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

Jia-Chen Gu
University of Science and Technology of China
gujc@mail.ustc.edu.cn

Zhen-Hua Ling
University of Science and Technology of China
zhling@ustc.edu.cn

Quan Liu
University of Science and Technology of China
iFLYTEK Research
quanliu@ustc.edu.cn

Abstract

In this paper, we propose an interactive matching network (IMN) to enhance the representations of contexts and responses at both the word level and sentence level for the multi-turn response selection task. First, IMN constructs word representations from three aspects to address the challenge of out-of-vocabulary (OOV) words. Second, an attentive hierarchical recurrent encoder (AHRE), which is capable of encoding sentences hierarchically and generating more descriptive representations by aggregating with an attention mechanism, is designed. Finally, the bidirectional interactions between whole multi-turn contexts and response candidates are calculated to derive the matching information between them. Experiments on four public datasets show that IMN significantly outperforms the baseline models by large margins on all metrics, achieving new state-of-the-art performance and demonstrating compatibility across domains for multi-turn response selection.

IMN从三个方面构建单词表示，以解决词汇表外(OOV)单词的挑战

一个带注意力的分层递归编码器设计了一种层次结构的句子编码方法(AHRE)

计算了多轮上下文与候选响应之间的双向交互，得到了它们之间的匹配信息

1 Introduction

Building a chatbot that can converse naturally with humans on open domain topics is a challenging yet intriguing problem in artificial intelligence. Recently, human-computer conversation has attracted increasing attention due to its promising potential and commercial value (Chen et al., 2017a; Young et al., 2017; Gu et al., 2018). Existing work on building chatbots includes generation-based methods (Shang et al., 2015; Serban et al., 2016) and retrieval-based methods (Lowe et al., 2015; Wu et al., 2017; Zhou et al., 2018; Zhang et al., 2018). Response selection, which aims to select the best-matched response from a set of candidates given the context of a conversation, is an important retrieval-based task for chatbots.

Many previous studies focus on single-turn dialogue, which takes the context of only the last utterance into consideration (Wang et al., 2013; Ji et al., 2014). However, extended multi-turn dialogue, which contains more information, is more practical and has attracted more attention. The existing work on multi-turn response selection can be categorized into two main architectures: concatenating all utterances in a conversation (Lowe et al., 2015; Kadlec et al., 2015; Lowe et al., 2017) and separating utterances followed by aggregation (Wu et al., 2017; Zhou et al., 2018; Zhang et al., 2018).

现有的多转向响应选择工作可以分为两种主要的体系结构:
(1)将会话中的所有语句连接起来;
(2)分离话语，然后聚合

The techniques of word embeddings and sentence embeddings are important to response selection as well as many other natural language processing (NLP) tasks. The context and the response must be projected to a vector space appropriately to capture their relationships, which are essential for the subsequent procedures. Recently, there has been growing interest in models for the word-level (Mikolov et al., 2013; Pennington et al., 2014; Dong and Huang, 2018) and sentence-level (Wang et al., 2017; Chen et al., 2017b) representations using neural networks, which have helped classification or inference algorithms to achieve better performance on many NLP tasks.

Another key technique to the response selection task lies in context-to-response matching. Modelling the semantic matching degree between two sentences is challenging. Chen et al. (2017b) showed that interactions between pairs of sentences can provide useful information to help matching. This type of matching method relies on only alignment and is fully computationally decomposable with respect to a pair of sentences.

句子对之间的相互作用可以提供有用的信息来帮助匹配。这种类型的匹配方法只依赖于对齐，并且完全可以对一对句子进行计算分解

In this paper, we propose a novel neural network architecture, called the interactive matching network (IMN), for multi-turn response selection in retrieval-based chatbots. To alleviate the issue of a large number of out-of-vocabulary (OOV) words, IMN constructs word representations with a combination of general pretrained word embedding vectors, those estimated on the task-specific training set and character-level embeddings vectors. Then, an attentive hierarchical recurrent encoder (AHRE) that is capable of encoding sentences hierarchically and generating more descriptive representations by aggregating with an attention mechanism is designed. Finally, IMN accounts for interactions between the context and the response by considering each whole multi-turn context as a single sequence. The context collects matching information from the response to enrich its representations, and the response does the same from the context. This global and bidirectional context-response interaction helps the context and response to capture information from each other to enhance the matching information.

IMN通过将整个多轮上下文看作一个序列来考虑上下文和响应之间的交互。

上下文从响应中收集匹配信息以丰富其表示，而响应也从上下文中收集匹配信息。这种全局和双向的上下文-响应交互帮助上下文和响应彼此捕获信息，从而增强匹配信息

We test our model on two English datasets, Ubuntu Dialogue Corpus V1 (Lowe et al., 2015) and Ubuntu Dialogue Corpus V2 (Lowe et al., 2017), and two Chinese datasets, Douban Conversation Corpus (Wu et al., 2017) and E-commerce Dialogue Corpus (Zhang et al., 2018), which are large-scale datasets that are publicly available for research on multi-turn conversation. The results show that our model can significantly outperform the baseline models by large margins on all metrics, achieving a new state-of-the-art performance and showing compatibility across domains for multi-turn response selection.

