**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∗ 1Institute of Computer Science and Technology, Peking University, Beijing, China 2Microsoft Corporation, Beijing, China

3Center 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 at- tention to retrieval-based dialogues in open- domain. In particular, people study the prob- lem by investigating context-response match- ing 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 per- formed 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 match- ing 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 re- sults on three benchmark data sets indicate that IoI can significantly outperform state-of-the- art methods in terms of various matching met- rics. Through further analysis, we also unveil how the depth of interaction affects the perfor- mance of IoI.

# Introduction

Building a chitchat style dialogue systems in open- domain for human-machine conversations has at- tracted increasing attention in the conversational artificial intelligence (AI) community. Generally speaking, there are two approaches to implement- ing such a conversational system. The first ap- proach leverages techniques of information re- trieval ([Lowe et al.](#_bookmark41), [2015](#_bookmark41); [Wu et al.](#_bookmark59), [2017](#_bookmark59); [Yan](#_bookmark63) [and Zhao](#_bookmark63), [2018](#_bookmark63)), and selects a proper response from an index; while the second approach di- rectly synthesizes a response with a natural lan-

∗Corresponding author: Rui Yan (ruiyan@pku.edu.cn).

1

guage generation model estimated from a large- scale conversation corpus ([Serban et al.](#_bookmark45), [2016](#_bookmark45); [Li](#_bookmark40) [et al.](#_bookmark40), [2017b](#_bookmark40)). In this work, we study the prob- lem of multi-turn response selection for retrieval- based dialogue systems where the input is a con- versation context consisting of a sequence of utter- ances. 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.](#_bookmark47), [2018](#_bookmark47)) from Mi- crosoft, and the e-commerce assistant AliMe As- sist from Alibaba Group ([Li et al.](#_bookmark36), [2017a](#_bookmark36)).

A key step in multi-turn response selection is to measure the matching degree between a con- versation context and a response candidate. State- of-the-art methods ([Wu et al.](#_bookmark59), [2017](#_bookmark59); [Zhou et al.](#_bookmark69), [2018b](#_bookmark69)) perform matching within a representation- interaction-aggregation framework ([Wu et al.](#_bookmark58), [2018b](#_bookmark58)) where matching signals in each utterance- response pair are distilled from their interaction based on their representations, and then are ag- gregated as a matching score. Although utterance- response interaction has proven to be crucial to the performance of the matching models ([Wu et al.](#_bookmark59), [2017](#_bookmark59)), it is executed in a rather shallow manner where matching between an utterance and a re- sponse candidate is determined only by one step of interaction on each type or each layer of rep- resentations. In this paper, we attempt to move from shallow interaction to deep interaction, and consider context-response matching with multi- ple steps of interaction where residual information from one time of interaction, which is generally ignored by existing methods, is leveraged for ad- ditional interactions. The underlying motivation is that if a model extracts some matching informa- tion from utterance-response pairs in one step of interaction, then by stacking multiple such steps, the model can gradually accumulate useful signals

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

for matching and finally capture the semantic rela- tionship between a context and a response candi- date in a more comprehensive way.

We propose an interaction-over-interaction net- work (IoI) for context-response matching, through which we aim to investigate: (1) how to make in- teraction 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 in- teraction are concatenated with the self-attention representations and then compressed to new rep- resentations of the utterance-response pair as the output of the block. Built on top of the interac- tion 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 oper- ations and let the utterance and the response inter- act with each other in an iterative way. Different blocks could distill different levels of matching in- formation in an utterance-response pair. To suffi- ciently leverage the information, a matching score is first calculated in each block through aggre- gating matching vectors of all utterance-response pairs, and then the block-wise matching scores are combined as the final matching degree of the con- text and the response candidate.

We conduct experiments on three benchmark data sets: the Ubuntu Dialogue Corpus ([Lowe](#_bookmark41) [et al.](#_bookmark41), [2015](#_bookmark41)), the Douban Conversation Corpus ([Wu et al.](#_bookmark59), [2017](#_bookmark59)), and the E-commerce Dialogue Corpus ([Zhang et al.](#_bookmark65), [2018b](#_bookmark65)). 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 R10@1 on the Ubuntu data, 2*.*3% absolute im- provement on MAP on the Douban data, and 3*.*7% absolute improvement on R10@1 on the E- commerce data. Through more quantitative anal- ysis, we also show that depth indeed brings im- provement 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 R10@1 between IoI and DAM is only 1*.*1%. Moreover, the improve- ment brought by depth mainly comes from short contexts.

