One Time of Interaction May Not Be Enough: Go Deep with an Interaction-over-Interaction Network for Response Selection in Dialogues

Chongyang Tao[1], **Wei Wu**[2], **Can Xu**[2], **Wenpeng Hu**[1], **Dongyan Zhao**[1,3] and **Rui Yan**[1,3*]
[1]Institute of Computer Science and Technology, Peking University, Beijing, China
[2]Microsoft Corporation, Beijing, China
[3]Center for Data Science, Peking University, Beijing, China
[1,3]{chongyangtao,wenpeng.hu,zhaody,ruiyan}@pku.edu.cn
[2]{wuwei,caxu}@microsoft.com

Abstract

Currently, researchers have paid great attention to retrieval-based dialogues in open-domain. In particular, people study the problem by investigating context-response matching for multi-turn response selection based on publicly recognized benchmark data sets. State-of-the-art methods require a response to interact with each utterance in a context from the beginning, but the interaction is performed in a shallow way. In this work, we let utterance-response interaction go deep by proposing an interaction-over-interaction network (IoI). The model performs matching by stacking multiple interaction blocks in which residual information from one time of interaction initiates the interaction process again. Thus, matching information within an utterance-response pair is extracted from the interaction of the pair in an iterative fashion, and the information flows along the chain of the blocks via representations. Evaluation results on three benchmark data sets indicate that IoI can significantly outperform state-of-the-art methods in terms of various matching metrics. Through further analysis, we also unveil how the depth of interaction affects the performance of IoI.

之前的方法中，候选答案
和上下文的交互比较
简单.

被加入

1 Introduction

Building a chitchat style dialogue systems in open-domain for human-machine conversations has attracted increasing attention in the conversational artificial intelligence (AI) community. Generally speaking, there are two approaches to implementing such a conversational system. The first approach leverages techniques of information retrieval (Lowe et al., 2015; Wu et al., 2017; Yan and Zhao, 2018), and selects a proper response from an index; while the second approach directly synthesizes a response with a natural language generation model estimated from a large-scale conversation corpus (Serban et al., 2016; Li et al., 2017b). In this work, we study the problem of multi-turn response selection for retrieval-based dialogue systems where the input is a conversation context consisting of a sequence of utterances. Compared with generation-based methods, retrieval-based methods are superior in terms of response fluency and diversity, and thus have been widely applied in commercial chatbots such as the social bot XiaoIce (Shum et al., 2018) from Microsoft, and the e-commerce assistant AliMe Assist from Alibaba Group (Li et al., 2017a).

A key step in multi-turn response selection is to measure the matching degree between a conversation context and a response candidate. State-of-the-art methods (Wu et al., 2017; Zhou et al., 2018b) perform matching within a representation-interaction-aggregation framework (Wu et al., 2018b) where matching signals in each utterance-response pair are distilled from their interaction based on their representations, and then are aggregated as a matching score. Although utterance-response interaction has proven to be crucial to the performance of the matching models (Wu et al., 2017), it is executed in a rather shallow manner where matching between an utterance and a response candidate is determined only by one step of interaction on each type or each layer of representations. In this paper, we attempt to move from shallow interaction to deep interaction, and consider context-response matching with multiple steps of interaction where residual information from one time of interaction, which is generally ignored by existing methods, is leveraged for additional interactions. The underlying motivation is that if a model extracts some matching information from utterance-response pairs in one step of interaction, then by stacking multiple such steps, the model can gradually accumulate useful signals

[*] Corresponding author: Rui Yan (ruiyan@pku.edu.cn).

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 1–11
Florence, Italy, July 28 - August 2, 2019.

for matching and finally capture the semantic relationship between a context and a response candidate in a more comprehensive way.

We propose an interaction-over-interaction network (IoI) for context-response matching, through which we aim to investigate: (1) how to make interaction go deep in a matching model; and (2) if the depth of interaction really matters in terms of matching performance. A key component in IoI is an interaction block. Taking a pair of utterance-response as input, the block first lets the utterance and the response attend to themselves, and then measures interaction of the pair by an attention-based interaction function. The results of the interaction are concatenated with the self-attention representations and then compressed to new representations of the utterance-response pair as the output of the block. Built on top of the interaction block, IoI initializes each utterance-response pair via pre-trained word embeddings, and then passes the initial representations through a chain of interaction blocks which conduct several rounds of representation-interaction-representation operations and let the utterance and the response interact with each other in an iterative way. Different blocks could distill different levels of matching information in an utterance-response pair. To sufficiently leverage the information, a matching score is first calculated in each block through aggregating matching vectors of all utterance-response pairs, and then the block-wise matching scores are combined as the final matching degree of the context and the response candidate.

We conduct experiments on three benchmark data sets: the Ubuntu Dialogue Corpus (Lowe et al., 2015), the Douban Conversation Corpus (Wu et al., 2017), and the E-commerce Dialogue Corpus (Zhang et al., 2018b). Evaluation results indicate that IoI can significantly outperform state-of-the-art methods with 7 interaction blocks over all metrics on all the three benchmarks. Compared with deep attention matching network (DAM), the best performing baseline on all the three data sets, IoI achieves 2.9% absolute improvement on $R_{10}@1$ on the Ubuntu data, 2.3% absolute improvement on MAP on the Douban data, and 3.7% absolute improvement on $R_{10}@1$ on the E-commerce data. Through more quantitative analysis, we also show that depth indeed brings improvement to the performance of IoI, as IoI with 1 interaction block performs worse than DAM on the Douban data and the E-commerce data, and on the Ubuntu data, the gap on $R_{10}@1$ between IoI and DAM is only 1.1%. Moreover, the improvement brought by depth mainly comes from short contexts.

Our contributions in this paper are three-folds: (1) proposal of a novel interaction-over-interaction network which enables deep-level matching with carefully designed interaction block chains; (2) empirical verification of the effectiveness of the model on three benchmarks; and (3) empirical study on the relationship between interaction depth and model performance.

