How to Represent Context Better? An Empirical Study on Context Modeling for Multi-turn Response Selection

Jiazhan Feng[1,2], **Chongyang Tao**[1], **Chang Liu**[1,3], **Rui Yan**[4] and **Dongyan Zhao**[1,3,5,6*]
[1]Wangxuan Institute of Computer Technology, Peking University
[2]School of Intelligence Science and Technology, Peking University
[3]Center for Data Science, Peking University
[4]Gaoling School of Artificial Intelligence, Renmin University of China
[5]Beijing Institute for General Artificial Intelligence
[6]State Key Laboratory of Media Convergence Production Technology and Systems
{fengjiazhan,chongyangtao,liuchang97,zhaody}@pku.edu.cn
ruiyan@ruc.edu.cn

Abstract

Building retrieval-based dialogue models that can predict appropriate responses based on the understanding of multi-turn context messages is a challenging problem. Early models usually concatenate all utterances or independently encode each dialogue turn, which may lead to an inadequate understanding of dialogue status. Although a few researchers have noticed the importance of context modeling in multi-turn response prediction, there is no systematic comparison to analyze how to model context effectively and no framework to unify those methods. In this paper, instead of configuring new architectures, we investigate how to improve existing models with a better context modeling method. Specifically, we heuristically summarize three categories of turn-aware context modeling strategies which model the context messages from the perspective of sequential relationship, local relationship, and query-aware manner respectively. A Turn-Aware Context Modeling (TACM) layer is explored to flexibly adapt and unify these context modeling strategies to several advanced response selection models. Evaluation results on three public data sets indicate that employing each individual context modeling strategy or multiple strategies can consistently improve the performance of existing models.

1 Introduction

Recently, building a dialogue system for open domain human-machine conversation is attracting more and more attention due to both availability of large-scale human conversation data and powerful models learned with neural networks. Existing work on building a conversational system includes generation-based methods and retrieval-based methods. A generation-based model directly synthesizes a response with a natural language generation method (Shang et al., 2015; Serban et al., 2016), while a retrieval-based model replies to a human input by selecting a proper response from a pre-built index (Lowe et al., 2015; Humeau et al., 2019). In this work, we study the problem of multi-turn response selection for retrieval-based dialogues, since retrieval-based systems are superior in terms of response fluency and informativeness, and play an important role in industrial products.

Real-world dialogues usually comprise multiple turns, where a retrieval model should select the most proper response by measuring the matching degree between multi-turn dialogue context and a number of response candidates. The key problem is how to make better use of multi-turn context information. Currently, there emerge two lines of research to represent the multi-turn dialogue context. One is to model each turn of utterance individually first and then aggregate a sequence of utterance-response matching features to get a final score (Wu et al., 2017; Zhou et al., 2018; Gu et al., 2019; Yang et al., 2018, 2020; Tao et al., 2019b), which are known as the *representation-matching-aggregation* paradigm. The other line is to concatenate all turns of utterances into a long sequence first and make them fully interact with each other by RNNs (Lowe et al., 2015; Zhou et al., 2016; Chen and Wang, 2019) or transformer layers (Humeau et al., 2019; Whang et al., 2020; Gu et al., 2020). In particular, recently models based on pre-trained language models (PLMs) such as BERT or SA-BERT (Gu et al., 2020) conduct multi-turn context modeling and response matching in a unified process.

These mainstream methods, including fully concatenating all utterances or independently encoding each dialogue turn, equally represent the information of each dialogue element and ignore the characteristics of multi-turn dialogue context, which may

*Corresponding author: Dongyan Zhao.

Findings of the Association for Computational Linguistics: EMNLP 2022, pages 7285–7298

lead to sub-optimal context representations and response matching features. Recently, researchers have begun to notice the importance of explicitly modeling the multi-turn dialogue context based on the characteristics of multi-turn dialogue context, including using of natural sequential relationship between dialogue turns (Zhou et al., 2016) or using the last turn of dialogue context to guide the modeling process of previous turns (Zhang et al., 2018; Yuan et al., 2019). However, there is no systematic comparison to analyze how to effectively model the multi-turn dialogue context considering characteristics of dialogue model and no framework to unify those methods for retrieval-based dialogues.

In this paper, instead of configuring new architectures, we investigate how to improve the performance of existing matching models with better context modeling methods. Following this idea, we heuristically summarize three categories of turn-aware context modeling strategies which model context messages from the perspective of sequential relationship, local relationship, and query-aware manner respectively. To compare those methods, we apply them on several representative response selection models through a Turn-Aware Context Modeling (TACM) layer, which allows different context modeling strategies to be flexibly applied to dialogue models.

To verify the effectiveness of the framework, we choose three representative multi-turn response selection models as our matching models, and conduct experiments on three public data sets including Ubuntu Dialogue Corpus (Lowe et al., 2015), Douban Conversation Corpus (Wu et al., 2017), and E-Commerce Dialogue Corpus (Zhang et al., 2018). Based on a series of experiments, we find query-aware context modeling is the best strategy and employing multiple context modeling strategies can consistently improve the performance of response selection. Besides, we also observe that our TACM layer can improve the capability of modeling long context. We hope our empirical comparison can shed light on future research on this line of work. Our contributions in this paper are four-fold:

- Three categories of turn-aware context modeling strategies inspired by inherent characteristics of multi-turn dialogues are summarized;
- A TACM layer is explored to flexibly adapt and unify these context modeling strategies to the advanced response selection models;
- A systematic comparison of different context modeling strategies and their combinations with representative response selection models on three benchmarks;
- Consistent improvements are brought to various response matching models without involving heavy-machinery, and are easy to generalize to downstream dialogue applications.

2 Related Works

Retrieval-based models design a discriminative model to measure the matching degree between a human input and a response candidate for response selection. Early studies mainly focus on single-turn context-response matching (Wang et al., 2013; Hu et al., 2014; Wang et al., 2015). Recently, researcher have devoted themselves to the multi-turn scenario. Several methods concatenate all turns of utterances into a long sequence first and then make them fully interact with each other by RNNs (Lowe et al., 2015; Zhou et al., 2016; Chen and Wang, 2019) or transformer layers (Humeau et al., 2019; Gu et al., 2020). In addition to these methods, some researchers construct dialogue models with a *representation-matching-aggregation* paradigm. Such approaches encode each turn of utterance individually first and then aggregate a sequence of utterance-response matching features to get a final score. Representative methods including sequential matching network (SMN) (Wu et al., 2017), deep attention matching network (DAM) (Zhou et al., 2018) and multi-hop selector network (MSN) (Yuan et al., 2019).

