Topic-Aware Multi-turn Dialogue Modeling

Yi Xu,[1,2,3] **Hai Zhao,** [1,2,3,*] **Zhuosheng Zhang** [1,2,3]

[1] Department of Computer Science and Engineering, Shanghai Jiao Tong University
[2] Key Laboratory of Shanghai Education Commission for Intelligent Interaction and
Cognitive Engineering, Shanghai Jiao Tong University, Shanghai, China
[3] MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University
xuyi_2019@sjtu.edu.cn, zhaohai@cs.sjtu.edu.cn, zhangzs@sjtu.edu.cn

Abstract

In the retrieval-based multi-turn dialogue modeling, it remains a challenge to select the most appropriate response according to extracting salient features in context utterances. As a conversation goes on, topic shift at discourse-level naturally happens through the continuous multi-turn dialogue context. However, all known retrieval-based systems are satisfied with exploiting local topic words for context utterance representation but fail to capture such essential global topic-aware clues at discourse-level. Instead of taking topic-agnostic n-gram utterance as processing unit for matching purpose in existing systems, this paper presents a novel topic-aware solution for multi-turn dialogue modeling, which segments and extracts topic-aware utterances in an unsupervised way, so that the resulted model is capable of capturing salient topic shift at discourse-level in need and thus effectively track topic flow during multi-turn conversation. Our topic-aware modeling is implemented by a newly proposed unsupervised topic-aware segmentation algorithm and Topic-Aware Dual-attention Matching (TADAM) Network, which matches each topic segment with the response in a dual cross-attention way. Experimental results on three public datasets show TADAM can outperform the state-of-the-art method, especially by 3.3% on E-commerce dataset that has an obvious topic shift.

Introduction

There are generally two ways to build a dialogue system, generation-based and retrieval-based. The former views conversation as a generation problem (Xing et al. 2017; Serban et al. 2017a,b; Zhou et al. 2017; Wu et al. 2018b), while the latter usually consists of a retrieval and matching process (Zhou et al. 2016; Wu et al. 2017; Zhou et al. 2018b; Zhu et al. 2018; Zhang et al. 2018; Zhang and Zhao 2018; Tao et al. 2019a; Gu, Ling, and Liu 2019; Zhang et al. 2020a,c).

The retrieval-based response selection task is aimed to select a most suitable response from a collection of candidate

Turns	Dialogue Text
Turn-1 A:	*Are there any discounts activities recently?*
Turn-2 B:	*No. Our product have been cheaper than before.*
Turn-3 A:	*Oh.*
Turn-4 B:	*Hum!*
Turn-5 A:	*I'll buy these nuts. Can you sell me cheaper?*
Turn-6 B:	*You can get some coupons on the homepage.*
Turn-7 A:	*Will you give me some nut clips?*
Turn-8 B:	*Of course we will.*
Turn-9 A:	*How many clips will you give?*

Table 1: A case in Yuan et al. (2019) from E-commerce. The topic has changed from "price" to "tool" after Turn-6.

answers according to a dialogue history. Early studies concatenate context together and then match with each candidate response (Lowe et al. 2015; Kadlec, Schmid, and Kleindienst 2015; Yan, Song, and Wu 2016; Tan et al. 2015; Wan et al. 2016; Wang and Jiang 2016). Recently, most works turn to explore the interaction between the response and each utterance, which is then fused for a final matching score (Zhou et al. 2016; Wu et al. 2017; Zhang et al. 2018; Zhou et al. 2018a,b; Tao et al. 2019a; Yuan et al. 2019). To be specific, recent models follow an architecture consisting of two parts: Encoder and Matching Network. The former either encodes each context utterance independently with traditional RNNs (Cho et al. 2014), or models the whole context with a pre-trained contextualized language model, and then splits out utterances for further matching (Zhu, Zhao, and Li 2020; Zhang et al. 2019). The latter matching modules vary in previous works, for example, MSN (Yuan et al. 2019) introduces a multi-hop selector which selects relevant utterances to reduce matching noise.

