Intra-/Inter-Interaction Network with Latent Interaction Modeling for Multi-turn Response Selection*

Yang Deng, Wenxuan Zhang, Wai Lam
The Chinese University of Hong Kong
{ydeng,wxzhang,wlam}@se.cuhk.edu.hk

Abstract

Multi-turn response selection has been extensively studied and applied to many real-world applications in recent years. However, current methods typically model the interactions between multi-turn utterances and candidate responses with iterative approaches, which is not practical as the turns of conversations vary. Besides, some latent features, such as user intent and conversation topic, are under-discovered in existing works. In this work, we propose Intra-/Inter-Interaction Network (I^3) with latent interaction modeling to comprehensively model multi-level interactions between the utterance context and the response. In specific, we first encode the intra- and inter-utterance interaction with the given response from both individual utterance and the overall utterance context. Then we develop a latent multi-view subspace clustering module to model the latent interaction between the utterance and response. Experimental results show that the proposed method substantially and consistently outperforms existing state-of-the-art methods on three multi-turn response selection benchmark datasets.

1 Introduction

Recent years have witnessed many successful real-world applications on chatbots and AI assistants, such as the XiaoIce (Shum et al., 2018) from Microsoft and the E-commerce assistant AliMe (Li et al., 2017) from Alibaba Group, which owe to the extensive researches on dialogue systems. Existing works on building conversational models mainly study generation-based (Wen et al., 2017; Xing et al., 2017) or retrieval-based methods (Lowe et al., 2015; Wu et al., 2017; Zhang et al., 2018). In this work, we focus on the problem of multi-turn response selection for retrieval-based dialogue systems, which aims at selecting appropriate responses from a set of candidates as the reply for the given multi-turn utterances.

Measuring the matching degree between the utterance context and the candidate response is the core of multi-turn response selection task. Recent works develop a variety of interaction model to enhance the utterance-response interaction from a broader (Zhou et al., 2018b; Tao et al., 2019a) or deeper perspective (Tao et al., 2019b; Wang et al., 2019; Yuan et al., 2019). Empirical evidences show that iterative architectures achieve state-of-the-art performance on multi-turn response selection, such as interaction-over-interaction (Tao et al., 2019b), iterated attentive matching (Wang et al., 2019), and multi-hop selector (Yuan et al., 2019).

Despite the effectiveness of these methods, multi-turn response selection task still remains some challenges when modeling the interaction between the utterance context and response: (i) In order to capture the interaction information between a candidate response and multi-turn utterances, most of existing iterative architectures may require deeper or more complex network structure along with the growth of the turns of conversations, which fall short to efficiently learn the multi-turn utterance representations. (ii) Existing methods mainly focus on measuring the semantic relevancy between the response and the given utterance context. Nevertheless, researchers observe that some latent features in the conversations, such as user intent (Wen et al., 2017; Perkins and Yang, 2019; Yang et al., 2020) or conversation topic (Xing et

[*] The work described in this paper is substantially supported by a grant from the Research Grant Council of the Hong Kong Special Administrative Region, China (Project Code: 14200719).

Proceedings of the 28th International Conference on Computational Linguistics, pages 4981–4992
Barcelona, Spain (Online), December 8-13, 2020

al., 2017; Yoon et al., 2018; Yoon et al., 2019), also attach great importance in dialogue systems, which have received little attention in recent multi-turn response selection studies.

In this work, we propose Intra-/Inter-Interaction Network (I^3) with latent interaction modeling to tackle the aforementioned issues. In specific, we adopt hierarchical structure instead of iterative structure to model the multi-level interactions in the multi-turn conversation, including the intra-utterance interaction between the response and each individual utterance, and the inter-utterance interaction among the response and the overall utterance context. Such comprehensive sentence representational learning enables each utterance to be encoded with rich information for mining the latent features. Besides, subspace clustering (Ji et al., 2017; Zhou et al., 2018a; Zhou et al., 2019), which aims to cluster the data into multiple subspaces and find a low-dimensional subspace for each class of data in an unsupervised manner, can be an effective approach to learn the latent feature representations without human-annotated labels. As for dialogue systems, the utterance context and the response can be regarded as two independent views of data (Perkins and Yang, 2019), and it is required to learn the latent representation from both views in a common space to model the coherency of their latent features. Inspired by latest mulit-view subspace clustering studies (Zhu et al., 2019; Zhang et al., 2020a), we propose two kinds of latent multi-view subspace clustering module, namely linear and generalized Latent Multi-view Subspace Clustering (lLMSC and gLMSC), to capture the latent features, which first encode the utterance and the response into view-specific latent representation respectively, and then project them to the same subspaces for multi-view clustering. Finally, we aggregate the three-level interaction information, including the intra-/inter-utterance interaction and latent feature matching information, to comprehensively measure the matching degree between the utterance context and candidate response.

To summarize, the main contributions of this work are as follows: (1) We propose a novel multi-turn response selection model, Intra-/Inter-Interaction Network (I^3), to capture the multi-level matching information by modeling the multi-turn conversations as a hierarchical structure; (2) We develop two kinds of latent multi-view subspace clustering module to model the latent feature coherency between the utterance and response; (3) Experimental results show that the proposed method substantially and consistently outperforms existing state-of-the-art methods on three multi-turn dialogue benchmark datasets.