As an important problem in dialogue systems, multi-turn context modeling has raised great interests in recent years. Especially for generation-based methods, various models adopt hierarchical encoder-decoder framework to model all context sentences (Serban et al., 2016, 2017; Xing et al., 2018; Chen et al., 2018). Tian et al. (2017) compare various methods to get a global representation for the context. Zhang et al. (2019) propose ReCoSa where attention weights between each context and response representations are computed and used in further decoding process. The problem is less explored in existing retrieval-based methods. Zhang et al. (2018) concatenate the last utterance to other turns, and then use Gated Self Attention to obtain utterance representation. Yuan et al. (2019) use multi-hop selectors to select useful information in dialogue history. These methods only consider

modeling one type of characteristic of multi-turn dialogue context. Besides, there is no systematic comparison to analyze how to model context effectively and no framework to unify those methods. Therefore, we consider exploring how to improve the existing models with a better context modeling method in this paper. Specifically, we summarize three categories of turn-aware context modeling strategies and conduct an empirical study on context modeling for multi-turn response selection.

3 Methodology

3.1 Problem Formalization

Given a data set $\mathcal{D} = \{(y, c, r)_z\}_{z=1}^{N}$ where $c = \{u_1, ..., u_{n_c}\}$ represents a n_c turns of conversation context with u_i the i-th turn, r is a response candidate, and $y \in \{0, 1\}$ denotes a label with $y = 1$ indicating r a proper response for c and otherwise $y = 0$. The goal of response selection is to learn a matching model $s(\cdot, \cdot)$ from $\mathcal{D}$. For any context-response pair (c, r), $s(c, r)$ gives a score that reflects the matching degree between c and r. According to $s(c, r)$, one can rank a set of candidates for response selection.

3.2 Matching with Turn-Aware Context Representation

Most of the representative context-response matching models follow a *representation-matching-aggregation* paradigm (Wu et al., 2017; Zhang et al., 2018; Zhou et al., 2018; Tao et al., 2019a; Wang et al., 2019; Yuan et al., 2019). The framework consists of (1) a *representation* layer to explicitly encode the utterance at each turn individually based on its word-level representations, where each utterance does not explicitly receive the contextual information from other turns of utterances, (2) a *matching* layer that lets the context and response interact based on their representations, (3) an *aggregation* layer incorporating the interaction features.

The idea of Turn-Aware Context Modeling (TACM) is to embed a new layer before the *representation* layer of a specific response selection model, where various modeling strategies are designed to make each utterance interact with other turns, so that the subsequent individually encoding of each utterance can be aware of the important contextual dialogue information from other utterances in the same session. It should be noted that we mainly focus on TACM layer in this paper, so the definition of the matching architecture follows

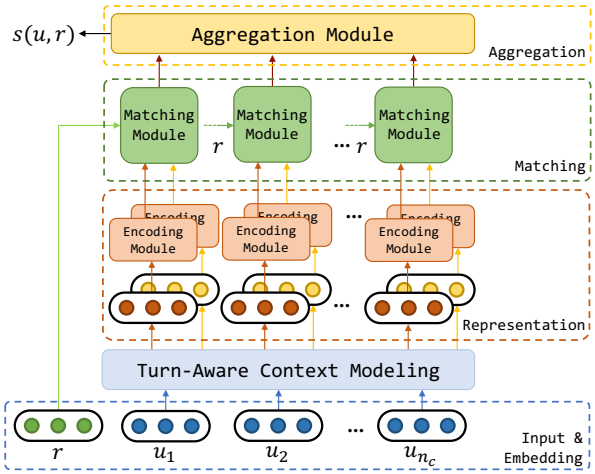


Figure 1: The framework architecture for matching with Turn-Aware Context Representation.

advanced architectures such as SMN, DAM and MSN. To facilitate clarification of our methods, we also depict a diagram of the turn-aware context modeling framework in Figure 1.

For each utterance $u_i = [w_{u_i,k}]_{k=1}^{n_{u_i}}$ in a context and its response candidate $r = [w_{r,k}]_{k=1}^{n_r}$, where n_{u_i} and n_r are the number of words in u_i and r respectively, we first represent u_i and r as sequences of word embeddings, namely $U_i^e = [e_{u_i,1}, e_{u_i,2}, ..., e_{u_i,n_{u_i}}]$ and $R^e = [e_{r,1}, e_{r,2}, ..., e_{r,n_r}]$, where $e \in \mathbb{R}^d$ denotes a d-dimension word embedding.

Then, we propose a Turn-Aware Context Modeling (TACM) layer which takes in the word embeddings of all turns of utterances $[U_i^e]_{i=1}^{n_c}$, and then models through several turns so that each utterance is fully interacted with other turns of utterances. Through different categories of TACM modules, each utterance representation can absorb contextual information from other turns of utterances in different semantic aspects. Suppose we have K sorts of TACM modules, the computation results of k-th module can be formalized as $\tilde{U}_i^k = \phi^k(U_i^e), \tilde{U}_i^k \in \mathbb{R}^{n_{u_i} \times d}$, where $\phi^k(\cdot)$ denotes the k-th TACM strategy.

Now, we can obtain a set of dialogue context representations as $\{\tilde{U}_i^k\}_{k=1}^{K}$, which serve as the input of the *representation-matching-aggregation* paradigm. Consistently with these models, $\forall i \in \{1, ..., n_c\}, \tilde{U}_i^k$ is encoded individually and achieves **I**ntra-**U**tterance **R**epresentation (IUR) as $\hat{U}_i^k = f_{\text{IUR}}(\tilde{U}_i^k)$ where $f_{\text{IUR}}(\cdot)$ stands for the representation function which can be a RNN (Wu et al., 2017), a self-attention module (Zhou et al., 2018)

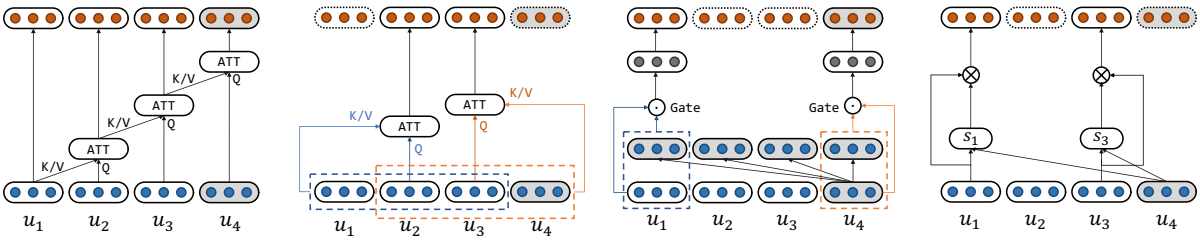


Figure 2: Sketches of four types of turn-aware context modeling strategies. Q, K and V denote the query sentence, the key sentence and the value sentence respectively. $u_{\{1,2,3,4\}}$ represent the utterances in the context in chronological order. For convenience, we only draw four turns of utterances. In other words, u_4 is the most recent turn, also referred as the query utterance. For the rolling representation, we only draw the forward rolling process.

or even a fusion network of multiple types of representation functions (Tao et al., 2019a). Similarly, R^e can also be processed into $\hat{R}$ as $\hat{R} = f_{\texttt{IUR}}(R^e)$.

