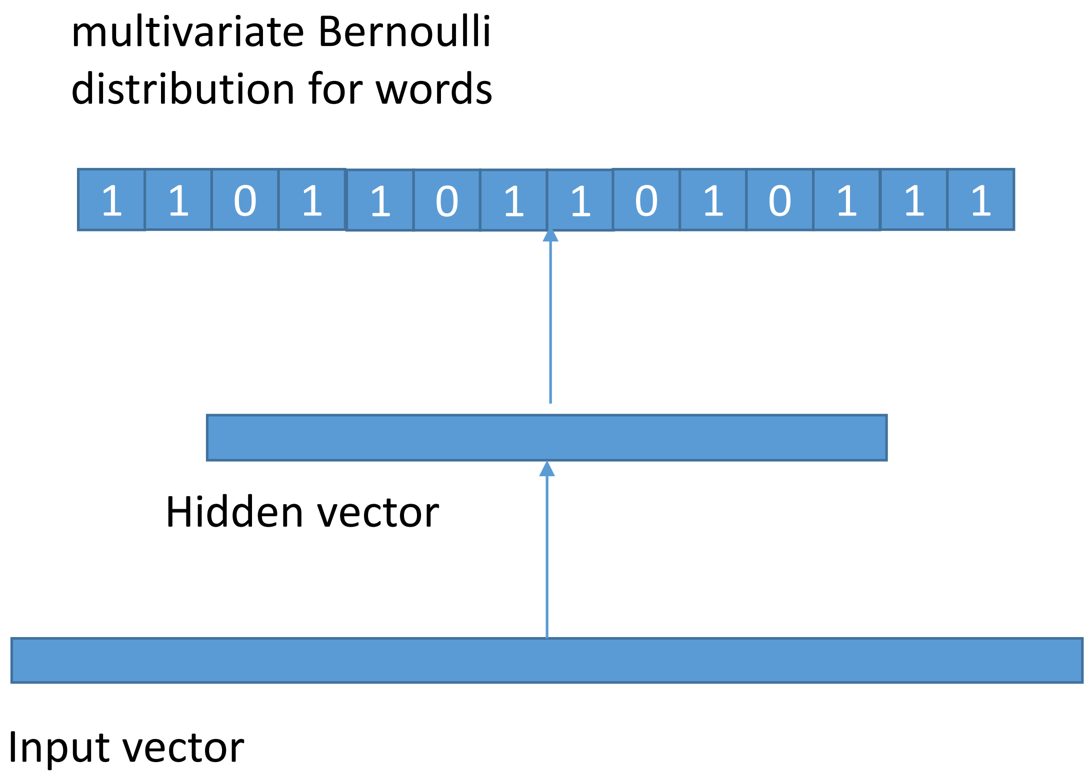


Generate content words

Static vocabulary contains function words

The word prediction model

- The input vector is given by the encoder LSTM
- MLP is employed to predict the vocabulary
- The word prediction loss is formulated as
$$P\left(w_{pos} = 1 \middle| X\right) + p(w_{neg} = 0|X)$$
where $\{w_{neg}\}$ are sampled by frequency, and $\{w_{pos}\}$ are words in the ground-truth response.



Time Complexity of decoding

Existing methods: $len_r \cdot m \cdot p + len_m \cdot m^2 \cdot len_r + len_r(m + p)|V|$

GRU Attention Projection

DVS2S: $len_r \cdot m \cdot p + len_m \cdot m^2 \cdot len_r + len_r(m + p)|T| + m \cdot |V|$

GRU Attention Projection Vocabulary Construction

$len_r(m + p)|V| > len_r(m + p)|T| + m \cdot |V|$, when $len_r > 1$

Model Training: integrate dynamic vocabulary as latent variables

- Objective Function:

$$\sum_{i=1}^{N} \log(p(Y_i|X_i)) = \sum_{i=1}^{N} \log(\sum_{T_i} p(Y_i|T_i, X_i)p(T_i|X_i)).$$

- Lower bound

$$\begin{aligned} L &= \sum_{i=1}^{N} \sum_{T_i} p(T_i|X_i) \log p(Y_i|T_i, X_i) \qquad (10) \\ &= \sum_{i=1}^{N} \sum_{T_i} [\prod_{j=1}^{|V|} p(t_{i,j}|X_i) \sum_{l=1}^{m} \log p(y_{i,l}|y_{i,<l}, T_i, X_i)] \\ &\leq \sum_{i=1}^{N} \log(\sum_{T_i} p(Y_i|T_i, X_i)p(T_i|X_i)) \\ &= \sum_{i=1}^{N} \log[p(Y_i|X_i)] \end{aligned}$$

