Interactive Matching Network for Multi-Turn Response Selection in Retrieval-Based Chatbots

Jia-Chen Gu[1], Zhen-Hua Ling[1] and Quan Liu[1,2]

[1]National Engineering Laboratory for Speech and Language Information Processing,
University of Science and Technology of China
[2]State Key Laboratory of Cognitive Intelligence, iFLYTEK Research

Outline

- **Introduction**
- Interactive Matching Network
- Experiments
- Conclusion

Introduction

- Multi-Turn Response Selection

  The task is to select the best-matched response from a set of candidates given the context of a conversation.

	Context	
Utterance 1	*Human*: How are you doing?	
Utterance 2	*ChatBot*: I am going to hold a drum class in Shanghai. Anyone wants to join?	
Utterance 3	*Human*: Interesting! Do you have coaches who can help me practice drum?	
Utterance 4	*ChatBot*: Of course.	
Utterance 5	*Human*: Can I have a free first lesson?	
	Response Candidates	
Response 1	Sure. Have you ever played drum before?	**(Correct)**
Response 2	What lessons do you want?	**(Wrong)**

Task and Notation Definition

Data: (context, response, label) triple as (c, r, y)

Context: $c = \{u_1, u_2, ..., u_n\}$

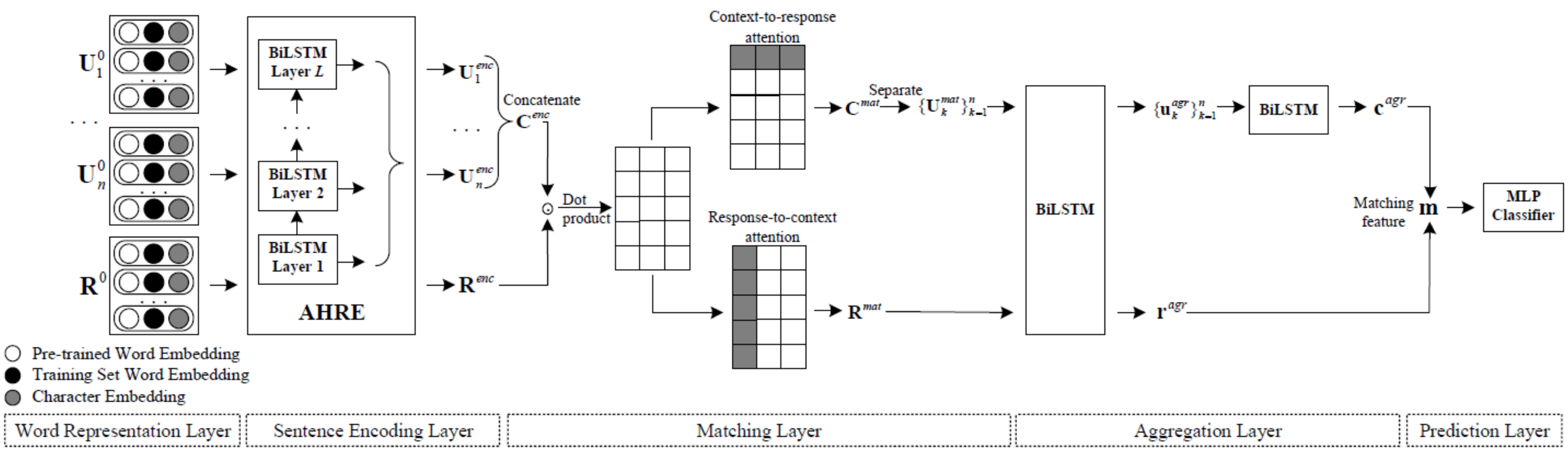
Response: r

Label: $y \in \{0, 1\}$

Outline

- Introduction
- **Interactive Matching Network**
- Experiments
- Conclusion

Interactive Matching Network



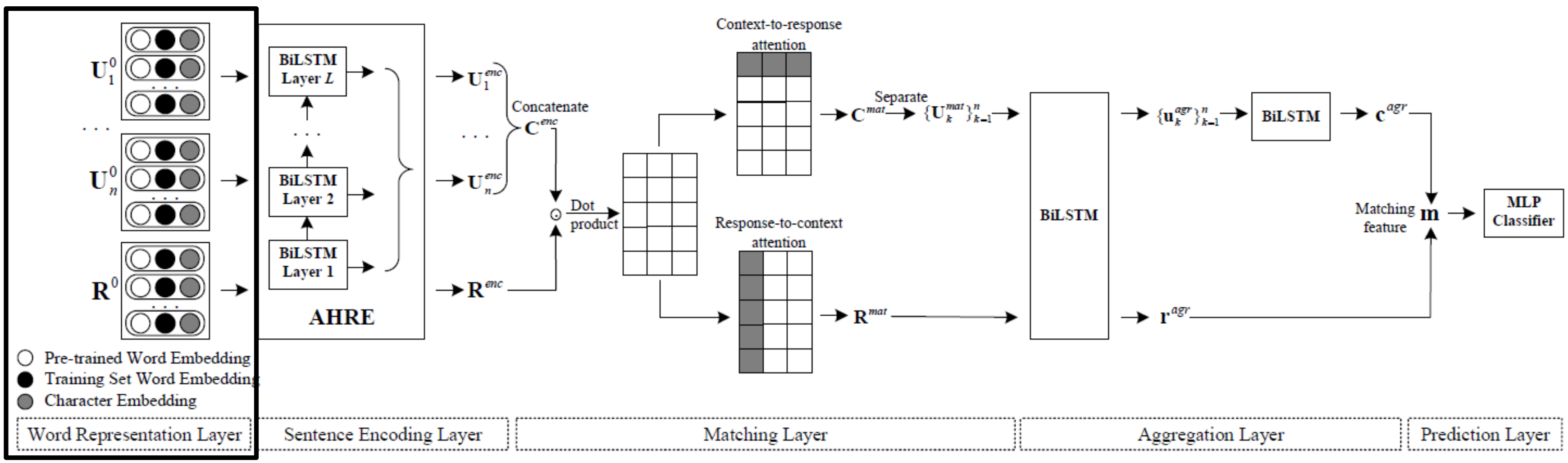
Three issues that IMN is designed to address:

1) The challenge of out-of-vocabulary (OOV) words.