In summary, our contributions in this paper are twofold. (1) This paper proposes a new model, named IMN, for multi-turn response selection in retrieval-based chatbots. (2) The empirical results show that our proposed model outperforms the baseline models by large margins in terms of all metrics on four datasets, achieving new state-of-the-art performance for multi-turn response selection.

2 Related Work

Chatbots aim to engage users in human-computer conversations in the open domain and are currently receiving increasing attention because they can target unstructured dialogue without a priori logical representation of the information exchanged during the conversation. Existing work on building chatbots includes generation-based methods (Shang et al., 2015; Serban et al., 2016) and retrieval-based methods (Lowe et al., 2015; Kadlec et al., 2015; Lowe et al., 2017; Wu et al., 2017; Zhou et al., 2018; Zhang et al., 2018). Generation-based models maximize the probability of generating a response given the previous dialogue. This approach enables the incorporation of rich context when mapping between consecutive dialogue turns. Retrieval-based chatbots have the advantage of informative and fluent responses because they select a proper response for the current conversation from a repository by means of response selection algorithms.

聊天机器人的目标是让用户在开放领域进行人机对话，目前正受到越来越多的关注，因为它们可以针对非结构化对话，而无需对对话过程中交换的信息进行先验的逻辑表示。

Previous work on retrieval-based chatbots focuses on response selection for single-turn conversation (Wang et al., 2013; Ji et al., 2014). Recently, researchers have extended the focus to multi-turn conversation, which is more practical for real applications. For example, Lowe et al. (2015), Kadlec et al. (2015) and Lowe et al. (2017) matched a response with the literal concatenation of context utterances. Zhou et al. (2016) improved multi-turn response selection with a multi-view model, including an utterance view and a word view. Wu et al. (2017) proposed the sequential matching network (SMN) to first match the response with each utterance and then to accumulate matching information by recurrent neural network (RNN). Zhang et al. (2018) refined utterance and employed self-matching attention to route the vital information in each utterance based on the SMN. Zhou et al. (2018) constructed representations at different granularities with stacked self-attention. Our proposed IMN model is based on the SMN (Wu et al., 2017) and improves it by (1) constructing word representations from three aspects, (2) enhancing sentence representations through AHRE and (3) capturing bidirectional interactions between contexts and responses.

(1)从三个方面构建词语表征，
(2)加强句子的表示法 AHRE
(3)捕捉上下文和响应之间的双向交互。

3 Interactive Matching Network

3.1 Problem Formalization

Given a dialogue dataset $\mathcal{D}$, an example of the dataset can be represented as (c, r, y). Specifically, $c = \{u_1, u_2, ..., u_n\}$ represents a conversation context with $\{u_k\}_{k=1}^{n}$ as the utterances. r is a response candidate, and $y \in \{0, 1\}$ denotes a label. $y = 1$ indicates that r is a proper response for c; otherwise, $y = 0$. Our goal is to learn a matching model $g(c, r)$ from $\mathcal{D}$. For any context-response pair (c, r), $g(c, r)$ measures the matching degree between c and r.

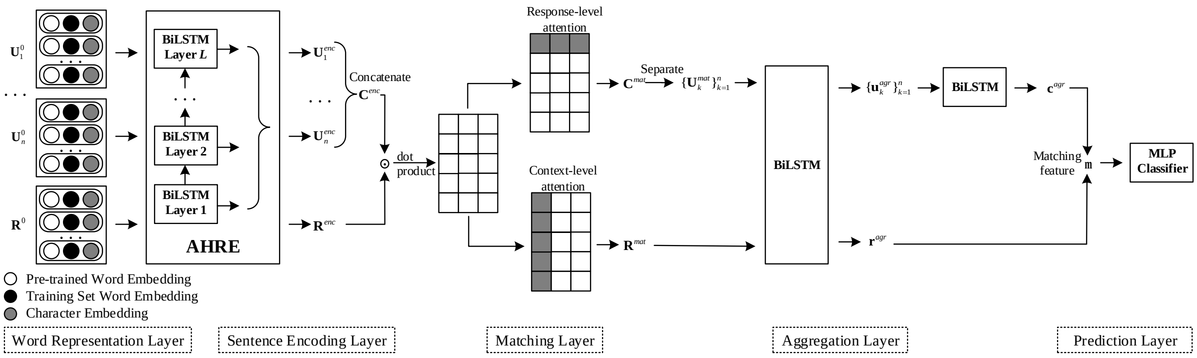

Figure 1: An overview of our proposed IMN model.

3.2 Model Overview

We present here our proposed IMN model, which is composed of five components: word representation layer, sentence encoding layer, matching layer, aggregation layer and prediction layer. Figure 1 shows an overview of the architecture.

IMN first constructs word representations with a combination of general pretrained word embeddings, those estimated on the task-specific training set and character-level embeddings. Then, utterances and a response are encoded with an attentive hierarchical recurrent encoder. Furthermore, IMN uses mutual context-level and response-level attention to collect matching information between context and response. Moreover, these high-order representations are fed into an RNN to obtain a set of utterance embeddings or response embeddings. The set of utterance embeddings are sent to another RNN following the chronological order of the utterances in the context to obtain the context embeddings. Finally, the context embeddings and the response embeddings are used to form the matching feature vectors, which are processed by a multi-layer perceptron to compute the matching score between the context utterances and the response.

