DEEP HYBRID NETWORKS BASED RESPONSE SELECTION FOR MULTI-TURN DIALOGUE SYSTEMS

Xishuo Li, Lijun Zhang, Wenge Rong, Baiwen Li, Li Qi

School of Computer Science and Engineering, Beihang University, Beijing 100191, China
{lixishuo, ljzhang, w.rong, lbw1406, liq_scse}@buaa.edu.cn

ABSTRACT

Proper response selection is an important challenge for a meaningful multi-turn dialogue. To this end, not only the coherence among the whole dialogue but also the interaction between utterance in adjacent turns need to be properly employed as the context for response selection. In this paper, we propose a deep hybrid network (DHN) to distill such contextual information. First, we match the response with each utterance and filter internal noises with recurrent neural networks. Second, several deep convolutional blocks perform as a feature extractor and output a matching vector to be fused into a final matching score. During this period, complex contextual information across the whole conversation can be thoroughly blended and captured. The empirical study on two commonly used public datasets has shown the proposed model's potential.

Index Terms— dialogue system, response selection, multi-turn conversation, contextual features extraction, deep hybrid network

1. INTRODUCTION

Implementing an intelligent dialogue system which could converse with human-beings coherently is one of the most important tasks for both humanism value and commercial interests. Currently, intelligent dialogue systems can be roughly classified into task-oriented systems (e.g. Cortana, Siri) and open domain ones (e.g. Microsoft Xiaoice) [1], while for both kinds of dialogue systems a fundamental challenge needs to be appropriately solved to maintain the meaningfulness and consistency of outputted responses [2, 3].

One of the promising solutions to this problem is the retrieval based approach, which intends to select the most relevant instances from a history repository with regard to the input in a conversation [4, 5]. This kind of approaches hypothesizes that similar utterances, or question and answering pairs, are probably happened in the previous conversations [2]. Based on this assumption, earlier methods employed the utterances in last input to retrieve possible responses in the repository [4, 6]. Though this kind of methods is easy in implementation, an intelligent dialogue system often has to cope with multiple turns of conversation. As such recent researchers have paid much attention to how to use all turns of inputs to further refine the set of possible instances, which is normally referred as multi-turn based response selection approaches [7, 2, 8].

When taking all utterances in the previous turns of the conversation into account, it is found that different utterances in different turns contribute unequally to final response selection [8]. As such it is important to not only comprehensively and properly model the relationship between utterances in each turn and response candidates, but also consider the correlation between utterances in adjacent turns in selecting candidates [7, 5]. In this research, inspired by recurrent neural networks' (RNN) capability to recognize sequential features and also convolutional neural networks' (CNN) ability to model local information, we propose a deep hybrid network (DHN) to match responses with regard to the all previous utterances.

The proposed solution first matches every response candidate against all the utterances in each turn by utilizing RNN to model sentences sequential information. After this step, for each candidate, there will be its multi-turn context representation. Next, deep convolutional blocks will be used to extract from such contextual representation and output a matching vector to indicate the possible match degree. Afterward, the matching vector will be further analyzed to generate a matching score. Finally, all response candidates' machine scores will be sorted, and then a suitable response will be identified. In the proposed framework, the combination of contextual features with different gratuities could be implemented via every convolution and pooling layer in CNN blocks, which allows fully capturing the contextual information. The experimental results on Ubuntu Corpus and Douban Conversation Corpus demonstrated the proposed methods' potential.

2. RELATED WORK

With the enormous progress made by modern neural-network-based natural language processing techniques, building an intelligent dialogue system which could converse with human-beings consistently is no longer just a fantasy [9, 4, 10]. Toward this end, various data-driven approaches have been

proposed [10, 4, 2, 5, 9, 11, 12], in which modeling multi-turn conversations are under the spotlight to utilize valuable contextual information. To obtain responses based on their informative context, Lowe et al. tried to concatenate the context together and measure the relevance between the encoded context and response representations [7]. Similarly Yan et al. reformulate the context by selecting utterances according to specific strategies [2], while Wu et al. match responses via constructing word-level and segment-level similarity matrixes between each utterance and response [8]. Different from previous approaches, we employ RNN to capture sequential information inside each sentence and deep CNNs to combine local semantics and accumulate contextual features with attentional weights simultaneously [13, 14].

3. METHODOLOGY

3.1. Overview

The architecture of DHN is illustrated in Fig. 1, which can be roughly divided into three parts, namely multi-turn context representation, context-aware matching, and aggregation.

