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

Jia-Chen Gu[1], Zhen-Hua Ling[1], Quan Liu[1,2]
[1]University of Science and Technology of China, Hefei, China
[2]State Key Laboratory of Cognitive Intelligence, iFLYTEK Research, Hefei, China
gujc@mail.ustc.edu.cn,zhling@ustc.edu.cn,quanliu@ustc.edu.cn

ABSTRACT

In this paper, we propose an interactive matching network (IMN) 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 outperforms the baseline models on all metrics, achieving a new state-of-the-art performance and demonstrating compatibility across domains for multi-turn response selection.

CCS CONCEPTS

• **Information systems → Retrieval models and ranking**.

KEYWORDS

Interactive matching network, multi-turn response selection, retrieval-based chatbot

ACM Reference Format:
Jia-Chen Gu[1], Zhen-Hua Ling[1], Quan Liu[1,2]. 2019. Interactive Matching Network for Multi-Turn Response Selection in Retrieval-Based Chatbots. In *The 28th ACM International Conference on Information and Knowledge Management (CIKM '19), November 3–7, 2019, Beijing, China*. ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3357384.3358140

1 INTRODUCTION

Building a chatbot that can converse naturally with humans on open domain topics is a challenging yet intriguing problem in artificial intelligence [1]. 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 approach for chatbots [4, 9, 12].

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. Typically, word embeddings established on the task-specific training set and a single-layer recurrent neural network are employed for the response selection task. Another key technique to the response selection task lies in context-response matching. Chen et al. [2] showed that interactions between pairs of sentences can provide useful information to help matching.

Wu et al. [9] proposed the sequential matching network (SMN) to match the response with each utterance and then to accumulate matching information by an RNN. Zhang et al. [11] refined utterance and employed self-matching attention to route the vital information in each utterance based on the SMN. Zhou et al. [12] proposed the deep attention matching network (DAM) to construct representations at different granularities with stacked self-attention.

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. Our proposed IMN is similar to SMN but has three main differences: (1) constructing word representations from three aspects to enhance the representations at the word-level, (2) enhancing sentence representations through an attentive hierarchical recurrent encoder to enhance the representations at the sentence-level and (3) capturing interactions between contexts and responses by collecting matching information bidirectionally to enrich the representations.

We test our model on Ubuntu Dialogue Corpus V1 [4], Ubuntu Dialogue Corpus V2 [5], Douban Conversation Corpus [9] and E-commerce Dialogue Corpus [11]. The results show that our model can outperform the baseline models on all metrics, achieving new state-of-the-art performance and showing compatibility across domains for multi-turn response selection.

In summary, our contributions in this paper are threefold. (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 in terms of all metrics on four datasets, achieving new state-of-the-art performance for multi-turn response selection. (3) This paper presents

CIKM '19, November 3–7, 2019, Beijing, China

ACM ISBN 978-1-4503-6976-3/19/11...$15.00
https://doi.org/10.1145/3357384.3358140

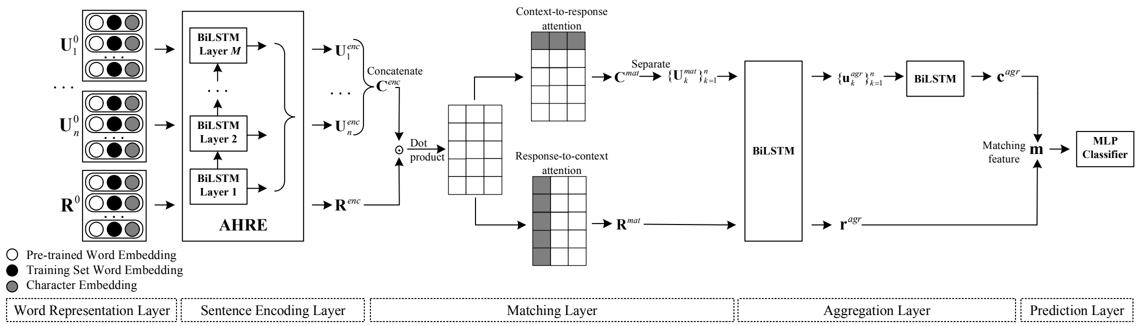

Figure 1: An overview of our proposed IMN model.

detailed experiments and discussions on contributions of each part to context-response pair matching.