2 Related Works

Existing methods for building intelligent dialogue systems can be categorized into retrieval-based methods (Lowe et al., 2015; Wu et al., 2017; Zhang et al., 2018), generation-based methods (Xing et al., 2017; Wen et al., 2017) and hybrid methods (Song et al., 2018; Yang et al., 2019). Besides, current studies on conversational systems have evolved from single-turn (Lowe et al., 2015; Kadlec et al., 2015) into multi-turn scenarios (Wu et al., 2017; Zhang et al., 2018). In this work, we focus on retrieval-based methods for multi-turn response selection.

The key to matching the response and the given utterance context is modeling the interaction between them, which is mainly addressed by deep learning models in current studies, like CNN (Kadlec et al., 2015), RNN (Lowe et al., 2015), and hybrid models (Yan et al., 2016). Based on these deep neural networks, some recent works further develop diverse and effective approaches to measure the relevance between the response and the utterances, such as integrating multi-view matching information (Zhou et al., 2016), modeling sequential utterance information (Wu et al., 2017), and refinement and aggregation scheme (Zhang et al., 2018). Inspired by recent progresses of transformer model (Vaswani et al., 2017), latest studies on multi-turn response selection step up to a new stage with carefully designed self-attention-based interaction networks, including deep attention matching network (Zhou et al., 2018b), multi-representation fusion network (Tao et al., 2019a), interaction-over-interaction network (Tao et al., 2019b), and multi-hop selector network (Yuan et al., 2019). In this work, we facilitate the interaction modeling by considering both intra-/inter-utterance interaction with a hierarchical encoder.

Apart from measuring the semantic and contextual relevancy, several efforts have been made on discovering some latent features in the conversations for modeling the intent or topic coherency between the utterance and the response. Yoon et al. (2018) and Yoon et al. (2019) incorporate latent clustering into context-based response/answer selection models to fetch latent topic information. Yang et al. (2018)

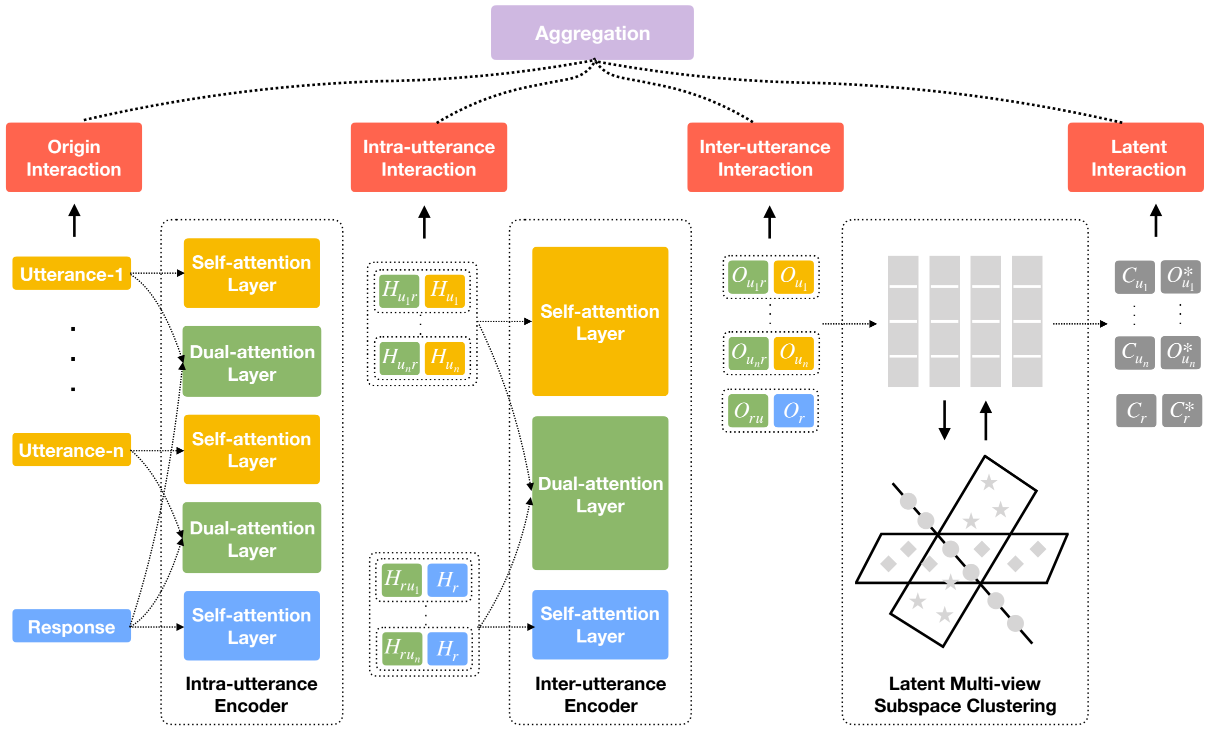

Figure 1: Intra-/Inter-Interaction Network (I^3) with Latent Multi-view Subspace Clustering (LMSC).

and Yang et al. (2020) leverage human-annotated conversational intent labels to model the user intent in information-seeking conversations to help response selection. In this paper, we study latent multi-view subspace clustering to measure the utterance-response coherency consistently in the latent subspace.