2) Hierarchical information when encoding the sentences.

3) Interactions between the context and response.

Interactive Matching Network



Word Representation Layer:

We construct word representations with a combination of

1) **general** pre-trained word embedding

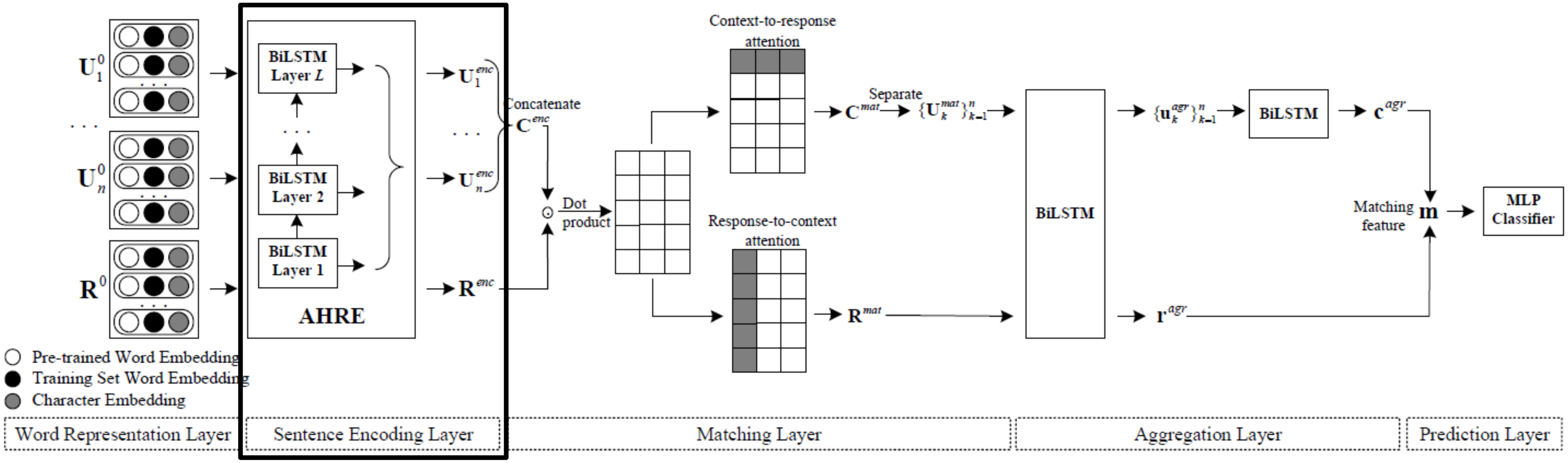
2) embedding estimated on the **task-specific** training set

3) **character-level** embeddings

The *k*-th utterance and the response are denoted as

$\mathbf{U}_k^0 = \{\mathbf{u}_{k,i}^0\}_{i=1}^{l_{u_k}}$ and $\mathbf{R}^0 = \{\mathbf{r}_j^0\}_{j=1}^{l_r}$ respectively.

Interactive Matching Network



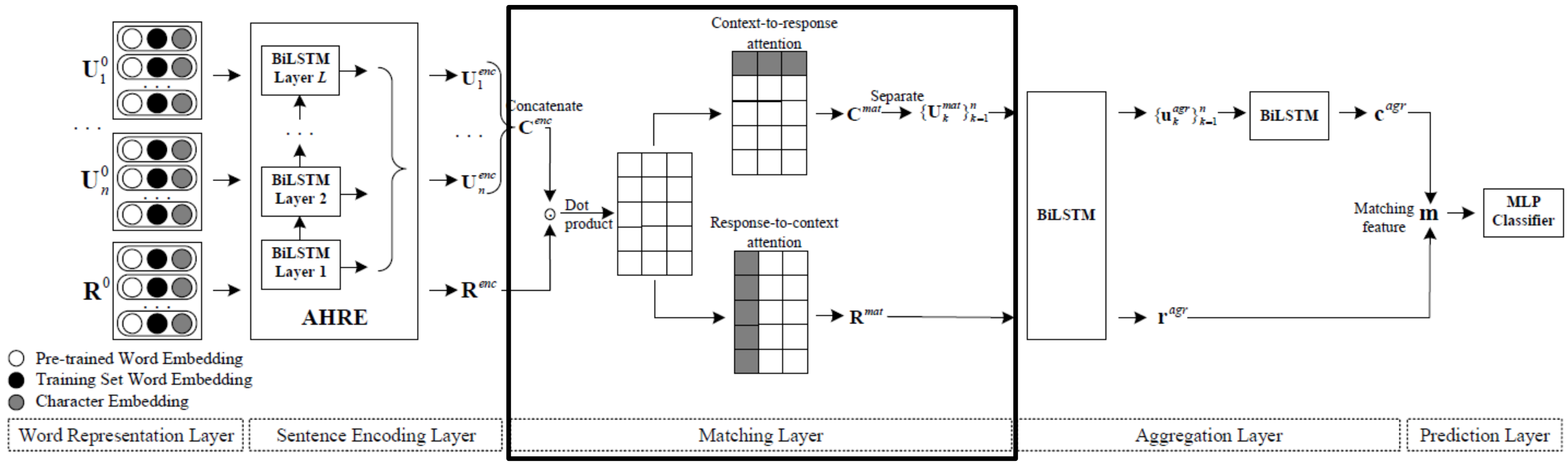
Sentence Encoding Layer:

An **attentive hierarchical recurrent encoder** (AHRE) is designed to aggregate the representations at all hidden layers.

