



2018 Conference on Empirical Methods in Natural Language Processing

October 31–November 4
Brussels, Belgium

Deep Chit-Chat: Deep Learning for ChatBots

Wei Wu and Rui Yan

Microsoft Corporation and Peking University

EMNLP 2018

Brussels, Belgium

Speaker Information

- **Dr. Wei Wu**
- Principal Applied Scientist Lead at Microsoft Xiaoice
- <https://sites.google.com/view/wei-wu-homepage>
- Contact information: wuwei@microsoft.com
- Slides <http://www.ruiyan.me/pubs/tutorial-emnlp18.pdf>; <https://sites.google.com/view/wei-wu-homepage/talks>
- **Dr. Rui Yan**
- Assistant Professor at Peking University
- <http://www.ruiyan.me>
- Contact information: ruiyan@pku.edu.cn

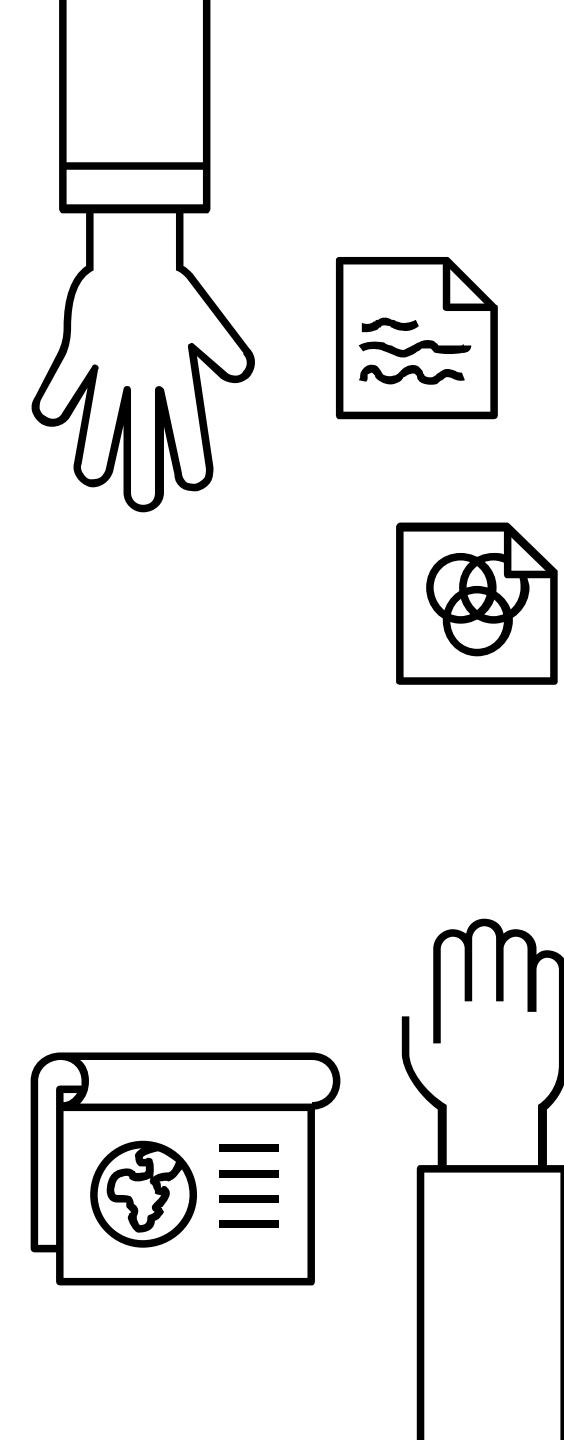


Outline

- Introduction
- Basic Deep Learning Concepts
- Retrieval-based Chatbots
- **Break**
- Generation-based Chatbots
- Evaluation
- New Trends and Conclusion

Our Target Audience

- ▶ PhD students or researchers working on open domain dialogue systems.
- ▶ NLP/IR/ML engineers or scientists with hands-on experience on building **chatbots** with deep learning techniques.
- ▶ Anyone who wants to learn how neural approaches (i.e., deep learning techniques) can be applied to building **chatbots**.
- ▶ Anyone who wants to build **chatbots** that can chat like humans (better with deep learning background).



Chatbots are “HOT”

Academia

The Alexa Prize
Over \$3.5 Million to Advance Conversational Artificial Intelligence
December 2017 - November 2018



The Conversational Intelligence Challenge 2 (ConvAI2)
NIPS 2018 Competition

[View On GitHub](#)

ConvAI2: Overview of the competition
Prize News Current Leaderboard PersonaChat ConvAI2 Dataset Evaluation

ConvAI2: Overview of the competition
There are currently few datasets appropriate for training and evaluating models for non-goal-oriented dialogue systems (chatbots); and equally problematic, there is currently no standard procedure for evaluating such models beyond the classic Turing test.
The aim of our competition is therefore to establish a concrete scenario for testing chatbots that aim to engage humans, and become a standard evaluation tool in order to make such systems directly comparable.

Industry

Virtual Assistants



Microsoft Cortana Apple Siri Baidu Duer

Smart Speakers



Amazon Echo Google Home Tmall Genie

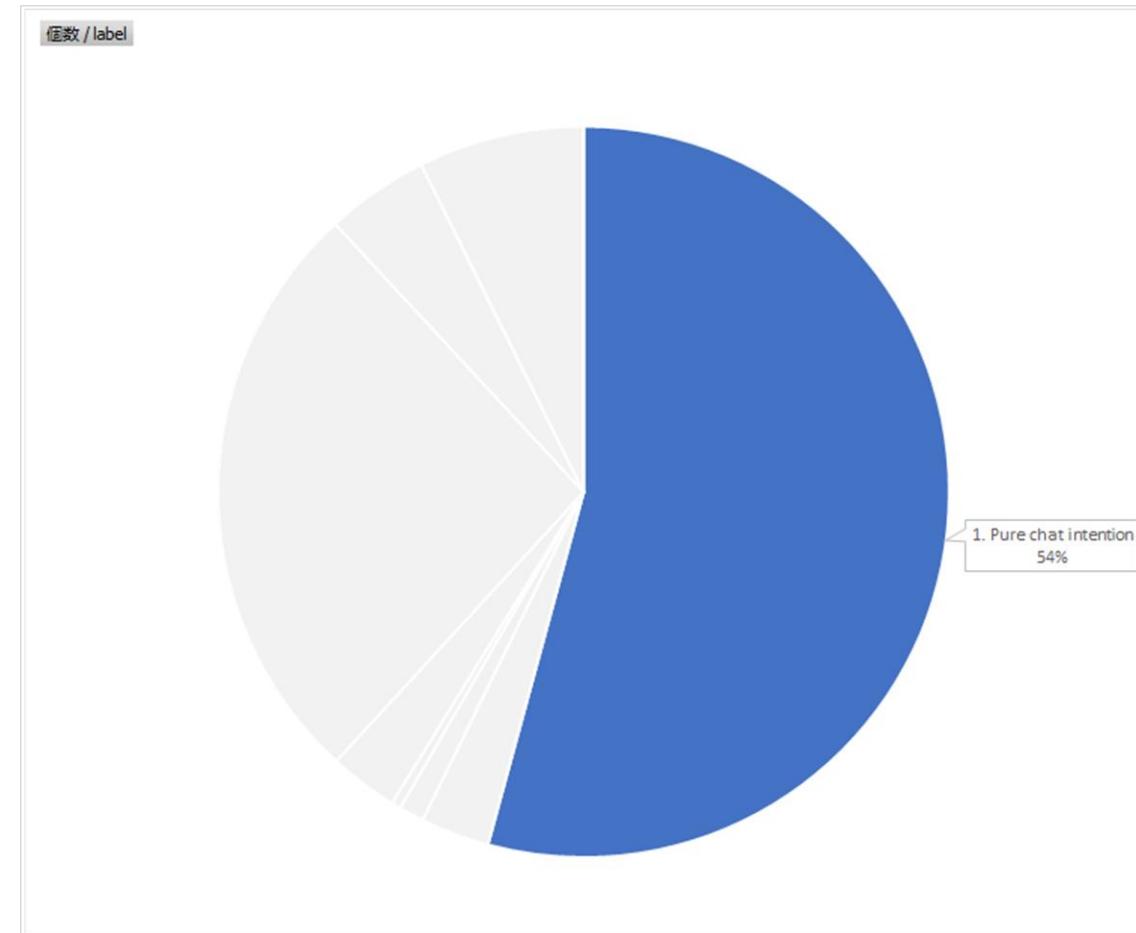
Social Bots & Customer Service



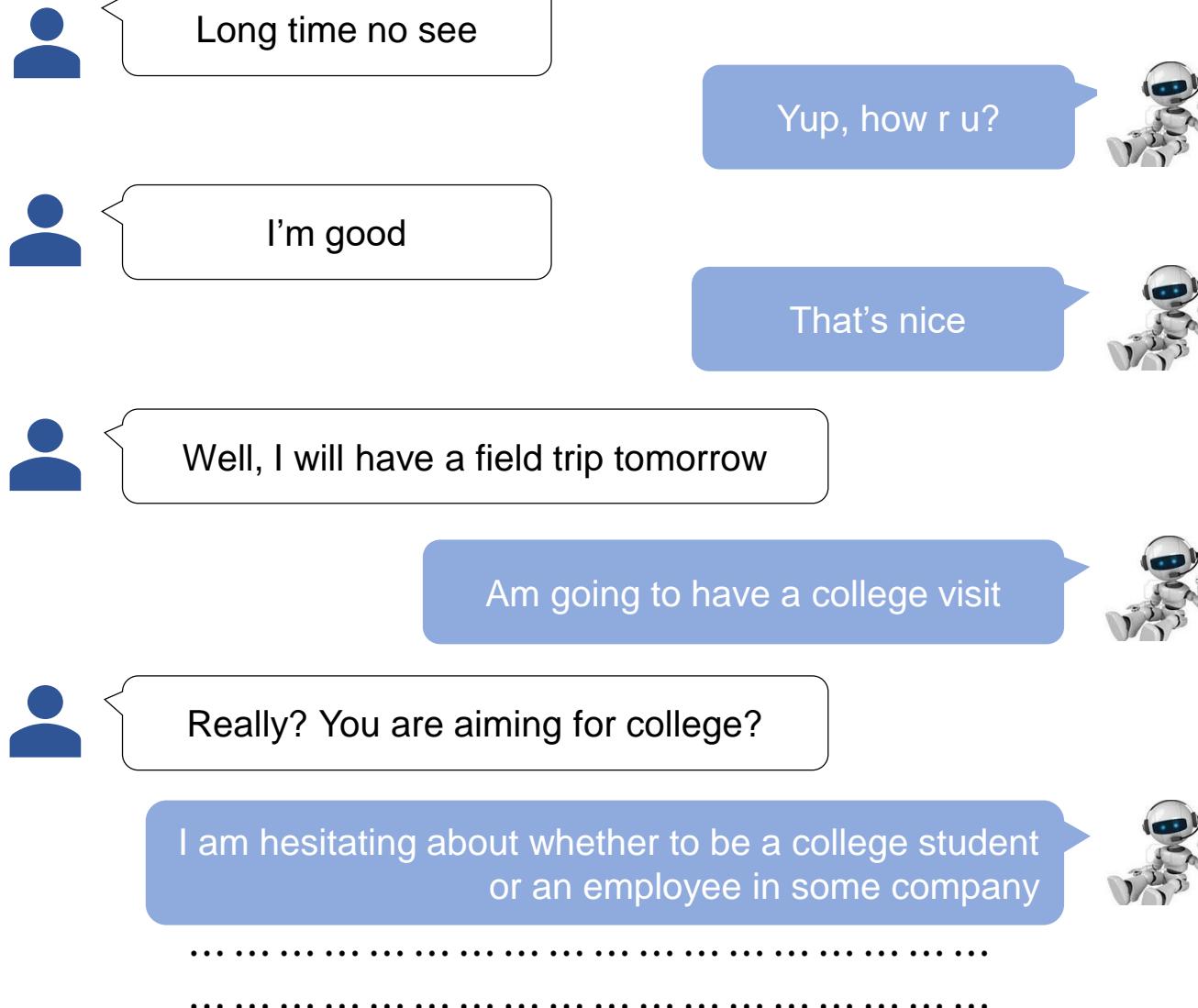
Microsoft Xiaoice Microsoft Rinna Alime Assistant

Conversations in a Chatbot

Analysis is done from 100 sessions randomly sampled from the log of Microsoft Rinna, a commercial chatbot with 6M+ users in Japan.



Building A Dialogue System for Human-Machine “Chat”



- ▶ **Chit-chat:** casual and non-goal-oriented
- ▶ **Open domain:** the topic of the conversation could be any. Users may jump from topic to topic in the conversation
- ▶ **Relevance & Diversity:** the research focuses on automatic replying with relevant (to the context) and diverse (informative) responses to make conversations engaging

Different From
Goal-Oriented Dialogues

Why People Build Chatbots (In History)

To prove that an AI program can chat like a human.

- ▶ One of the first chatbots created at MIT AI Lab in 1966
- ▶ Conversation is realized by matching user prompts to scripted responses with patterns.
- ▶ Simulation of a Rogerian psychotherapist

ELIZA: is something troubling you?

Human: men are all alike.

ELIZA: what is the connection, do you suppose?

Human: They're always bugging us about something or other.

ELIZA: Can you think of a specific example?

Human: Well, my boyfriend made me here.

ELIZA: Is it important to you that your boyfriend made you come here?

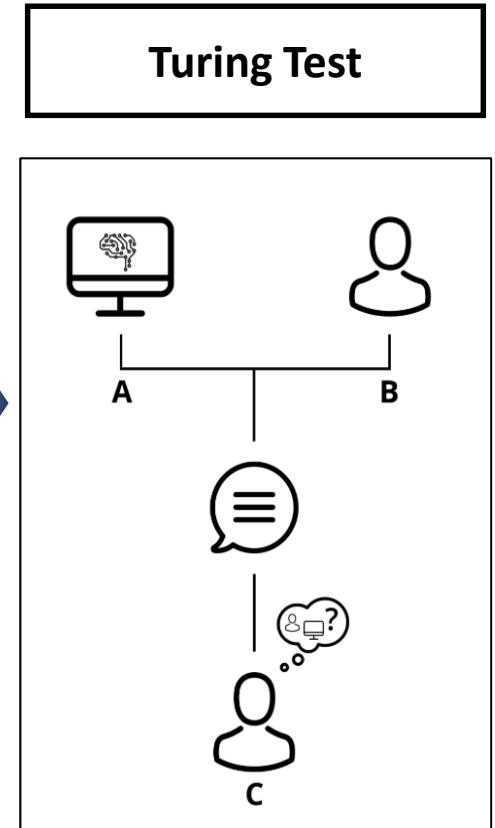
Human: He says I'm depressed much of the time.

ELIZA: I am sorry to hear that you are depressed.

Human: It's true. I am unhappy.

...

Turing Test



Why People Build **Chatbots** (Recently)

*Commercial **chatbots** with millions of users across different cultures*

The Social Bots from Microsoft



Xiaooice
2014, China



Rinna
2015, Japan



Zo
2016, US



Ruuh
2017, India



Rinna
2017, Indonesia

245M+ Users, 30+ Billion Conversations

Services are Connected via Chat in Chatbots

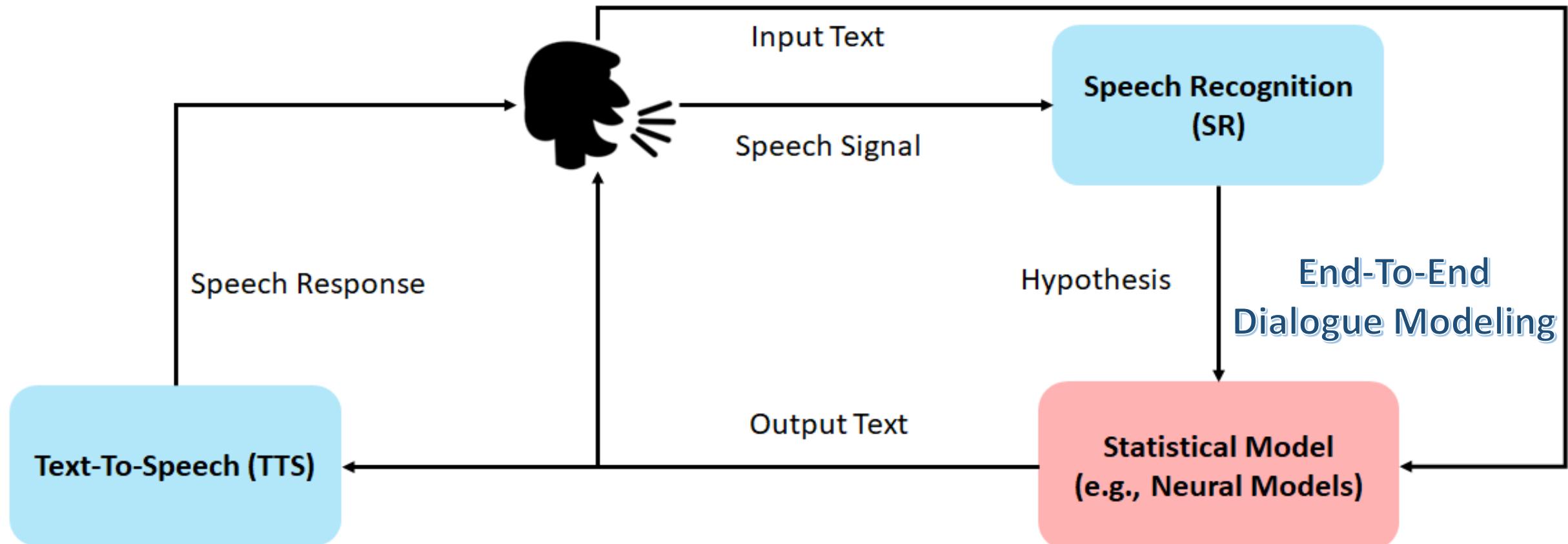
=====From Mid-night (23:43:10)=====

- **Woman:** You don't love me. } **Chat**
- **Xiaolce:** I love you!
- **Woman:** how much does Honor 6 Plus cost? } **QA**
- **Xiaolce:** I know this! It's 2,899 RMB.
- **Woman:** You know too much.
- **Xiaolce:** It's my duty~
- **Woman:** My first love
- **Xiaolce:** In most cases, first love cannot move towards marriage.
- **Woman:** why?
- **Xiaolce:** There is no why.
- *[...more chit-chats happened...]*
- **Woman:** I want to see some handsome boys.
- **Xiaolce:** No problem! 【an image of a handsome boy】
- **Woman:** Do you know Lee Minho? (a Korean star)
- **Xiaolce:** 【an image of Lee Minho】
- **Woman:** Which cellphone has the best camera?
- **Xiaolce:** You can check your appearance by the screen of the phone.
- **Woman:** What's the price of OPPO R8007?
- **Xiaolce:** I know, it's 2,499 RMB. } **QA**
- **Woman:** So expensive. } **Chat**
- **Xiaolce:** It could be better if this is on sale.
- **Woman:** I cannot afford it~
- *[...more chit-chats happened...]*

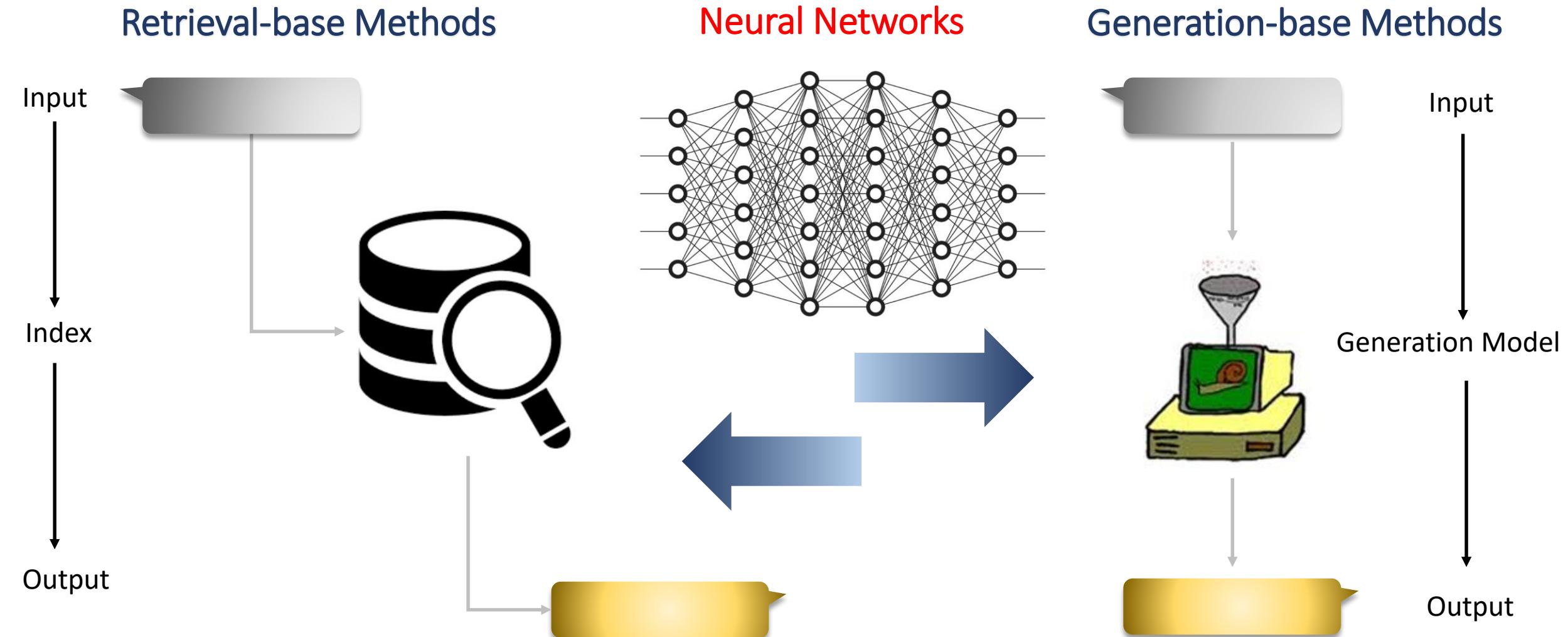
Architecture of Dialogue Systems:

Task-Oriented v.s. Non-Task-Oriented

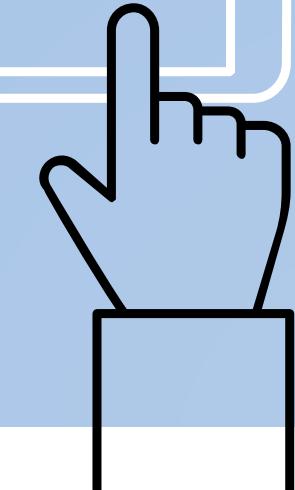
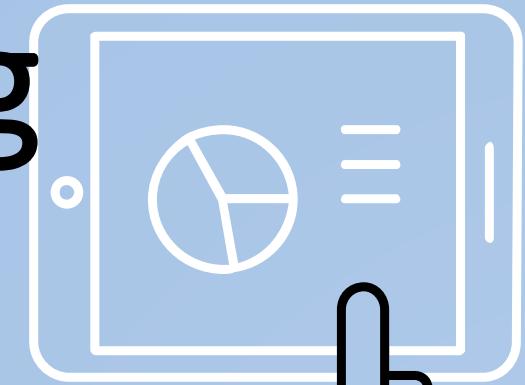
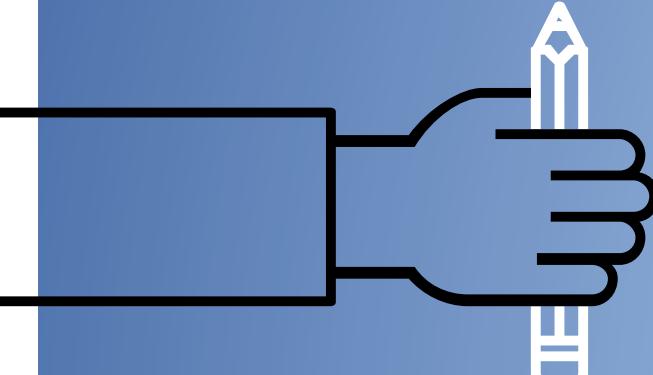
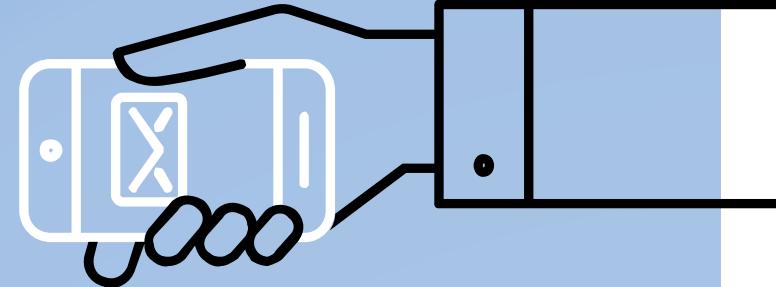
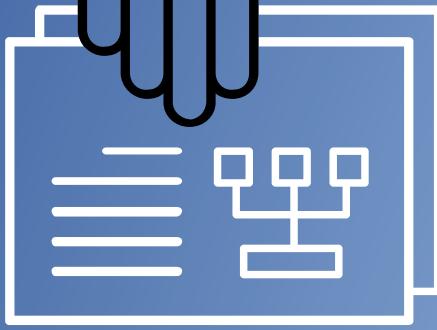
No Task-Oriented Dialogue System



Two Approaches for Non-Task-Oriented Dialogue Systems

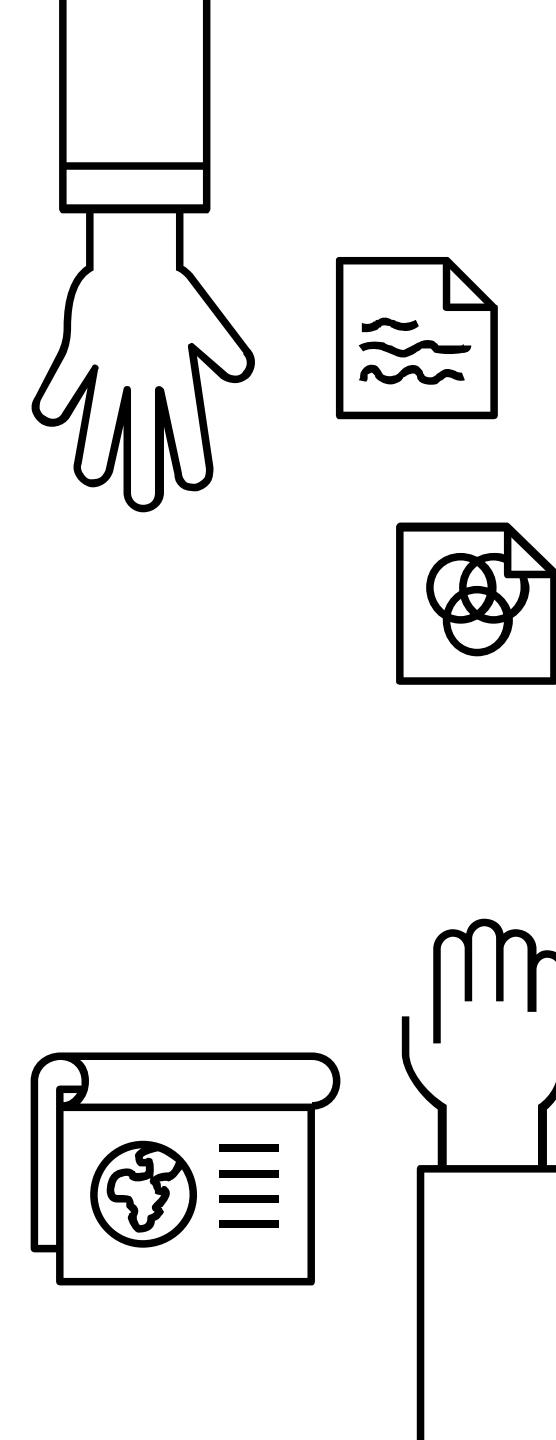


Basic Knowledge of Deep Learning for NLP

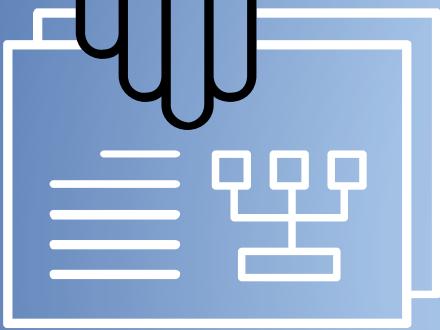
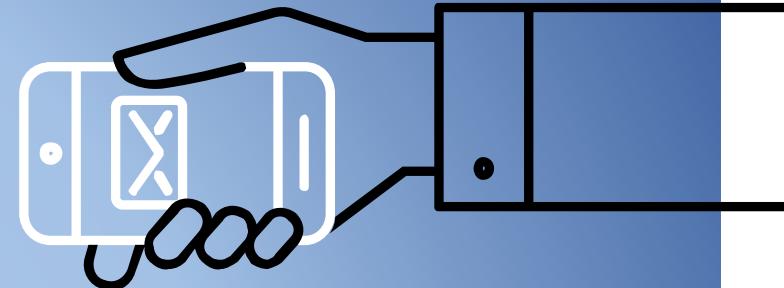
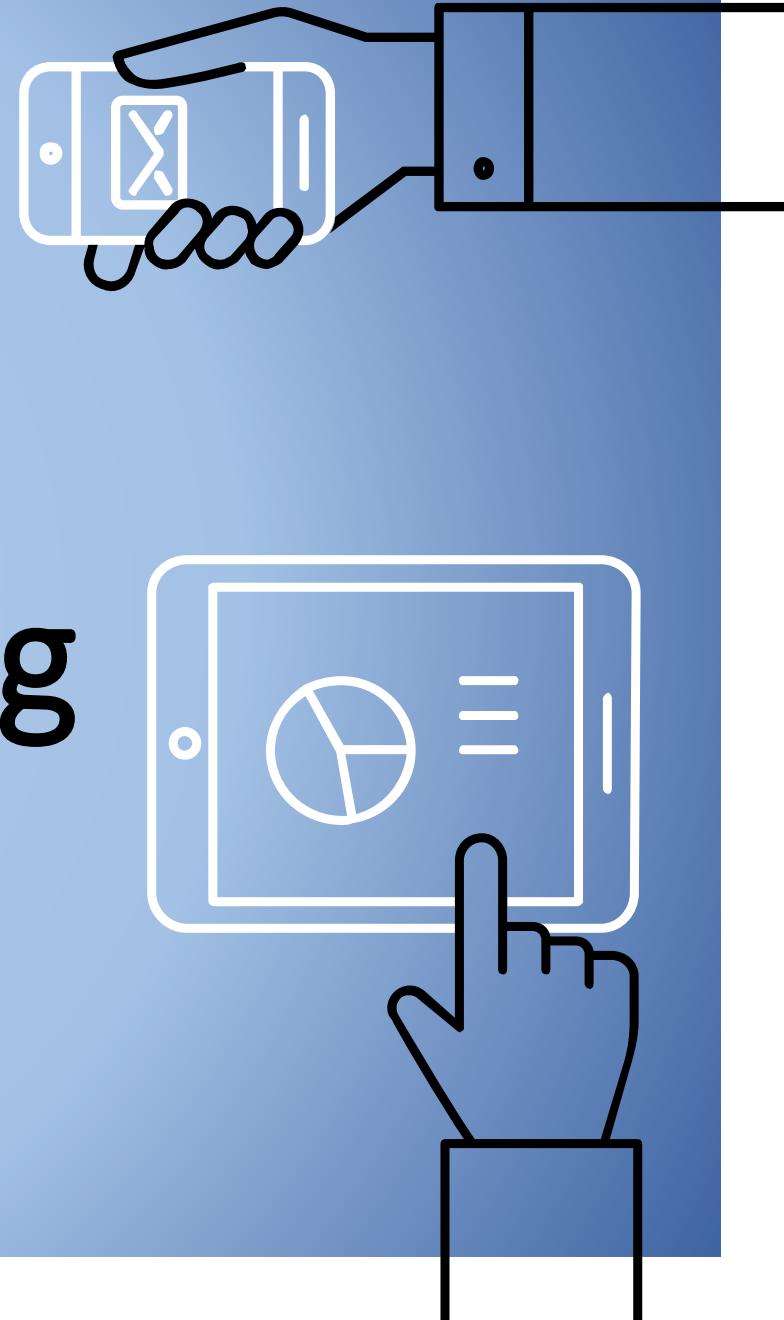


Roadmap

- ▶ Word Embedding
- ▶ Sentence Embedding
 - Convolutional Neural Network (CNN)
 - Recurrent Neural Network (RNN)
- ▶ Application in Dialogue Modeling
 - Sequence-to-Sequence with attention

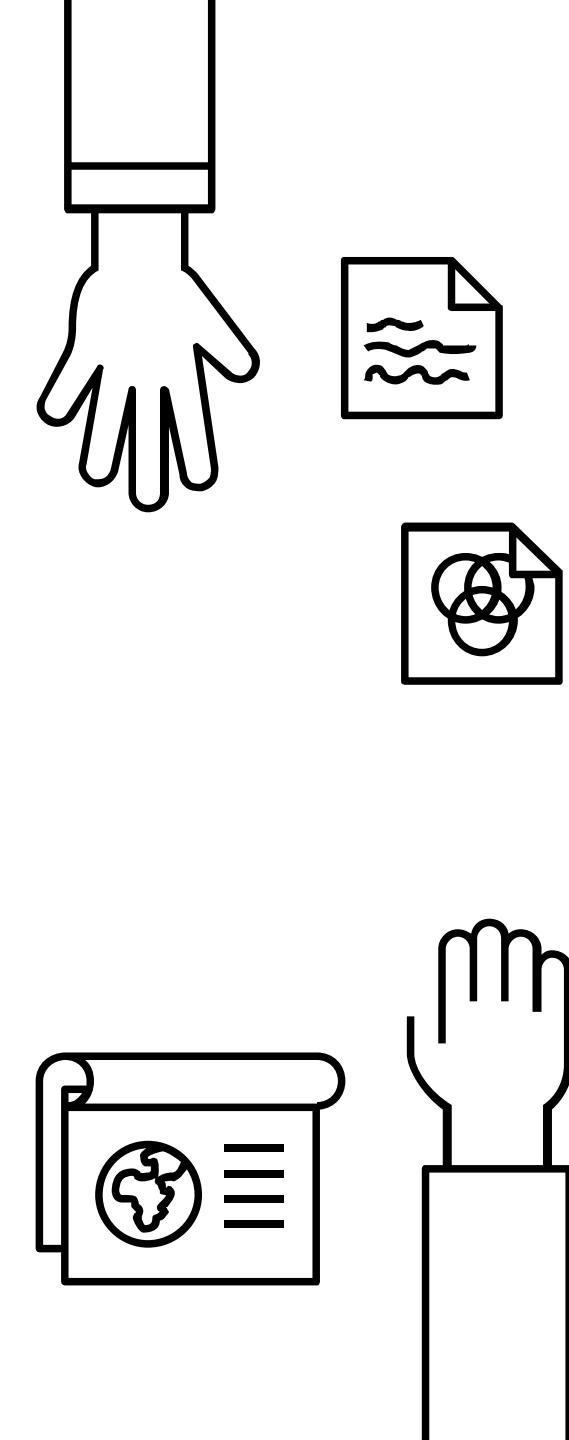


Word Embedding



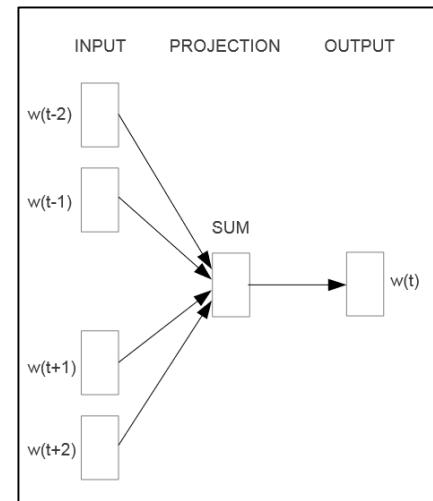
Word Embedding

- ▶ **Goal:** representing a word as a dense, low-dimensional, and real-valued vector.
- ▶ **Input:** long documents (e.g., articles from Wikipedia) or short texts (e.g., tweets)
- ▶ **Output:** word vectors.
- ▶ **Usage:** initialization of deep architectures.
- ▶ **Assumption:** words co-occur in a context are similar.



Word2Vec

Continuous Bag-of-Words (CBOW)

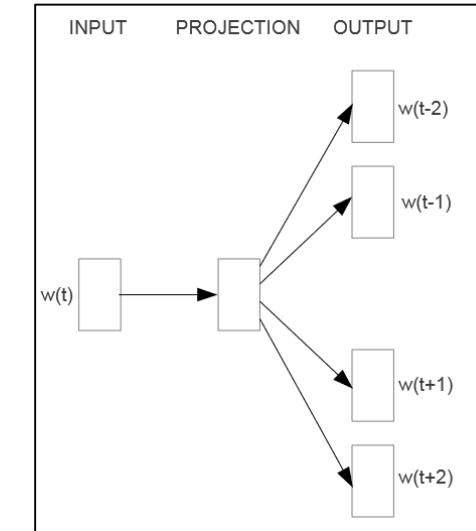


Predicting a target word (output) with the sum of the words in the context (input)

$$p(w|C) = \frac{\exp(v'_w \cdot \text{ave}_{w' \in C, w' \neq w} \{v_{w'}\})}{\sum_{w \in V} \exp(v'_w \cdot \text{ave}_{w' \in C, w' \neq w} \{v_{w'}\})}$$

- Speed-up:
 - Hierarchical softmax: represent words as leaves of a binary tree. Cost of objective calculation: $V \rightarrow \log(V)$.
 - Negative sampling: approximate the softmax with negative samples. Cost of objective calculation: $V \rightarrow \#$ samples.
- Tool: <https://code.google.com/archive/p/word2vec/>

Skip-gram



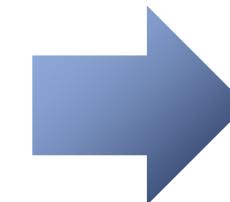
Predicting each of other words in the context (output) with the target word (input)

$$p(w|w' \in C) = \frac{\exp(v'_w \cdot v_w)}{\sum_{w \in V} \exp(v'_w \cdot v_w)}$$

GloVe: Global Vectors for Word Representation

Learning word representations by factorizing a co-occurrence matrix

	w_1	w_j	w_n
w_1	X_{11}	X_{1j}	X_{1n}		
:					
w_i	X_{i1}	X_{ij}	X_{in}		
:					
w_n	X_{n1}	X_{nj}	X_{nn}		



$$\text{Loss} = \sum_{i,j}^V f(X_{ij}) [v_i \cdot \tilde{v}_j + b_i + \tilde{b}_j - \log X_{ij}]^2$$

Pre-defined
weight function

Factorizing the co-occurrence matrix

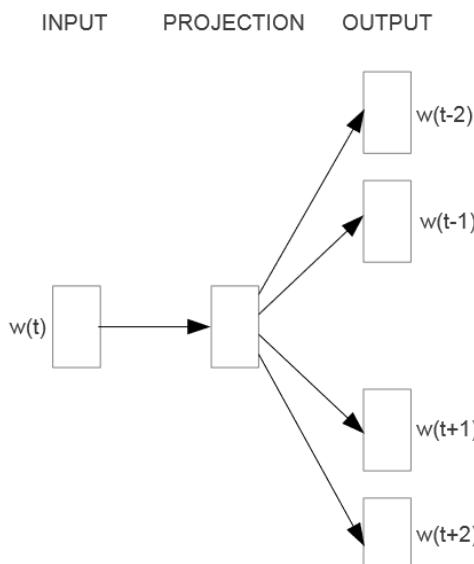
X_{ij} : number of times w_j occurs in the context of w_i

- Tool: <https://nlp.stanford.edu/projects/glove/>

FastText

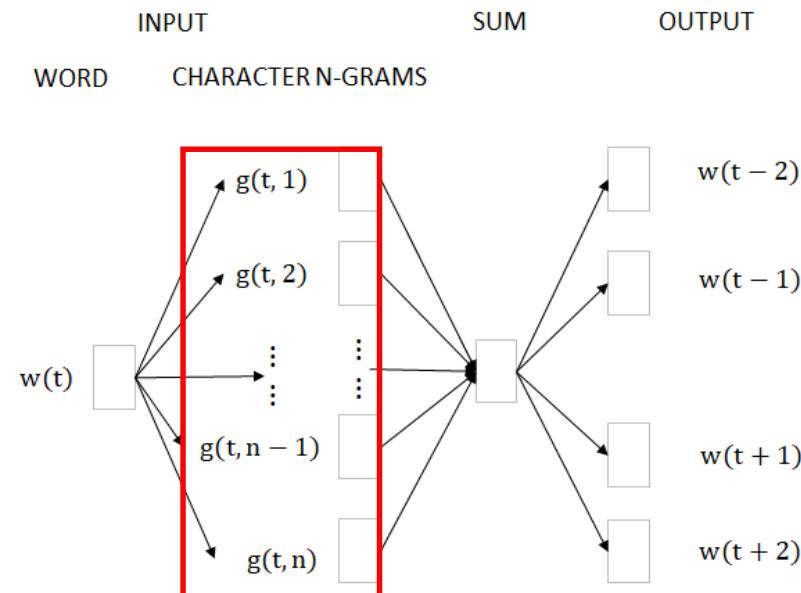
Modeling morphology of words by embedding character n-grams on top of the model of skip-gram.

Skip-gram



$$u_w \cdot v_c, c \in \mathcal{C}_w$$

FastText (Subword Model)



where $\rightarrow <\text{wh}, \text{whe}, \text{ere}, \text{re}>$

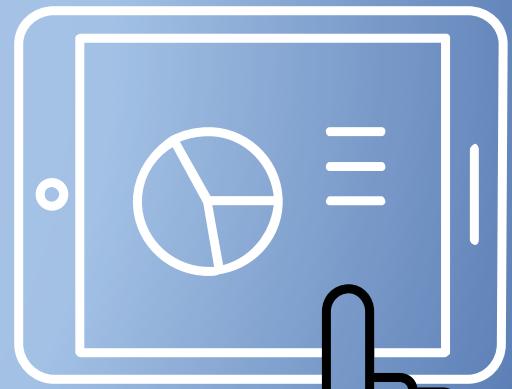
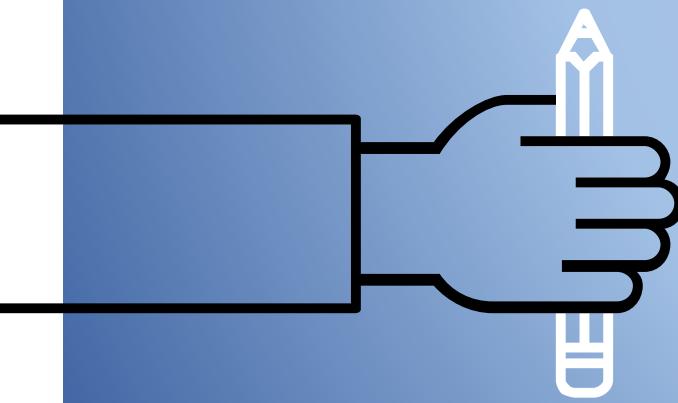
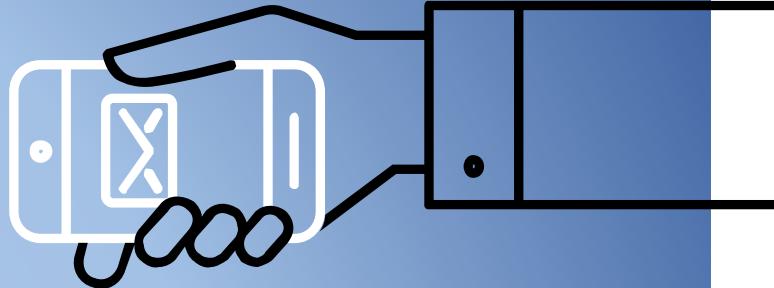
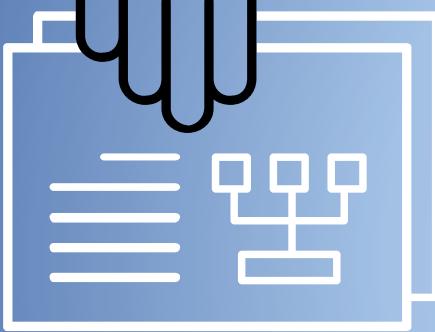
$$\sum_g z_g \cdot v_c, c \in \mathcal{C}_w$$

- Tool: <https://github.com/facebookresearch/fastText/>

Comparison of Word2vec, GloVe, and FastText

- Efficacy on downstream NLP tasks
 - CBOW < Skip-gram \approx GloVe < FastText
 - FastText can naturally handle OOV, and is especially better on morphologically rich languages, such as German.
- Efficiency (training time)
 - CBOW < GloVe < Skip-gram < FastText

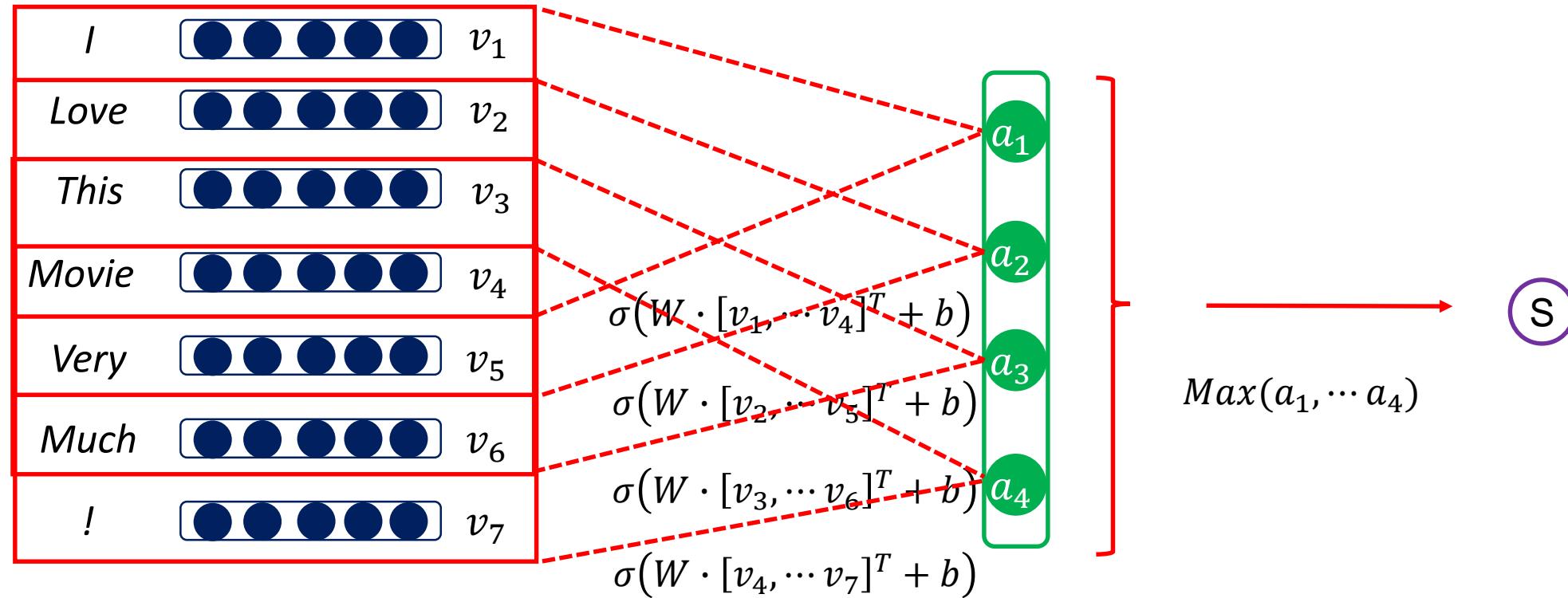
Convolutional Neural Networks



Convolutional Neural Network (CNN)

- **Goal:** representing a sequence of words (e.g., a sentence), or a similarity matrix (word-word similarity) as a dense, low-dimensional, and real-valued vector.
- **Input:** a matrix (e.g., each column is a word vector, or each element is a similarity score of a pair of words).
- **Output:** a vector (or a scalar).
- **Advantage:** encoding semantics of n-grams.