2 INTERACTIVE MATCHING NETWORK

We present here our proposed IMN model, which is composed of five layers.Figure 1 shows an overview of the architecture.

2.1 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 $g(c, r)$, which provides the matching degree between c and r by minimizing the sigmoid cross entropy from $\mathcal{D}$.

2.2 Word Representation Layer

One challenge of large dialogue corpora is the large number of OOV words. To address this issue, we propose to construct word representations with a combination of general pretrained word embedding, those estimated on the task-specific training set and character-level embeddings.

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 $\mathbf{R}^0 = \{\mathbf{r}_j^0\}_{j=1}^{l_r}$. $\mathbf{u}_{k,i}^0$ and $\mathbf{r}_j^0 \in \mathbb{R}^d$ are embeddings of a d-dimensional vector. l_{u_k} and l_r are the numbers of words in $\mathbf{U}_k^0$ and $\mathbf{R}^0$ respectively.

2.3 Sentence Encoding Layer

Typically, the outputs of the top layer in a multi-layer RNNs 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. Motivated by the method of ELMo [6], we propose a new sentence encoder, called the attentive hierarchical recurrent encoder (AHRE) to make full use of the representations at all hidden layers.

BiLSTMs [3] are employed as our basic building blocks. In an M-layer RNN, each m^{th} layer takes the output of the $m-1^{th}$ layer as its input.

Finally, we obtain a set of M representations $\{\mathbf{U}_k^1, ..., \mathbf{U}_k^M\}$ and $\{\mathbf{R}^1, ..., \mathbf{R}^M\}$ for the k-th utterance in a conversation and a response candidate through the M-layer RNNs, where $\mathbf{U}_k^m = \{\mathbf{u}_{k,i}^m\}_{i=1}^{l_{u_k}}$ and $\mathbf{R}^m = \{\mathbf{r}_j^m\}_{j=1}^{l_r}$, $l \in \{1, ..., M\}$. Here, we propose to combine the set of representations to obtain the enhanced representations $\mathbf{u}_{k,i}^{enc}$ and $\mathbf{r}_j^{enc}$ by learning the attention weights of all the layers. Mathematically, we have

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

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 AHRE are expected to fuse multi-level characteristics of sentences.

2.4 Matching Layer

Unlike previous work, which matches responses with each utterance in a context separately in an utterance-response manner [9, 11, 12], IMN matches the response with the whole context in a global context-response way, i.e., considering the whole context as a single sequence. The global context-response matching can help select the most relevant parts of the whole context and neglect the irrelevant parts.

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

Then, an attention-based alignment is employed to collect information between two sequences by computing the attention weight between each tuple as $e_{ij} = (\mathbf{c}_i^{enc})^T \cdot \mathbf{r}_j^{enc}$.

For a word in the response, its response-to-context relevant representation is composed as

$$\bar{\mathbf{r}}_j^{enc} = \sum_{i=1}^{l_c} \frac{exp(e_{ij})}{\sum_{k=1}^{l_c} exp(e_{kj})} \mathbf{c}_i^{enc}, j \in \{1, ..., l_r\}, \tag{2}$$

where $\bar{\mathbf{R}}^{enc} = \{\bar{\mathbf{r}}_j^{enc}\}_{j=1}^{l_r}$, $\bar{\mathbf{r}}_j^{enc}$ is a weighted summation of $\{\mathbf{c}_i^{enc}\}_{i=1}^{l_c}$. The same calculation is performed for each word in

Table 1: Evaluation results of IMN and previous methods on Ubuntu Dialogue Corpus V1 and V2.