Our contributions in this paper are three-folds:

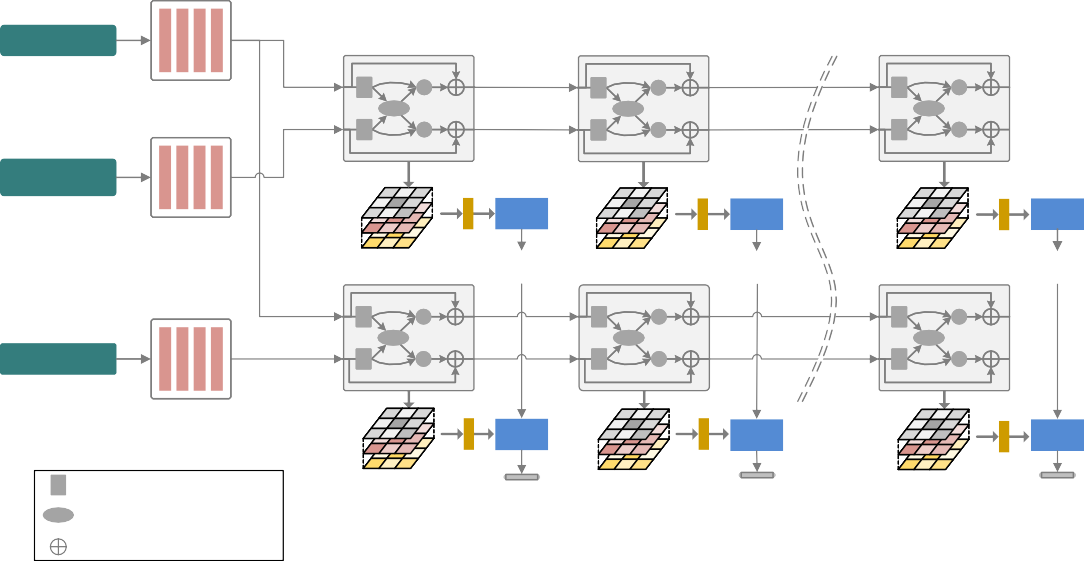
(1) proposal of a novel interaction-over-interaction network which enables deep-level matching with carefully designed interaction block chains; (2) empirical verification of the effectiveness of the model on three benchmarks; and (3) empiri- cal study on the relationship between interaction depth and model performance.

# Related Work

Existing methods for building an open-domain di- alogue system can be categorized into two groups. The first group learns response generation mod- els under an encoder-decoder framework. On top of the basic sequence-to-sequence with attention architecture ([Vinyals and Le](#_bookmark52), [2015](#_bookmark52); [Shang et al.](#_bookmark46), [2015](#_bookmark46); [Tao et al.](#_bookmark50), [2018](#_bookmark50)), various extensions have been made to tackle the “safe response” problem ([Li et al.](#_bookmark37), [2015](#_bookmark37); [Mou et al.](#_bookmark43), [2016](#_bookmark43); [Xing et al.](#_bookmark60), [2017](#_bookmark60); [Zhao et al.](#_bookmark66), [2017](#_bookmark66); [Song et al.](#_bookmark48), [2018](#_bookmark48)); to gener- ate responses with specific personas or emotions ([Li et al.](#_bookmark38), [2016a](#_bookmark38); [Zhang et al.](#_bookmark64), [2018a](#_bookmark64); [Zhou et al.](#_bookmark67), [2018a](#_bookmark67)); and to pursue better optimization strate- gies ([Li et al.](#_bookmark40), [2017b](#_bookmark40), [2016b](#_bookmark39)).

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 se- lection by setting the human input as a single message ([Wang et al.](#_bookmark54), [2013](#_bookmark54); [Hu et al.](#_bookmark29), [2014](#_bookmark29); [Wang et al.](#_bookmark55), [2015](#_bookmark55)), and moves to context-response matching for multi-turn response selection re- cently. Representative methods include the dual LSTM model ([Lowe et al.](#_bookmark41), [2015](#_bookmark41)), the deep learn- ing to respond architecture ([Yan et al.](#_bookmark62), [2016](#_bookmark62)), the multi-view matching model ([Zhou et al.](#_bookmark68), [2016](#_bookmark68)), the sequential matching network ([Wu et al.](#_bookmark59), [2017](#_bookmark59), [2018b](#_bookmark58)), and the deep attention matching net- work ([Zhou et al.](#_bookmark69), [2018b](#_bookmark69)). Besides model design, some attention is also paid to the learning prob- lem of matching models ([Wu et al.](#_bookmark57), [2018a](#_bookmark57)). 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.



**Response**

**Interaction Block 1**

**Interaction Block 2**

**Interaction Block L**

**Utterance-1**

**GRU**

**GRU**

**GRU**

*v11 v12 v1L*

*T11 T12 T1L*

**Utterance-n**

**Initial Representation**

**GRU**

**GRU**

**GRU**

*vn1 vn2 vnL*

**: Self-attention**

**: Interaction Operation**

**: Add Operation**

*Tn1*

*Tn2*

*TnL*

g(c,r)

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

**...**

**...**

**...**

**...**

Encouraged by the big success of deep neural architectures such as Resnet ([He et al.](#_bookmark28), [2016](#_bookmark28)) and inception ([Szegedy et al.](#_bookmark49), [2015](#_bookmark49)) in computer vi- sion, 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 vi- sion, they have proven effective in a few tasks such as text classification ([Conneau et al.](#_bookmark26), [2017](#_bookmark26)), natu- ral language inference ([Kim et al.](#_bookmark33), [2018](#_bookmark33); [Tay et al.](#_bookmark51), [2018](#_bookmark51)), and question answering ([Tay et al.](#_bookmark51), [2018](#_bookmark51); [Kim et al.](#_bookmark33), [2018](#_bookmark33)), etc. In this work, we attempt to improve the accuracy of multi-turn response se- lection in retrieval-based dialogue systems by in- creasing the depth of context-response interaction in matching. Through extensive studies on bench- marks, we show that depth can bring significant improvement to model performance on the task.