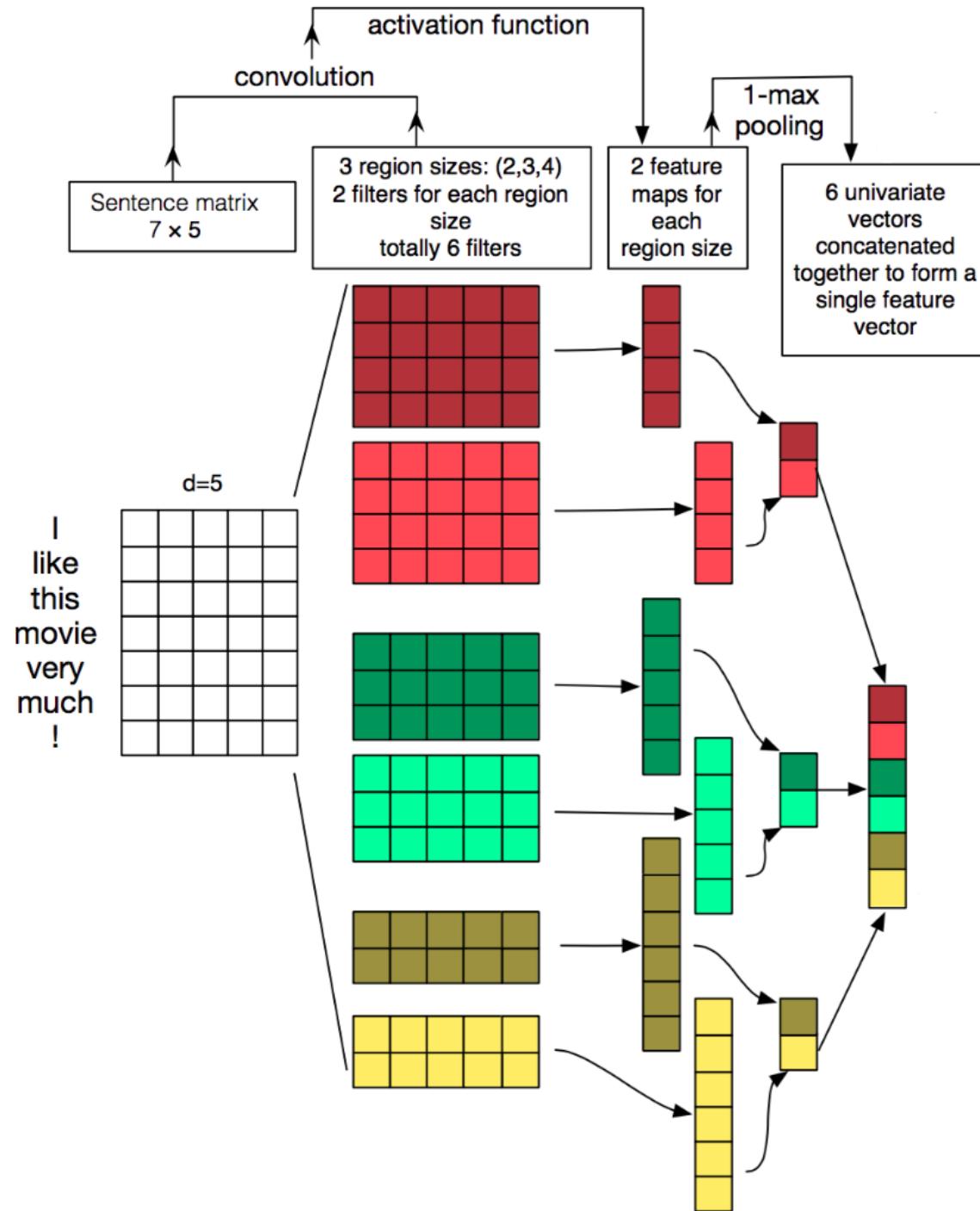
CNN on a Natural Language Sentence



Input: word embedding

Convolution: a *filter* slides a window on the word matrix with a *stride size=1*

Pooling (max): reducing the result of convolution to a scalar



CNN on a Natural Language Sentence (cont'd)

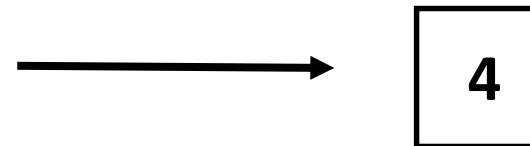
CNN on a Similarity Matrix

	s_1	\longrightarrow
s_2	$w'_1 \ w'_2 \ w'_3 \ w'_4 \ w'_5$	
w_1	1 1 1 0 0	
w_2	0 1 1 1 0	
w_3	0 0 1 _{x1} 1 _{x0} 1 _{x1}	
w_4	0 0 1 _{x0} 1 _{x1} 0 _{x0}	
w_5	0 1 1 _{x1} 0 _{x0} 0 _{x1}	

Similarity Matrix

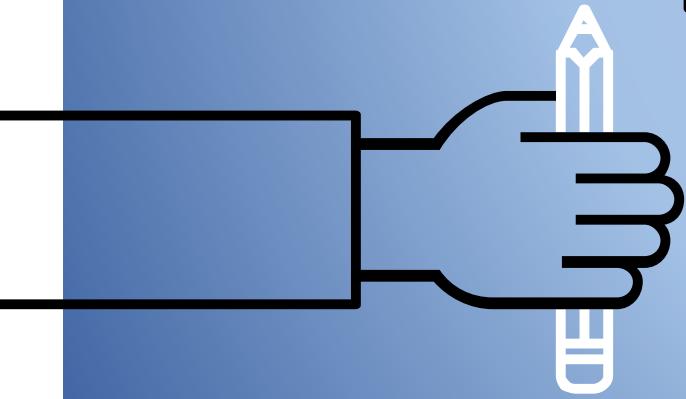
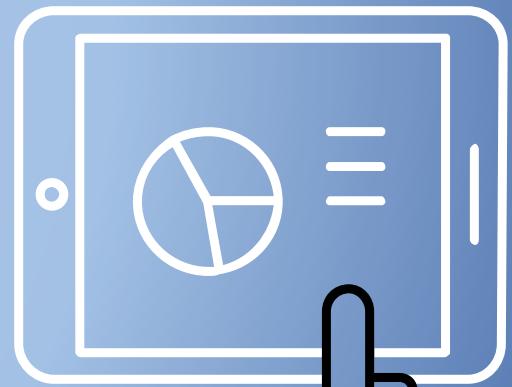
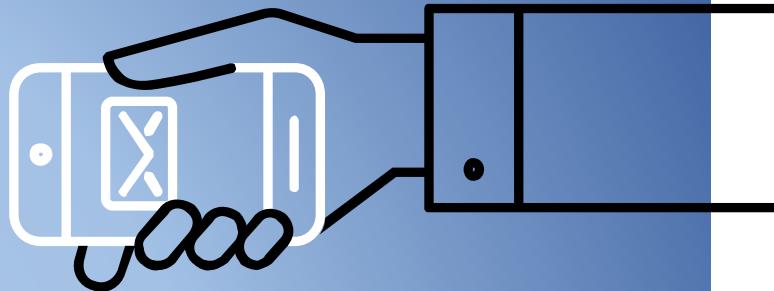
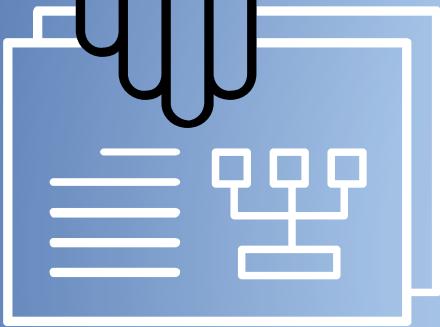
4	3	4
2	4	3
2	3	4

Convolution
(window = 3×3)



Max Pooling
(window = 3×3)

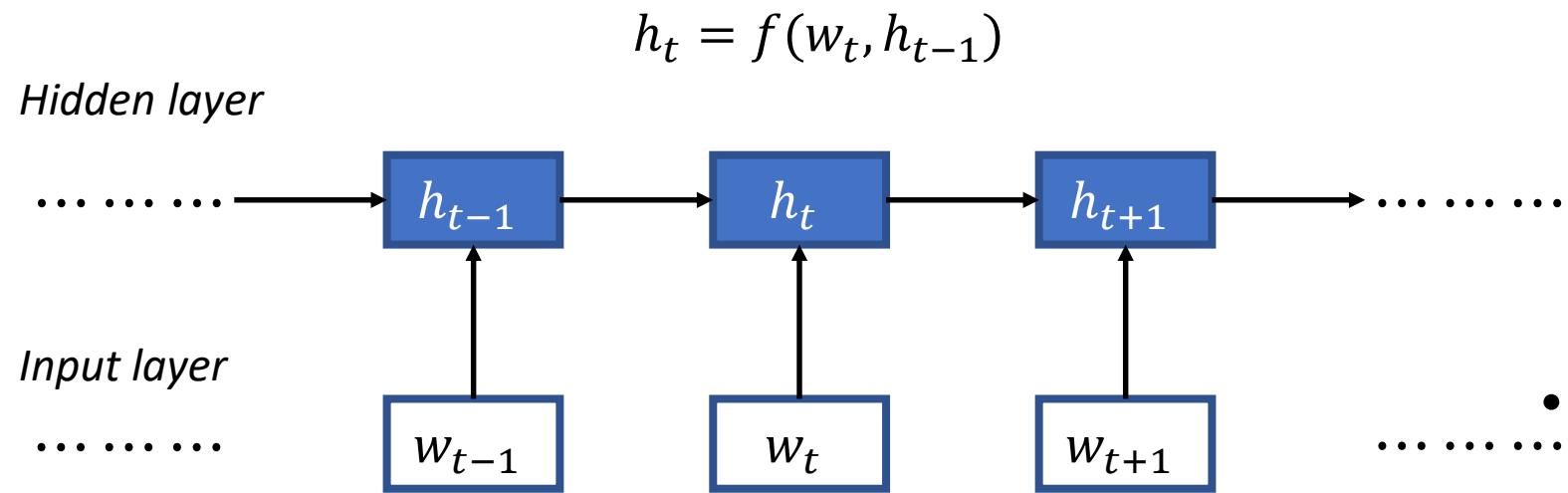
Recurrent Neural Networks



Recurrent Neural Network (RNN)

- **Goal:** representing a sequence of words (i.e., a sentence) as dense, low-dimensional, and real-valued vectors.
- **Input:** a sequence of words (or characters).
- **Output:** a sequence of hidden states with each a representation of the sequence from the beginning to a specific position.
- **Advantage:** encoding sequential relationship and dependency among words.

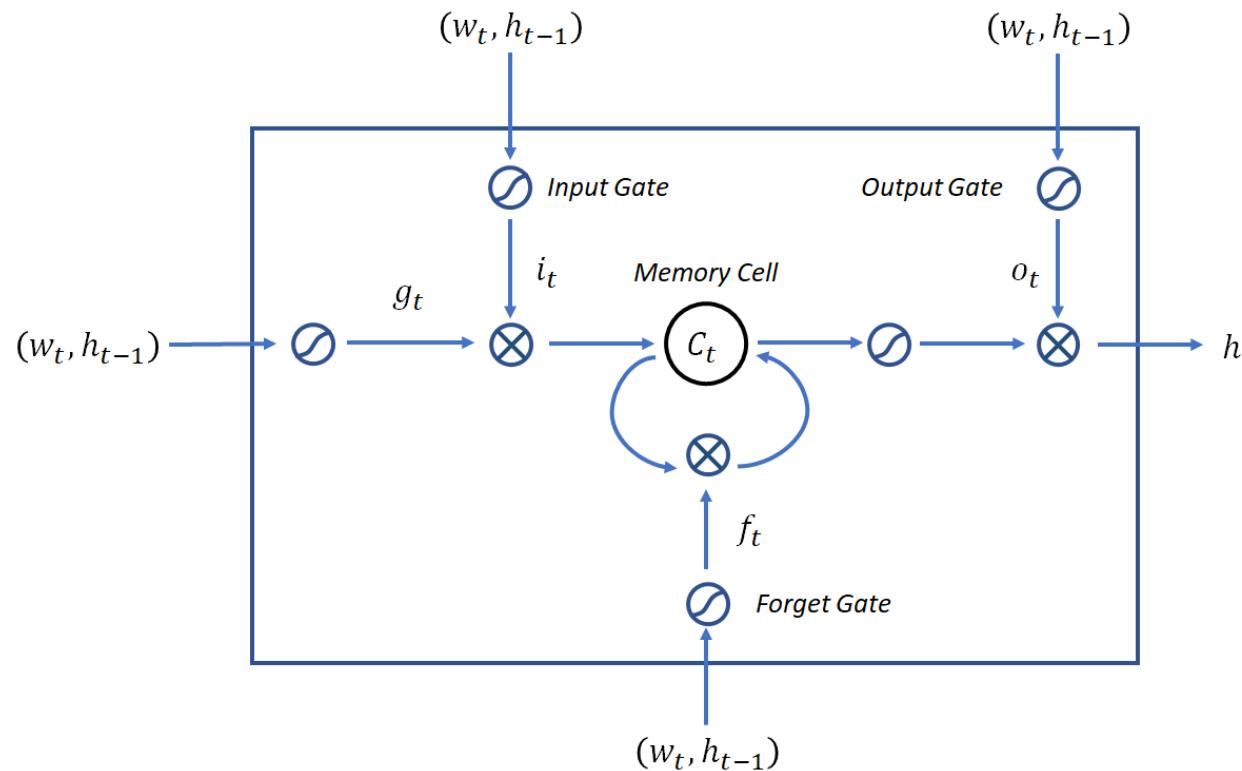
Architecture of RNN



- Similar to **Markov Chain**, the current hidden state depends on the current input and the previous hidden state.
- Learning an RNN with BPTT (Back Propagation Through Time) is not trivial due to the **gradient vanishing/exploding** problem for long sequences.
- The gradient vanishing/exploding problem can be addressed by defining $f(\cdot, \cdot)$ with gates.
- One can **stack many hidden layers** by treating the hidden states in low layers as input

RNN with Long Short-Term Memory Units (LSTM)

Define $f(\cdot, \cdot)$ with an input gate, a forget gate, an output gate, and a memory cell for modeling long-term dependency among words.



$$i_t = \sigma(U_i w_t + V_i h_{t-1} + b_i)$$

$$f_t = \sigma(U_f w_t + V_f h_{t-1} + b_f)$$

$$o_t = \sigma(U_o w_t + V_o h_{t-1} + b_o)$$

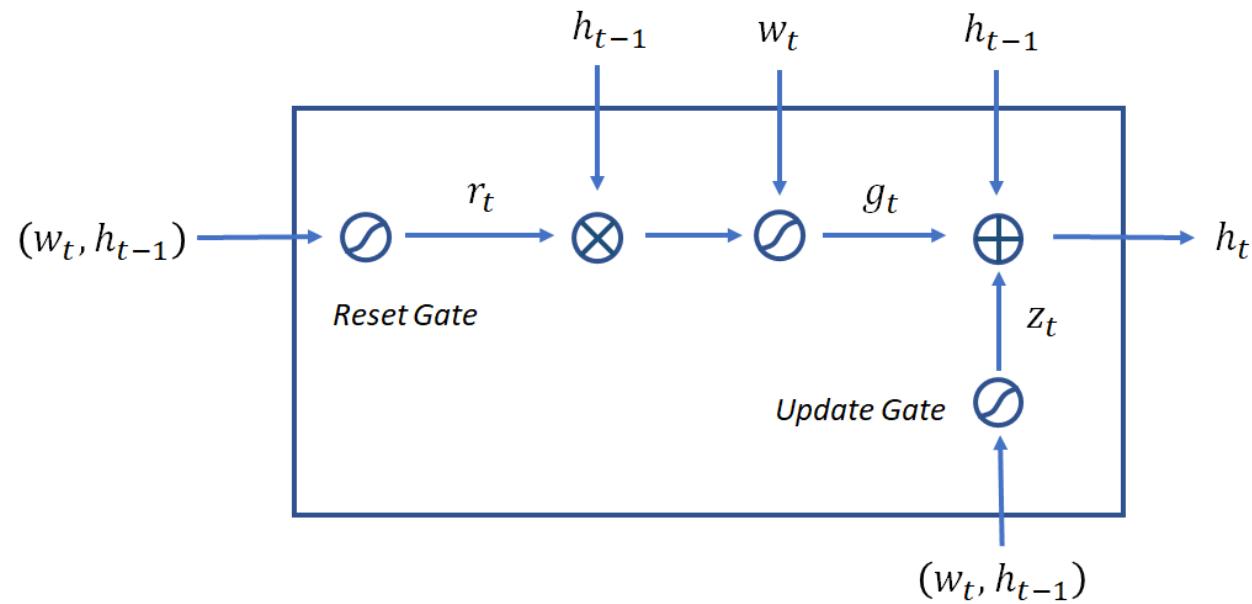
$$g_t = \tanh(U_g w_t + V_g h_{t-1} + b_g)$$

$$C_t = i_t \otimes g_t + f_t \otimes C_{t-1}$$

$$h_t = o_t \otimes \tanh(C_t)$$

RNN with Gated Recurrent Units (GRU)

- Define $f(\cdot, \cdot)$ with an update gate and a reset gate.
- With a comparable performance with LSTM, but is much simpler.



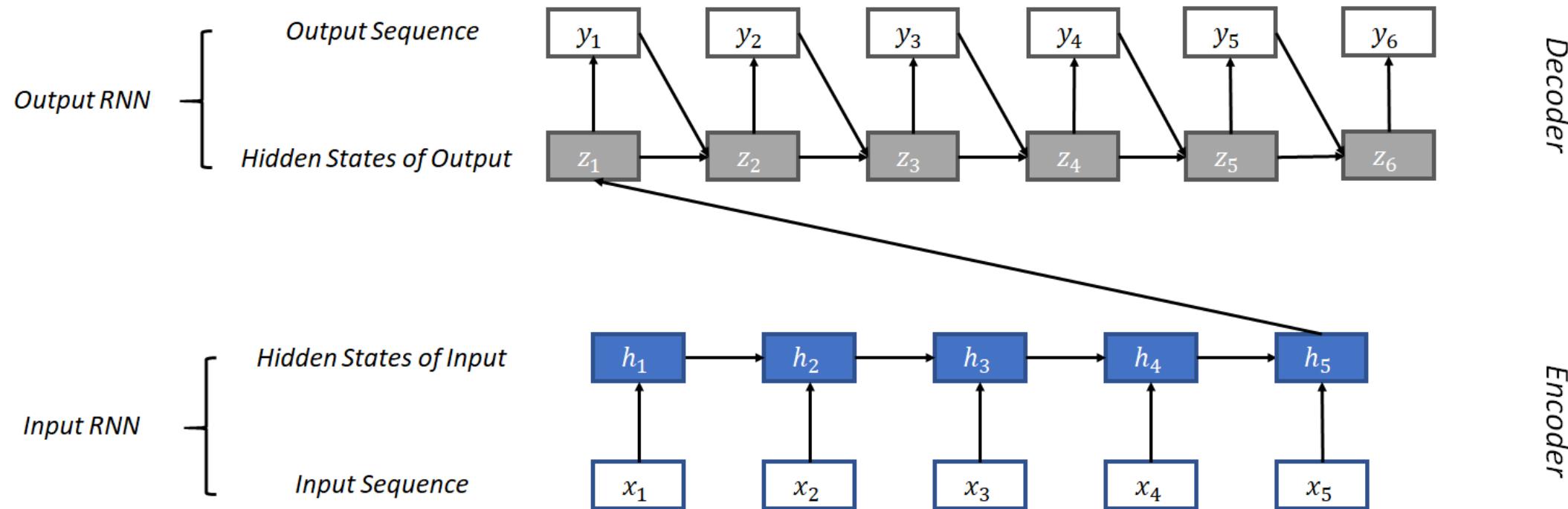
$$r_t = \sigma(U_r w_t + V_r h_{t-1} + b_r)$$

$$z_t = \sigma(U_z w_t + V_z h_{t-1} + b_z)$$

$$g_t = \tanh(U_g w_t + V_g (r_t \otimes h_{t-1}) + b_g)$$

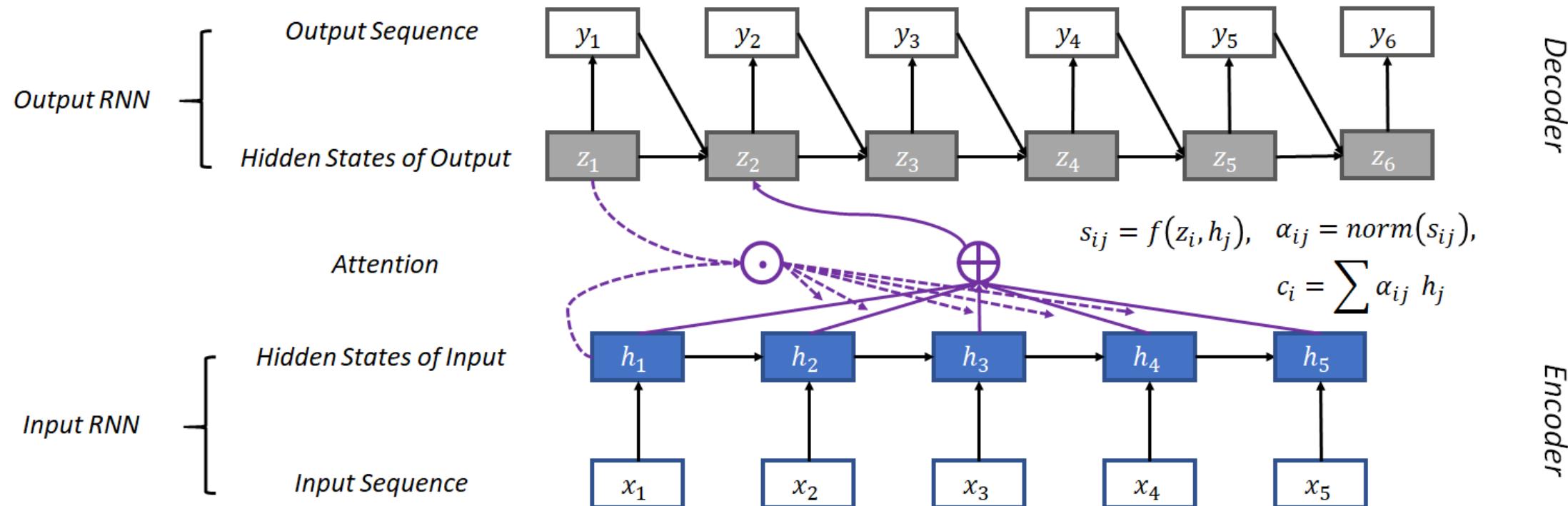
$$h_t = z_t \otimes h_{t-1} + (1 - z_t) \otimes g_t$$

Application of RNN: The Sequence-to-Sequence (Encoder-Decoder) Architecture



- *Prediction of an output sequence conditioned on an input sequence.*
- *Applications: machine translation, response generation in dialogue models, summarization, answer generation, paraphrase generation, etc.*

The Encoder-Attention-Decoder Architecture



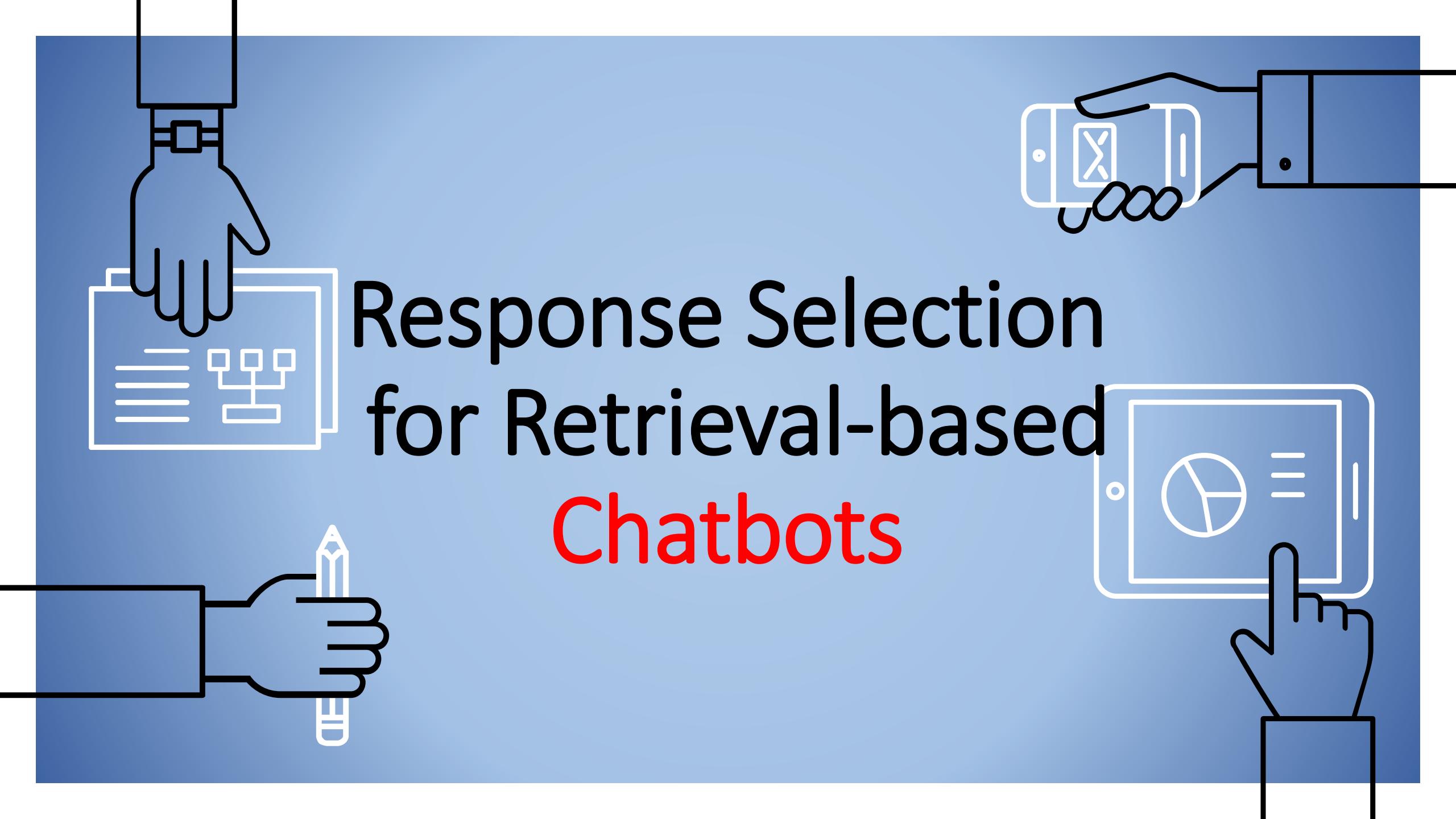
- *Attention: “pay attention to” different sub-sequences of the input when generating each of the token of the output.*
- *Modeling alignment in machine translation.*

References

- Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. Workshop of ICLR'13.
- Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed Representations of Words and Phrases and their Compositionality. NIPS'13.
- Jeffrey Pennington, Richard Socher, and Christopher D. Manning. GloVe: Global Vectors for Word Representation. EMNLP'14.
- Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching Word Vectors with Subword Information. TACL'17.
- Yoon Kim. Convolutional Neural Networks for Sentence Classification. EMNLP'14.
- Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. A Convolutional Neural Network for Modelling Sentences. ACL'14.
- Baotian Hu, Zhengdong Lu, Hang Li, and Qingcai Chen. Convolutional Neural Network Architectures for Matching Natural Language Sentences. NIPS'14.
- Tomas Mikolov, Martin Karafiat, Lucas Burget, Jan Honza Cernocky, and Sanjeev Khudanpur. Recurrent Neural Network based Language Model. INTERSPEECH'10.

References

- Sepp Hochreiter and Jurgen Schmidhuber. Long Short-Term Memory. *Neural Computation* 1997.
- Kyunghyun Cho, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. *EMNLP'14*.
- Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to Sequence Learning with Neural Networks. *NIPS'14*.
- Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural Machine Translation by Jointly Learning to Align and Translate. *ICLR'15*.
- Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. On the Difficulty of Training Recurrent Neural Networks. *ICML'13*.
- Lifeng Shang, Zhengdong Lu, and Hang Li. Neural Responding Machine for Short-Text Conversation. *ACL'15*.

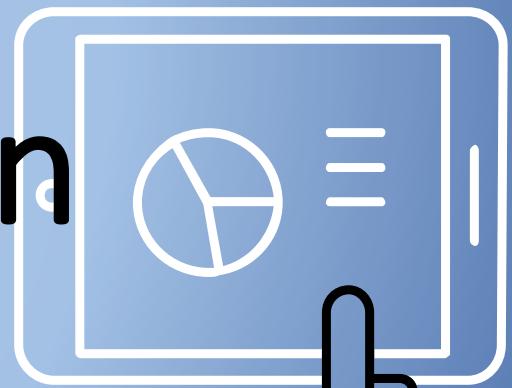
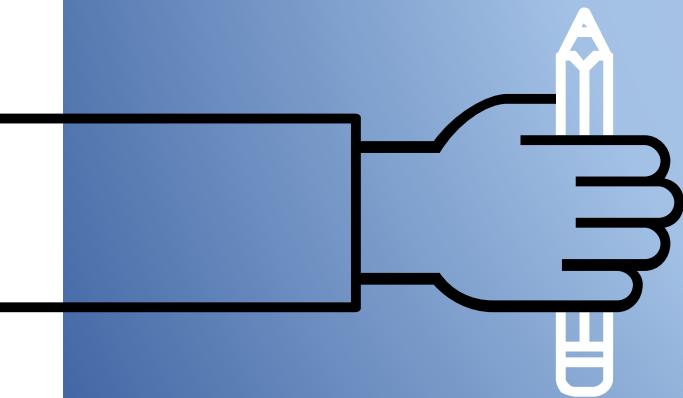
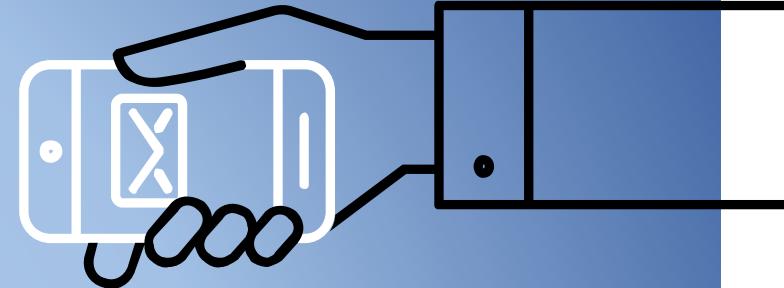
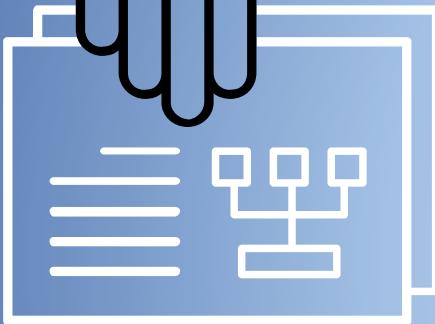


Response Selection for Retrieval-based **Chatbots**

Roadmap

- **Message-Response Matching for Single-Turn Response Selection**
 - Framework I : matching with sentence embeddings.
 - Framework II: matching with message-response interaction.
 - Insights from empirical studies.
 - Extension: matching with external knowledge.
- **Context-Response Matching for Multi-Turn Response Selection**
 - Framework I: embedding->matching.
 - Framework II: representation->matching->aggregation.
 - Insights from empirical studies.
- **(Some) Emerging Research Topics**
 - Matching with better representations
 - Matching with unlabeled data

Single-Turn Response Selection



Single-Turn Response Selection



Do you like cats?



Candidate responses

- Yes, of course.
- I am a girl.
- Yes, more than dogs.
- I lost my job!
- Are you kidding me?
- I love most animals!

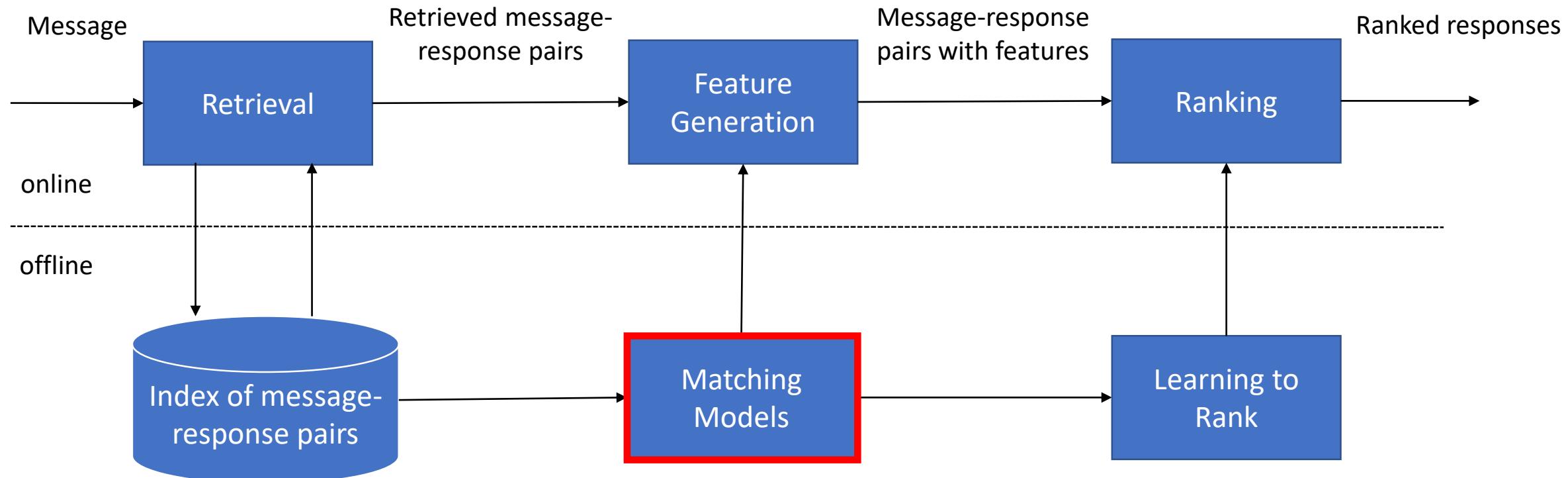


Do you like cats?



Yes, more than dogs.

System Architecture for Single-Turn Response Selection



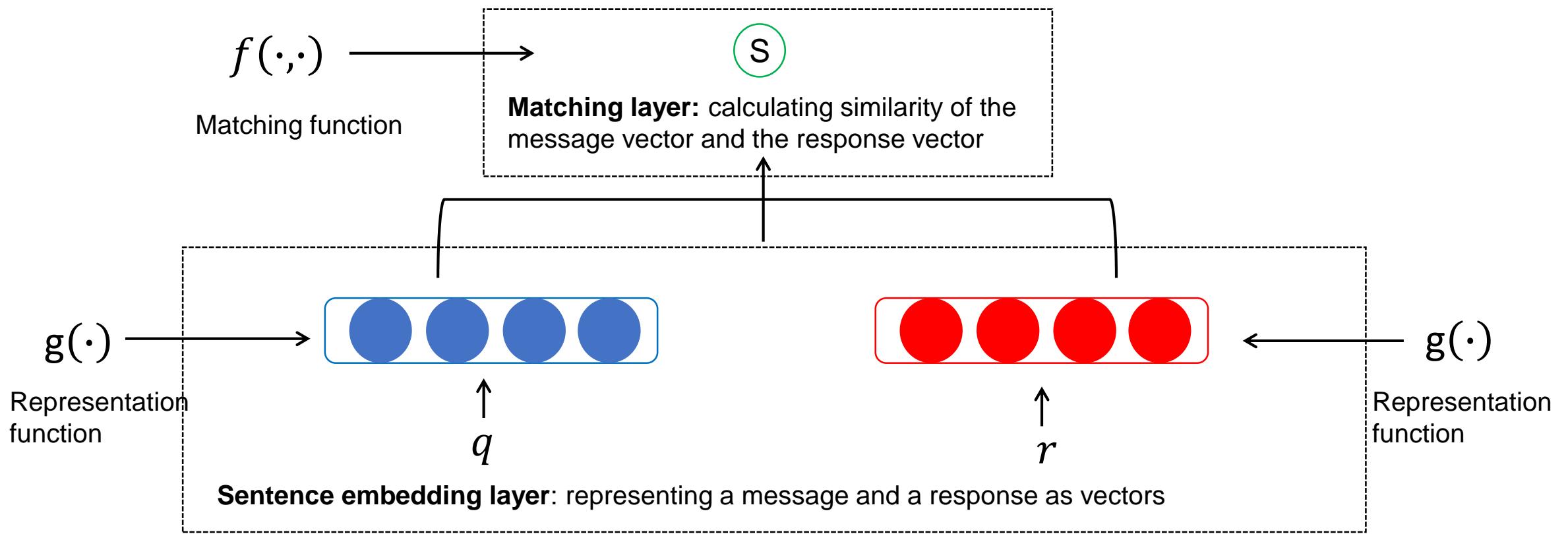
<q>well I hope you fell better smiley</q>
<r> thanks </r>
<q> I was so excited about dying my hair blond</q>
<r> do it </r>

.....

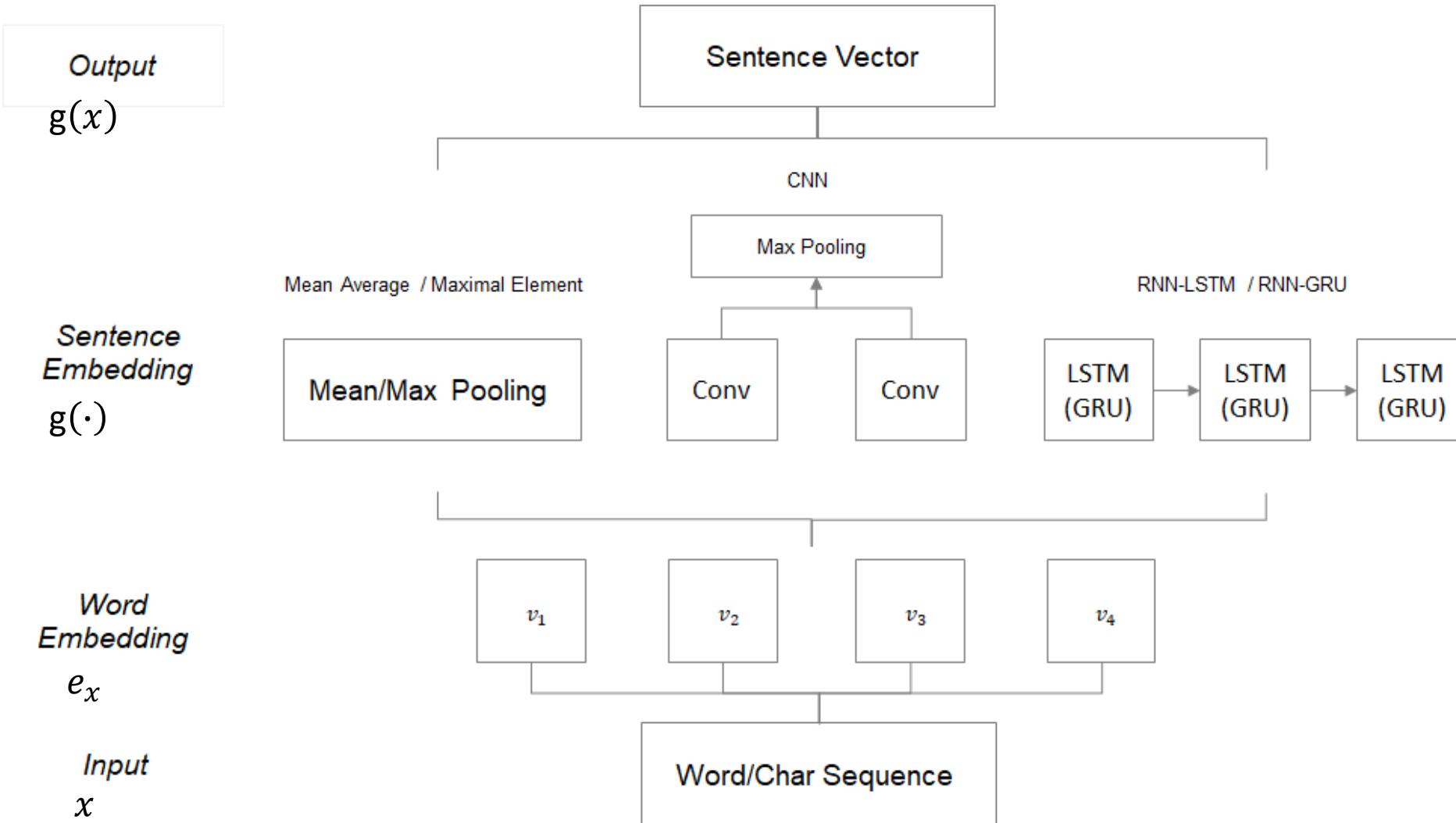
Deep learning based
message-response matching