In multi-turn dialogue modeling, topic-aware clues have been more or less considered as it is certain that a long enough multi-turn dialogue may have multiple topics as the conversation goes on and topic shift naturally happens by all means. Table 1 shows an example that there is topic change after Turn-6. Even though existing work like Yuan et al. (2019) select semantic-relevant information or like Wu et al. (2018a) use topic information at word-level for bet-

ter matching, all known systems keep using topic-agnostic or topic-mixed n-gram utterances as a whole for matching context. Instead, this work explicitly extracts topic segments from the dialogue history as basic units for further matching, which is capable of tracking global topic flow throughout the entire multi-turn dialogue history at discourse-level. The proposed topic-aware modeling method groups topic-similar utterances so as to capture salient information and keep robust to irrelevant contextual noise at most degree.

To implement the proposed topic-aware modeling, we present a novel response selection model, named TADAM (Topic-Aware Dual-Attention Matching) network. For encoding, segments and response are concatenated and fed into an encoder which usually adopts a pre-trained language model, and then separated for further matching. For the matching network, we use selectors to weight segments based on relevance with the response in both word and segment levels, and then use cross-attention in a dual way for the multi-turn matching. Especially, we emphasize the last segment, which is the closest to the answer. Finally, we fuse these matching vectors for a final score.

As to our best knowledge, this is the first attempt of handling multi-turn dialogue model in a topic-aware segmentation way. Thus, for lack of multi-turn dialogue datasets with labels for dialogue topic boundaries, we label or splice two datasets in Chinese and English, respectively[1].

Our model especially fits dialogue scenes which have obvious topic shift. Experiments show our proposed model achieves state-of-the-art performance on three benchmark datasets.

Related Work

As our work concerns about topic, we have to consider how to well segment text with multiple sentences into topic-aware units. For such a purpose, previous methods vary in how they represent sentences and how they measure the lexical similarity between sentences (Joty, Carenini, and Ng 2013). TextTiling (Hearst 1997) proposes pseudo-sentences and applies cosine-based lexical similarity on term frequency. Based on TextTiling, LCSeg (Galley et al. 2003) introduces lexical chains (Morris and Hirst 1991) to build vectors. To alleviate the data sparsity from term frequency vector representation, Choi, Wiemer-Hastings, and Moore (2001) employ Latent Semantic Analysis (LSA) for representation and Song et al. (2016) further use word embeddings to enhance TextTiling.

For the concerned task in this paper, response selection in multi-turn dialogues, early studies conduct single-turn match, which directly concatenates all context utterances and then match with the candidate response (Lowe et al. 2015; Kadlec, Schmid, and Kleindienst 2015; Yan, Song, and Wu 2016; Tan et al. 2015; Wan et al. 2016; Wang and Jiang 2016). Recent works tend to explore relationship sequentially among all utterances, which generally follow the representation-matching-aggregation framework (Zhou et al. 2016; Wu et al. 2017; Zhang et al. 2018; Zhou et al.

2018a,b; Tao et al. 2019a; Yuan et al. 2019). In representation, there are two main approaches. One is to encode each utterance separately while the other is to encode the whole context using pre-trained contextualized language models (Devlin et al. 2018; Lan et al. 2020; Liu et al. 2019; Clark et al. 2020; Zhang et al. 2020b; Zhang, Yang, and Zhao 2021), and then split out each utterance. For matching networks, there are various forms. For example, DAM (Zhou et al. 2018b) proposes to match a response with its multi-turn context-based entirely on attention. MSN (Yuan et al. 2019) filters out irrelevant context to reduce noise. In the aggregation process, recent studies use GRU (Cho et al. 2014) with utterance vectors as input and obtain a matching score based on the last hidden state. Recently, many works such as SA-BERT (Gu et al. 2020), BERT-VFT (Whang et al. 2020), DCM (Li, Li, and Ji 2020) have applied pre-training on the domain dataset, and get much improvement with their settings like speaker embeddings.