# Problem Formalization

Suppose that there is a conversation data set D = {(*yi, ci, ri*)}*N* . ∀*i* ∈ {1*, . . . , N* }, *ci* = *ui,*1*, . . . , ui,li* represents a conversation context with *ui,k* the *k*-th turn, *ri* is a response candidate, and *yi* 0*,* 1 denotes a label with *yi* = 1 indicating *ri* a proper response for *ci*, otherwise *yi* = 0. The task is to learn a matching model *g*( *,* ) from , and thus for a new context-response pair (*c, r*), *g*(*c, r*) measures the matching degree

*i*=1

{ }

∈ { }

* · D

between *c* and *r*.

In the following sections, we will elaborate how to define *g*( *,* ) to achieve deep interaction be- tween *c* and *r*, and how to learn such a deep model from D.

· ·

# Interaction-over-Interaction Network

We define *g*( *,* ) as an interaction-over-interaction network (IoI). Figure [1](#_bookmark1) illustrates the architecture of IoI. The model pairs each utterance in a con- text with a response candidate, and then aggre- gates matching information from all the pairs as a matching score of the context and the response candidate. For each pair, IoI starts from initial rep- resentations of the utterance and the response, and then feeds the pair to stacked interaction blocks. Each block represents the utterance and the re- sponse by letting them interact with each other based on the interactions before. Matching signals are first accumulated along the sequence of the ut- terances in each block, and then combined along the chain of blocks as the final matching score. Be- low we will describe details of components of IoI and how to learn the model with D.

· ·

## Initial Representations

Given an utterance *u* in a context *c* and a re- sponse candidate *r*, *u* and *r* are initialized as **Eu** = [**e***u,*1*, ,* **e***u,m*] and **Er** = [**e***r,*1*, ,* **e***r,n*] respec-

· · · · · ·

tively. *i* 1*, . . . , m* and *j* 1*, . . . , n* ,

∀ ∈ { } ∀ ∈ { }

**e***u,i* and **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](#_bookmark42) [et al.](#_bookmark42), [2013](#_bookmark42)) on . **Eu** and **Er** are then processed by stacked interaction blocks that model different levels of interaction between *u* and *r* and generate matching signals.

D

## Interaction Block

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 re- sponse, and a compression module that condenses the results of the first two modules into representa- tions 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

round of residual connection and layer normaliza- tion. For ease of presentation, we denote the entire attention mechanism as *fAT T* (**Q***,* **K**).

Let **U***k*−1 and **R***k*−1 be the input of the *k*-th

block where **U**0 = **Eu** and **R**0 = **Er**, then the self-attention module is defined as

**U**ˆ *k* = *f*ATT(**U***k*−1*,* **U***k*−1)*,* (4)

**R**ˆ *k* = *f*ATT(**R***k*−1*,* **R***k*−1)*.* (5)

The interaction module first lets **U***k*−1 and **R***k*−1

attend to each other by

**U***k* = *f*ATT(**U***k*−1*,* **R***k*−1)*,* (6)

**R***k* = *f*ATT(**R***k*−1*,* **U***k*−1)*.* (7)

Then **U***k*−1 and **R***k*−1 further interact with **U***k*

and **R***k* respectively, which can be formulated as

**U**˜ *k* = **U***k*−1 0 *k*

lays a foundation for the self-attention module and the interaction module. Let **Q** ∈ R*nq* ×*d* and **K** ∈ R*nk* ×*d* be a query and a key respectively,

**R**˜ *k* = **R***k*−1 0

**U** *,* (8)

**R***k,* (9)

where *nq* and *nk* denote numbers of words and *d*

is the embedding size, then attention from **Q** to **K** is defined as

where denotes element-wise multiplication. Fi-

nally, the compression module updates **U***k*−1 and

0

**R***k*−1 to **U***k* and **R***k* as the output of the block.

Suppose that **e***k*

and **e***k*

are the *i*-th entries of

**Q**ˆ = *S*(**Q***,* **K**) · **K***,* (1)

and

*u,i*

*r,i*

and

*u,i*

are cal-

**e**

**U***k* **R***k* respectively, then **e***k*

*k*

*r,i*

where *S*( *,* ) is a function for attention weight cal-

· ·

culation. Here, we exploit the symmetric function

in ([Huang et al.](#_bookmark31), [2017b](#_bookmark31)) as *S*(·*,* ·) which is given

culated by

**e***k*−1

*u,i*

by:

*k*  **e**ˆ*k* 

**e**˜

*k*−1

**e***u,i* = ReLU(**w***p*  *u,i*  + **b***p*) + **e**

 **e**

*S*(**Q***,* **K**) = softmax(*f* (**QW**)**D***f* (**KW**)T)*.* (2)

*k u,i k u,i*



*u,i*

 

*,*(10)

In Equation ([2](#_bookmark7)), *f* is a ReLU activation function,

**D** is a diagonal matrix, and both **D** ∈ R*d*×*d* and *k*

∈

**e***k*−1

 **e**ˆ*k* 

*r,i*

**e**˜

*k*−1

**e***r,i* = ReLU(**w***p*  *r,i*  + **b***p*) + **e**

 **e**



*,*(11)

**W** R*d*×*d* are parameters to estimate from train-

ing data. Intuitively, in Equation ([1](#_bookmark6)), each entry of

**K** is weighted by an importance score defined by

*k*

*r,i*

*k*

*r,i*

*r,i*

the similarity of an entry of **Q** and an entry of **K**.