3 Method

3.1 Problem Definition

Suppose that there is a conversation data set $\mathbb{D} = \{(U_t, r_t, y_t)\}_{t=1}^{N_D}$, where $U_t = \{u_t^1, u_t^2, ..., u_t^i\}_{i=1}^{N}$ represents a conversation context with u_t^i as the i-th turn utterance in the t-th sample. r_t and y_t are the response candidate and the corresponding label, i.e., whether r_t is an appropriate response given U_t. The goal is to learn a model $g(\cdot)$ with $\mathbb{D}$ to measure the matching degree between U_t and r_t. For simplicity, we omit t in the following notations.

We propose an Intra-/Inter-Interaction Network (I^3) with latent interaction modeling to model $g(\cdot)$. The overview of the proposed model is depicted in Figure 1.

3.2 Intra-/Inter-Interaction Network

3.2.1 Attention Module

Following the former success on multi-turn response selection (Zhou et al., 2018b; Yuan et al., 2019), we employ the Attentive Module proposed by Zhou et al. (2018b) as the basic component of the proposed hierarchical transformer encoder, which is a variant of original transformer block (Vaswani et al., 2017).

The Attention Module is denoted as $\texttt{Attention}(Q, K, V)$, with three input vectors: the query vectors $Q \in \mathbb{R}^{l_q \times d}$, the key vectors $K \in \mathbb{R}^{l_k \times d}$, and the value vectors $V \in \mathbb{R}^{l_v \times d}$, where l_q, l_k, and l_v denote the length of each input and d is the dimension of the embedding. The Attention Module first conducts Scale Dot-Product Attention to apply attention weights upon the value vectors:

$$V_{att} = \textbf{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V. \tag{1}$$

Then, V_{att} and Q are added up together and passed through a layer normalization operation. A feed-forward network (FFN) with ReLU activation is applied upon the normalization result x, and the output of FFN will be residually added to x. Finally, another layer normalization will be applied to obtain the final output:

$$x = \mathbf{LayerNorm}(Q + V_{att}), \tag{2}$$

$$\mathbf{FFN}(x) = W_2\mathbf{ReLU}(W_1x + b_1) + b_2, \tag{3}$$

$$\mathbf{Attention}(Q, K, V) = \mathbf{LayerNorm}(x + \mathbf{FFN}(x)), \tag{4}$$

where W_1, b_1, W_2, b_2 are parameters to be learned.

3.2.2 Intra-utterance Encoder

The intra-utterance encoder is used to encode the individual utterance and response information. The intra-utterance encoder layer in I^3 consists of two kinds of attention module, Self-attention Layer and Dual-attention Layer:

Self-attention Layer is exploited to attend the important word-level information from each individual utterance and response sentence:

$$H_{u_i} = \mathbf{Attention}(E_{u^i}, E_{u^i}, E_{u^i}) \in \mathbb{R}^{L\times d}, \quad H_r = \mathbf{Attention}(E_r, E_r, E_r) \in \mathbb{R}^{L\times d}, \tag{5}$$

where L denotes the length of a sentence, and E_u and E_r are the embeddings of input sequences.

Dual-attention Layer is used to capture the relevant information between each utterance and the response sentence:

$$H_{u_ir} = \mathbf{Attention}(E_{u^i}, E_r, E_r) \in \mathbb{R}^{L\times d}, \quad H_{ru_i} = \mathbf{Attention}(E_r, E_{u^i}, E_{u^i}) \in \mathbb{R}^{L\times d}. \tag{6}$$

3.2.3 Inter-utterance Encoder

The inter-utterance encoder layer is used to learn the overall contextual information across multiple utterance. A mean pooling layer is applied over the local sentence representation for each sentence for obtaining the context sequence:

$$\hat{H}_u = \mathbf{MeanPool}([H_{u_1}, H_{u_2}, ..., H_{u_n}]) \in \mathbb{R}^{N\times d}, \quad \hat{H}_r = \mathbf{MeanPool}([H_r]) \in \mathbb{R}^{1\times d}. \tag{7}$$

The same self-attention and dual-attention layers are applied upon the context sequence $\hat{H}_u$ and $\hat{H}_r$ to capture inter-interaction among utterances and between the utterance context and the response:

$$O_u = \mathbf{Attention}(\hat{H}_u, \hat{H}_u, \hat{H}_u) \in \mathbb{R}^{N\times d}, \quad O_r = \mathbf{Attention}(\hat{H}_r, \hat{H}_r, \hat{H}_r) \in \mathbb{R}^{1\times d}, \tag{8}$$

$$O_{ur} = \mathbf{Attention}(\hat{H}_u, \hat{H}_r, \hat{H}_r) \in \mathbb{R}^{N\times d}, \quad O_{ru} = \mathbf{Attention}(\hat{H}_r, \hat{H}_u, \hat{H}_u) \in \mathbb{R}^{N\times d}, \tag{9}$$

where $O_u = \{o_{u_1}, o_{u_2}, ..., o_{u_n}\}$ and O_r are the self-attentive sentence representations. $O_{ur} = \{o_{u_1r}, o_{u_2r}, ..., o_{u_nr}\}$ and $O_{ru} = \{o_{ru_1}, o_{ru_2}, ..., o_{ru_n}\}$ are the dual-attentive sentence representations.

3.2.4 Intra-/Inter-utterance Interaction Matching

We derive the matching feature by combining dot product and cosine similarity between the utterance and response representations as Zhou et al. (2018b) and Yuan et al. (2019).

The first matching feature matrix M_1 is derived from the original word embeddings of the input utterance U and response r:

$$M_1 = [UA_1r^T : cos(U, r)] \in \mathbb{R}^{N\times 2\times L\times L}, \tag{10}$$

where $A_1 \in \mathbb{R}^{d\times d}$ is a similarity parameter matrix to be learned.