	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
CompAgg [7]	0.884	0.631	0.753	0.927	0.895	0.641	0.776	0.937
BiMPM [8]	0.897	0.665	0.786	0.938	0.877	0.611	0.747	0.921
HRDE-LTC [10]	0.916	0.684	0.822	0.960	0.915	0.652	0.815	0.966
SMN [9]	0.926	0.726	0.847	0.961	-	-	-	-
DUA [11]	-	0.752	0.868	0.962	-	-	-	-
DAM [12]	0.938	0.767	0.874	0.969	-	-	-	-
IMN	**0.946**	**0.794**	**0.889**	**0.974**	**0.945**	**0.771**	**0.886**	**0.979**
IMN(Ensemble)	**0.951**	**0.807**	**0.900**	**0.978**	**0.950**	**0.791**	**0.899**	**0.982**

Table 2: Evaluation results of IMN and previous methods on the Douban Conversation Corpus and E-commerce Corpus.

	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
SMN [9]	0.529	0.569	0.397	0.233	0.396	0.724	0.453	0.654	0.886
DUA [11]	0.551	0.599	0.421	0.243	0.421	0.780	0.501	0.700	0.921
DAM [12]	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**

a context to form context-to-response representation $\bar{\mathbf{C}}^{enc} = \{\bar{\mathbf{c}}_i^{enc}\}_{i=1}^{l_c}$.

To further enhance the collected information, the matching matrices are formed as

$$\mathbf{C}^{mat} = [\mathbf{C}^{enc}; \bar{\mathbf{C}}^{enc}; \mathbf{C}^{enc} - \bar{\mathbf{C}}^{enc}; \mathbf{C}^{enc} \odot \bar{\mathbf{C}}^{enc}], \quad (3)$$

$$\mathbf{R}^{mat} = [\mathbf{R}^{enc}; \bar{\mathbf{R}}^{enc}; \mathbf{R}^{enc} - \bar{\mathbf{R}}^{enc}; \mathbf{R}^{enc} \odot \bar{\mathbf{R}}^{enc}]. \quad (4)$$

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

2.5 Aggregation Layer

The aggregation layer converts the matching matrices of separated utterances and responses into a final matching vector.

First, the set of utterance embeddings $\mathbf{U}^{agr} = \{\mathbf{u}_k^{agr}\}_{k=1}^{n}$ and the response embeddings $\mathbf{r}^{agr}$ are obtained by composing the enhanced local matching information $\mathbf{U}_k^{mat}$ and $\mathbf{R}^{mat}$ with a BiLSTM, and a combination of max pooling and last hidden state pooling.

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

The final matching feature vector is the concatenation of the context embeddings and the response embeddings as $\mathbf{m} = [\mathbf{c}^{agr}; \mathbf{r}^{agr}]$.

2.6 Prediction Layer

We then input the matching feature vector $\mathbf{m}$ into a multi-layer perceptron (MLP) classifier. The MLP returns a score to denote the matching degree of a context-response pair.

3 EXPERIMENTS

3.1 Datasets

We tested IMN on Ubuntu Dialogue Corpus V1 [4], Ubuntu Dialogue Corpus V2 [5], Douban Conversation Corpus [9] and E-commerce Dialogue Corpus [11].

3.2 Evaluation Metrics

We used the same evaluation metrics as those used in previous work [4, 9, 11]. We calculated the recall of the true positive replies among the k selected responses from n available candidates, denoted as $\mathbf{R}_n@k$. In addition, mean average precision (**MAP**), mean reciprocal rank (**MRR**) and precision-at-one (**P**@1), are especially considered for the Douban corpus, following the settings of previous work.

3.3 Experimental Results

Table 1 and Table 2 present the evaluation results of IMN and previous methods. All the results except ours are from the existing literature. IMN outperforms 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 Corpus includes multiple correct candidates for a context in its test set. Hence, **MAP** and **MRR** are recommended for reference.

