Recent Advances towards Dialogue Systems in Open Domain

Dr. Rui Yan

Wangxuan Institute of Computer Technology, Peking University

ruiyan@pku.edu.cn

www.ruiyan.me

Conversational AI

- Human-computer conversation has been attracting increasing attention.
- Conversational agent (ChatBot)
  - e.g., Xiaoice (Microsoft), Turing Robot
- Virtual personal assistant
  - e.g., Cortana (Microsoft), Siri (Apple), Now (Google)
- E-commerce customer service robot
  - e.g., Alime (Alibaba), Jimi (JingDong)

Taxonomy of Dialogue Systems

- Domain
  - Vertical domain (Task driven)
    - Complete domain-specific tasks (e.g., hotel booking, weather enquiries, etc)
  - Open domain (Non-task driven)
    - Naturally and meaningfully converse with humans on any open domain topics
- Technique
  - Templated-based
  - Retrieval-based
  - Generation-based
  - Ensemble-based

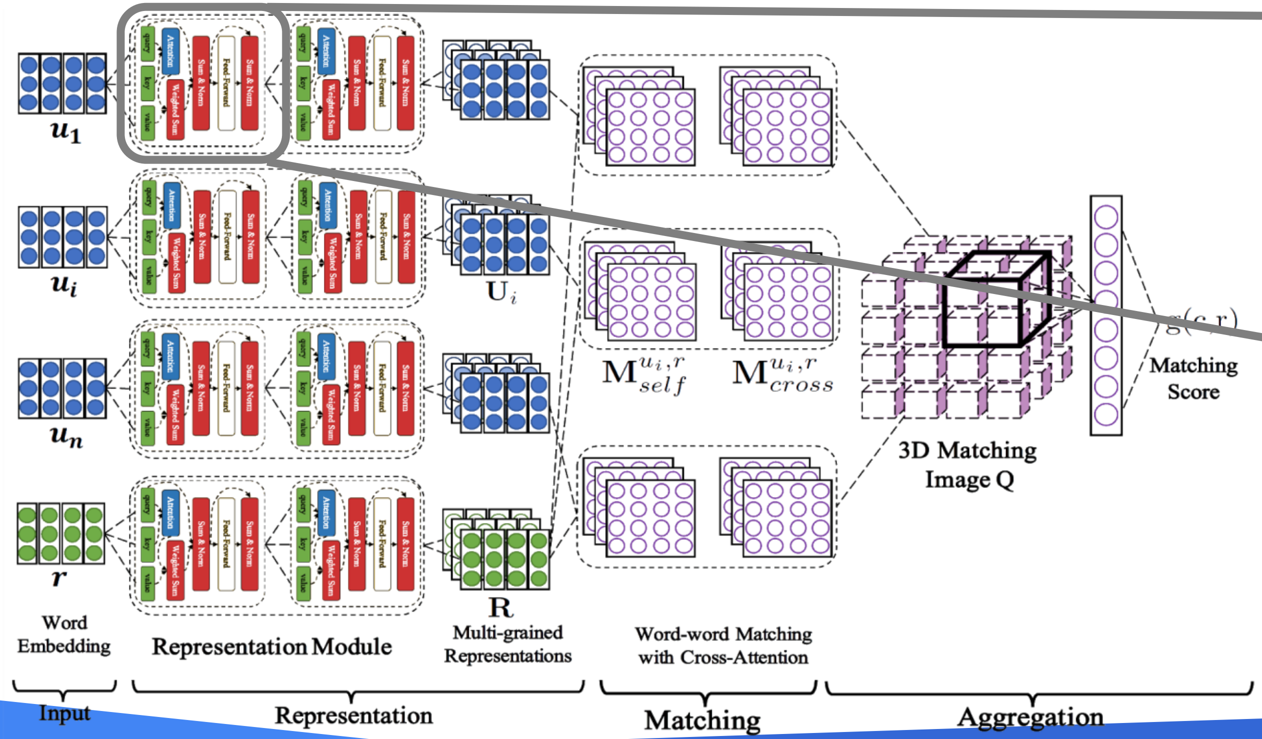
Retrieval-based Approaches

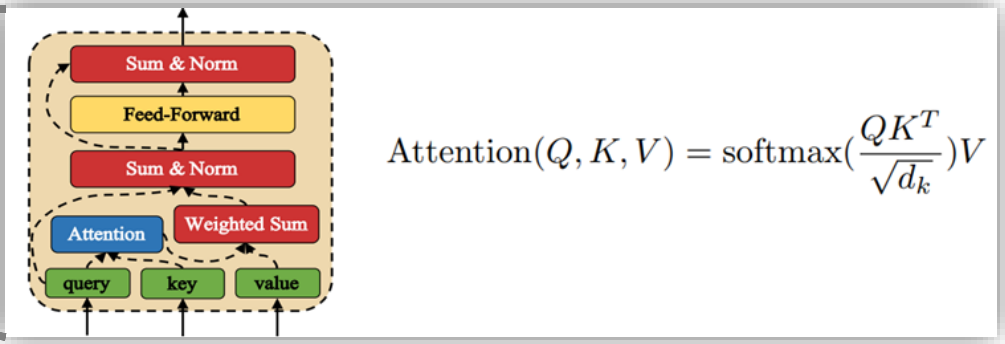
Matching with Better Representation

I. Representations Go Deep

[Zhou et al., ACL 2018]

- Deep Attention Matching Network (DAM)
  - Representing utterances and responses by stacking multiple attention modules





- **Self-attention**: let an utterance attend itself
- **Cross-attention**: let an utterance and a response attend each other

Modeling long-term dependency among words

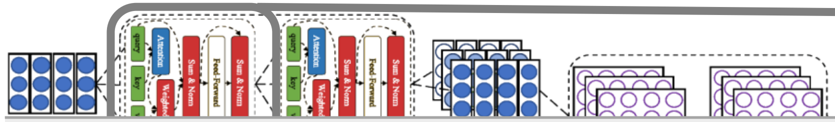
Zhou et al., Multi-turn Response Selection for Chatbots with Deep Attention Matching Network. *ACL'18*