2 Related Work

Existing methods for building an open-domain dialogue system can be categorized into two groups. The first group learns response generation models under an encoder-decoder framework. On top of the basic sequence-to-sequence with attention architecture (Vinyals and Le, 2015; Shang et al., 2015; Tao et al., 2018), various extensions have been made to tackle the "safe response" problem (Li et al., 2015; Mou et al., 2016; Xing et al., 2017; Zhao et al., 2017; Song et al., 2018); to generate responses with specific personas or emotions (Li et al., 2016a; Zhang et al., 2018a; Zhou et al., 2018a); and to pursue better optimization strategies (Li et al., 2017b, 2016b).

The second group learns a matching model of a human input and a response candidate for response selection. Along this line, the focus of research starts from single-turn response selection by setting the human input as a single message (Wang et al., 2013; Hu et al., 2014; Wang et al., 2015), and moves to context-response matching for multi-turn response selection recently. Representative methods include the dual LSTM model (Lowe et al., 2015), the deep learning to respond architecture (Yan et al., 2016), the multi-view matching model (Zhou et al., 2016), the sequential matching network (Wu et al., 2017, 2018b), and the deep attention matching network (Zhou et al., 2018b). Besides model design, some attention is also paid to the learning problem of matching models (Wu et al., 2018a). Our work belongs to the second group. The proposed interaction-over-interaction network is unique in that it performs matching by stacking multiple interaction blocks, and thus extends the shallow interaction in state-of-the-art methods to a deep

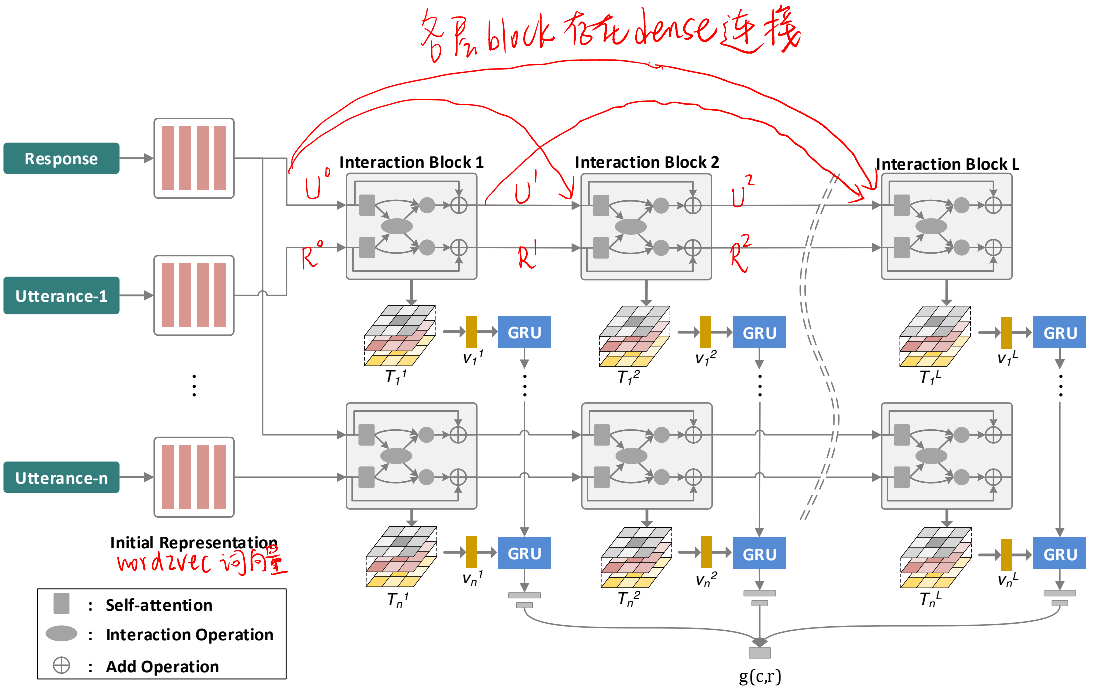


Figure 1: Architecture of interaction-over-interaction network.

form. As far as we know, this is the first architecture that realizes deep interaction for multi-turn response selection.

Encouraged by the big success of deep neural architectures such as Resnet (He et al., 2016) and inception (Szegedy et al., 2015) in computer vision, researchers have studied if they can achieve similar results with deep neural networks on NLP tasks. Although deep models have not yet brought breakthroughs to NLP as they do to computer vision, they have proven effective in a few tasks such as text classification (Conneau et al., 2017), natural language inference (Kim et al., 2018; Tay et al., 2018), and question answering (Tay et al., 2018; Kim et al., 2018), etc. In this work, we attempt to improve the accuracy of multi-turn response selection in retrieval-based dialogue systems by increasing the depth of context-response interaction in matching. Through extensive studies on benchmarks, we show that depth can bring significant improvement to model performance on the task.

3 Problem Formalization

Suppose that there is a conversation data set $\mathcal{D} = \{(y_i, c_i, r_i)\}_{i=1}^N$. $\forall i \in \{1, \ldots, N\}$, $c_i = \{u_{i,1}, \ldots, u_{i,l_i}\}$ represents a conversation context with $u_{i,k}$ the k-th turn, r_i is a response candidate, and $y_i \in \{0, 1\}$ denotes a label with $y_i = 1$ indicating r_i a proper response for c_i, otherwise $y_i = 0$. The task is to learn a matching model $g(\cdot, \cdot)$ from $\mathcal{D}$, and thus for a new context-response pair (c, r), $g(c, r)$ measures the matching degree between c and r.

In the following sections, we will elaborate how to define $g(\cdot, \cdot)$ to achieve deep interaction between c and r, and how to learn such a deep model from $\mathcal{D}$.

4 Interaction-over-Interaction Network

总体思路

We define $g(\cdot, \cdot)$ as an interaction-over-interaction network (IoI). Figure 1 illustrates the architecture of IoI. The model pairs each utterance in a context with a response candidate, 每个block and then aggregates 每层交互的Matching matching information from all the pairs as a matching score of the context and the response candidate. For each pair, IoI starts from initial representations of the utterance and the 多层堆叠的blocks response, and then feeds the pair to stacked interaction blocks. Each block represents the utterance and the response by letting them interact with each other based on the interactions before. Matching signals 每层交互的Matching are first accumulated along the sequence of the utterances in each block, and then combined along 总Matching the chain of blocks as the final matching score. Below we will describe details of components of IoI and how to learn the model with $\mathcal{D}$.