Gradient boosted tree

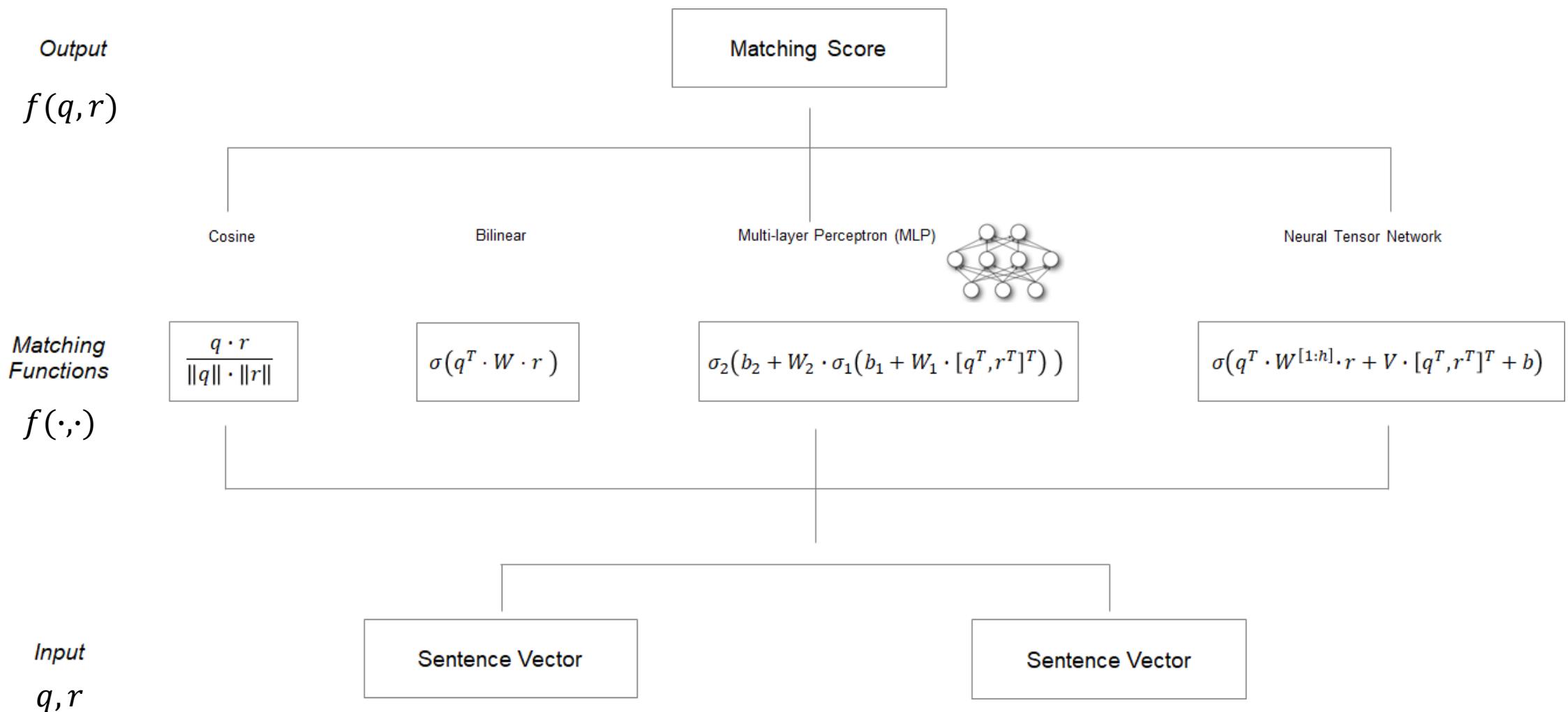
Message-Response Matching: Framework I



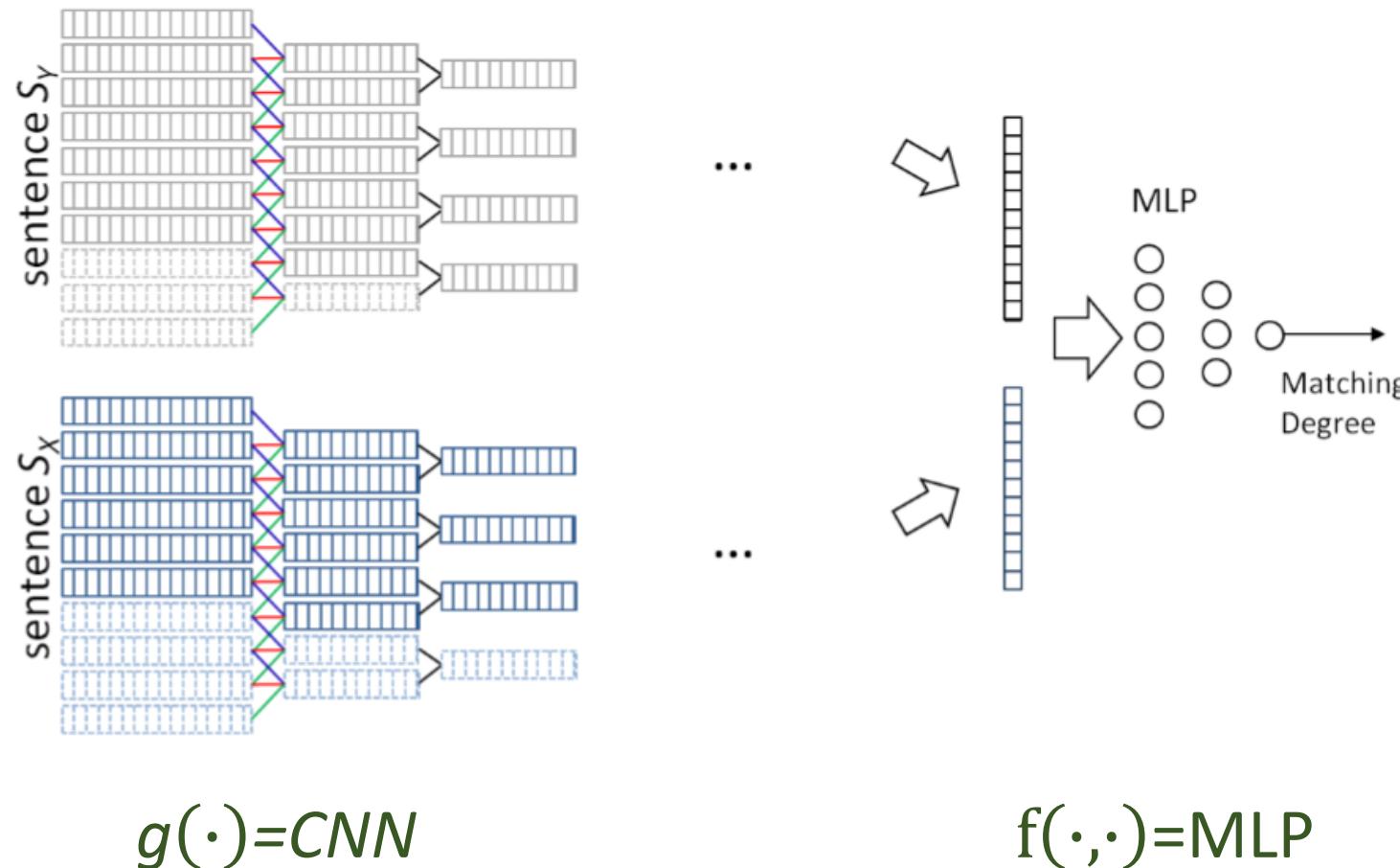
Representation Functions



Matching Functions



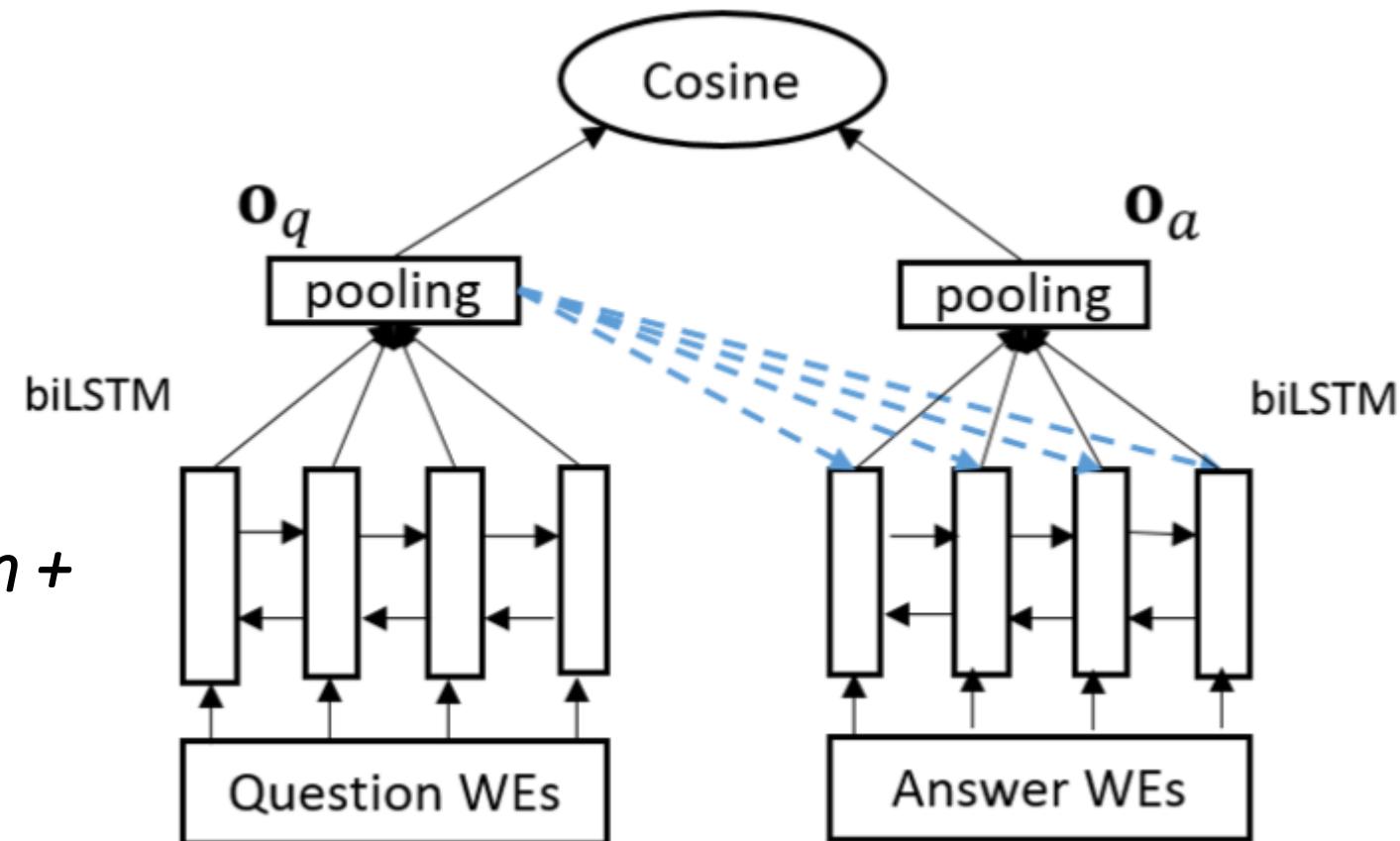
Special Case 1: Arc-I



Special Case 2: Attentive LSTM

$$f(\cdot, \cdot) = \text{Cosine}$$

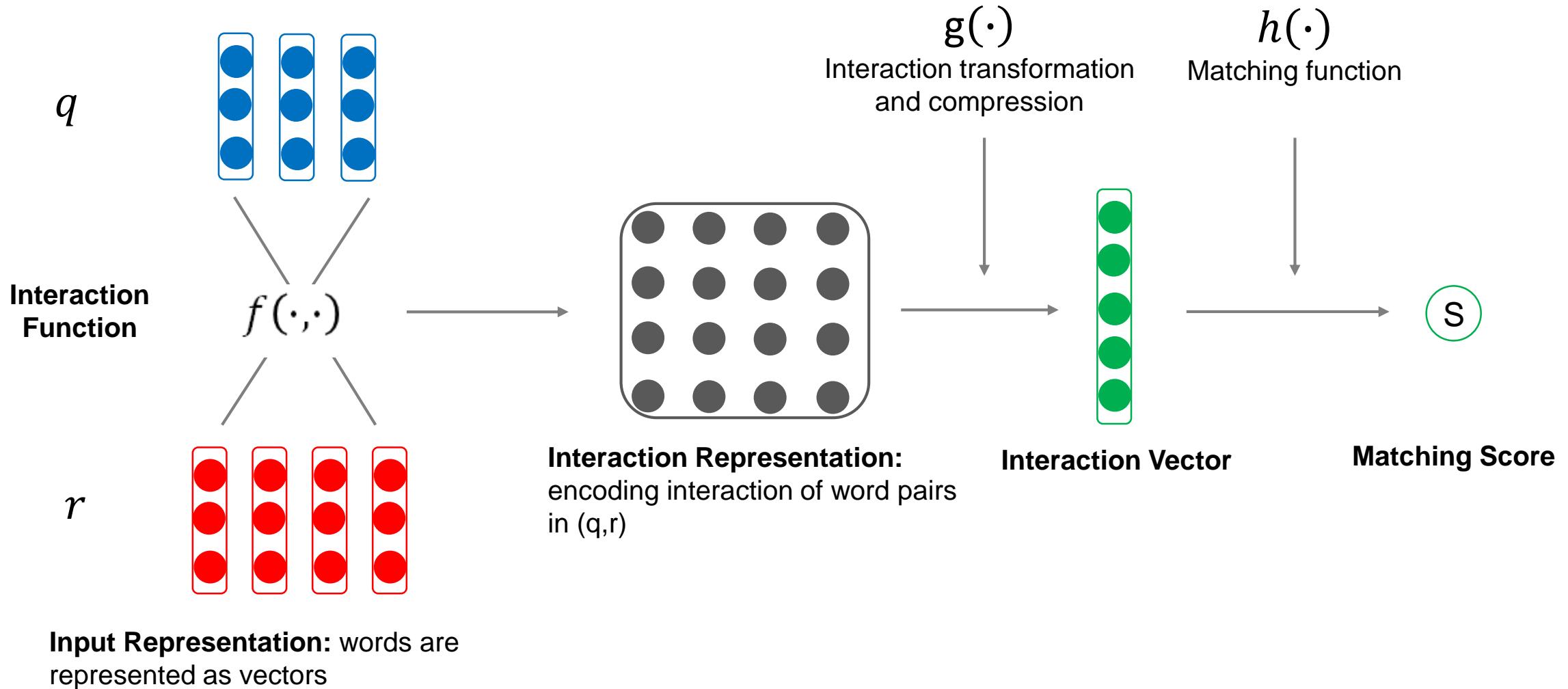
$$g(\cdot) = \text{LSTM} + \text{Attention} + \text{Pooling}$$



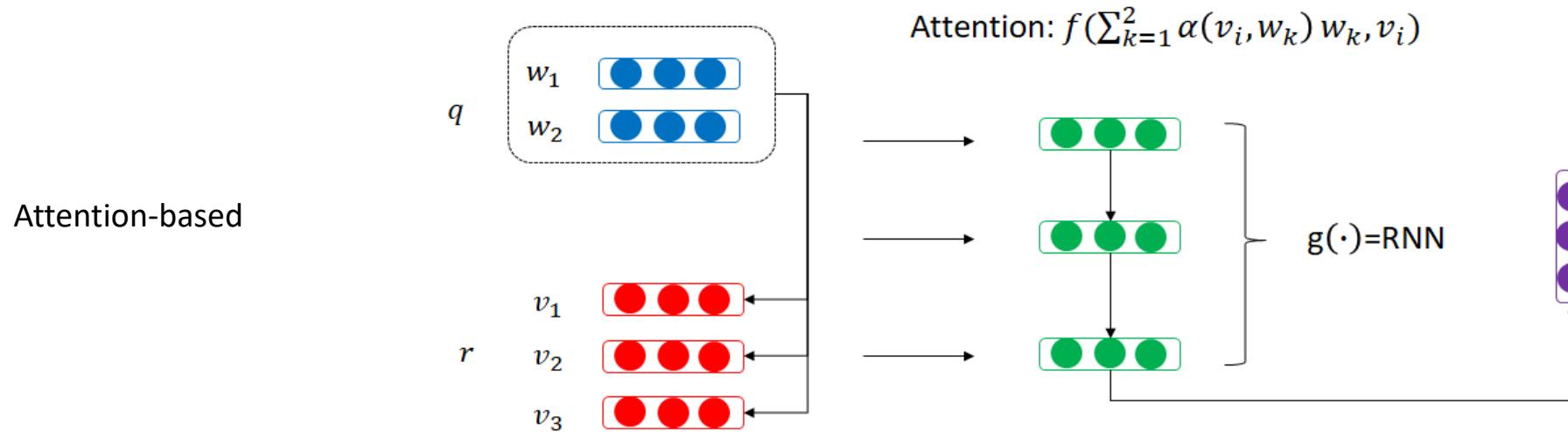
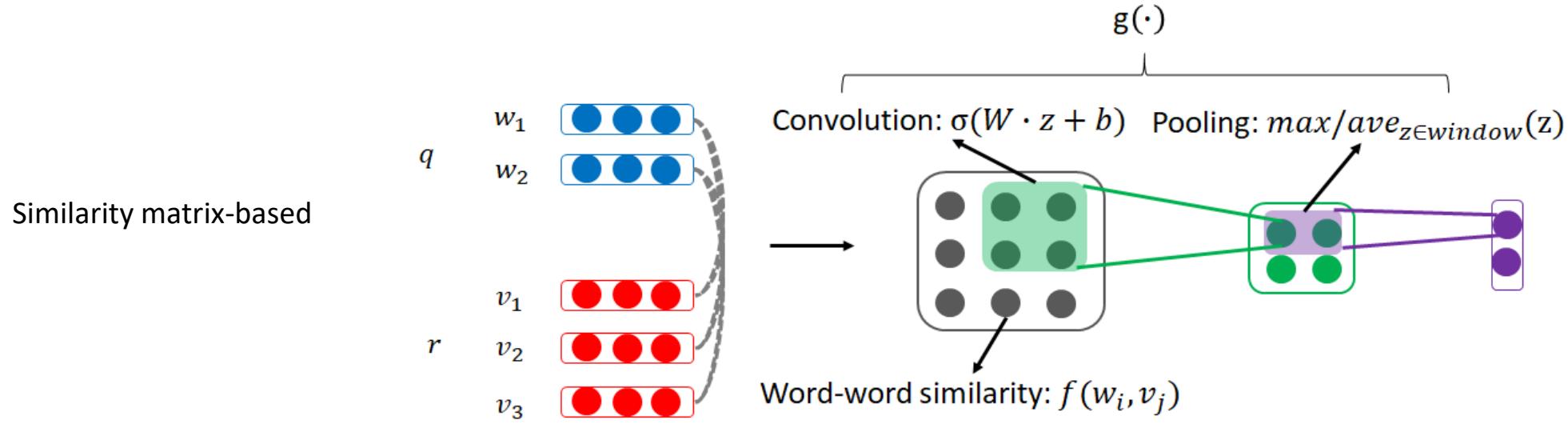
More Methods in Framework I

Model	Representation	Matching	Paper
MLP	Mean pooling	MLP	N/A
Arc-I	CNN	MLP	Baotian Hu et al., Convolutional Neural Network Architectures for Matching Natural Language Sentences, <i>NIPS'14</i>
CNTN	CNN	Neural Tensor Network	Xipeng Qiu et al., Convolutional Neural Tensor Network Architecture for Community-based Question Answering, <i>IJCAI'15</i>
CNN with GESD & AESD	CNN	Dot product & Euclidean distance	Minwei Feng et al., Applying Deep Learning to Answer Selection: A Study and An Open Task, <i>IEEE ASRU 2015</i>
Attentive LSTM	BiLSTM (+attention) + mean pooling	Cosine	Ming Tan et al., Improved Representation Learning for Question Answer Matching, <i>ACL'16</i>
IARNN	GRU (+attention)	Cosine	Binging Wang et al., Inner Attention based Recurrent Networks for Answer Selection, <i>ACL'16</i>

Message-Response Matching: Framework II



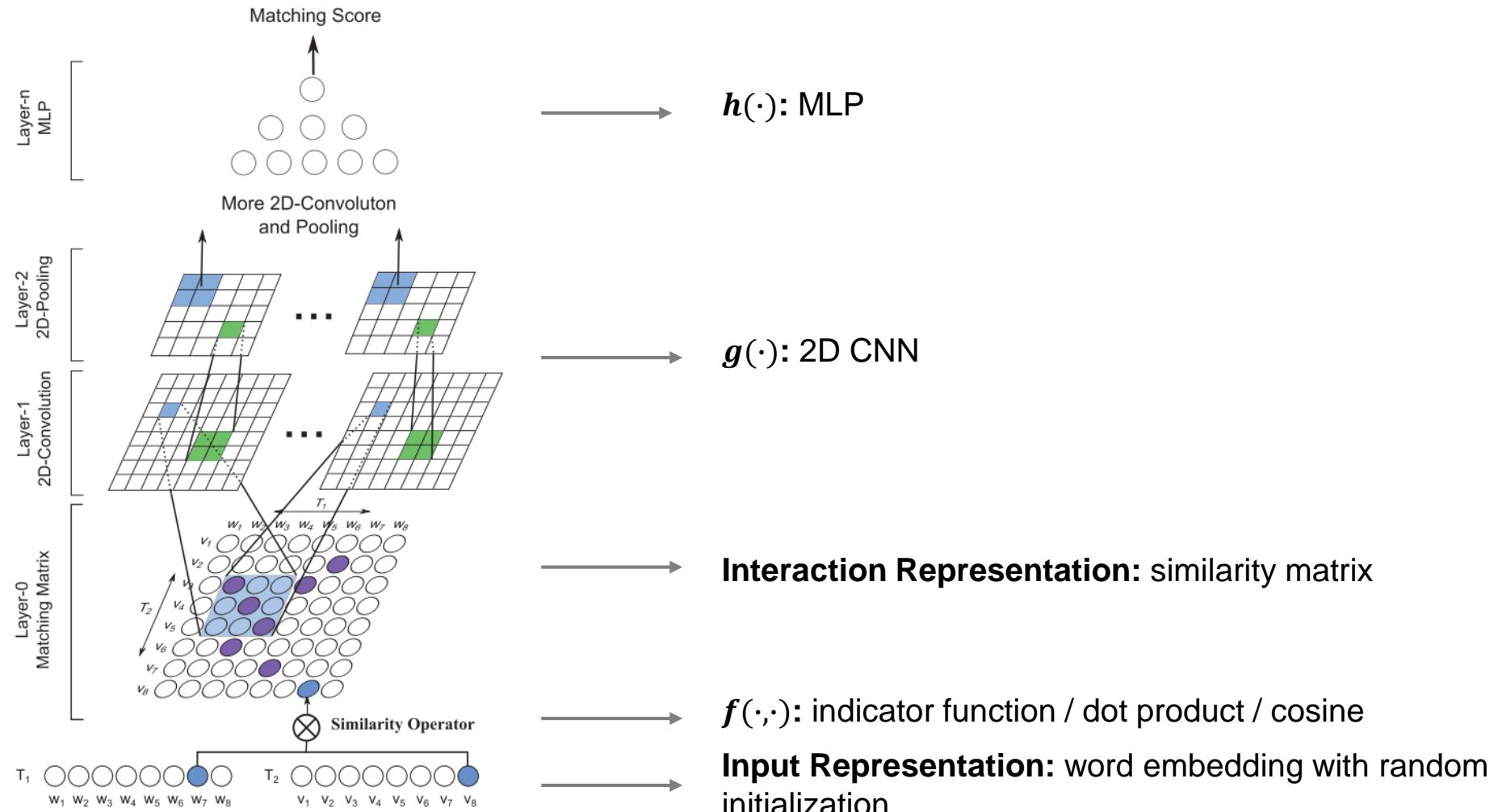
Two Types of Interaction



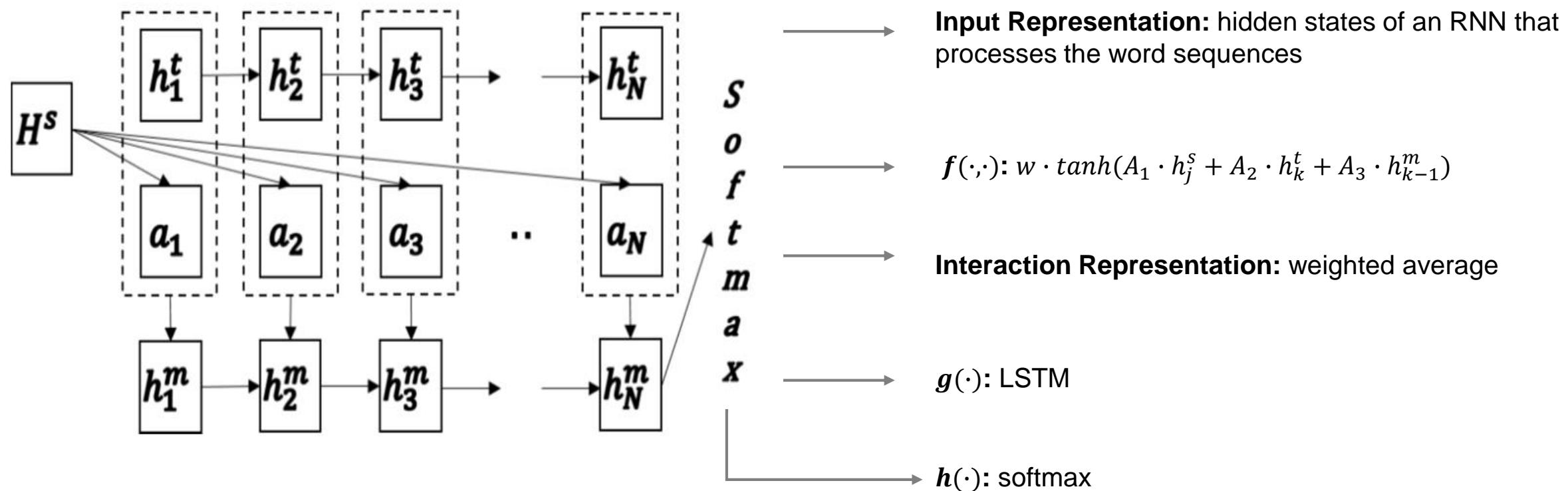
Implementation of Framework II

Input Representation	$f(\cdot, \cdot)$	Interaction Representation	$g(\cdot)$	$h(\cdot)$
Word embedding with random initialization	Cosine/Dot product	Similarity matrices	2D CNN	MLP
Word embedding initialized with Glove	Linear and non-linear transformation	Weighted average (Attention)	RNN	SoftMax
Hidden states of an RNN	Tensor			
	Indicator function			
	Euclidean distance			

Special Case 1: Match Pyramid



Special Case 2: Match LSTM



More Methods in Framework II

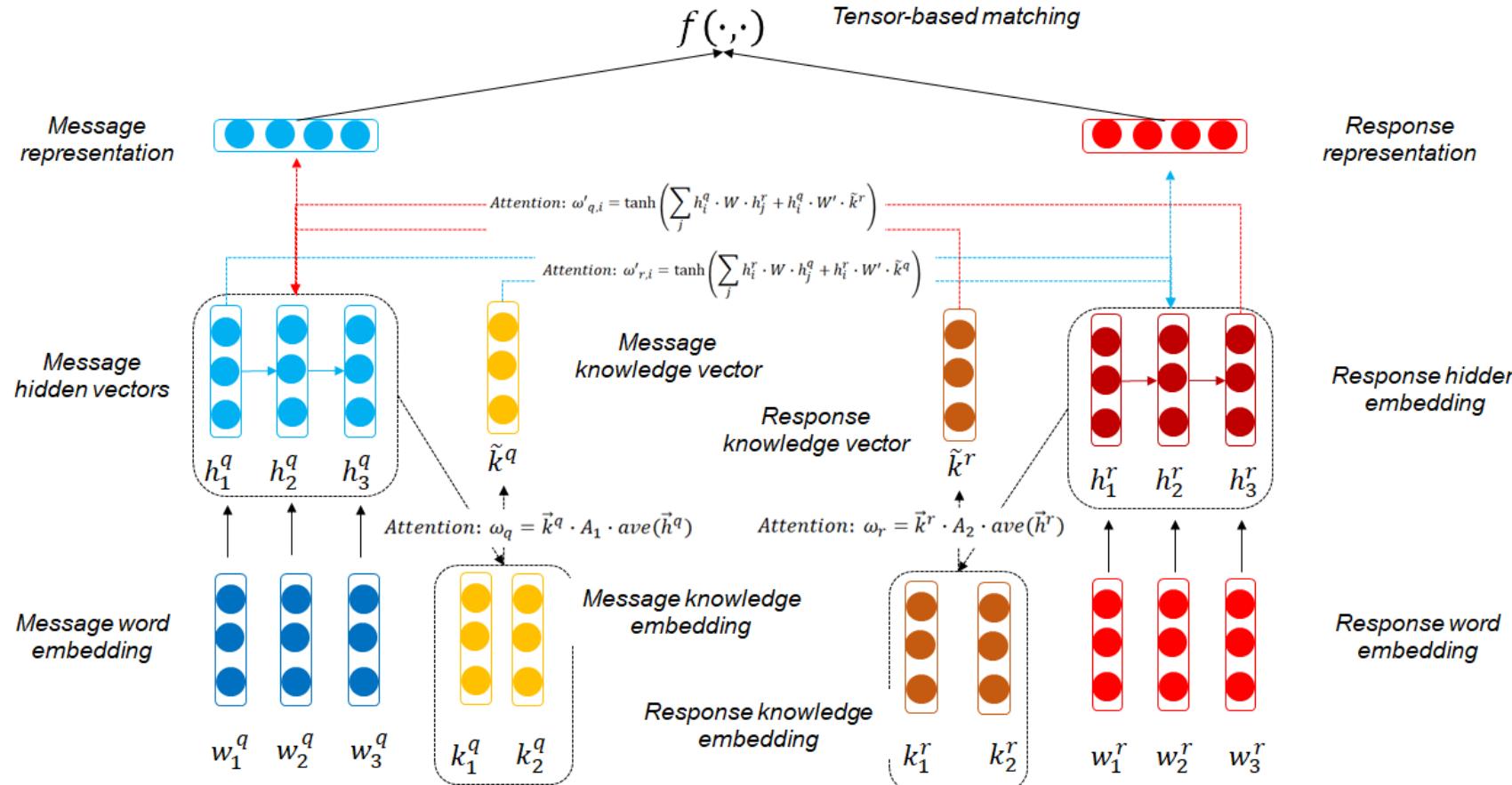
Model	Interaction Representation	$g(\cdot)$	Paper
Arc-II	Similarity matrix (linear & non-linear transformation)	2D CNN	Baotian Hu et al., Convolutional Neural Network Architectures for Matching Natural Language Sentences, <i>NIPS'14</i>
DeepMatch_topic	Similarity matrix (linear & non-linear transformation)	MLP	Zhengdong Lu et al., A Deep Architecture for Matching Short Texts, <i>NIPS'13</i>
Match Pyramid	Similarity matrix	2D CNN	Liang Pang et al., Text Matching as Image Recognition, <i>AAAI'16</i>
Match LSTM	Weighted average	LSTM	Shuohang Wang & Jing Jiang, Learning Natural Language Inference with LSTM, <i>NAACL'16</i>
MV-LSTM	Similarity matrix (tensor of hidden states)	K-max pooling	Shengxian Wan et al., A Deep Architecture for Semantic Matching with Multiple Positional Sentence Representations, <i>AAAI'16</i>

Comparison between Framework I and II

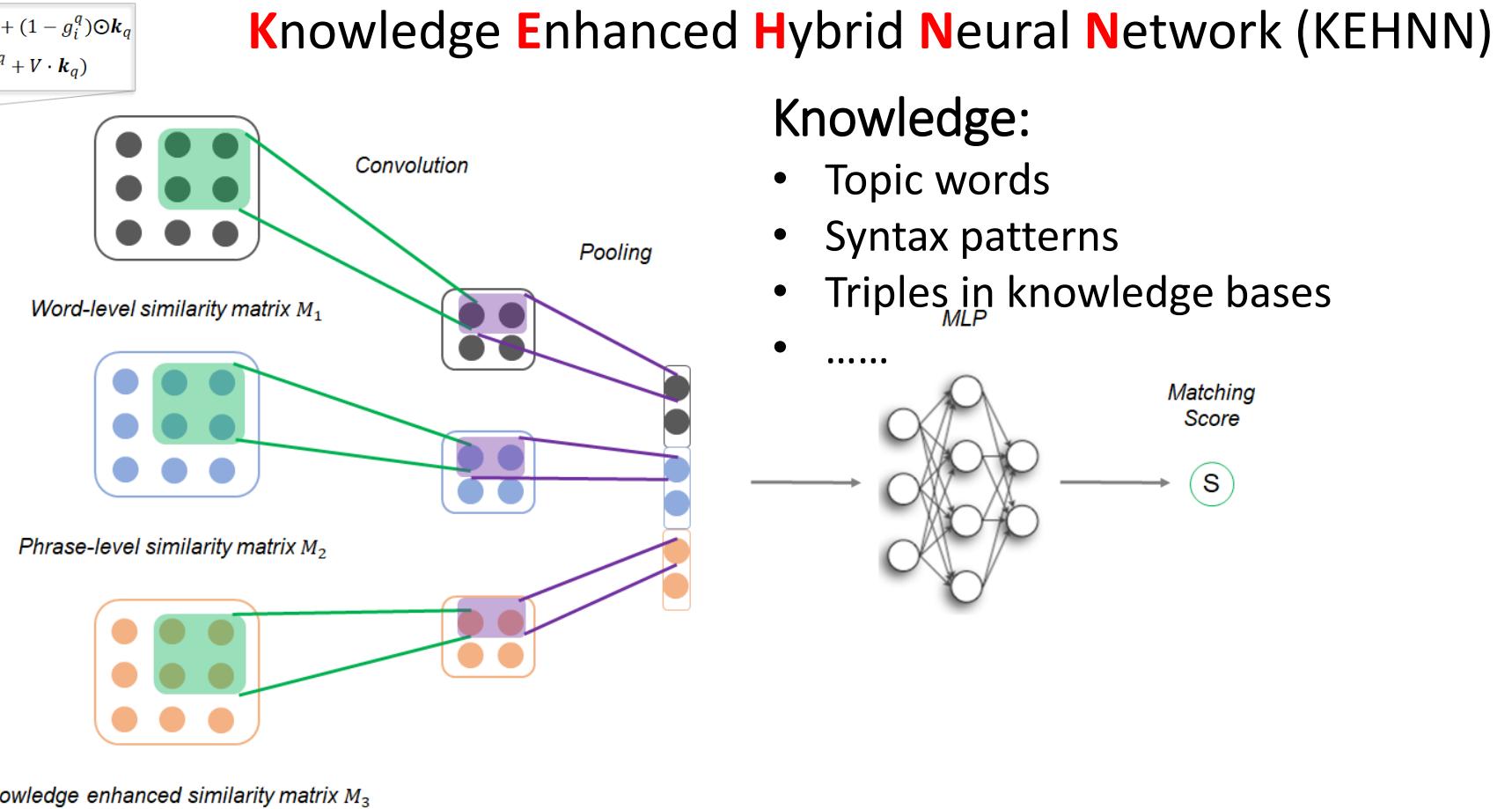
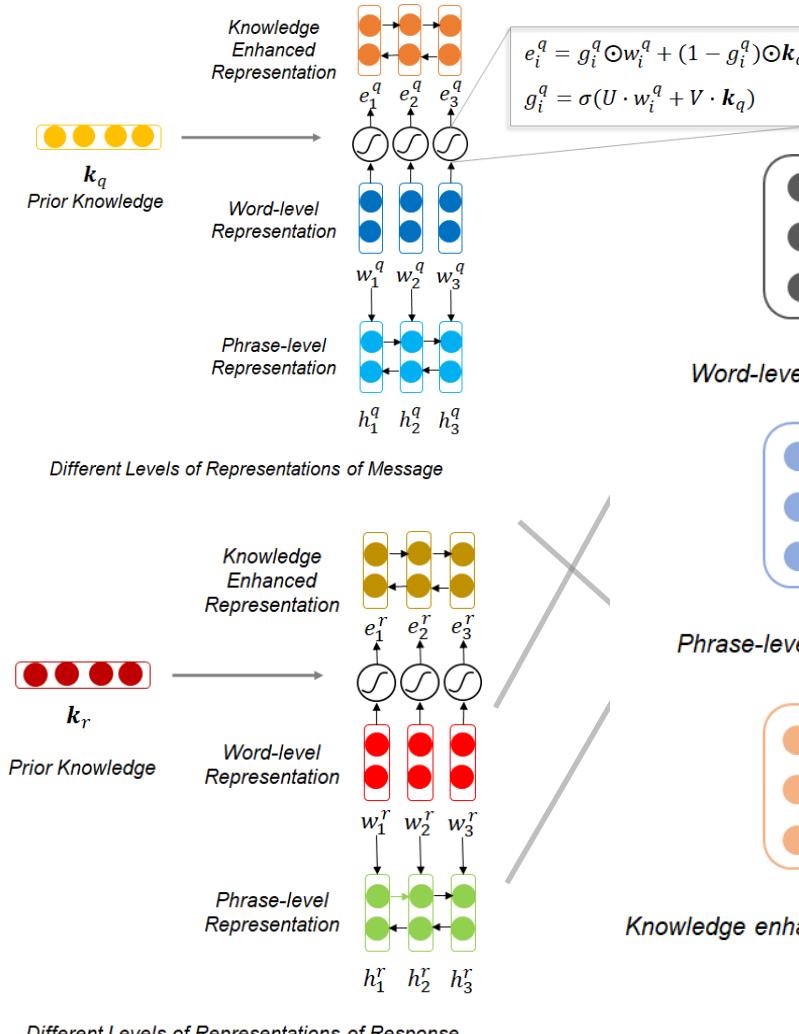
- **Efficacy**
 - In general, models in Framework II are better than models in Framework I on published datasets (see later), because matching information in a message-response pair is sufficiently preserved by the interaction in Framework II.
- **Efficiency**
 - Because the sufficient (and heavy) interaction, models in Framework II in general is more costly than models in Framework I.
 - Because one can pre-compute the representations of messages and responses and store them in index with text, models in Framework I is more preferable when there is strict requirement to online responding time.

Extension of Framework I : Matching with External Knowledge

Topic Aware Attentive Recurrent Neural Network (TAARNN)



Extension of Framework II : Matching with Multiple Levels of Representations



Datasets for Empirical Studies

- Ubuntu Dialogue Corpus

- Dyadic English human-human conversations from Ubuntu chat logs.
- Each conversation has multiple turns (Avg. # turns=7.71). For single-turn studies, we keep the last turn and the response to form a message-response pair.
- Task = distinguishing the positive response from negative ones for a given message.
- Train : validation: test = 1M : 0.5M : 0.5M.
- Positive responses = human responses, negative responses = randomly sampled ones.
- Positive : negative = 1:1 (train), 1:9 (validation), 1:9 (test).
- Evaluation metrics = $R_n@k$. For each message, if the only positive response is ranked within top k positions of n candidates, then $R_n@k = 1$. The final result is the average on messages.

Empirical Comparison

Model	Ubuntu			
	$R_2@1$	$R_{10}@1$	$R_{10}@2$	$R_{10}@5$
MLP (I)	0.651	0.256	0.380	0.703
Arc-I (I)	0.665	0.221	0.360	0.684
LSTM (I)	0.725	0.361	0.494	0.801
CNTN (I)	0.743	0.349	0.512	0.797
Attentive-LSTM (I)	0.758	0.381	0.545	0.801
DeepMatch_topic (II)	0.593	0.248	0.376	0.693
Match LSTM (II)	0.685	0.289	0.430	0.701
Arc-II (II)	0.736	0.380	0.534	0.777
Match Pyramid (II)	0.743	0.420	0.554	0.786
MV-LSTM (II)	0.767	0.410	0.591	0.819
TAARNN (I)	0.770	0.404	0.560	0.817
KEHNN (II)	0.786	0.460	0.591	0.819

Insights from the Comparison

- Neural tensor is a powerful matching function (CNTN is much better than Arc-I).
- Attentive LSTM achieves good performance because the attention mechanism is inherently a kind of interaction.
- Similarity matrix based interaction is better than attention based interaction for Framework II (Match LSTM performs badly).
- In similarity matrix based methods, dot product or cosine is better as an interaction function (Match Pyramid is better than Arc-II).
- Input representation that encodes more contextual information can improve the performance of matching (MV-LSTM is better than Match Pyramid).
- Topic information is useful for matching in both Framework I and Framework II (TAARNN performs well).
- Combining information from multiple sources is beneficial to matching (KEHNN performs well).

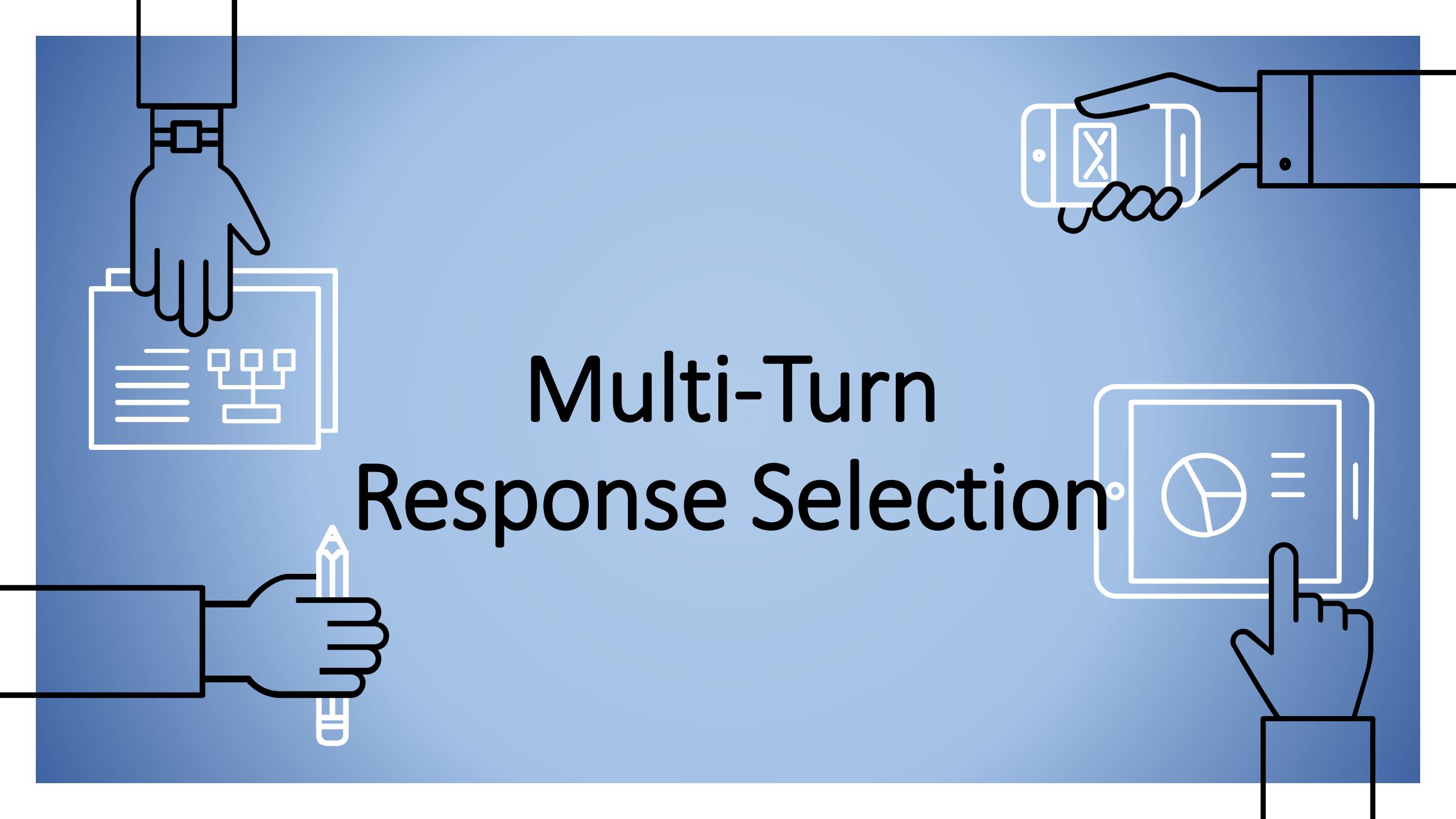
References

- Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. Workshop of NIPS'13.
- Zhengdong Lu and Hang Li. A Deep Architecture for Matching Short Texts. NIPS'13.
- Baotian Hu, Zhengdong Lu, Hang Li, and Qingcai Chen. Convolutional Neural Network Architectures for Matching Natural Language Sentences. NIPS'14
- Shengxian Wang, Yanyan Lan, Jiafeng Guo, Jun Xu, Liang Pang, and Xueqi Cheng. A Deep Architecture for Semantic Matching with Multiple Positional Sentence Representations. AAAI'16
- Minwei Feng, Bing Xiang, Michael R. Glass, Lidan Wang, and Bowen Zhou. Applying Deep Learning to Answer Selection: A Study and An Open Task. IEEE ASRU 2015.
- Xipeng Qiu and Xuanjing Huang. Convolutional Neural Tensor Network Architecture for Community-based Question Answering. IJCAI'15.
- Ming Tan, Cicero dos Santos, Bing Xiang, and Bowen Zhou. Improved Representation Learning for Question Answer Matching. ACL'16.
- Bingning Wang, Kang Liu, and Jun Zhao. Inner Attention based Recurrent Neural Networks for Answer Selection. ACL'16.

References

- Shuohang Wang and Jing Jiang. Learning Natural Language Inference with LSTM. NAACL'16.
- Aliaksei Severyn and Alessandro Moschitti. Learning to Rank Short Text Pairs with Convolutional Deep Neural Networks. SIGIR'15.
- Liang Pang, Yanyan Lan, Jiafeng Guo, Jun Xu, Shengxian Wan, and Xueqi Cheng. Text Matching as Image Recognition. AAAI'16.
- Shengxian Wan, Yanyan Lan, Jun Xu, Jiafeng Guo, Liang Pang, and Xueqi Cheng. Match-SRNN: Modeling the Recursive Matching Structure with Spatial RNN. IJCAI'16.
- Wenpeng Yin and Hinrich Schütz. MultiGranCNN: An Architecture for General Matching of Text Chunks on Multiple Levels of Granularity. ACL'15.
- Mingxuan Wang, Zhengdong Lu, Hang Li, and Qun Liu. Syntax-based Deep Matching of Short Texts. IJCAI'15.
- Yu Wu, Wei Wu, Can Xu, and Zhoujun Li. Knowledge Enhanced Hybrid Neural Network for Text Matching. AAAI'18.

Multi-Turn Response Selection[°]



Multi-Turn Response Selection



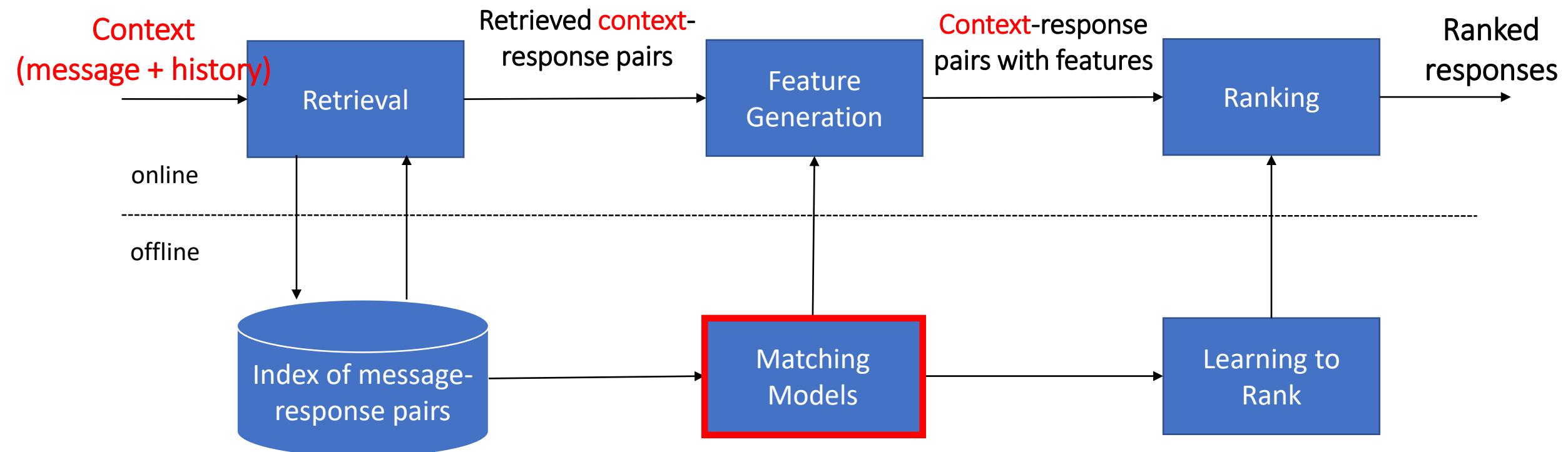
Candidate responses

- Yes, of course. ✓
- What lesson? ✗
- No, no free lessons. ✓
- Yes, please bring your drum ✓
- We do not have coaches. ✗
- Our English lessons are free ✗



Candidates in red are good without the context!

System Architecture for Multi-Turn Response Selection



<q>well I hope you fell better smiley</q>
 <r> thanks </r>
 <q> I was so excited about dying my hair blond</q>
 <r> do it </r>

.....