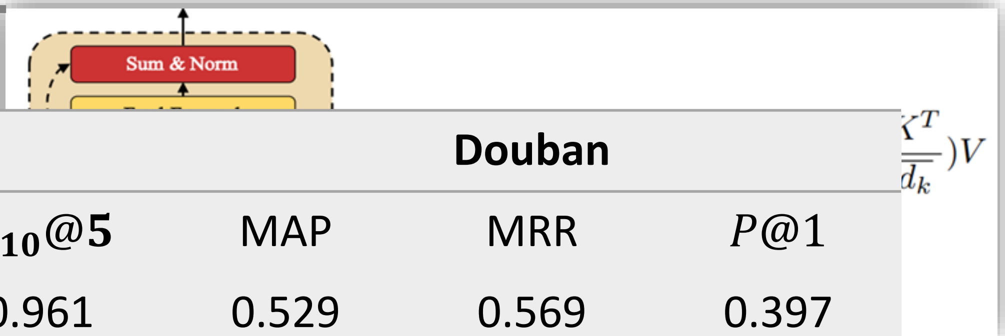
Matching with Better Representation

I. Representations Go Deep

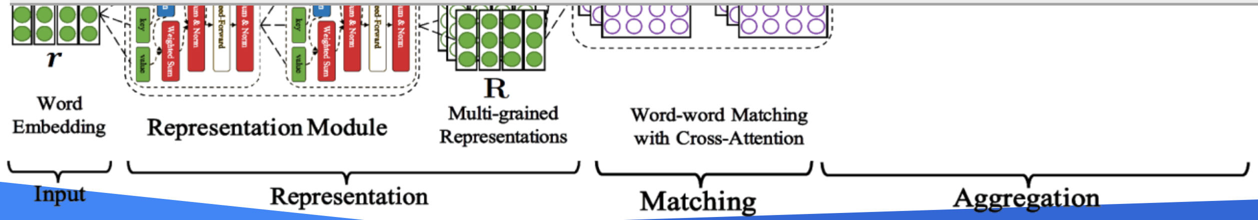
[Zhou et al., ACL 2018]

- Deep Attention Matching Network (DAM)
  - Representing utterances and responses by stacking multiple attention modules





Model	Ubuntu				Douban		
	$R_2@1$	$R_{10}@1$	$R_{10}@2$	$R_{10}@5$	MAP	MRR	$P@1$
SMN	0.926	0.726	0.847	0.961	0.529	0.569	0.397
DAM	0.938	0.767	0.874	0.969	0.550	0.601	0.427
	+1.2%	+4.1%	+2.7%	+0.8%	+2.1%	+3.1%	+3.0%



g-

- **Cross-attention**: let an utterance and a response attend each other

term dependency among words

Zhou et al., Multi-turn Response Selection for Chatbots with Deep Attention Matching Network. *ACL'18*

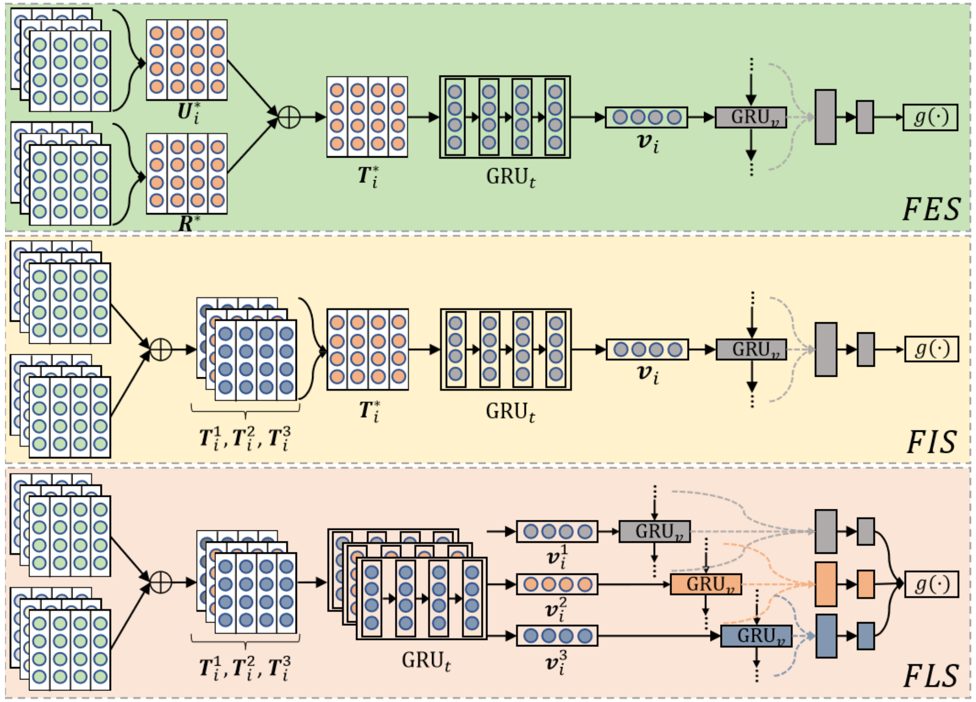
Matching with Better Representation

II. Representations Go Wide

[Tao et al., WSDM 2019]

- Multi-Representation Fusion Network (MRFN)
  - Fusing multiple types of representations are helpful, but how to fuse matters.

- Word2Vec
- Char-based Embedding
- CNN
- RNN
- Self-attention
- Cross-attention



Fusing **before** interaction

Fusing **after** interaction, but **before** aggregation

Fusing **in the end**

Matching with Better Representation

II. Representations Go Wide **[Tao et al., WSDM 2019]**

- Multi-Representation Fusion Network (MRFN)
  - Fusing multiple types of representations are helpful, but how to fuse matters.