∈

where **w***p* ∈ R4*d*×*d* and **b***p* are learnable projec-

The entries of **K** are then linearly combined with

tion weights and biases, **e**ˆ*k*

{*u,r*}*,i*

, **e***k*

{*u,r*}*,i*

, **e**˜*k* ,

{*u,r*}*,i*

the weights to form a new representation of **Q**.

A residual connection ([He et al.](#_bookmark28), [2016](#_bookmark28)) and a

and **e***k*−1

{*u,r*}*,i*

{ } { } { }

are the *i*-th entries of {**U**ˆ *,* **R**ˆ }*k*,

layer normalization ([Ba et al.](#_bookmark24), [2016](#_bookmark24)) are then ap-

plied to **Q**ˆ as **Q**˜ . After that, **Q**˜ is fed to a feed

forward network which is formulated as ReLU(**Q˜W**1 + **b**1)**W**2 + **b**2*,* (3)

where **W**{1*,*2} R*d*×*d* and **b**{1*,*2} are parame-

ters. The output of the attention mechanism is de-

fined with the result of Equation ([3](#_bookmark10)) after another

**U***,* **R** *k*, **U**˜ *,* **R**˜ *k*, and **U***,* **R** *k*−1, respectively.

Inspired by [Huang et al.](#_bookmark30) ([2017a](#_bookmark30)), we also intro- duce direct connections from initial representa- tions to all their corresponding subsequent blocks.

## Matching Aggregation

Suppose that *c* = (*u*1*, . . . , ul*) is a conversation context with *ui* the *i*-th utterance, then in the *k*- th interaction block, we construct three similarity

matrices by

*k*

**U***k*−1 · (**R***k*−1)T

# Learning Methods

We consider two strategies to learn an IoI model

**M***i,*1 = *i ,*

√

*d*

from the training data D. The first strategy es-

**M***k* =

**U**ˆ *k* · (**R**ˆ *k*)T

(12)

timates the parameters of IoI (denoted as Θ) by

minimizing a global loss function that is formu-

*i,*2 *d*

*i* √ *,*

lated as

**U***k* · *k* T

*k*

*i*

(**R**

)

*N*

**M***i,*3 =

√*d ,*

− 区［*y*

log(*g*(*c , r* ))+(1−*y* ) log(1−*g*(*c , r* ))－*.*

where **U***k*−1 and **R***k*−1 are the input of the *k*-th

*i*

*i i i i*

*i*=1

*i i*

(17)

block, **U**ˆ *k* and **R**ˆ *k* are defined by Equations ([4](#_bookmark2)-[5](#_bookmark3)),

*i*

*i*

and **U***k* and **R***k* are calculated by Equations ([6](#_bookmark4)-[7](#_bookmark5)). The three matrices are then concatenated into a 3- D matching tensor **T***k* R*mi*×*n*×3 which can be written as

**T***k* = **M***k* ⊕ **M** ⊕ **M** *,* (13)

*i*

∈

*k*

*k*

⊕

*i*

*i,*1

*i,*2

*i,*3

In the second strategy, we construct a local loss function for each block and minimize the summa- tion of the local loss functions. By this means, each block can be directly supervised by the la- bels in during learning. The learning objective is then defined as

*L*

*N*

where denotes a concatenation operation, and *mi* and *n* refer to numbers of words in *ui* and *r* respectively.

− 区区［*yi* log(*gk*(*ci, ri*))

+ (1 − *yi*) log(1 − *gk*(*ci, ri*))－*.*

*k*=1 *i*=1

D

(18)

We exploit a convolutional neural net-

work ([Krizhevsky et al.](#_bookmark35), [2012](#_bookmark35)) to extract matching features from **T***k*. The output of the final feature maps are flattened and mapped to a *d*-dimensional matching vector **v***k* with a linear transformation.

*i*

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](#_bookmark34)

*k k i* [and Ba](#_bookmark34), [2015](#_bookmark34)).

(**v**1 *, ,* **v***l* ) is then fed to a GRU ([Chung et al.](#_bookmark25), [2014](#_bookmark25)) to capture temporal relationship among

· · ·

(*u*1*, . . . , ul*). *i* 1*, . . . , l* , the *i*-th hidden state of the GRU model is given by

∀ ∈ { }

**h***k* = GRU(**v***k,* **h***k* )*,* (14)

*i i i*−1

where **h***k* is randomly initialized. A matching score for context *c* and response candidate *r* in the *k*-th block is defined as

0

*gk*(*c, r*) = *σ*(**h***k* · **w***o* + **b***o*)*,* (15)

*l*

where **w***o* and **b***o* are parameters, and *σ*( ) is a sig- moid function. Finally, *g*(*c, r*) is defined by

·

*L*

区

*g*(*c, r*) = *gk*(*c, r*)*,* (16)

*k*=1

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 mo- tivated by (1) only using the last block will make training of IoI difficult due to the gradient van- ishing/exploding problem; and (2) different blocks may capture different levels of matching informa- tion in (*c, r*), and thus leveraging all of them could enhance matching accuracy.