Then, we match the intra-utterance information and inter-utterance information with the candidate response by using the local sentence representations from Section 3.2.2 and the global sentence representations from Section 3.2.3, respectively:

$$M_2 = [H_uA_2{H_r}^T : cos(H_u, H_r), \quad M_3 = [H_{ur}A_3{H_{ru}}^T : cos(H_{ur}, H_{ru})] \in \mathbb{R}^{N\times 2\times L\times L}, \tag{11}$$

$$M_4 = [O_uA_4{O_r}^T : cos(O_u, O_r)], \quad M_5 = [O_{ur}A_5{O_{ru}}^T : cos(O_{ur}, O_{ru})] \in \mathbb{R}^{N\times 2\times N}, \tag{12}$$

where $A_2, A_3, A_4, A_5 \in \mathbb{R}^{d\times d}$ are also similarity parameter matrices to be learned.

3.3 Latent Interaction Modeling

In addition to the intra- and inter-utterance interactions, we develop a latent multi-view subspace clustering approaches for the representational learning of latent features in the dialog content to capture the latent interaction between the utterance context and the candidate response, in which the utterance context and the response are regarded as two different views of dialog content.

3.3.1 Multi-view Latent Representation Learning

Let X_u, X_r denote the inputs of two different views, where $X_u = \{o_{u_i}\}, X_r = \{o_{r_i}\} \in \mathbb{R}^{n \times d_x}$, n and d_x are the number of samples and the dimension of the embedding.

As shown in the Figure 1, we first encode the inputs of each view into the latent representation C_v, where C_v is a common notation of different views, i.e., C_u and C_r, by using a view-specific linear encoder, namely **Linear Multi-view Latent Clustering**. Then the latent representation is self-represented by a self-attentive weighted sum of a common clustering memory matrix across different views:

$$C_v = W_v^{(1)} X_v + b_v^{(1)}, \quad C_v^* = \mathbf{softmax}(C_v Z^T) Z, \tag{13}$$

where $W_v^{(1)}$ and $b_v^{(1)}$ are linear projection parameters to be learned. $Z \in \mathbb{R}^{n_c \times d_c}$ is a common self-representation matrix for all views, and n_c, d_c are the pre-defined number of clusters and the dimension of self-representation matrix, which connects the latent representations C_u and C_r. And C_v^*, i.e., C_u^* and C_r^*, are the clustering representations in the subspace, which are used for measuring the latent feature coherency between the utterance and response. After self-representation operation, the clustering representations are reconstructed by the view-specific decoders:

$$X_v^* = W_v^{(2)} C_v^* + b_v^{(2)}, \tag{14}$$

where $W_v^{(2)}$ and $b_v^{(2)}$ are linear projection parameters to be learned.

The above approach assumes a linear relationship between the latent representation and the features from each view, which also leads to a linear relationship among the features from different views. As one may expect, the relationship among the features from different views is likely to be non-linear, thus, we also study the non-linear situation, namely **Generalized Multi-view Latent Clustering**. The only difference between linear and generalized multi-view latent representation learning is that the generalized form adopts non-linear encoder-decoder in the projection and reconstruction of the latent representation. In this work, we adopt basic Multi-Layer Perceptron (MLP) as the non-linear encoder-decoder:

$$C_v = \mathbf{MLP_1}(X_v), \quad X_v^* = \mathbf{MLP_2}(C_v^*). \tag{15}$$

3.3.2 Latent Interaction

After the multi-view latent representation learning, we obtain the latent clustering representations, i.e., C_u^* and C_r^*, and the reconstructed sentence representations, i.e., O_u^* and O_r^*, which are exploited to match the coherency of latent features between the utterance and response, with the same matching formula as Section 3.2.4:

$$M_6 = [C_u^* A_6 C_r^{*T} : cos(C_u^*, C_r^*)], \quad M_7 = [O_u^* A_7 O_r^{*T} : cos(O_u^*, O_r^*)] \in \mathbb{R}^{N \times 2 \times N}, \tag{16}$$

where $A_6, A_7 \in \mathbb{R}^{d \times d}$ are coherence parameter matrices to be learned.

3.3.3 Loss Function of Multi-view Latent Clustering

The loss function of the multi-view latent representation learning module consists of two parts, information preservation loss and self-representation loss:

$$\mathbb{L}_c = \sum_{v \in \{u,r\}} \alpha_v \left(\underbrace{||X_v^* - X_v||_F^2}_{\text{information preservation}} + \underbrace{\lambda ||C_v^* - C_v||_F^2}_{\text{self-representation}} \right), \tag{17}$$

Dataset (train/dev/test)	Ubuntu	Douban	E-Commerce
#samples	1M/500K/500K	1M/50K/50K	1M/10K/10K
Avg #candidates	2/10/10	2/2/10	2/2/10
Avg #turns	10.1/10.1/10.1	6.7/6.8/6.5	5.5/5.5/5.6
Avg #words	11.4/11.3/11.4	18.6/18.5/20.7	7.0/7.0/7.1

Table 1: Statistics of datasets

where α_v and λ are the hyper-parameters that balance the weight of different views and losses. The information preservation loss ensures that the information from the contextual representation is encoded into the latent representation for each view, while the self-representation loss aims to minimize the differences between the common clustering representation and the view-specific latent representation and alleviate the bias among different views.