4.1 Initial Representations

Given an utterance u in a context c and a response candidate r, u and r are initialized as $\mathbf{E_u} = [\mathbf{e}_{u,1}, \cdots, \mathbf{e}_{u,m}]$ and $\mathbf{E_r} = [\mathbf{e}_{r,1}, \cdots, \mathbf{e}_{r,n}]$ respectively. wordzvec $\forall i \in \{1, \ldots, m\}$ and $\forall j \in \{1, \ldots, n\}$, $\mathbf{e}_{u,i}$ and $\mathbf{e}_{r,j}$ are representations of the i-th word of u and the j-th word of r respectively which are obtained by pre-training Word2vec (Mikolov et al., 2013) on $\mathcal{D}$. $\mathbf{E_u}$ and $\mathbf{E_r}$ are then processed by stacked interaction blocks that model different levels of interaction between u and r and generate matching signals.

4.2 Interaction Block

每个block由三个module组成

The stacked interaction blocks share the same internal structure. In a nutshell, each block is composed of a self-attention module that captures long-term dependencies within an utterance and a response, an interaction module that models the interaction between the utterance and the response, and a compression module that condenses the results of the first two modules into representations of the utterance and the response as output of the block. The output is then utilized as the input of the next block.

Before diving to details of the block, we first generally describe an attention mechanism that lays a foundation for the self-attention module and the interaction module. Let $\mathbf{Q} \in \mathbb{R}^{n_q \times d}$ and $\mathbf{K} \in \mathbb{R}^{n_k \times d}$ be a query and a key respectively, where n_q and n_k denote numbers of words and d is the embedding size, then attention from $\mathbf{Q}$ to $\mathbf{K}$ is defined as

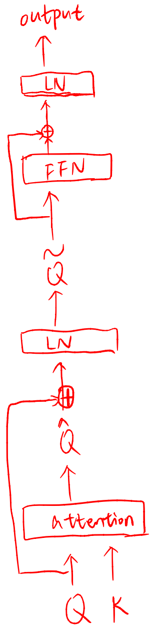
定义 $f_{ATT}(Q,K)$

$$\hat{\mathbf{Q}} = S(\mathbf{Q},\mathbf{K}) \cdot \mathbf{K}, \in n_q \times d \qquad (1)$$

where $S(\cdot,\cdot)$ is a function for attention weight calculation. Here, we exploit the symmetric function in (Huang et al., 2017b) as $S(\cdot,\cdot)$ which is given by:

$$S(\mathbf{Q},\mathbf{K}) \overset{\in n_q \times n_k}{=} \mathrm{softmax}(\overset{\text{ReLU}}{\textcircled{f}}(\mathbf{QW})\overset{\text{对角矩阵}}{\textcircled{\mathbf{D}}} f(\underset{n_k \times d}{\underline{\mathbf{K}}}\underset{d\times d}{\underline{\mathbf{W}}})^\top). \qquad (2)$$

In Equation (2), f is a ReLU activation function, $\mathbf{D}$ is a diagonal matrix, and both $\mathbf{D} \in \mathbb{R}^{d \times d}$ and $\mathbf{W} \in \mathbb{R}^{d \times d}$ are parameters to estimate from training data. Intuitively, in Equation (1), each entry of $\mathbf{K}$ is weighted by an importance score defined by the similarity of an entry of $\mathbf{Q}$ and an entry of $\mathbf{K}$. The entries of $\mathbf{K}$ are then linearly combined with the weights to form a new representation of $\mathbf{Q}$.



A residual connection (He et al., 2016) and a layer normalization (Ba et al., 2016) are then applied to $\hat{\mathbf{Q}}$ as $\tilde{\mathbf{Q}}$. After that, $\tilde{\mathbf{Q}}$ is fed to a feed forward network which is formulated as

残差+layerNorm: $\tilde{Q} = LN(Q + \hat{Q})$

$$\text{2层FFN:}\quad \mathrm{ReLU}(\tilde{\mathbf{Q}}\mathbf{W}_1 + \mathbf{b}_1)\mathbf{W}_2 + \mathbf{b}_2, \qquad (3)$$

where $\mathbf{W}_{\{1,2\}} \in \mathbb{R}^{d \times d}$ and $\mathbf{b}_{\{1,2\}}$ are parameters. The output of the attention mechanism is defined with the result of Equation (3) after another round of residual connection and layer normalization. For ease of presentation, we denote the entire attention mechanism as $f_{ATT}(\mathbf{Q},\mathbf{K})$.

Let $\mathbf{U}^{k-1}$ and $\mathbf{R}^{k-1}$ be the input of the k-th block where $\mathbf{U}^0 = \mathbf{E_u}$ and $\mathbf{R}^0 = \mathbf{E_r}$, then the self-attention module is defined as

self-attention:

$$\hat{\mathbf{U}}^k = f_{\text{ATT}}(\mathbf{U}^{k-1}, \mathbf{U}^{k-1}), \qquad (4)$$

$$\hat{\mathbf{R}}^k = f_{\text{ATT}}(\mathbf{R}^{k-1}, \mathbf{R}^{k-1}). \qquad (5)$$

The interaction module first lets $\mathbf{U}^{k-1}$ and $\mathbf{R}^{k-1}$ attend to each other by

interaction:

$$\overline{\mathbf{U}}^k = f_{\text{ATT}}(\mathbf{U}^{k-1}, \mathbf{R}^{k-1}), \qquad (6)$$

$$\overline{\mathbf{R}}^k = f_{\text{ATT}}(\mathbf{R}^{k-1}, \mathbf{U}^{k-1}). \qquad (7)$$