Given a dataset $\{\langle s, r, y\rangle_i\}_{i=1}^{N}$, where s represents a conversation context consisting of $\{u_j\}_{j=1}^{n}$ as utterances and r is the response candidate. The label $y \in \{0, 1\}$ represents the matching degree of s and r. We aim to learn a matching discriminator $g(s, r)$, which can point out the relevance between any context and response.

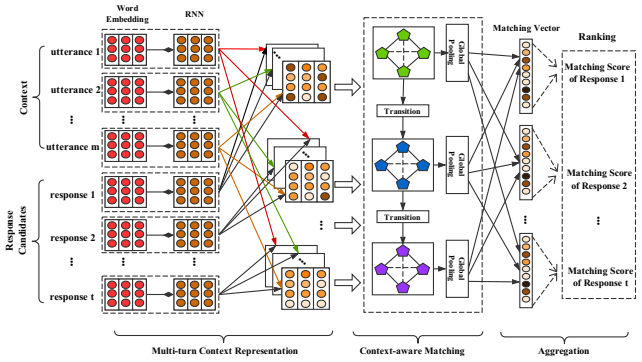

Fig. 1. Architecture of DHN

For each utterance in the context and a response candidate, DHN first represents them as word embedding matrixes, namely $U_i = \left[e_{u_i}^1, e_{u_i}^2, \cdots, e_{u_i}^{n_{u_i}}\right]$ and $R = \left[e_r^1, e_r^2, \cdots, e_r^{n_r}\right]$, where e is a d-dimensional word embedding vector. Then DHN constructs the multi-turn context representation $M \in \mathbb{R}^{L\times P\times P}$ via a representation module, where L is relevant to the representation method, and P is the maximum sentence length (zero-padding and truncating strategies are applied here to convert all utterances to the same length). Besides, each channel of M corresponds to the similarity relationship between the relevant utterance and response.

The candidate's context representation matrix M is then fed into the context-aware matching module. Finally, the matching vector m distilled by the matching module is aggregated into a real-value number $md \in [0, 1]$ through a single layer perceptron, which indicates the matching degree of the conversation context and the response candidate. We employ RNN to model the sequential structure and reduce noises inside each speech in the representation module and deep CNNs to capture local semantics with various granularities and combine contextual information flow across the conversation with attentional weights.

3.2. Multi-turn Context Representation

In this part, the zero-padding and truncating strategies are utilized to convert the size of U_i and R to the $P \times d$ size (P is the padded length and d is the word embedding's dimensionality). Then a matching matrix is built for each utterance-response pair via the following operations:

$$M_{u_i,r} = U_i \cdot S_1 \cdot R^T \tag{1}$$

We set the coefficient matrix S_1 to be identity matrix in our experiments. Note that $M_{u_i,r} \in \mathbb{R}^{P\times P}$. Besides, we apply a RNN to capture internal coherence and filter noises in U_i and R, and transform them into $\bar{U}_i$ and $\bar{R}$. And $\bar{U}_i = \left[z_{u_i}^1, z_{u_i}^2, \cdots, z_{u_i}^{n_{u_i}}\right]$ could be formulated as:

$$z_{u_i}^t = W_1 \cdot \sigma\left(W_2 e_{u_i}^t + W_3 z_{u_i}^{t-1} + b_1\right) + b_2 \tag{2}$$

here W_1, W_2, W_3 and b_1, b_2 are trainable parameters, and $\sigma(\cdot)$ is a sigmoid activation function. $\bar{R}$ is also represented in a similar way. Then the matching between $\bar{U}_i$ and $\bar{R}$ is computed via:

$$\bar{M}_{u_i,r} = \bar{U}_i \cdot S_2 \cdot \bar{R}^T \tag{3}$$

Note that S_2 is a trainable linear transforming matrix. Then all matching matrixes are stacked following the order of the speeches in the conversation to build the multi-turn context representation $M \in \mathbb{R}^{L\times P\times P}$:

$$M = \left[M_{u_1,r}; \cdots M_{u_n,r}; \bar{M}_{u_1,r}; \cdots \bar{M}_{u_n,r}\right] \tag{4}$$

Notice that here L is equal to two times of the maximum dialogue length (we truncate longer conversations and pad zeros into shorter ones to convert all conversations to the same length). In this manner, each channel of M corresponds to an utterance with supervision from response ($M_{u_i,r}$ or $\bar{M}_{u_i,r}$) (Besides this method, we explore other representation approaches which concatenate or add $U_i, \bar{U}_i$ and $R, \bar{R}$ together with weights but the results are relatively worse).