Model	Ubuntu				Douban		
	$R_2@1$	$R_{10}@1$	$R_{10}@2$	$R_{10}@5$	MAP	MRR	$P@1$
SMN	0.926	0.726	0.847	0.961	0.529	0.569	0.397
DAM	0.938	0.767	0.874	0.969	0.550	0.601	0.427
MRFN(FES)	0.930	0.742	0.857	0.963	0.538	0.583	0.405
MRFN(FIS)	0.936	0.762	0.870	0.967	0.558	0.605	0.438
MRFN(FLS)	0.945	0.786	0.886	0.976	0.571	0.617	0.448
	+0.7%	**+1.9%**	**+1.2%**	**+0.7%**	**+2.1%**	**+1.6%**	**+2.1%**

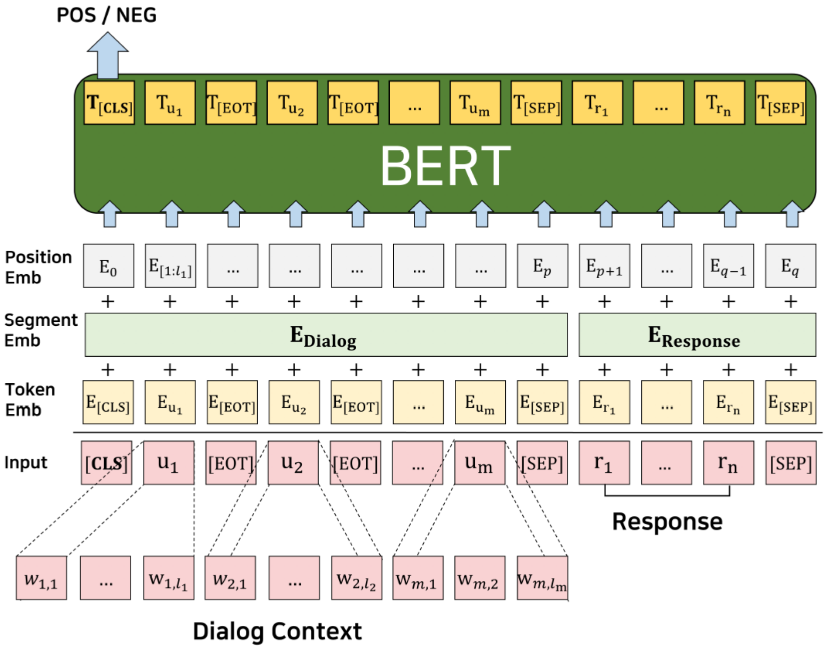
Tao et al., Multi-Representation Fusion Network for Multi-turn Response Selection in Retrieval-based Chatbots. *WSDM'19*

Matching with Better Representation

III. Representations from Pre-Training

[Whang et al., arXiv]

- Pre-training neural networks on large scale data sets as representations significantly improves the existing models.



Bi-directional Encoder Representations from Transformer (BERT)

Model	$R_{10}@1$	$R_{10}@2$	$R_{10}@5$
MultiView	0.662	0.801	0.951
DL2R	0.626	0.783	0.944
AK-DE-biGRU	0.747	0.868	0.972
$\text{SMN}_{dynamic}$	0.726	0.847	0.961
DUA	0.752	0.868	0.962
DAM	0.767	0.874	0.969
IMN	0.777	0.888	0.974
ESIM	**0.796**	**0.894**	0.975
MRFN_{FLS}	0.786	0.886	**0.976**
BERT_{base}	0.817	0.904	0.977
BERT-DPT	0.851	0.924	0.984
BERT-VFT	**0.855**	**0.928**	**0.985**
BERT-VFT(DA)	**0.858**	**0.931**	**0.985**

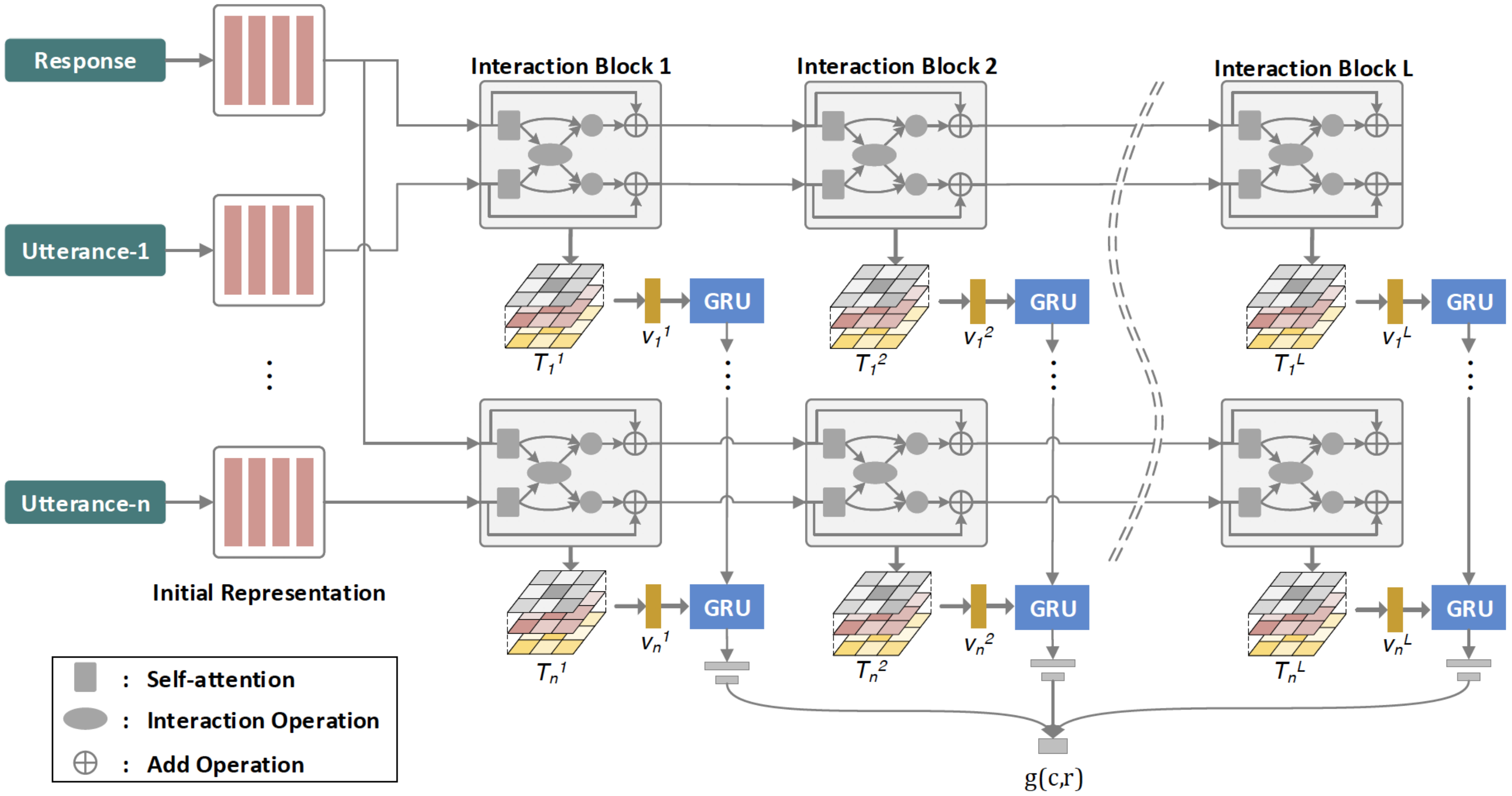
Table 1: Model comparison on Ubuntu Corpus V1.