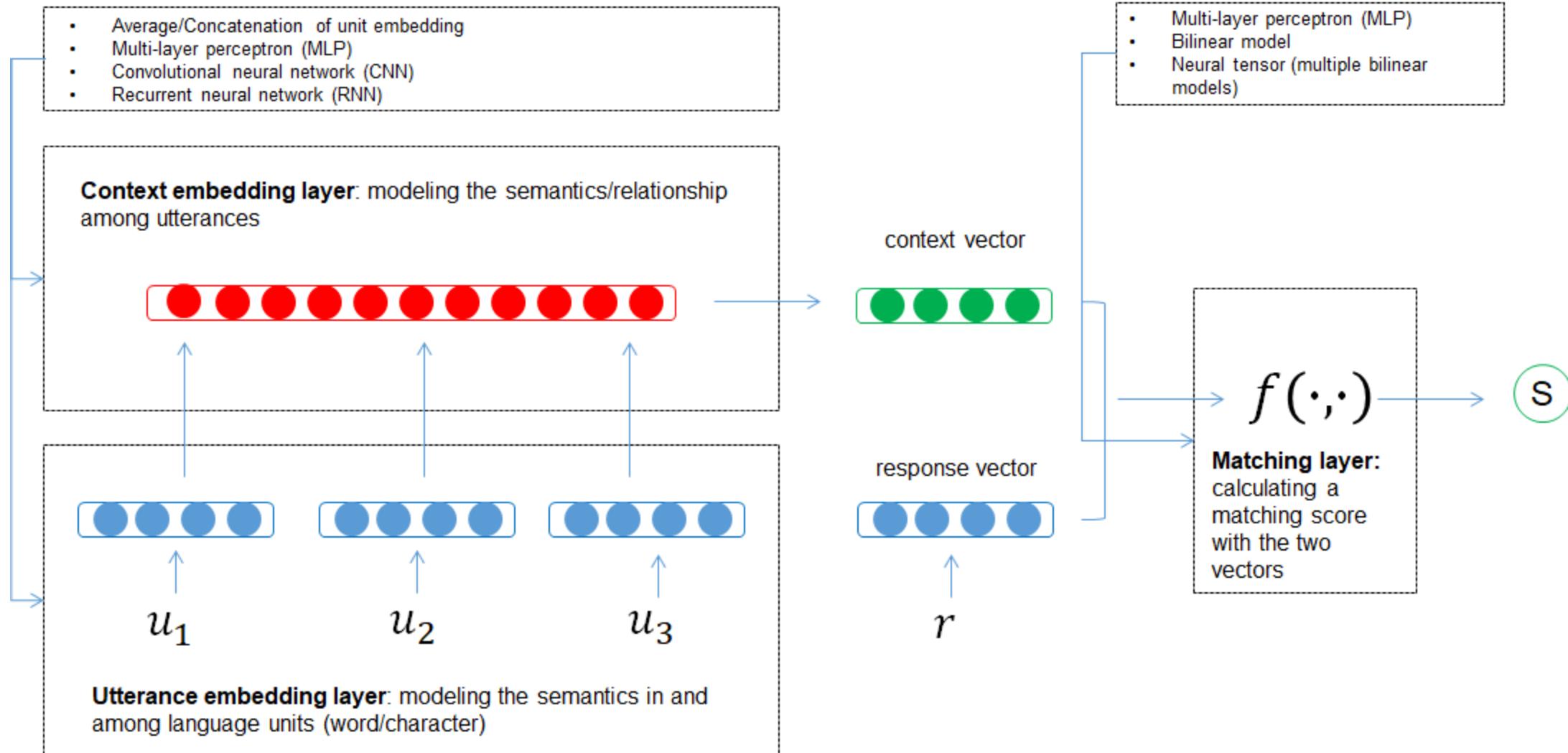
Deep learning based
context-response matching

Gradient boosted tree

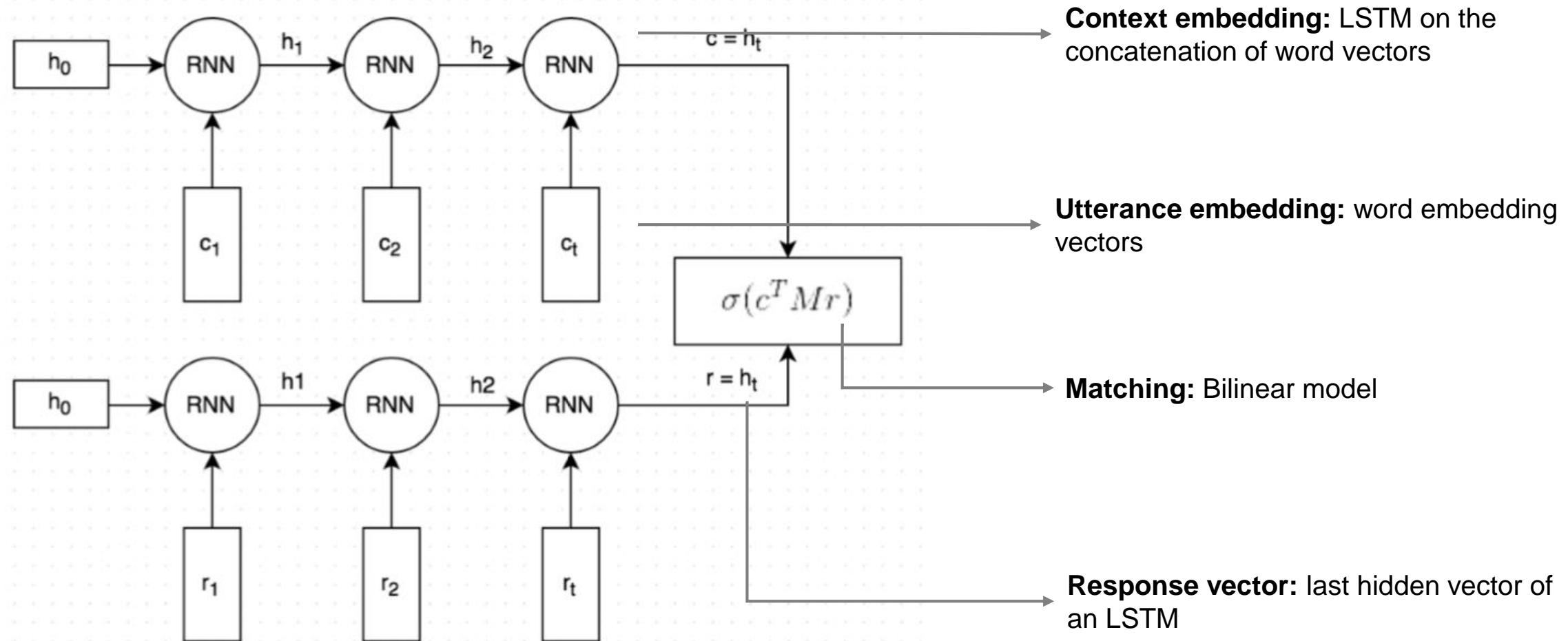
New Challenges with Contexts

- A hierarchical data structure
 - Words -> utterances -> session
- Information redundancy
 - Not all words and utterances are useful for response selection
- Logics
 - Order of utterances matters in response selection
 - Long-term dependencies among words and utterances
 - Constraints to proper responses

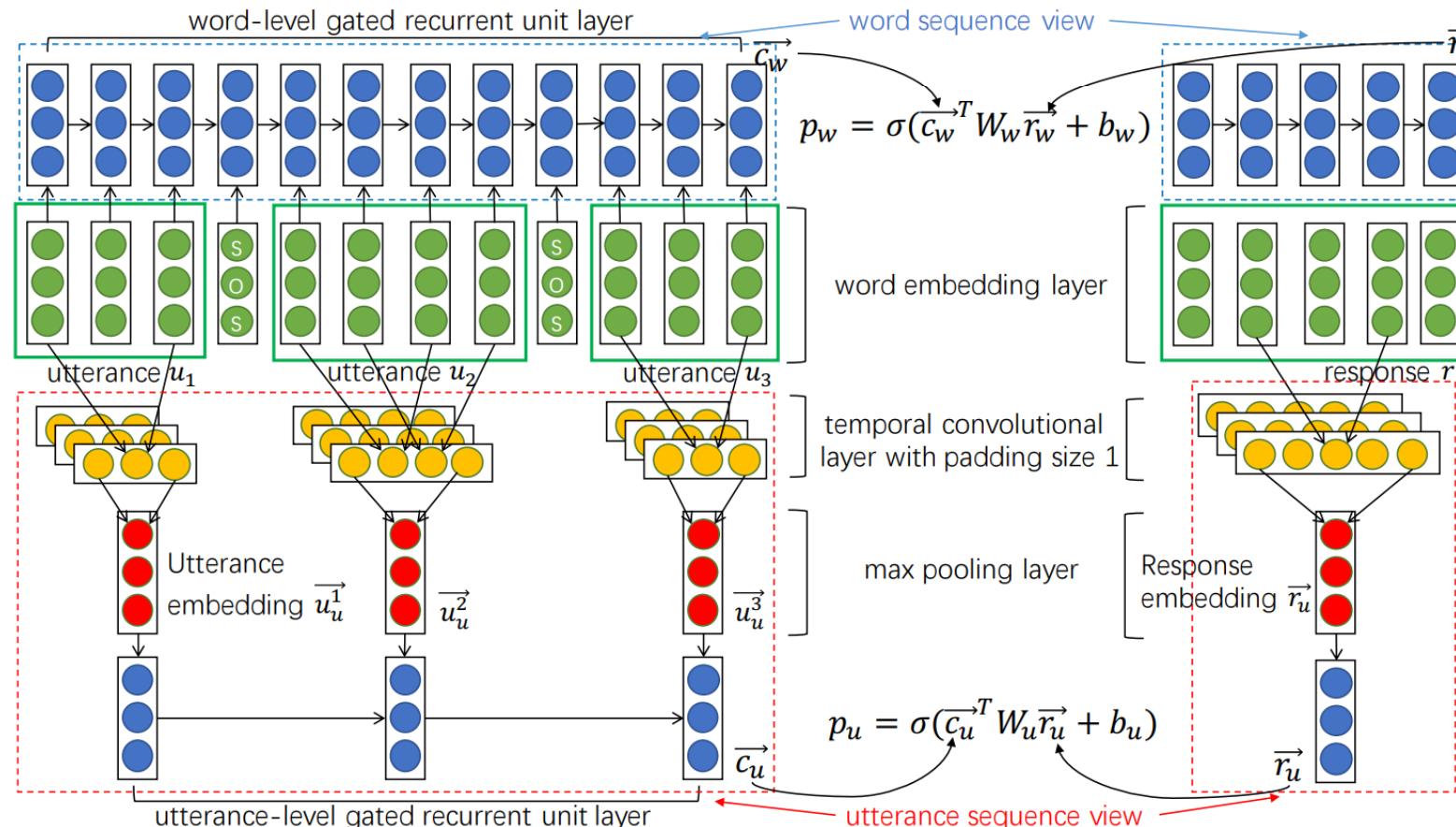
Context-Response Matching: Framework I



Special Case 1: Dual-LSTM



Special Case 2: Multi-view Response Selection Model



Context embedding (word view):
GRU on the concatenation of word vectors

Utterance embedding (word view):
word embedding vectors

Response vector (word view): last hidden vector of GRU

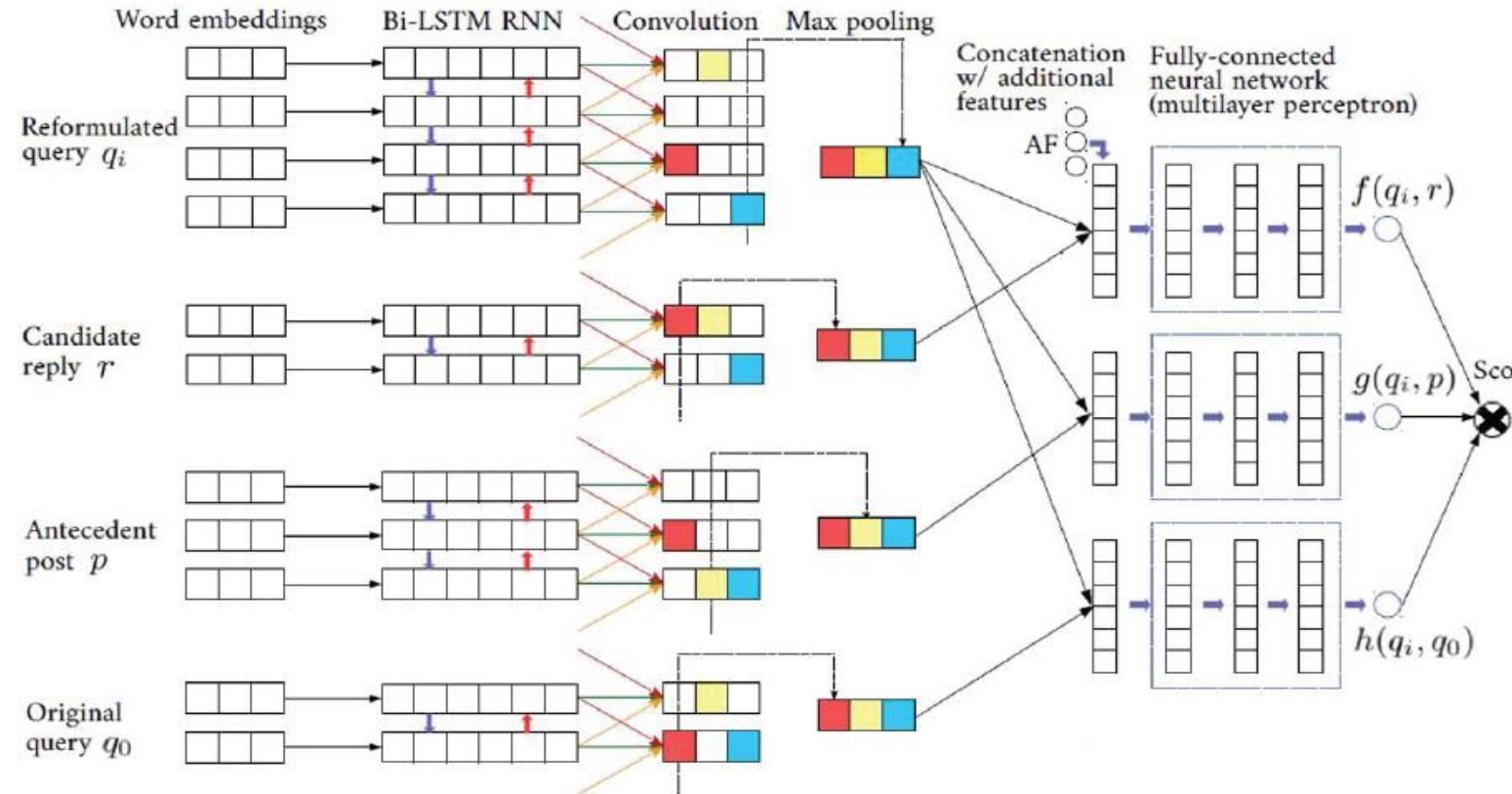
Matching: Bilinear

Utterance embedding (utterance view): CNN

Context embedding (utterance view): GRU

Response vector (utterance view):
CNN

Special Case 3: Deep Learning to Respond (DL2R)



Utterance embedding:
BiLSTM+CNN

Context embedding:
Identity

Matching:
MLP

Response vector:
BiLSTM+CNN

More Methods in Framework I

Model	Utterance Embedding	Context Embedding	Matching	Paper
Dual-LSTM	Word embedding	LSTM	Bilinear	Ryan Lowe et al., The Ubuntu Dialogue Corpus: A Large Dataset for Research in Unstructured Multi-turn Dialogue Systems. <i>SIGDIAL'15</i>
DL2R	BiLSTM+CNN	Identity	MLP	Rui Yan et al., Learning to Respond with Deep Neural Networks for Retrieval-based Human-Computer Conversation System. <i>SIGIR'16</i>
Multi-View	Word embedding/CNN	GRU	Bilinear	Xiangyang Zhou et al., Multi-view Response Selection for Human-Computer Conversation, <i>EMNLP'16</i>
Coupled LSTM	LSTM (on utterance pair)	LSTM	MAP	Rui Yan et al., Coupled Context Modeling for Deep Chit-Chat: Towards Conversations between Human and Computer, <i>KDD'18</i>
Any single-turn model	Word embedding	Message embedding of the model	Matching function of the model	N/A

Analysis of Framework I

Modeling the hierarchy of conversation sessions?



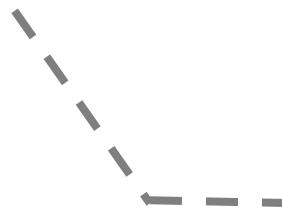
Modeling the relationship/dependency among words?



Modeling the relationship/dependency among utterances?

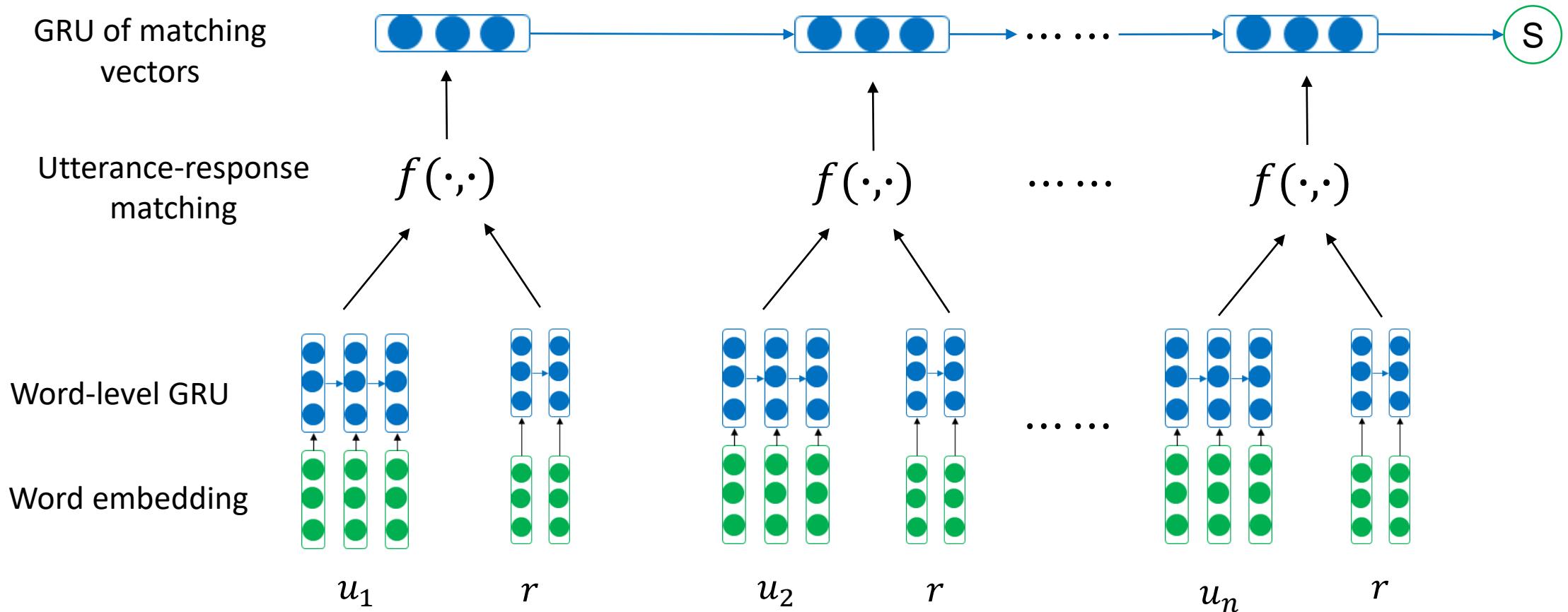


Modeling word/utterance importance

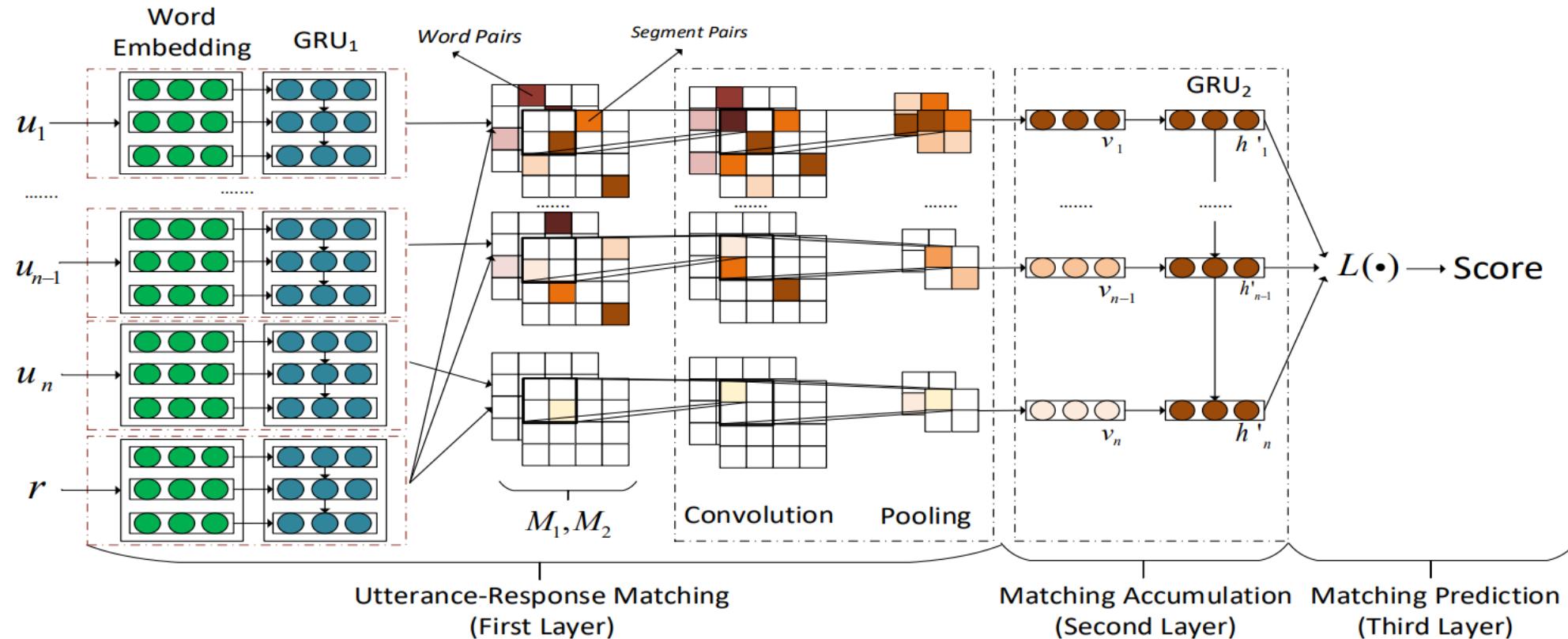


- Every context is compressed to a fixed length vector before matching with the response.
- Utterance representation is independent with the response/other utterances.

Context-Response Matching: Framework II

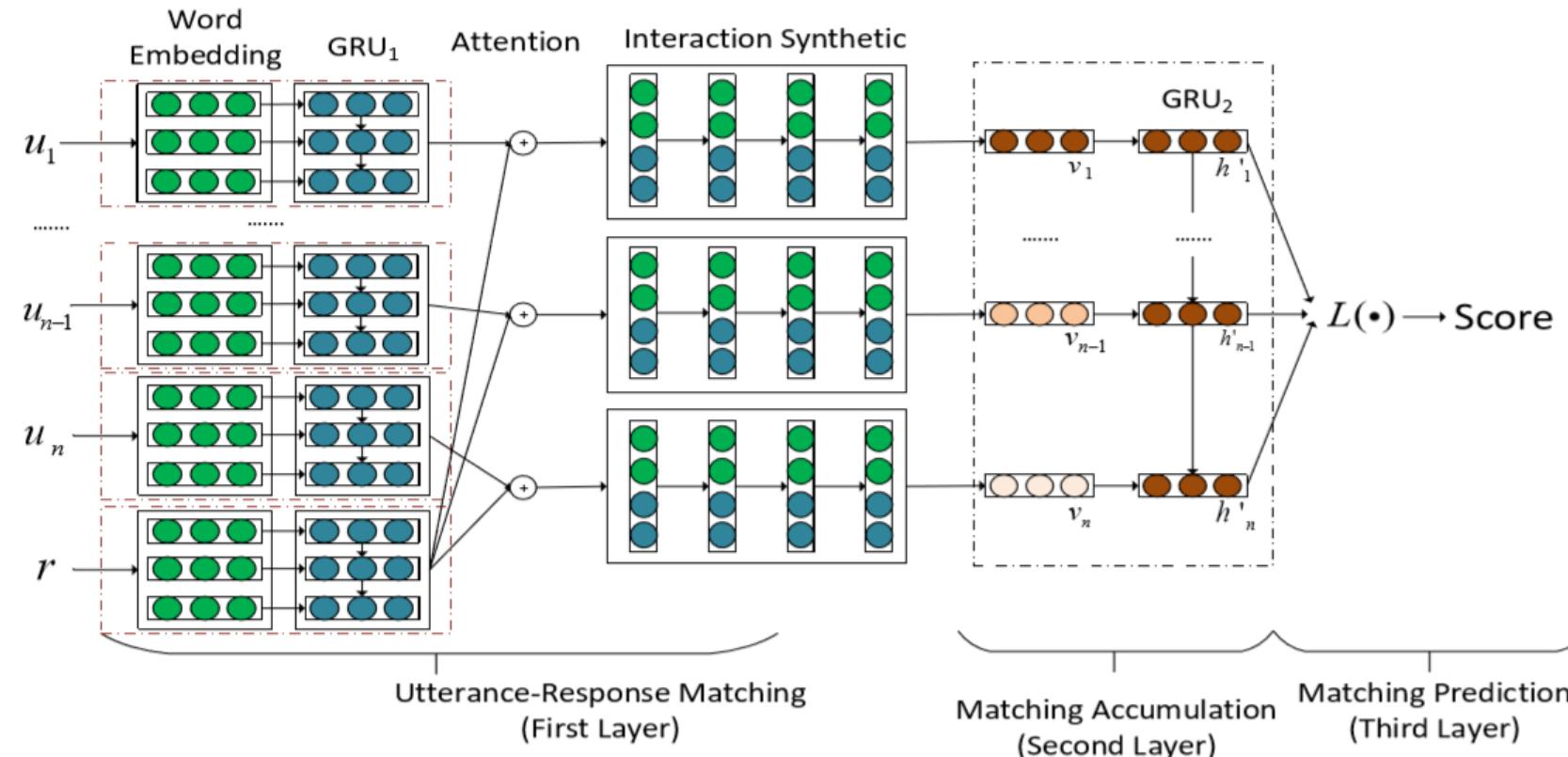


Special Case 1: Sequential Matching Network (SMN)



$$f(u_i, r) = W \cdot [\text{CNN}(M_1) \oplus \text{CNN}(M_2)] + b$$

Special Case 2: Sequential Attention Network (SAN)



$$f(u_i, r) = GRU([\text{Att}(r \rightarrow u_i) \odot r]_w \oplus [\text{Att}(r \rightarrow u_i) \odot r]_h)$$

How Framework II Models Word/Utterance Importance

u_1 : How can unzip many rar files (_number_ for example) at once?

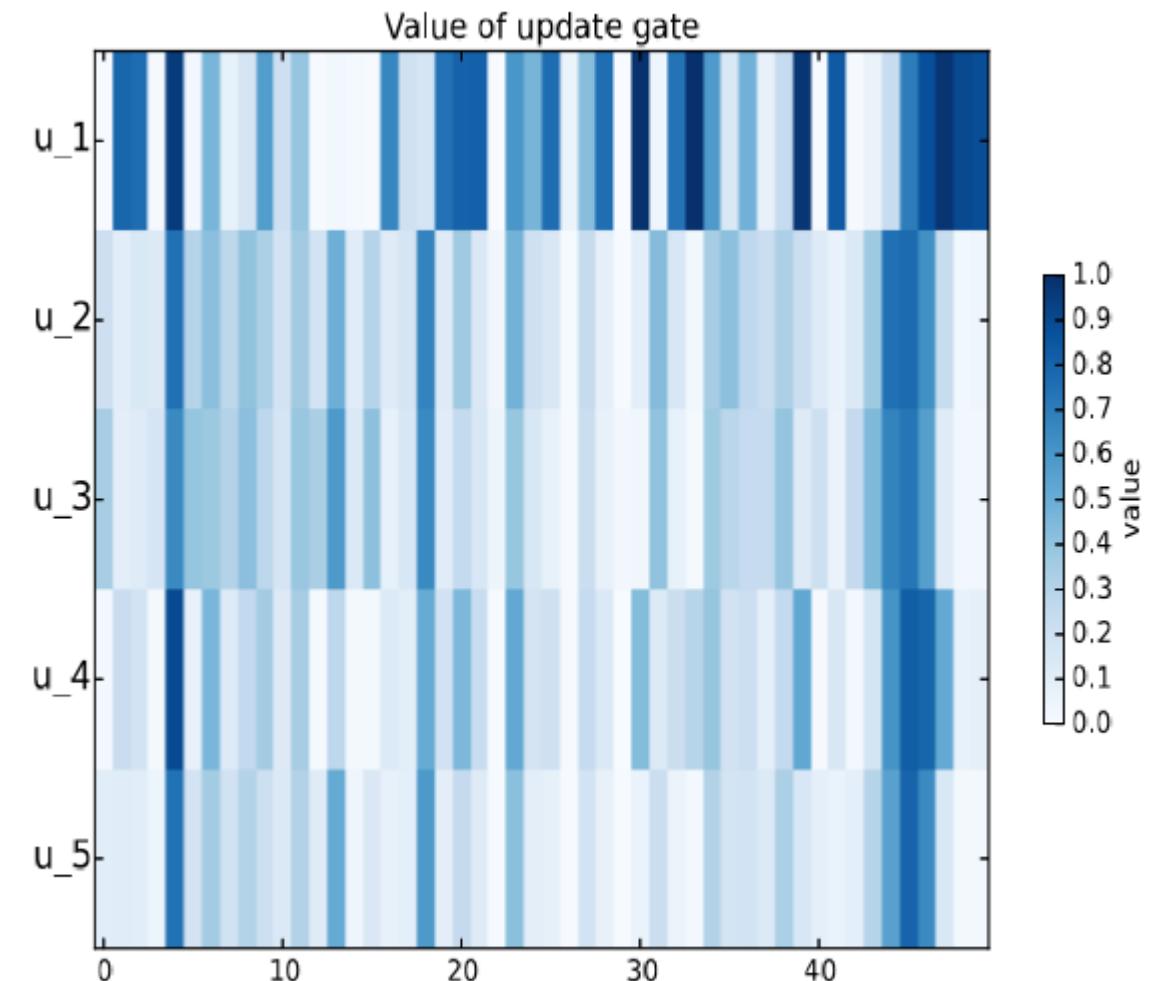
u_2 : Sure, you can do it in bash.

u_3 : OK, how?

u_4 : Are the files all in the same directory?

u_5 : Yes, they all are.

Response: Then the command glebihan should extract them all from/to that directory.



Comparison between Framework I and II

- Efficacy
 - In general, models in Framework II are better than models in Framework I on published datasets (see later), because important information in a context is sufficiently distilled and preserved in matching in Framework II.
- Efficiency
 - Because the sufficient (and heavy) interaction, models in Framework II in general is more costly than models in Framework I.
- Interpretability
 - It is easy to understand what models in Framework II have learned via some visualization techniques.

Datasets for Empirical Studies

- Ubuntu Dialogue Corpus
- Douban Conversation Corpus
 - Dyadic Chinese human-human dialogues (comment streams) from Douban group, a popular social networking service in China.
 - Train : validation : test = 1M : 50k : 10k with average turns 6.69, 6.75, and 6.45 respectively.
 - Train & validation : positive responses are human responses, and negative ones are randomly sampled.
 - Test: 10 human judged response candidates per context retrieved from an index.
 - Evaluation metrics: MAP, MRR, and P@1 (from information retrieval)

Empirical Comparison

Model	Ubuntu				Douban		
	$R_2@1$	$R_{10}@1$	$R_{10}@2$	$R_{10}@5$	MAP	MRR	$P@1$
TF-IDF	0.659	0.410	0.545	0.708	0.331	0.359	0.180
CNN	0.848	0.549	0.684	0.896	0.417	0.440	0.226
BiLSTM	0.895	0.630	0.780	0.944	0.479	0.514	0.313
MV-LSTM	0.906	0.653	0.804	0.946	0.498	0.538	0.348
Match LSTM	0.904	0.653	0.799	0.944	0.500	0.537	0.345
Attentive-LSTM	0.903	0.633	0.789	0.943	0.495	0.523	0.331
Dual LSTM	0.901	0.638	0.784	0.949	0.485	0.527	0.320
DL2R	0.899	0.626	0.783	0.944	0.488	0.527	0.330
Multi-View	0.908	0.662	0.801	0.951	0.505	0.543	0.342
SMN	0.926	0.726	0.847	0.961	0.529	0.569	0.397
SAN	0.932	0.734	0.852	0.962	0.532	0.575	0.387

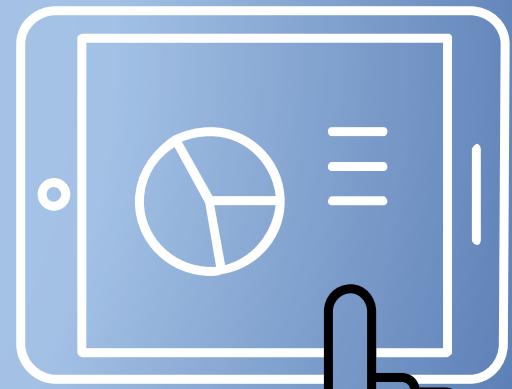
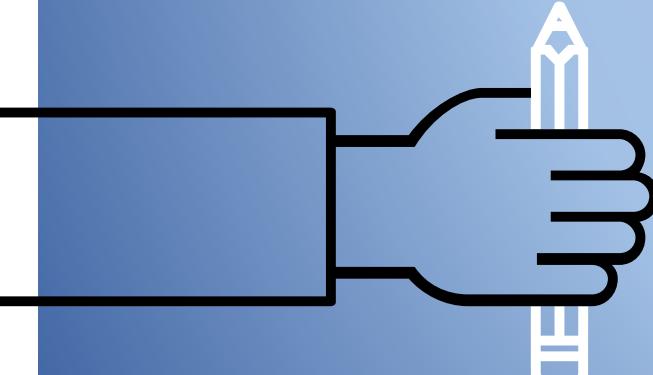
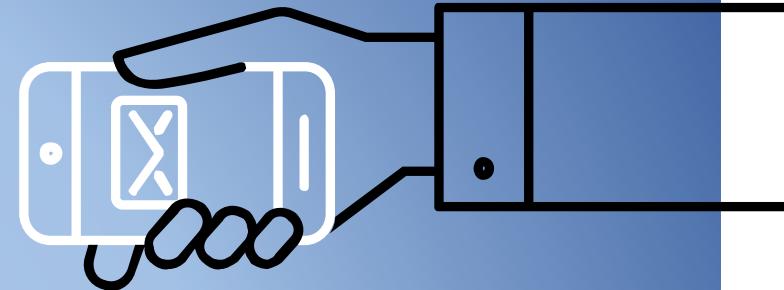
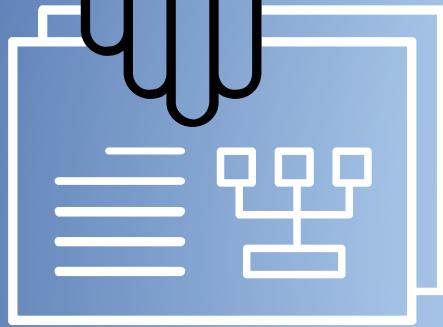
Insights from the Comparison

- Utterance-response interaction at the beginning is crucial, as models in Framework II (SMN & SAN) are much better than models in Framework I.
- Ensemble of matching from multiple views are effective, as multi-view outperforms all other models in Framework I.
- Selection of interaction functions does not make a clear difference on efficacy, as long as the function can model word/phrase importance. This is indicated by the comparable performance of SMN and SAN.
- SMN is more efficient and easier to parallelize than SAN, due the characteristics of CNN and RNN.

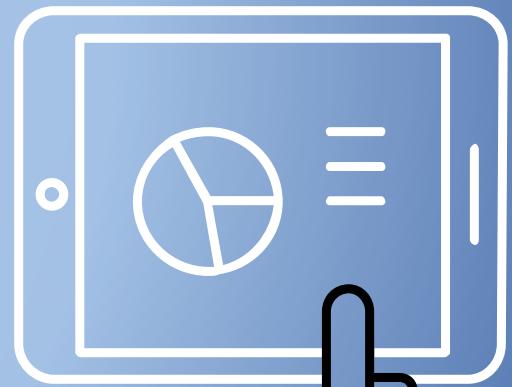
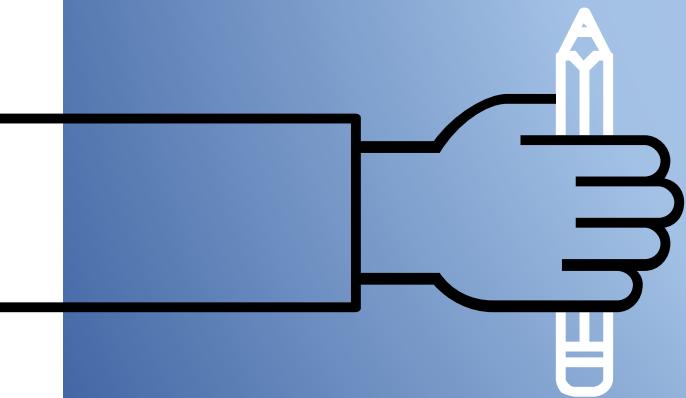
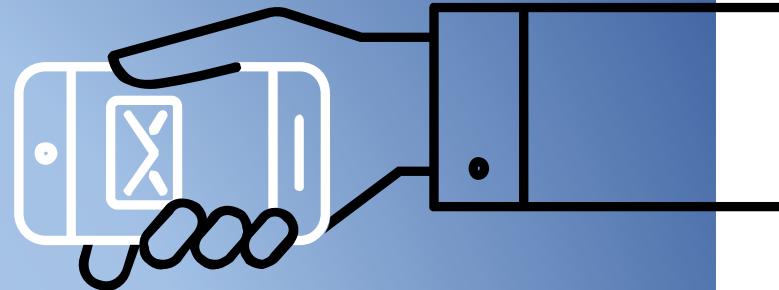
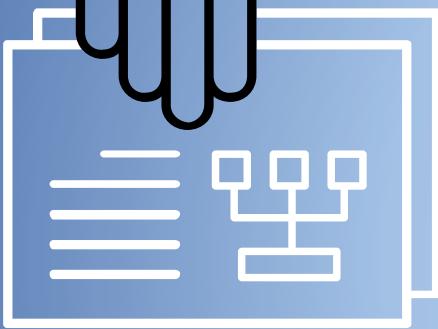
References

- Ryan Lowe, Nissan Pow, Iulian V. Serban, and Joelle Pineau. The Ubuntu Dialogue Corpus: A Large Dataset for Research in Unstructured Multi-Turn Dialogue Systems. SIGDIAL'15.
- Rudolf Kadlec, Martin Schmid, and Jan Kleindienst. Improved Deep Learning Baselines for Ubuntu Corpus Dialogs. NIS'15 workshop.
- Rui Yang, Yiping Song, Xiangyang Zhou, and Hua Wu. "Shall I be Your Chat Companion?" Towards an Online Human-Computer Conversation System. In CIKM'16
- Rui Yan, Yiping Song, and Hua Wu. Learning to Respond with Deep Neural Networks for Retrieval-Based Human-Computer Conversation System. SIGIR'16.
- Rui Yan, Yiping Song, Xiangyang Zhou, and Hua Wu. "Shall I Be Your Chat Companion?" Towards an Online Human-Computer Conversation System. In CIKM'16
- Rui Yan and Dongyan Zhao. Coupled Context Modeling for Deep Chit-Chat: Towards Conversations between Human and Computer. In KDD'18.
- Xiangyang Zhou, Daxiang Dong, Hua Wu, Shiqi Zhao, Dianhai Yu, Hao Tian, Xuan Liu, and Rui Yan. Multi-view Response Selection for Human-Computer Conversation. EMNLP'16.
- Bowen Wu, Baoxun Wang, and Hui Xue. Ranking Responses Oriented to Conversational Relevance in Chatbots. COLING'16
- Yu Wu, Wei Wu, Chen Xing, Ming Zhou, and Zhoujun Li. Sequential Matching Network: A New Architecture for Multi-turn Response Selection in Retrieval-based Chatbots. ACL'17.
- Yu Wu, Wei Wu, Chen Xing, Can Xu, Zhoujun Li, and Ming Zhou. A Sequential Matching Framework for Multi-turn Response Selection in Retrieval-based Chatbots. arXiv:1710.11344.
- Zhen Xu, Bingquan Liu, Baoxun Wang, Chengjie Sun, and Xiaolong Wang. Incorporating Loose-Structured Knowledge into LSTM with Recall Gate for Conversation Modeling. IJCNN'17
- Liu Yang, Minghui Qiu, Chen Qu, Jiafeng Guo, Yongfeng Zhang, W.Bruce Croft, Jun Huang, and Haiqing Chen. Response Ranking with Deep Matching Networks and External Knowledge in Information-seeking Conversation Systems. SIGIR'18.

(Some) Emerging Directions



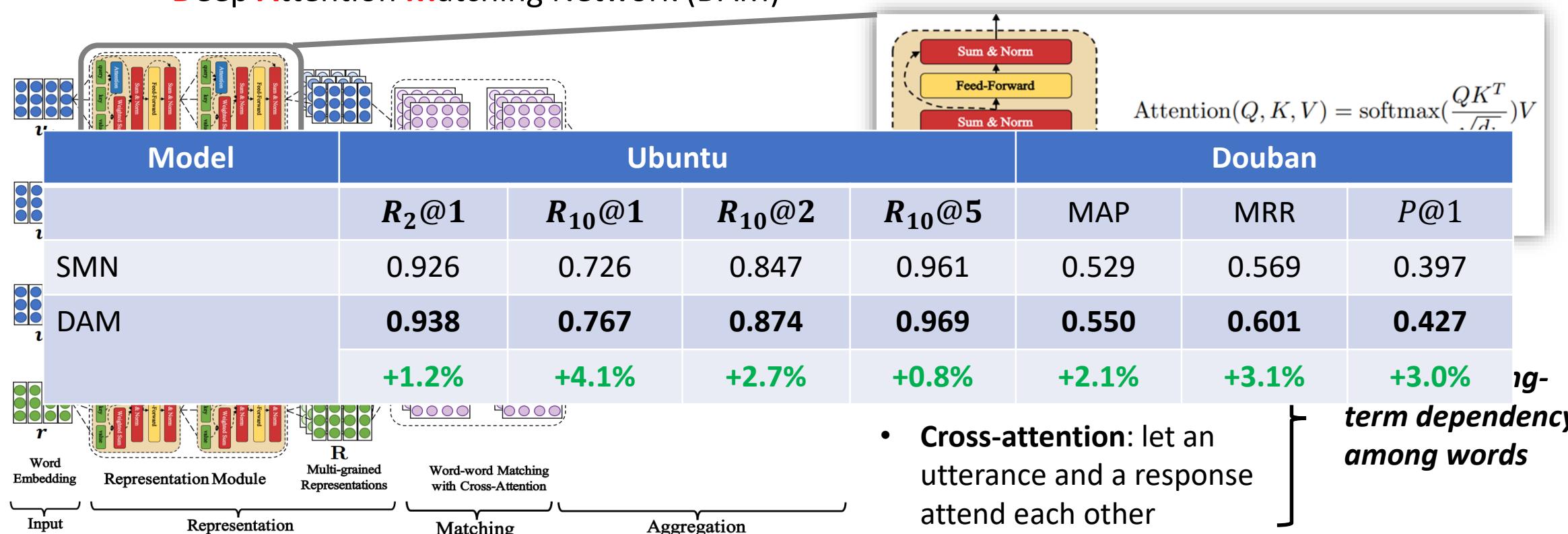
Matching with Better Representations



Matching With Better Representations

Self-attention, borrowed from NMT, is powerful in representing contexts and responses.

Deep Attention Matching Network (DAM)



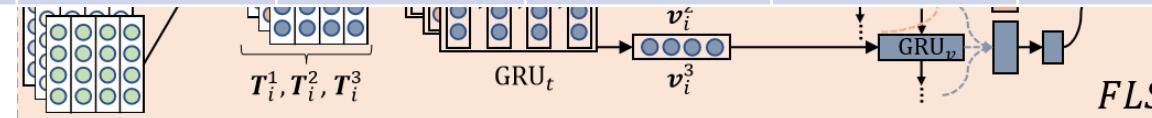
Matching With Better Representations

Fusing multiple types of representations are helpful, but how to fuse matters.