Then $\mathbf{U}^{k-1}$ and $\mathbf{R}^{k-1}$ further interact with $\overline{\mathbf{U}}^k$ and $\overline{\mathbf{R}}^k$ respectively, which can be formulated as

interaction:

$$\tilde{\mathbf{U}}^k = \mathbf{U}^{k-1} \odot \overline{\mathbf{U}}^k, \qquad (8)$$

$$\tilde{\mathbf{R}}^k = \mathbf{R}^{k-1} \odot \overline{\mathbf{R}}^k, \qquad (9)$$

where $\odot$ denotes element-wise multiplication. Finally, the compression module updates $\mathbf{U}^{k-1}$ and $\mathbf{R}^{k-1}$ to $\mathbf{U}^k$ and $\mathbf{R}^k$ as the output of the block. Suppose that $\mathbf{e}^k_{u,i}$ and $\mathbf{e}^k_{r,i}$ are the i-th entries of $\mathbf{U}^k$ and $\mathbf{R}^k$ respectively, then $\mathbf{e}^k_{u,i}$ and $\mathbf{e}^k_{r,i}$ are calculated by

compression:

$$\mathbf{e}^k_{u,i} = \mathrm{ReLU}(\mathbf{w}_p \begin{bmatrix} \mathbf{e}^{k-1}_{u,i} \\ \hat{\mathbf{e}}^k_{u,i} \\ \overline{\mathbf{e}}^k_{u,i} \\ \tilde{\mathbf{e}}^k_{u,i} \end{bmatrix} + \mathbf{b}_p) + \mathbf{e}^{k-1}_{u,i}, \qquad (10)$$

$$\mathbf{e}^k_{r,i} = \mathrm{ReLU}(\mathbf{w}_p \begin{bmatrix} \mathbf{e}^{k-1}_{r,i} \\ \hat{\mathbf{e}}^k_{r,i} \\ \overline{\mathbf{e}}^k_{r,i} \\ \tilde{\mathbf{e}}^k_{r,i} \end{bmatrix} + \mathbf{b}_p) + \mathbf{e}^{k-1}_{r,i}, \qquad (11)$$

where $\mathbf{w}_p \in \mathbb{R}^{4d \times d}$ and $\mathbf{b}_p$ are learnable projection weights and biases, $\hat{\mathbf{e}}^k_{\{u,r\},i}$, $\overline{\mathbf{e}}^k_{\{u,r\},i}$, $\tilde{\mathbf{e}}^k_{\{u,r\},i}$, and $\mathbf{e}^{k-1}_{\{u,r\},i}$ are the i-th entries of $\{\hat{\mathbf{U}}, \hat{\mathbf{R}}\}^k$, $\{\overline{\mathbf{U}}, \overline{\mathbf{R}}\}^k$, $\{\tilde{\mathbf{U}}, \tilde{\mathbf{R}}\}^k$, and $\{\mathbf{U}, \mathbf{R}\}^{k-1}$, respectively. Inspired by Huang et al. (2017a), we also introduce direct connections from initial representations to all their corresponding subsequent blocks. ???

4.3 Matching Aggregation

Suppose that $c = (u_1, \dots, u_l)$ is a conversation context with u_i the i-th utterance, then in the k-th interaction block, we construct three similarity matrices by

Matching:

$$
\begin{aligned}
\mathbf{M}_{i,1}^{k} &= \frac{\overset{m\times d}{\mathbf{U}_i^{k-1}} \cdot \overset{d\times n}{(\mathbf{R}^{k-1})^\top}}{\sqrt{d}} \in m\times n \\
\mathbf{M}_{i,2}^{k} &= \frac{\hat{\mathbf{U}}_i^{k} \cdot (\hat{\mathbf{R}}^{k})^\top}{\sqrt{d}}, \in m\times n \qquad (12) \\
\mathbf{M}_{i,3}^{k} &= \frac{\overline{\mathbf{U}}_i^{k} \cdot (\overline{\mathbf{R}}^{k})^\top}{\sqrt{d}}, \in m\times n
\end{aligned}
$$

where $\mathbf{U}_i^{k-1}$ and $\mathbf{R}^{k-1}$ are the input of the k-th block, $\hat{\mathbf{U}}_i^k$ and $\hat{\mathbf{R}}^k$ are defined by Equations (4-5), and $\overline{\mathbf{U}}_i^k$ and $\overline{\mathbf{R}}^k$ are calculated by Equations (6-7). The three matrices are then concatenated into a 3-D matching tensor $\mathbf{T}_i^k \in \mathbb{R}^{m_i \times n \times 3}$ which can be written as

$$\mathbf{T}_i^k = \mathbf{M}_{i,1}^k \oplus \mathbf{M}_{i,2}^k \oplus \mathbf{M}_{i,3}^k, \in m\times n\times 3 \qquad (13)$$

第i个上下文表示: V_i^k = 对T_i^k进行卷积 + flatten + 映射到d维向量.

where $\oplus$ denotes a concatenation operation, and m_i and n refer to numbers of words in u_i and r respectively.

We exploit a convolutional neural network (Krizhevsky et al., 2012) to extract matching features from $\mathbf{T}_i^k$. The output of the final feature maps are flattened and mapped to a d-dimensional matching vector $\mathbf{v}_i^k$ with a linear transformation. $(\mathbf{v}_1^k, \cdots, \mathbf{v}_l^k)$ is then fed to a GRU (Chung et al., 2014) to capture temporal relationship among $(u_1, \ldots, u_l)$. $\forall i \in \{1, \ldots, l\}$, the i-th hidden state of the GRU model is given by

$$\underline{\mathbf{h}_i^k = \mathrm{GRU}(\mathbf{v}_i^k, \mathbf{h}_{i-1}^k),} \qquad (14)$$

where $\mathbf{h}_0^k$ is randomly initialized. A matching score for context c and response candidate r in the k-th block is defined as

第k层交互得分: $$\underline{g^k(c,r) = \overset{\text{sigmoid}}{\sigma}(\mathbf{h}^k_{\underset{\text{总对话上下文数}}{l}} \cdot \mathbf{w}_o + \mathbf{b}_o),} \qquad (15)$$

where $\mathbf{w}_o$ and $\mathbf{b}_o$ are parameters, and $\sigma(\cdot)$ is a sigmoid function. Finally, $g(c,r)$ is defined by

总得分: $$\underline{g(c,r) = \sum_{k=1}^{\overset{\text{总交互层数}}{L}} g^k(c,r),} \qquad (16)$$

where L is the number of interaction blocks in IoI. Note that we define $g(c,r)$ with all blocks rather than only with the last block. This is motivated by (1) only using the last block will make training of IoI difficult due to the gradient vanishing/exploding problem; and (2) different blocks may capture different levels of matching information in (c,r), and thus leveraging all of them could enhance matching accuracy.