首先，IMN可以显著缓解大量的OOV问题。

AHRE的设计目的是对句子进行分层编码，并使用注意机制进行聚合，以生成更具描述性的表示

双向收集匹配信息有助于丰富上下文和响应的表示

IMN has the following characteristics for multi-turn response selection. First, IMN can significantly alleviate the issue of a large number of OOV words. Second, AHRE is designed to encode the sentence hierarchically and aggregate with an attention mechanism to generate more descriptive representations. Third, collecting matching information bidirectionally can help to enrich the representations of the context and the response. In summary, the characteristics benefit the final feature vectors for response selection.

For the matching part of a model, in contrast to the ESIM model (Chen et al., 2017b), which matches a sentence with a sentence, IMN performs matching between a sentence (i.e., response) and a sequence of sentences (i.e., multi-turn contexts). Moreover, in contrast to the SMN model (Wu et al., 2017), which computes word-level and sentence-level similarities directly to distill a matching vector, IMN calculates bidirectional interactions between responses and whole contexts to collect matching information. IMN also improves the word representation and sentence encoding components, while both ESIM and SMN are constructed with only pretrained word embeddings and a single-layer RNN encoder.

IMN计算响应和整个上下文之间的双向交互，以收集匹配信息。

Details about each layer are provided in the following sections.

3.3 Word Representation Layer

One challenge of large dialogue corpora is the large number of OOV words. To address this issue, Dong and Huang (2018) proposed an algorithm that combines the general pretrained word embeddings with those estimated on a task-specific training set. To further enhance the word embeddings, a convolutional neural network (CNN) was employed to model the morphology information at the character-level (Lee et al., 2017) and has shown its effectiveness at addressing OOV words.

Formally, the embeddings of the k-th utterance in a conversation and a response candidate at this layer are denoted as $\mathbf{U}_k^0 = \{\mathbf{u}_{k,i}^0\}_{i=1}^{l_{u_k}}$ and $\mathbf{R}^0 = \{\mathbf{r}_j^0\}_{j=1}^{l_r}$, respectively. $\mathbf{u}_{k,i}^0$ and $\mathbf{r}_j^0 \in \mathbb{R}^d$ are embeddings of a d-dimensional vector.

3.4 Sentence Encoding Layer

The recurrent neural network (RNN) (Mikolov et al., 2010) has been proven to be good at modelling chronological relationships in language sequences, and multi-layer RNNs have achieved good performance in many NLP tasks, such as neural machine translation (NMT) (Bahdanau et al., 2014). Encoding sequences with deep neural networks can help to capture deeper and more useful information. Typically, the outputs of the top RNN layer are regarded as the final sentence representations, and the other layers are neglected. However, the lower layers can also provide useful sentence descriptions, such as part-of-speech tagging and syntax-related information (Hashimoto et al., 2017).

较低层还可以提供有用的句子描述，如词性标注和与语法相关的信息

To make full use of the representations at all hidden layers, we propose a new sentence encoder, called the attentive hierarchical recurrent encoder (AHRE). This encoder is motivated by the method of embeddings from language models (ELMo) (Peters et al., 2018), which combines the internal states of multi-layer RNNs. Specifically, an AHRE learns a linear combination of the vectors stacked above each input word, which improves the performance compared to using only the top RNN layer.

具体地说，一个AHRE学习了堆叠在每个输入单词上方的向量的线性组合，这与只使用顶部RNN相比提高了性能

Furthermore, bidirectional LSTMs (BiLSTMs) (Hochreiter and Schmidhuber, 1997) are employed as our basic building blocks. In an L-layer RNN, each l^{th} layer takes the output of the $l-1^{th}$ layer as its input. We denote the calculations as the follows,

$$\mathbf{u}_{k,i}^{l} = \textbf{BiLSTM}(\mathbf{U}_k^{l-1}, i), i \in \{1, ..., l_{u_k}\}, \quad (1)$$

$$\mathbf{r}_{j}^{l} = \textbf{BiLSTM}(\mathbf{R}^{l-1}, j), j \in \{1, ..., l_r\}. \quad (2)$$

where $\mathbf{U}_k^l = \{\mathbf{u}_{k,i}^l\}_{i=1}^{l_{u_k}}$ and $\mathbf{R}^l = \{\mathbf{r}_j^l\}_{j=1}^{l_r}$, $l \in \{1, ..., L\}$. The weights for these two BiLSTMs are shared in our implementation. Due to space limitations, we omit the description of basic chain LSTMs; the reader can refer to Hochreiter and Schmidhuber (1997) for details.