3.3. Context-aware Matching

The CliqueNet architecture [15] introduces recurrent feedbacks into the DenseNet architecture [16] to further enhance

the gradient flow among layers. Derived from CliqueNet, we design the context-aware matching module which consists of several convolutional blocks. Each block could be further divided into two stages. Layers in the first stage take the concatenation of former layers' output as input feature-maps, while the output feature-maps of i^{th} layer in the second stage can be formulated as:

$$x_{i,2} = g\left(\sum_{l<i} W_{li} * x_{l,2} + \sum_{m>i} W_{mi} * x_{m,1}\right) \quad (5)$$

were g is a non-linear transformation function which implements the joint operation $BN - ReLU - Conv^k_{3\times3}$, $*$ represents convolutional operation with parameters matrix W_{ij}, and $,+$ denotes concatenation operation. $BN, ReLU$ and $Conv^k_{3\times3}$ refer to the batch normalization operation [17], ReLU activation function [18] and a convolutional layer with k kernels of 3×3 size respectively. Additionally, layers in the first stage also feed their inputs to the function $g(\cdot)$. Note that in the second stage, every layer is updated based on former layers in the second stage and following layers in the first stage, and W_{ij} decides weights between the i^{th} and j^{th} layer. With this method, a spatial attention mechanism is brought into the deep convolutional blocks. Namely, the most important contextual information in different turns is left behind with higher weights.

We implement three identical blocks in our experiments. Each block concatenates its input and second stage feature-maps as block features. All block features form the matching vector m after global pooling. Every second stage feature-maps function as input to the next block after being processed by a transition layer, which consists of $BN - ReLU - AP - Conv^k_{1\times1}$ operations (AP is an average pooling layer).

3.4. Aggregation

In this part, a fully-connected layer with softmax is employed to translate the matching vector m into a probability distribution. The values of the probability distribution denote the scores of "not matching" and "matching", respectively. The "matching" score is the final matching degree $md \in [0, 1]$. Thus, we utilize the binary cross-entropy (a special case of multi-class cross-entropy) as our loss function:

$$L = -\sum_{i=1}^{N} \left[y_i log\left(g\left(s_i, r_i\right)\right) + \left(1 - y_i\right) log\left(1 - g\left(s_i, r_i\right)\right)\right] \quad (6)$$

where y_i is the true label of samples, and $g\left(s_i, r_i\right)$ is the predicted matching degree (matching score).

4. EXPERIMENTAL STUDY

4.1. Datasets

To evaluate the proposed method's potential, we conduct experiment study on two public datasets, i.e., Douban Conversation Corpus and Ubuntu Corpus, which are typical open-domain and domain-specific datasets, respectively.

Douban Conversation Corpus is a Chinese corpus and published in 2017 [8], which consists of 1 million training data, 50 thousand validation data, and 10 thousand test data. The negative responses in training and validation set are randomly sampled, and the ratio of positive over negative responses is $1 : 1$. In test set, there are 1 positive candidate and 9 negative ones for each context, and negative responses are crawled from Sina Weibo via Lucene (`https://lucenenet.apache.org/`), and their labels are manually annotated.

Ubuntu Corpus is crawled from the Ubuntu Forum [7]. The Ubuntu Corpus (English dataset) is composed of one million training samples, 0.5 million instances in both validation and test sets. The ratio of positive over negative responses is $1 : 1$ for the training set and $1 : 9$ for validation and testing, and all the negative responses are randomly sampled.

4.2. Experimental Settings

To evaluate the proposed method, we follow the configuration in [8] and employ $R_n@k$ (recall at position k in n candidates), MAP (Mean Average Precision), MRR (Mean Reciprocal Rank) and P@1 (Precision-at-one) to evaluate our experimental results.

The proposed model is implemented on Tensorflow [19] and we employ word2vec [20] to initialize the word embedding with the dimensionality of 200. Parameters are updated by stochastic gradient descent with Adam algorithm [21]. The initial learning rate, β_1 and β_2 of Adam are 0.001, 0.9 and 0.999 respectively. The training batch size is set to 60 in our experiments. The maximum dialogue and sentence length are set to 10 and 50 for both datasets respectively. Zero-padding and truncating strategies are applied. Other structural parameters, including the kernels per layer k and the number of total layers L, are tuned to select the best model. All models are trained and tested on a GTX 1060 platform. We select and compare our framework against the most representative baselines, including single-turn approaches: LSTM and biLSTM, MV-LSTM [22], Match-LSTM [6] and Attentive-LSTM [23]. Besides, other advanced multi-turn frameworks are also employed, e.g., DL2R [2], Multi-view [5], and SMN [8].