Then, an **U**tterance-**R**esponse **M**atching (URM) layer follows, where $\forall k \in \{1, ..., K\}$, $\hat{U}_i^k$ interacts with $\hat{R}$ and finally matched into matrices with several channels which can be formalized as $M_i = f_{\texttt{URM}}(\{\hat{U}_i^k\}_{k=1}^K, \hat{R})$, where $f_{\texttt{URM}}(\cdot,\cdot)$ represents the matching function, which can be a similarity function or an attention-based interaction function (Tao et al., 2019a), $M_i \in \mathbb{R}^{(K*n_M)\times n_{u_i}\times n_r}$ with n_M the number of channels of the matching matrices in the original response selection models[1].

Finally, an **Agg**regation (AGG) layer is employed to fuse or aggregate $\{M_i\}_{i=1}^{n_c}$ defined as $\hat{y} = f_{\texttt{AGG}}(\{M_i\}_{i=1}^{n_c})$, where $\hat{y}$ is final logits. This aggregation process $f_{\texttt{AGG}}(\cdot)$ also depends on specific multi-turn response selection model, which may be a 3D convolutional neural network (Zhou et al., 2018) or a 2D convolutional neural network followed by a recurrent neural network (Wu et al., 2017; Yuan et al., 2019) to model the dependencies among different turns on interaction features, instead of the utterance representations.

3.3 Turn-Aware Context Modeling Strategies

We consider the following three categories of strategies that cover sequential context modeling, local context modeling and query-aware context modeling, which are depicted succinctly in Figure 2.

Sequential Context Modeling: Due to the natural sequential relationship between dialogue turns, understanding of subsequent turn of utterance requires dialogue information flow[2]. Therefore, we propose a *"Rolling"* strategy to model such temporal relationships and directly connecting representations from previous or following turns of utterances into the current utterance, which is similar to a hierarchical transformer-based architecture. It can capture intra-sentence and inter-sentence connections in a structured and dynamic sequential manner.

As shown in Figure 2(a), at turn $i \in \{1, ..., n_c\}$, we compute the rolling representation as:

$$\begin{aligned}
\overrightarrow{U_i^S} &= f_{\texttt{ATT}}(U_i^e, \overrightarrow{U_{i-1}^S}, \overrightarrow{U_{i-1}^S}) \\
\overleftarrow{U_i^S} &= f_{\texttt{ATT}}(U_i^e, \overleftarrow{U_{i+1}^S}, \overleftarrow{U_{i+1}^S}) \\
\overrightarrow{U_1^S} &= U_1^e, \quad \overleftarrow{U_{n_c}^S} = U_{n_c}^e \\
U_i^S &= \texttt{ReLU}([\overrightarrow{U_i^S}; \overleftarrow{U_i^S}] \cdot W_s + b_s)
\end{aligned} \tag{1}$$

where $f_{\texttt{ATT}}(Q, K, V)$ is a transformer layer (Vaswani et al., 2017), which takes in the query sentence Q, key sentence K and value sentence V. W_s, b_s are learnt parameters, $[\cdot;\cdot]$ is a concatenation operation, $\overrightarrow{U_i^S}$ and $\overleftarrow{U_i^S}$ denote the forward and backward rolling representations respectively. U_i^S has the same dimensions as U_i^e.

Local Context Modeling: Another viewpoint is that in most dialogue context, data display a great deal of locality of reference. In this phenomenon, a large amount of information about an utterance can be derived from its neighboring turns. For example, on account of the possibility of topic shifts in dialogues, adjacent turns of utterances may have relevant conversation content. In summary, to capture these local structures in the context, we empirically put forward a *"Window"* strategy to capture local context features under a sliding window attention as demonstrated in Figure 2(b).

$\forall i \in \{1, ..., n_c\}$, the window representation of i-th utterance is calculated by $f_{\texttt{ATT}}$ between the current utterance and local context. The window

[1]Here, n_M is 2 in SMN, 12 in DAM, and 3 in MSN.

[2]We consider the information flow from both directions.

representation and local context is defined as:

$$U_i^L = f_{\texttt{ATT}}(U_i^e, \mathcal{C}_i, \mathcal{C}_i)$$
$$\mathcal{C}_i = [U_\alpha^e; ...; U_{i-1}^e; U_i^e; U_{i+1}^e; ...; U_\beta^e] \quad (2)$$

where $\alpha = \texttt{max}(1, i-\gamma), \beta = \texttt{min}(n_c, i+\gamma)$ denote both sides of the window respectively and γ is the offset, which is a hyper-parameter to be tuned by us. In this equation, $\mathcal{C}_i \in \mathbb{R}^{(\sum_{j=\alpha}^{\beta} n_{u_j}) \times d}$ is the concatenation of all utterances around the current i-th utterance.

Query-Aware Context Modeling: Apart from the above two strategies, we also grasp the importance of the last turn of utterance u_{n_c} which is often considered as a dialogue query, since most of the response candidates are directly respond to it[3]. For this purpose, we intuitively utilize the dialogue query to capture relevant utterance information in conversation history. Towards measuring the necessity of each turn of utterances to replenish u_{n_c}, we propose two different strategies named *"Highway"* and *"Weighted"* illustrated in Figure 2(c,d). Both of them can identify important utterances and capture the implicit relationship of the whole context. The difference is that the former calculates a specific weight for each entry of the representation vector (namely word-level), while the latter only assigns a weight to utterance-level representation.

For *"Highway"* representation, firstly, $\forall i \in \{1, ..., n_c\}$, i-th turn of utterance U_i^e is concatenated with the dialogue query $U_{n_c}^e$ and obtain a concatenated representation $U_i^c = [U_i^e; U_{n_c}^e]$ where $U_i^c \in \mathbb{R}^{(n_{u_i}+n_{u_{n_c}}) \times d}$. Then, inspired by Srivastava et al. (2015), this concatenated representation U_i^c and the current utterance representation U_i^e are then fed into a Highway Network to fuse both features, which is processed as follows:

$$o_i = \sigma(U_i^e \cdot W_g + b_g)$$
$$U_i^r = \texttt{GELU}(U_i^c \cdot W_r + b_r) \quad (3)$$
$$U_i^H = o_i \odot U_i^r + (1 - o_i) \odot U_i^e$$

where W_g, W_r, b_g, b_r are learnt parameters and $\odot$ denotes the element-wise multiplication. GELU (Hendrycks and Gimpel, 2016) is an activation function. The gating unit $o_i \in \mathbb{R}^{n_{u_i} \times d}$ is learnt to regulate the flow of query-aware information U_i^r. Similarly, U_i^H and U_i^e have the same dimensions.