一个双向LSTM

Finally, we obtain a set of L representations $\{\mathbf{U}_k^1, ..., \mathbf{U}_k^L\}$ and $\{\mathbf{R}^1, ..., \mathbf{R}^L\}$ for the k-th utterance in a conversation and a response candidate through the L-layer RNNs. Typically, $\mathbf{U}_k^L$ or $\mathbf{R}^L$, i.e., the outputs of the top layer, are used as the final encoded vectors. Here, we propose to combine the set of representations to obtain enhanced representations $\mathbf{u}_{k,i}^{enc}$ and $\mathbf{r}_j^{enc}$ by learning the attention weights of all the layers. Mathematically, we have

$$\mathbf{u}_{k,i}^{enc} = \sum_{l=1}^{L} w_l \mathbf{u}_{k,i}^{l}, i \in \{1, ..., l_{u_k}\}, \quad (3)$$

归一化权重让模型自己学 对每个层级的softmax归一化权重

$$\mathbf{r}_{j}^{enc} = \sum_{l=1}^{L} w_l \mathbf{r}_{j}^{l}, j \in \{1, ..., l_r\}. \quad (4)$$

where $\mathbf{U}_k^{enc} = \{\mathbf{u}_{k,i}^{enc}\}_{i=1}^{l_{u_k}}$, $\mathbf{R}^{enc} = \{\mathbf{r}_j^{enc}\}_{j=1}^{l_r}$ and w_l are the softmax-normalized weights shared between utterances and responses, which need to be estimated during the training process. As a result, representations given by our encoder are expected to capture and fuse multi-level characteristics of sentences.

我们的编码器给出的表示形式有望捕获和融合句子的多层次特征

3.5 Matching Layer

Interactions between the context and the response provide important information for deciding the matching degree between them. Unlike previous work, which matches the response with each utterance in the context separately in an utterance-to-response manner (Wu et al., 2017; Zhou et al., 2018; Zhang et al., 2018), IMN matches the response with the whole context in a global context-to-response way, i.e., considering the whole context as a single sequence. The goal of utterance-to-response matching is to collect the relevant parts in each utterance while neglecting the possible premise that the whole utterance is irrelevant to the response. Collecting any part of an irrelevant utterance introduces noise for the matching process. Instead, global context-to-response matching can help to select the most relevant parts of the whole context and neglect the irrelevant parts.

上下文和响应之间的交互为决定它们之间的匹配程度提供了重要的信息。

话语-反应匹配的目的是收集每个话语中的相关部分，而忽略整个话语与反应无关的可能前提。收集不相关话语的任何部分都会为匹配过程带来噪声。全局上下文-响应匹配可以帮助选择整个上下文中最相关的部分，而忽略不相关的部分

First, the context representation $\mathbf{C}^{enc} = \{\mathbf{c}_i^{enc}\}_{i=1}^{l_c}, l_c = \sum_{k=1}^{n} l_{u_k}$ is formed by concatenating the set of utterance representations $\{\mathbf{U}_k^{enc}\}_{k=1}^{n}$.

$$\mathbf{C}^{enc} = Concatenate(\{\mathbf{U}_k^{enc}\}_{k=1}^{n}) \quad (5)$$

Then, an attention-based alignment is employed to collect information between two sequences by computing the attention weight between each representation tuple $\{\mathbf{c}_i^{enc}, \mathbf{r}_j^{enc}\}$ as

$$e_{ij} = (\mathbf{c}_i^{enc})^T \cdot \mathbf{r}_j^{enc} \quad (6)$$

Furthermore, local inference is determined by the attention weights e_{ij} computed above to obtain the local relevance between a context and a

局部推理由上述计算的注意权值eij确定，得到两者之间的局部关联

response bidirectionally. For a word in the context, its response-level relevant representation carried by the response is identified and composed using e_{ij} as

$$\bar{\mathbf{c}}_i^{enc} = \sum_{j=1}^{l_r} \frac{exp(e_{ij})}{\sum_{k=1}^{l_r} exp(e_{ik})} \mathbf{r}_j^{enc}, i \in \{1, ..., l_c\}, \tag{7}$$

where $\bar{\mathbf{C}}^{enc} = \{\bar{\mathbf{c}}_i^{enc}\}_{i=1}^{l_c}$ and $\bar{\mathbf{c}}_i^{enc}$ is a weighted summation of $\{\mathbf{r}_j^{enc}\}_{j=1}^{l_r}$. Intuitively, the contents in $\{\mathbf{r}_j^{enc}\}_{j=1}^{l_r}$ that are relevant to $\mathbf{c}_i^{enc}$ are selected to form $\bar{\mathbf{c}}_i^{enc}$. The same calculation is performed for each word in the response to form the context-level representations as

$$\bar{\mathbf{r}}_j^{enc} = \sum_{i=1}^{l_c} \frac{exp(e_{ij})}{\sum_{k=1}^{l_c} exp(e_{kj})} \mathbf{c}_i^{enc}, j \in \{1, ..., l_r\}. \tag{8}$$

where $\bar{\mathbf{R}}^{enc} = \{\bar{\mathbf{r}}_j^{enc}\}_{j=1}^{l_r}$. To further enhance the collected information, we compute the differences and the element-wise products between $\{\mathbf{C}^{enc}, \bar{\mathbf{C}}^{enc}\}$ and between $\{\mathbf{R}^{enc}, \bar{\mathbf{R}}^{enc}\}$. The differences and element-wise products are then concatenated with the original vectors to obtain the enhanced representations, as follows,

$$\mathbf{C}^{mat} = [\mathbf{C}^{enc}; \bar{\mathbf{C}}^{enc}; \mathbf{C}^{enc} - \bar{\mathbf{C}}^{enc}; \mathbf{C}^{enc} \odot \bar{\mathbf{C}}^{enc}] \tag{9}$$