4.3. Results and Discussion

Table 1 lists the evaluation results and marks the best result in boldface. As shown in this table, our model outperforms all of the baselines, especially on the Douban Conversation Corpus. Moreover, due to the storage limit of our device, models with higher structural parameters (with better performance generally) cannot be trained. Nevertheless, the performance of our framework illustrates the significance of utilizing multi-turn context and modeling the supervision from responses to ut-

Table 1. Comparison results on two public datasets

Model	Ubuntu Corpus			Douban Conversation Corpus					
	$R_{10}@1$	$R_{10}@2$	$R_{10}@5$	MAP	MRR	P@1	$R_{10}@1$	$R_{10}@2$	$R_{10}@5$
LSTM	0.638	0.784	0.949	0.485	0.537	0.320	0.187	0.343	0.720
biLSTM	0.630	0.780	0.944	0.479	0.514	0.313	0.184	0.330	0.716
MV-LSTM	0.653	0.804	0.946	0.498	0.538	0.348	0.202	0.351	0.710
Match-LSTM	0.653	0.799	0.944	0.500	0.537	0.345	0.202	0.348	0.720
Attentive-LSTM	0.633	0.789	0.943	0.495	0.523	0.331	0.192	0.328	0.718
DL2R	0.626	0.783	0.944	0.488	0.527	0.330	0.193	0.342	0.705
Multi-view	0.662	0.801	0.951	0.505	0.543	0.342	0.202	0.350	0.729
SMN	0.726	0.847	**0.961**	0.529	0.569	0.397	0.233	0.396	0.724
Our proposed	**0.733**	**0.852**	**0.961**	**0.562**	**0.621**	**0.432**	**0.241**	**0.407**	**0.754**

terances (compared with single-turn approaches) and the validity of employing RNN-CNN hybrid structure to filter internal noises, distill local semantics and fully capture contextual information flow (compared with multi-turn methods).

In the following paragraphs we would like to analyze the influence of different hyperparameters on our model's performance to help understand the internal mechanism of DHN. Then we discuss the major reasons which lead to wrong predictions to designate the directions of future researches.

Quantity analysis We investigate how our model performs across different maximum dialogue and sentence length. Figure 2 illustrates the changes of $R_n@k$ on the Douban Conversation Corpus. As demonstrated, our model performs stably with both long and short context. Additionally, longer context results to a relatively better performance generally, which proves the effectiveness of richer contextual information. Thus, we set the maximum dialogue length to 10 to balance performance and computation burden. Similarly, our model performance increases steadily with longer sentences, but we set the maximum sentence length to 50 in our experiments to maintain both efficiency and effectivity.

Table 2. Results with various structural parameters

Setting	$R_{10}@1$	$R_{10}@2$	$R_{10}@5$
L=6 k=12	0.217	0.371	0.716
L=12, k=12	0.232	0.385	0.706
L=12, k=24	**0.241**	**0.407**	**0.754**

Structural parameters tuning As shown in Table 2, higher structural parameters result in better performance generally. However, increasing *L=6, k=12* to *L=12, k=12* leads to a slight reduction on the $R_{10}@5$ metric, which implies that maintaining the balance between k and L is also of great significance. Due to the limit of our device, models with the size bigger than *L=12, k=24* could not be trained. Thus, we select the best configurations (*L=12, k=24*) in our experiments.

Error analysis To further improve the performance of DHN, we carefully analyze the failing cases and classify them into two categories. 1) Improper responses: there are multiple suitable response candidates or no appropriate candidates, which lead to the instability of the $R_n@k$ metrics. 2) Logical error: the predicted response may be logically contradictory about former information. Introducing adversarial samples and expert labeling into the training procedure would help resolve the two issues. However, this is generally not low-cost. Hence more automatic techniques need to be explored.

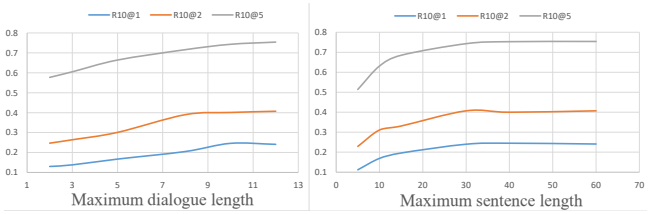

Fig. 2. Changes of performance across different maximum dialogue and sentence length

5. CONCLUSION AND FUTURE WORK

In this paper, we investigate the task of selecting suitable response candidates for the multi-turn cpnversation via the deep hybrid network, which use RNN to construct a candidate's context representation and then use CNN to further study the candidate's suitability from adjacent utterance's perspective. Experimental study on two public datasets demonstrated that our proposed method can outperform all baselines and the comprehensive discussion about our method's mechanism is discussed. In the future, we would study how to model contextually logical consistency and further explore better frameworks which perform closer to human evaluation without expensive expert labeling in the training procedure.