As to incorporate topic issues into the dialogue system, all existing work were satisfied with introducing local topic-aware features at word-level. Wu et al. (2018a) introduce two topic vectors in the matching step, which are linear combinations of topic words of the context and the response respectively. The idea of topic words are also used to generate informative responses in generation-base chatbots (Xing et al. 2017). Besides, Chen et al. (2018) construct an open-domain dialogue system Gunrock, which does 'topic' classification for each randomly segmented text-piece and then assigns domain-specific dialogue module to generate response. However, 'topic' quoted by Chen et al. (2018) actually refers to *intent*, which is very coarse across all domains, such as music, animal. When multi-turn dialogue naturally consists of multiple topics from a perspective of discourse, all known systems and existing studies ignore such essential global topic-aware clues and keep handling multi-turn dialogue in terms of topic-agnostic segments and the only adopted topic-aware information is introduced at word-level.

Different from all the previous studies, this work for the first time proposes a novel topic-aware solution which explicitly extracts topic segments from multi-turn dialogue history and thus is capable of globally handling topic-aware clues at discourse-level. Our topic-aware model accords with realistic dialogue scenes where topic shift is a common fact as a conversation goes on.

Topic-Aware Multi-turn Dialogue Modeling

Our model consists of two parts: (1)Topic-aware Segmentation, which segments the multi-turn dialogue and extracts topic-aware utterances in an unsupervised way. (2) TADAM Network, which matches candidate responses with topic-aware segments to select a most appropriate one.

Topic-aware Segmentation

Considering that a continuous multi-turn dialogue history D with n utterances, where $D = \{u_1, ..., u_n\}$, we design an unsupervised topic-aware segmentation algorithm to determine segmentation points from intervals between u_i as shown in Algorithm 1, where $E(\cdot)$ is to encode text sequence and more details is in Subsection Results.

Algorithm 1 Topic-aware Segmentation Algorithm

Input: Dialogue $D = \{u_1, u_2, ..., u_n\}$
Output: Topic segment list S
1: S=[ ], start index $i = 1$
2: **while** $i \leq n$ **do**
3:     $l = u_{i-d} \oplus \cdots \oplus u_{i-1}$
4:     $j = 1, c_0 = $ ",
5:     **while** $i + j \leq n + 1$ **and** $j \leq R$ **do**
6:         $c_j = c_{j-1} \oplus u_{i+j-1}$
7:         **if** $j \bmod K == 0$ **then**
8:             $r_j = u_{i+j} \oplus \cdots \oplus u_{i+j+d-1}$
9:             $cost_{c_j} = \max(\ \text{sim}(E(c_j), E(l)),$
                    $\text{sim}(E(c_j), E(r_j)))$
10:         $j += 1$
11:     $c_j^* = \min_{c_j} cost_{c_j}$
12:     $S.$append$(c_j^*), i = |c_j^*| + 1$
    return S

The proposed segmentation algorithm greedily checks each interval between two adjacent utterances to determine a segmentation that lets two resulted segments mostly differ. In detail, for each candidate interval, we concatenate utterances from the previous segmentation point as center segment c and set a sliding window d to get left and right segments l, r. We expect to get segment c, which is the least similar to l, r.

Besides, considering the number of utterances in each topic segment, we set a maximum range R within which segmentation may happen from last point. In addition, to avoid yielding too many fragments, interval check is actually performed skipping every K ones. We also set a threshold θ_{cost} in Algorithm 1, i.e., if $cost > \theta_{cost}$, then there is no segmentation so as to further control fragmentation.

TADAM Network

After topic-aware segmentation, dialogue history is segmented into topic segments, which are then as context input of TADAM to keep self-informative and robust to irrelevant contextual information.

We denote the dataset as $D = \{(C, R, Y)_k\}_{k=1}^N$, where C is dialogue context, and R is the candidate response, and $Y \in \{0, 1\}$ is the label indicating whether R is a proper response for C. Besides, $C = \{S_1, ..., S_n\}$ and $S_i, 1 \leq i \leq n$ is the i-th topic segment in context C.

Figure 1 shows the architecture of TADAM. For encoding, all segments and response are concatenated as an input sequence for an encoder and then separated out according to position information. For matching, we first use response to weight each segment at both word and segment levels in order to pay more attention to those which are more related. Then, the response is used to match each weighted segment in a dual cross-attention way for deep matching of each pair of segment and response, which outputs a matching vector for each segment. Those matching vectors are fed into GRU for a single one. Ultimately, we fuse the single vector with the matching vector of the last segment for a score.