双向注意力机制

$$\mathbf{R}^{mat} = [\mathbf{R}^{enc}; \bar{\mathbf{R}}^{enc}; \mathbf{R}^{enc} - \bar{\mathbf{R}}^{enc}; \mathbf{R}^{enc} \odot \bar{\mathbf{R}}^{enc}] \tag{10}$$

Thus far, we have collected the relevant information between the context and the response; now, we have to convert the concatenated context back to separate utterances.

$$\{\mathbf{U}_k^{mat}\}_{k=1}^{n} = Separate(\mathbf{C}^{mat}) \tag{11}$$

3.6 Aggregation Layer

通常使用RNN和池操作作为聚合方法来组合和获取句子嵌入。

An RNN followed by a pooling operation is typically employed as the aggregation method to compose and obtain the sentence embeddings. Here, BiLSTM and a combination of max pooling and last hidden state pooling are employed to obtain the utterance and response embeddings.

最后使用隐式状态池获取语句和响应嵌入。

First, the utterance and response embeddings are composed by the enhanced local matching information $\mathbf{U}_k^{mat}$ and $\mathbf{R}^{mat}$ as

$$\mathbf{u}_{k,i} = \textbf{BiLSTM}(\mathbf{U}_k^{mat}, i), i \in \{1, ..., l_{u_k}\}, \tag{12}$$

$$\mathbf{r}_j = \textbf{BiLSTM}(\mathbf{R}^{mat}, j), j \in \{1, ..., l_r\}. \tag{13}$$

The weights for these two BiLSTMs are shared in our implementation. Then, the aggregated embeddings are calculated by pooling operations as

$$\mathbf{u}_k^{agr} = [\mathbf{u}_{k,max}; \mathbf{u}_{k,l_{u_k}}], k \in \{1, ..., n\}, \tag{14}$$

$$\mathbf{r}^{agr} = [\mathbf{r}_{max}; \mathbf{r}_{l_r}]. \tag{15}$$

Furthermore, the set of utterance inference vectors $\{\mathbf{u}_k^{agr}\}_{k=1}^{n}$ is fed into another BiLSTM in chronological order of the utterances in the context as

$$\mathbf{c}_k = \textbf{BiLSTM}(\mathbf{U}^{agr}, k), k \in \{1, ..., n\}. \tag{16}$$

Another pooling operation is performed to obtain the aggregated context embeddings as

$$\mathbf{c}^{agr} = [\mathbf{c}_{max}; \mathbf{c}_n]. \tag{17}$$

The final matching feature vector is the concatenation of the context embeddings and the response embeddings as

$$\mathbf{m} = [\mathbf{c}^{agr}; \mathbf{r}^{agr}]. \tag{18}$$

3.7 Prediction Layer

We then input the matching feature vector $\mathbf{m}$ into a multi-layer perceptron (MLP) classifier. An MLP is a feedforward neural network that is estimated in a supervised manner using examples of features together with known labels. Here, the MLP is designed to predict whether a context and response pair match appropriately according to the matching feature $\mathbf{m}$. Finally, the MLP returns a score to denote the degree of matching.

MLP是一种前馈神经网络，它以一种有监督的方式估计，使用特征的例子和已知的标签。

3.8 Training Criteria

We learn $g(c, r)$, which provides the probability that r is a proper candidate to c by minimizing the sigmoid cross entropy from $\mathcal{D}$. Let Θ denote the parameters of IMN; then, the objective function $\mathcal{L}(\mathcal{D}, \Theta)$ of learning can be formulated as

$$\mathcal{L}(\mathcal{D}, \Theta) = -\sum_{(c,r,y)\in\mathcal{D}} [ylog(g(c,r)) + (1-y)log(1-g(c,r))] \tag{19}$$

Dataset	Ubuntu V1			Ubuntu V2			Douban			E-commerce		
	Train	Valid	Test	Train	Valid	Test	Train	Valid	Test	Train	Valid	Test
pairs	1M	356k	355k	1M	195k	189k	1M	50k	10k	1M	10k	10k
positive:negative	1: 1	1: 9	1: 9	1: 1	1: 9	1: 9	1: 1	1: 1	1: 9	1: 1	1: 1	1: 9
positive/context	1	1	1	1	1	1	1	1	1.18	1	1	1
turns/context	8.44	2.66	2.65	6.29	5.86	6.03	6.69	6.75	5.95	5.51	5.48	5.64
words/utterance	20.38	21.16	21.17	14.06	15.28	15.28	18.56	18.50	20.74	7.02	6.99	7.11

Table 1: Statistics of the datasets that our model is tested on.