# Experiments

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

## Experimental Setup

The first data we use is the Ubuntu Dialogue Cor- pus ([Lowe et al.](#_bookmark41), [2015](#_bookmark41)) which is a multi-turn En- glish conversation data set constructed from chat logs of the Ubuntu forum. We use the version provided by [Xu et al.](#_bookmark61) ([2017](#_bookmark61)). 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.](#_bookmark41) ([2015](#_bookmark41)), we em- ploy recall at position *k* in *n* candidates (R*n*@k) as evaluation metrics.

The second data set is the Douban Conversation Corpus ([Wu et al.](#_bookmark59), [2017](#_bookmark59)) that consists of multi- turn Chinese conversations collected from Douban group[1](#_bookmark14). 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 re- sponse is randomly sampled. For each context in the test set, 10 response candidates are retrieved from an index and their appropriateness regard- ing to the context is annotated by human labelers. Following [Wu et al.](#_bookmark59) ([2017](#_bookmark59)), we employ R*n*@ks, mean average precision (MAP), mean reciprocal rank (MRR) and precision at position 1 (P@1) as evaluation metrics.

Finally, we choose the E-commerce Dialogue Corpus ([Zhang et al.](#_bookmark65), [2018b](#_bookmark65)) as an experimen- tal data set. The data consists of multi-turn real- world conversations between customers and cus- tomer service staff in Taobao[2](#_bookmark15), which is the largest e-commerce platform in China. It contains 1 mil- lion context-response pairs for training, and 10 thousand pairs for validation and test. Positive re- sponses in this data are real human responses, and negative candidates are automatically constructed by ranking the response corpus based on conver- sation history augmented messages using Apache Lucene[3](#_bookmark16). The ratio of the positive and the neg- ative is 1:1 in training and validation, and 1:9 in test. Following ([Zhang et al.](#_bookmark65), [2018b](#_bookmark65)), we employ R10@1, R10@2, and R10@5 as evaluation metrics.

## Baselines

We compare IoI with the following models:

**Single-turn Matching Models:** these models, including RNN ([Lowe et al.](#_bookmark41), [2015](#_bookmark41)), CNN ([Lowe](#_bookmark41) [et al.](#_bookmark41), [2015](#_bookmark41)), LSTM ([Lowe et al.](#_bookmark41), [2015](#_bookmark41)), BiL- STM ([Kadlec et al.](#_bookmark32), [2015](#_bookmark32)), MV-LSTM ([Wan et al.](#_bookmark53), [2016](#_bookmark53)) and Match-LSTM ([Wang and Jiang](#_bookmark56), [2016](#_bookmark56)), perform context-response matching by concate- nating all utterances in a context into a single long document and calculating a matching score be- tween the document and a response candidate.

**Multi-View** ([Zhou et al.](#_bookmark68), [2016](#_bookmark68)): the model cal- culates matching degree between a context and a response candidate from both a word sequence view and an utterance sequence view.

**DL2R** ([Yan et al.](#_bookmark62), [2016](#_bookmark62)): the model first refor- mulates 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.](#_bookmark59), [2017](#_bookmark59)): the model lets each ut- terance in a context interact with a response can- didate at the beginning, and then transforms inter- action matrices into a matching vector with CNN. The matching vectors are finally accumulated with an RNN as a matching score.

**DUA** ([Zhang et al.](#_bookmark65), [2018b](#_bookmark65)): the model considers the relationship among utterances within a context by exploiting deep utterance aggregation to form a fine-grained context representation. Each re- fined utterance then matches with a response can- didate, and their matching degree is finally calcu- lated through an aggregation on turns.

**DAM** ([Zhou et al.](#_bookmark69), [2018b](#_bookmark69)): the model lets each utterance in a context interact with a response can- didate at different levels of representations ob- tained 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 avail- able in [Zhou et al.](#_bookmark69) ([2018b](#_bookmark69)), we directly copy the numbers from the paper. For the E-commerce data, [Zhang et al.](#_bookmark65) ([2018b](#_bookmark65)) report performance of all baselines except DAM. Thus, we copy all avail- able numbers from the paper and implement DAM with the published code[4](#_bookmark17). In order to conduct sta- tistical tests, we also run the code of DAM on the Ubuntu data and the Douban data.