5 Learning Methods

两种训练策略

We consider two strategies to learn an IoI model from the training data $\mathcal{D}$. The first strategy estimates the parameters of IoI (denoted as Θ) by minimizing a global loss function that is formulated as

最小化一个全局损失函数:

$$-\sum_{i=1}^{N} \left[y_i \log(g(c_i, r_i)) + (1-y_i) \log(1-g(c_i, r_i)) \right]. \qquad (17)$$

In the second strategy, we construct a local loss function for each block and minimize the summation of the local loss functions. By this means, each block can be directly supervised by the labels in $\mathcal{D}$ during learning. The learning objective is then defined as

每层交互有一个局部损失函数

$$
\begin{aligned}
-\sum_{k=1}^{L} \sum_{i=1}^{N} \big[& y_i \log(g^k(c_i, r_i)) \\
& + (1-y_i) \log(1-g^k(c_i, r_i)) \big].
\end{aligned} \qquad (18)
$$

We compare the two learning strategies through empirical studies, as will be reported in the next section. In both strategies, Θ are optimized using back-propagation with Adam algorithm (Kingma and Ba, 2015).

6 Experiments

We test the proposed IoI on three benchmark data sets for multi-turn response selection.

6.1 Experimental Setup

The first data we use is the Ubuntu Dialogue Corpus (Lowe et al., 2015) which is a multi-turn English conversation data set constructed from chat logs of the Ubuntu forum. We use the version provided by Xu et al. (2017). The data contains 1 million context-response pairs for training, and 0.5 million pairs for validation and test. In all the three sets, positive responses are human responses, while negative ones are randomly sampled. The ratio of the positive and the negative is 1:1 in the training set, and 1:9 in both the validation set and the test set. Following Lowe et al. (2015), we employ recall at position k in n candidates (R_n@k) as evaluation metrics. 评估指标

The second data set is the Douban Conversation Corpus (Wu et al., 2017) that consists of multi-turn Chinese conversations collected from Douban group[1]. There are 1 million context-response pairs

[1] https://www.douban.com/group

for training, 50 thousand pairs for validation, and 6,670 pairs for testing. In the training set and the validation set, the last turn of each conversation is taken as a positive response and a negative response is randomly sampled. For each context in the test set, 10 response candidates are retrieved from an index and their appropriateness regarding to the context is annotated by human labelers. Following Wu et al. (2017), we employ R_n@ks, mean average precision (MAP), mean reciprocal 指标 rank (MRR) and precision at position 1 (P@1) as evaluation metrics.

Finally, we choose the E-commerce Dialogue Corpus (Zhang et al., 2018b) as an experimental data set. The data consists of multi-turn real-world conversations between customers and customer service staff in Taobao[2], which is the largest e-commerce platform in China. It contains 1 million context-response pairs for training, and 10 thousand pairs for validation and test. Positive responses in this data are real human responses, and negative candidates are automatically constructed by ranking the response corpus based on conversation history augmented messages using Apache Lucene[3]. The ratio of the positive and the negative is 1:1 in training and validation, and 1:9 in test. Following (Zhang et al., 2018b), we employ 指标 R_{10}@1, R_{10}@2, and R_{10}@5 as evaluation metrics.

6.2 Baselines

We compare IoI with the following models:

Single-turn Matching Models: these models, including RNN (Lowe et al., 2015), CNN (Lowe et al., 2015), LSTM (Lowe et al., 2015), BiLSTM (Kadlec et al., 2015), MV-LSTM (Wan et al., 2016) and Match-LSTM (Wang and Jiang, 2016), perform context-response matching by concatenating all utterances in a context into a single long document and calculating a matching score between the document and a response candidate.

Multi-View (Zhou et al., 2016): the model calculates matching degree between a context and a response candidate from both a word sequence view and an utterance sequence view.

DL2R (Yan et al., 2016): the model first reformulates the last utterance with previous turns in a context with different approaches. A response candidate and the reformulated message are then represented by a composition of RNN and CNN. Finally, a matching score is computed with the concatenation of the representations.

SMN (Wu et al., 2017): the model lets each utterance in a context interact with a response candidate at the beginning, and then transforms interaction matrices into a matching vector with CNN. The matching vectors are finally accumulated with an RNN as a matching score.

DUA (Zhang et al., 2018b): the model considers the relationship among utterances within a context by exploiting deep utterance aggregation to form a fine-grained context representation. Each refined utterance then matches with a response candidate, and their matching degree is finally calculated through an aggregation on turns.

DAM (Zhou et al., 2018b): the model lets each utterance in a context interact with a response candidate at different levels of representations obtained by a stacked self-attention module and a cross-attention module.

For the Ubuntu data and the Douban data, since results of all baselines under fine-tuning are available in Zhou et al. (2018b), we directly copy the numbers from the paper. For the E-commerce data, Zhang et al. (2018b) report performance of all baselines except DAM. Thus, we copy all available numbers from the paper and implement DAM with the published code[4]. In order to conduct statistical tests, we also run the code of DAM on the Ubuntu data and the Douban data.

6.3 Implementation Details

In IoI, we set the size of word embedding as 200. For the CNN in matching aggregation, we set the window size of convolution and pooling kernels as 两层卷积 $(3,3)$, and the strides as $(1,1)$ and $(3,3)$ respectively. The number of convolution kernels is 32 in the first layer and 16 in the second layer. The dimension of the hidden states of GRU is set as 200. Following Wu et al. (2017), we limit the length of a context to 10 turns and the length of an utterance (either from a context or from a response candidate) to 50 words. Truncation or zero-padding is applied to a context or a response candidate when necessary. We gradually increase the number of interaction blocks (i.e., L) in IoI, and finally set $L = 7$ in comparison with the baseline models. In optimization, we choose 0.2 as a dropout rate, and 50 as the size of mini-batches. The learning rate is initialized as 0.0005, and exponentially decayed