Whang et al., Domain Adaptive Training BERT for Response Selection, 2019 Arxiv

Matching with Better Interaction

- Interaction-over-interaction network
  - ➢ Representations-[Interaction]^K -Aggregation

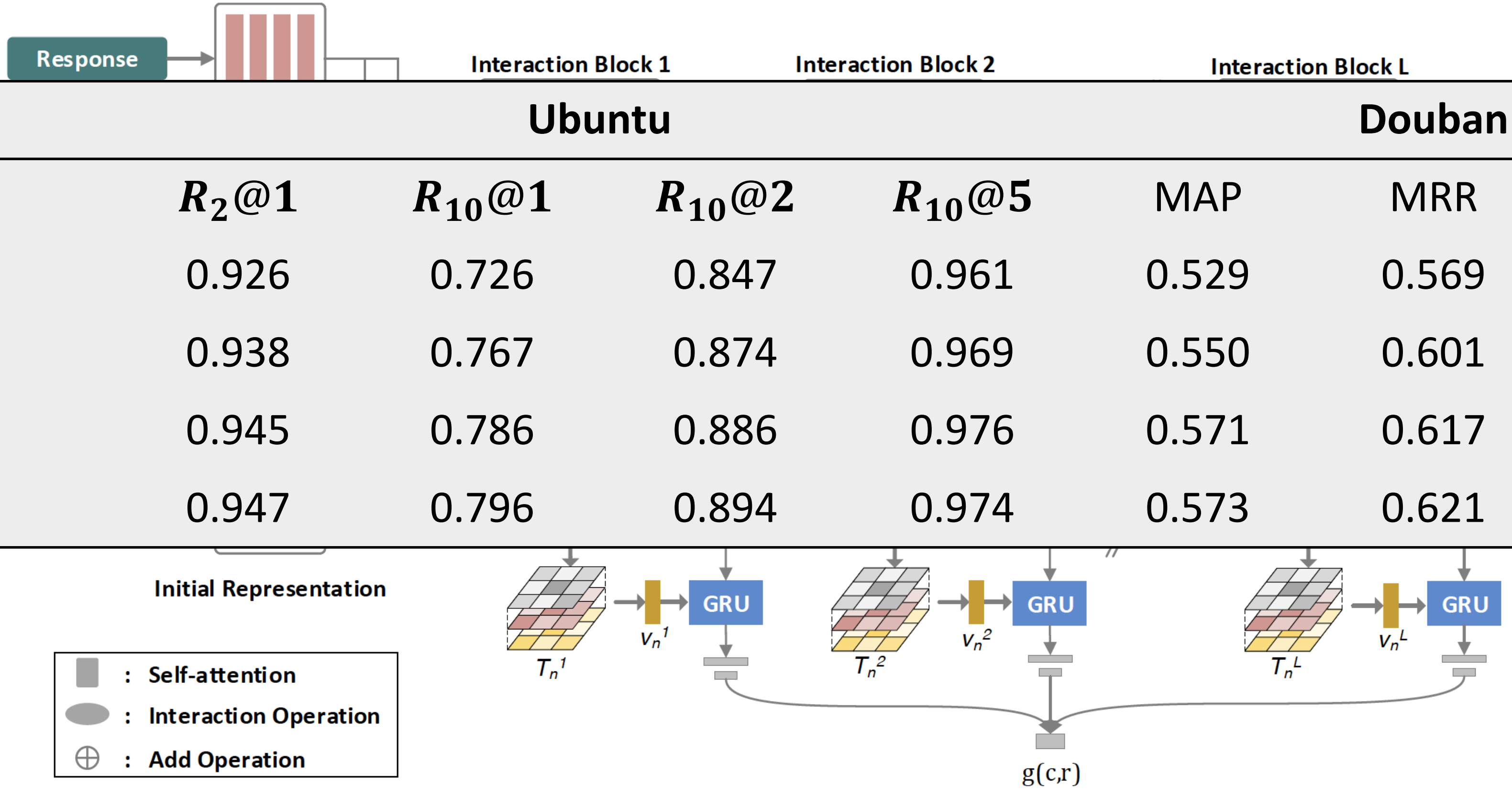
[Tao et al., ACL 2019]



Matching with Better Interaction

- Interaction-over-interaction network
  - Representations-[Interaction]^{K} -Aggregation

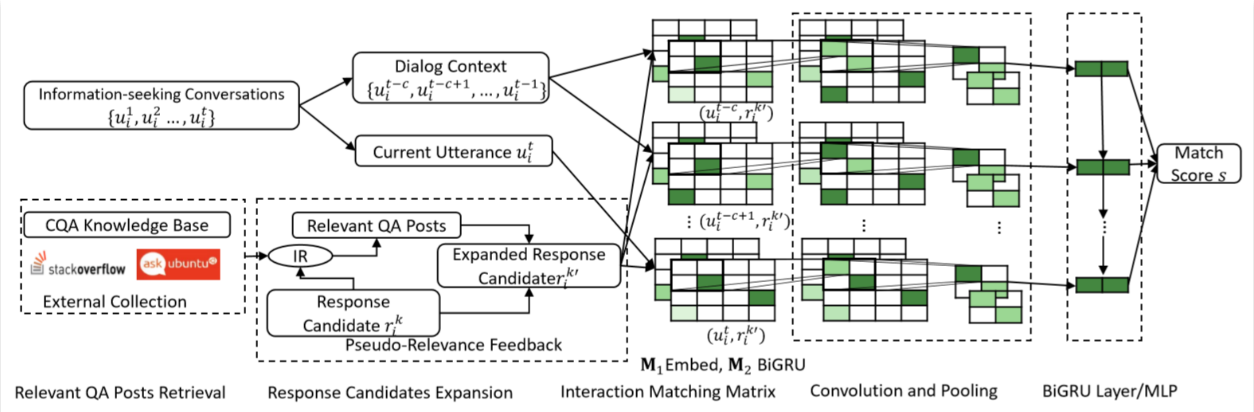
[Tao et al., ACL 2019]