Our proposed model outperforms the present state-of-the-art methods on the respective datasets by a margin of 2.6% in terms of $\mathbf{R}_{10}$@1 on Ubuntu V1; 11.9% in terms of $\mathbf{R}_{10}$@1 on Ubuntu V2; 2.0% in terms of **MAP** and 1.4% in terms of **MRR** on Douban Corpus; and 12.0% in terms of $\mathbf{R}_{10}$@1 on E-commerce Corpus, achieving a new state-of-the-art

Table 3: Ablation tests on Ubuntu V2 test set.

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

Table 4: Layer-wise weights of a three-layer AHRE.

	Layer 1	Layer 2	Layer 3
Weights	0.4938	0.2181	0.2881

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. Our code has been published at *https://github.com/JasonForJoy/IMN* to help replicate our results.

4 ABLATIONS AND ANALYSIS

To demonstrate the importance of each component in our proposed model, various parts of the architecture were ablated, as shown in Table 3.

AHRE. The number of layers in the AHRE was set to 3. The AHRE can be considered as a generalized recurrent encoder that degenerates into a single-layer RNN when the number of layers in the AHRE is set to 1. The softmax-normalized weights of layers in the AHRE are listed in Table 4, which indicates that each layer of the multi-layer RNNs contributes to the embeddings.

Char emb. The character embeddings in the word representation layer were ablated, which resulted in a performance decrease. Additionally, we found that the lowest layer of the RNN in the AHRE constituted the highest weight, as shown in Table 4. These two results may be explained by the importance of morphology information to the response selection.

Match. The decreased performance indicates that interactions between contexts and responses are beneficial for matching. We conduct a case study and visualize the response-to-context weights used in Eq. 2 to demonstrate their ability to select relevant parts as shown in Figure 2. Some important words such as "*connect*", "*router*" and "*ethernet*" in the response can select their relevant words in the context, and some unimportant words such as "*tried*", "*channels*" and "*the*" in the context occupy small weights when forming representations.

5 CONCLUSION

In this paper, we propose an interactive matching network for the response selection task. An empirical study on four public datasets shows that our proposed model outperforms

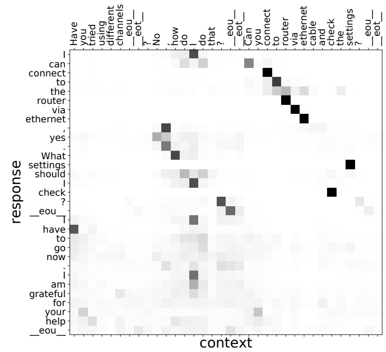

Figure 2: Response-to-context attention weights for a sample. The darker units mean larger values.

the baseline models on all metrics, achieving new state-of-the-art performance and showing compatibility across domains for multi-turn response selection.

ACKNOWLEDGEMENTS

We thank anonymous reviewers for their valuable comments.

REFERENCES

[1] Hongshen Chen, Xiaorui Liu, Dawei Yin, and Jiliang Tang. 2017. A Survey on Dialogue Systems: Recent Advances and New Frontiers. *SIGKDD Explorations* 19, 2 (2017), 25–35.

[2] Qian Chen, Xiaodan Zhu, Zhenhua Ling, Si Wei, Hui Jiang, and Diana Inkpen. 2016. Enhanced lstm for natural language inference. *arXiv preprint arXiv:1609.06038* (2016).

[3] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation* 9, 8 (1997), 1735–1780.

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

[5] 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 (2017), 31–65.

[6] Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of NAACL-HLT*. 2227–2237.

[7] Shuohang Wang and Jing Jiang. 2016. A Compare-Aggregate Model for Matching Text Sequences. *CoRR* abs/1611.01747 (2016).

[8] 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*. AAAI Press, 4144–4150.

[9] 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)*. 496–505.

[10] 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 NAACL-HLT*. 1575–1584.

[11] 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*. 3740–3752.

[12] Xiangyang Zhou, Lu Li, Daxiang Dong, Yi Liu, Ying Chen, Wayne Xin Zhao, Dianhai Yu, and Hua Wu. [n. d.]. Multi-turn response selection for chatbots with deep attention matching network. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*.