## 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 (3*,* 3), and the strides as (1*,* 1) and (3*,* 3) respec- tively. The number of convolution kernels is 32 in the first layer and 16 in the second layer. The di- mension of the hidden states of GRU is set as 200. Following [Wu et al.](#_bookmark59) ([2017](#_bookmark59)), we limit the length of a context to 10 turns and the length of an utterance (either from a context or from a response candi- date) 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](https://www.taobao.com/)

3<http://lucene.apache.org/> 4 <https://github.com/baidu/Dialogue>

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Metrics  Models | **Ubuntu Corpus** | | | | **Douban Corpus** | | | | | |
| R2@1 | R10@1 | R10@2 | R10@5 | MAP | MRR | P@1 | R10@1 | R10@2 | R10@5 |
| RNN ([Lowe et al., 2015](#_bookmark41)) | 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](#_bookmark41)) | 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](#_bookmark41)) | 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](#_bookmark32)) | 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](#_bookmark62)) | 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](#_bookmark53)) | 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](#_bookmark56)) | 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](#_bookmark68)) | 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](#_bookmark59)) | 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](#_bookmark65)) | - | 0.752 | 0.868 | 0.962 | 0.551 | 0.599 | 0.421 | 0.243 | 0.421 | 0.780 |
| DAM ([Zhou et al., 2018b](#_bookmark69)) | 0.938 | 0.767 | 0.874 | 0.969 | 0.550 | 0.601 | 0.427 | 0.254 | 0.410 | 0.757 |
| IoI-global  IoI-local | **0.941**  **0.947** | **0.778**  **0.796** | **0.879**  **0.894** | 0.970  **0.974** | **0.566**  **0.573** | 0.608  **0.621** | 0.433  0.444 | 0.263  **0.269** | **0.436**  **0.451** | **0.781**  **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).

0.80

78

R10@1

|  |  |  |  |
| --- | --- | --- | --- |
| Metrics  Models | R10@1 | R10@2 | R10@5 |
| RNN ([Lowe et al., 2015](#_bookmark41)) | 0.325 | 0.463 | 0.775 |
| CNN ([Lowe et al., 2015](#_bookmark41)) | 0.328 | 0.515 | 0.792 |
| LSTM ([Lowe et al., 2015](#_bookmark41)) | 0.365 | 0.536 | 0.828 |
| BiLSTM ([Kadlec et al., 2015](#_bookmark32)) | 0.355 | 0.525 | 0.825 |
| DL2R ([Yan et al., 2016](#_bookmark62)) | 0.399 | 0.571 | 0.842 |
| MV-LSTM ([Wan et al., 2016](#_bookmark53)) | 0.412 | 0.591 | 0.857 |
| Match-LSTM ([Wang and Jiang, 2016](#_bookmark56)) | 0.410 | 0.590 | 0.858 |
| Multi-View ([Zhou et al., 2016](#_bookmark68)) | 0.421 | 0.601 | 0.861 |
| SMN ([Wu et al., 2017](#_bookmark59)) | 0.453 | 0.654 | 0.886 |
| DUA([Zhang et al., 2018b](#_bookmark65)) | 0.501 | 0.700 | 0.921 |
| DAM ([Zhou et al., 2018b](#_bookmark69)) | 0.526 | 0.727 | 0.933 |
| IoI-global  IoI-local | **0.554**  **0.563** | 0.747  **0.768** | 0.942  **0.950** |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 0.7 | 0.7  89 | 93 0.7 | 94 0.7 | 95 0.7 | 95 0.7 | 96  0.7 | 94 |
| 0.7 |  |  |  |  |  | U  E-  D | buntu Commerc ouban | e |

0.79

0.78

0.55

R10@1

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  | 0.5 | 54 0.5 | 63 0.5 | 63 0.5 | 61 |
|  | 0.5 | 0.5  16 | 28 0.5 | 37 |  |  |  |  |
| 0.4 | 67 |  |  |  |  |  |  |  |

0.50

0.45

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  | 0.4 | 41 0.4 | 0.4  40 | 44 0.4 | 41 |
|  | 0.4 | 0.4  21 | 30 0.4 | 32 |  |  |  |  |
| 0.4 | 02 |  |  |  |  |  |  |  |

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).

0.44

0.42

P@1

0.40

1 2 3 4 5 6 7 8

# Interaction Blocks

during training.

## Evaluation Results

Table [1](#_bookmark18) and Table [2](#_bookmark19) report evaluation results on the three data sets where IoI-global and IoI-local rep- resent models learned with Objective ([17](#_bookmark11)) and Ob- jective ([18](#_bookmark13)) respectively. We can see that both IoI- local and IoI-global outperform the best perform- ing baseline, and improvements from IoI-local on all metrics and from IoI-global on a few met- rics 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 struc- ture than optimizing the final matching model.

## Discussions

In this section, we make some further analysis with IoI-local to understand (1) how depth of in-

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

teraction affects the performance of IoI; (2) how context length affects the performance of IoI; and

1. importance of different components of IoI with respect to matching accuracy.

**Impact of interaction depth.** Figure [2](#_bookmark20) illus- trates how the performance of IoI changes with re- spect to the number of interaction blocks on test sets of the three data. From the chart, we ob- serve 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 match- ing accuracy. With shallow interaction (*L* = 1), IoI performs worse than DAM on the Douban data and the E-commerce data. Only after the interac- tion goes deep (*L* ≥ 5), improvement from IoI

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Metrics  Models | **Ubuntu data** | | | **Douban data** | | | **E-commerce data** | | |
| R2@1 | R10@1 | R10@2 | MAP | MRR | P@1 | R10@1 | R10@2 | R10@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-*E*ˆ | 0.946 | 0.790 | 0.888 | 0.565 | 0.613 | 0.433 | 0.557 | 0.749 | 0.941 |
| IoI-*E* | 0.947 | 0.793 | 0.890 | 0.566 | 0.613 | 0.439 | 0.560 | 0.754 | 0.943 |
| IoI-*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.