Acknowledgement

This work was supported by Graduate Innovation Practice Fund of Beihang University (No. YCSJ-02-2018-01).

6. REFERENCES

[1] Iulian Vlad Serban, Alessandro Sordoni, Yoshua Bengio, Aaron C Courville, and Joelle Pineau, "Building end-to-end dialogue systems using generative hierarchical neural network models.," in *AAAI*, 2016, vol. 16, pp. 3776–3784.

[2] Rui Yan, Yiping Song, and Hua Wu, "Learning to respond with deep neural networks for retrieval-based human-computer conversation system," in *SIGIR*, 2016, pp. 55–64.

[3] Tsung-Hsien Wen, David Vandyke, Nikola Mrksic, Milica Gasic, Lina M Rojas-Barahona, Pei-Hao Su, Stefan Ultes, and Steve Young, "A network-based end-to-end trainable task-oriented dialogue system," *arXiv preprint arXiv:1604.04562*, 2016.

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

[5] Xiangyang Zhou, Daxiang Dong, Hua Wu, Shiqi Zhao, Dianhai Yu, Hao Tian, Xuan Liu, and Rui Yan, "Multi-view response selection for human-computer conversation," in *EMNLP*, 2016, pp. 372–381.

[6] Shuohang Wang and Jing Jiang, "Learning natural language inference with lstm," *arXiv preprint arXiv:1512.08849*, 2015.

[7] Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau, "The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems," *arXiv preprint arXiv:1506.08909*, 2015.

[8] Yu Wu, Wei Wu, Chen Xing, Can Xu, Zhoujun Li, and Ming Zhou, "A sequential matching framework for multi-turn response selection in retrieval-based chatbots," *arXiv preprint arXiv:1710.11344*, 2017.

[9] Alan Ritter, Colin Cherry, and William B Dolan, "Data-driven response generation in social media," in *EMNLP*, 2011, pp. 583–593.

[10] Lasguido Nio, Sakriani Sakti, Graham Neubig, Tomoki Toda, Mirna Adriani, and Satoshi Nakamura, "Developing non-goal dialog system based on examples of drama television," in *Natural Interaction with Robots, Knowbots and Smartphones*, pp. 355–361. Springer, 2014.

[11] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan, "A diversity-promoting objective function for neural conversation models," *arXiv preprint arXiv:1510.03055*, 2015.

[12] Chen Xing, Wei Wu, Yu Wu, Jie Liu, Yalou Huang, Ming Zhou, and Wei-Ying Ma, "Topic aware neural response generation.," in *AAAI*, 2017, vol. 17, pp. 3351–3357.

[13] Xiaoqiang Zhou, Baotian Hu, Qingcai Chen, and Xiaolong Wang, "Recurrent convolutional neural network for answer selection in community question answering," *Neurocomputing*, vol. 274, pp. 8–18, 2018.

[14] Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V Le, "Qanet: Combining local convolution with global self-attention for reading comprehension," *arXiv preprint arXiv:1804.09541*, 2018.

[15] Yibo Yang, Zhisheng Zhong, Tiancheng Shen, and Zhouchen Lin, "Convolutional neural networks with alternately updated clique," in *CVPR*, 2018, pp. 2413–2422.

[16] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger, "Densely connected convolutional networks," in *CVPR*, 2017, pp. 2261–2269.

[17] Sergey Ioffe and Christian Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.

[18] Xavier Glorot and Yoshua Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *AISTATS*, 2010, pp. 249–256.

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

[20] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean, "Distributed representations of words and phrases and their compositionality," in *NIPS*, 2013, pp. 3111–3119.

[21] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[22] Shengxian Wan, Yanyan Lan, Jun Xu, Jiafeng Guo, Liang Pang, and Xueqi Cheng, "Match-srnn: Modeling the recursive matching structure with spatial rnn," *arXiv preprint arXiv:1604.04378*, 2016.

[23] Ming Tan, Cicero dos Santos, Bing Xiang, and Bowen Zhou, "Lstm-based deep learning models for non-factoid answer selection," *arXiv preprint arXiv:1511.04108*, 2015.