[2] https://www.taobao.com

[3] http://lucene.apache.org/

[4] https://github.com/baidu/Dialogue

Metrics Models	Ubuntu Corpus				Douban Corpus					
	$R_2@1$	$R_{10}@1$	$R_{10}@2$	$R_{10}@5$	MAP	MRR	P@1	$R_{10}@1$	$R_{10}@2$	$R_{10}@5$
RNN (Lowe et al., 2015)	0.768	0.403	0.547	0.819	0.390	0.422	0.208	0.118	0.223	0.589
CNN (Lowe et al., 2015)	0.848	0.549	0.684	0.896	0.417	0.440	0.226	0.121	0.252	0.647
LSTM (Lowe et al., 2015)	0.901	0.638	0.784	0.949	0.485	0.527	0.320	0.187	0.343	0.720
BiLSTM (Kadlec et al., 2015)	0.895	0.630	0.780	0.944	0.479	0.514	0.313	0.184	0.330	0.716
DL2R (Yan et al., 2016)	0.899	0.626	0.783	0.944	0.488	0.527	0.330	0.193	0.342	0.705
MV-LSTM (Wan et al., 2016)	0.906	0.653	0.804	0.946	0.498	0.538	0.348	0.202	0.351	0.710
Match-LSTM (Wang and Jiang, 2016)	0.904	0.653	0.799	0.944	0.500	0.537	0.345	0.202	0.348	0.720
Multi-View (Zhou et al., 2016)	0.908	0.662	0.801	0.951	0.505	0.543	0.342	0.202	0.350	0.729
SMN (Wu et al., 2017)	0.926	0.726	0.847	0.961	0.529	0.569	0.397	0.233	0.396	0.724
DUA(Zhang et al., 2018b)	-	0.752	0.868	0.962	0.551	0.599	0.421	0.243	0.421	0.780
DAM (Zhou et al., 2018b)	0.938	0.767	0.874	0.969	0.550	0.601	0.427	0.254	0.410	0.757
IoI-global	**0.941**	**0.778**	**0.879**	0.970	**0.566**	0.608	0.433	0.263	**0.436**	**0.781**
IoI-local	**0.947**	**0.796**	**0.894**	**0.974**	**0.573**	**0.621**	0.444	**0.269**	**0.451**	**0.786**

Table 1: Evaluation results on the Ubuntu data and the Douban data. Numbers in bold mean that the improvement to the best performing baseline is statistically significant (t-test with p-value < 0.05).

Metrics Models	$R_{10}@1$	$R_{10}@2$	$R_{10}@5$
RNN (Lowe et al., 2015)	0.325	0.463	0.775
CNN (Lowe et al., 2015)	0.328	0.515	0.792
LSTM (Lowe et al., 2015)	0.365	0.536	0.828
BiLSTM (Kadlec et al., 2015)	0.355	0.525	0.825
DL2R (Yan et al., 2016)	0.399	0.571	0.842
MV-LSTM (Wan et al., 2016)	0.412	0.591	0.857
Match-LSTM (Wang and Jiang, 2016)	0.410	0.590	0.858
Multi-View (Zhou et al., 2016)	0.421	0.601	0.861
SMN (Wu et al., 2017)	0.453	0.654	0.886
DUA(Zhang et al., 2018b)	0.501	0.700	0.921
DAM (Zhou et al., 2018b)	0.526	0.727	0.933
IoI-global	**0.554**	0.747	0.942
IoI-local	**0.563**	**0.768**	**0.950**

Table 2: Evaluation results on the E-commerce data. Numbers in bold mean that the improvement to the best performing baseline is statistically significant (t-test with p-value < 0.05).

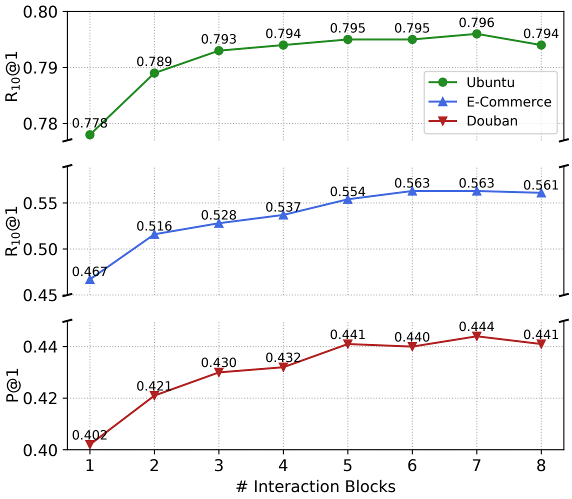


Figure 2: Performance of IoI under different numbers of the interaction blocks.

during training.

6.4 Evaluation Results

Table 1 and Table 2 report evaluation results on the three data sets where IoI-global and IoI-local represent models learned with Objective (17) and Objective (18) respectively. We can see that both IoI-local and IoI-global outperform the best performing baseline, and improvements from IoI-local on all metrics and from IoI-global on a few metrics are statistically significant (t-test with p-value < 0.05). IoI-local is consistently better than IoI-global over all metrics on all the three data sets, demonstrating that directly supervising each block in learning can lead to a more optimal deep structure than optimizing the final matching model.

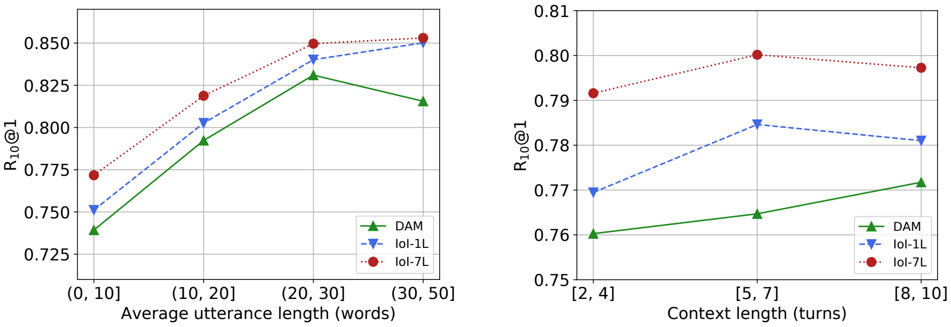
6.5 Discussions

In this section, we make some further analysis with IoI-local to understand (1) how depth of interaction affects the performance of IoI; (2) how context length affects the performance of IoI; and (3) importance of different components of IoI with respect to matching accuracy.