Model Training: integrate dynamic vocabulary as latent variables

- Gradient: $$\sum_{T_i} p(T_i|X_i)[\frac{\partial \log p(Y_i|T_i, X_i)}{\partial \Theta} + \log(Y_i|T_i, X_i)\frac{\partial \log p(T_i|X_i)}{\partial \Theta}]$$
- Approximate gradient: $$\frac{1}{S}\sum_{s=1}^{S}[\frac{\partial \log p(Y_i|\widetilde{T}_{i,s}, X_i)}{\partial \Theta} + \log(Y_i|\widetilde{T}_{i,s}, X_i)\frac{\partial \log p(\widetilde{T}_{i,s}|X_i)}{\partial \Theta}],$$
- Reduce variance: $$\frac{\partial L_i(\Theta)}{\partial \Theta} \approx \frac{1}{S}\sum_{s=1}^{S}[\frac{\partial \log p(Y_i|\widetilde{T}_{i,s}, X_i)}{\partial \Theta} + ((\frac{1}{m}\sum_{j=1}^{m} \log p(y_{i,j}|y_{i,<j}, |\widetilde{T}_{i,s}, X_i) - b_k)\frac{\partial \log p(\widetilde{T}_{i,s}|X_i)}{\partial \Theta}],$$

Algorithm 1: Optimization Algorithm

Input: $\mathcal{D}$, V, initial learning rate lr, MaxEpoch
Init: Θ
  Pretrain a Seq2Seq model with $\mathcal{D}$.
  Fix the encoder, and pre-train $\{W_c, b_c\}$ in Equation (8) by maximizing $\sum_{i=1}^{N} \sum_{j=1}^{|V|} \log[p(t_{i,j}|X_i)]$
while $e <$ *MaxEpoch* ***and*** *perplexity does not increase in 2 successive epchos* **do**
  foreach *mini-batch k* **do**
    Compute the sampling probability $\{\beta_i\}^{|V|}$ with Equation (8)
    for $s < S$ **do**
      Sample a $\widetilde{T}_s \sim$ multivariate Bernoulli$(\{\beta_i\}^{|V|})$
      Compute loss according to Equation (10)
      Compute gradient according to Equation (13)
    end
    Update b_k according to Equation (14)
    Update parameter Θ with AdaDelta algorithm
  end
  if *perplexity increases* **then**
    $lr = lr/2$
  end
end
Output: Θ

Outline

- Task, challenges, and ideas
- Our approach
  - Dynamic vocabulary for S2S learning
- Experiment
  - Experiment setup: data set and baseline methods
  - Evaluation and analysis

Dataset: Baidu Tieba data

	train	val	test
message-response pairs	5M	10000	10000
Vocabulary Size	30000		
Vocabulary Coverage	98.8% words in messages, and 98.3% words in responses		

Baseline Methods

- S2SA: the standard S2S model with an attention machenism. We use the implementation with Blocks https://github.com/mila-udem/blocks
- S2SA-MMI: the model proposed by Li et al. (Li et al.2015). We implement this baseline by the code publishedby the authors at https://github.com/jiweil/Neural-Dialogue-Generation.
- TA-S2S: the topic-aware sequence-to-sequence modelproposed in (Xing et al. 2016). We implement this base-line by the code published by the authors at https://github.com/LynetteXing1991/TAJA-Seq2Seq.
- CVAE: recent work for response generation with a con-ditional variational auto-encoder (Zhao, Zhao, and Esḱenazi). We use the published code at https://github.com/snakeztc/NeuralDialog-CVAE

Evaluation Metrics

- Until now, how to evaluate generated response automatically is still an open problem.

- Word overlap based method: BLEU, ROUGE …
- Embedding based metrics: Embedding Average (Average), Embedding Extrema (Extrema), and Embedding Greedy (Greedy)
- Diversity Evaluation: **Distinct-ngram**, entropy
- Toward Turning Test: employ a discriminator