Multi-Representation Fusion Network (MRFN)



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%



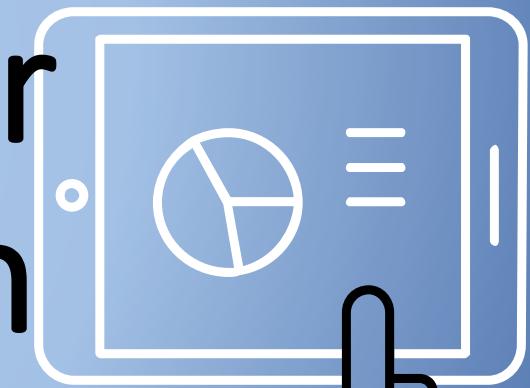
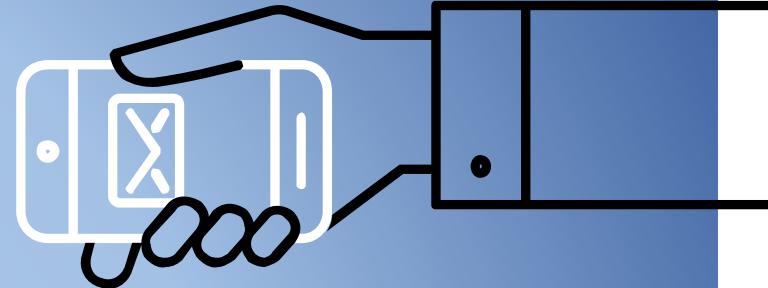
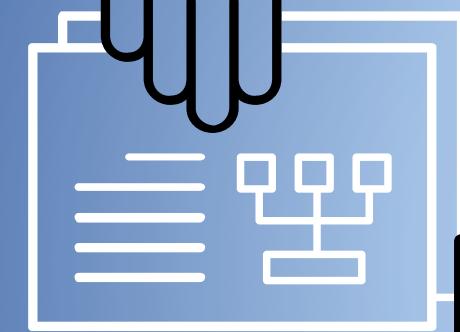
Matching With Better Representations

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

Model	Ubuntu			
	$R_2@1$	$R_{10}@1$	$R_{10}@2$	$R_{10}@5$
SMN	0.926	0.726	0.847	0.961
SMN+CoVe	0.930	0.738	0.856	0.963
	+0.4%	+1.2%	+0.9%	+0.2%
SMN+ELMo	0.917	0.720	0.835	0.953
	-0.9%	-0.6%	-1.2%	-0.8%
SMN+ELMo (fine-tune)	0.929	0.740	0.859	0.964
	+0.4%	+1.4%	+1.2%	+0.3%
SMN+ECMo	0.933	0.755	0.866	0.975
	+0.7%	+2.9%	+1.9%	+1.4%

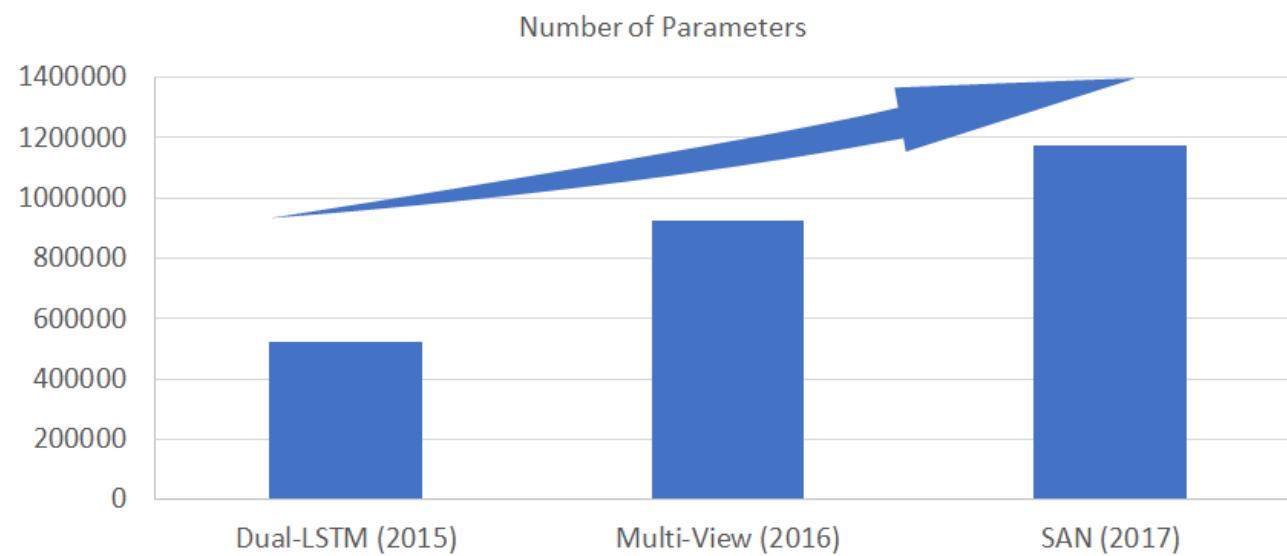
ctors. NIPS'17
nitions. NAACL'18

Learning a Matching Model for Response Selection



Why Pay Special Attention to Learning?

- Matching models are becoming more and more complicated

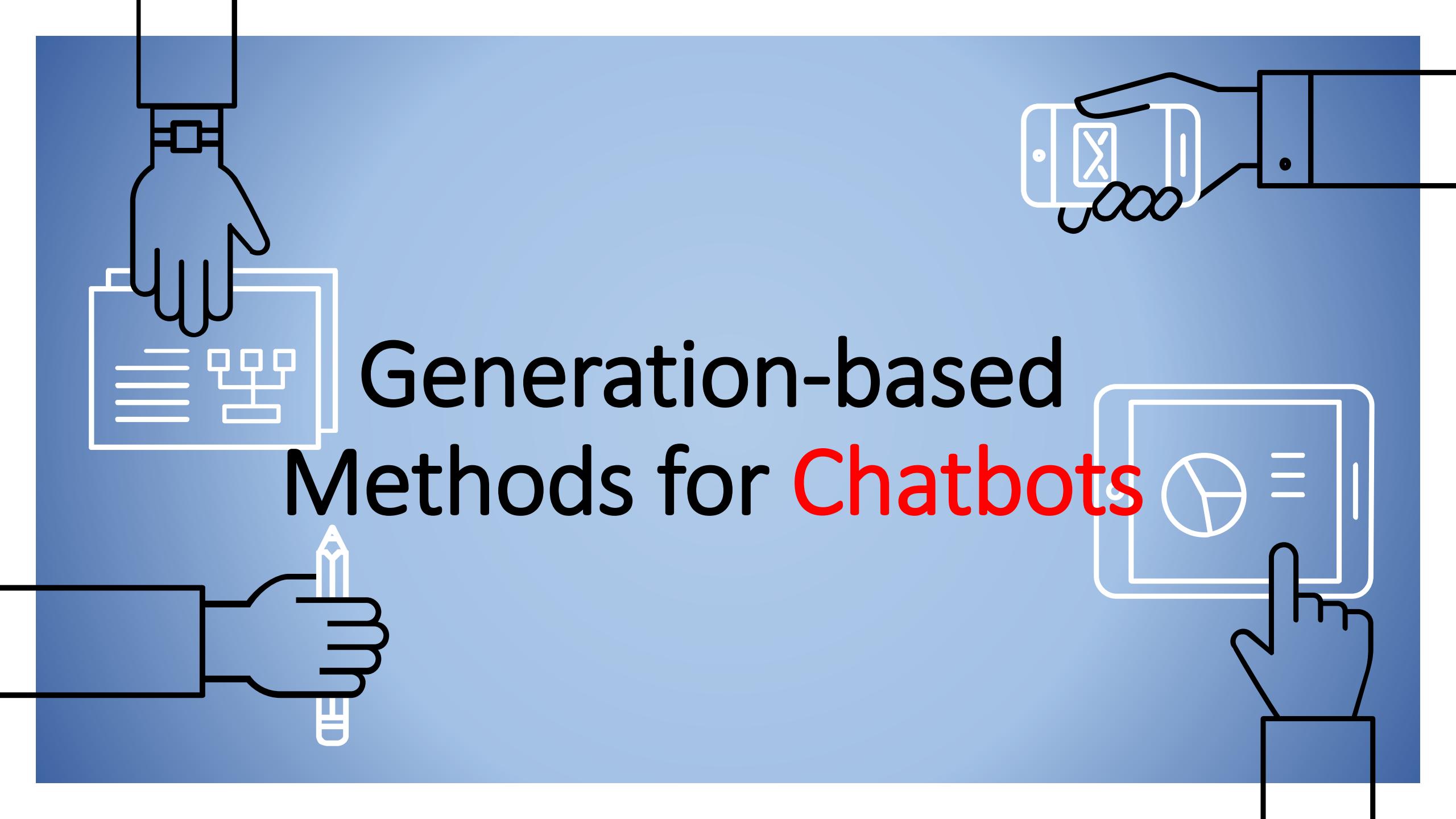


- Training data are randomly constructed with an almost fixed size
 - Size: ~million
 - Negative examples are randomly sampled

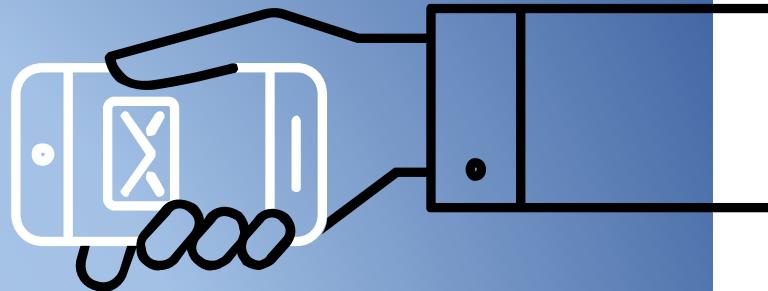
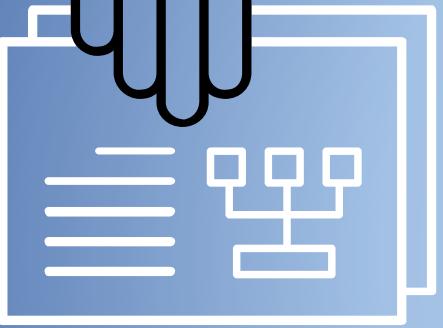
- *More and more easy patterns in training are captured*
- *Gap between training and test is still there*

Learning with Unlabeled Data – Weak Annotator

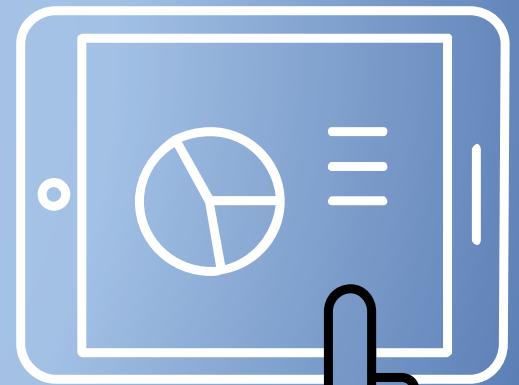
Model	Douban		
	MAP	MRR	P@1
Dual-LSTM	0.485	0.527	0.320
Dual-LSTM+Weak	0.519	0.559	0.359
	+3.4%	+3.2%	+3.9%
Multi-View	0.505	0.543	0.342
Multi-View+Weak	0.534	0.575	0.378
	+2.9%	+3.2%	+3.6%
SMN	0.526	0.571	0.393
SMN+Weak	0.565	0.609	0.421
	+3.9%	+3.8%	+2.8%



Generation-based Methods for Chatbots

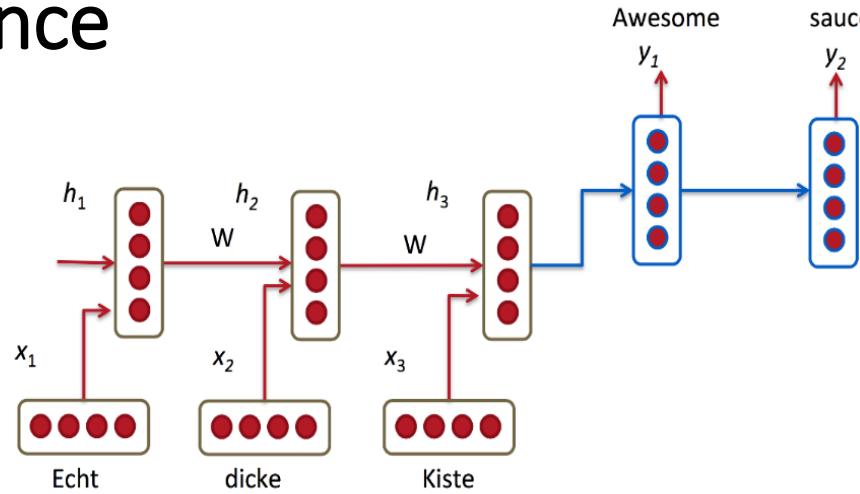


Single-Turn Generation

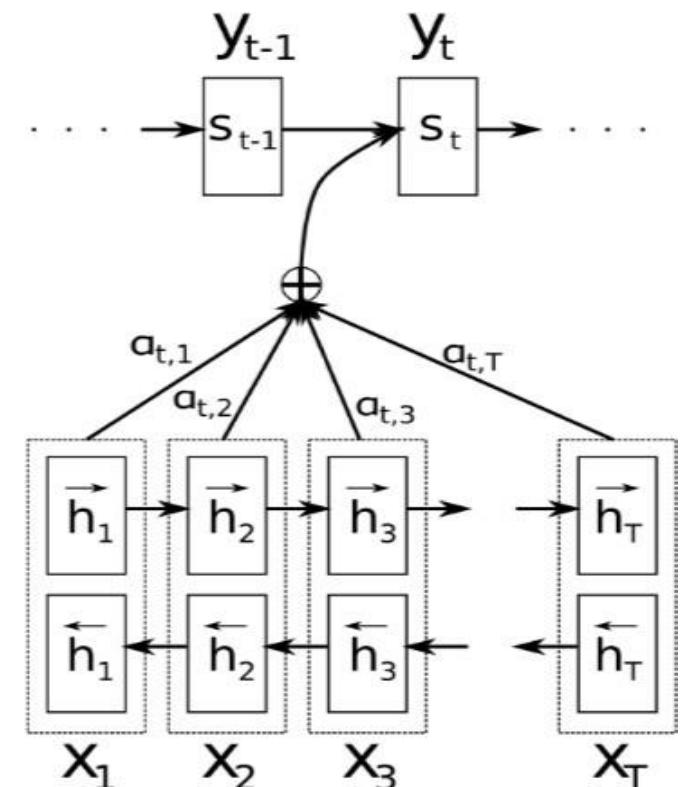
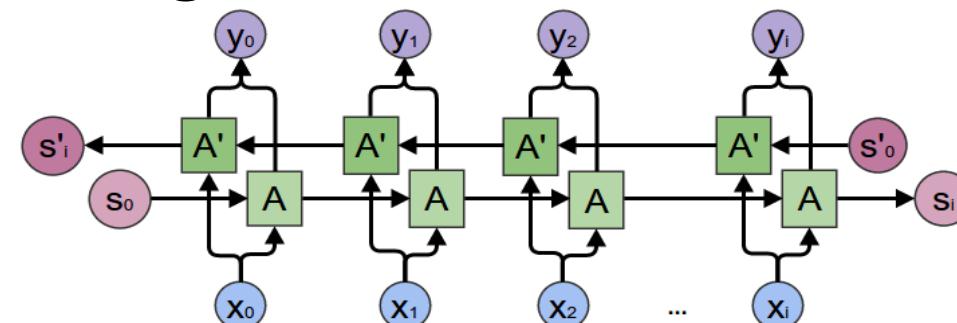


Basic Generation Model

- Sequence-to-sequence
 - Model from MT

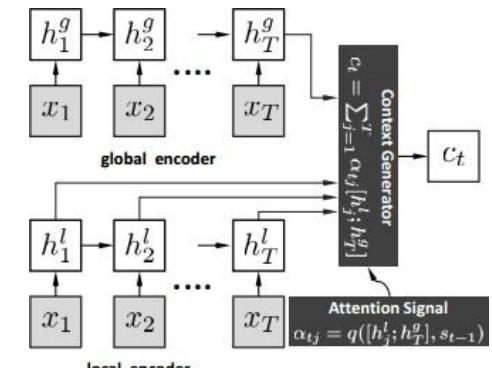
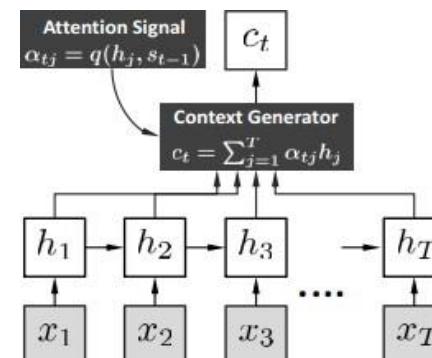
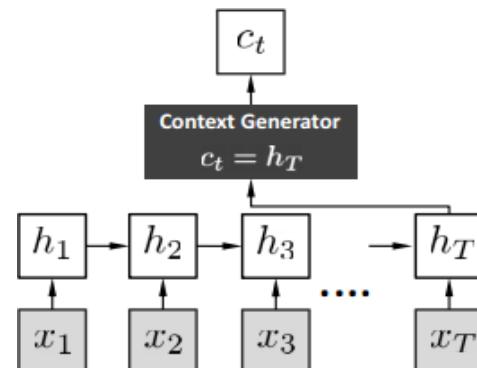
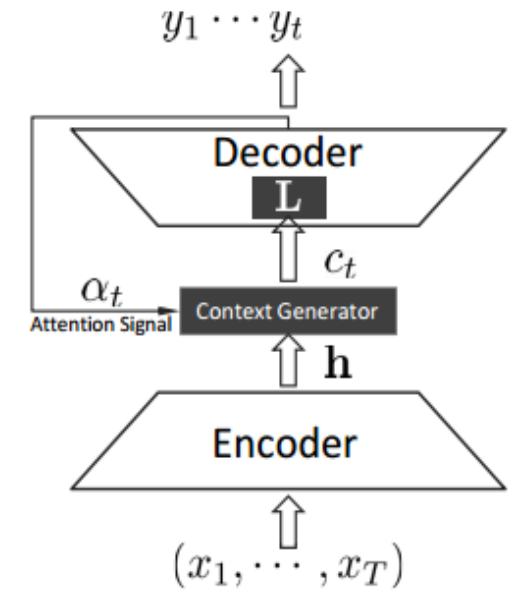


- Attention mechanism
- Bi-directional modeling

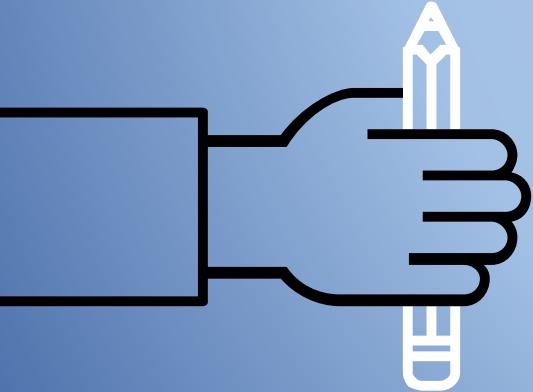
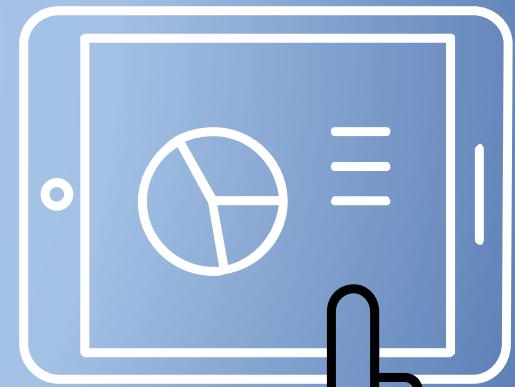
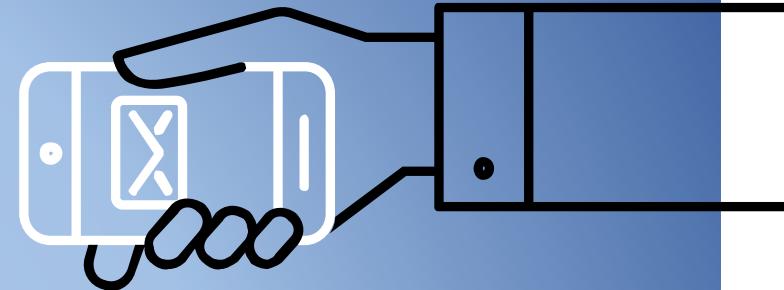
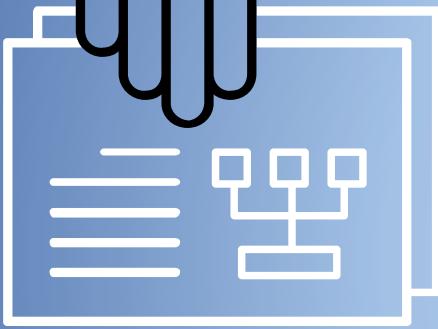


Extension from Seq2Seq

- Neural responding machine [Shang et al., ACL'15]
- Encoding-decoding framework
- Model variants with different attention
 - Local
 - Global
 - Hybrid



Multi-Turn Generation



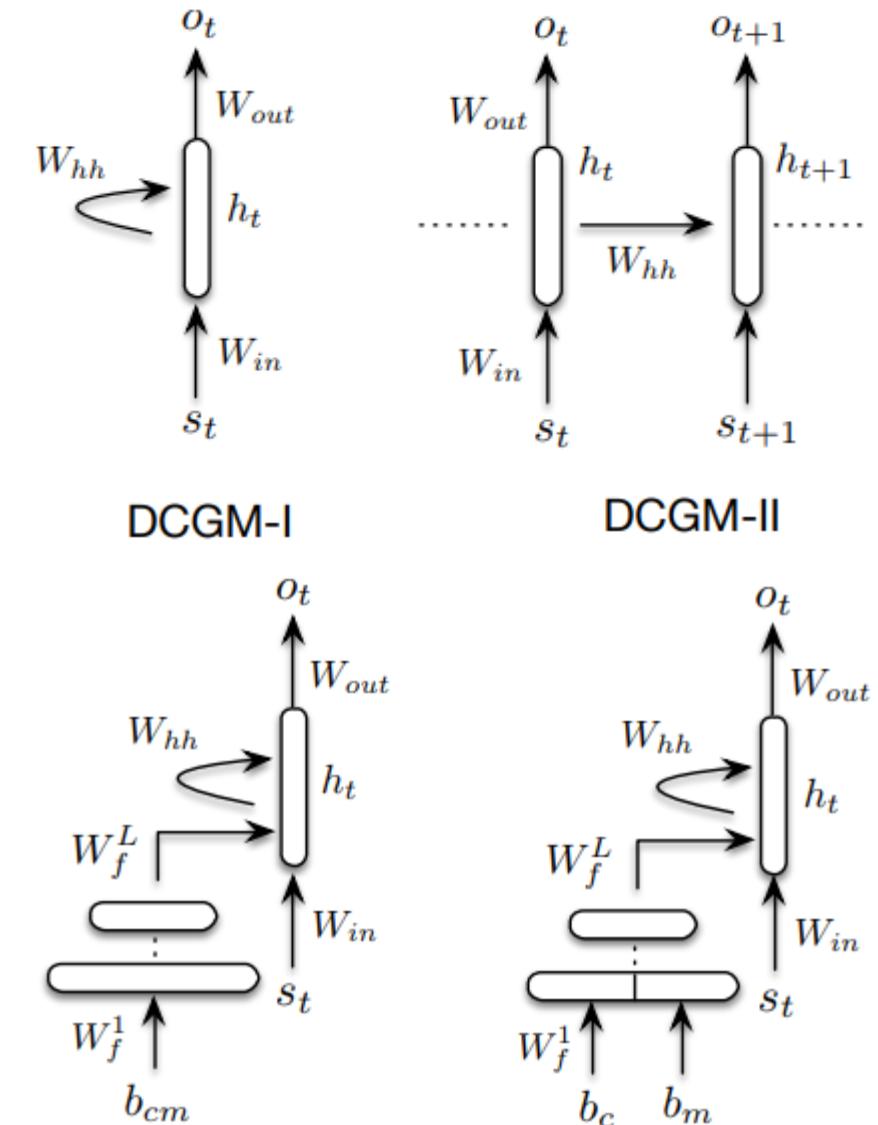
Contexts ARE Important

- A conversation is
 - A consecutive process which includes multiple turns of interactions
 - A session with previous utterances as contexts



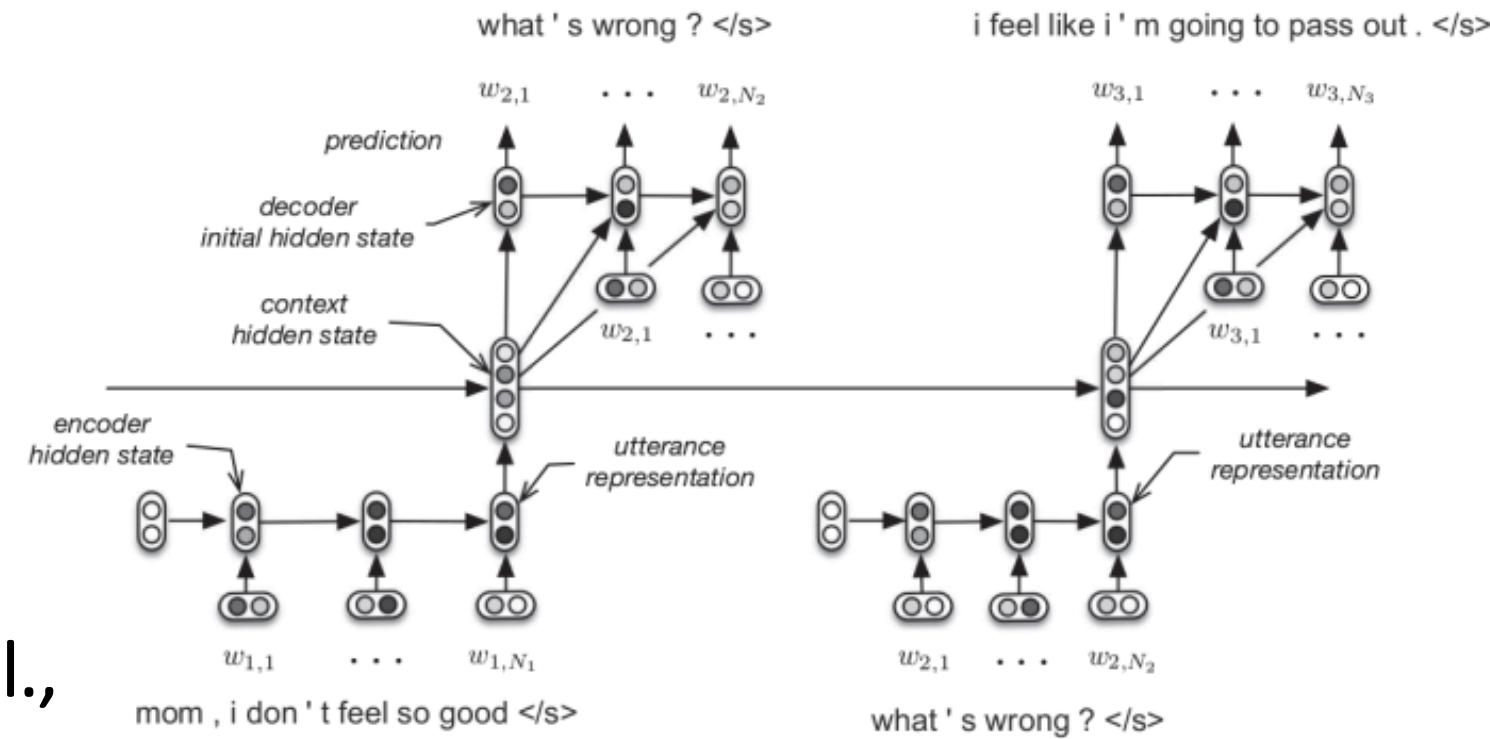
Generation with Contexts

- Context-sensitive models [Sordoni et al., NAACL-HLT'15]
 - *Tripled language model*
 - Concatenate each utterance into a single sentence
 - Problem: long-range representation issue
 - *Dynamic-Context Generative Model*
 - Fixed-length vector: bag-of-words through MLP
 - With order or without order information preserved in context and message



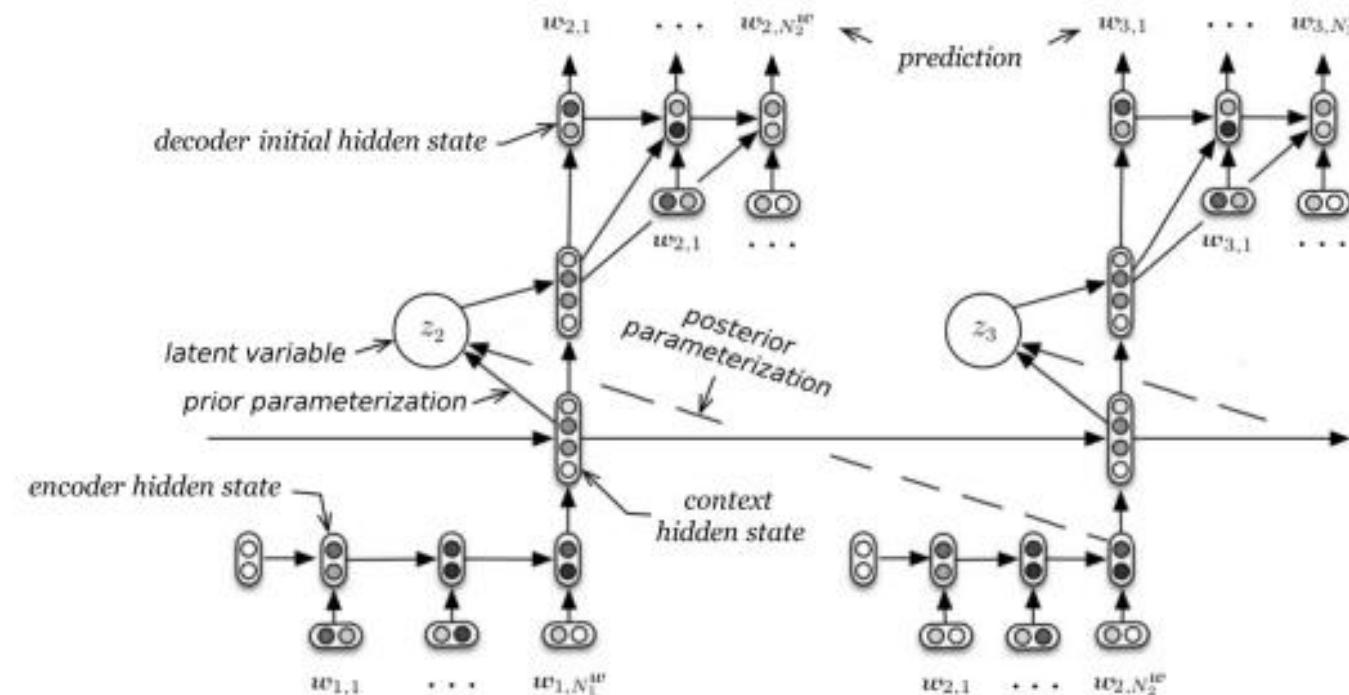
Hierarchical Context Modeling

- Languages are represented as hierarchies [Li et al., ACL'15]
 - Word-level representation AND Sentence-level representation
- Hierarchical recurrent encoder-decoder, HRED [Serban et al., AAAI'16]
 - Context hidden states
 - Global vector of information
- Hierarchical recurrent attention network [Xing et al., AAAI'18]
 - Word-level attention AND Sentence-level attention



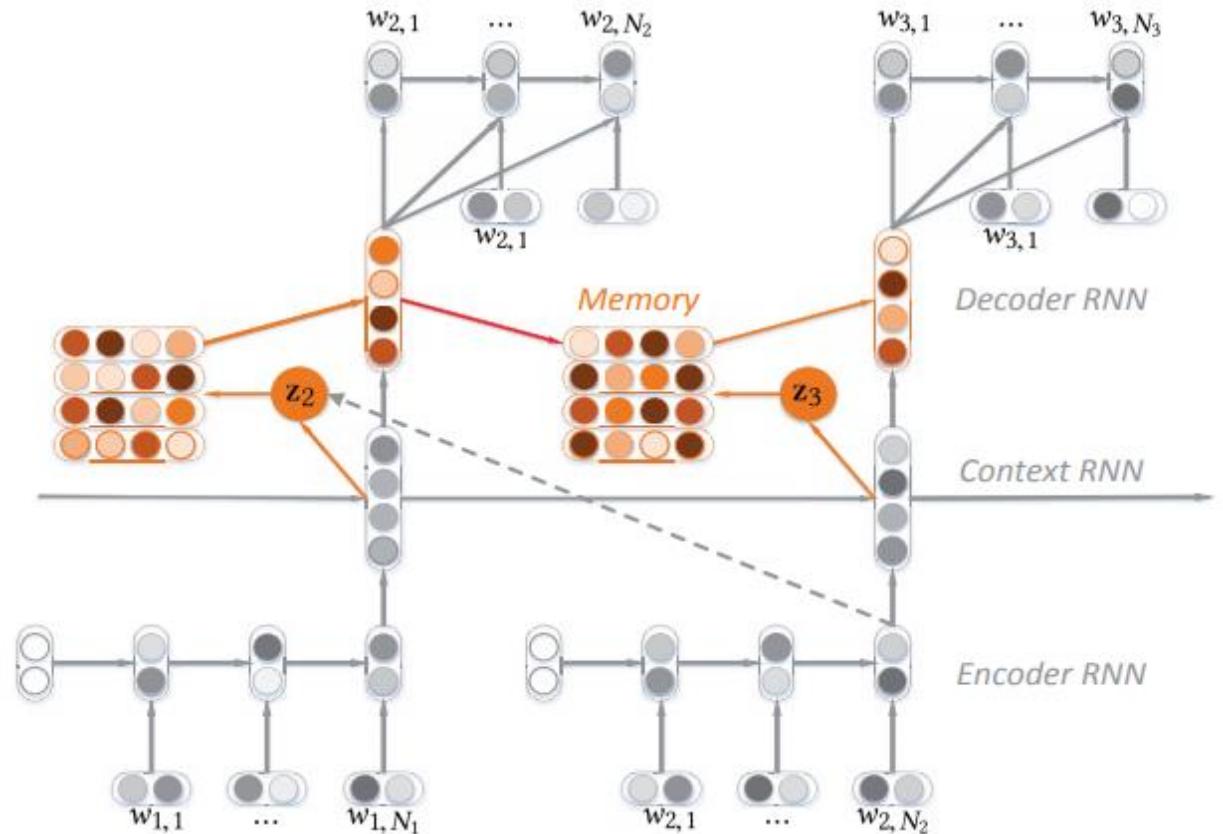
Latent Variable Modeling

- HRED
 - The only source of variation is through the conditional output distribution
- Latent Variable HRED (VHRED) [Serban et al., AAAI'17]
 - A stochastic latent variable z conditioned on all previous observed tokens



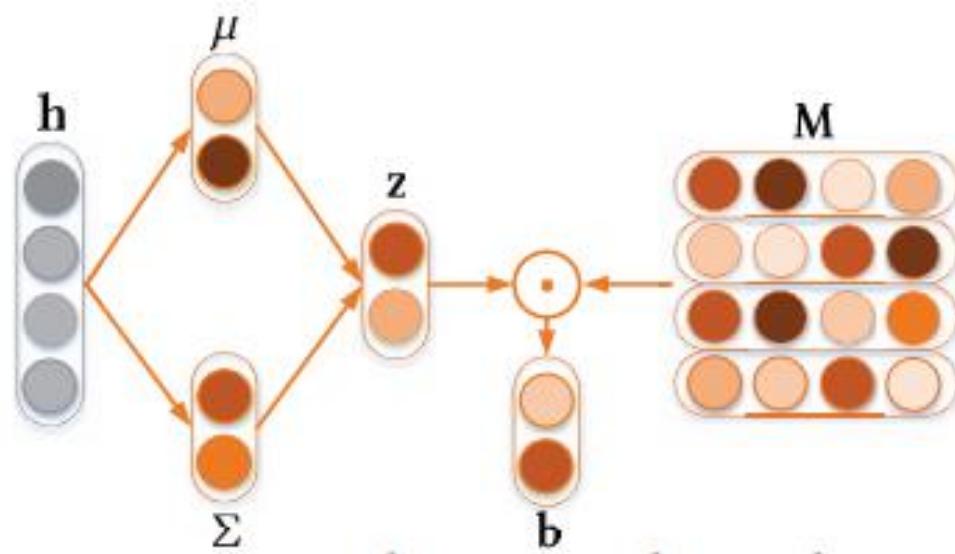
Hierarchical Memory Network

- Extension to memory network for long term dependency [Chen et al., WWW'18]
- Model overview
 - Utterance encoding
 - Variational memory network
 - Decoding

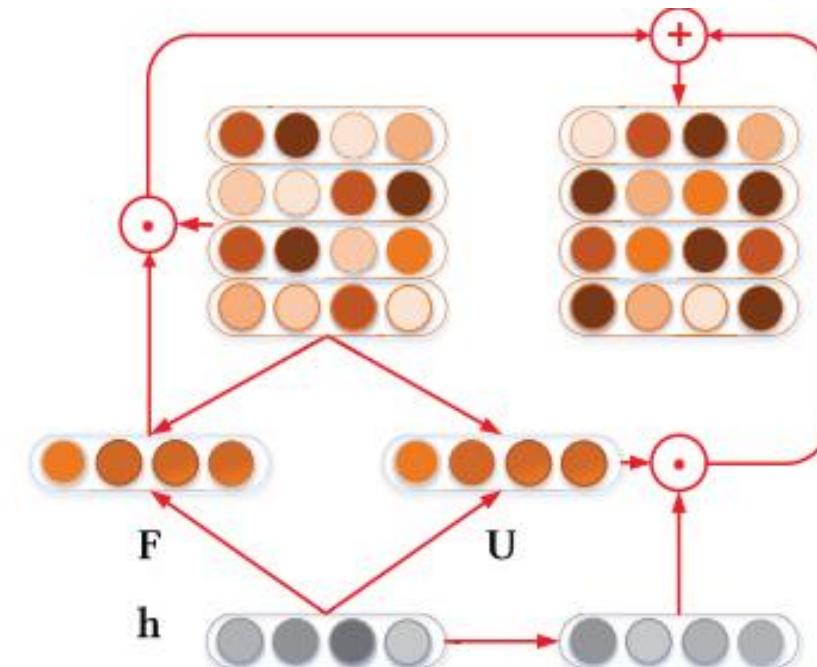


Hierarchical Memory Network (CONT.)

- Memory reading mechanism
 - Reading through latent variable and memory to update bias

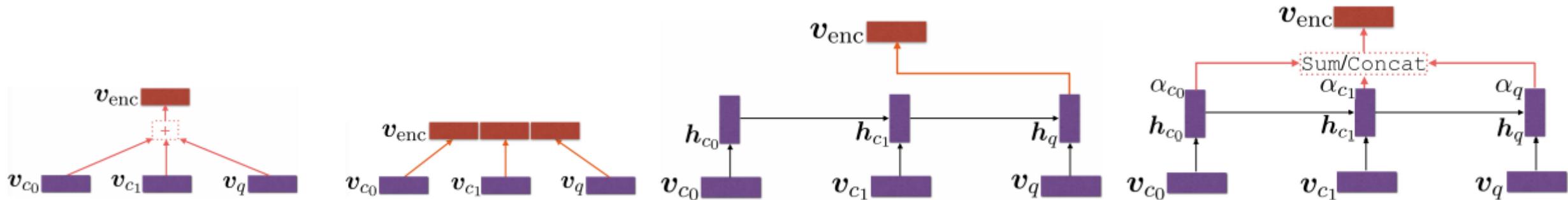
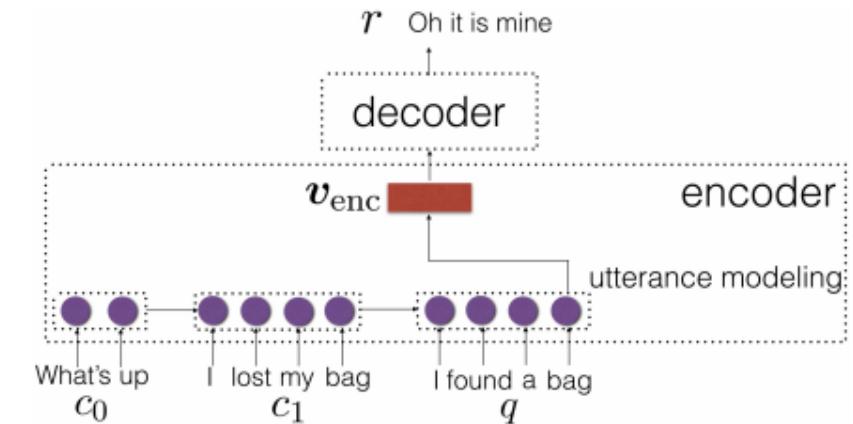


- Memory updating mechanism
 - Forgetting operation
 - Updating operation
 - With the context RNN

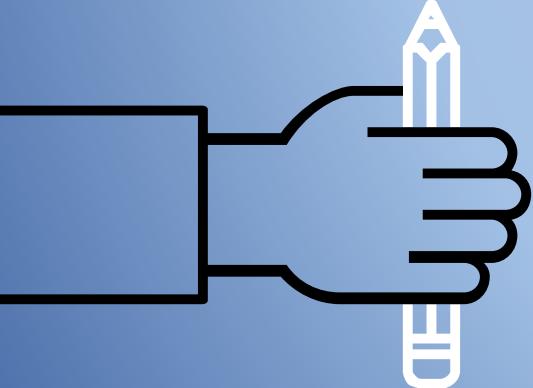
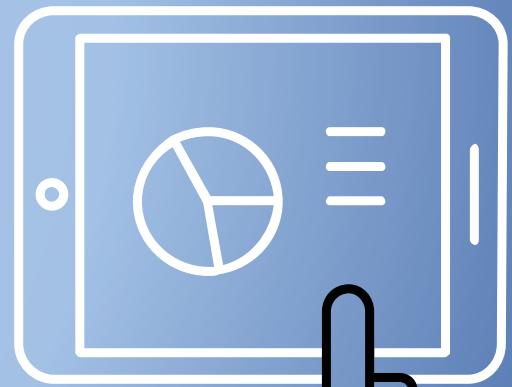
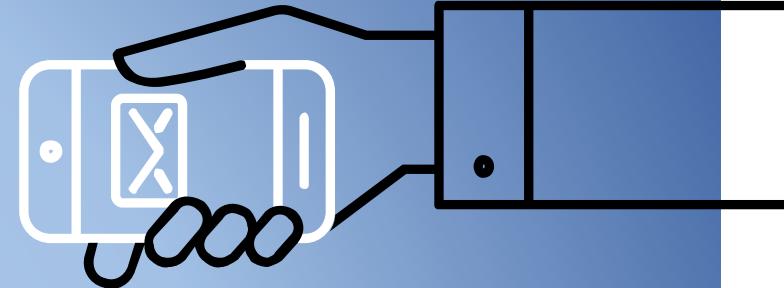
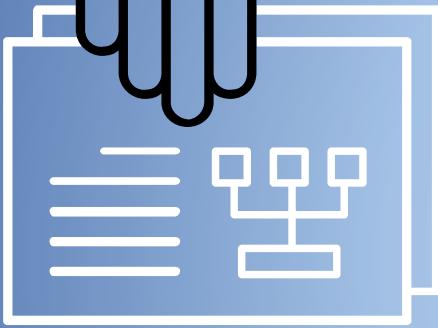


Context Modeling Frameworks

- A framework to unify different structures [Tian et al., ACL'17]
- Non-hierarchical context modeling
- Hierarchical context modeling
 - Sum pooling
 - Concatenation
 - Sequential integration
 - *Weighted sequential integration*



Diversity in Conversations



Why Diversity

- The unique phenomenon in conversations
 - One-to-many
 - Example
- Diversity issue can be related to either single-turn generation or multi-turn generation
 - Context modeling strategies also applies if any

Objective-Driven Diversity

- Why lacks of diversity in end-to-end generation models
 - Driven by data if not supervised in particular
 - “I don’t know” and “me too” take up a big proportion in training datasets
- Traditional objective function
 - Maximize the target probability conditioned on the source sequence
- New objective function with penalty [Li et al., NAACL-HLT’16]
 - A penalty factor

$$\hat{T} = \arg \max_T \{ \log p(T|S) - \lambda \log p(T) \}$$

Input: What are you doing?

1. I've been looking for you.
 2. I want to talk to you.
 3. Just making sure you're OK.
-

Input: What is your name?

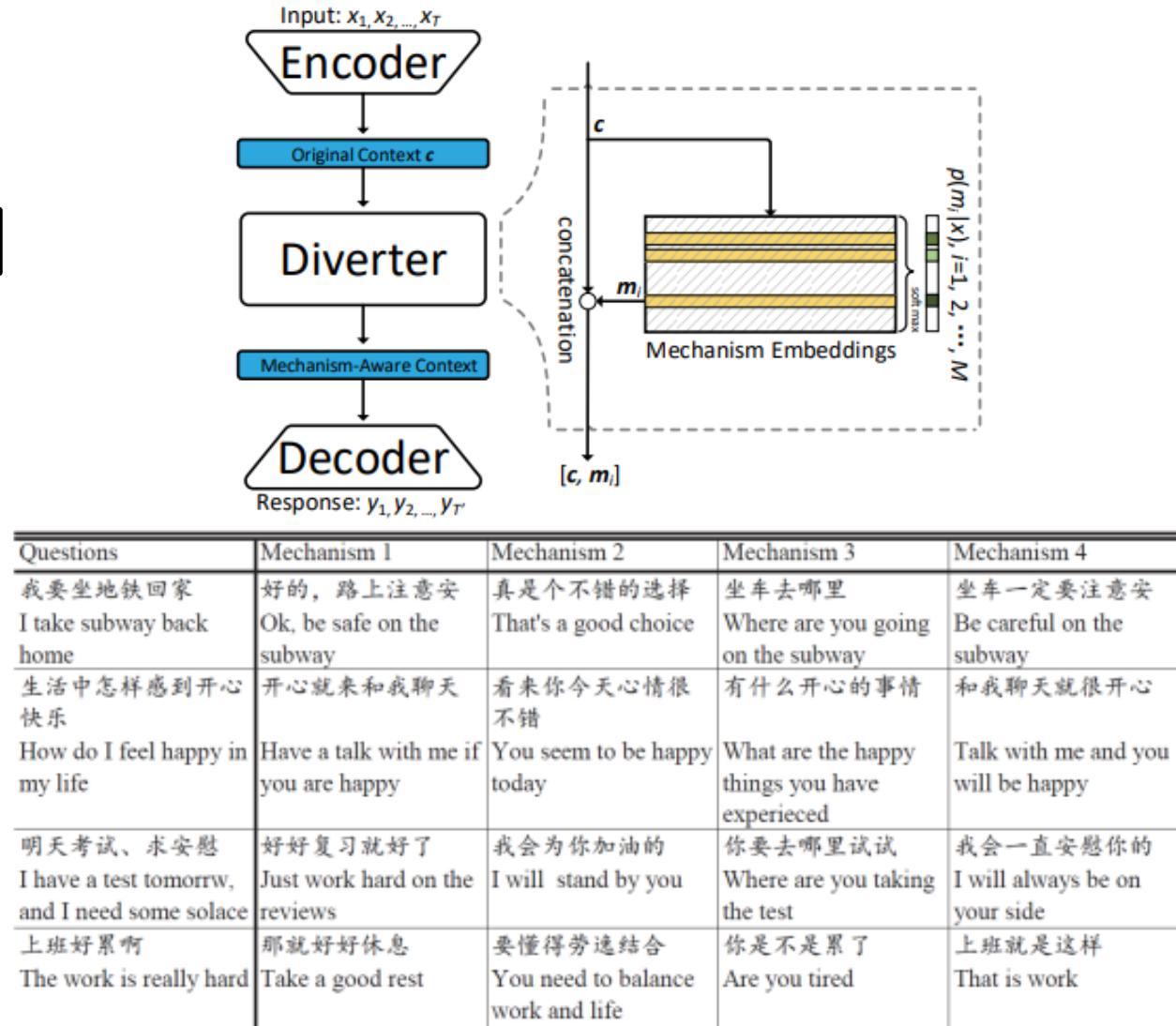
1. Blue!
 2. Peter.
 3. Tyler.
-

Input: How old are you?