Impact of interaction depth. Figure 2 illustrates how the performance of IoI changes with respect to the number of interaction blocks on test sets of the three data. From the chart, we observe a consistent trend over the three data sets: there is significant improvement during the first few blocks, and then the performance of the model becomes stable. The results indicate that depth of interaction indeed matters in terms of matching accuracy. With shallow interaction ($L = 1$), IoI performs worse than DAM on the Douban data and the E-commerce data. Only after the interaction goes deep ($L \geq 5$), improvement from IoI

Models \ Metrics	Ubuntu data			Douban data			E-commerce data		
	R_2@1	R_{10}@1	R_{10}@2	MAP	MRR	P@1	R_{10}@1	R_{10}@2	R_{10}@5
IoI	0.947	0.796	0.894	0.573	0.621	0.444	0.563	0.768	0.947
IoI-E	0.947	0.794	0.891	0.568	0.616	0.438	0.559	0.762	0.943
IoI-$\hat{E}$	0.946	0.790	0.888	0.565	0.613	0.433	0.557	0.749	0.941
IoI-$\overline{E}$	0.947	0.793	0.890	0.566	0.613	0.439	0.560	0.754	0.943
IoI-$\tilde{E}$	0.947	0.795	0.891	0.571	0.616	0.441	0.562	0.740	0.944
IoI-M_1	0.946	0.793	0.890	0.568	0.611	0.436	0.557	0.743	0.943
IoI-M_2	0.944	0.788	0.886	0.562	0.605	0.427	0.551	0.739	0.942
IoI-M_3	0.946	0.793	0.889	0.567	0.615	0.438	0.558	0.748	0.946

Table 3: Evaluation results of the ablation study on the three data sets.



(a) R_{10}@1 vs. Average utterance length

(b) R_{10}@1 vs. Number of turns

Figure 3: Performance of IoI across contexts with different lengths on the Ubuntu data.

to DAM on the two data becomes significant. On the Ubuntu data, improvement to DAM from the deep model ($L = 7$) is more than twice as much as that from the shallow model ($L = 1$). The performance of IoI becomes stable earlier on the Ubuntu data than it does on the other two data. This may stem from the different nature of test sets of the three data. The test set of the Ubuntu data is in large size and built by random sampling, while the test sets of the other two data are smaller and constructed through response retrieval.

Impact of context length. Context length is measured by (1) number of turns in a context and (2) average length of utterances in a context. Figure 3 shows how the performance of IoI varies across contexts with different lengths, where we bin test examples of the Ubuntu data into buckets and compare IoI ($L = 7$) with its shallow version ($L = 1$) and DAM. We find that (1) IoI, either in a deep form or in a shallow form, is good at dealing with contexts with long utterances, as the model achieves better performance on longer utterances; (2) overall, IoI performs well on contexts with more turns, although too many turns (e.g., ≥ 8) is still challenging; (3) a deep form of our model is always better than its shallow form, no matter how we measure context length, and the gap between the two forms is bigger on short contexts than it is on long contexts, indicating that depth mainly improves matching accuracy on short contexts; and (4) trends of DAM in both charts are consistent with those reported in (Zhou et al., 2018b), and on both short contexts and long contexts, IoI is superior to DAM.

Ablation study. Finally, we examine how different components of IoI affects its performance. First, we remove $\mathbf{e}^{k-1}_{u,i}$ ($\mathbf{e}^{k-1}_{r,i}$), $\hat{\mathbf{e}}^{k}_{u,i}$ ($\hat{\mathbf{e}}^{k}_{r,i}$), $\overline{\mathbf{e}}^{k}_{u,i}$ ($\overline{\mathbf{e}}^{k}_{r,i}$), and $\tilde{\mathbf{e}}^{k}_{u,i}$ ($\tilde{\mathbf{e}}^{k}_{r,i}$) one by one from Equation (10) and Equation (11), and denote the models as IoI-E, IoI-$\hat{E}$, IoI-$\overline{E}$, and IoI-$\tilde{E}$ respectively. Then, we keep all representations in Equation (10) and Equation (11), and remove $\mathbf{M}^k_{i,1}$, $\mathbf{M}^k_{i,2}$, and $\mathbf{M}^k_{i,3}$ one by one from Equation (13). The models are named IoI-M_1, IoI-M_2, and IoI-M_3 respectively. Table 3 reports the ablation results[5]. We conclude that (1) all representations are useful in representing the information flow along the chain of interaction blocks and capturing the matching information between an utterance-response pair within the blocks, as removing any component gener-

[5] Due to space limitation, we only report results on main metrics.

ally causes performance drop on all the three data sets; and (2) in terms of component importance, $\hat{E} > \overline{E} > E > \tilde{E}$ and $M_2 > M_1 \approx M_3$, meaning that self-attention (i.e., $\hat{E}$) and cross-attention (i.e., $\overline{E}$) are more important than others in information flow representation, and self-attention (i.e., those used for calculating M_2) convey more matching signals. Note that these results are obtained with IoI ($L = 7$). We also check the ablation results of IoI ($L = 1$) and do not see much difference on overall trends and relative gaps among different ablated models.

7 Conclusions and Future Work

We present an interaction-over-interaction network (IoI) that lets utterance-response interaction in context-response matching go deep. Depth of the model comes from stacking multiple interaction blocks that execute representation-interaction-representation in an iterative manner. Evaluation results on three benchmarks indicate that IoI can significantly outperform baseline methods with moderate depth. In the future, we plan to integrate our IoI model with models like ELMo (Peters et al., 2018) and BERT (Devlin et al., 2018) to study if the performance of IoI can be further improved.

Acknowledgement

We would like to thank the anonymous reviewers for their constructive comments. This work was supported by the National Key Research and Development Program of China (No. 2017YFC0804001), the National Science Foundation of China (NSFC Nos. 61672058 and 61876196).

References

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450*.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.

Alexis Conneau, Holger Schwenk, Loïc Barrault, and Yann Lecun. 2017. Very deep convolutional networks for text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1107–1116.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Baotian Hu, Zhengdong Lu, Hang Li, and Qingcai Chen. 2014. Convolutional neural network architectures for matching natural language sentences. In *Advances in Neural Information Processing Systems*, pages 2042–2050.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. 2017a. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708.

Hsin-Yuan Huang, Chenguang Zhu, Yelong Shen, and Weizhu Chen. 2017b. FusionNet: Fusing via fully-aware attention with application to machine comprehension. In *International Conference on Learning Representations*.