3.4 Aggregation and Training

We concatenate the word-level matching matrices together, i.e., $M = [M_1 : M_2 : M_3] \in \mathbb{R}^{N\times 6\times L\times L}$, and extract the corresponding utterance-level features $F_w \in \mathbb{R}^{N\times d_f}$ via a convolutional layer, where d_f is the dimension of the feature size. Then all the utterance-level matching features $F = [F_w : M_4 : M_5 : M_6 : M_7] \in \mathbb{R}^{N\times(d_f+6N)}$ are aggregate by the GRU layer. Finally, the output of GRU is passed through a single-layer perceptron to obtain the matching score $g(U_t, r_t)$.

The overall model is trained to minimize the cross-entropy loss function and the latent multi-view subspace clustering loss function:

$$\mathbb{L}_s = -\sum\nolimits_{t=1}^{N_D} \left[y_t \log g(U_t, r_t) + (1 - y_t) \log\left(1 - g(U_t, r_t)\right)\right], \tag{18}$$

$$\mathbb{L} = \mathbb{L}_s + \mathbb{L}_c. \tag{19}$$

4 Experiment

4.1 Datasets & Evaluation Metrics

We evaluate the proposed method on three multi-turn response selection benchmark datasets, including (1) **Ubuntu** Dialogue Corpus (Lowe et al., 2015) contains multi-turn conversations about technical support issues from the Ubuntu Forum[1], (2) **Douban** Conversation Corpus (Wu et al., 2017) collects conversation content from the Douban group[2] which is a social networking website, and (3) **E-commerce** Dialogue Corpus (Zhang et al., 2018) is a conversation dataset in E-commerce scenario, which is collected from Taobao[3]. The statistics of these datasets are shown in Table 1.

Following previous works (Wu et al., 2017; Zhang et al., 2018; Yuan et al., 2019), we adopt recall at position k in n candidates, i.e, $R_n@k$, as the evaluation metrics. As for Douban dataset, we also adopt MAP (Mean Average Precision), MRR (Mean Reciprocal Rank), and P@1 (Precision@1) for evaluation, since there are more than one ground-truth responses in the Douban Corpus.

4.2 Baseline Models

Single-turn Matching Models: Lowe et al. (2015) and Kadlec et al. (2015) employ RNN, CNN, LSTM, and BiLSTM for response selection tasks by regarding the given context as a whole for matching the candidate responses.

Multi-turn Matching Models: We further separate existing multi-turn matching models into two groups, Pre-transformer Models and Post-transformer Models. **Pre-transformer Models** combine or hybrid RNN and CNN models with carefully designed matching strategies, including DL2R (Yan et al., 2016), Multi-View(Zhou et al., 2016), SMN (Wu et al., 2017), and DUA (Zhang et al., 2018). **Post-transformer Models** leverage improved and adaptive self-attention mechanism to enhance the

[1] https://ubuntuforums.org/

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

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

Model	Ubuntu Corpus			Douban Corpus						E-Commerce Corpus		
	R_{10}@1	R_{10}@2	R_{10}@5	MAP	MRR	P@1	R_{10}@1	R_{10}@2	R_{10}@5	R_{10}@1	R_{10}@2	R_{10}@5
RNN (Lowe et al., 2015)	40.3	54.7	81.9	39.0	42.2	20.8	11.8	22.3	58.9	32.5	46.3	77.5
CNN (Kadlec et al., 2015)	54.9	68.4	89.6	41.7	44.0	22.6	12.1	25.2	64.7	32.8	51.5	79.2
LSTM (Kadlec et al., 2015)	63.8	78.4	94.9	48.5	53.7	32.0	18.7	34.3	72.0	36.5	53.6	82.8
BiLSTM (Kadlec et al., 2015)	63.0	78.0	94.4	47.9	51.4	31.3	18.4	33.0	71.6	35.5	52.5	82.5
DL2R (Yan et al., 2016)	62.6	78.3	94.4	48.8	52.7	33.0	19.3	34.2	70.5	39.9	57.1	84.2
Multi-View (Zhou et al., 2016)	66.2	80.1	95.1	50.5	54.3	34.2	20.2	35.0	72.9	42.1	60.1	86.1
SMN (Wu et al., 2017)	72.6	84.7	96.1	52.9	56.9	39.7	23.3	39.6	72.4	45.3	65.4	88.6
DUA (Zhang et al., 2018)	75.2	86.8	96.2	55.1	59.9	42.1	24.3	42.1	78.0	50.1	70.0	92.1
DAM (Zhou et al., 2018b)	76.7	87.4	96.9	55.0	60.1	42.7	25.4	41.0	75.7	-	-	-
MRFN (Tao et al., 2019a)	78.6	88.6	97.6	57.1	61.7	44.8	27.6	43.5	78.3	-	-	-
IACMN (Wang et al., 2019)	78.2	88.6	97.3	57.1	62.1	44.8	26.9	45.3	78.3	-	-	-
IoI (Tao et al., 2019b)	79.6	89.4	97.4	57.3	62.1	44.4	26.9	45.1	78.6	56.3	76.8	95.0
MSN (Yuan et al., 2019)	80.0	89.9	97.8	58.7	63.2	47.0	29.5	45.2	78.8	60.6	77.0	93.7
I^3	80.1	89.9	97.8	58.7	63.3	46.7	29.1	46.0	79.5	61.0	78.7	95.1
I^3-lLMSC	80.2	**90.1**	**97.8**	**59.2**	**64.0**	**47.9**	**29.8**	45.4	**80.3**	**62.0**	**80.0**	94.9
I^3-gLMSC	**80.6**	**90.1**	**97.8**	58.7	63.4	46.7	28.5	**46.4**	79.7	61.1	79.6	**95.6**

Table 2: Experimental results

interaction between the utterance and response during the representational learning process, including DAM (Zhou et al., 2018b), MRFN (Tao et al., 2019a), IACMN (Wang et al., 2019), IoI (Tao et al., 2019b), and MSN (Yuan et al., 2019).