For *"Weighted"* representation, we first integrate each turn of dialogue information into an utterance-level representation vector $\bar{U}_i$. Then, we calculate the semantic similarity between each turn and dialogue query through cosine similarity to obtain the relevant score of each utterance. Finally, all turns of utterances are weighted by relevant score to get a new weighted representation. At turn $i \in \{1, ..., n_c\}$, we can formulate this procedure as:

$$\bar{U}_i = \texttt{MEAN}(f_{\texttt{ATT}}(U_i^e, U_i^e, U_i^e))$$
$$s_i = \frac{\bar{U}_i \cdot \bar{U}_{n_c}^\top}{\|\bar{U}_i\|_2 \cdot \|\bar{U}_{n_c}\|_2} \quad (4)$$
$$U_i^W = s_i * U_i^e$$

where MEAN($\cdot$) represents mean pooling operation over self-attended word embeddings, s_i is the weight scalar for the i-th turn[4]. Darker area means larger value. The weighted representation U_i^W has the same dimension as U_i^e.

4 Experiments

4.1 Datasets and Evaluation Metrics

We evaluate our methods on three public data sets: Ubuntu Dialogue Corpus (Lowe et al., 2015), Douban Conversation Corpus (Wu et al., 2017), and E-commerce Dialogue Corpus (Zhang et al., 2018).

The first data set we adopt is Ubuntu Dialogue Corpus (Lowe et al., 2015) which is a multi-turn English conversation data set constructed from chat logs of the Ubuntu forum. We use the version provided by Xu et al. (2017). The data contain 1 million context-response pairs for training, and 0.5 million pairs for validation and test respectively. In all three sets, positive responses are human responses, while negative ones are randomly sampled. The ratio of the positive and the negative is 1:1 in the training set, and 1:9 in both the validation set and the test set. Following Lowe et al. (2015), we employ recall at position k in n candidates ($\text{R}_n@k$) as evaluation metrics.

The second data set is Douban Conversation Corpus (Wu et al., 2017) which is a multi-turn Chinese conversation data set crawled from Douban group[5].

[3] We acknowledge that there may still be topic shifts in the query. In preliminary experiments, we randomly selected 100 conversations from Douban dataset for human annotation and found that only 6 samples (6%) had topic shifts in the query. Therefore, we can conclude that topic shifts would not affect the importance of the query to match response in most cases.

[4] s_{n_c} achieves the largest value since the dialogue query u_{n_c} attends to itself.

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

Metrics / Models	Ubuntu Corpus			Douban Corpus						E-commerce Corpus		
	$R_{10}@1$	$R_{10}@2$	$R_{10}@5$	MAP	MRR	P@1	$R_{10}@1$	$R_{10}@2$	$R_{10}@5$	$R_{10}@1$	$R_{10}@2$	$R_{10}@5$
Multi-View (Zhou et al., 2016)	0.662	0.801	0.951	0.505	0.543	0.342	0.202	0.350	0.729	0.421	0.601	0.861
DUA (Zhang et al., 2018)	0.752	0.868	0.962	0.551	0.599	0.421	0.243	0.421	0.780	0.501	0.700	0.921
MRFN (Tao et al., 2019a)	0.786	0.886	0.976	0.571	0.617	0.448	0.276	0.435	0.783	-	-	-
IoI (Tao et al., 2019b)	0.796	0.894	0.974	0.573	0.621	0.444	0.269	0.451	0.786	0.563	0.768	0.950
BERT (Gu et al., 2020)	0.808	0.897	0.975	0.591	0.633	0.454	0.280	0.470	0.828	0.610	0.814	0.973
SMN (Wu et al., 2017)	0.726	0.847	0.961	0.529	0.569	0.397	0.233	0.396	0.724	0.453	0.654	0.886
SMN + TACM (Ours)	0.789*	0.889*	0.973*	0.564*	0.618*	0.451*	0.271*	0.429*	0.776*	0.521*	0.701*	0.920*
DAM (Zhou et al., 2018)	0.767	0.874	0.969	0.550	0.601	0.427	0.254	0.410	0.757	0.526	0.727	0.933
DAM + TACM (Ours)	0.803*	0.899*	0.979*	0.569*	0.610*	0.437*	0.269*	0.434*	0.775*	0.535*	0.732	0.934
MSN (Yuan et al., 2019)	0.800	0.899	0.978	0.587	0.632	0.470	0.295	0.452	0.788	0.606	0.770	0.937
MSN + TACM (Ours)	0.811*	0.904*	0.979	0.594*	0.640*	0.482*	0.303*	0.457	0.789	0.616*	0.793*	0.955*

Table 1: Evaluation results on three data sets. Numbers marked with $*$ mean that the improvement is statistically significant compared with corresponding baseline (t-test with p-value < 0.05).

The data set consists of 1 million context-response pairs for training, 50 thousand pairs for validation, and 6,670 pairs for test. In the training set and the validation set, the last turn of each conversation is regarded as a positive response and negative responses are randomly sampled. The ratio of the positive and the negative is 1:1 in training and validation set. In the test set, each context has 10 response candidates retrieved from an index whose appropriateness regarding to the context is judged by human annotators. The average number of positive responses per context is 1.18. Following Wu et al. (2017), we employ $R_{10}@1$, $R_{10}@2$, $R_{10}@5$, mean average precision (MAP), mean reciprocal rank (MRR), and precision at position 1 (P@1) as evaluation metrics.

Apart from the above two data sets, we also choose E-commerce Dialogue Corpus (Zhang et al., 2018). The data consist of real-world conversations between customers and customer service staffs in Taobao[6], which is the largest e-commerce platform in China. There are 1 million context-response pairs in the training set, and 10 thousand pairs in the validation set and test set respectively. Each context in the training and validation set corresponds to one positive response candidate and one negative response candidate, while in the test set, the number of response candidates per context is 10 with only one of them positive. Human responses are treated as positive responses, and negative ones are automatically collected by ranking the response corpus based on conversation history augmented messages using Apache Lucene[7]. Following Zhang et al. (2018), we employ $R_{10}@1$, $R_{10}@2$, and $R_{10}@5$ as evaluation metrics.

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

[7] http://lucene.apache.org/

4.2 Baselines and Matching Models

4.2.1 Referenced Models

Since the task of retrieval-based dialogues was proposed, many impressive models have emerged. Therefore, we choose these models as referenced baselines including the multi-view matching model (Multi-View) (Zhou et al., 2016), the deep utterance aggregation model (DUA) (Zhang et al., 2018), the multi-representation fusion network (MRFN) (Tao et al., 2019a), the interaction-over-interaction network (IOI) (Tao et al., 2019b) and BERT for response selection (Gu et al., 2020).

4.2.2 Selected Matching Models

Since our proposed TACM layer can be adapted to the existing multi-turn context-response matching models, we choose the following three representative models to verify its effectiveness.

SMN: Wu et al. (2017) first lets each turn of utterance interact with the response, and forms a matching vector for the pair through CNNs. Then, all of the matching vectors are aggregated with a RNN as a matching score. We select the model as it is a representative in the framework of *representation-matching-aggregation*, where the $f_{\texttt{IUR}}$ is a RNN encoder, $f_{\texttt{URM}}$ is an inner-product similarity function and $f_{\texttt{AGG}}$ is a 2D CNN followed by an RNN.