M-layer RNNs output a set of representations $\{\mathbf{U}_k^1, ..., \mathbf{U}_k^M\}$ and $\{\mathbf{R}^1, ..., \mathbf{R}^M\}$, and then aggregate them with attention mechanism.

$$\mathbf{u}_{k,i}^{enc} = \sum_{m=1}^{M} w_m \mathbf{u}_{k,i}^m, \ \ \mathbf{r}_j^{enc} = \sum_{m=1}^{M} w_m \mathbf{r}_j^m$$

Interactive Matching Network

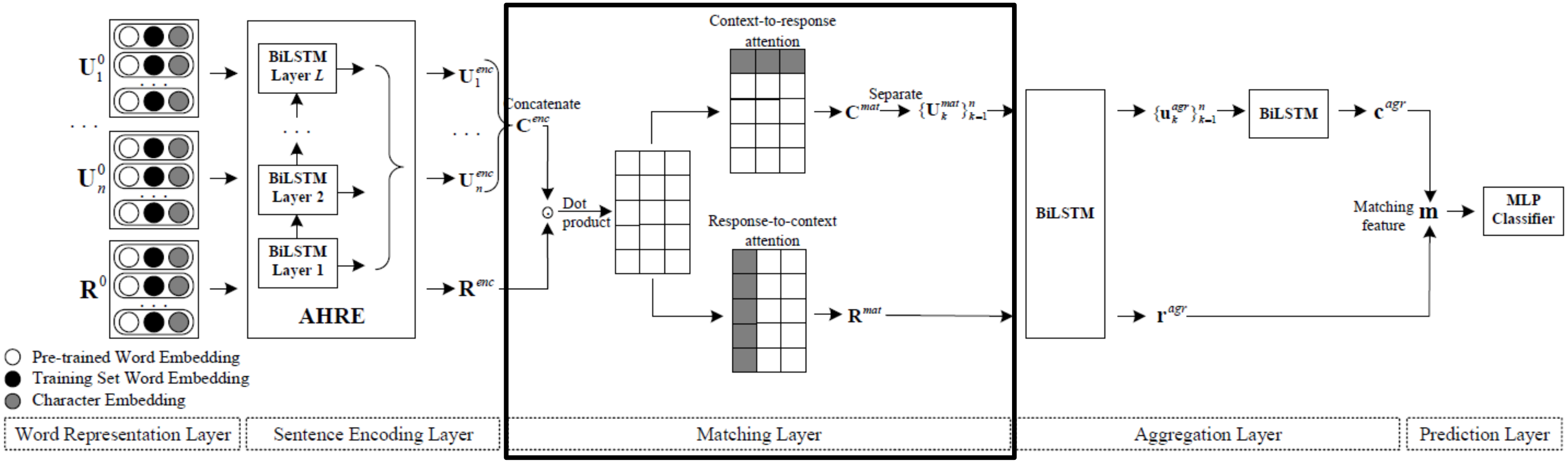


Matching Layer:

Previous work matches in a **local utterance-response** way.

IMN matches in a **global and bidirectional context-response** way.

Interactive Matching Network



Matching Layer:

Concatenate all utterances $\{\mathbf{U}_k^{enc}\}_{k=1}^{n}$ to obtain the context $\mathbf{C}^{enc}$.