4.3 Implementation Details

For a fair comparison, we follow previous works (Wu et al., 2017; Zhou et al., 2018b; Yuan et al., 2019) to adopt Word2Vec (Mikolov et al., 2013) word embedings with the dimension of 200, which is pre-trained on the training data without extra materials for pre-training. For the hyper-parameters settings of I^3, the number of all attention layers is set to be 1. In the aggregation, three 2-D convolutional layers are used to extract matching features with 16 [3,3], 32 [3,3], and 64 [3,3] filters, respectively. The dimension of the hidden states in GRU is set to be 300. In LMSC module, we observe similar performances when varying the number of clusters and the weights of different view of clustering. Thus, the number of clusters is fixed to be 10. α_v and λ are also set to 1. Specifically for gLMSC, the encoder-decoder MLPs are two-layer and the hidden size of them is set to be 300. The maximum length of sentence and the maximum number of utterance turns are set to be 50 and 10. The learning rate and the dropout rate are set to be 0.001 and 0.2, and all datasets are trained on a mini-batch of 200.

4.4 Results

Table 2 presents the evaluation results over different methods on three datasets. Obviously, multi-turn methods outperform single-turn methods to a large margin, and it is needless to emphasize the necessity of multi-turn response selection studies. Compared with pre-transformer methods, post-transformer methods have a better performance on multi-turn response selection, which demonstrates the effectiveness of self-attention mechanism on capture the interaction between texts.

As for the proposed models, we observe that the basic I^3 model achieves state-of-the-art performance on 10 out of 12 metrics. More importantly, different with latest iterative interaction based models, i.e., IACMN, IoI, and MSN, the depth of network for I^3 will be fixed and not be affected by the growth of conversation turns. As is reported in their works, IoI achieves the best performance on these datasets with 7 times of iterative interaction blocks, and MSN with 3-hops selector. In another word, I^3 can decently achieve competitive or even better performance with a single layer of interaction, regardless of various number of conversation turns. In addition, by adding the latent multi-view subspace clustering modules, I^3-lMVLC and I^3-gMVLC further improve the performance with a noticeable margin. For instance, there is an additional improvement of about 1% on E-commerce Corpus by adding the lLMSC module. By comparing lLMSC and gLMSC, we observe that these two kinds of LMSC modules perform

Model	Douban Corpus						E-Commerce Corpus			Average Scores
	MAP	MRR	P@1	R_{10}@1	R_{10}@2	R_{10}@5	R_{10}@1	R_{10}@2	R_{10}@5	
I^3	58.7	63.3	46.7	29.1	46.0	79.5	61.0	78.7	95.1	66.1
w/o local transformer	56.8	61.9	44.9	27.4	44.2	76.9	53.5	72.6	92.4	62.4
w/o global transformer	57.9	62.7	45.5	27.8	45.2	78.9	53.6	73.5	92.3	63.1
w/o self-attention	57.7	62.3	44.9	27.3	45.6	78.9	54.1	73.5	93.6	63.3
w/o dual-attention	57.7	62.3	44.8	27.4	45.5	79.7	54.6	74.6	93.2	63.5
w/o origin-interaction	58.3	63.0	46.6	28.8	45.0	79.1	58.9	76.5	93.7	64.9
w/o intra-interaction	57.5	62.5	45.4	27.5	45.1	79.1	53.7	72.5	93.3	63.0
w/ LTC	58.6	63.1	47.3	29.2	46.0	79.3	60.7	78.6	94.7	66.0
w/ LC	58.3	63.5	47.2	28.9	45.0	79.6	59.4	78.4	94.7	65.6
w/ lLMSC	**59.2**	**64.0**	**47.9**	**29.8**	45.4	**80.3**	**62.0**	**80.0**	94.9	**66.7**
w/ gLMSC	58.7	63.4	46.7	28.5	**46.4**	79.7	61.1	79.6	**95.6**	66.3

Table 3: Ablation study and comparisons of clustering strategies

differently on different datasets. This situation is common in clustering methods (Zhang et al., 2020a), as it is difficult to determine whether the relationship among different samples is linear or non-linear.

5 Discussion

5.1 Ablation Study

In order to validate the effectiveness of different modules in the proposed I^3 network, we conduct several ablation studies on Douban Corpus and E-commerce Corpus in terms of discarding different components. Apart from the original metrics, we also report the Average Scores, which is the mean of all the scores in two datasets, to integrally observe the difference. As is presented in the first part of Table 3, there are several notable observations: (i) As for the hierarchical transformer encoder, both local and global transformer contribute to the final performance to a large extent, which validates the effectiveness of encoding multi-level utterance information. (ii) By leaving only self-attention or dual-attention as the functional module in hierarchical transformer, we observe that these two kinds of attention modules guarantee the superiority of the performance. (iii) Under the matching-aggregate framework, we also evaluate the contribution of each matching feature. Note that we omit the "w/o inter-interaction" result, since it will be the same model as "w/o global transformer". From the results, we observe that origin-interaction contributes far less than the other two matching features.