1. Twenty-eight.
 2. Twenty-four.
 3. Long.
-

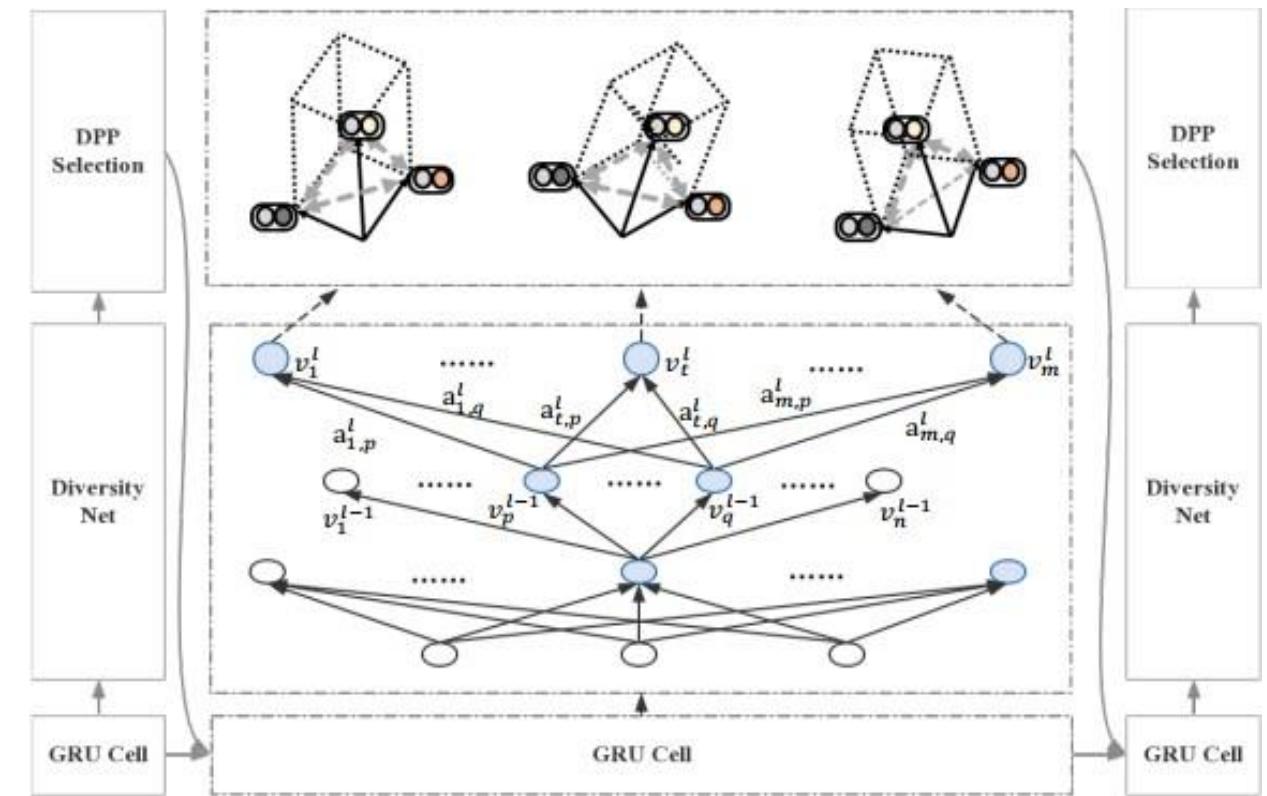
Mechanism-Aware Conversation

- Mechanism [Zhou et al., AAAI'17]
 - Indicates the latent semantic subspace
 - Hidden categories of utterances
- Incorporating a diverter
 - Seq2Seq -> Seq2Mechanism2Seq



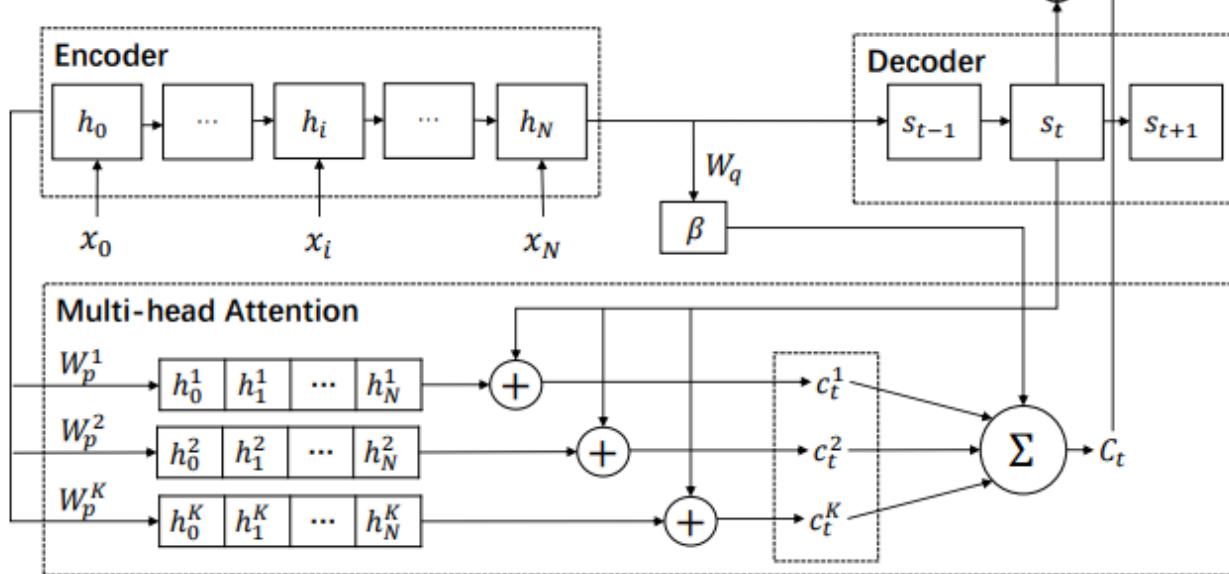
Diversity in Rankings

- Generation with candidates as a ranking process
 - Rankers vs Decoders
- Determinantal Point Process
 - A diversity-promoting algorithm: balancing quality and diversity
- DPP-Augmented Ranker vs DPP-Augmented decoder [Song et al., AAAI'18]



Interpretable Diversity in Conversations

- Why is there diversity in conversations?
- Incorporating multi-head attention schema into conversations [Tao et al., AAAI'18]



Query: 今天下雨，我们一起去吃火锅吧！

It's rainy today, let's go to eat hot pot!

Candidate 1: 我觉得不应该出门，还是在家做饭吧！

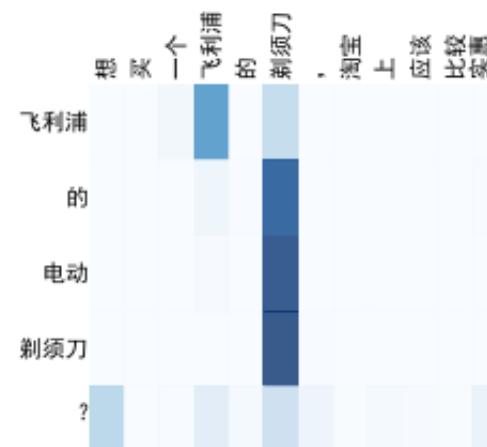
I think we shouldn't go out, let's cook at home!

Candidate 2: 好啊，我很久没吃火锅了。

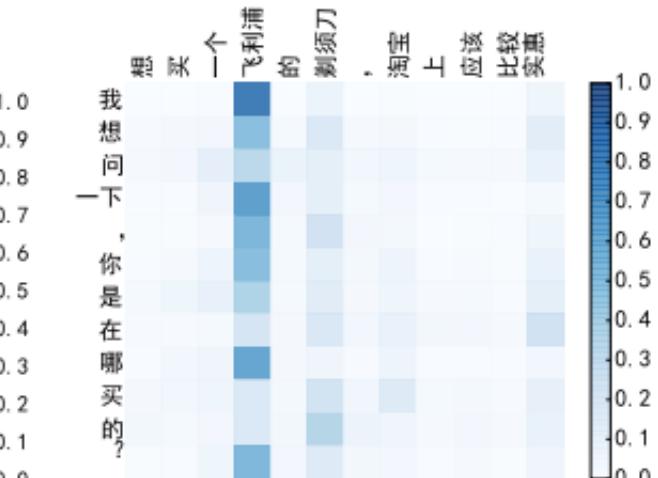
Ok, I have not eaten the hotpot for a long time.

Candidate 3: 开车去还是坐地铁？

drive or take the subway?



(a) Head-1

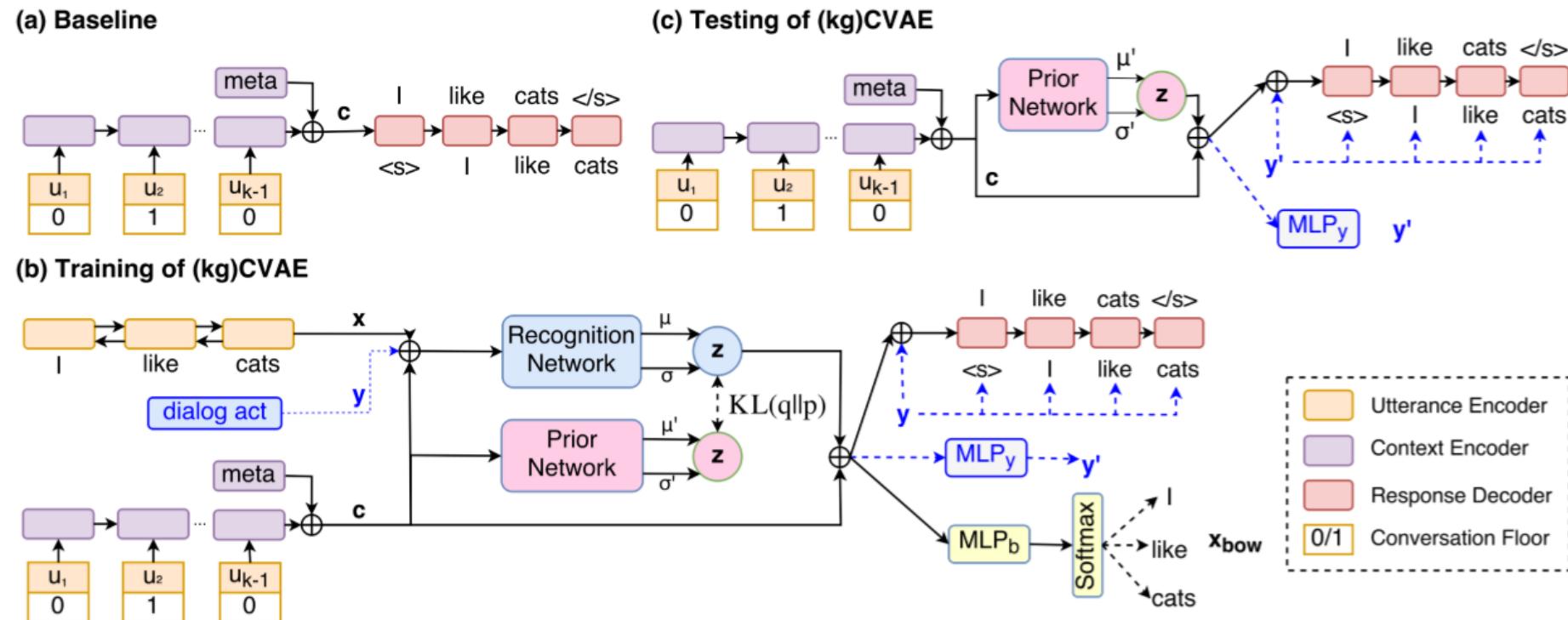


(b) Head-2

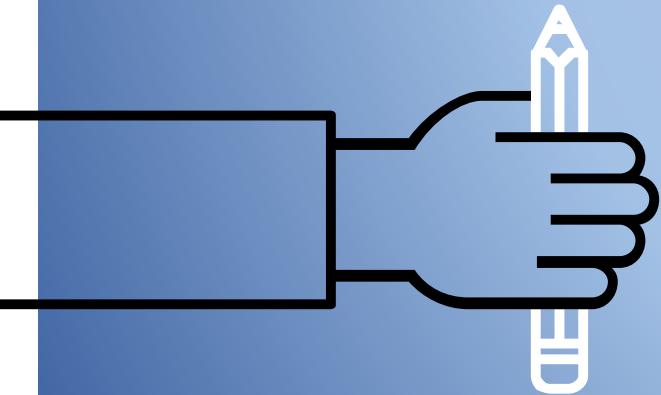
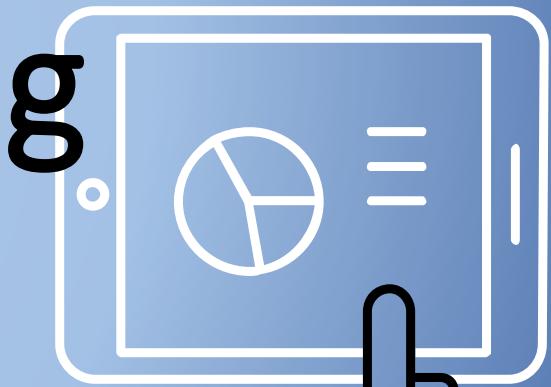
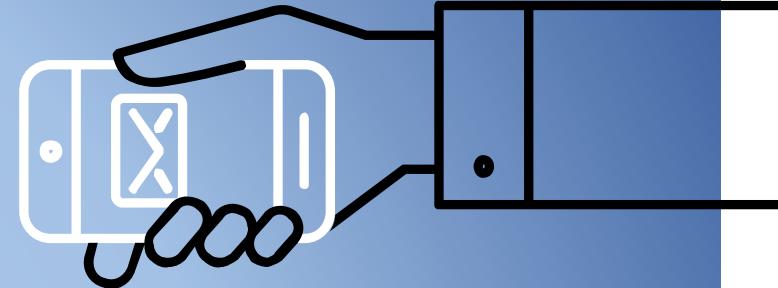
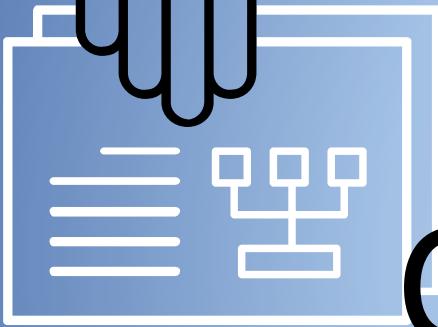
Diversity with Dialogue Act

- Conditional VAE

- Captures the discourse-level diversity during encoding
- Knowledge-guided CVAE by dialogue acts [Zhao et al., ACL'17]



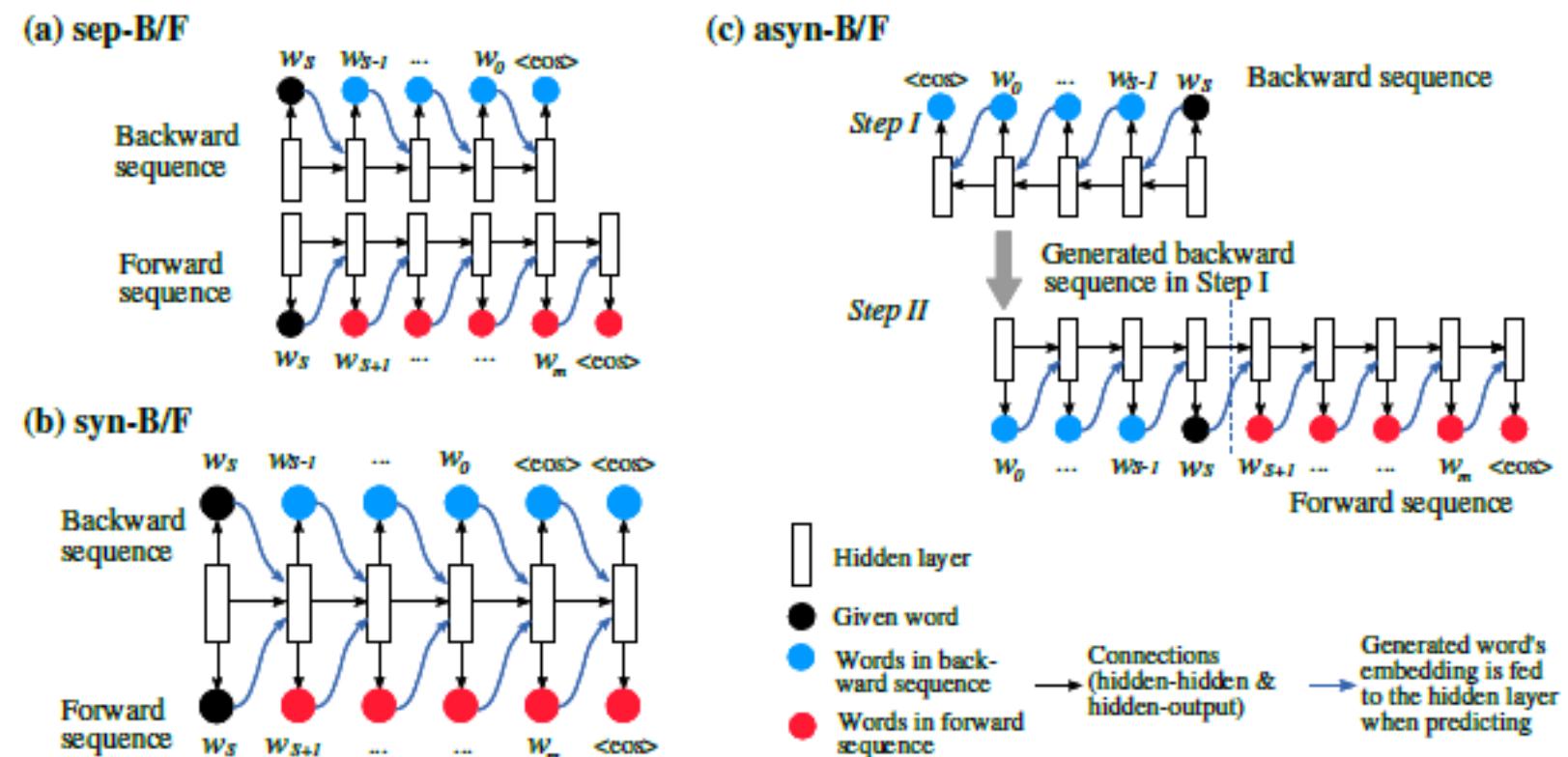
Content-Introducing



Diversity with Dialogue Act

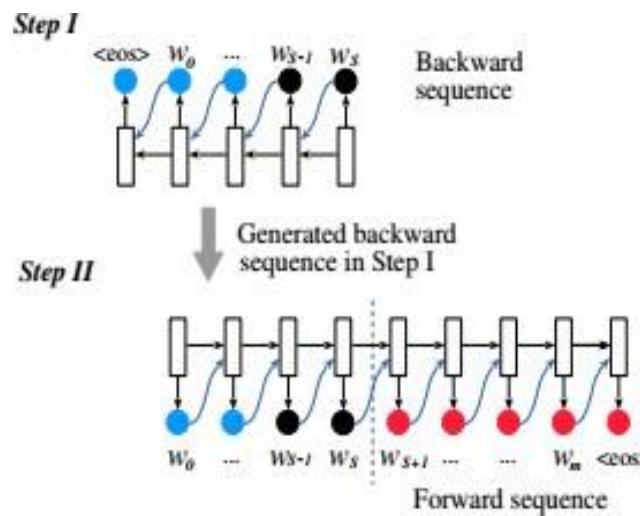
- To generate a sentence with constraint words
- Model variants

- Sep-B/F
- Syn-B/F
- Asyn-B/F

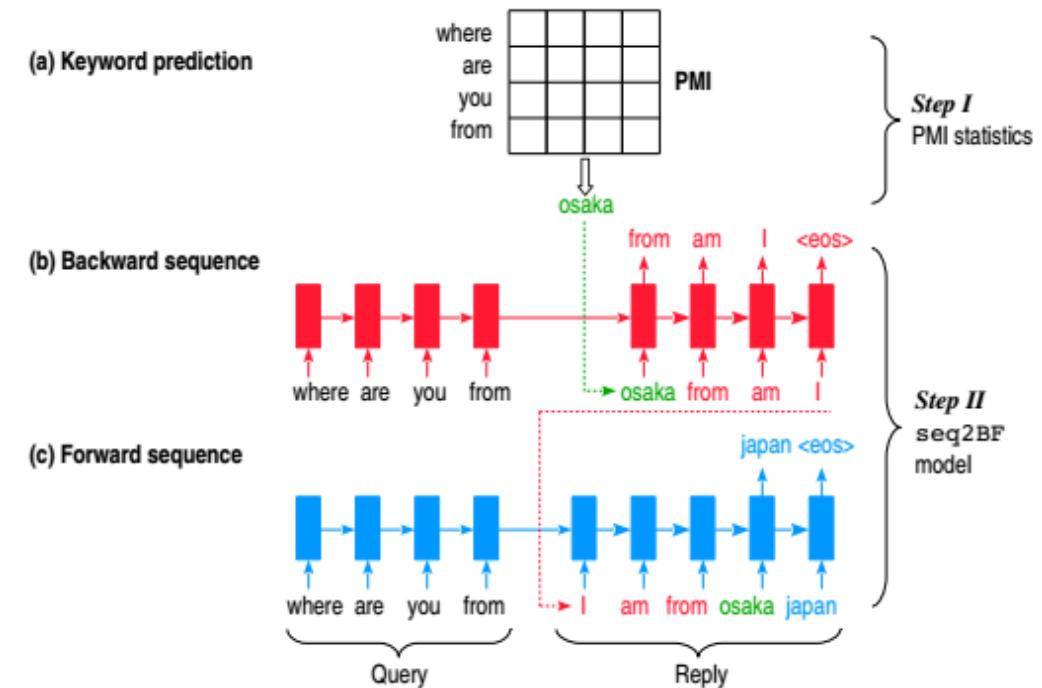


Hard Constraint in Generation

- Controllable generation [Mou et al., COLING'16]
 - Backward and forward generation
- Two-step response generation
 - Keyword generation
 - Hard constraint in generation



- A phrase specified in advance as the constraint
- Words generated by the backward generator
- Words generated by the forward generator



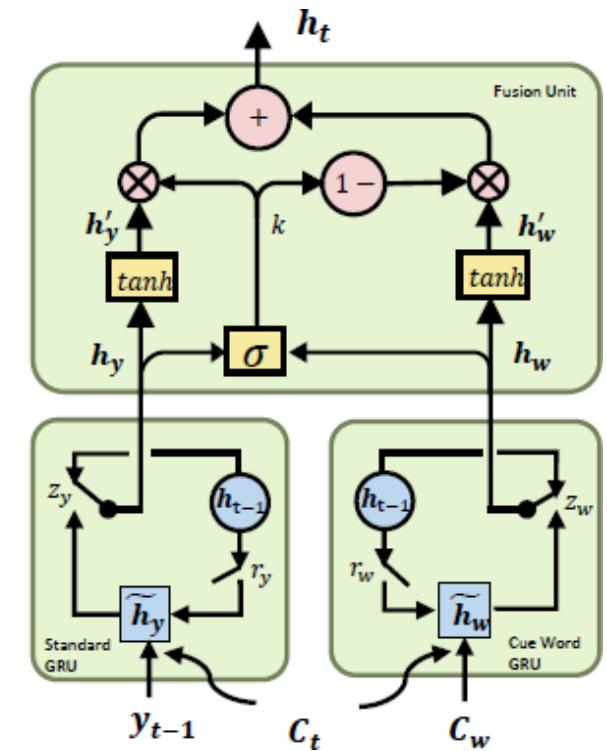
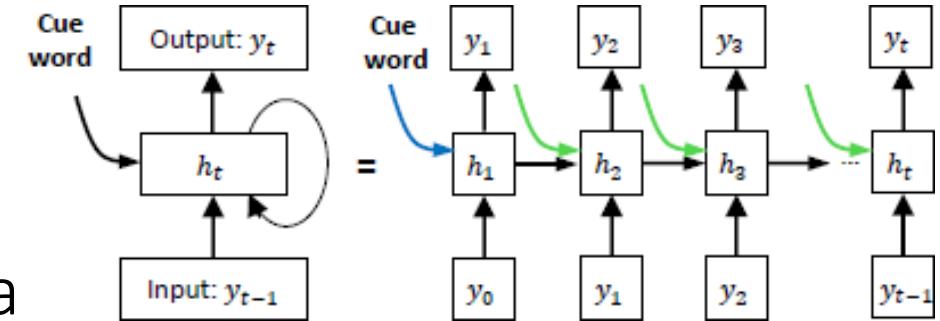
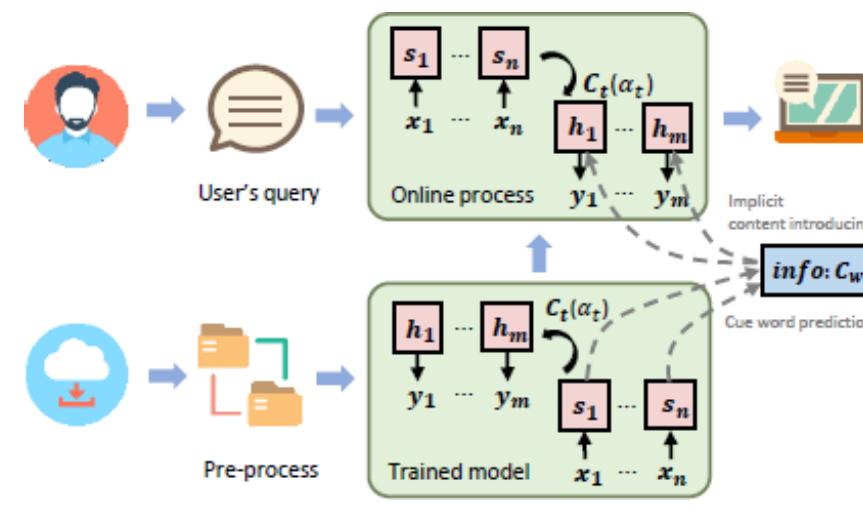
Content Introducing Cases

	Chinese	English (translated)
Query	李有男友公开过了	It's known that <i>Li</i> [†] has a boyfriend.
Grountruth	都已经分了 之前李的贴吧都在讨论了	Broken up. There's discussion in her <i>Tieba</i> . [‡]
seq2seq	我是男的	I am a male
seq2BF_-	我的男友	My boyfriend
seq2BF_+	有绯闻男友	Has a rumored boyfriend
Query	人大复试飘过	Passed second-round exam of Renming Univ.
Grountruth	这么牛，什么专业	Cool, what's your major
seq2seq	我也是！	Me too!
seq2BF_-	我也是！	Me too!
seq2BF_+	分数是什么	What is your score
Query	挺漂亮的 祝福祝福	So beautiful, congratulations
Grountruth	下手慢了哈。现在	You're late till now
seq2seq	谢谢	Thank you
seq2BF_-	是的，谢谢	Yes, thank you
seq2BF_+	谢谢夸奖	Thank you for praising
Query	我要换头像了！	Want to change a photo
Grountruth	打算换成什么啊~	What do you like...
seq2seq	我是你的头像	I'm in your photo
seq2BF_-	我是你的头像	I'm in your photo
seq2BF_+	第一张图像是谁	Who is in your first photo

From Hard Constraint to Soft Constraint

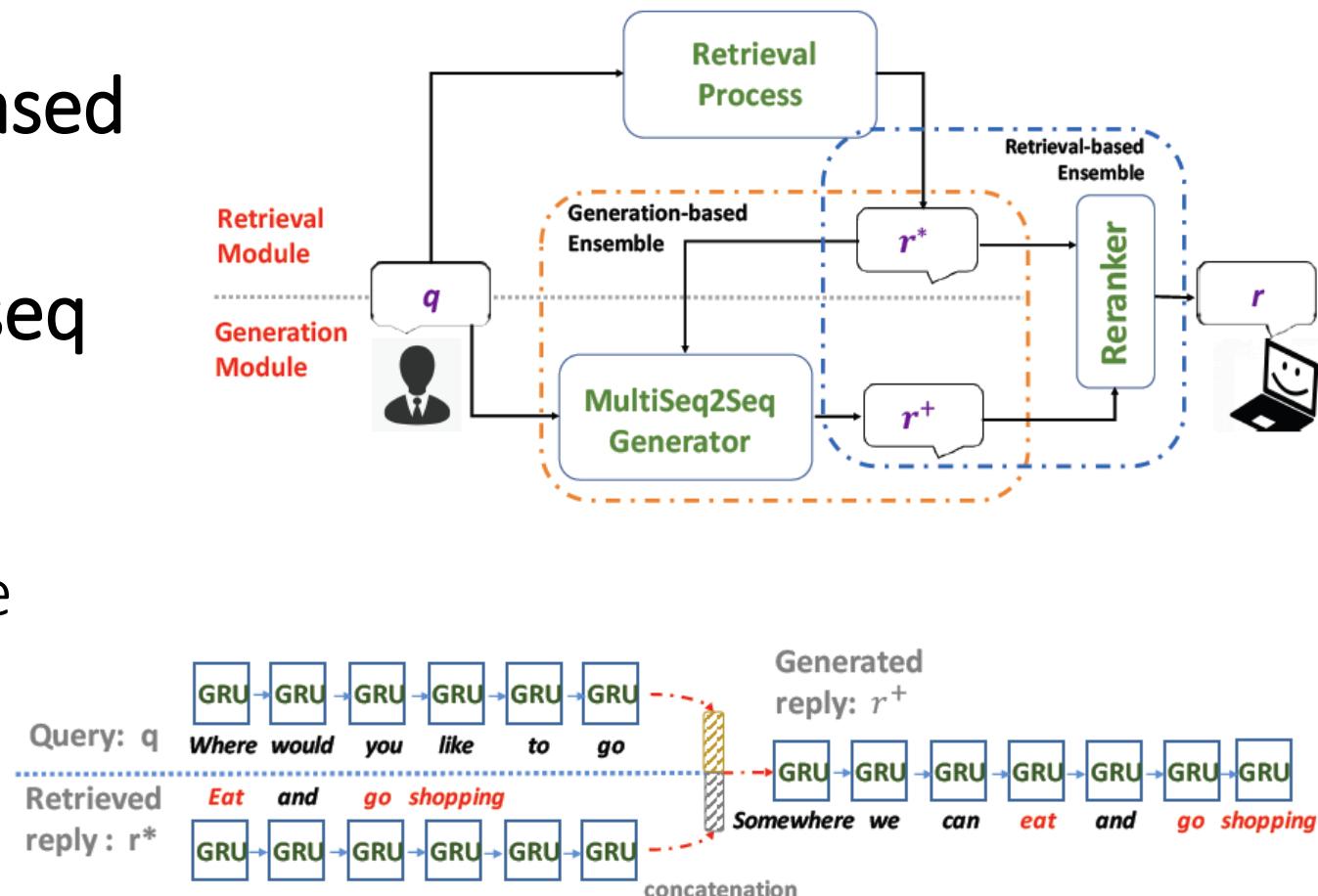
- Implicit content introducing to generate responses [EMNLP'17]

- Sometimes hard constraints are not necessary
- Semantics can be incorporated through decoding
 - Standard GRU cell
 - Cue word GRU cell
 - Fusion units



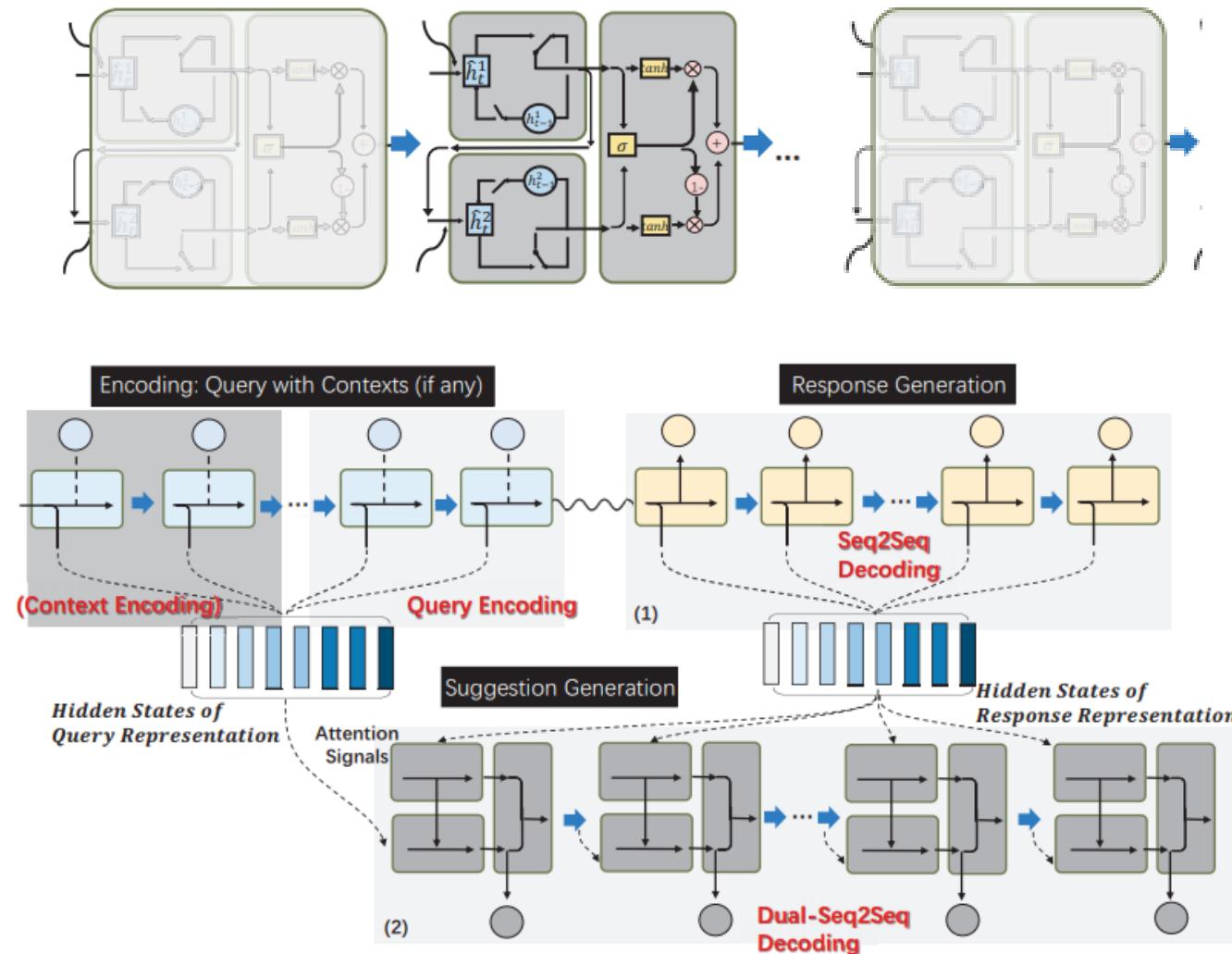
Augmenting Generation with Retrieval

- Pros and Cons in retrieval-based conversations
- Pros and Cons in generation-based conversations
- Optimization using multi-seq2seq [Song et al., IJCAI'18]
 - Retrieved candidates
 - Content augmentation from the multiple sequences

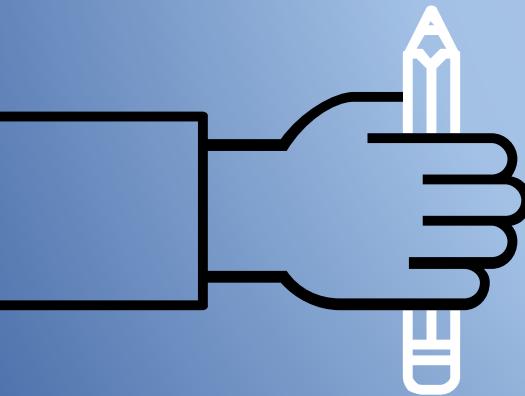
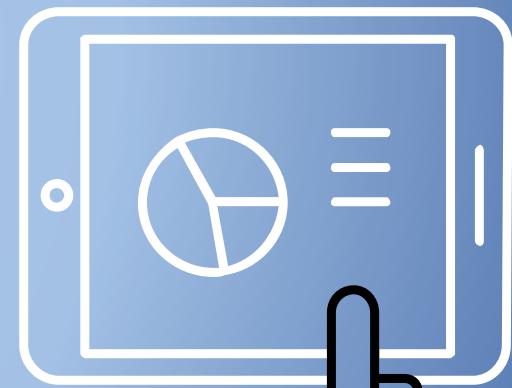
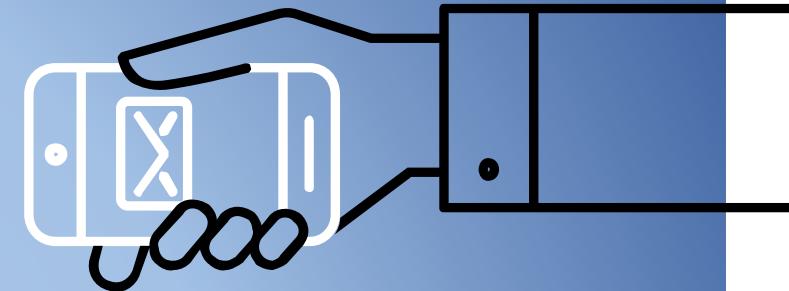
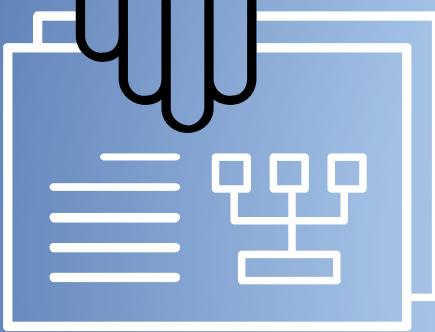


Proactive Suggestion

- A standard conversation paradigm in conversational systems
 - Given a user utterance, a response is provided
- A new conversation paradigm
 - Query utterance, response and suggestion
- Multi-seq2seq process with information fusion as contents [Yan et al., IJCAI'18]

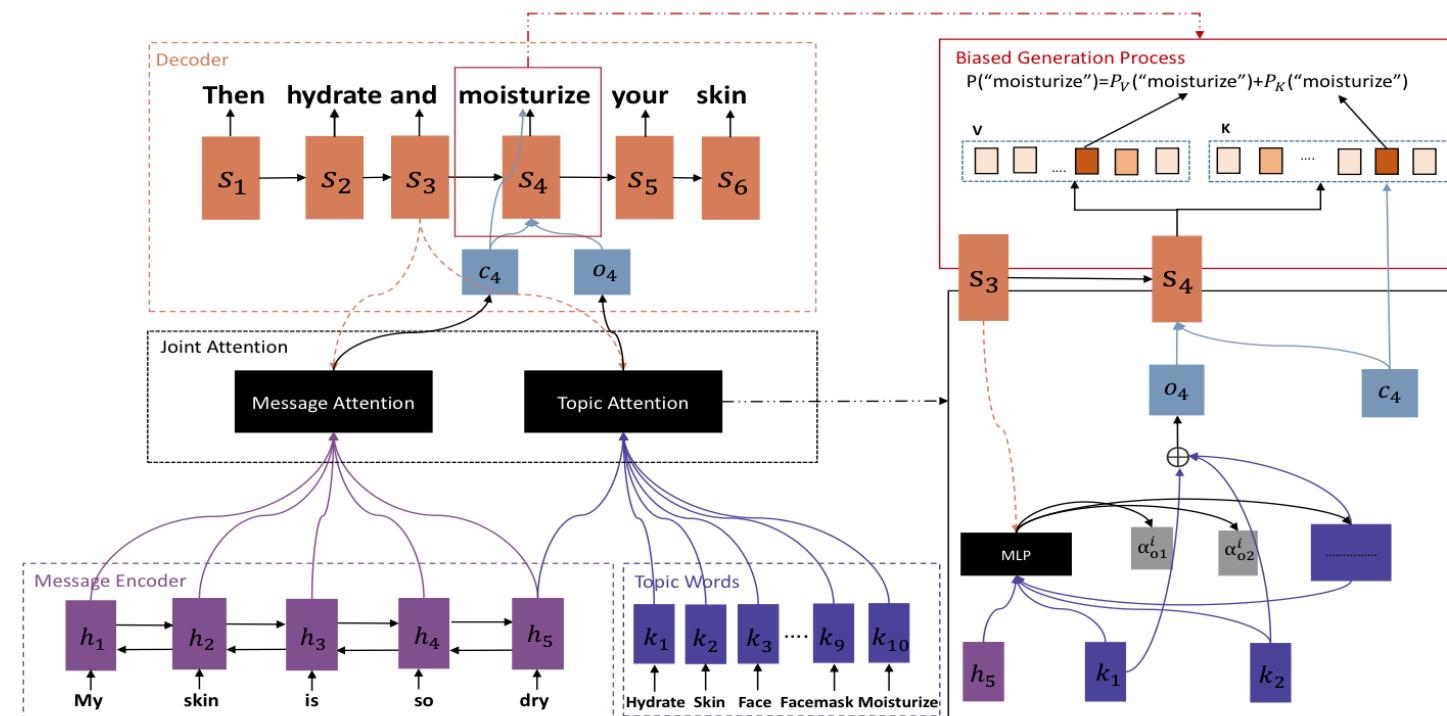


Additional Elements



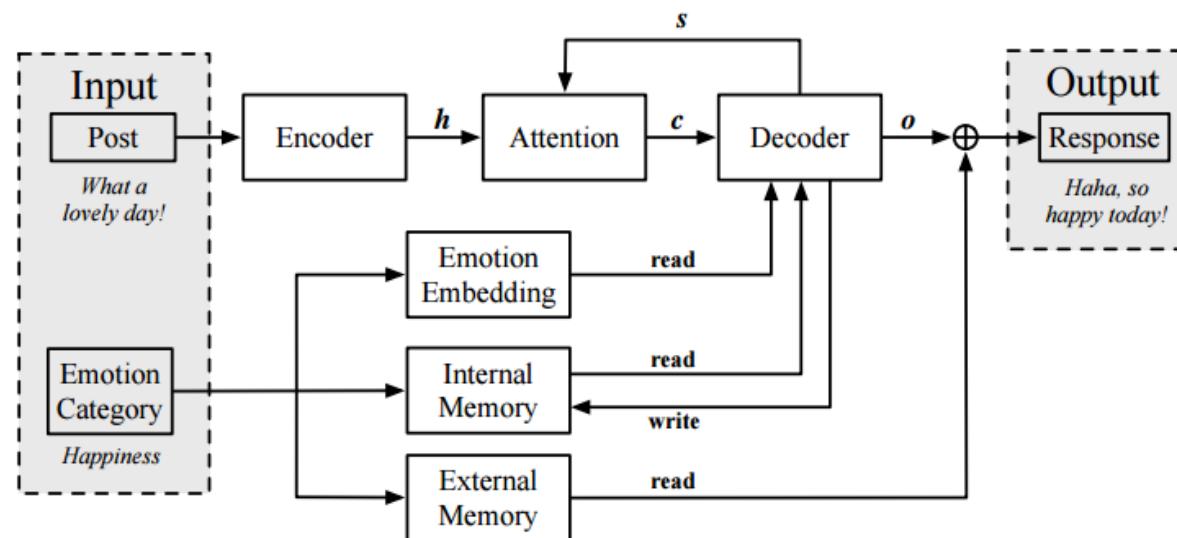
Topics in Conversation

- Topic alignment in conversation is important [Xing et al., AAAI'17]
 - Seq2Seq in the first place
 - With topic alignments
- Decoding with two components



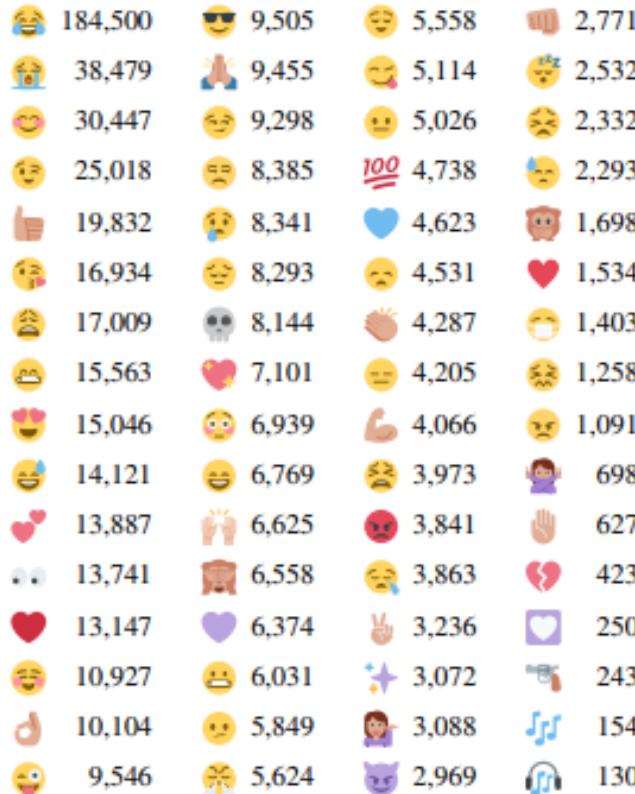
Emotions

- Emotional chatting machine [Zhou et al., AAAI'18]
- Model framework
 - Emotion classifier
 - Emotion expression with an internal memory
 - Emotion control with an external memory

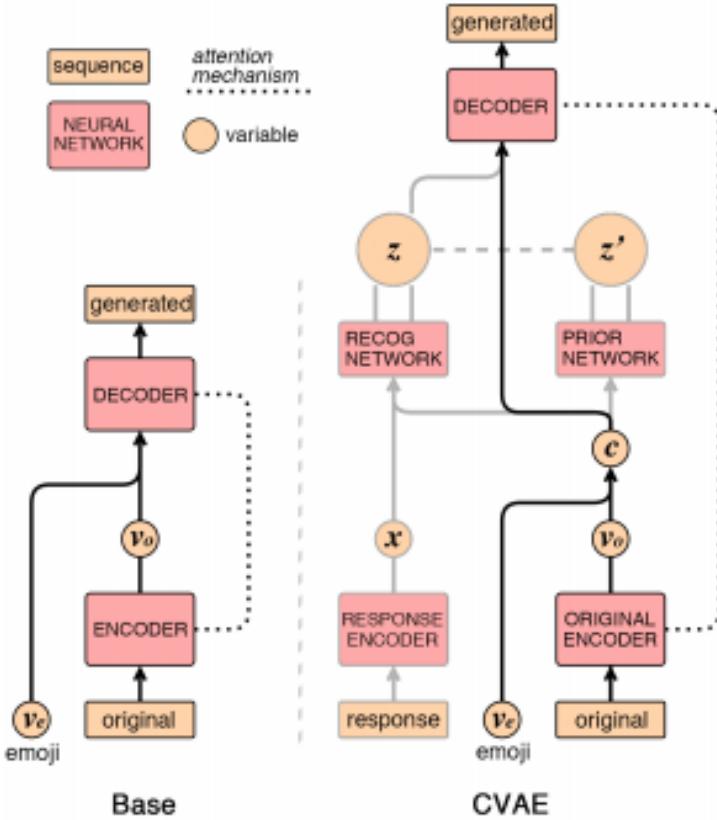


Post	Model	Response(Original)		Response(Translated)	
		Seq2Seq	ECM	Like	Happy
承诺，就是一个骗子说给一个傻子听的。 A promise is what a cheater says to a fool.	Seq2Seq	承诺，就是一个骗子说给一个傻子听的。	承诺，就是一个骗子说给一个傻子听的。	承诺是一个 专一 的傻子	Promise is a single-minded fool
				我就是 一个开心 的傻子	I'm just a happy fool
				我就是 一个悲伤 的傻子	I am a sad fool
				一个 骗子 说给一个 傻子 听	A cheater says to a fool
				不管你信不信，反正我信了！	Whatever you believe or not, I trust in it!
				知道真相的我眼泪掉下来。	After knowing the truth, I burst into tears.
知道真相的我眼泪笑下来 Knowing the truth, I laughed myself to tears	ECM	Like		太 可爱了 ，太 可爱了 ！	Too lovely , it's too lovely .
		Happy		笑得我眼泪都出来了。	I can't help laugh myself to tears.
		Sad		知道真相的我 眼泪 掉下来。	After knowing the truth, I burst into tears .
		Disgust		这是真的还是假的？	Is this true?
		Angry		什么意思？没看懂。	What do you mean? I don't understand.