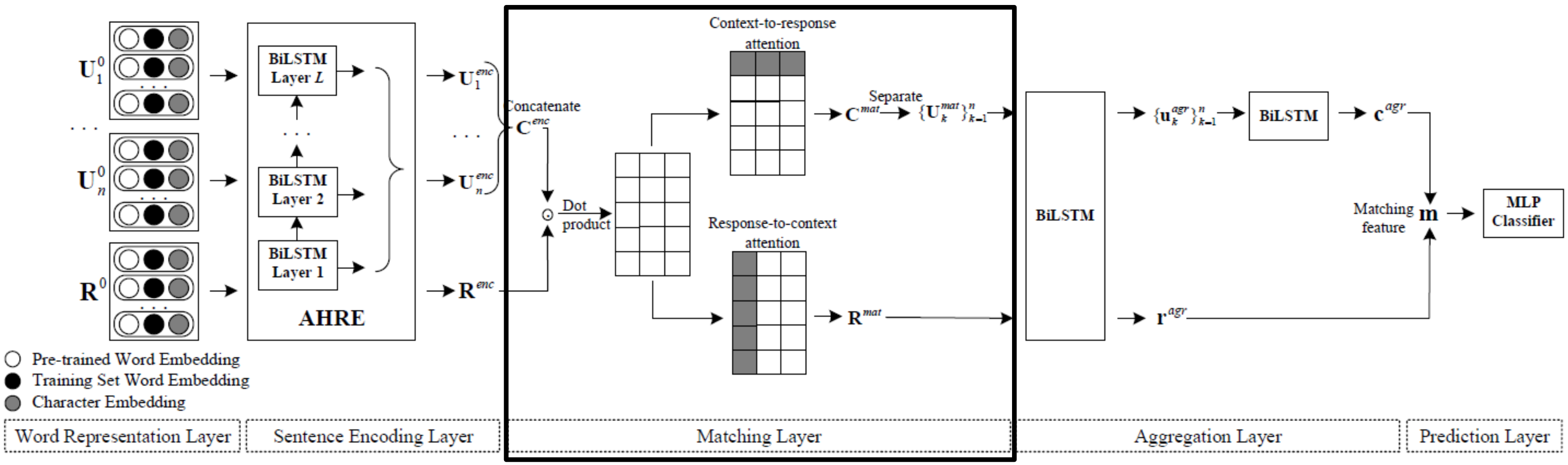
For the response,

$$e_{ij} = (\mathbf{c}_i^{enc})^T \cdot \mathbf{r}_j^{enc}$$

$$\bar{\mathbf{r}}_j^{enc} = \sum_{i=1}^{l_c} \frac{exp(e_{ij})}{\sum_{k=1}^{l_c} exp(e_{kj})} \mathbf{c}_i^{enc}, j \in \{1, ..., l_r\}$$

Similar operations are performed for the context in reverse.

Interactive Matching Network



Matching Layer:

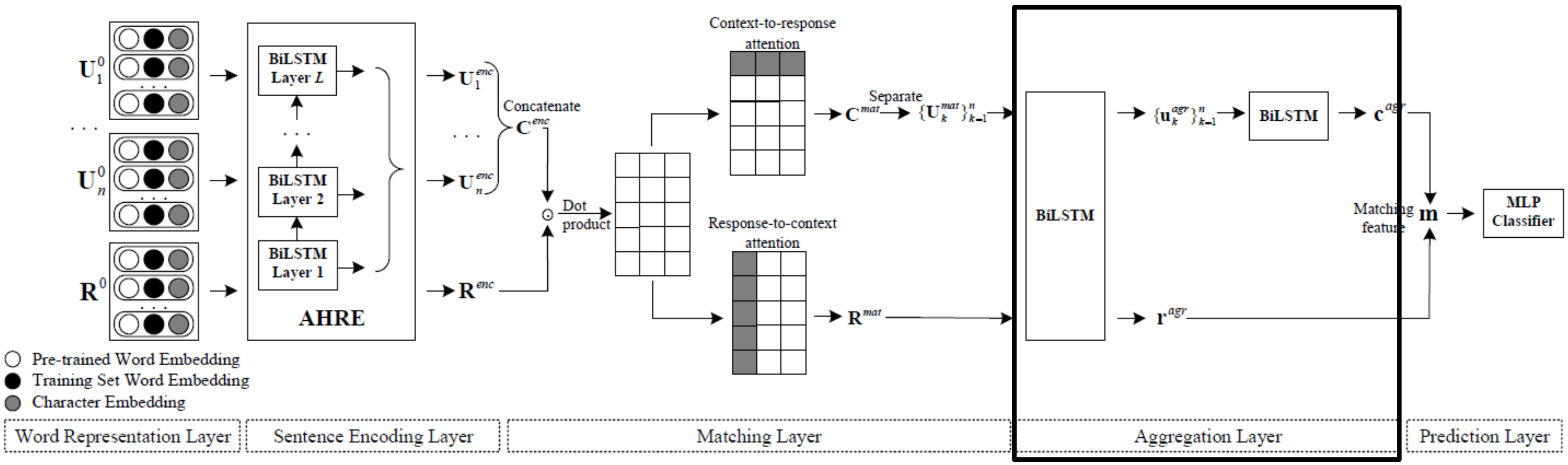
To further enhance the collected information.

$$\mathbf{C}^{mat} = [\mathbf{C}^{enc}; \bar{\mathbf{C}}^{enc}; \mathbf{C}^{enc} - \bar{\mathbf{C}}^{enc}; \mathbf{C}^{enc} \odot \bar{\mathbf{C}}^{enc}],$$

$$\mathbf{R}^{mat} = [\mathbf{R}^{enc}; \bar{\mathbf{R}}^{enc}; \mathbf{R}^{enc} - \bar{\mathbf{R}}^{enc}; \mathbf{R}^{enc} \odot \bar{\mathbf{R}}^{enc}]$$

Finally, the concatenated context $\mathbf{C}^{mat}$ need to be converted to separate utterances $\{\mathbf{U}_k^{mat}\}_{k=1}^{n}$.