More details: 中国计算机学会通讯 > 2017年第9期: 对话系统评价技术进展及展望

Quantitative Evaluation

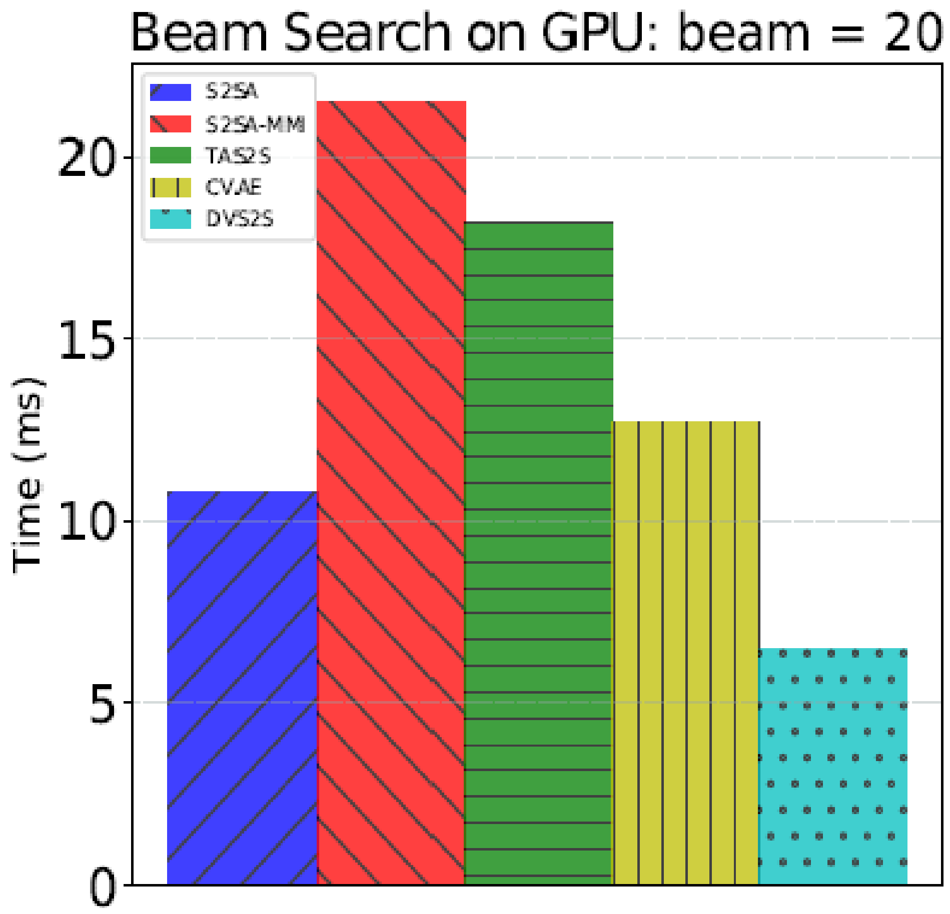
Table 1: Automatic evaluation results. Numbers in bold mean that improvement from the model on that metric is statistically significant over the baseline methods (t-test, p-value < 0.01).

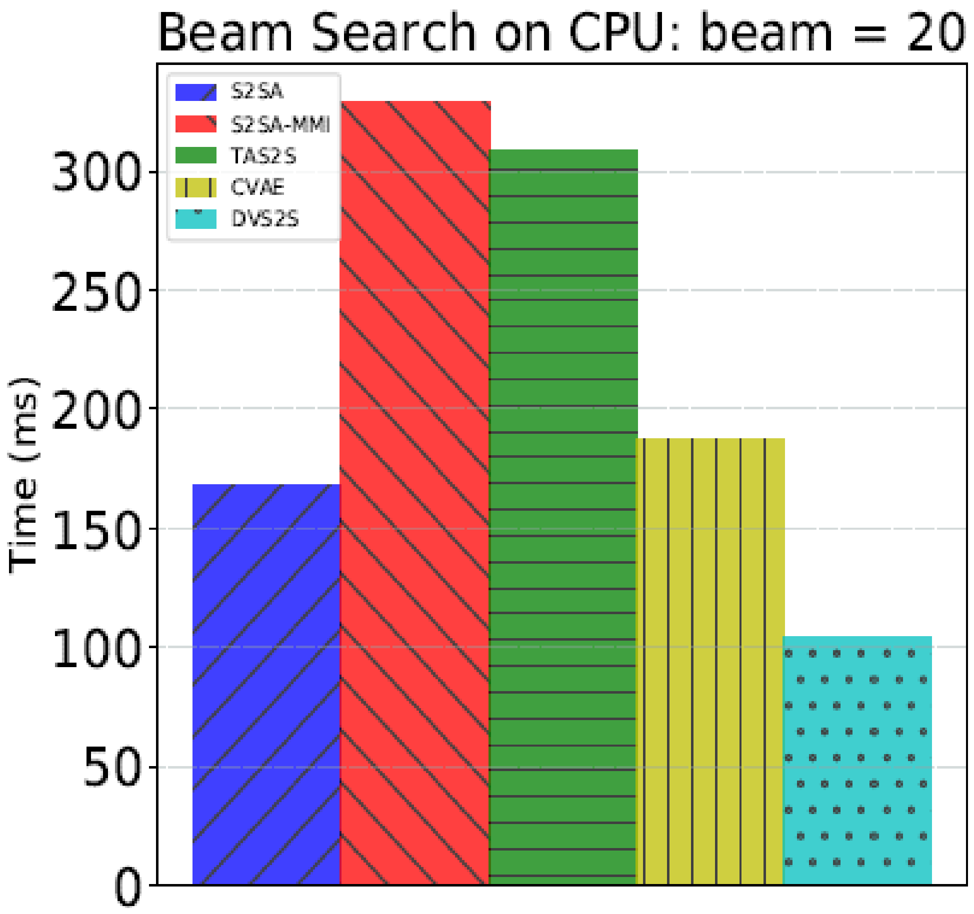
	BLEU-1	BLEU-2	BLEU-3	Average	Extrema	Greedy	Distinct-1	Distinct-2
S2SA	4.96	1.96	0.81	25.32	11.70	24.73	0.091	0.207
S2SA-MMI	5.82	1.47	0.70	27.16	14.96	25.89	0.151	0.378
TAS2S	6.26	2.11	0.98	27.92	15.86	26.29	0.161	0.401
CVAE	6.33	1.86	0.55	28.92	15.01	26.13	0.143	0.346
S-DVS2S	**8.01**	2.94	0.93	**32.41**	**20.15**	**29.89**	**0.221**	**0.601**
DVS2S	**9.89**	3.51	1.33	**34.05**	**22.72**	**31.61**	**0.233**	**0.632**