Encoding and Separation We take well pre-trained language model as our encoder component. Following the encoding manner of pre-trained language models such as BERT and ALBERT, we concatenate all segments $\{S_i\}_{i=1}^n$ and response R with special tokens: $X = [\text{CLS}]\, S_1\, [\text{SEP}] ... [\text{SEP}]\, S_n\, [\text{SEP}]\, R\, [\text{SEP}]$, which is fed into the encoder. Then we split the segments and response according to position information. Because the number of segments n is different among all contexts, and the length of each segment varies, we set the maximum number of segments in each context as T, and the maximum length of segments as L. Thus after separation, we can get context representation $C \in \mathbb{R}^{T \times L \times d}$, where d is dimension of each token in the encoder output. Response representation $R \in \mathbb{R}^{L \times d}$ and C contains T segment representation $\{S_i \in \mathbb{R}^{L \times d}\}_{i=1}^T$. More encoding modes are discussed in Section Analysis.

Segment Weighting Yuan et al. (2019) use the last k utterances of context to select relevant and meaningful context utterances at both word and utterance granularities. Considering our process unit is segments, which are much longer than single utterances, it will lead to information scarcity if we delete the whole unselected segments. Instead, we use the candidate response as a key utterance to give each segment a weight at both word and segment granularities.
• **Word-level Weights:** At the word level, we build a matching feature map between each segment S_i and response R, which is formulated as:

$$M^1 = \text{transfer}(C, W, R),$$
$$M = \frac{1}{\sqrt{d}}\tanh(M^1)V,$$
$$M^{pool} = [\max_{\text{dim}=2} M, \max_{\text{dim}=3} M],$$

where $M^1_{xyuv} = \sum_k C_{xyk}W_{kkv}R_{uk}$. $W \in \mathbb{R}^{d \times d \times h}, V \in \mathbb{R}^h$ are learnable parameters, and $M^1 \in \mathbb{R}^{T \times L \times L \times h}$, $M \in \mathbb{R}^{T \times L \times L}$, $M^{pool} \in \mathbb{R}^{T \times 2L}$. The matching map M is max-pooled in both row and column. M^{pool} indicates the relevance between response R and T segments at the word level. Then we transfer matching features M^{pool} into weights for T segments through a linear layer:

$$s_1 = \text{softmax}(M^{pool}W' + b),$$

where $s_1 \in \mathbb{R}^T$ is segment weights at word level, and $W' \in \mathbb{R}^{2L}, b \in \mathbb{R}^T$ are learnable parameters.
• **Segment-level Weights:** At the segment level, we build segment representation by mean-pooling token vectors so that the segment is represented by a vector. Then we use cosine similarity with the response R to get a weight for each segment:

$$C' = \text{mean}(C)_{\text{dim}=2},$$
$$s_2 = \cos(C', R),$$

where $C' \in \mathbb{R}^{T \times d}$, $s_2 \in \mathbb{R}^T$. s_2 is segment weights at the segment level, which catches the overall semantic similarity between the response and segment.
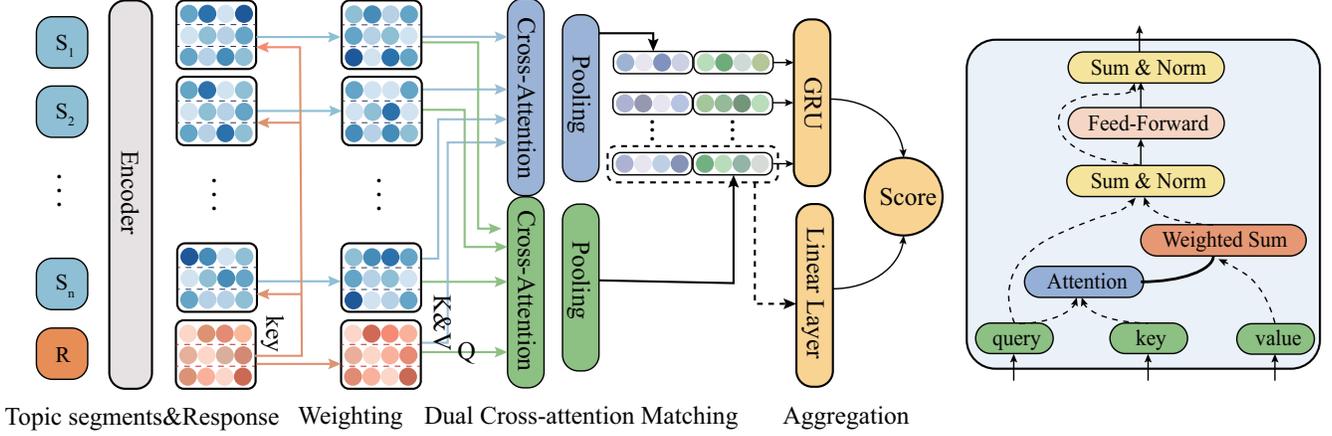
Figure 1: Architecture of our TADAM network (left) and Attentive Module (right).