0.850

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  | DAM  IoI-1 | L |
|  |  |  | IoI-7 | L |

0.825

0.800

R10@1

0.775

0.750

0.725

(0, 10] (10, 20] (20, 30] (30, 50]

Average utterance length (words)

* + 1. R10@1 vs. Average utterance length

0.81

0.80

IoI-1

IoI-7

L L

DAM

0.79

R10@1

0.78

0.77

0.76

0.75

[2, 4] [5, 7] [8, 10]

Context length (turns)

* + 1. R10@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 perfor- mance 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 con- structed through response retrieval.

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 im- proves matching accuracy on short contexts; and

1. trends of DAM in both charts are consistent with those reported in ([Zhou et al.](#_bookmark69), [2018b](#_bookmark69)), and on both short contexts and long contexts, IoI is supe- rior to DAM.

**Ablation study.** Finally, we examine how dif- ferent components of IoI affects its performance. First, we remove **e***k*−1 (**e***k*−1), **e**ˆ*k* (**e**ˆ*k* ), **e***k*

*u,i*

*r,i*

*u,i*

*r,i*

*u,i*

**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. Fig- ure [3](#_bookmark22) shows how the performance of IoI varies

(**e***k* ), and **e**˜*k* (**e**˜*k* ) one by one from Equation

([10](#_bookmark8)) and Equation ([11](#_bookmark9)), and denote the models as IoI-*E*, IoI-*E*ˆ, IoI-*E*, and IoI-*E*˜ respectively. Then,

*r,i*

*u,i*

*r,i*

we keep all representations in Equation ([10](#_bookmark8)) and

*i,*1

*i,*3

across contexts with different lengths, where we

Equation ([11](#_bookmark9)), and remove **M***k*

*k*

*i,*2

, **M**

, and **M***k*

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

≥

one by one from Equation ([13](#_bookmark12)). The models are named IoI-*M*1, IoI-*M*2, and IoI-*M*3 respectively. Table [3](#_bookmark21) reports the ablation results[5](#_bookmark23). We conclude that (1) all representations are useful in represent- ing the information flow along the chain of inter- action blocks and capturing the matching infor- mation between an utterance-response pair within the blocks, as removing any component gener-

5Due 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,

*E*ˆ *> E > E > E*˜ and *M*2 *> M*1 *M*3, meaning

≈

that self-attention (i.e., *E*ˆ) and cross-attention (i.e.,

*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 differ- ent ablated models.

# Conclusions and Future Work

We present an interaction-over-interaction net- work (IoI) that lets utterance-response inter- action in context-response matching go deep. Depth of the model comes from stacking multi- ple 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.](#_bookmark44), [2018](#_bookmark44)) and BERT ([Devlin](#_bookmark27) [et al.](#_bookmark27), [2018](#_bookmark27)) to study if the performance of IoI can be further improved.

# Acknowledgement

We would like to thank the anonymous re- viewers for their constructive comments. This work was supported by the National Key Re- search and Development Program of China (No. 2017YFC0804001), the National Science Foun- dation of China (NSFC Nos. 61672058 and 61876196).

# References

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hin- ton. 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 model- ing. *arXiv preprint arXiv:1412.3555*.

Alexis Conneau, Holger Schwenk, Lo¨ıc Barrault, and Yann Lecun. 2017. [Very deep convolutional net-](https://www.aclweb.org/anthology/E17-1104) [works for text classification](https://www.aclweb.org/anthology/E17-1104). In *Proceedings of the 15th Conference of the European Chapter of the As- sociation 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 under- standing. *arXiv preprint arXiv:1810.04805*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recog- nition. 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 archi- tectures for matching natural language sentences. In *Advances in Neural Information Processing Sys- tems*, 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 recogni- tion*, pages 4700–4708.

Hsin-Yuan Huang, Chenguang Zhu, Yelong Shen, and Weizhu Chen. 2017b. FusionNet: Fusing via fully- aware attention with application to machine compre- hension. 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 No- jun Kwak. 2018. Semantic sentence matching with densely-connected recurrent and co-attentive infor- mation. *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 Hin- ton. 2012. Imagenet classification with deep con- volutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105.

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

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2015. [A diversity-promoting objec-](https://doi.org/10.18653/v1/N16-1014) [tive function for neural conversation models](https://doi.org/10.18653/v1/N16-1014). *Pro- ceedings of the 2016 Conference of the North Amer- ican Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119.

Jiwei Li, Michel Galley, Chris Brockett, Georgios Sp- ithourakis, Jianfeng Gao, and Bill Dolan. 2016a. [A](https://doi.org/10.18653/v1/P16-1094) [persona-based neural conversation model](https://doi.org/10.18653/v1/P16-1094). In *Asso- ciation for Computational Linguistics*, pages 994– 1003.