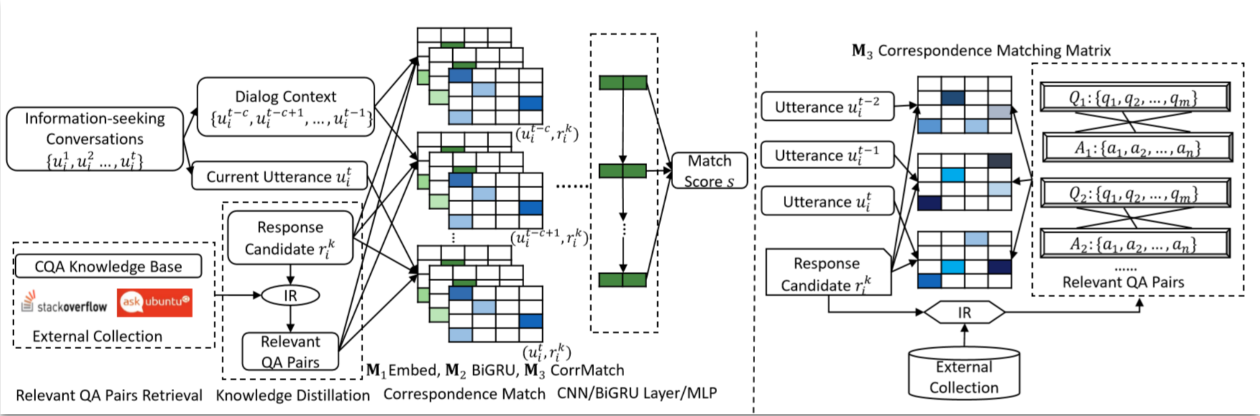
Model	Ubuntu				Douban		
	$R_2@1$	$R_{10}@1$	$R_{10}@2$	$R_{10}@5$	MAP	MRR	$P@1$
SMN	0.926	0.726	0.847	0.961	0.529	0.569	0.397
DAM	0.938	0.767	0.874	0.969	0.550	0.601	0.427
MRFN(FLS)	0.945	0.786	0.886	0.976	0.571	0.617	0.448
IOI	0.947	0.796	0.894	0.974	0.573	0.621	0.444

Matching with External Knowledge

[Yang et al., SIGIR 2018]



Knowledge is incorporated into matching through ***Pseudo Relevance Feedback***

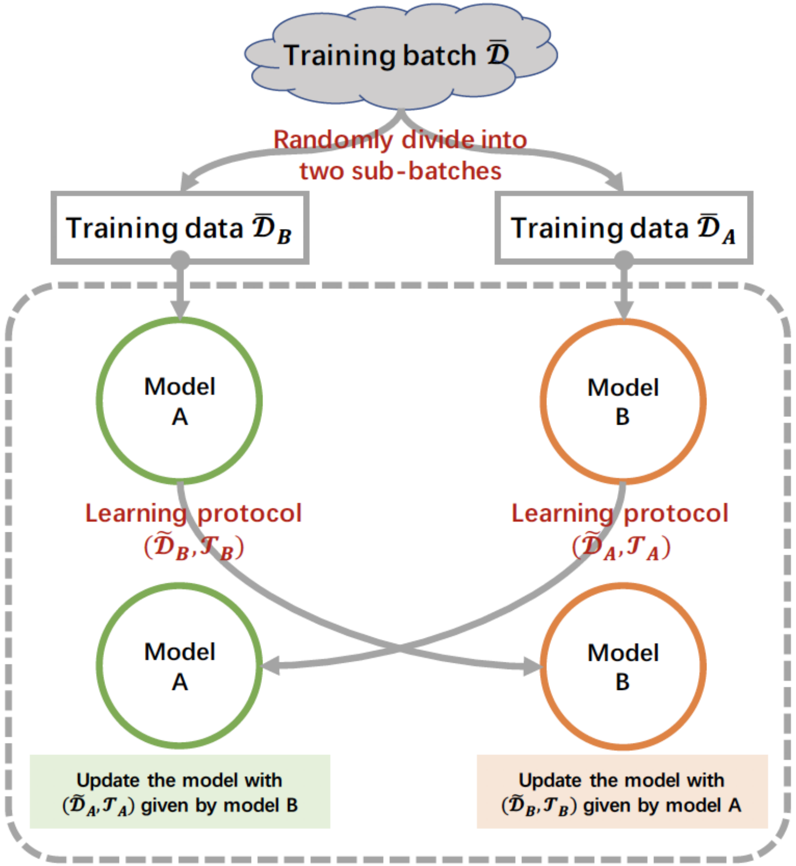


Knowledge is incorporated into matching through an ***Extra Matching Channel***

Learning a Better Matching model

- Learning with Co-Teaching – Denoising with Your Peer

[Feng et al., ACL 2019]



Key Ideas

- Teaching: two models judge quality of training examples mutually. The knowledge is transferred between the two models through learning protocols.
- Learning: two models learn from their peers via the transferred learning protocols.
- Co-evolving: through teaching and learning, the two models get improved together.
- Resemble: two peer students who learn from different but related materials inspire each other during learning through knowledge exchange.