• **Combination:** In order to fuse both two levels, we set a hyper-parameter α and sum up s_1, s_2. Then we multiply s and C to give segments different degree of relevance based on the response:

$$s = \alpha s_1 + (1 - \alpha)s_2,$$
$$\tilde{C} = s \odot C,$$

where $s \in \mathbb{R}^T, \tilde{C} \in \mathbb{R}^{T \times L \times d}$.

Dual Cross-attention Matching We also use the Attentive Module in DAM (Zhou et al. 2018b) which is a unit of transformer (Vaswani et al. 2017) to encode the interaction between two sequences. Here we apply cross-attention to the segment and response in a dual way, which is fused for further process.

• **Attentive Module:** The architecture of Attentive Module is shown in Figure 1, which takes three sequences as input: query sequence $Q \in \mathbb{R}^{n_q \times d}$, key sequence $K \in \mathbb{R}^{n_k \times d}$ and value sequence $V \in \mathbb{R}^{n_v \times d}$, where n_q, n_k, n_v are the number of tokens respectively, and d is embedding dimension.

Attentive Module first takes each word in the query sentence to attend to words in the key sentence via Scaled Dot-Product Attention (Vaswani et al. 2017), then applies those attention results upon the value sentence, which is defined as:

$$V_{att} = \mathrm{softmax}(\frac{QK^T}{\sqrt{d}})V,$$

where $V_{att} \in \mathbb{R}^{n_q \times d}$. Then layer normalization (Ba, Kiros, and Hinton 2016) takes V_{att} as input and we denote the output as $V'_{att} \in \mathbb{R}^{n_q \times d}$, which is then fed into a feed-forward network FFN with RELU (LeCun, Bengio, and Hinton 2015) activation:

$$FFN(V'_{att}) = \max(0, V'_{att}W_1 + b_1)W_2 + b_2,$$

where W_1, b_1, W_2, b_2 are learnable parameters and $FFN(V'_{att}) \in \mathbb{R}^{n_q \times d}$. We denote the whole Attentive Module as:

$$\mathrm{AttentiveModule}(Q, K, V),$$

• **Dual Cross-Attention Matching:** For each segment $\tilde{S}_i \in \mathbb{R}^{L \times d}$ in weighted context $\tilde{C} \in \mathbb{R}^{T \times L \times d}$, we use cross-attention to build S_i^{cross}, each token of which reflects the relevance with response R, and R_i^{cross} about relevance with $\tilde{S}_i$:

$$S_i^{cross} = \mathrm{AttentiveModule}(\tilde{S}_i, R, R),$$
$$R_i^{cross} = \mathrm{AttentiveModule}(R, \tilde{S}_i, \tilde{S}_i),$$

where $S_i^{cross}, R_i^{cross} \in \mathbb{R}^{L \times d}$. All the S_i^{cross}, R_i^{cross} construct $C_S^{cross} \in \mathbb{R}^{T \times L \times d}, C_R^{cross} \in \mathbb{R}^{T \times L \times d}$, which are then mean-pooled and concatenated as C^{cross} (other pooling and fusion methods are discussed in Section Analysis).

$$C_S^{cross} = \{S_i^{cross}\}_{i=1}^T, \tag{1}$$
$$C_R^{cross} = \{R_i^{cross}\}_{i=1}^T, \tag{2}$$
$$C^{cross} = [\underset{\dim=2}{\mathrm{mean}}(C_S^{cross}), \underset{\dim=2}{\mathrm{mean}}(C_R^{cross})], \tag{3}$$

where $C^{cross} \in \mathbb{R}^{T \times 2d}$ indicates the matching features for each segment based on the candidate response.