DAM: Zhou et al. (2018) constructs representations of utterances in the context and the response with stacked self-attention and cross-attention. We select the model as it is a representative context-response matching model based on Transformer architecture (Vaswani et al., 2017), where the $f_{\texttt{IUR}}$ is an Attentive Module, $f_{\texttt{URM}}$ is a similarity function over representations and $f_{\texttt{AGG}}$ is a 3D CNN.

MSN: Yuan et al. (2019) firstly utilizes a multi-hop selector to select relevant utterances as context.

Then, the model matches the filtered context with candidate response and obtains a matching score. We choose the model as it is the best performing multi-turn context-response matching model without PLMs on three benchmarks, where $f_{\texttt{IUR}}$ is a multi-hop selector network, $f_{\texttt{URM}}$ is the ensemble of inner-product and cosine similarity functions over self-attention and cross-attention representations and $f_{\texttt{AGG}}$ is a 2D CNN followed by an RNN.

It is worth noting that we do not adopt PLMs as backbone in our main experiments because they concatenate multi-turn context and treat the problem in single-turn scenario. Take BERT as an example, it conducts full interaction over the whole dialogue turns of utterances for context comprehension (Gu et al., 2020). This interaction is direct, but there may be redundant calculations for multi-turn context, resulting in a large amount of parameters. Instead, we put forward series of heuristic strategies to conduct turn-aware interactions for multi-turn dialogue context. In subsequent experiments, we find that our model can achieve comparable performance to BERT[8] with one third of parameters.

4.3 Implementation Details

We implement all models with PyTorch (Paszke et al., 2017). Word embedding is pre-trained with Word2Vec (Mikolov et al., 2013) on the training set of each corpus, and the dimension of word vectors is 200. For fair comparison, we limit the maximum number of utterances in each context as 10 and the maximum number of words in each utterance and response as 50 following Wu et al. (2017); Zhou et al. (2018); Yuan et al. (2019). Truncation or zero-padding is applied to a context or a response candidate when necessary. All other settings such as the kernel size of CNN in matching module and the dimension of hidden states of RNN in aggregation layer are consistent with the original papers. The batch size and initial learning rate are also consistent with the default setting of the proposed baselines (SMN, DAM, MSN). We used their public code to reproduce their models, and the results were similar to those reported in the original papers. For more detailed settings of baselines, please refer to Appendix A.1. The parameters were updated by Adam (Kingma and Ba, 2015). In *"Window"* strategy, γ is set as 1. Early stopping on the validation data is adopted as a regularization strategy.

[8]Without loss of generality, we adapt the idea of TACM to standard BERT (Devlin et al., 2019) and do not consider any self-supervised post-training (Whang et al., 2020; Xu et al., 2021) or data augmentation strategies (Han et al., 2021) in this paper.

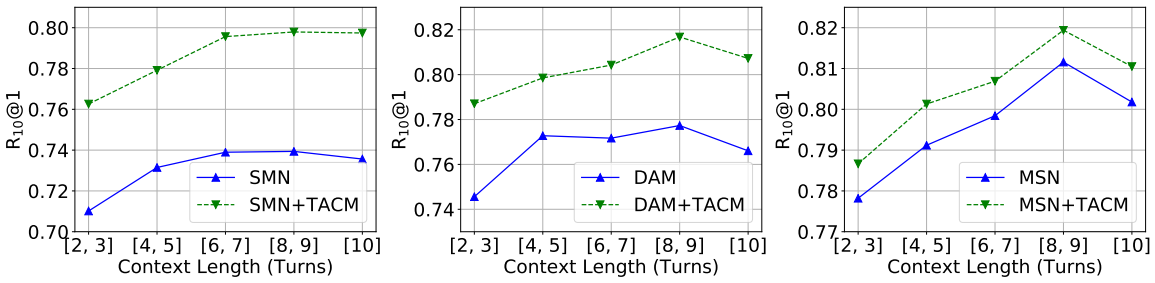
4.4 Evaluation Results

Table 1 reports the evaluation results of training with turn-aware context modeling with *"Rolling"*, *"Window"*, *"Highway"* and *"Weighted"* strategies. We can see that all modeling strategies can consistently improve the original matching models on all three data sets. The improvement from the corresponding baselines is statistically significant (t-test with p-value < 0.05) on $\text{R}_{10}@1$ (the most important evaluation metric in retrieval-based chatbots) and many other metrics. In particular, as SMN and DAM both use non-turn-aware representation, the improvement also shows the effectiveness of turn-awareness. Furthermore, we can observe that as the performance of the original model enhances (that is, from SMN to DAM to MSN), the improvement brought by TACM layer gradually decreases. This may due to the increasing complexity of the original model's utterance-response matching ($f_{\texttt{URM}}$) and feature fusion ($f_{\texttt{AGG}}$), which alleviates the missing semantic relationship among different turns of utterances. Besides, it is interesting to find that a simple SMN with TACM even performs better than DAM (encoding the context with five self-attention layers) on the Ubuntu and Douban data, although DAM is in a more complicated structure.

In addition, we are surprised to find that MSN+TACM achieves comparable or even better performance in most metrics to BERT but uses fewer parameters. It should be noted that the number of parameters of SMN + TACM , DAM + TACM and MSN + TACM are 35.1M, 35.7M, 39.6M respectively, which is almost $1/3$ of BERT (110M). Such results indicate that the heuristic TACM strategies are lighter and more effective than the full-interaction in multiple transformer layers (such as BERT), which may contains amounts of redundant semantic interactions among the dialogue turns, thus greatly increasing the complexity. The above experimental phenomenon suggests that a dynamic interaction strategy among dialogue turns in PLMs can be explored in future work.

4.5 Further Discussions

Ablation Study. We also conducted additional comprehensive ablation experiments to explore the improvement of the model brought by the above four TACM strategies on Ubuntu data with SMN, DAM and MSN respectively as demonstrated in Ta-



(a) SMN (b) DAM (c) MSN

Figure 3: Performance of models (with or without TACM) across different length of contexts on Ubuntu.