5.2 Comparison on Latent Clustering Strategy

We compare the proposed lLMSC and gLMSC module with other two latent clustering modules proposed for response/answer selection, including LTC (Yoon et al., 2018) and LC (Yoon et al., 2019). LTC (Yoon et al., 2018) is a latent topic clustering module to extract semantic information from target samples, which only clusters the information from the view of response. LC (Yoon et al., 2019) further applies the latent topic clustering module for both question and answer separately. Different from these two strategies, LMSC not only projects both utterances and the response into the same subspace for coherence measurement, but also applies specific loss functions to control the information preservation during the clustering process. The results are presented in the second part of Table 3. Despite the improvement on some of the metrics, there is not much difference on the overall performance for these two clustering strategy. However, lLMSC and gLMSC effectively improve the overall performance.

5.3 Case Study of Latent Subspace Clustering

The Latent Multi-view Subspace Clustering module is proposed to extract latent features for measuring the coherency between the utterance and response. To facilitate further investigation of the latent subspace clustering, we derive the probability of words in each cluster, and rank by their frequency. After filtering the stop words, the results of clusters and words for E-commerce Corpus are presented in Table 4. Note that the category of cluster is conjectured from the cluster results, since there is no ground-truth label for the latent cluster. From the clustering result, we observe an obvious inclination for each cluster.

	Cluster 1	**Cluster 2**	**Cluster 3**	**Cluster 4**	...
Ubuntu Corpus	**Solution**	**App**	**Network**	**Disk**	...
	Googling, Solution, Solve, Searching, Link, Forums, Explain	Emerald, Metacity, VLC, Compiz, Applet, Plugin, Firefox	Nameserver, Addr, Inet, IP, Subnet, Hosts, TCP, Localhost	RAID, FSCK, SATA, BIOS, Memory, Kernel, Disk, Drives	
E-commerce Corpus	**Payment**	**Refund**	**Discount**	**Free shipping**	...
	Delivery, EMS, Change price, Payment, Package, Order, Check, Default	Contact, Refund, Shipping fee, Apply, Repost, Reject, Return, Price difference	Discount, Gift, Coupon, Price, Activity, Cheap, Free shipping, Cost-effective	Address, Free shipping, Delivery, Ningxia, Tibet, Qinghai, Xinjiang, Hainan	

Table 4: Top ranked words for each cluster on Ubuntu and E-commerce Corpus

For example, in the E-commerce Corpus, the conversation topics are clustered into different groups, such as free shipping, refund, payment, discount, etc. One one hand, latent multi-view subspace clustering can assist the measurement of the latent representation coherency, leading to a better utterance-response matching result. On the other hand, such clustering approach provides an unsupervised way to discover the latent features of the dialogue.

5.4 Error Analysis

To better understand the failure modes of the proposed methods, we analyze 100 failure cases, and find the error cases could be classified into the following categories for later further improvement.

Information Imbalance ($\approx 45\%$): Some conversation samples suffer a great imbalance on the provided information from the utterance context and the response, leading to the difficulties in matching the utterance and response. Among them, about 70% of them give a short and simple response, such as "Sure.", "I see.", etc. While the rest only provide little information in the utterance context, for which even human cannot determine the true response. One possible way to address this kind of failures is to introduce background information to balance the information from both the utterance and the response.

Mislabeling or Misspelling ($\approx 25\%$): We attribute these failures to the data issues. For instance, the ground-truth response is "Yes, we will.", while there are some negative candidates that contains both the true response but also some extra information, like "Yes, we will address it as soon as possible.". However, this negative candidate is also supposed to be a good or better response to the given utterances. Besides, some ground-truth responses are misspelled.

Inconsistency of Fact ($\approx 20\%$): There are some conversations concerning factoid issues, such as the date, the place, the size, etc. However, the proposed method lacks of the ability to verify whether the information provide in the response is fact of not. To address the problem of the inconsistency of fact in the response, it would be better to incorporate some supporting knowledge (Deng et al., 2018) and consider the interrelationship (Zhang et al., 2020b) among all the candidate responses.

Multiple Intents/Topics ($\approx 10\%$): Compared with the error analysis provided in Zhang et al. (2018), the issues related to user intent or conversation topic have been alleviated to a great extent. However, there are still some cases involved multiple intents/topics remaining to be tackled by further studying the latent clustering representational learning.

6 Conclusion

In this work, we propose Intra-/Inter-Interaction Network (I^3) with latent interaction modeling for multi-turn response selection. We propose a hierarchical transformer encoder to capture the intra- and inter-utterance interaction with the candidate response from both individual utterance and the overall utterance context. Besides, we develop a latent multi-view subspace clustering module to model the latent feature coherency between the utterance and response. Experimental results show that the proposed method substantially and consistently outperforms existing SOTA methods on three benchmark datasets.

References

Yang Deng, Ying Shen, Min Yang, Yaliang Li, Nan Du, Wei Fan, and Kai Lei. 2018. Knowledge as A bridge: Improving cross-domain answer selection with external knowledge. In *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*, pages 3295–3305.