4 Experiments

4.1 Datasets

We tested IMN on two English public multi-turn response selection datasets, Ubuntu Dialogue Corpus V1 (Lowe et al., 2015) and Ubuntu Dialogue Corpus V2 (Lowe et al., 2017), and two Chinese datasets, Douban Conversation Corpus (Wu et al., 2017) and E-commerce Dialogue Corpus (Zhang et al., 2018). Ubuntu Dialogue Corpus V1 and V2 contain multi-turn dialogues about Ubuntu system troubleshooting in English. V2 is an updated version of V1, in that V2 separates the training/validation/testing sets by time, which more closely mimics the real-life implementation of training a model on past data to predict future data. In both of the Ubuntu corpora, the positive responses are true responses from humans, and the negative responses are randomly sampled. The Douban Conversation Corpus was crawled from a Chinese social network on open-domain topics. It was constructed in a similar way to the Ubuntu corpus. The Douban Conversation Corpus collected responses via a small inverted-index system, and labels were manually annotated. The E-commerce Dialogue Corpus collected real-world conversations between customers and customer service staff from the largest e-commerce platform in China. Some statistics of these datasets are provided in Table 1.

4.2 Evaluation Metrics

We used the same evaluation metrics as those used in previous work (Wu et al., 2017). Each model was tasked with selecting the k best-matched responses from n available candidates for the given conversation context c, and we calculated the recall of the true positive replies among the k selected responses, denoted as $\mathbf{R}_n@k$, as the main evaluation metric. In addition to $\mathbf{R}_n@k$, we considered the mean average precision (**MAP**) (Baeza-Yates et al., 1999), mean reciprocal rank (**MRR**) (Voorhees et al., 1999) and precision-at-one (**P**@1), especially for the Douban corpus, following the settings of previous work.

4.3 Training Details

The Adam method (Kingma and Ba, 2014) was employed for optimization, with a batch size of 96 for the two English datasets and 128 for the two Chinese datasets. The initial learning rate was 0.001 and was exponentially decayed by 0.96 every 5000 steps. Dropout (Srivastava et al., 2014) with a rate of 0.2 was applied to the word embeddings and all hidden layers. The word embeddings for the English datasets were concatenations of the 300-dimensional GloVe embeddings (Pennington et al., 2014), 100-dimensional embeddings estimated on the training set using the Word2Vec algorithm (Mikolov et al., 2013) and 150-dimensional character-level embeddings with window sizes of {3, 4, and 5}, each consisting of 50 filters. The word embeddings for the Chinese datasets were concatenations of the 200-dimensional embeddings from Song et al. (2018) and the 200-dimensional embeddings estimated on the training set using the Word2Vec algorithm. Character-level embeddings were not employed for the two Chinese datasets due to the large number of Chinese characters. The word embeddings were not updated during training. All hidden states of the LSTM had 200 dimensions. The number of BiLSTM layers in the AHRE was 3. The MLP at the prediction layer had a hidden unit size of 256 with ReLU (Nair and Hinton, 2010) activation. The maximum word length was set to 18, the maximum utterance length was set to 50, and the maximum context length was set to 10. We padded with zeros if the number of utterances in a context was less than 10; otherwise, we kept the last 10 utterances. We used the development dataset to set the stop condition to select the best model for testing.

All codes were implemented in the TensorFlow framework (Abadi et al., 2016) and will be published to help replicate our results after paper

	Ubuntu Corpus V1				Ubuntu Corpus V2			
	$\mathbf{R}_{2}$@1	$\mathbf{R}_{10}$@1	$\mathbf{R}_{10}$@2	$\mathbf{R}_{10}$@5	$\mathbf{R}_{2}$@1	$\mathbf{R}_{10}$@1	$\mathbf{R}_{10}$@2	$\mathbf{R}_{10}$@5
TF-IDF (Lowe et al., 2015, 2017)	0.659	0.410	0.545	0.708	0.749	0.488	0.587	0.763
RNN (Lowe et al., 2015, 2017)	0.768	0.403	0.547	0.819	0.777	0.379	0.561	0.836
LSTM (Lowe et al., 2015, 2017)	0.878	0.604	0.745	0.926	0.869	0.552	0.721	0.924
Multi-View (Zhou et al., 2016)	0.908	0.662	0.801	0.951	-	-	-	-
DL2R (Yan et al., 2016)	0.899	0.626	0.783	0.944	-	-	-	-
MV-LSTM (Wan et al., 2016)	0.906	0.653	0.804	0.946	-	-	-	-
Match-LSTM (Wang and Jiang, 2016b)	0.904	0.653	0.799	0.944	-	-	-	-
RNN-CNN (Baudiš et al., 2016)	-	-	-	-	0.911	0.672	0.809	0.956
CompAgg (Wang and Jiang, 2016a)	0.884	0.631	0.753	0.927	0.895	0.641	0.776	0.937
BiMPM (Wang et al., 2017)	0.897	0.665	0.786	0.938	0.877	0.611	0.747	0.921
HRDE-LTC (Yoon et al., 2018)	0.916	0.684	0.822	0.960	0.915	0.652	0.815	0.966
SMN (Wu et al., 2017)	0.926	0.726	0.847	0.961	-	-	-	-
DUA (Zhang et al., 2018)	-	0.752	0.868	0.962	-	-	-	-
DAM (Zhou et al., 2018)	0.938	0.767	0.874	0.969	-	-	-	-
IMN	**0.945**	**0.777**	**0.880**	**0.974**	**0.945**	**0.771**	**0.886**	**0.979**
IMN(Ensemble)	**0.950**	**0.794**	**0.893**	**0.978**	**0.950**	**0.791**	**0.899**	**0.982**

Table 2: Evaluation results of IMN and previous methods on Ubuntu Dialogue Corpus V1 and Ubuntu Dialogue Corpus V2.