Models \ Metrics	R_{10}@1	R_{10}@2	R_{10}@5
SMN	0.726	0.847	0.961
SMN + TACM$_{\texttt{rolling}}$	0.778	0.882	0.972
SMN + TACM$_{\texttt{window}}$	0.771	0.878	0.970
SMN + TACM$_{\texttt{highway}}$	**0.780**	**0.883**	**0.971**
SMN + TACM$_{\texttt{weighted}}$	0.768	0.874	0.970
SMN + TACM$_{\texttt{top2}}$	0.782	0.885	0.973
SMN + TACM$_{\texttt{all}}$	0.789	0.889	0.973
DAM	0.767	0.874	0.969
DAM + TACM$_{\texttt{rolling}}$	0.779	0.883	0.971
DAM + TACM$_{\texttt{window}}$	0.793	0.892	0.975
DAM + TACM$_{\texttt{highway}}$	**0.797**	**0.894**	**0.976**
DAM + TACM$_{\texttt{weighted}}$	0.787	0.890	0.974
DAM + TACM$_{\texttt{top2}}$	0.799	0.896	0.977
DAM + TACM$_{\texttt{all}}$	0.803	0.899	0.979
MSN	0.800	0.899	0.978
MSN + TACM$_{\texttt{rolling}}$	0.805	0.900	0.978
MSN + TACM$_{\texttt{window}}$	0.804	0.899	0.977
MSN + TACM$_{\texttt{highway}}$	0.806	0.901	0.978
MSN + TACM$_{\texttt{weighted}}$	**0.807**	**0.901**	**0.978**
MSN + TACM$_{\texttt{top2}}$	0.808	0.903	0.979
MSN + TACM$_{\texttt{all}}$	0.811	0.904	0.979

Table 2: Evaluation results of model ablation on Ubuntu data. Numbers in bold indicate the best strategies for the corresponding models.

ble 2. We denote the models using matching model X and modeling strategy Y as X+TACM$_Y$. Specifically, we define X+TACM$_{all}$ as all four strategies are used. From the table, we can clearly find that different strategies have different performance improvements on different matching models. For the case of SMN or DAM as the matching model, the best strategy is *"Highway"*. With respect to MSN, *"Highway"* and *"Weighted"* show comparative best performance. Moreover, as we can see, all strategies can independently improve the performance of the original matching models.

We also expose the experimental parameters of each matching model when using the best strategy in Table 3, where the training environment and hyper-parameters are strictly consistent. The results show that we can obtain a significant performance improvement with a considerable increase of model parameters. Comparing the performance of the model that uses the best strategy and the model that uses all strategies, we can find that if all strategies are exploited, the performance of the matching model will be further improved. Intuitively, the representation features obtained by the *"Rolling"* and *"Window"* strategies have a certain degree of redundancy, since the effect of *"Window"* strategy might be covered by that of *"Rolling"*, because of the update mechanism of recurrent attention. Besides, *"Highway"* and *"Weighted"* can also capture similar query-aware features since the query-aware representation of *"Highway"* strategy is based on word-level attention, which may cover the utterance-level weighting mechanism of *"Weighted"* strategy. To verify our assumption, we conducted an additional group of experiments: For each matching model, we selected two strategies X, Y and equipped them, where X represents a strategy with better performance chosen from *"Rolling"* and *"Window"*, and Y stands for another strategy with better performance selected from *"Highway"* and *"Weighted"*. We denote the model as X+TACM$_{top2}$ and demonstrate the results in Table 2. It is interesting to find that uti-

Models \ Metrics	Parameters	R_{10}@1
SMN + TACM$_{\texttt{best}}$	↑6.68%	↑5.4%
DAM + TACM$_{\texttt{best}}$	↑12.3%	↑3.0%
MSN + TACM$_{\texttt{best}}$	↑5.95%	↑0.7%

Table 3: Experimental parameter statistics when the model uses the best strategy on Ubuntu data. ↑ stands for the growth rate.

lizing two better strategies (such as *"Rolling"* and *"Highway"*) leads to a slight performance drop compared to X+TACM$_{all}$, though each representation is useful. In real application, we can choose one of them for multi-turn response selection.

Impact of Context Length We further study how the number of turns influences the performance of different models when the TACM layer is incorporated. Figure 3 shows how the performance of the models changes with respect to different numbers of turns in contexts. We observe a similar trend for all models: they first increase monotonically until context length reaches a certain value (9 for all three matching models), and then drop when context length keeps increasing. The reason might be that when only a few utterances are available in contexts, the model could not capture enough information for matching, but when the contexts become long enough, noise will be brought to matching as utterances in early history could be irrelevant to the query utterance. Despite the fact that long context (Turn=10) is still challenging, the gap between the two forms is bigger on long contexts than it is on short contexts, indicating that our TACM layers can improve the capability of modeling long context and demonstrate higher improvement of matching accuracy on long contexts. It is noted that the performance gap between MSN and MSN with TACM layer does not widen obviously as the number of turns increases. The reason might be that the architecture of MSN is complex and it introduces query-aware features for context-response matching. Nonetheless, MSN with TACM still significant outperforms MSN, which confirms the effectiveness of our framework.

We provide more empirical studies of TACM including comparison between PLM-based interaction and heuristic interaction, case visualization and analysis of hyper-parameter sensitivity in Appendix A.2, A.3 and A.4 respectively.

5 Conclusion

This paper investigates how to improve the performance of existing matching models with a better context modeling method. Empirical results on three benchmarks indicate that query-aware context modeling is the best strategy and employing multiple context modeling strategies can consistently improve the performance of existing response selection models. Additionally, our TACM layer can improve the capability of modeling long context.

Limitations

Besides its merits, our framework still has a few limitations could be further explored in future works. On the one hand, although we try our best to summarize the existing context modeling strategies into three categories, there may still be hybrid or complex methods that cannot be directly categorized; On the other hand, although our methods have been shown to be effective for retrieval-based dialogue models, it also seems reasonable for generative approaches, which needs to be investigated in future work.

We hope our results could encourage future work on addressing these limitations to further explore context modeling for multi-turn response selection.

Ethics Statement

This paper investigates how to improve the performance of existing matching models with a better context modeling method. There will not be any ethical problems or negative social consequences from the research. The data in this paper are all publicly available and are widely adopted by researchers. The proposed method does not introduce ethical/social bias in the data.

Acknowledgements

We would like to thank the anonymous reviewers for their constructive comments and suggestions. This work was supported by the National Key R&D Program of China under Grant No. 2020AAA0106600. Dongyan Zhao is the corresponding author.

References

Hongshen Chen, Zhaochun Ren, Jiliang Tang, Yihong Eric Zhao, and Dawei Yin. 2018. Hierarchical variational memory network for dialogue generation. In *Proceedings of the 2018 World Wide Web Conference*, pages 1653–1662.

Qian Chen and Wen Wang. 2019. Sequential matching model for end-to-end multi-turn response selection. In *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7350–7354. IEEE.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for*

Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jia-Chen Gu, Tianda Li, Quan Liu, Zhen-Hua Ling, Zhiming Su, Si Wei, and Xiaodan Zhu. 2020. Speaker-aware bert for multi-turn response selection in retrieval-based chatbots. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 2041–2044.

Jia-Chen Gu, Zhen-Hua Ling, and Quan Liu. 2019. Interactive matching network for multi-turn response selection in retrieval-based chatbots. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 2321–2324.

Janghoon Han, Taesuk Hong, Byoungjae Kim, Youngjoong Ko, and Jungyun Seo. 2021. Fine-grained post-training for improving retrieval-based dialogue systems. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1549–1558, Online. Association for Computational Linguistics.

Dan Hendrycks and Kevin Gimpel. 2016. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*.

Baotian Hu, Zhengdong Lu, Hang Li, and Qingcai Chen. 2014. Convolutional neural network architectures for matching natural language sentences. In *Advances in Neural Information Processing Systems*, pages 2042–2050.