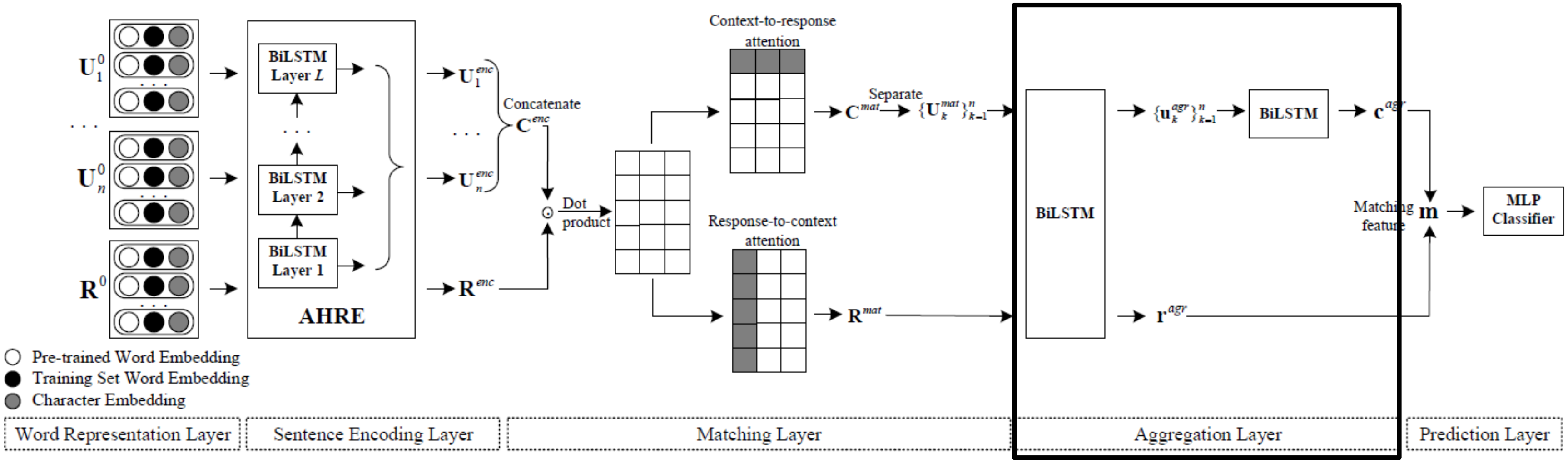
Interactive Matching Network



Aggregation Layer:

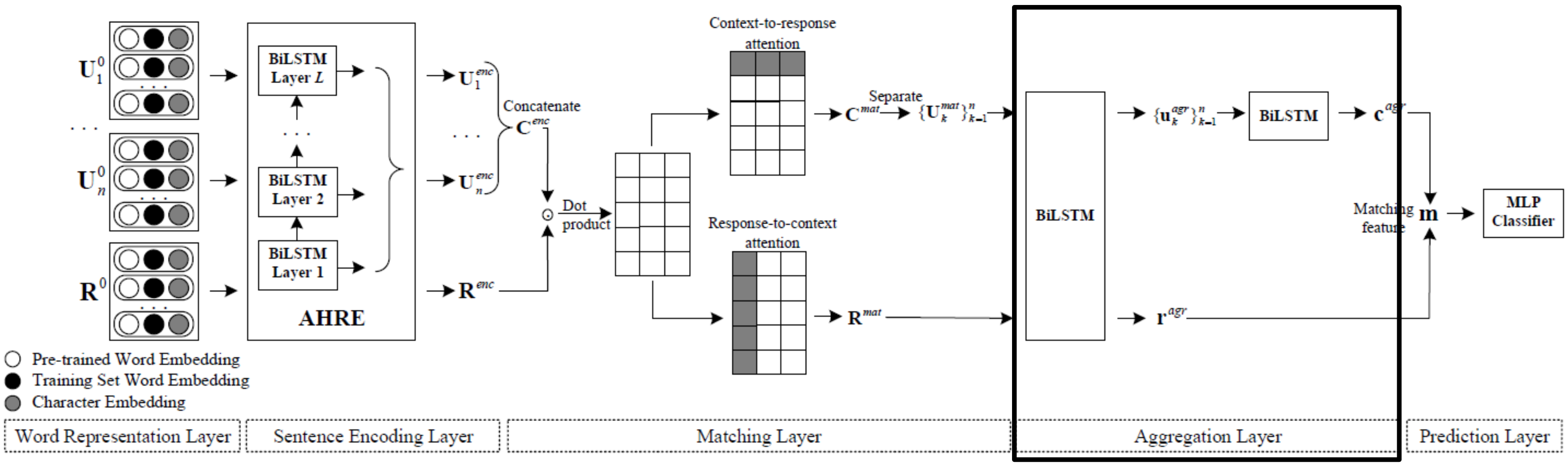
To convert the matching matrices of separated utterances and responses into a final matching vector.

Interactive Matching Network



Aggregation Layer:
Composing the enhanced local matching information $\{\mathbf{U}_k^{mat}\}_{k=1}^n$ and $\mathbf{R}^{mat}$ with a **BiLSTM**, and a combination of **max pooling** and **last hidden state pooling** to obtain a set of utterance embeddings $\mathbf{U}^{agr} = \{\mathbf{u}_k^{agr}\}_{k=1}^n$ and the response embeddings $\mathbf{r}^{agr}$.