Aggregation We use GRU to model the relation of segments as SMN (Wu et al. 2017). Considering the last segment which is the closest to the response may contain more relevant information, we add C_T^{cross}, which is the dual matching between the last segment S_T^{cross} and the response R_T^{cross}. Furthermore, we add a linear layer to C_T^{cross}:

$$\hat{H} = \mathrm{GRU}(C^{cross}),$$
$$C_T^{cross} = [\underset{\dim=1}{\mathrm{mean}}(S_T^{crose}), \underset{\dim=1}{\mathrm{mean}}(R_T^{crose})], \tag{4}$$
$$\hat{C}_T^{cross} = W_3 C_T^{cross} + b_3,$$

where $\hat{H}, C_T^{cross}, \hat{C}_T^{cross} \in \mathbb{R}^{2d}$. $W_3 \in \mathbb{R}^{2d \times 2d}, b_3 \in \mathbb{R}^{2d}$ are learnable parameters. Finally, we get the score for context C and its candidate response r using a linear layer, where $W_4 \in \mathbb{R}^{4d}, b_4 \in \mathbb{R}$ are learnable parameters (other fusion methods are discussed in Section Analysis):

$$score = \mathrm{sigmoid}(W_4^T[\hat{H}, \hat{C}_T^{cross}] + b_4). \tag{5}$$

Experiments

Dataset

For topic-aware segmentation, our proposed algorithm is evaluated in two newly built datasets as it is the first attempt for topic-aware multi-turn dialogues. **For Chinese**, we annotate a dataset including 505 phone records of customer service on banking consultation. **For English**, we build dataset including 711 dialogues by joining dialogues from existing multi-turn dialogue datasets: MultiWOZ Corpus[2] (Budzianowski et al. 2018) and Stanford Dialog Dataset (Eric et al. 2017), where each dialogue is about one topic.

For response selection, TADAM is tested on three widely used public datasets.(1) **Ubuntu Corpus** (Lowe et al. 2015): consists of English multi-turn conversations about technical support collected from chat logs of the Ubuntu forum. (2) **Douban Corpus** (Wu et al. 2017): consists of multi-turn conversations from the Douban group, which is a popular social networking service in China. (3) **E-commerce Corpus** (Zhang et al. 2018): includes conversations between customers and shopkeepers from the largest e-commerce platform Taobao in China. The E-commerce Corpus has an obvious topic shift, including commodity consultation, logistics express, recommendation, and chitchat.

Evaluation

For topic-aware segmentation, we adopt three metrics: (1) MAE adopted by (Takanobu et al. 2018), is defined as $\frac{1}{|T|}\sum_{D \in T}|N_{pred}(D) - N_{ref}(D)|$, where T is a dialogue, and $N_{pred}(D), N_{ref}(D)$ denote the prediction and reference number of segments in dialogue T. (2) WindowDiff (WD) adopted from (Pevzner and Hearst 2002). WD moves a short window through the dialogue from beginning to end, and if the number of segmentation in prediction and reference is not identical, a penalty of 1 is added. The window size in our experiments is set to 4, and we report the mean WD for all experiments. (3) F_1 score is the harmonic average of recall and precision of segmentation points.

For response selection, we use the same metric $R_n@k$ as previous works, which selects k best matchable candidate responses among n and calculates the recall of the true ones. Besides, because Douban corpus has more than one correct candidate response, we also use MAP (Mean Average Precision), MRR (Mean Reciprocal Rank), and Precision-at-one P@1 as previous works.

Experimental Settings

For both tasks, we use pre-trained BERT[3] (Devlin et al. 2018) (bert-base-uncased & bert-base-chinese) as encoder.