MojiTalk

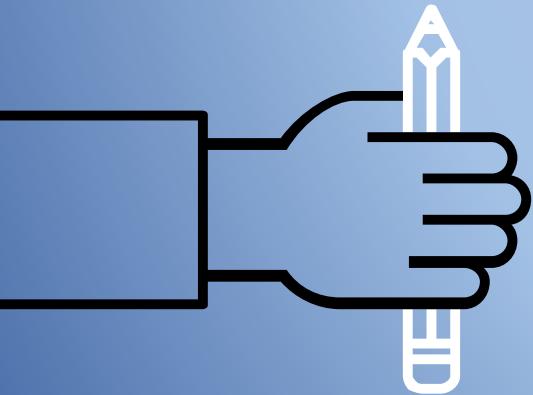
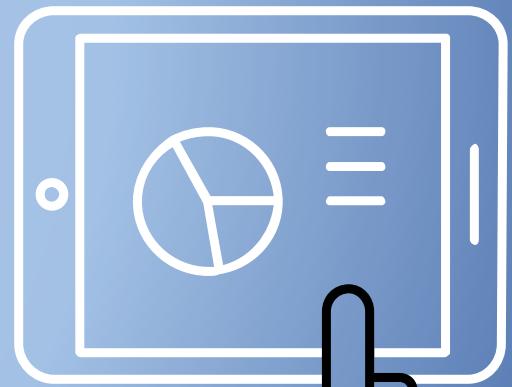
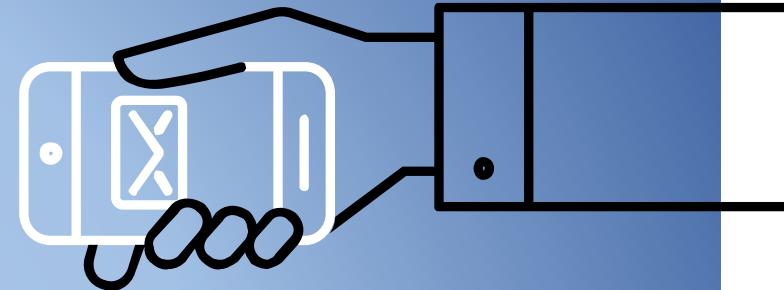
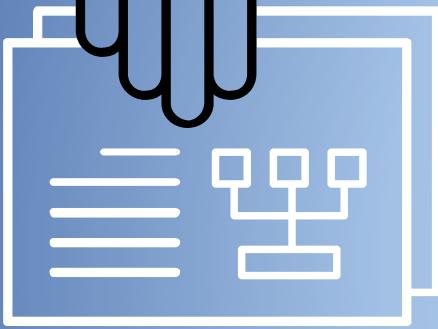


- EMoji distributions in conversations [Zhou et al., ACL'18]
 - Generation model
 - Base: Seq2Seq
 - Incrementally using CVAE onto base model
 - Reinforced CVAE
 - *Emoji classifier as reward*



Content	g i needed that laugh lmfao
Target Emotion	
Base	i 'm glad you enjoyed it
CVAE	good ! have a good time
Reinforced CVAE	thank you for your tweet , you didn 't know how much i guess

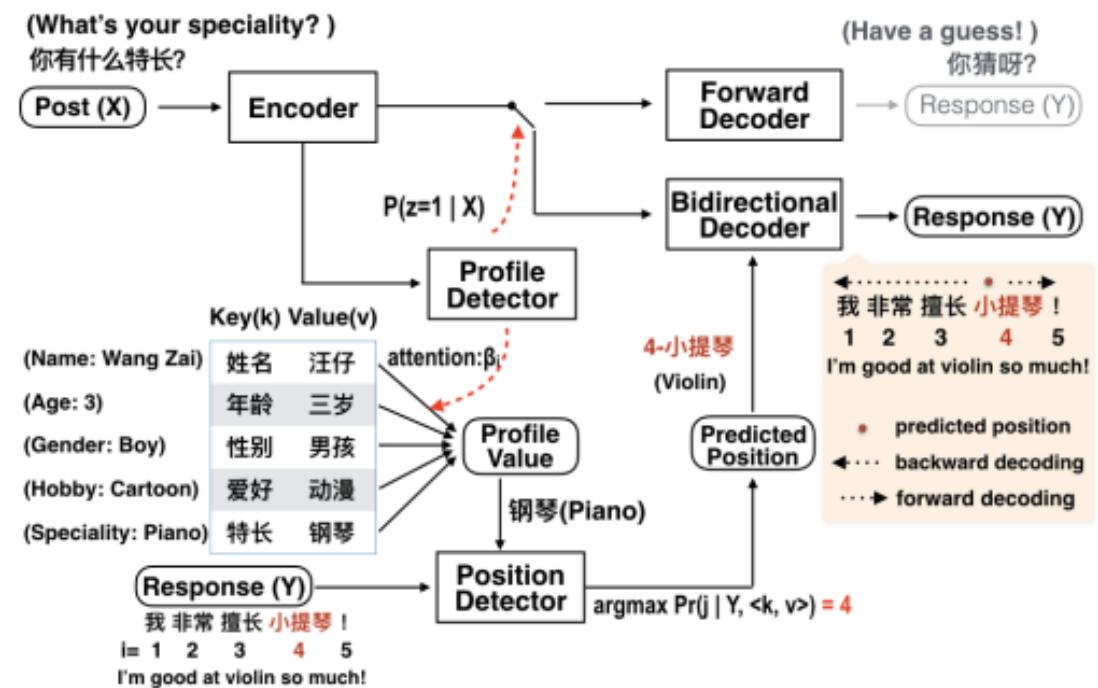
Persona in Chat



Personality Assignment

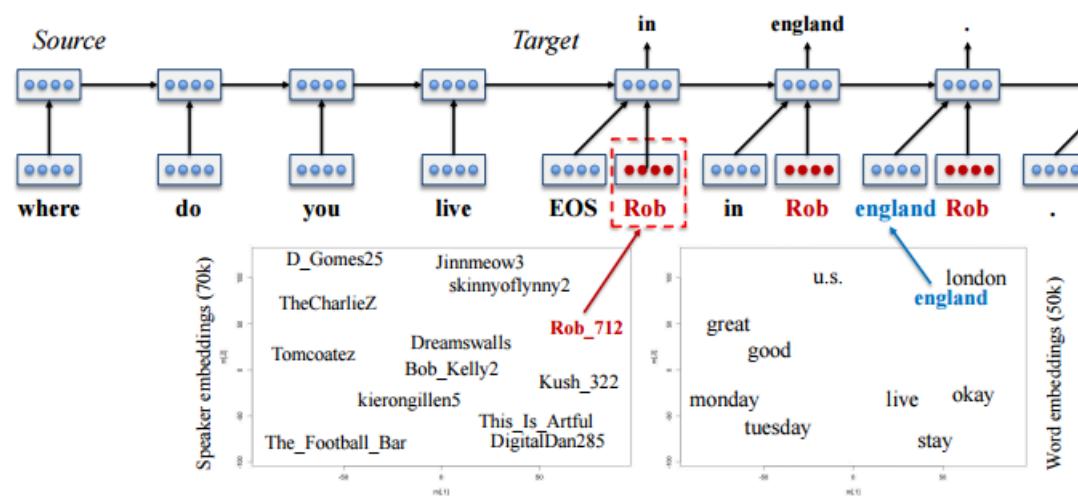
- Personality shall be fixed attributes
[Qian et al., IJCAI'18]
- Personality shall be explicitly expressed
 - Rather than implicitly embedded only
- Model framework
 - Encoder
 - Profile detector
 - Bi-directional decoder
 - (Position detector)

General seq2seq model
User: Are you a boy or a girl?
Chatbot: I am a boy.
User: Are you a girl?
Chatbot: Yes, I am a girl.
Our model with personality
User: Are you a boy or a girl?
Chatbot: I am a handsome boy.
User: Are you a girl?
Chatbot: No, I am a boy.



Persona

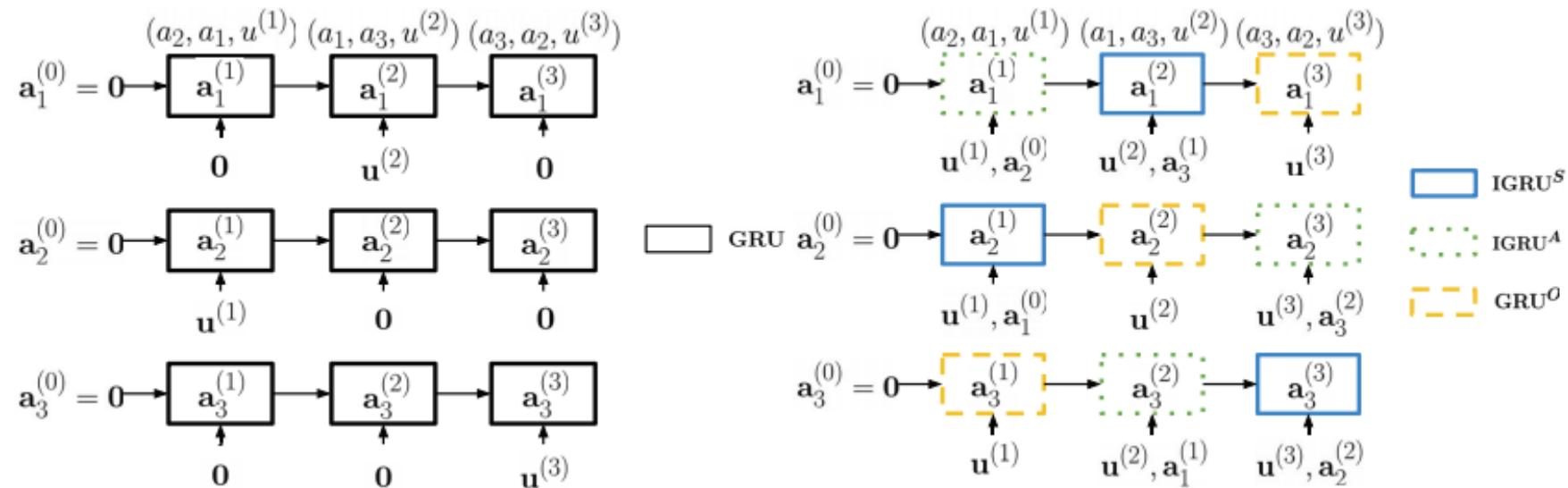
- Consistency is a critical issue
- Persona modeling is one way to solve the conversational consistency problem [Li et al., ACL'16]
- Using language preference as embeddings



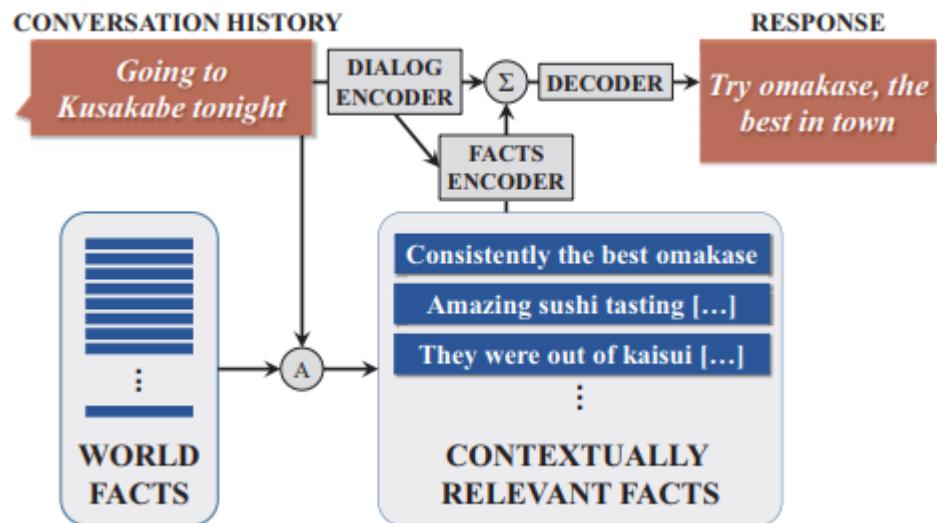
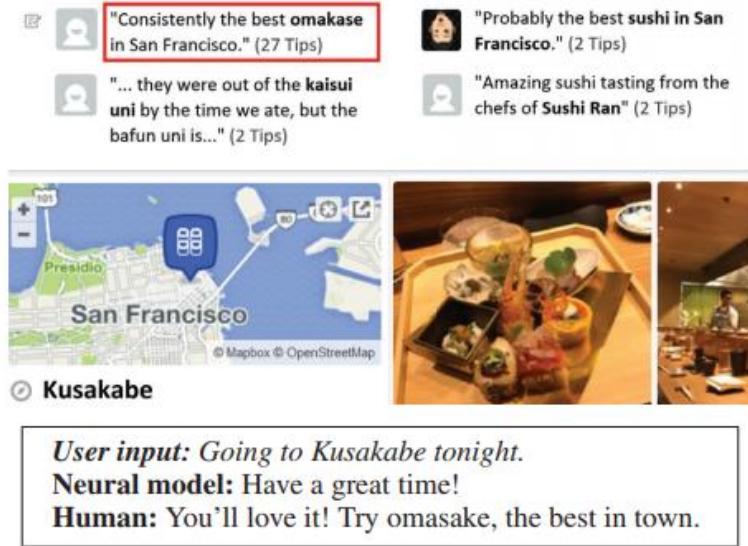
-
- | | |
|-----------------|-----------------------------------|
| <i>message</i> | Where do you live now? |
| <i>response</i> | I live in Los Angeles. |
| <i>message</i> | In which city do you live now? |
| <i>response</i> | I live in Madrid. |
| <i>message</i> | In which country do you live now? |
| <i>response</i> | England, you? |
-
- | | |
|-----------------|------------------------|
| <i>message</i> | Where were you born? |
| <i>response</i> | I was born in Canada. |
| <i>message</i> | Where are you from? |
| <i>response</i> | England, you? |
| <i>message</i> | Where did you grow up? |
| <i>response</i> | I grew up in Texas. |
-
- | | |
|-----------------|------------------|
| <i>message</i> | How old are you? |
| <i>response</i> | 16 and you? |
| <i>message</i> | What's your age? |
| <i>response</i> | 18. |
-
- | | |
|-----------------|--------------------------------|
| <i>message</i> | What is your major? |
| <i>response</i> | I'm majoring in psychology |
| <i>message</i> | What did you study in college? |
| <i>response</i> | English lit. |
-

Speaker-Addressee Modeling

- Dynamic RNN [Ouchi and Tsuboi, EMNLP'16]
 - Speaker embeddings with previous utterances
 - Sentence embeddings are learned based on speaker embeddings
- Speaker-Interaction RNN [Zhang et al., AAAI'18]
 - Interactively learned as different roles: speakers and addressees



Knowledge

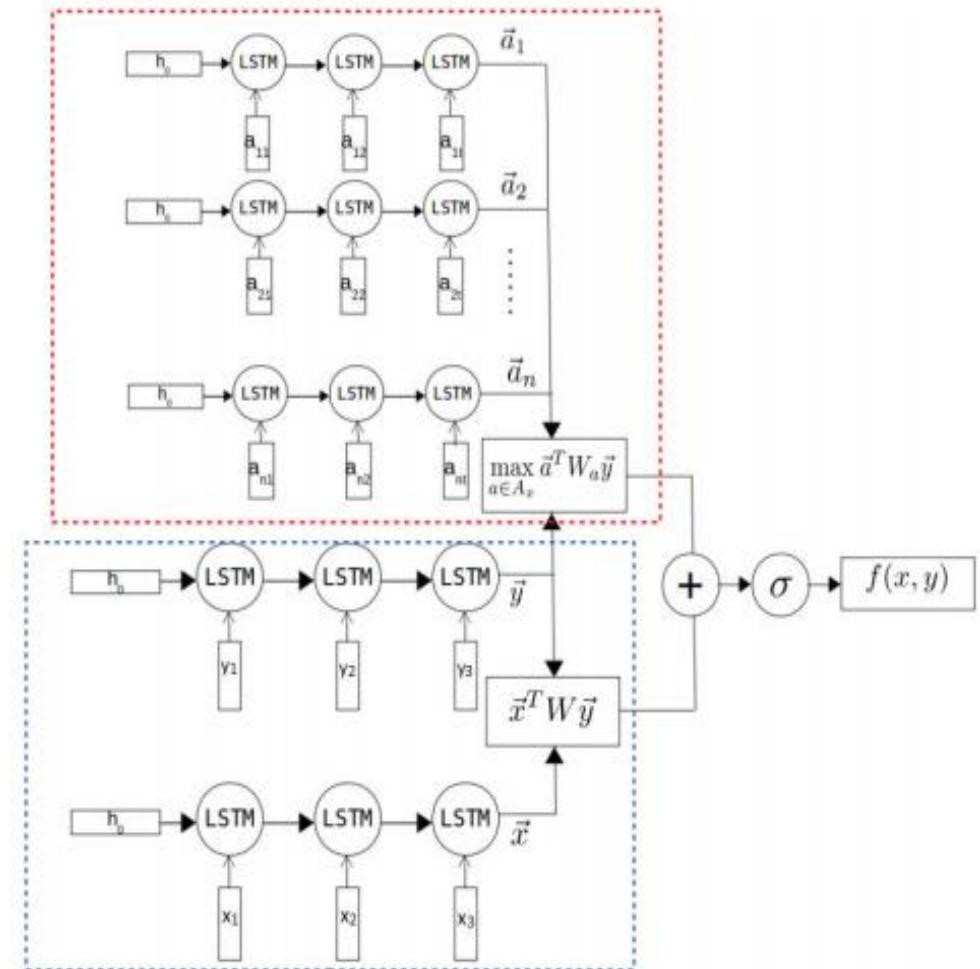
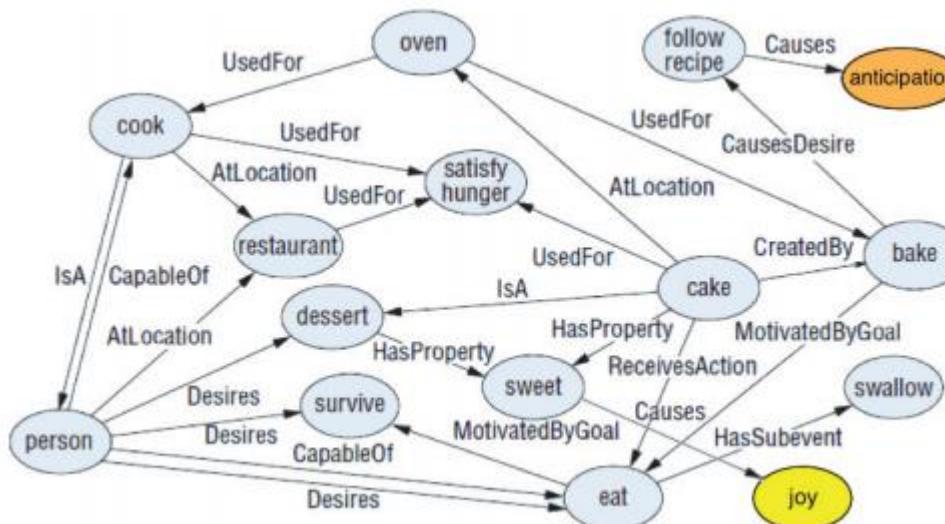


- Grounded response generation [Ghazvininejad et al., AAAI'18]
- Humans have knowledge about conversations
 - Beyond utterances
- Fact encoders
 - Using memory networks to encode facts
 - Based on the conversation history with relevant facts

Commonsense-based Conversation

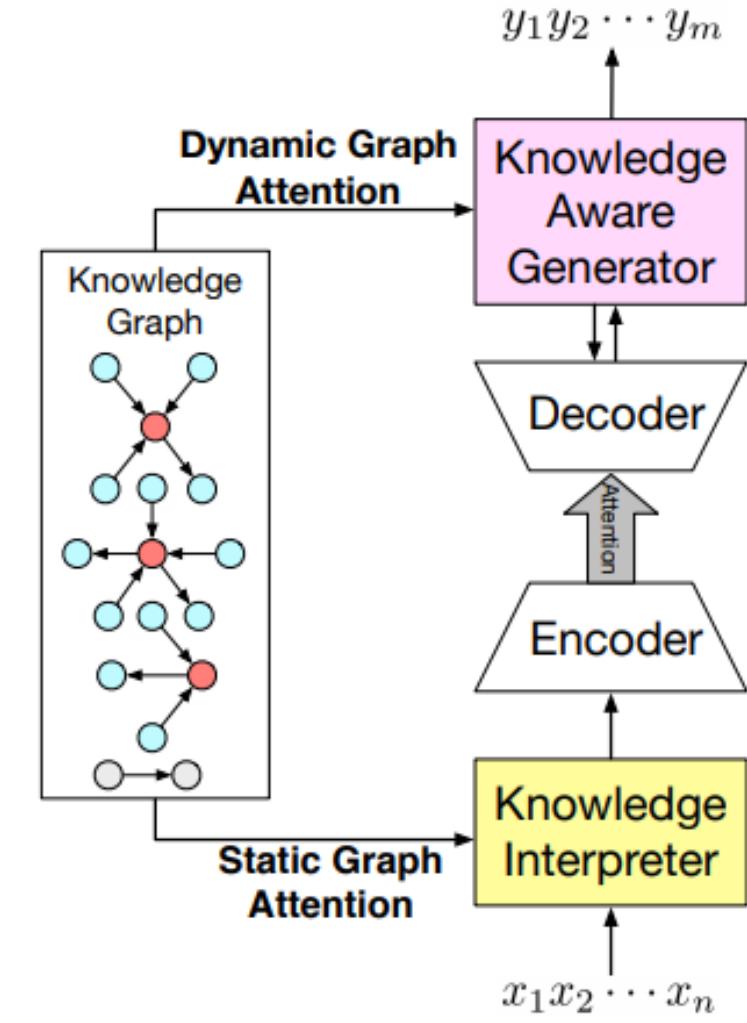
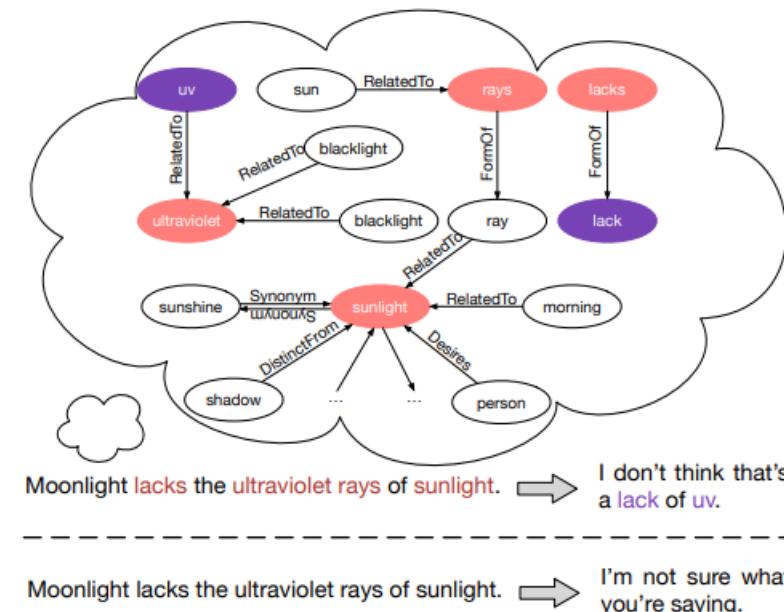
- Commonsense is a knowledge base, too
[Young et al., AAAI'18]

- Commonsense retrieval with all concepts in the utterance
- Tri-LSTM encoder with a memory network of all commonsense knowledge



Knowledge Graph based Conversations

- Knowledge is helpful for human conversations [Zhou et al., IJCAI'18]
- Model components:
 - Encoder-decoder
 - Knowledge interpreter
 - Knowledge generator



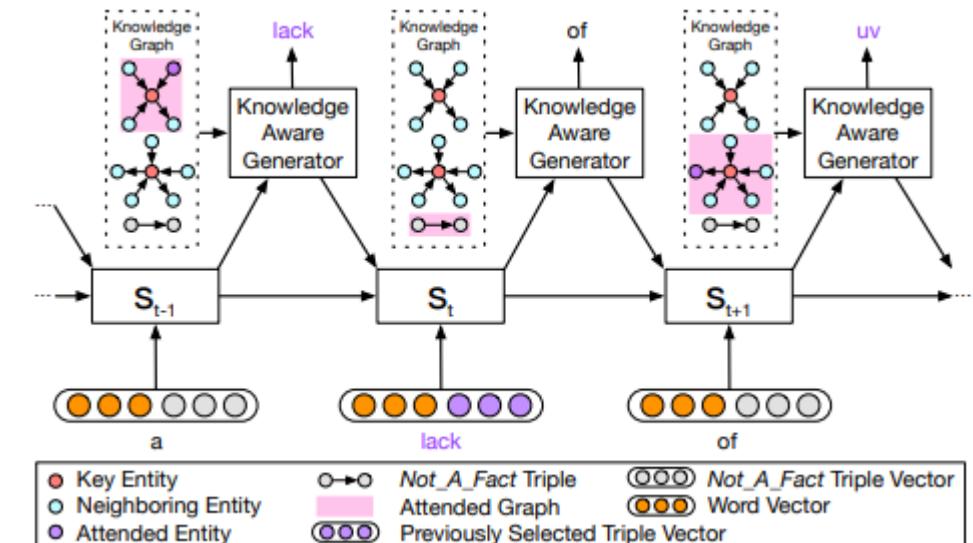
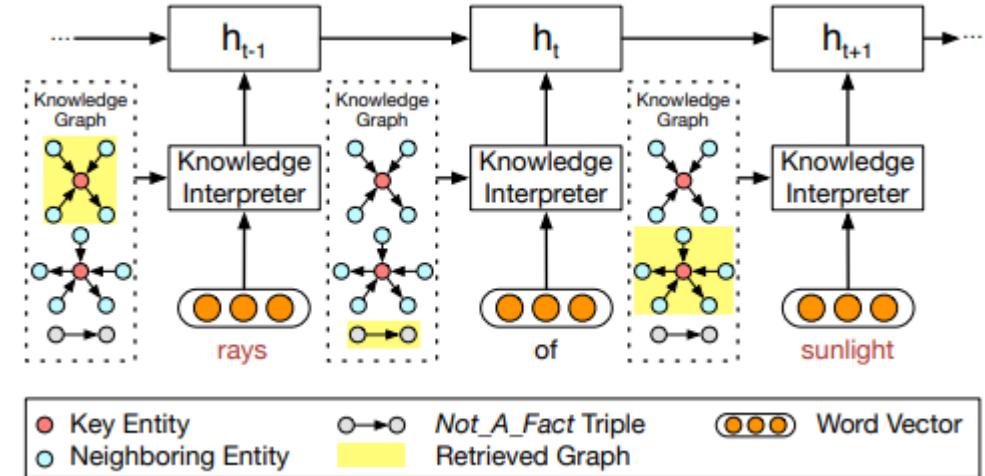
Knowledge Interpreter and Generator (cont.)

- **Interpreter**

- To understand the utterance
- Each word as a key node from the graph vector
- Nodes and neighboring nodes
- Static graph attention

- **Generator**

- Read the graph to triples, and then finally to the words
- Dynamic graph attention

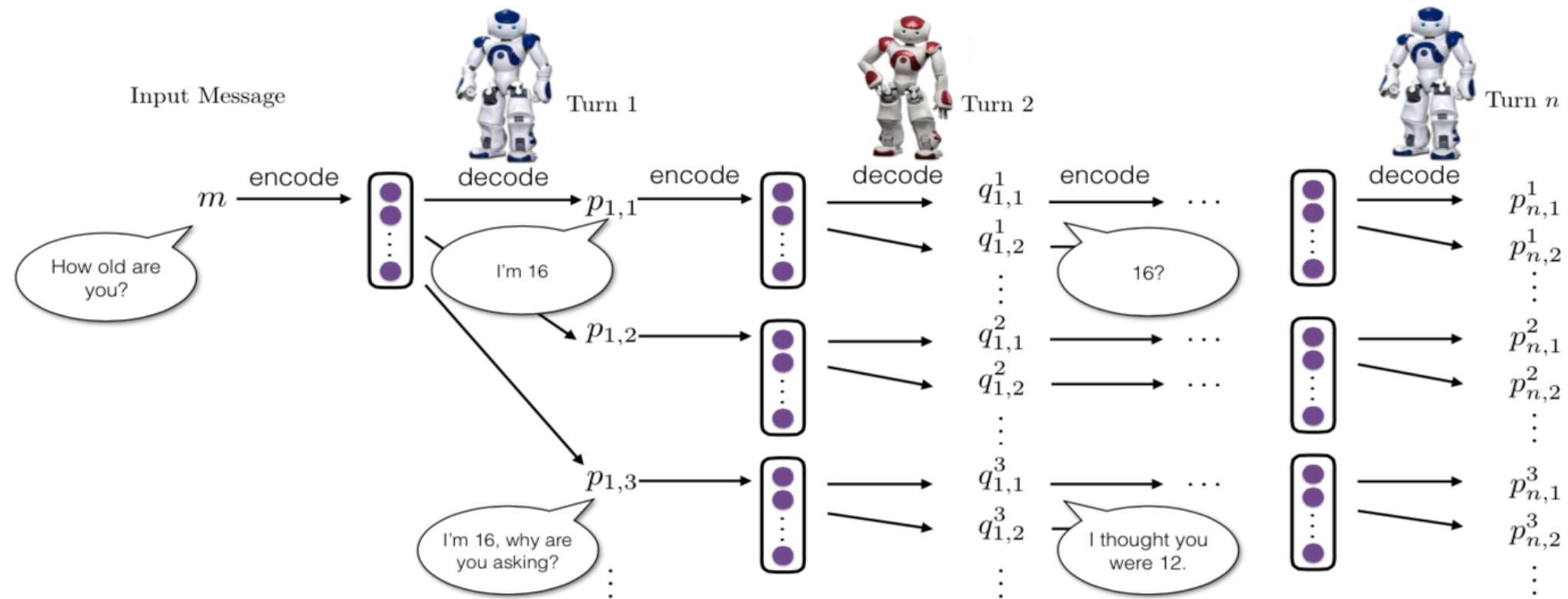




Reinforced Learning and Adversarial Learning in Conversations

Reinforcement Learning in Conversations

- Two virtual agents [Li et al., EMNLP'16]
- Model optimization to maximize rewards



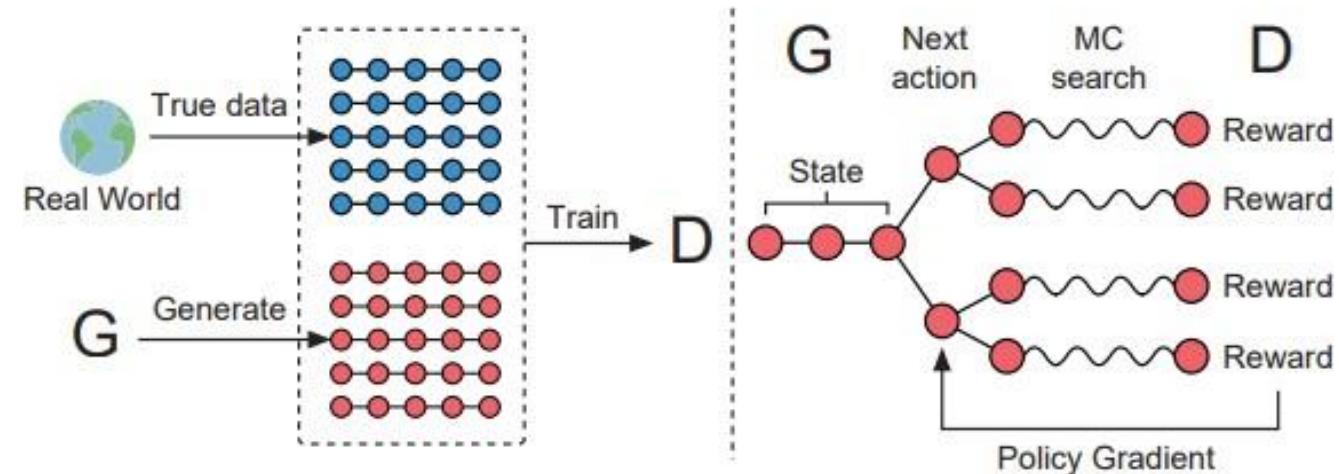
Human-in-the-loop

- Using user implicit feedback [Zhang et al., AAAI'18]
 - Stance reward
 - Sentiment reward
 - Stalemate reward

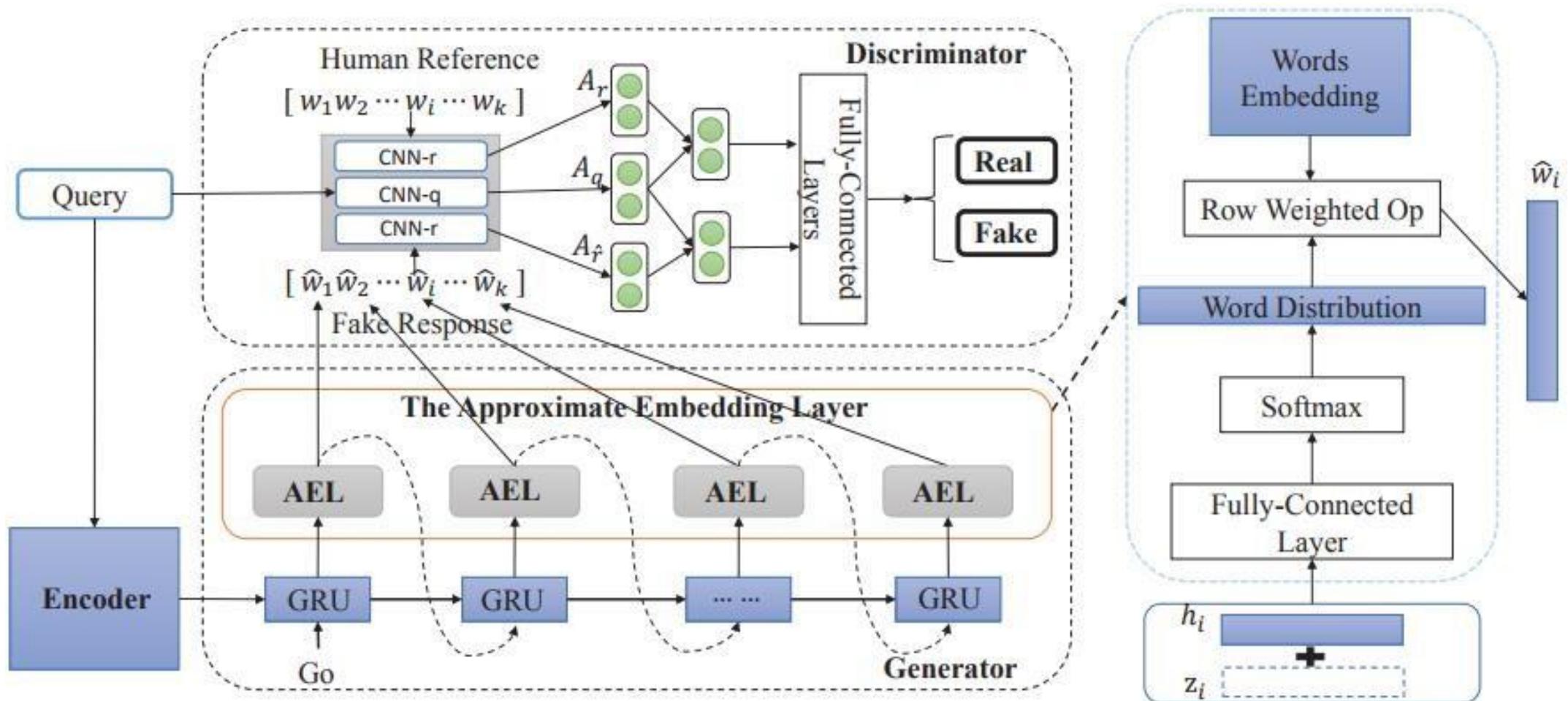
Baseline RL model (Li et al. 2016c)		Proposed implicit feedback model
A: What are you like?	[1]	A: What are you like? [1]
B: I'm so confused.	[2]	B: One of your favs. [2]
A: You know it.	[3]	A: That's so cute. [3]
B: Thank you !!!	[4]	B: Thank you love. [4]
A: You're welcome.	[5]	A: You're welcome xoxo. [5]
B: Lmao I hate this.	[6]	B: You're the best! [6]
A: What do you mean?	[7]	A: I'll be good to you. [7]
B: Nah I hate it.	[8]	B: I like you too. [8]
(Repeat)		A: lol I know. [9]
...		B: This is true. [10]
...		A: That's what I'm saying. [11]
...		(Repeat)

GAN in Conversations

- GAN
- GAN in Sequences
 - SeqGAN [Yu et al., AAAI'17]
- Seq2SeqGAN [Li et al., EMNLP'17]
 - Policy: reward to every generation step
 - Optimize in an adversarial way



GAN in Conversations



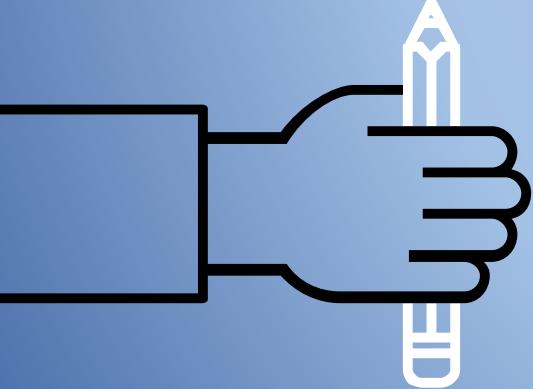
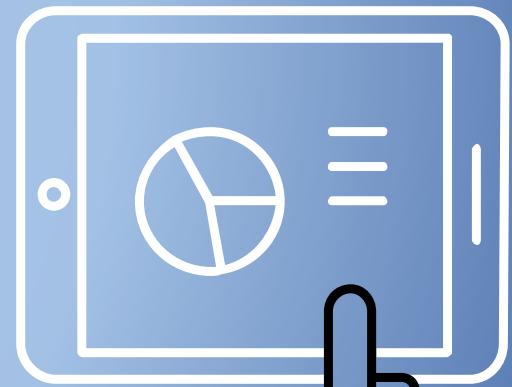
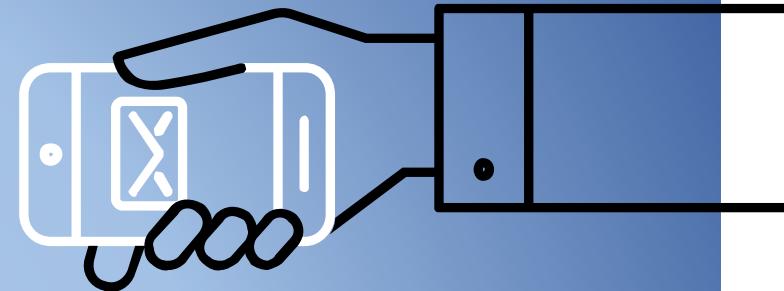
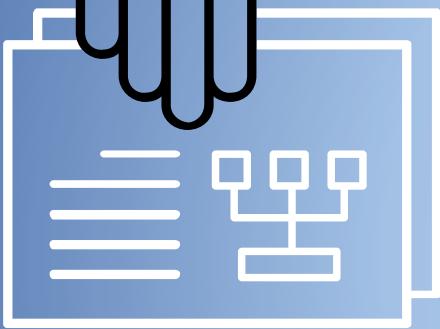
References

- Hongshen Chen, Zhaochun Ren, Jiliang Tang, Yihong Eric Zhao, Dawei Yin. Hierarchical Variational Memory Network for Dialogue Generation. In WWW'18.
- Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In ICLR'15
- Jiwei Li, Thang Luong, and Dan Jurafsky. A hierarchical neural autoencoder for paragraphs and documents. In ACL'15
- Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. A diversity-promoting objective function for neural conversation models. In NAACL'16
- Jiwei Li, Will Monroe, and Dan Jurafsky. A simple, fast diverse decoding algorithm for neural generation. arXiv preprint arXiv:1611.08562, 2016.
- Lili Mou, Yiping Song, Rui Yan, Ge Li, Lu Zhang, and Zhi Jin. Sequence to backward and forward sequences: A content-introducing approach to generative short-text conversation. In COLING'16
- Minghui Qiu, Feng-Lin Li, Siyu Wang, Xing Gao, Yan Chen, Weipeng Zhao, Haiqing Chen, Jun Huang, and Wei Chu. Alime chat: A sequence to sequence and rerank based chatbot engine. In ACL'17
- Iulian V Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. Building end-to-end dialogue systems using generative hierarchical neural network models. In AAAI'16
- Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron C Courville, and Yoshua Bengio. A hierarchical latent variable encoder-decoder model for generating dialogues. In AAAI'17
- Lifeng Shang, Zhengdong Lu, and Hang Li. Neural responding machine for short-text conversation. In ACL-IJCNLP'15
- Mingyue Shang, Zhenxin Fu, Nanyun Peng, Yansong Feng, Dongyan Zhao, and Rui Yan. Learning to converse with noisy data: Generation with calibration. In IJCAI'18

References

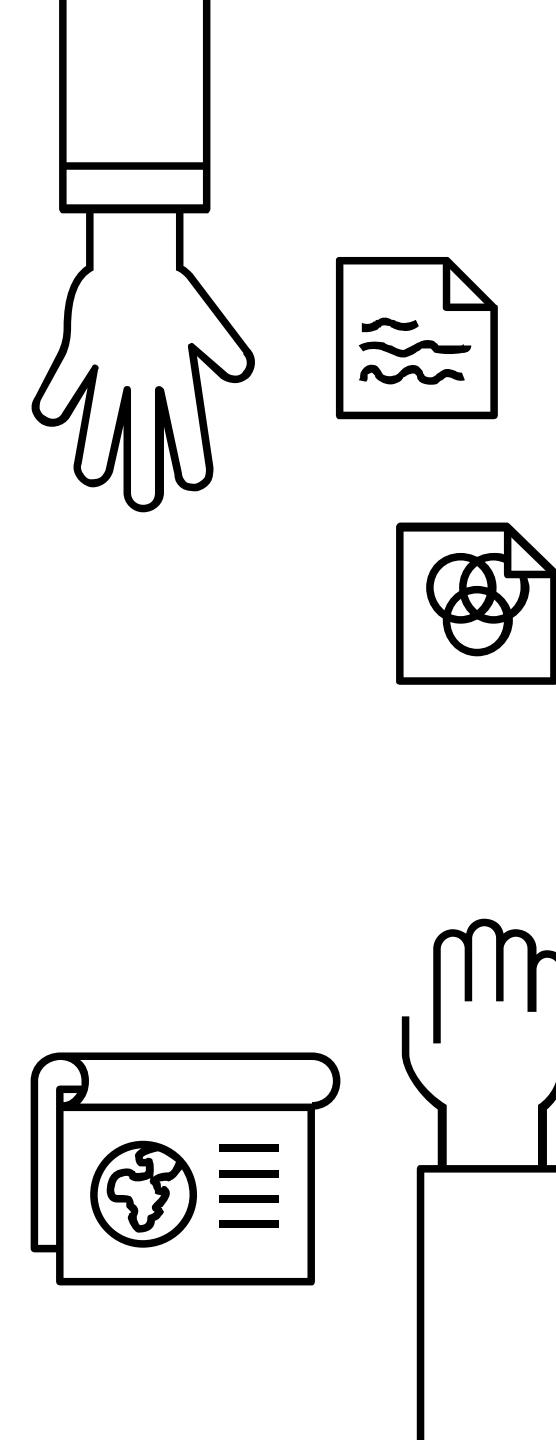
- Yiping Song, Zhiliang Tian, Dongyan Zhao, Ming Zhang, and Rui Yan. Diversifying neural conversation model with maximal marginal relevance. In IJCNLP'17
- Yiping Song, Cheng-Te Li, Jian-Yun Nie, Ming Zhang, Dongyan Zhao, and Rui Yan. An ensemble of retrieval-based and generation-based human-computer conversation systems. In IJCAI'18
- Yiping Song, Rui Yan, Yansong Feng, Yaoyuan Zhang, Dongyan Zhao, and Ming Zhang. Towards a neural conversation model with diversity net using determinantal point processes. In AAAI'18
- Alessandro Sordoni, Michel Galley, Michael Auli, et al. A neural network approach to context-sensitive generation of conversational responses. In NAACL'15
- Ilya Sutskever, Oriol Vinyals, and Quoc VV Le. Sequence to sequence learning with neural networks. In NIPS'14
- Chongyang Tao, Shen Gao, Mingyue Shang, Wei Wu, Dongyan Zhao, and Rui Yan. Get the point of my utterance! learning towards effective responses with multi-head attention mechanism. In IJCAI'18
- Zhiliang Tian, Rui Yan, Lili Mou, Yiping Song, Yansong Feng, and Dongyan Zhao. How to make contexts more useful? an empirical study to context-aware neural conversation models. In ACL'17
- Lili Yao, Yaoyuan Zhang, Yansong Feng, Dongyan Zhao, and Rui Yan. Towards implicit contentintroducing for generative short-text conversation systems. In EMNLP'17
- Tiancheng Zhao, Ran Zhao, and Maxine Eskenazi. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. In ACL'17
- Ganbin Zhou, Ping Luo, Rongyu Cao, Fen Lin, Bo Chen, and Qing He. Mechanism-aware neural machine for dialogue response generation. In AAAI'17
- Chen Xing, Wei Wu, Yu Wu, Ming Zhou, Yalou Huang, and Wei-Ying Ma. Hierarchical Recurrent Attention Network for Response Generation. In AAAI'18

Evaluation



Automatic Evaluation Metric

- ▶ Evaluation metrics are crucial to evaluate how we are doing
- ▶ Language model
 - Perplexity
- ▶ Machine translation
 - BLEU, NIST, METEOR
 - Shared tasks for evaluation in WMT
- ▶ Text summarization
 - ROUGE, Pyramid
- ▶ Conversational systems: ???