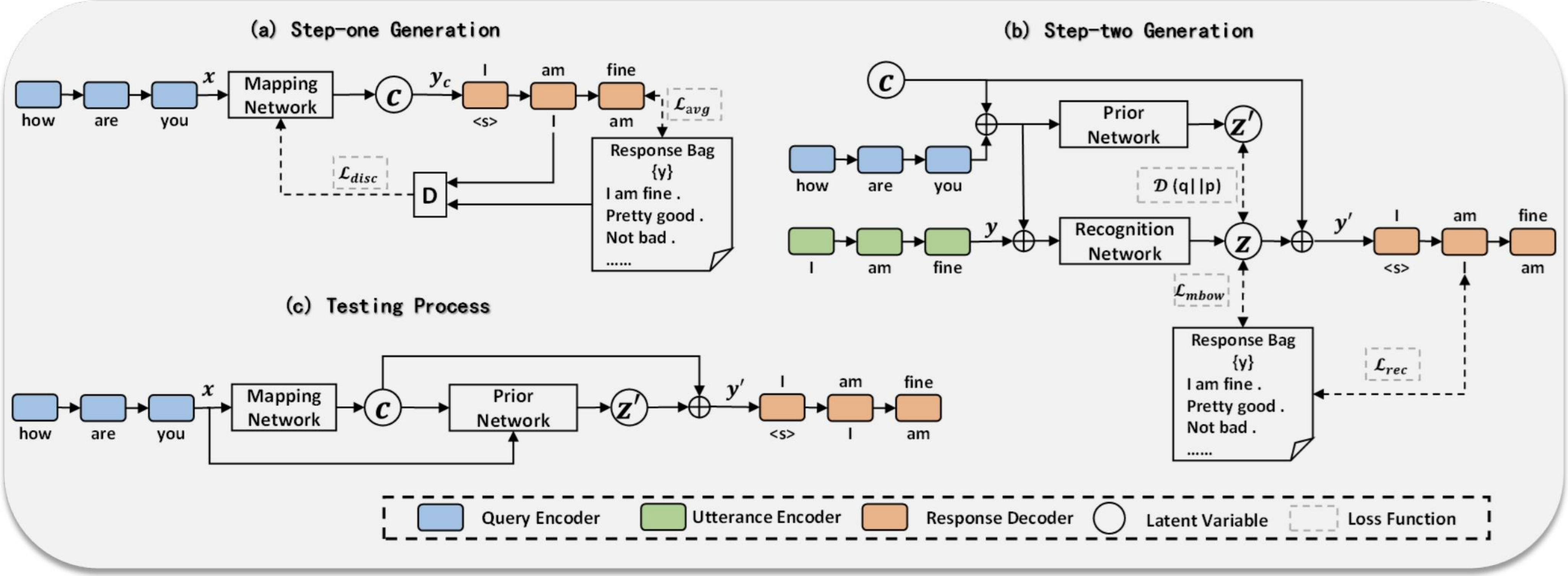
Feng et al., Learning a Matching Model with Co-Teaching for Multi-turn Response Selection in Retrieval-based Dialogue Systems. *ACL'19*

Generation-based Approaches

Response Diversity

[Xu et al., ACL 2019]

- Modeling the 1-to-n mapping by considering the correlation of different valid responses.



Response Diversity

[Xu et al., ACL 2019]

- Controlling multiple attributes in response generation (customize responses by tailoring the set of attributes)

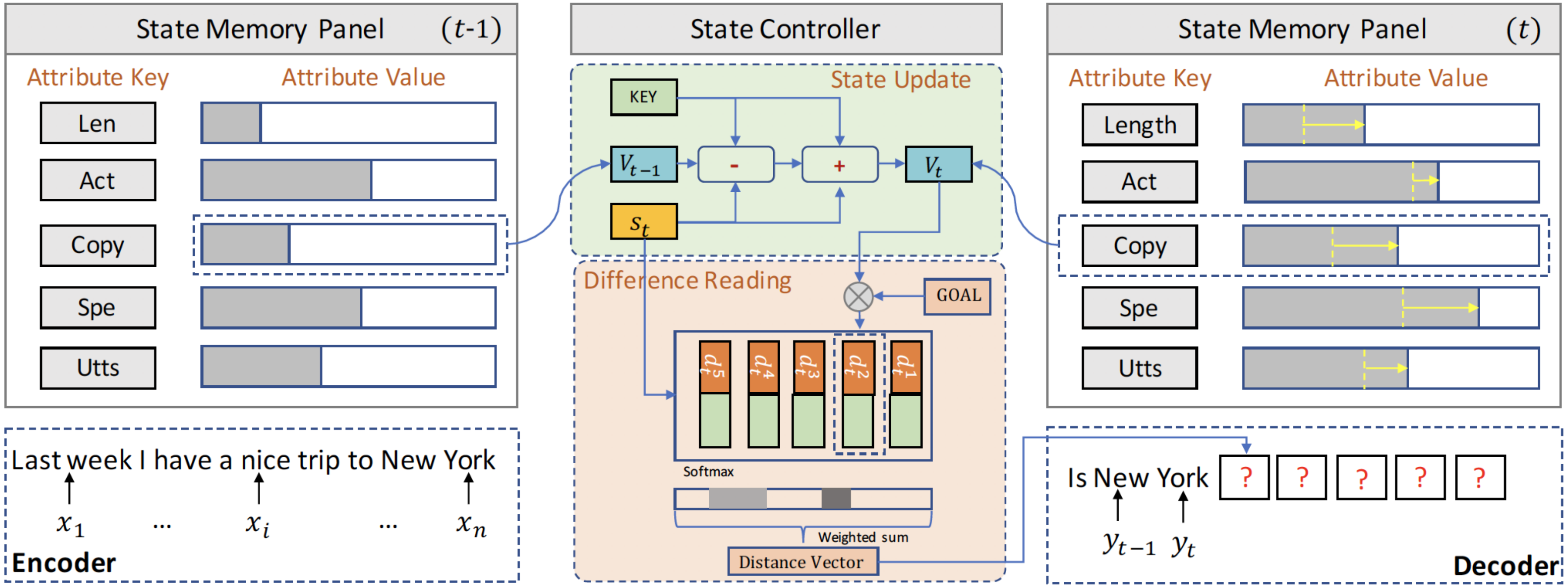
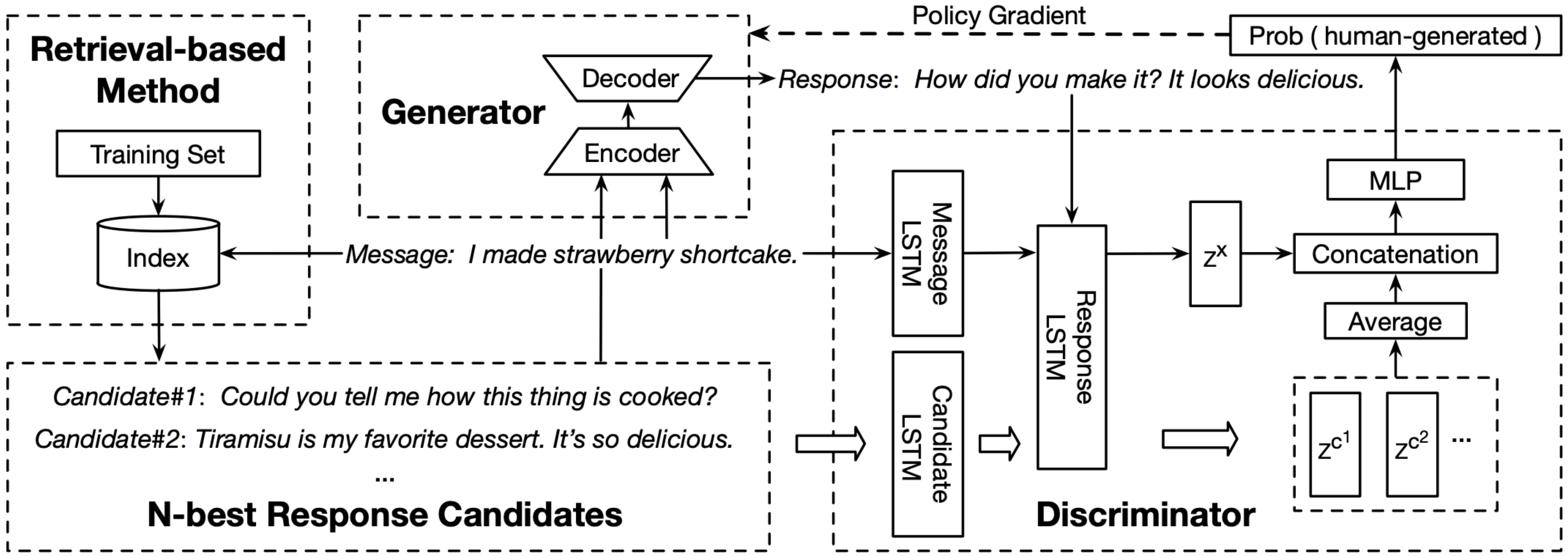