For topic-aware segmentation: In both datasets, we set range $R = 8$, jump step $K = 2$ (value of 1 will lead to fragmentation), window size $d = 2$ and threshold θ_{cost}=0.6. For TextTiling, the length of a pseudo sentence is set to 20 in the Chinese dataset and 10 in the English dataset, which is close to the mean length of utterances in both datasets. Window size and block size for TextTiling are all set to 6.

[2]https://doi.org/10.17863/CAM.41572

[3]https://github.com/huggingface/transformers

For response selection: We apply topic-aware segmentation algorithm to Ubuntu, Douban and E-commerce with range $R = 2, 2, 6$ after trying different values. As to our model, the max input sequence length is set to 350 after WordPiece tokenization and the max number of segments is 10. We set the learning rate as 2e-5 using BertAdam with a warmup proportion of 10%. Our model is trained with batch size of $\{20,32,20\}$ and epoch of $\{3,3,4\}$ for Ubuntu, Douban and E-commerce. Besides, the α of word-level weights is set to 0.5. As for the baseline BERT, we finetuned it. The epoch is set as $\{3, 2, 3\}$ for three datasets, and other settings are the same as our model.

Results

Topic-aware segmentation: To evaluate the performance of our method, we compare it with a typical text segmentation algorithm, TextTiling (Hearst 1997), which is a classic text segmentation algorithm using term frequency vectors to represent text. Besides, TextTiling+Embedding (Song et al. 2016) applies word embedding to compute similarity between texts. Therefore, we also compare TextTiling with our segmentation algorithm in three representation methods: $BERT_{CLS}$, which uses "$[CLS]$" embedding to directly encode the entire text, and $BERT_{mean}$, GloVe: just use mean vector of all words in the text.

Method	Chinese			English		
	MAE	WD	F_1	MAE	WD	F_1
TextTiling	1.90	0.45	0.52	10.08	0.83	0.34
TextTiling+GloVe	2.0	0.45	0.52	6.38	0.75	0.33
TextTiling+$BERT_{mean}$	6.50	0.60	0.45	9.64	0.81	0.32
TextTiling+$BERT_{CLS}$	6.51	0.60	0.45	9.78	0.82	0.33
Our algo.+ GloVe	3.83	0.61	0.48	3.48	0.59	0.56
Our algo.+$BERT_{mean}$	2.95	0.52	0.51	2.98	**0.52**	**0.61**
Our algo.+$BERT_{CLS}$	**0.79**	**0.34**	**0.61**	**1.04**	0.54	0.44

Table 2: Topic-aware segmentation results.

Results are shown in Table 2. We can find that, just except for GloVe in the Chinese dataset, our proposed segmentation algorithm for multi-turn dialogues surpasses TextTiling in all metrics. TextTiling tends to have larger MAE, because it ignores the number of turns in a topic round. In our following response selection part, we use our algorithm with $BERT_{CLS}$ for topic-aware segmentation.

Response selection: First, we apply the segmentation algorithm to dialogues on Ubuntu, Douban and E-commerce datasets. We use the smallest number larger than average length of dialogues as the initial value of R. Then different R are tried around the initial value and we select the best one.

Besides, we concatenate the context and candidate response as input for BERT as a basic sequence classification baseline. Results in Table 3 show that our model outperforms all public works and especially gets much improvement (3.3% in $R_{10}@1$) over the strong pre-trained contextualized language model in E-commerce dataset, which shows the effectiveness of our topic-aware models in dialogues with topic shifting scenes.