Jiwei Li, Will Monroe, Alan Ritter, Dan Jurafsky, Michel Galley, and Jianfeng Gao. 2016b. [Deep rein-](https://doi.org/10.18653/v1/D16-1127) [forcement learning for dialogue generation](https://doi.org/10.18653/v1/D16-1127). In *Pro- ceedings of the 2016 Conference on Empirical Meth- ods in Natural Language Processing*, pages 1192– 1202.

Jiwei Li, Will Monroe, Tianlin Shi, Se˙bastien Jean, Alan Ritter, and Dan Jurafsky. 2017b. [Adversarial](http://aclweb.org/anthology/D17-1230) [learning for neural dialogue generation](http://aclweb.org/anthology/D17-1230). In *Proceed- ings 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](https://doi.org/10.18653/v1/W15-4640) [dataset for research in unstructured multi-turn dia-](https://doi.org/10.18653/v1/W15-4640) [logue systems](https://doi.org/10.18653/v1/W15-4640). 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 Cor- rado, and Jeff Dean. 2013. Distributed representa- tions of words and phrases and their compositional- ity. 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 for-](http://www.aclweb.org/anthology/C16-1316) [ward sequences: A content-introducing approach to](http://www.aclweb.org/anthology/C16-1316) [generative short-text conversation](http://www.aclweb.org/anthology/C16-1316). In *Proceedings of COLING 2016, the 26th International Confer- ence on Computational Linguistics: Technical Pa- pers*, pages 3349–3358.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. [Deep contextualized word rep-](https://doi.org/10.18653/v1/N18-1202) [resentations](https://doi.org/10.18653/v1/N18-1202). In *Proceedings of the 2018 Confer- ence of the North American Chapter of the Associ- ation for Computational Linguistics: Human Lan- guage Technologies, Volume 1 (Long Papers)*, pages 2227–2237.

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

Lifeng Shang, Zhengdong Lu, and Hang Li. 2015. [Neural responding machine for short-text conversa-](https://doi.org/10.3115/v1/P15-1152) [tion](https://doi.org/10.3115/v1/P15-1152). 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 opportuni- ties 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 ensem- ble 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, Du- mitru Erhan, Vincent Vanhoucke, and Andrew Ra- binovich. 2015. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vi- sion 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](https://www.aclweb.org/anthology/D18-1479) [attention refinement for matching text sequences](https://www.aclweb.org/anthology/D18-1479). In *Proceedings of the 2018 Conference on Empiri- cal Methods in Natural Language Processing*, pages 4492–4502.

Oriol Vinyals and Quoc Le. 2015. A neural conversa- tional 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 spa- tial rnn. In *IJCAI*, pages 2922–2928.

Hao Wang, Zhengdong Lu, Hang Li, and Enhong Chen. 2013. [A dataset for research on short-text](http://aclweb.org/anthology/D13-1096) [conversations](http://aclweb.org/anthology/D13-1096). In *Proceedings of the 2013 Confer- ence 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 Inter- national Joint Conference on Artificial Intelligence*, pages 1354–1361.

Shuohang Wang and Jing Jiang. 2016. [Learning nat-](https://doi.org/10.18653/v1/N16-1170) [ural language inference with LSTM](https://doi.org/10.18653/v1/N16-1170). In *Proceed- ings of the 2016 Conference of the North Ameri- can 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](https://www.aclweb.org/anthology/P18-2067) [for response selection in retrieval-based chatbots](https://www.aclweb.org/anthology/P18-2067). In *Proceedings of the 56th Annual Meeting of the As- sociation 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 match-](https://doi.org/10.1162/coli_a_00345) [ing framework for multi-turn response selection in](https://doi.org/10.1162/coli_a_00345) [retrieval-based chatbots](https://doi.org/10.1162/coli_a_00345). *Computational Linguis- tics*, 45(1):163–197.

Yu Wu, Wei Wu, Chen Xing, Ming Zhou, and Zhou- jun Li. 2017. [Sequential matching network: A](https://doi.org/10.18653/v1/P17-1046) [new architecture for multi-turn response selection](https://doi.org/10.18653/v1/P17-1046) [in retrieval-based chatbots](https://doi.org/10.18653/v1/P17-1046). In *Proceedings of the 55th Annual Meeting of the Association for Compu- tational 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 *SI- GIR*, 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](https://www.aclweb.org/anthology/P18-1205) [have pets too?](https://www.aclweb.org/anthology/P18-1205) In *Proceedings of the 56th An- nual 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-](http://aclweb.org/anthology/C18-1317) [turn conversation with deep utterance aggregation](http://aclweb.org/anthology/C18-1317). 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 autoen- coders. In *Proceedings of the 55th Annual Meet- ing 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 ma- chine: Emotional conversation generation with in- ternal 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](https://doi.org/10.18653/v1/D16-1036) [conversation](https://doi.org/10.18653/v1/D16-1036). In *Proceedings of the 2016 Confer- ence 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](http://aclweb.org/anthology/P18-1103) [with deep attention matching network](http://aclweb.org/anthology/P18-1103). In *Proceed- ings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Pa- pers)*, volume 1, pages 1118–1127.