Figure 1: Architecture of goal tracking memory enhanced sequence-to-sequence model.

Retrieval-then-Generation

[Zhu et al., ACL 2019]

- Retrieval-Enhanced Adversarial Training (REAT) method for neural response generation.



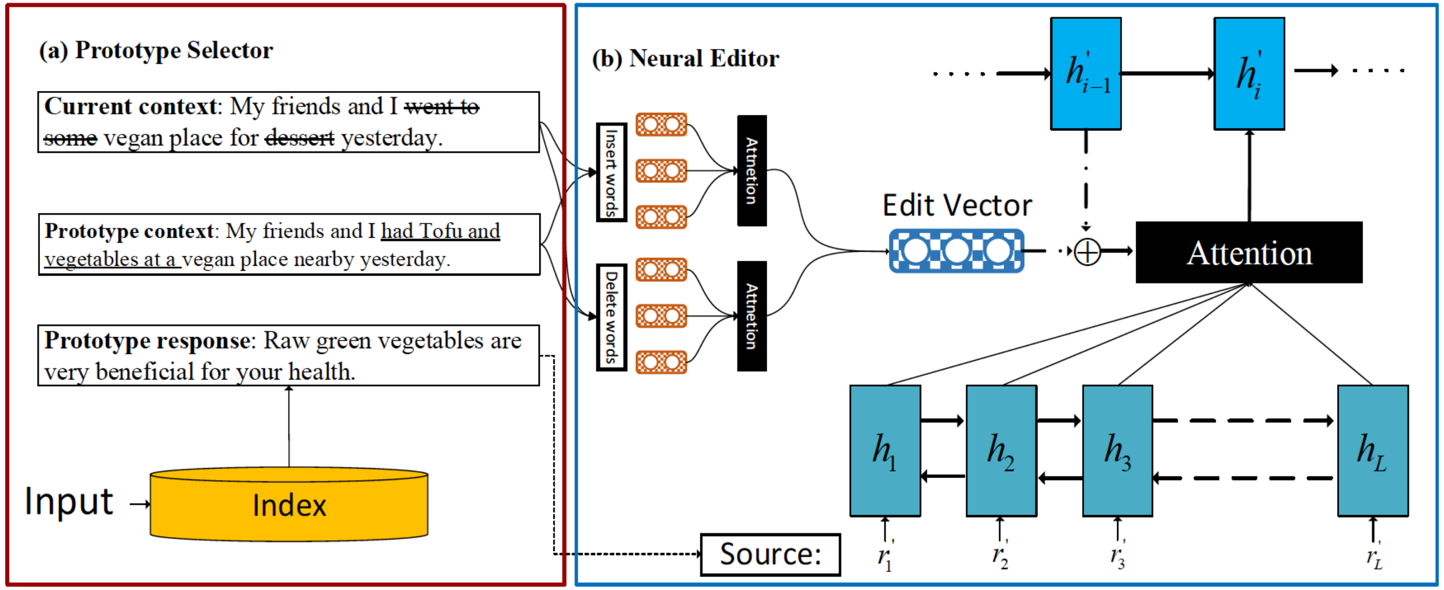
Retrieval-then-Generation

[Wu et al., AAAI 2019]

- A prototype-then-edit paradigm for response generation

Context	My friends and I ~~went to some~~ vegan place for ~~dessert~~ yesterday.
Prototype context	My friends and I had Tofu and vegetables at a vegan place nearby yesterday.
Prototype response	**Raw green vegetables** are very **beneficial** for your health.
Revised response	**Desserts** are very **bad** for your health.