Samuel Humeau, Kurt Shuster, Marie-Anne Lachaux, and Jason Weston. 2019. Poly-encoders: Architectures and pre-training strategies for fast and accurate multi-sentence scoring. In *International Conference on Learning Representations*.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*.

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

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in pytorch.

Iulian Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30.

Iulian Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron Courville, and Yoshua Bengio. 2017. A hierarchical latent variable encoder-decoder model for generating dialogues. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

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)*, pages 1577–1586.

Rupesh K Srivastava, Klaus Greff, and Jürgen Schmidhuber. 2015. Training very deep networks. In *Advances in neural information processing systems*, pages 2377–2385.

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 '19, page 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 annual meeting of the association for computational linguistics*, pages 1–11.

Zhiliang Tian, Rui Yan, Lili Mou, Yiping Song, Yansong Feng, and Dongyan Zhao. 2017. How to make context more useful? an empirical study on context-aware neural conversational models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 231–236.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

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.

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*, pages 1081–1090.

Mingxuan Wang, Zhengdong Lu, Hang Li, and Qun Liu. 2015. Syntax-based deep matching of short texts. In *Proceedings of the 24th International Conference on Artificial Intelligence*, IJCAI'15, page 1354–1361. AAAI Press.

Taesun Whang, Dongyub Lee, Chanhee Lee, Kisu Yang, Dongsuk Oh, and HeuiSeok Lim. 2020. An effective domain adaptive post-training method for bert in response selection. In *Proc. Interspeech 2020*.

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)*, pages 496–505.

Chen Xing, Yu Wu, Wei Wu, Yalou Huang, and Ming Zhou. 2018. Hierarchical recurrent attention network for response generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Ruijian Xu, Chongyang Tao, Daxin Jiang, Xueliang Zhao, Dongyan Zhao, and Rui Yan. 2021. Learning an effective context-response matching model with self-supervised tasks for retrieval-based dialogues. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14158–14166.

Zhen Xu, Bingquan Liu, Baoxun Wang, Chengjie Sun, and Xiaolong Wang. 2017. Incorporating loose-structured knowledge into conversation modeling via recall-gate lstm. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 3506–3513. IEEE.

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. In *Proceedings of The Web Conference 2020*, pages 2592–2598.

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*, pages 245–254.

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)*, pages 111–120.

Hainan Zhang, Yanyan Lan, Liang Pang, Jiafeng Guo, and Xueqi Cheng. 2019. Recosa: Detecting the relevant contexts with self-attention for multi-turn dialogue generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3721–3730.

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)*, pages 1118–1127.

A Appendix

A.1 The Detailed Settings of the Experiments for Baselines

We used public codes to reproduce all three baselines (SMN, DAM, MSN), and the results were similar to those reported in the original papers. Specifically, we limited the maximum number of utterances in each context as 10 and the maximum number of words in each utterance and response as 50. Following Wu et al. (2017); Zhou et al. (2018); Yuan et al. (2019), we padded zeros if the number of turns in a context is less than 10, otherwise we kept the last 10 turns. If the length of each utterance or each response candidate exceeded the limitation, we only kept the first tokens because we assume that the most important part will be given first, otherwise we padded zeros behind. Word embeddings were all initialized by the results of Word2Vec (Mikolov et al., 2013) which ran on the training data, and the dimension of word vectors is 200. Adam (Kingma and Ba, 2015) algorithm was used in all baselines. For SMN, the window size of CNN was $(3, 3)$ and the initial learning rate was 0.001. The batch size was 200. The hidden size of the two GRUs was 200 and 50. For DAM, the number of stacked self-attention layers was 5. The learning rate was initialized as $1e-3$ and gradually decreased during training, and the batch size was 256. For MSN, the dimension of the hidden states of GRU was 300. The learning rate was also initialized as $1e-3$ and gradually decreased during training. The batch size was 200, 150, 200 on Ubuntu, Douban and E-commerce Corpus respectively.

A.2 PLM-based Interaction v.s. Heuristic Interaction

We are also curious about how to adapt turn-aware context modeling strategies to existing PLMs. Take BERT as an example, all utterances in context and response candidate are concatenated as a single consecutive token sequence with special tokens separating them, which converts multi-turn context understanding into a single-turn scenario and makes context interaction non-turn-aware (Gu et al., 2020). Thus, we cannot directly use the several strategies introduced in Section 3.3. But we can still borrow the idea and validate the effectiveness of the aforementioned context interaction patterns on PLMs by masking the partial input sequence at the turn level in each transformer layer. Figure 4 depicts the visualization of turn-aware masking.

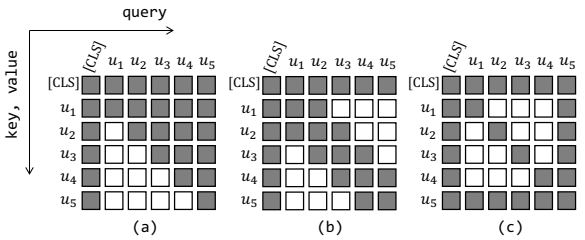


Figure 4: Transformer attention masks for different turn-aware context modeling strategies in BERT, and white color indicates absence of attention. (a) Sequential Context Modeling; (b) Local Context Modeling; (c) Query-Aware Context Modeling. For convenience, we only draw five turns of utterances. In other words, $n_c = 5$ and u_5 is the query utterance. For the Sequential Context Modeling, we only test the forward rolling process. As for the Local Context Modeling, we only consider $\gamma = 1$ consistent with the *"Window"* strategy.

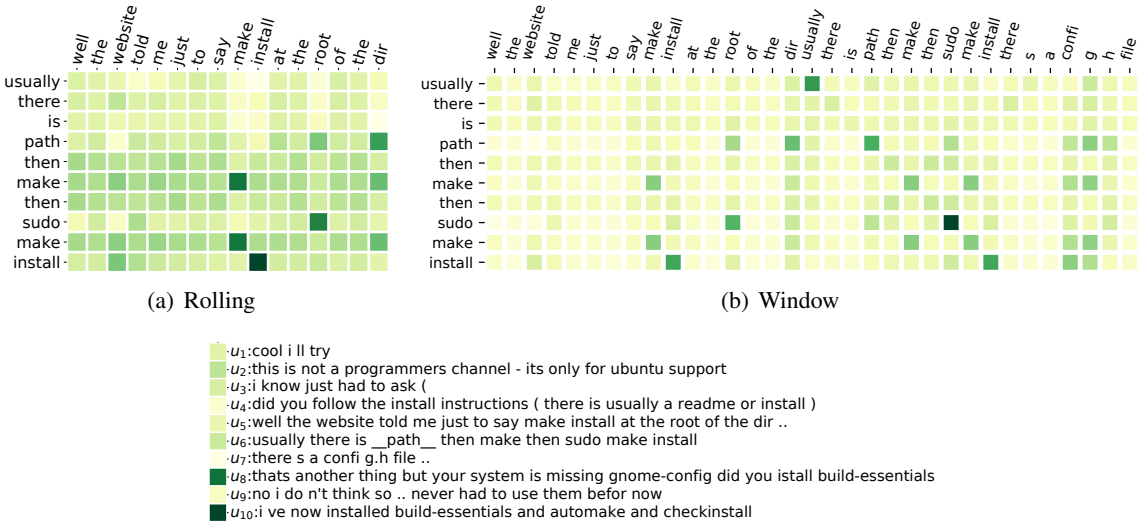
Models \ Metrics	$R_{10}@1$	$R_{10}@2$	$R_{10}@5$
BERT*	0.808	0.897	0.975
BERT + $TACM_{Sequential}$	0.816	0.901	0.979
BERT + $TACM_{Local}$	0.818	0.904	0.979
BERT + $TACM_{Query-Aware}$	0.820	0.904	0.980

Table 4: Performance of the BERT models with heuristic context interaction on Ubuntu data. $*$ means the results copied from Gu et al. (2020).