Interactive Matching Network

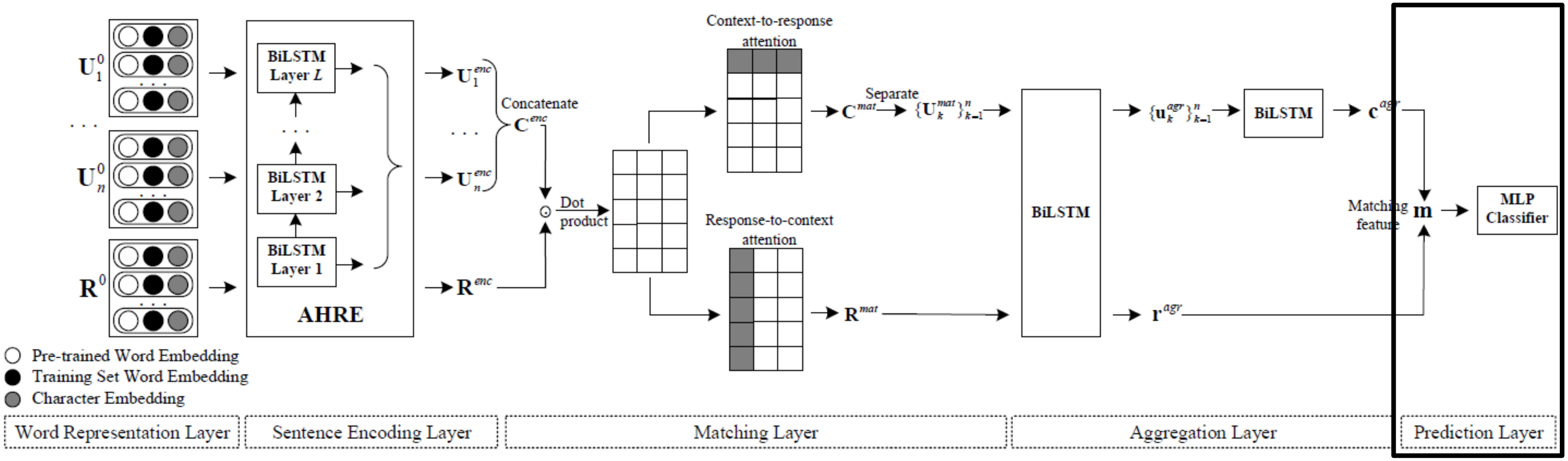


Aggregation Layer:

The set of utterance embeddings $\mathbf{U}^{agr} = \{\mathbf{u}_k^{agr}\}_{k=1}^{n}$ is fed into another BiLSTM in chronological order followed by another pooling operation to obtain the aggregated context embeddings $\mathbf{c}^{agr}$.

The final matching feature vector is $\mathbf{m} = [\mathbf{c}^{agr}; \mathbf{r}^{agr}]$.

Interactive Matching Network



Prediction Layer:

A multi-layer perceptron (MLP) classifier to return a score denoting the matching degree.

Experiments

- Datasets
- Overall Performance
- Analysis

Experiments

- Datasets

Datasets		Train	Valid	Test
Ubuntu V1	pairs	1M	0.5M	0.5M
	positive : negative	1:1	1:9	1:9
	positive/context	1	1	1
Ubuntu V2	pairs	1M	195k	189k
	positive : negative	1:1	1:9	1:9
	positive/context	1	1	1
Douban	pairs	1M	50k	10k
	positive : negative	1:1	1:1	1:9
	positive/context	1	1	1.18
E-commerce	pairs	1M	10k	10k
	positive : negative	1:1	1:1	1:9
	positive/context	1	1	1

Experiments

- Overall Performance

	Ubuntu Corpus V1				Ubuntu Corpus V2			
	R$_2$@1	**R**$_{10}$@1	**R**$_{10}$@2	**R**$_{10}$@5	**R**$_2$@1	**R**$_{10}$@1	**R**$_{10}$@2	**R**$_{10}$@5
TF-IDF (Lowe et al., 2015, 2017)	0.659	0.410	0.545	0.708	0.749	0.488	0.587	0.763
RNN (Lowe et al., 2015, 2017)	0.768	0.403	0.547	0.819	0.777	0.379	0.561	0.836
LSTM (Lowe et al., 2015, 2017)	0.878	0.604	0.745	0.926	0.869	0.552	0.721	0.924
DL2R (Yan et al., 2016)	0.899	0.626	0.783	0.944	-	-	-	-
Match-LSTM (Wang and Jiang, 2016b)	0.904	0.653	0.799	0.944	-	-	-	-
MV-LSTM (Wan et al., 2016)	0.906	0.653	0.804	0.946	-	-	-	-
Multi-View (Zhou et al., 2016)	0.908	0.662	0.801	0.951	-	-	-	-
RNN-CNN (Baudis and Sedivý, 2016)	-	-	-	-	0.911	0.672	0.809	0.956
CompAgg (Wang and Jiang, 2016a)	0.884	0.631	0.753	0.927	0.895	0.641	0.776	0.937
BiMPM (Wang et al., 2017)	0.897	0.665	0.786	0.938	0.877	0.611	0.747	0.921
HRDE-LTC (Yoon et al., 2018)	0.916	0.684	0.822	0.960	0.915	0.652	0.815	0.966
SMN (Wu et al., 2017)	0.926	0.726	0.847	0.961	-	-	-	-
DUA (Zhang et al., 2018)	-	0.752	0.868	0.962	-	-	-	-
DAM (Zhou et al., 2018)	0.938	0.767	0.874	0.969	-	-	-	-
IMN	**0.946**	**0.794**	**0.889**	**0.974**	**0.945**	**0.771**	**0.886**	**0.979**
IMN(Ensemble)	**0.951**	**0.807**	**0.900**	**0.978**	**0.950**	**0.791**	**0.899**	**0.982**

IMN outperforms all other models by a large margin, which shows its effectiveness.