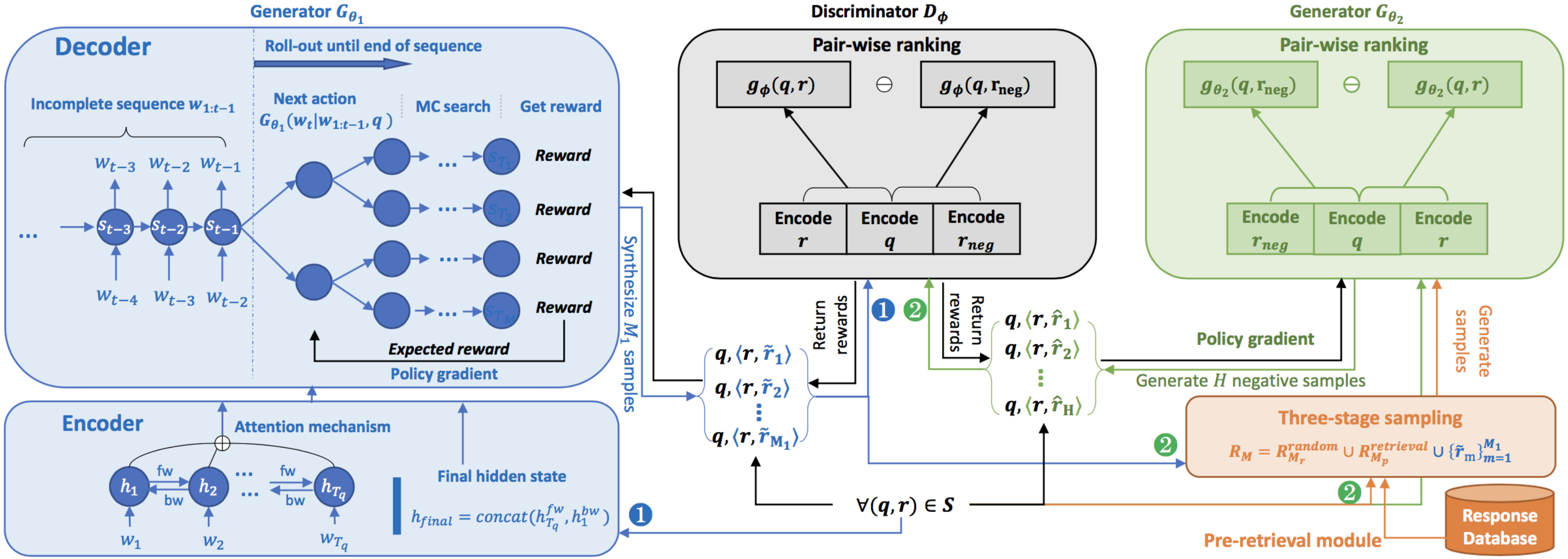
Table 1: An example of context-aware prototypes editing. Underlined words mean they do not appear in the original context, while ~~words with strikethrough~~ mean they are not in the prototype context. Words in bold represent they are modified in the revised response.



Retrieval-Generation Ensemble

[Zhang et al., SIGIR 2019]

- An adversarial learning framework for enhancing a retrieval-generation ensemble model (mutual enhanced)



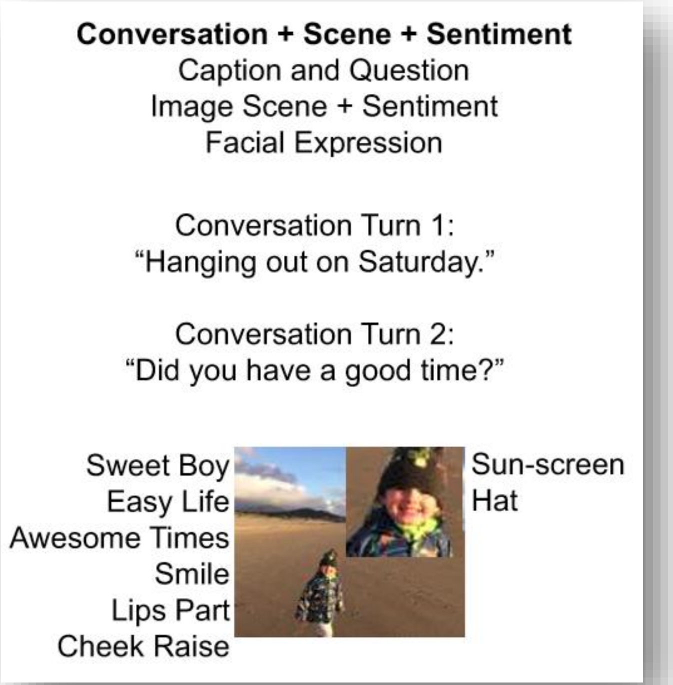
More Challenges to Dialogue System

- Larger context (especially in multi-turn dialogue)
  - How to encode long context information
  - Dose the neural model understand the context?
- How to select/generate a meaningful and coherent reply?
- How to incorporate domain knowledge, world knowledge?
- How to design fast (approximate) algorithms for efficient training and inference.
  - Current dialog models are clumsy and require long training time and huge computational power.

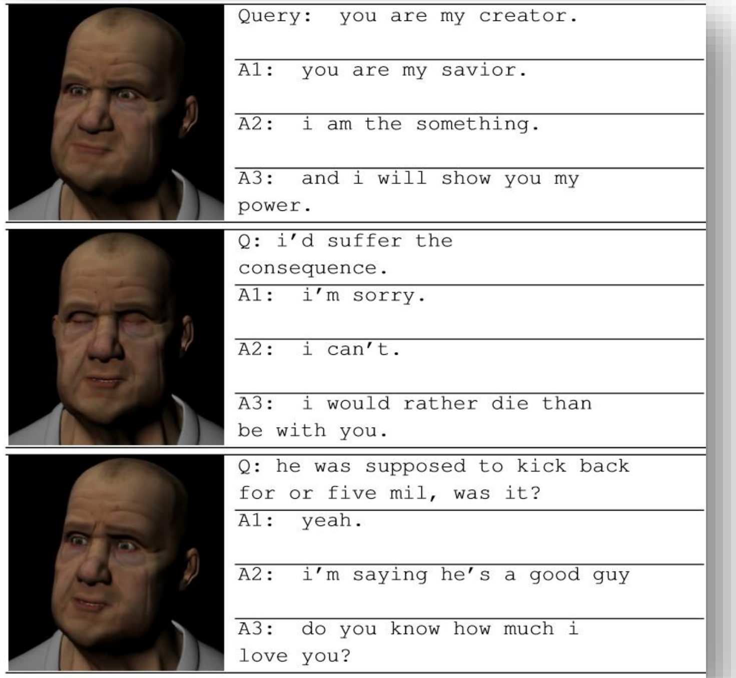
Looking Forward

- Model Design -> Model Learning
- Single Modality -> Multiple Modalities
- Big Data -> Small Data
  - Big data: mixed intentions, styles, characters, ……
  - Small data: specific intention, style, character, ……

Single Modality to Multiple Modalities



Huber et al., Emotional Dialogue Generation using Image-Grounded Language Models. CHI'18



Chu et al., A Face-to-Face Neural Conversation Model. CVPR'18

Q&A

ruiyan@pku.edu.cn