Specifically, for all strategies, $[CLS]$ can be aware of any other words, $[SEP]$ is consistent with the utterance it follows. To simplify our exposition, we operate on the attention matrix $A \in [0, 1]^{n_s \times n_s}$ of the self-attention mechanism, where n_s is the length of concatenated input sequence of BERT. In order to distinguish the query, key and value in the attention mechanism of BERT from the dialogue query u_{n_c}, here we denote the query, key, and value in the attention mechanism as Q, K and V. As for the Sequential Context Modeling, each turn of utterances can only see the previous turns (here we only test the forward rolling process) and the other turns are masked out. The value of i-th Q word and j-th K/V word in the attention mask matrix is:

$$A_{ij} = \begin{cases} 1, & T(i) \geq T(j) \\ 0, & \text{otherwise} \end{cases} \quad (5)$$

where $T(i), T(j)$ stand for the turn id of i-th Q word and j-th K/V word respectively. As for the Local Context Modeling, each turn of utterances can only be aware of the two adjacent turns of utterance (here we only consider $\gamma = 1$ consistent



(a) Rolling (b) Window

(c) Weighted

Figure 5: Visualization of different strategies in DAM. Experiments are conducted on Ubuntu data. (a) Attention weights between U_6^e and $\overrightarrow{U_5^S}$ in *"Rolling"* strategy, (b) Attention weights between U_6^e and $\mathcal{C}_6$ in *"Window"* strategy, (c) Weights between dialogue query u_{10} and dialogue context $\{u_1, \cdots, u_{10}\}$ in *"Weighted"* strategy. Darker square means larger value in the heatmap.

with the *"Window"* strategy) and the other turns are masked out. The value of i-th Q word and j-th K/V word in the attention mask matrix is:

$$A_{ij} = \begin{cases} 1, & |T(i) - T(j)| \leq 1 \\ 0, & \text{otherwise} \end{cases} \quad (6)$$

As for the Query-Aware Context Modeling, in each transformer layer, we force every word $w_{u_i,k}$ in u_i to be able to pay attention to dialogue query u_{n_c}, but u_i cannot see other turns of utterance except query and itself. This process is achieved by masking out the words of other turns of utterances. The value of i-th Q word and j-th K/V word in the attention mask matrix is formulated as:

$$A_{ij} = \begin{cases} 1, & \begin{array}{r} T(i) = T(j) \text{ or } T(i) = n_c \\ \text{or } T(j) = n_c \end{array} \\ 0, & \text{otherwise} \end{cases} \quad (7)$$

We conduct experiments on BERT on Ubuntu data and the results (shown in Table 4) indicate that by adapting the idea of TACM to BERT, the performance of the model can be improved, which means our TACM strategies are superior to the full interaction in BERT-based models. Among three categories of TACM strategies, Query-Aware Context Modeling is the best, which is consistent with traditional multi-turn response matching models.

A.3 Case Visualization

For a better insight into how four TACM strategies capture turn-aware context information, we perform a case study by visualizing the attention weights between different turns of utterances in DAM. The example is shown in Figure 5, which comes from the test sets of Ubuntu Corpus, and the model successfully selected the best response candidate for it while the corresponding baseline failed. Figure 5(a) gives the visualization results of the attention weights in formation of unidirectional rolling representation of u_6 (calculated by $f_{\text{ATT}}(U_6^e, \overrightarrow{U_5^S}, \overrightarrow{U_5^S})$ in formula 1), denoted as $\overrightarrow{U_6^S}$. We can see that between u_6 and u_5, associated pairs like "path" & "dir", "sudo" & "root" are successfully identified, indicates that *"Rolling"* strategy is useful to recognize such temporal relationships between dialogue turns. In Figure 5(b), intersection areas between the segment "make install" in u_6 and word pieces "confi g" in $\mathcal{C}_6$ got larger matching scores, reveals that the *"Window"* strategy is beneficial to capture the semantic association between adjacent turns of utterances. It is worth noting that the *"Window"* strategy can also grab the correlated word pairs "sudo" & "root", "path" & "dir", which confirms our conjecture once again that there may be redundancy between *"Rolling"* and *"Window"* representations. To examine whether the *"Weighted"* strategy helps to recognize the different corre-

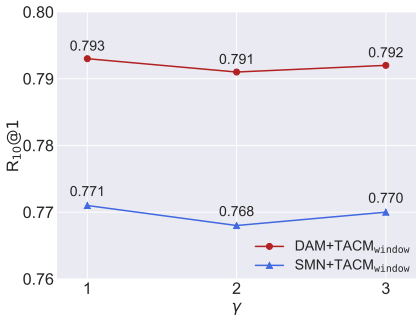


Figure 6: Effects of γ to *"Window"* strategy. Experiments are conducted on Ubuntu data.

lations of multi-turn query-aware history utterances for selecting the response, Figure 5(c) illustrates the correlated weights distribution of each turn of utterance in context $\{u_i\}_1^{n_c}$ (calculated by the cosine similarity in formula 4), denoted as $\{s_i\}_1^{n_c}$. As demonstrated, the model significantly identifies the informative query-aware utterances like u_8, and chooses to discard u_9, which is unrelated to the topic and has little information compared with dialogue query u_{n_c} ($n_c = 10$ in our experiments). Since the *"Highway"* strategy adopt gate mechanism for each entry of the representation vector of u_i, it is difficult for us to visualize it intuitively with an attention matrix, so we only depict the *"Weighted"* strategy to demonstrate the query-aware context modeling. To sum up, this example explains why TACM works well.

A.4 Analysis of Hyper-parameter Sensitivity

We also check the effect of hyper-parameters in *"Window"* strategy. Figure 6 illustrates how the performance of *"Window"* strategy varies under different offsets γ in Ubuntu data with matching model SMN and DAM since the absolute improvement on these two models exceeds 1%. We can observe that the performance of the model fluctuates less with γ changes. We guess this is because the farther utterances are less semantically dependent on the current utterance. Thus, we conclude that the *"Window"* strategy is not sensitive and robust to the choice of offset γ.