Table 2: Human evaluation results. The ratios are calculated by combining the annotations from the three judges together.

	0	1	2	Kappa
S2SA	0.321	0.564	0.115	0.43
S2SA-MMI	0.302	0.555	0.143	0.42
TAS2S	0.249	0.571	0.180	0.40
CVAE	0.252	0.563	0.185	0.41
S-DVS2S	0.232	0.484	0.284	0.38
DVS2S	0.094	0.581	0.325	0.45

Efficiency comparison





Case Study

Message	DVS2S	TAS2S	S2SA
我买了辆新奥迪。 I bought a new Audi.	奥迪的性价比很高。 Audi is cost-effective. **keyword**: 奥迪(Audi), 奔驰(Benz), 价格(price) ...	哈哈，你赢了 haha, you win.	在哪儿 Where is it?
奢侈品销售收入还不错。 Luxury salesmen have a good payment.	做销售的都收入不高。 Salesmen do not have high salary. **keyword**: 销售(Salesman), 薪水(Salary),金钱(Money) ...	有什么推荐的吗? Do you have something to recommand?	有兴趣的话可以来看看 Come here if you want.
为什么这个药店被敲诈了? Why the pharmacy was extorted?	药店有卖假货的。 The pharmacy sold fake products. **keyword**:药店(pharmacy), 药(medicine),发烧(fever) ...	我也有这样的问题 I have the same question.	这是个问题 That is a problem.

Summary

- (1) proposal of changing the static vocabulary mechanism to a dynamic vocabulary mechanism in the response generation for chatbots.
- (2) proposal of a dynamic vocabulary sequence-to-sequence model and derivation of a learning approach that can jointly optimize word selection and response generation.
- (3) empirical verification of the effectiveness and efficiency of the proposed model on large scale conversation data.

Reference

- Li, J.; Galley, M.; Brockett, C.; Gao, J.; and Dolan, B. 2015. A diversity-promoting objective function for neural conversation models. NAACL 2016
- Li, J.; Monroe, W.; Ritter, A.; Galley, M.; Gao, J.; and Jurafsky,D. 2016b. Deep reinforcement learning for dialogue generation. EMNLP 2016
- Li, J.; Monroe, W.; Shi, T.; Ritter, A.; and Jurafsky, D. 2017. Adversarial learning for neural dialogue generation. EMNLP 2017
- Serban, I. V.; Sordoni, A.; Lowe, R.; Charlin, L.; Pineau, J.; Courville, A. C.; and Bengio, Y. 2017. A hierarchical latent variable encoder-decoder model for generating dialogues. In AAAI, February 4-9, 2017, San Francisco, California, USA., 3295–3301.
- Vinyals, O., and Le, Q. 2015. A neural conversational model. arXiv preprint arXiv:1506.05869.
- Xing, C.; Wu, W.; Wu, Y.; Liu, J.; Huang, Y.; Zhou, M.; and Ma, W.-Y. Topic aware neural response generation. AAAI 2017.
- Xu, Z.; Liu, B.; Wang, B.; SUN, C.; Wang, X.; Wang, Z.; and Qi,C. 2017. Neural response generation via gan with an approximate embedding layer. In EMNLP 2017.
- Coherent Dialogue with Attention-based Language Models Hongyuan Mei, Mohit Bansal, Matthew R. Walter Proceedings of AAAI 2017, San Francisco, California.
- Ji, Zongcheng, Zhengdong Lu, and Hang Li. "An information retrieval approach to short text conversation." *arXiv preprint arXiv:1408.6988* (2014).
- Chen, Hongshen, et al. "A Survey on Dialogue Systems: Recent Advances and New Frontiers." *arXiv preprint arXiv:1711.01731* (2017).
- Mou, L.; Song, Y.; Yan, R.; Li, G.; Zhang, L.; and Jin, Z. 2016. Sequence to backward and forward sequences: A content-introducing
- approach to generative short-text conversation. COLING 2016

THANKS!