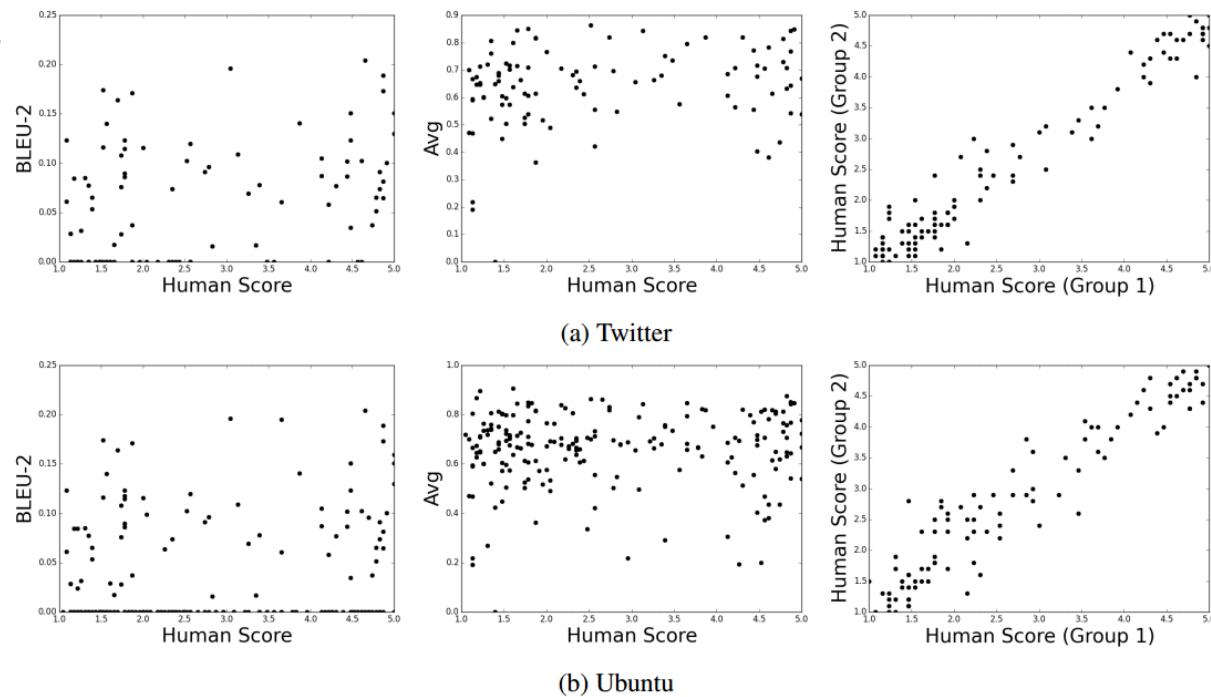


Evaluation Metrics for Conversations

- Human evaluation
 - Applicable to almost every NLP task
 - Point-wise evaluation
 - Pair-wise evaluation
- Automatic evaluation metrics
 - BLEU [Ritter et al., 2011; Li et al., 2015; Sordoni et al., 2015; Song et al., 2016]
 - Information: entropy, perplexity [Serban et al. 2016 and Mou et al. 2016]
 - Diversity: distinct-1, distinct-2 [Li et al., 2015]
 - Average response length [Serban et al. 2016, Mou et al. 2016]

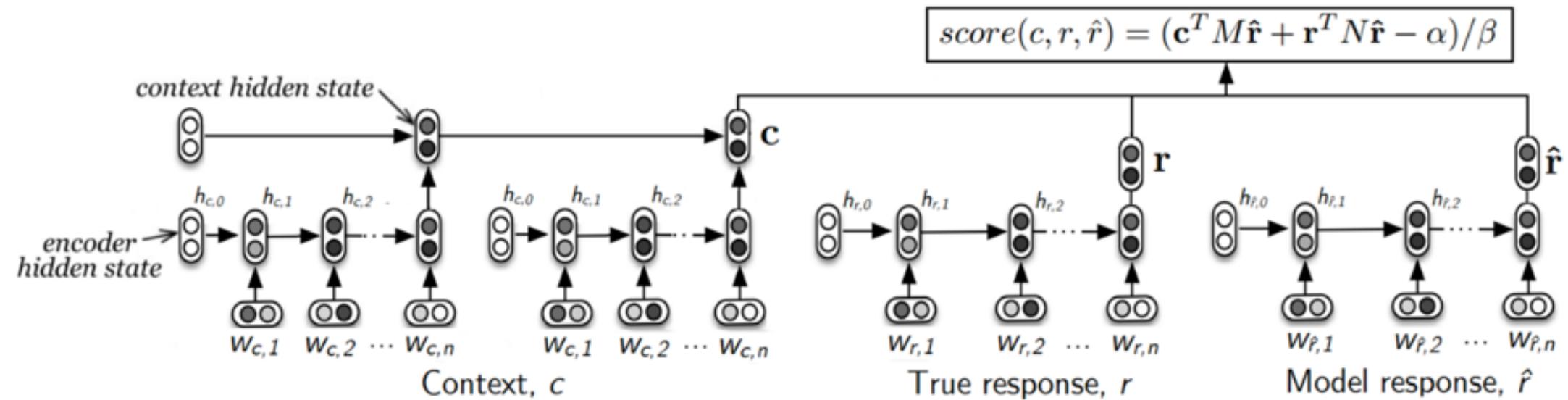
How **NOT** to Evaluate Conversations

- Weak correlation between human evaluation and automatic evaluation metrics [Liu et al., EMNLP'16]
 - Significant diversity in the space of valid replied to a given input.
 - Utterances are typically short and casual in open-domain dialog systems.



ADEM Model

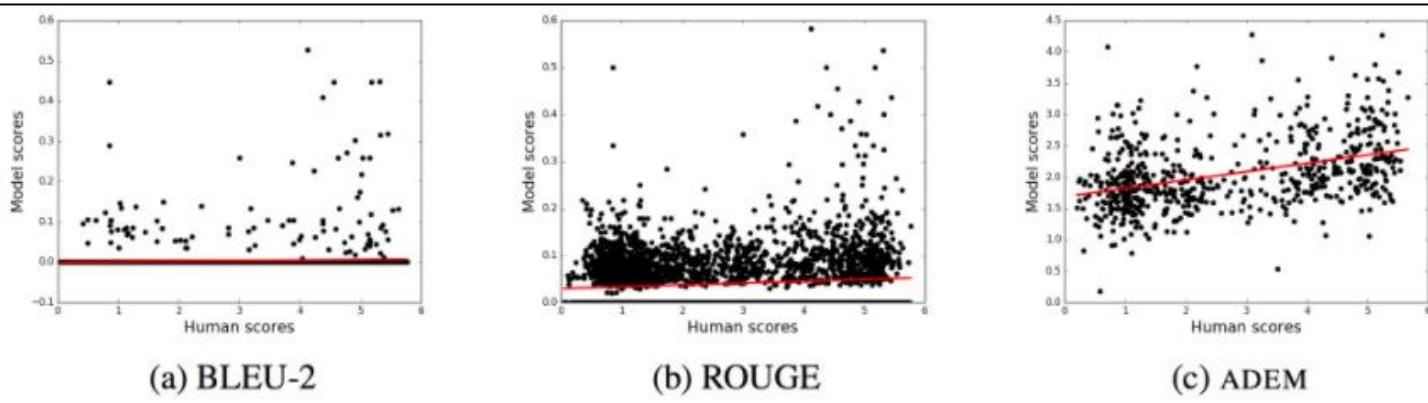
- ADEM: learnable metric [Lowe et al., ACL'17]
 - Human labeling first
 - Predict a score of a reply given its query (context) and a ground truth reply
 - It requires massive human-annotated scores to train the network



Higher Correlation

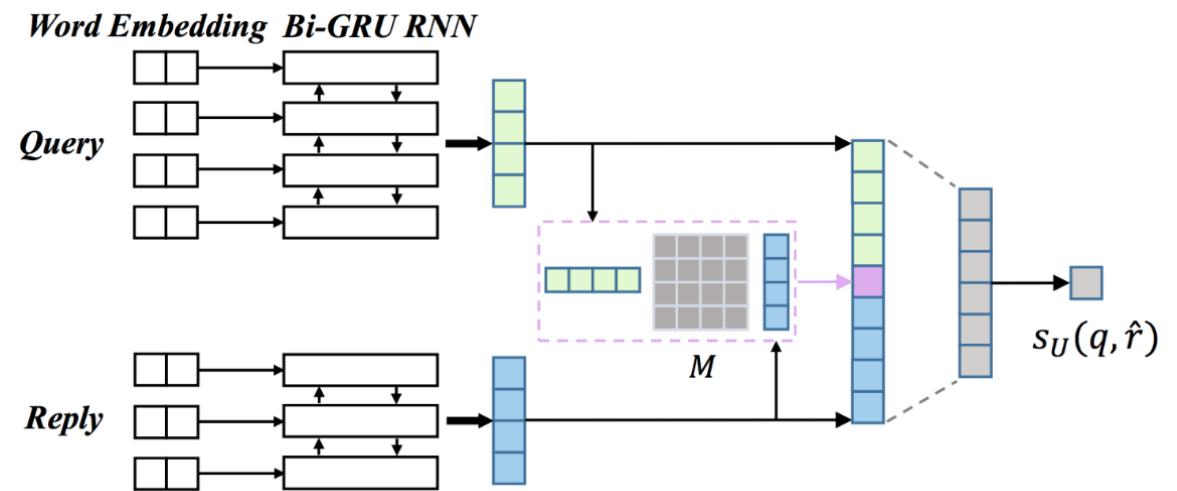
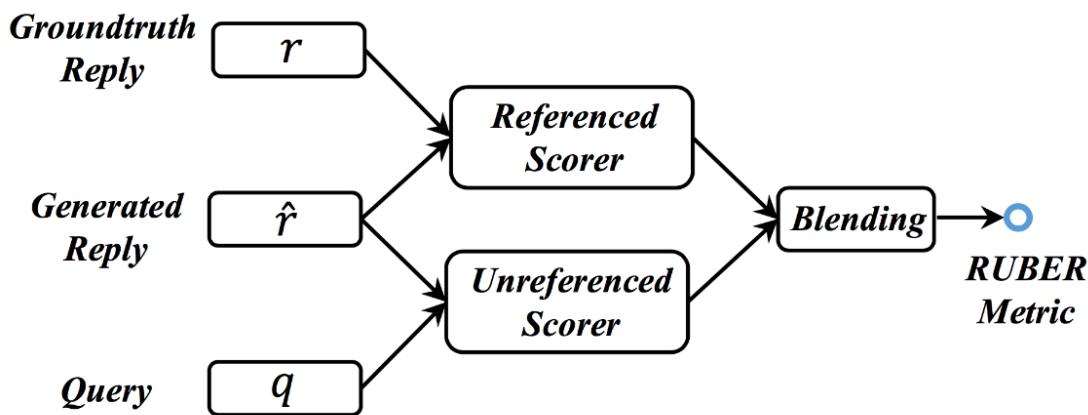
Better performance
than **BLEU** and **ROUGE**

Context	Reference response	Model responses	Human score	ADEM score
photo to see my television debut go to - some. some on <url> - hehe <url> → it really was you? i thought ppl were recognizing someone who looked like you! were the oysters worth the wait?	yeah it was me . haha i 'd kinda forgotten about it it was filmed a while ago	1) i'm not sure. i just don't know what to do with it. 2) you heard the horsepower productions remix of lee scratch perry's 'exercising' off his 'mighty upsetter' album? 3) you wont chug a fuzzy peach navel 4) they were!	3 1 1 5	1.602 1.513 1.744 3.274
just beat call of duty!! → want a cookie? → yes!! → come get it	im in kenmore at the moment	1) i'm gonna get a new phone some moro 2) no way man. 3) wow i just got a free pizza coupon! get yours before theres no more! <url> 4) i'm going to go to the mall.	1 5 1 1	1.848 4.265 0.921 2.634
am i out of twitter jail yet? testing → yeah. i posted bail → thanks. i am a right chatter tweetbox on sundays. same happened last sunday lol	any news on meeting our user ? i go to the us on friday and i don 't want to miss anything arranged	1) i'm not sure if i'm going to be able to get it. 2) good to see another mac user in the leadership ranks 3) awww poor baby hope u get to feeling better soon. maybe some many work days at piedmont 4) did you tweet too much?	3 4 2 5	1.912 1.417 1.123 2.539



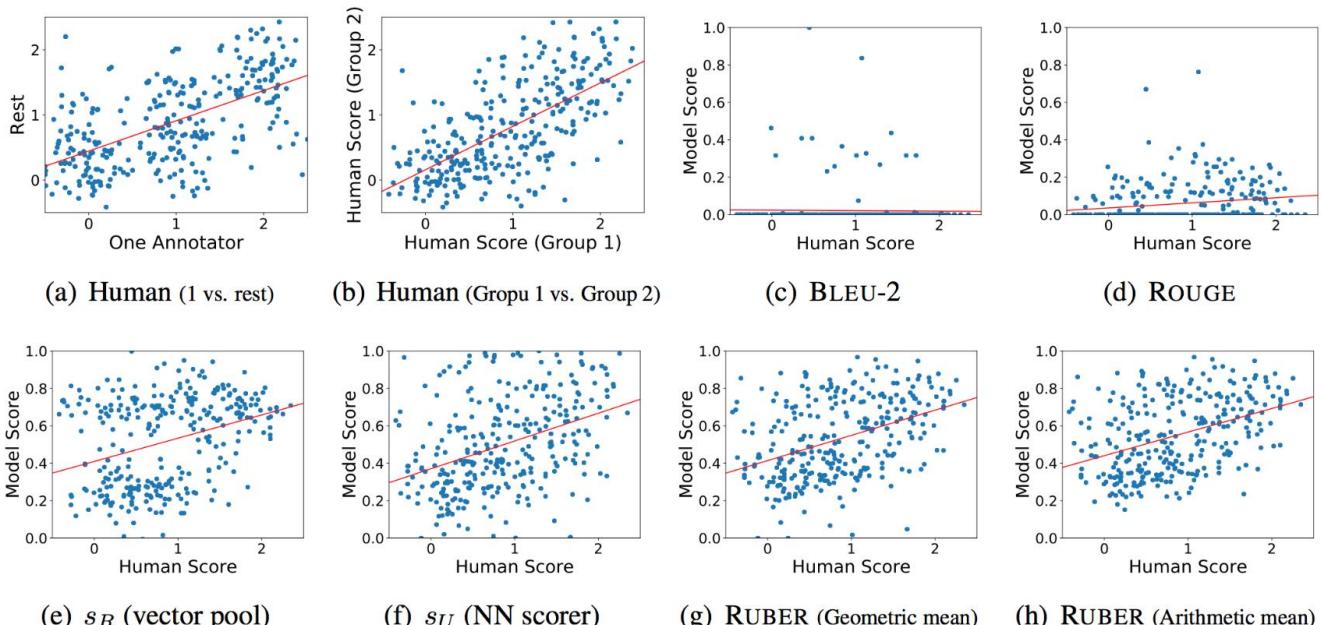
RUBER

RUBER: a Referenced metric and Unreferenced metric Blended Evaluation Routine for open-domain dialog systems [Tao et al., AAAI'18]



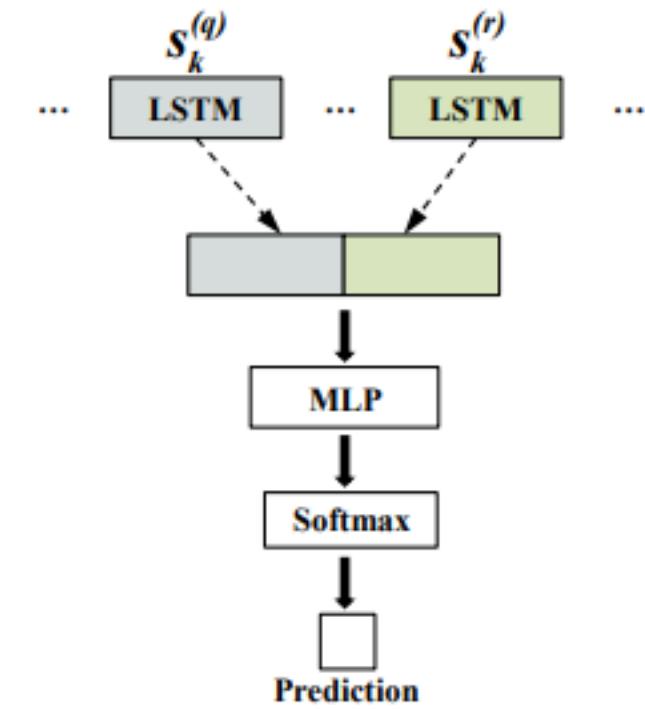
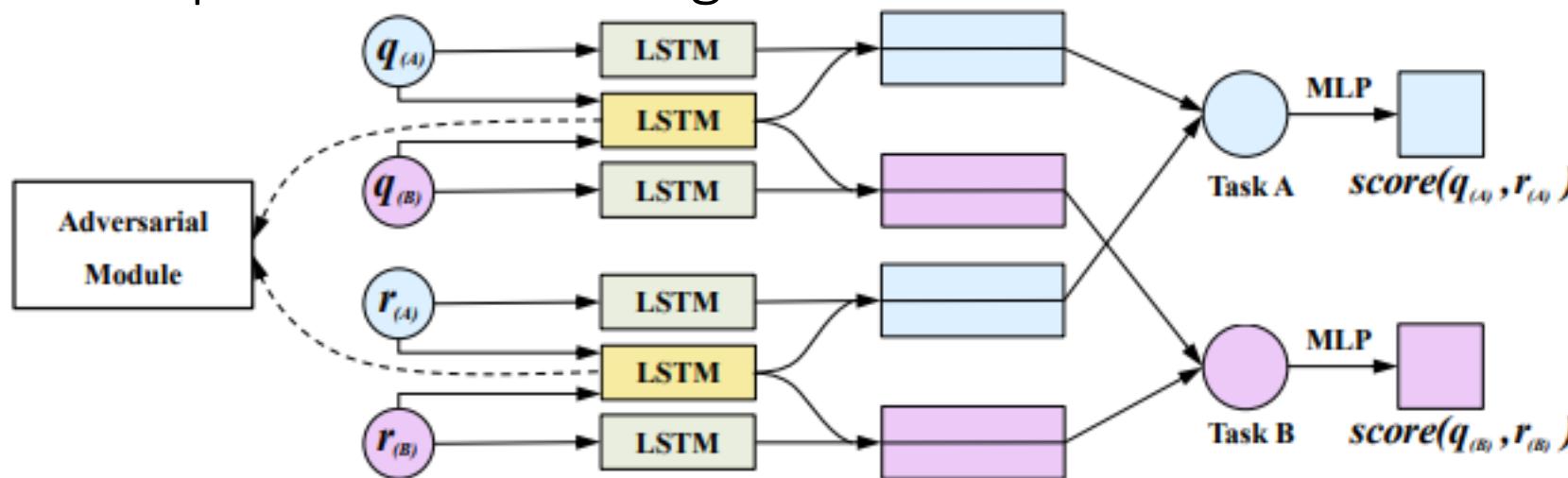
RUBER Results

Metrics		Retrieval (Top-1)		Seq2Seq (w/ attention)	
		Pearson(<i>p</i> -value)	Spearman(<i>p</i> -value)	Pearson(<i>p</i> -value)	Spearman(<i>p</i> -value)
Inter-annotator	Human (Avg)	0.4927(<0.01)	0.4981(<0.01)	0.4692(<0.01)	0.4708(<0.01)
	Human (Max)	0.5931(<0.01)	0.5926(<0.01)	0.6068(<0.01)	0.6028(<0.01)
Referenced	BLEU-1	0.2722(<0.01)	0.2473(<0.01)	0.1521(<0.01)	0.2358(<0.01)
	BLEU-2	0.2243(<0.01)	0.2389(<0.01)	-0.0006(0.9914)	0.0546(0.3464)
	BLEU-3	0.2018(<0.01)	0.2247(<0.01)	-0.0576(0.3205)	-0.0188(0.7454)
	BLEU-4	0.1601(<0.01)	0.1719(<0.01)	-0.0604(0.2971)	-0.0539(0.3522)
	ROUGE	0.2840(<0.01)	0.2696(<0.01)	0.1747(<0.01)	0.2522(<0.01)
	Vector pool (s_R)	0.2844(<0.01)	0.3205(<0.01)	0.3434(<0.01)	0.3219(<0.01)
Unreferenced	Vector pool	0.2253(<0.01)	0.2790(<0.01)	0.3808(<0.01)	0.3584(<0.01)
	NN scorer (s_U)	0.4278(<0.01)	0.4338(<0.01)	0.4137(<0.01)	0.4240(<0.01)
RUBER	Min	0.4428(<0.01)	0.4490(<0.01)	0.4527 (<0.01)	0.4523 (<0.01)
	Geometric mean	0.4559(<0.01)	0.4771(<0.01)	0.4523(<0.01)	0.4490(<0.01)
	Arithmetic mean	0.4594 (<0.01)	0.4906 (<0.01)	0.4509(<0.01)	0.4458(<0.01)
	Max	0.3263(<0.01)	0.3551(<0.01)	0.3868(<0.01)	0.3623(<0.01)



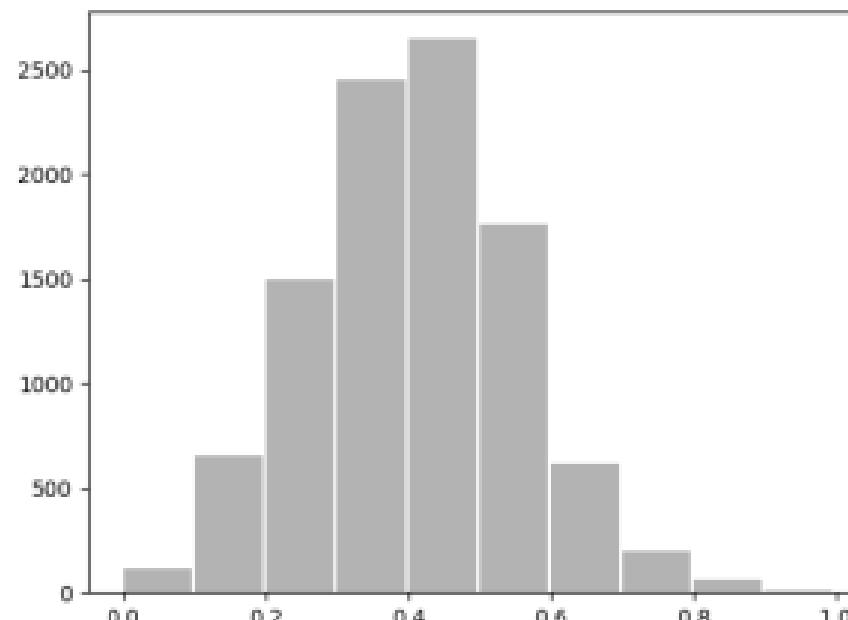
RUBER Plus

- Multi-linguality in conversations [Tong et al., IJCAI'18]
 - A natural solution with multi-task learning
- Responding patterns are language-insensitive
- Adversarial learning to understand the language-insensitive parts
 - Share-private embeddings



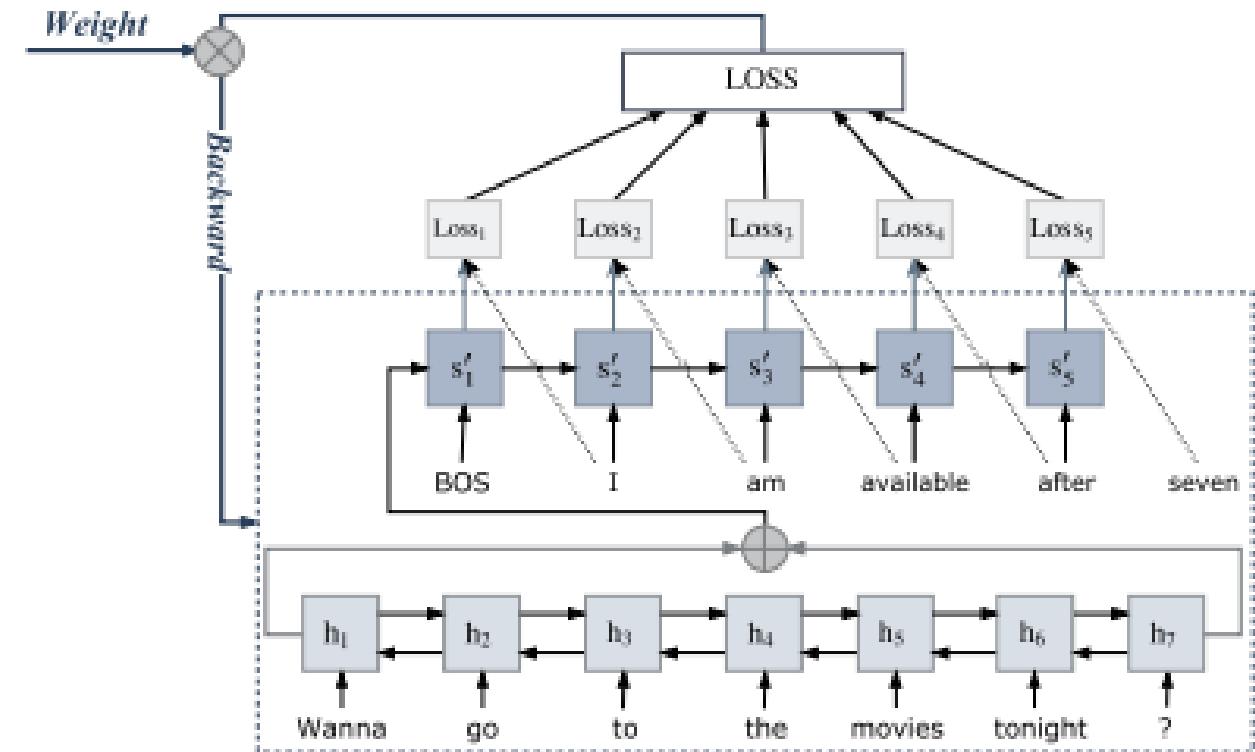
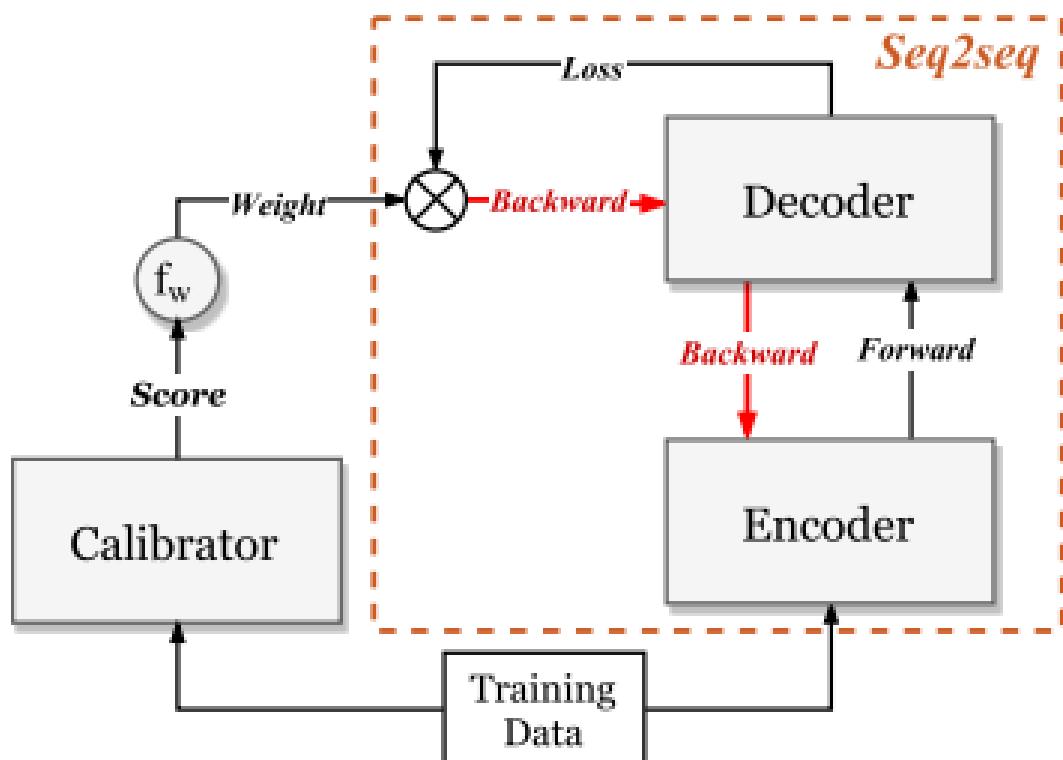
Evaluation-based Generation

- Learning to converse with noisy data [Shang et al., IJCAI'18]
 - Not all training samples are equally important
 - Conversational data are noisy
 - Generation with calibration: using RUBER as the evaluation metric



Evaluation-based Generation (cont.)

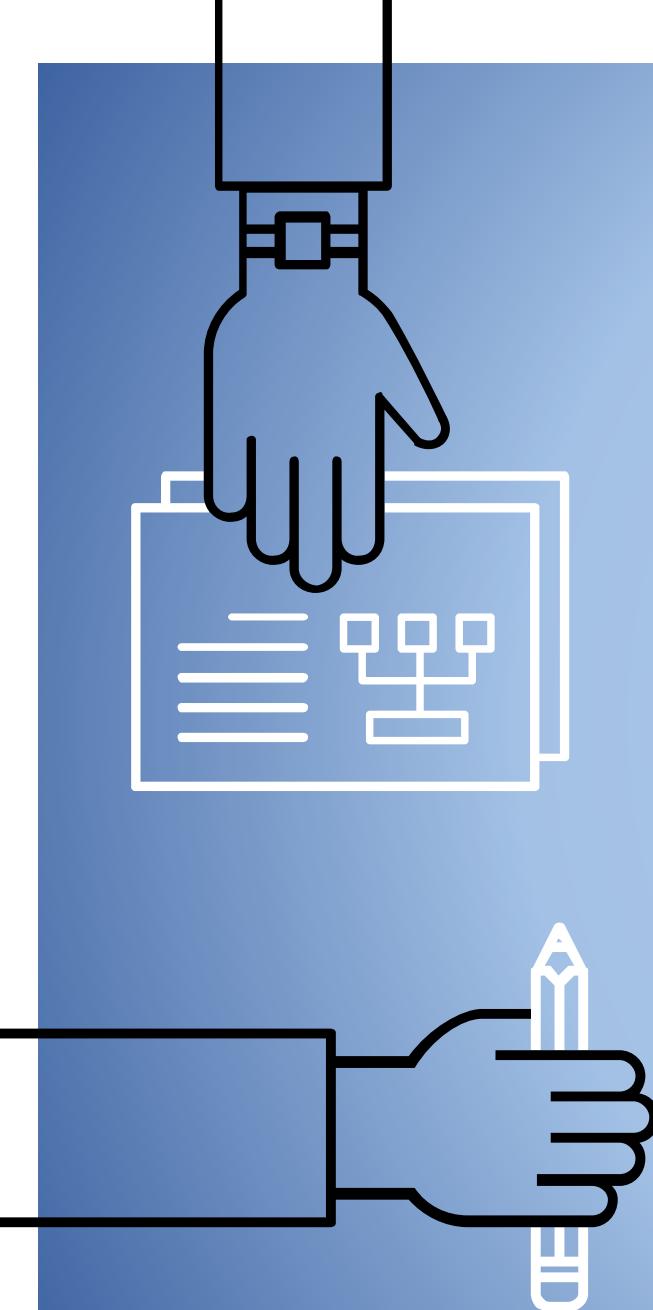
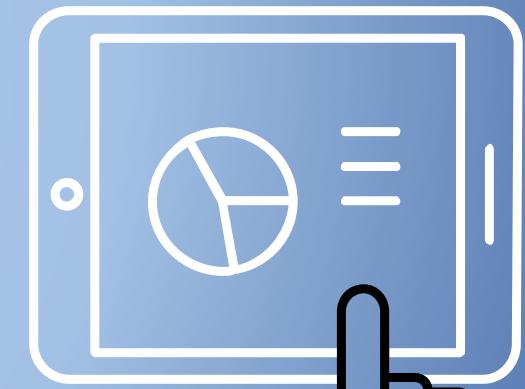
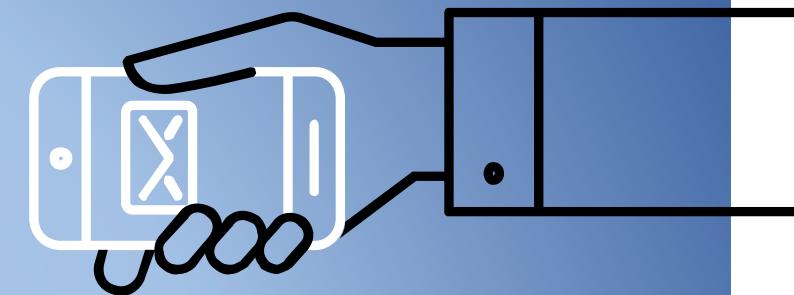
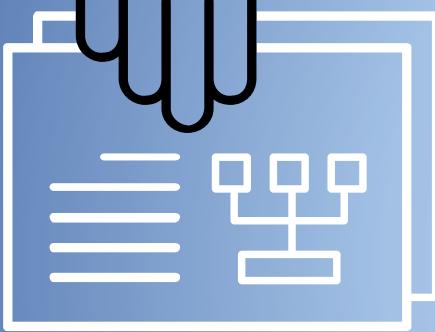
- Learning to fit good training samples
 - Instance weighting
- The weights are propagated through loss



References

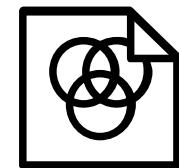
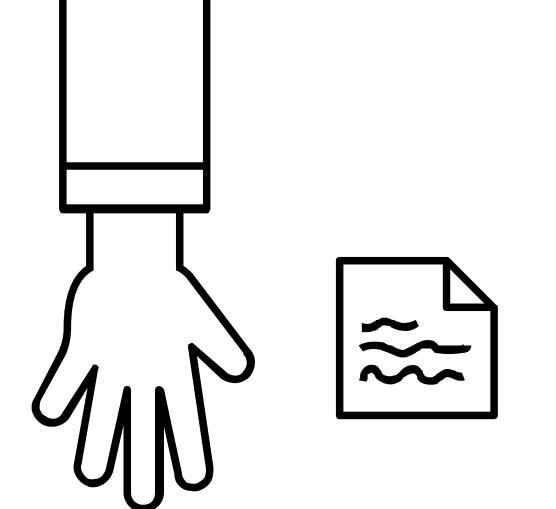
- Chia-Wei Liu, Ryan Lowe, et al. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In EMNLP'16
- Ryan Lowe, Michael Noseworthy, Iulian Vlad Serban, Nicolas Angelard-Gontier, Yoshua Bengio, and Joelle Pineau. Towards an automatic turing test: Learning to evaluate dialogue responses. In ACL'17
- Ani Nenkova and Rebecca Passonneau. Evaluating content selection in summarization: The pyramid method. In HLT-NAACL'04
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In ACL'02
- Mingyue Shang, Zhenxin Fu, Nanyun Peng, Yansong Feng, Dongyan Zhao, and Rui Yan. Learning to converse with noisy data: Generation with calibration. In IJCAI'18
- Chongyang Tao, Lili Mou, Dongyan Zhao, and Rui Yan. Ruber: An unsupervised method for automatic evaluation of open-domain dialog systems. In AAAI'18
- Xiaowei Tong, Zhenxin Fu, Mingyue Shang, Dongyan Zhao, and Rui Yan. One “ruler” for all languages: Multi-lingual dialogue evaluation with adversarial multi-task learning. In IJCAI'18

Future Trends

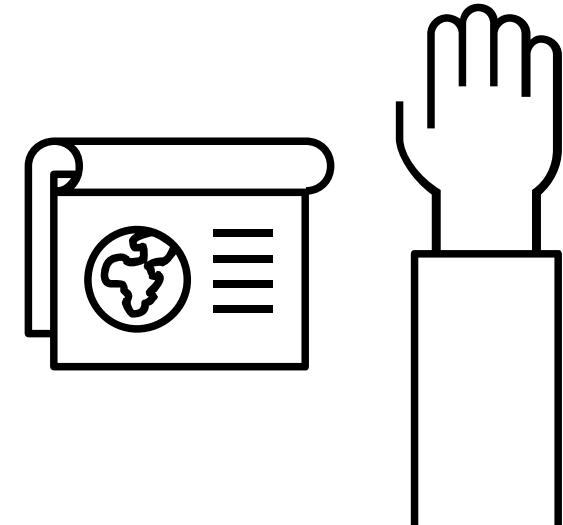


Reminder of Mentioned Trends

- ▶ Learning Methods
- ▶ Representations



146



Reasoning in Dialogues

- ▶ **Context-based Reasoning**
 - “John picked up the apple.”
 - “John walked into the living room.”
 - Where is the apple
- ▶ **Knowledge**
 - Knowledge graph
- ▶ **Commonsense**
 - “I had milk this morning.”

X-Grounded Dialogues

- ▶ **Multi-Modal**
 - Vision: image
 - Video
- ▶ **Labels**
 - Emotion
- ▶ **Texts**
 - Profile
 - Wikipedia
 - Knowledge Graph

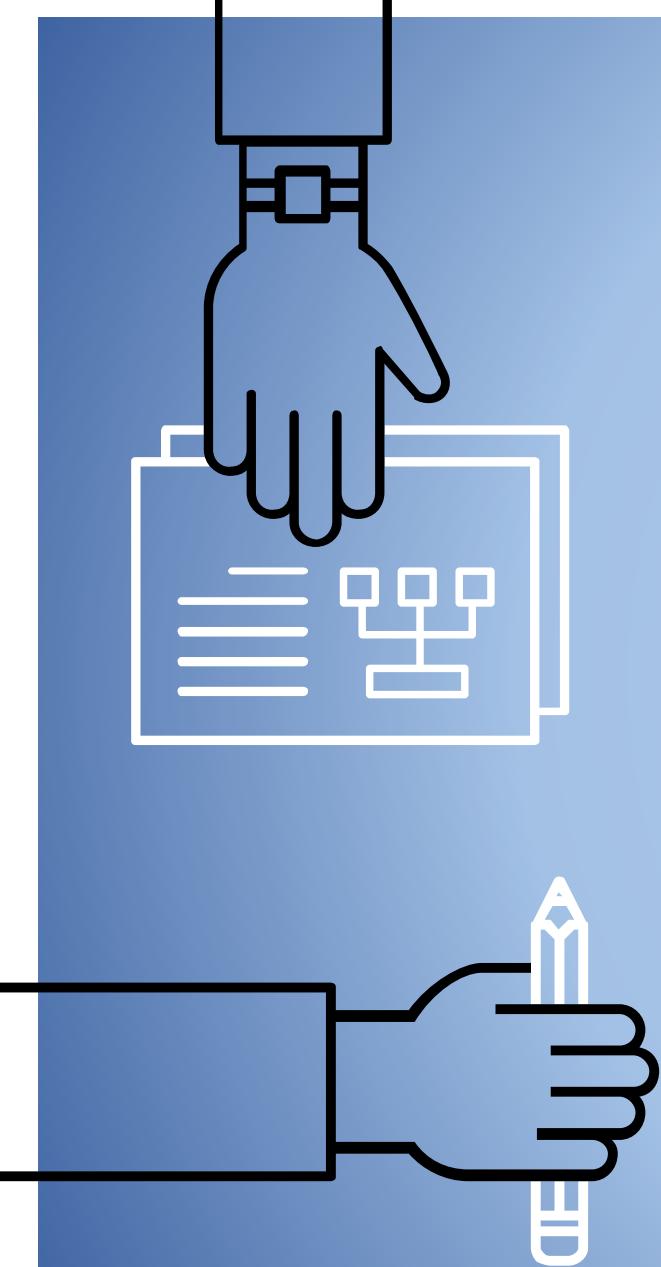
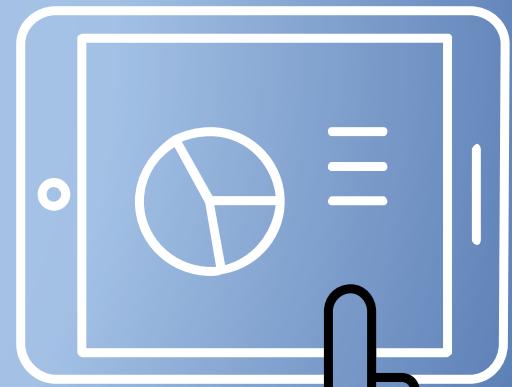
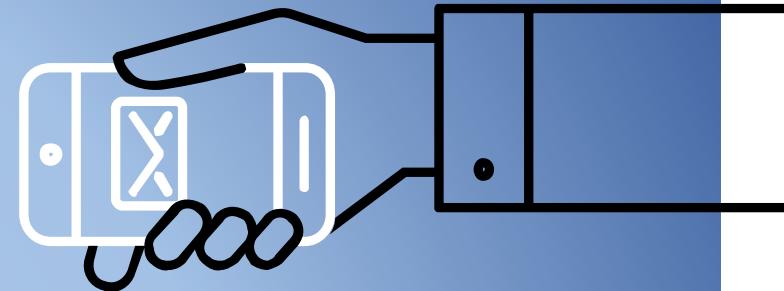
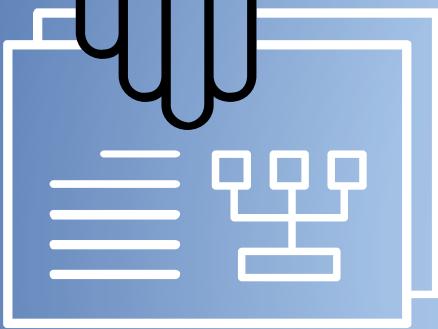
Dialogue Management in Open Domain

- ▶ **Keyword-based Planning**
 - Content-introducing in dialogues
- ▶ **Dialogue act-based management**
- ▶ **Targets**
 - Length
 - Consistency

Evaluation & Benchmark Datasets

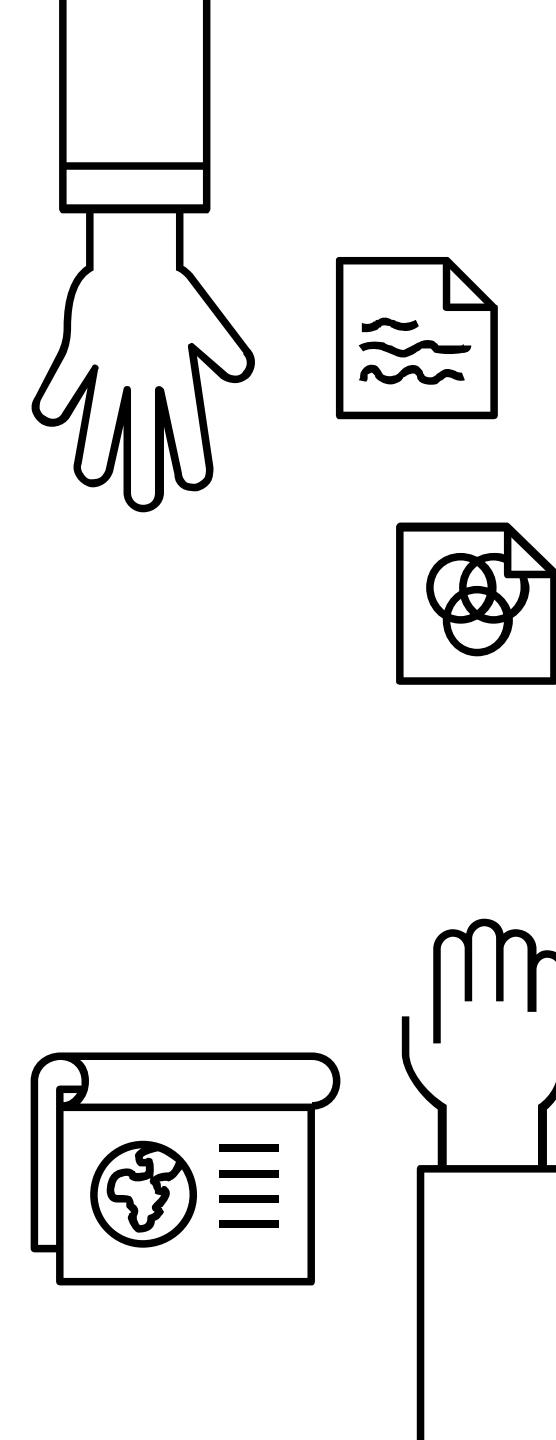
- ▶ **Towards better evaluation metrics**
 - From BLEU, ROUGE, to RUBER and ADAM
 - How do you evaluate informativeness, interestingness
- ▶ **Benchmark datasets for dialogue generation**
 - Can we have 1 set for all methods?

Conclusion



Take-Home Messages

- ▶ Background knowledge about dialogues
- ▶ Retrieval-based methods
- ▶ Generation-based methods
 - Diversity, additional elements
- ▶ Evaluations
 - Learnable metrics
- ▶ Future trends
 - Grounding, Learning, Reasoning





Thank you for listening

Q&A