Experiments

- Overall Performance

	Douban Conversation Corpus						E-commerce Corpus		
	MAP	**MRR**	**P**@1	$\mathbf{R}_{10}$@1	$\mathbf{R}_{10}$@2	$\mathbf{R}_{10}$@5	$\mathbf{R}_{10}$@1	$\mathbf{R}_{10}$@2	$\mathbf{R}_{10}$@5
TF-IDF	0.331	0.359	0.180	0.096	0.172	0.405	0.159	0.256	0.477
RNN	0.390	0.422	0.208	0.118	0.223	0.589	0.325	0.463	0.775
LSTM	0.485	0.527	0.320	0.187	0.343	0.720	0.365	0.536	0.828
Multi-View	0.505	0.543	0.342	0.202	0.350	0.729	0.421	0.601	0.861
DL2R	0.488	0.527	0.330	0.193	0.342	0.705	0.399	0.571	0.842
MV-LSTM	0.498	0.538	0.348	0.202	0.351	0.710	0.412	0.591	0.857
Match-LSTM	0.500	0.537	0.345	0.202	0.348	0.720	0.410	0.590	0.858
SMN	0.529	0.569	0.397	0.233	0.396	0.724	0.453	0.654	0.886
DUA	0.551	0.599	0.421	0.243	0.421	0.780	0.501	0.700	0.921
DAM	0.550	0.601	0.427	0.254	0.410	0.757	-	-	-
IMN	**0.570**	**0.615**	**0.433**	**0.262**	**0.452**	**0.789**	**0.621**	**0.797**	**0.964**
IMN(Ensemble)	**0.576**	**0.618**	**0.441**	**0.268**	**0.458**	**0.796**	**0.672**	**0.845**	**0.970**

IMN outperforms all other models by a large margin, which shows its effectiveness.

Experiments

- Ablation Study

-

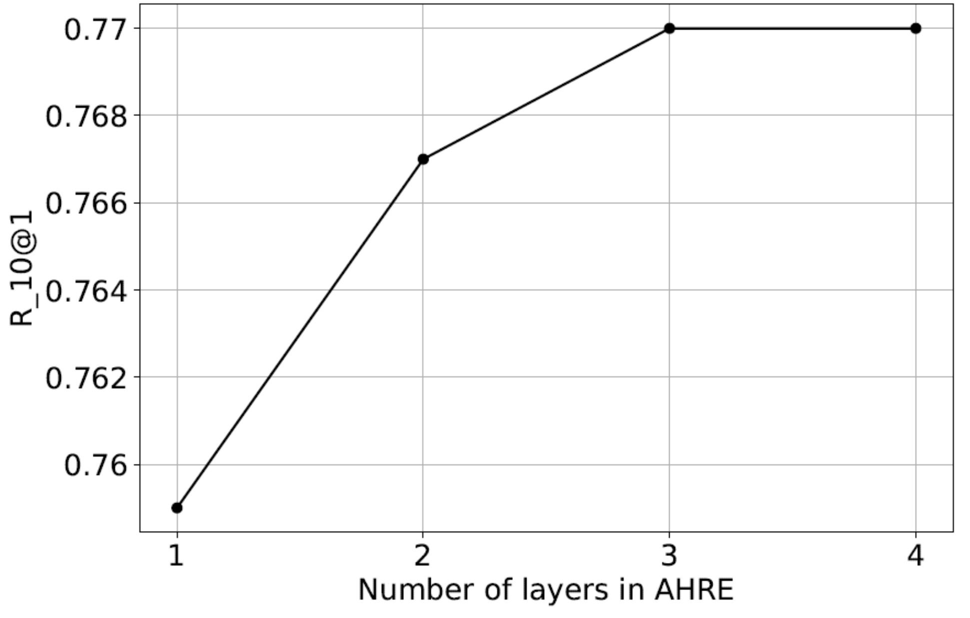
	$\mathbf{R}_2$@1	$\mathbf{R}_{10}$@1	$\mathbf{R}_{10}$@2	$\mathbf{R}_{10}$@5
IMN	0.945	0.771	0.886	0.979
- AHRE	0.940	0.758	0.874	0.974
- Char emb	0.941	0.762	0.878	0.976
- Match	0.904	0.613	0.792	0.958

Ablation tests on the Ubuntu V2 dataset.

Experiments

- AHRE

•



The AHRE proposed in this paper can be considered as a generalized recurrent encoder that degenerates into a single-layer RNN when the number of layers in the AHRE is set to 1.

Experiments

- AHRE
-

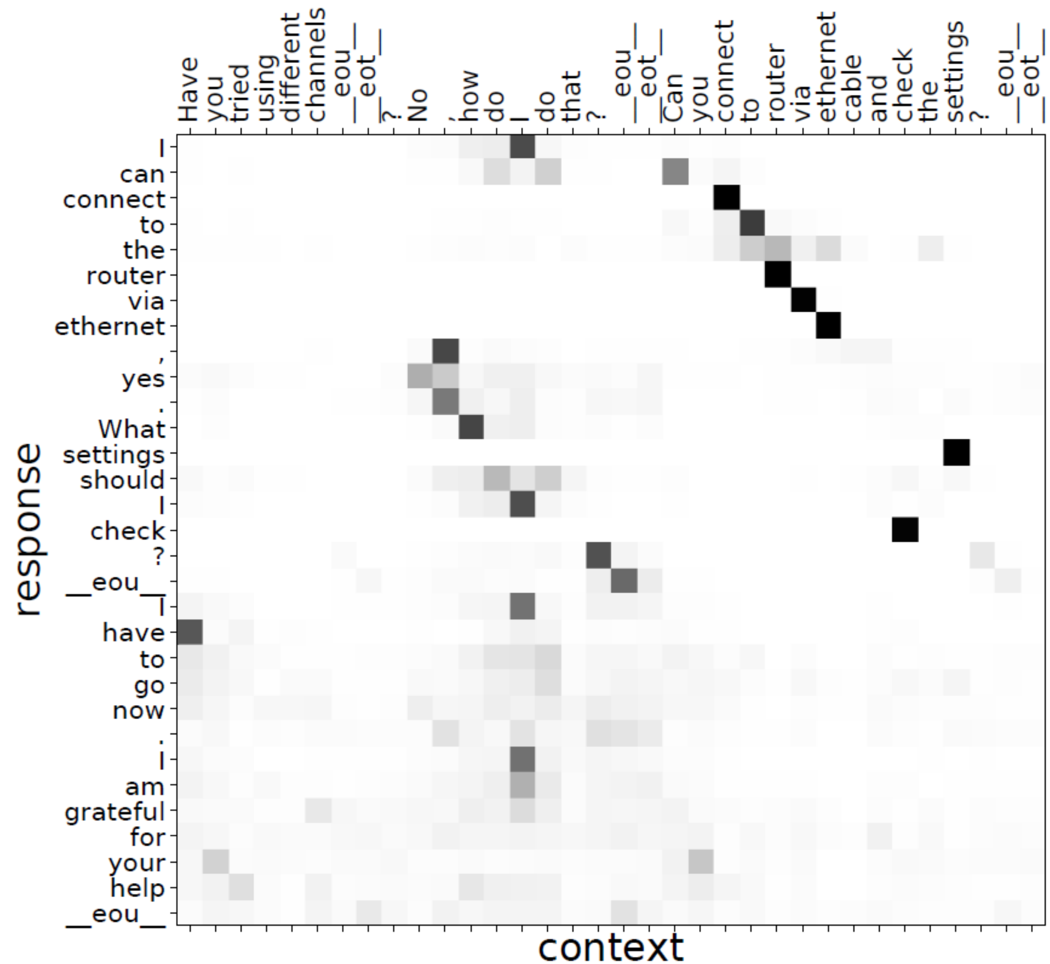
	Layer 1	Layer 2	Layer 3
Weights	0.4938	0.2181	0.2881