Pan Ji, Tong Zhang, Hongdong Li, Mathieu Salzmann, and Ian D. Reid. 2017. Deep subspace clustering networks. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 24–33.

Rudolf Kadlec, Martin Schmid, and Jan Kleindienst. 2015. Improved deep learning baselines for ubuntu corpus dialogs. *CoRR*, abs/1510.03753.

Feng-Lin Li, Minghui Qiu, Haiqing Chen, Xiongwei Wang, Xing Gao, Jun Huang, Juwei Ren, Zhongzhou Zhao, Weipeng Zhao, Lei Wang, Guwei Jin, and Wei Chu. 2017. *AliMe Assist* : An intelligent assistant for creating an innovative e-commerce experience. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, CIKM 2017, Singapore, November 06 - 10, 2017*, pages 2495–2498.

Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. 2015. The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. In *Proceedings of the SIGDIAL 2015 Conference, The 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue, 2-4 September 2015, Prague, Czech Republic*, pages 285–294.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pages 3111–3119.

Hugh Perkins and Yi Yang. 2019. Dialog intent induction with deep multi-view clustering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 4014–4023.

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

Yiping Song, Cheng-Te Li, Jian-Yun Nie, Ming Zhang, Dongyan Zhao, and Rui Yan. 2018. An ensemble of retrieval-based and generation-based human-computer conversation systems. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, pages 4382–4388.

Chongyang Tao, Wei Wu, Can Xu, Wenpeng Hu, Dongyan Zhao, and Rui Yan. 2019a. Multi-representation fusion network for multi-turn response selection in retrieval-based chatbots. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining, WSDM 2019, Melbourne, VIC, Australia, February 11-15, 2019*, pages 267–275.

Chongyang Tao, Wei Wu, Can Xu, Wenpeng Hu, Dongyan Zhao, and Rui Yan. 2019b. One time of interaction may not be enough: Go deep with an interaction-over-interaction network for response selection in dialogues. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 1–11.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 5998–6008.

Heyuan Wang, Ziyi Wu, and Junyu Chen. 2019. Multi-turn response selection in retrieval-based chatbots with iterated attentive convolution matching network. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management, CIKM 2019, Beijing, China, November 3-7, 2019*, pages 1081–1090.

Tsung-Hsien Wen, Yishu Miao, Phil Blunsom, and Steve J. Young. 2017. Latent intention dialogue models. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, pages 3732–3741.

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, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, 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 *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 3351–3357.

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, SIGIR 2016, Pisa, Italy, July 17-21, 2016*, pages 55–64.

Liu Yang, Minghui Qiu, Chen Qu, Jiafeng Guo, Yongfeng Zhang, W. Bruce Croft, Jun Huang, and Haiqing Chen. 2018. Response ranking with deep matching networks and external knowledge in information-seeking conversation systems. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018, Ann Arbor, MI, USA, July 08-12, 2018*, pages 245–254.

Liu Yang, Junjie Hu, Minghui Qiu, Chen Qu, Jianfeng Gao, W. Bruce Croft, Xiaodong Liu, Yelong Shen, and Jingjing Liu. 2019. A hybrid retrieval-generation neural conversation model. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management, CIKM 2019, Beijing, China, November 3-7, 2019*, pages 1341–1350.

Liu Yang, Minghui Qiu, Chen Qu, Cen Chen, Jiafeng Guo, Yongfeng Zhang, W. Bruce Croft, and Haiqing Chen. 2020. IART: intent-aware response ranking with transformers in information-seeking conversation systems. *CoRR*, abs/2002.00571.

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, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 1575–1584.

Seunghyun Yoon, Franck Dernoncourt, Doo Soon Kim, Trung Bui, and Kyomin Jung. 2019. A compare-aggregate model with latent clustering for answer selection. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management, CIKM 2019, Beijing, China, November 3-7, 2019*, pages 2093–2096.

Chunyuan Yuan, Wei Zhou, Mingming Li, Shangwen Lv, Fuqing Zhu, Jizhong Han, and Songlin Hu. 2019. Multi-hop selector network for multi-turn response selection in retrieval-based chatbots. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 111–120.

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, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*, pages 3740–3752.

Changqing Zhang, Huazhu Fu, Qinghua Hu, Xiaochun Cao, Yuan Xie, Dacheng Tao, and Dong Xu. 2020a. Generalized latent multi-view subspace clustering. *IEEE Trans. Pattern Anal. Mach. Intell.*, 42(1):86–99.

Wenxuan Zhang, Yang Deng, and Wai Lam. 2020b. Answer ranking for product-related questions via multiple semantic relations modeling. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*, pages 569–578.

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, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 372–381.

Pan Zhou, Yunqing Hou, and Jiashi Feng. 2018a. Deep adversarial subspace clustering. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 1596–1604.

Xiangyang Zhou, Lu Li, Daxiang Dong, Yi Liu, Ying Chen, Wayne Xin Zhao, Dianhai Yu, and Hua Wu. 2018b. Multi-turn response selection for chatbots with deep attention matching network. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 1118–1127.

Lei Zhou, Xiao Bai, Dong Wang, Xianglong Liu, Jun Zhou, and Edwin R. Hancock. 2019. Latent distribution preserving deep subspace clustering. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 4440–4446.

Pengfei Zhu, Binyuan Hui, Changqing Zhang, Dawei Du, Longyin Wen, and Qinghua Hu. 2019. Multi-view deep subspace clustering networks. *CoRR*, abs/1908.01978.