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

Table 3: Evaluation results of IMN and previous methods on the Douban Conversation Corpus and E-commerce Corpus. All the results except ours are copied from Wu et al. (2017); Zhang et al. (2018); Zhou et al. (2018).

acceptance[1].

4.4 Experimental Results

Table 2 and Table 3 present the evaluation results of IMN and previous methods. All the results except ours are from the existing literature. IMN significantly outperforms the other models on all metrics and datasets, which demonstrates its ability to select the best-matched response and its compatibility across domains (system troubleshooting, social network and e-commerce). The Douban Conversation Corpus is different from the other three datasets in that it includes multiple correct candidates for a context in the test set, which leads to low $\mathbf{R}_n@ks$, e.g., if there are 3 correct responses, the maximum $\mathbf{R}_{10}$@1 is 0.33. Hence, **MAP** and **MRR** are recommended for reference. Our proposed model outperforms the baseline model SMN by a large margin of 5.1% in terms of $\mathbf{R}_{10}$@1 on Ubuntu Dialogue Corpus V1; 4.1% in terms of **MAP** and 4.6% in terms of **MRR** on the Douban Conversation Corpus; and 16.8% in terms of $\mathbf{R}_{10}$@1 on the E-commerce Dialogue Corpus. Moreover, our proposed model outperforms the present state-of-the-art methods on the respective datasets by a margin of 1.0% in terms of $\mathbf{R}_{10}$@1 on Ubuntu Dialogue Corpus V1; 11.9% in terms of $\mathbf{R}_{10}$@1 on Ubuntu Dialogue Corpus V2; 2.0% in terms of **MAP** and 1.4% in terms of **MRR** on the Douban Conversation Corpus; and 12.0% in terms of $\mathbf{R}_{10}$@1 on the E-commerce Dialogue Corpus, achieving new state-of-the-art performance on all datasets. Furthermore, we provide ensemble models built by averaging the outputs of four single models with identical architectures and different random initializations. The ensemble models further improves the response selection performance.

豆瓣会话语料库与其他三个数据集的不同之处在于，它包含了测试集中上下文的多个正确候选项，这导致了较低的Rn@ks.

因此，推荐MAP和MRR作为参考

我们提供了通过平均四个具有相同架构和不同随机初始化的单个模型的输出来构建的集成模型。

[1] https://github.com/JasonForJoy/IMN

	Ubuntu Corpus V1			
	$\mathbf{R}_2$@1	$\mathbf{R}_{10}$@1	$\mathbf{R}_{10}$@2	$\mathbf{R}_{10}$@5
IMN	0.945	0.777	0.880	0.974
- AHRE	0.941	0.767	0.874	0.972
- Char emb	0.934	0.749	0.863	0.969
- Match	0.938	0.763	0.868	0.970

证明 char-level 很重要

Table 4: Ablation tests of a single model on the Ubuntu Dialogue Corpus V1 test set.

5 Ablations and Analysis

To demonstrate the importance of each component in our proposed model, various parts of the architecture were ablated, and the results were reported for the test set of Ubuntu Dialogue Corpus V1, as shown in Table 4.

AHRE The number of layers in the AHRE was tuned on the validation set, as shown in Fig 2, and was set to 3. The AHRE proposed in this paper can be considered to be a generalized recurrent encoder that degenerates into a single-layer RNN when the number of layers in the AHRE is set to 1. We found that IMNp with a single-layer RNN encoder outperformed SMN by a large margin of 4.1% in terms of $\mathbf{R}_{10}$@1 and achieved slightly better performance than DAM, while DAM requires a multi-layer self-attention encoder. The softmax-normalized weights of every layer in the AHRE are listed in Table 5, which indicates that each layer of the multi-layer RNNs contributes to the sentence embeddings.

一定程度说明层级的有效性 层级 不一定需要很高

Char emb The character embeddings in the word representation layer were ablated, which resulted in a performance decrease. Additionally, we found that the lower layers of the RNN in the AHRE constitute the most weight, as shown in Table 5. These two results lead to the conclusion that morphology information is very important to the response selection task, possibly because morphology information can help to match similar words literally. A response with more literally similar words may be more appropriate.

这两个结果得出的结论是，形态学信息对响应选择任务非常重要，这可能是因为形态学信息可以帮助匹配字面上相似的单词。用更接近字面意思的词来回答可能更合适

Match The matching layer in IMN was replaced with the method used in the SMN, that is, computing word-level and sequence-level similarities, followed by distilling information through an alternation of convolution and pooling operations to form a matching vector. The decreased performance indicates that bidirectional interactions between the context and the response are beneficial for collecting matching information and making decisions on whether the context and response match.

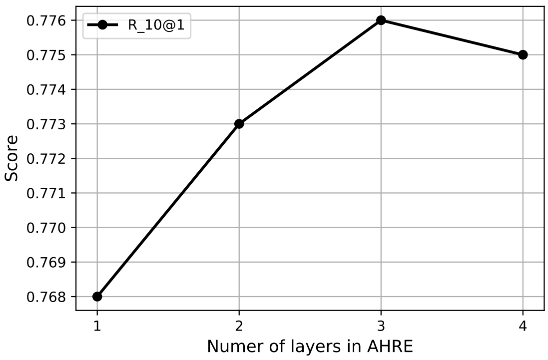

Figure 2: Performance of IMN for different numbers of layers in the AHRE.