The softmax-normalized weights of every layer in the AHRE are listed in Table, which indicates that each layer of the multi-layer RNNs contributes to the sentence embeddings.

Experiments

- Context-Response Matching

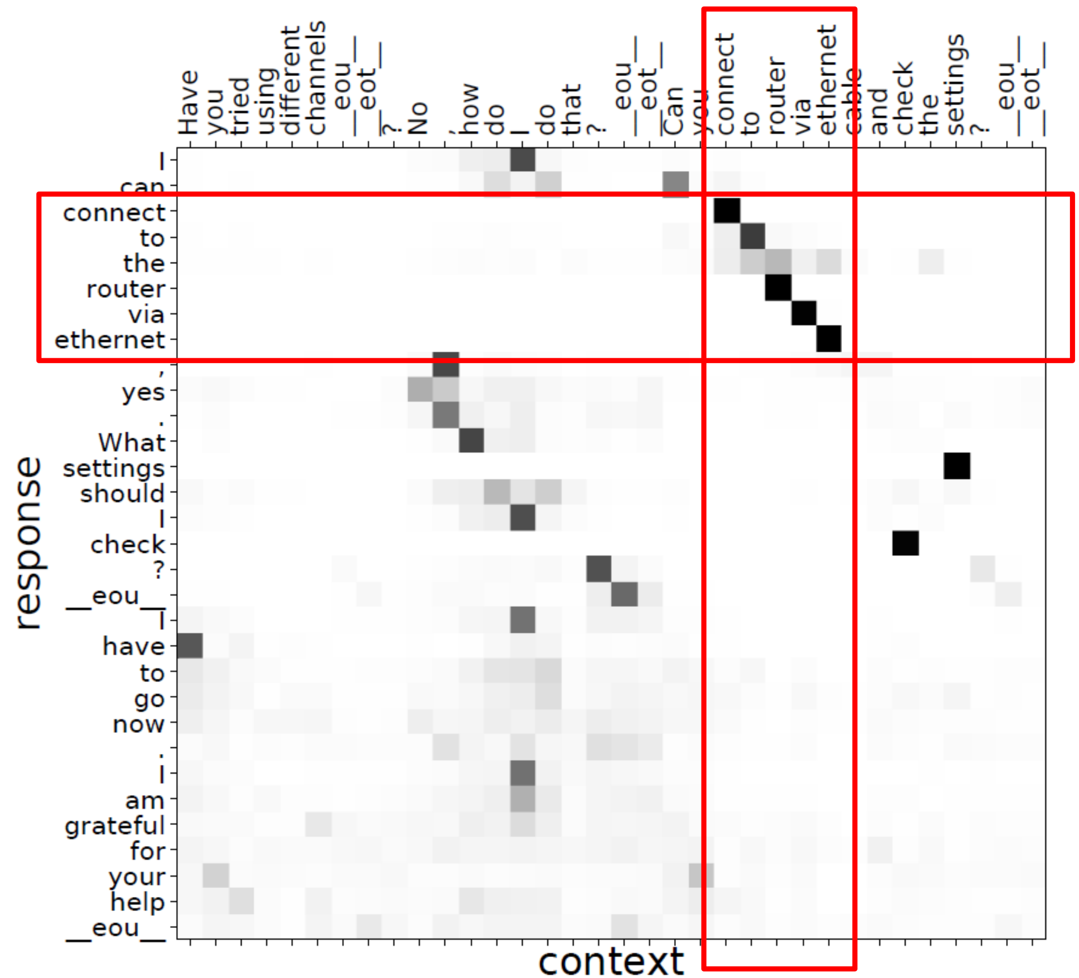
The attention-based matching is able to capture some matching information from the other.



Experiments

- Context-Response Matching

The attention-based matching is able to capture some matching information from the other.



Conclusion

- The representations of the context and response at both the word-level and sentence-level are important for the down-stream matching task.
- Bidirectional and global context-response interactions can help capture the matching information from each other.

Thanks!

Source code

https://github.com/JasonForJoy/IMN