Rudolf Kadlec, Martin Schmid, and Jan Kleindienst. 2015. Improved deep learning baselines for ubuntu corpus dialogs. *arXiv preprint arXiv:1510.03753*.

Seonhoon Kim, Jin-Hyuk Hong, Inho Kang, and Nojun Kwak. 2018. Semantic sentence matching with densely-connected recurrent and co-attentive information. *arXiv preprint arXiv:1805.11360*.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105.

Feng-Lin Li, Minghui Qiu, Haiqing Chen, Xiongwei Wang, Xing Gao, Jun Huang, Juwei Ren, Zhongzhou Zhao, Weipeng Zhao, Lei Wang, et al. 2017a. AliMe assist: An intelligent assistant for creating an innovative e-commerce experience. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 2495–2498.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2015. A diversity-promoting objective function for neural conversation models. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119.

Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and Bill Dolan. 2016a. A persona-based neural conversation model. In *Association for Computational Linguistics*, pages 994–1003.

Jiwei Li, Will Monroe, Alan Ritter, Dan Jurafsky, Michel Galley, and Jianfeng Gao. 2016b. Deep reinforcement learning for dialogue generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202.

Jiwei Li, Will Monroe, Tianlin Shi, Sèbastien Jean, Alan Ritter, and Dan Jurafsky. 2017b. Adversarial learning for neural dialogue generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2157–2169.

Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. 2015. The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 285–294.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Lili Mou, Yiping Song, Rui Yan, Ge Li, Lu Zhang, and Zhi Jin. 2016. Sequence to backward and forward sequences: A content-introducing approach to generative short-text conversation. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3349–3358.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237.

Iulian Vlad Serban, Alessandro Sordoni, Yoshua Bengio, Aaron C. Courville, and Joelle Pineau. 2016. End-to-end dialogue systems using generative hierarchical neural network models. In *AAAI*, pages 3776–3784.

Lifeng Shang, Zhengdong Lu, and Hang Li. 2015. Neural responding machine for short-text conversation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 1577–1586.

Heung-Yeung Shum, Xiaodong He, and Di Li. 2018. From Eliza to XiaoIce: Challenges and opportunities with social chatbots. *Frontiers of IT & EE*, 19(1):10–26.

Yiping Song, Rui Yan, Cheng-Te Li, Jian-Yun Nie, Ming Zhang, and Dongyan Zhao. 2018. An ensemble of retrieval-based and generation-based human-computer conversation systems. In *IJCAI*, pages 4382–4388.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9.

Chongyang Tao, Shen Gao, Mingyue Shang, Wei Wu, Dongyan Zhao, and Rui Yan. 2018. Get the point of my utterance! learning towards effective responses with multi-head attention mechanism. In *IJCAI*, pages 4418–4424.

Yi Tay, Luu Anh Tuan, and Siu Cheung Hui. 2018. Co-stack residual affinity networks with multi-level attention refinement for matching text sequences. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4492–4502.

Oriol Vinyals and Quoc Le. 2015. A neural conversational model. *arXiv preprint arXiv:1506.05869*.

Shengxian Wan, Yanyan Lan, Jun Xu, Jiafeng Guo, Liang Pang, and Xueqi Cheng. 2016. Match-srnn: Modeling the recursive matching structure with spatial rnn. In *IJCAI*, pages 2922–2928.

Hao Wang, Zhengdong Lu, Hang Li, and Enhong Chen. 2013. A dataset for research on short-text conversations. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 935–945.

Mingxuan Wang, Zhengdong Lu, Hang Li, and Qun Liu. 2015. Syntax-based deep matching of short texts. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence*, pages 1354–1361.

Shuohang Wang and Jing Jiang. 2016. Learning natural language inference with LSTM. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1442–1451.

Yu Wu, Wei Wu, Zhoujun Li, and Ming Zhou. 2018a. Learning matching models with weak supervision for response selection in retrieval-based chatbots. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 420–425.

Yu Wu, Wei Wu, Chen Xing, Can Xu, Zhoujun Li, and Ming Zhou. 2018b. A sequential matching framework for multi-turn response selection in retrieval-based chatbots. *Computational Linguistics*, 45(1):163–197.

Yu Wu, Wei Wu, Chen Xing, Ming Zhou, and Zhoujun Li. 2017. Sequential matching network: A new architecture for multi-turn response selection in retrieval-based chatbots. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 496–505.

Chen Xing, Wei Wu, Yu Wu, Jie Liu, Yalou Huang, Ming Zhou, and Wei-Ying Ma. 2017. Topic aware neural response generation. In *AAAI*, pages 3351–3357.

Zhen Xu, Bingquan Liu, Baoxun Wang, Chengjie Sun, and Xiaolong Wang. 2017. Incorporating loose-structured knowledge into LSTM with recall gate for conversation modeling. In *Proceedings of the 2017 International Joint Conference on Neural Networks*, pages 3506–3513.

Rui Yan, Yiping Song, and Hua Wu. 2016. Learning to respond with deep neural networks for retrieval-based human-computer conversation system. In *SIGIR*, pages 55–64.

Rui Yan and Dongyan Zhao. 2018. Coupled context modeling for deep chit-chat: towards conversations between human and computer. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2574–2583. ACM.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018a. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213.

Zhuosheng Zhang, Jiangtong Li, Pengfei Zhu, Hai Zhao, and Gongshen Liu. 2018b. Modeling multi-turn conversation with deep utterance aggregation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3740–3752. Association for Computational Linguistics.

Tiancheng Zhao, Ran Zhao, and Maxine Eskenazi. 2017. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 654–664.

Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. 2018a. Emotional chatting machine: Emotional conversation generation with internal and external memory. In *The Thirty-Second AAAI Conference on Artificial Intelligence*, pages 730–738.

Xiangyang Zhou, Daxiang Dong, Hua Wu, Shiqi Zhao, Dianhai Yu, Hao Tian, Xuan Liu, and Rui Yan. 2016. Multi-view response selection for human-computer conversation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 372–381.

Xiangyang Zhou, Lu Li, Daxiang Dong, Yi Liu, Ying Chen, Wayne Xin Zhao, Dianhai Yu, and Hua Wu. 2018b. Multi-turn response selection for chatbots with deep attention matching network. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1118–1127.