	Layer 1	Layer 2	Layer 3
Weights	0.4912	0.2234	0.2854

说明浅层的特征的重要程度
其他的层也有一定的补充

Table 5: Layer-wise weights of the three-layer AHRE used in our experiments.

6 Conclusion

In this paper, we propose an interactive matching network for the multi-turn response selection task. This model enhances the representations of the context and the response at both the word level and sentence level. It also establishes bidirectional and global context-to-response interactions to help capture matching information. An empirical study on four public datasets shows that our proposed model significantly outperforms the baseline models by a large margin on all metrics, achieving new state-of-the-art performance and showing compatibility across domains for multi-turn response selection. However, given a long response composed of multiple utterances, we have neglected the relationships between utterances. Modelling a response consisting of multiple utterances and establishing relationships between them will be the focus of our future work.

References

Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. 2016. Tensorflow: a system for large-scale machine learning. In *OSDI*, volume 16, pages 265–283.

Ricardo Baeza-Yates, Berthier Ribeiro-Neto, et al. 1999. *Modern information retrieval*, volume 463. ACM press New York.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Petr Baudiš, Jan Pichl, Tomáš Vyskočil, and Jan Šedivỳ. 2016. Sentence pair scoring: Towards unified framework for text comprehension. *arXiv preprint arXiv:1603.06127*.

Hongshen Chen, Xiaorui Liu, Dawei Yin, and Jiliang Tang. 2017a. A survey on dialogue systems: Recent advances and new frontiers. *ACM SIGKDD Explorations Newsletter*, 19(2):25–35.

Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, Hui Jiang, and Diana Inkpen. 2017b. Enhanced lstm for natural language inference. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1657–1668.

Jianxiong Dong and Jim Huang. 2018. Enhance word representation for out-of-vocabulary on ubuntu dialogue corpus. *arXiv preprint arXiv:1802.02614*.

Jia-Chen Gu, Zhen-Hua Ling, Yu-Ping Ruan, and Quan Liu. 2018. Building sequential inference models for end-to-end response selection. *arXiv preprint arXiv:1812.00686*.

Kazuma Hashimoto, Yoshimasa Tsuruoka, Richard Socher, et al. 2017. A joint many-task model: Growing a neural network for multiple nlp tasks. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1923–1933.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Zongcheng Ji, Zhengdong Lu, and Hang Li. 2014. An information retrieval approach to short text conversation. *arXiv preprint arXiv:1408.6988*.

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

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. End-to-end neural coreference resolution. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 188–197.

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

Ryan Thomas Lowe, Nissan Pow, Iulian Vlad Serban, Laurent Charlin, Chia-Wei Liu, and Joelle Pineau. 2017. Training end-to-end dialogue systems with the ubuntu dialogue corpus. *Dialogue & Discourse*, 8(1):31–65.

Tomáš Mikolov, Martin Karafiát, Lukáš Burget, Jan Černockỳ, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *Eleventh Annual Conference of the International Speech Communication Association*.

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.

Vinod Nair and Geoffrey E Hinton. 2010. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Matthew 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)*, volume 1, pages 2227–2237.

Iulian Vlad Serban, Alessandro Sordoni, Yoshua Bengio, Aaron C Courville, and Joelle Pineau. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In *AAAI*, volume 16, 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 (Volume 1: Long Papers)*, volume 1, pages 1577–1586.

Yan Song, Shuming Shi, Jing Li, and Haisong Zhang. 2018. Directional skip-gram: Explicitly distinguishing left and right context for word embeddings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language*

Technologies, Volume 2 (Short Papers), volume 2, pages 175–180.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958.

Ellen M Voorhees et al. 1999. The trec-8 question answering track report. In *Trec*, volume 99, pages 77–82.

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 *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, pages 2922–2928. AAAI Press.

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.

Shuohang Wang and Jing Jiang. 2016a. A compare-aggregate model for matching text sequences. *arXiv preprint arXiv:1611.01747*.

Shuohang Wang and Jing Jiang. 2016b. 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.

Zhiguo Wang, Wael Hamza, and Radu Florian. 2017. Bilateral multi-perspective matching for natural language sentences. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 4144–4150. AAAI Press.

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: Long Papers)*, volume 1, pages 496–505.

Rui Yan, Yiping Song, and Hua Wu. 2016. Learning to respond with deep neural networks for retrieval-based human-computer conversation system. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 55–64. ACM.

Seunghyun Yoon, Joongbo Shin, and Kyomin Jung. 2018. Learning to rank question-answer pairs using hierarchical recurrent encoder with latent topic clustering. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, volume 1, pages 1575–1584.

Tom Young, Erik Cambria, Iti Chaturvedi, Minlie Huang, Hao Zhou, and Subham Biswas. 2017. Augmenting end-to-end dialog systems with commonsense knowledge. *arXiv preprint arXiv:1709.05453*.

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

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. 2018. 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.