Model	Ubuntu			Douban						E-commerce		
	$R_{10}@1$	$R_{10}@2$	$R_{10}@5$	MAP	MRR	P@1	$R_{10}@1$	$R_{10}@2$	$R_{10}@5$	$R_{10}@1$	$R_{10}@2$	$R_{10}@5$
TF-IDF (Lowe et al. 2015)	41.0	54.5	70.8	33.1	35.9	18.0	9.6	17.2	40.5	15.9	25.6	47.7
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, Schmid, and Kleindienst 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, Schmid, and Kleindienst 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, Schmid, and Kleindienst 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, Song, and Wu 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
Atten-LSTM (Tan et al. 2015)	63.3	78.9	94.3	49.5	52.3	33.1	19.2	32.8	71.8	40.1	58.1	84.9
MV-LSTM (Wan et al. 2016)	65.3	80.4	94.6	49.8	53.8	34.8	20.2	35.1	71.0	41.2	59.1	85.7
Match-LSTM (Wang and Jiang 2016)	65.3	79.9	94.4	50.0	53.7	34.5	20.2	34.8	72.0	41.0	59.0	85.8
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	-	-	-
IMN (Gu, Ling, and Liu 2019)	79.4	88.9	97.4	57.0	61.5	44.3	26.2	45.2	78.9	62.1	79.7	96.4
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
BERT	81.9	90.4	97.8	58.7	62.7	45.1	27.6	45.8	82.7	62.7	82.2	96.2
TADAM (Ours)	**82.1**	**90.6**	**97.8**	**59.4**	**63.3**	45.3	28.2	**47.2**	**82.8**	**66.0**	**83.4**	**97.5**

Table 3: Response selection results on on Ubuntu, Douban and E-commerce datasets.

Through observation, we find that the fact of topic shift is negligible in Ubuntu and Douban where a whole dialogue is almost about one topic. This is why improvement of the model in Douban is not as obvious as that in E-commerce with multiple topics. However, our work is not supposed to work best in all scenarios, but especially focuses on the case of topic shift which are common in the more challenging e-commerce/banking situations. Results show that it does work in specific application scenarios where topic shift is ambiguous, which right verifies the motivation of this work.

Analysis

Ablation Studies

In order to investigate the performance of each part of our model, we conduct serious ablation experiments from three angles and results are shown in Table 4. First, we explore the influence of the segment weighting part (in Subsection TADAM Network) by moving word or segment level weights, or both of them (line 3-5). Second, in the aggregation part (in Subsection TADAM Network), we concatenate the multi-turn matching result $\hat{H}$ and last segment matching result $\hat{C}_T^{cross}$ to get a score. Hence we remove either of both each time (line 6-7). Third, we do dual cross-attention matching which includes cross-attentioned segments C_S^{cross} and cross-attentioned responses C_R^{cross} in Equation 1, 2 respectively, and we explore single matching methods in line 8-9.

As shown in Table 4, for E-commerce, removing word or segment level weights all perform worse. Besides, adding extra last segment match does make sense as the traditional multi-turn matching method with GRU. Moreover, the single match either uses segments or response as the query sequence leads to much decrease, which shows the effective-

Model	E-commerce			Ubuntu		
	$R_{10}@1$	$R_{10}@2$	$R_{10}@5$	$R_{10}@1$	$R_{10}@2$	$R_{10}@5$
TADAM	66.0	83.4	97.5	82.1	90.6	97.8
w/o word weights	62.0	82.2	96.3	81.9	90.6	97.8
w/o seg. weights	63.1	82.7	97.0	81.8	90.5	97.8
w/o weights	62.2	82.2	97.5	81.9	90.6	97.7
w/o last seg. match	64.0	82.9	96.7	82.0	90.5	97.8
w/o multi-turn match	63.8	82.5	96.4	81.9	90.5	97.8
single match (seg.)	62.7	83.3	97.4	81.9	90.6	97.8
single match (res.)	62.4	83.2	97.6	81.6	90.3	97.8

Table 4: Ablation study on E-commerce and Ubuntu.

ness of our dual matching design. Results of Ubuntu are not so obvious as that of E-commerce, which can be attributed to that our work especially focuses on the case of topic shift but dialogues in Ubuntu are almost about one topic.

Exploring Other Input Modes

We encode the context and candidate response by concatenating cut segments and the response with special tokens, which is fed into the encoder and then split by position information. In this part, we investigate other input modes in two respects and results are shown in Table 5 (line 3,4).

First, we encode each segment and response separately, indicating that the segment or response itself just focuses on its own meaning without interaction with other contexts. Results in line 3 show that this input mode causes a great loss, which means that although the segment itself can be encoded purely, it leads to information scarcity and will be more sensitive to segmentation error. Table 6 shows a topic-aware segmentation case from E-commerce Corpus. We can

Model	$R_{10}@1$	$R_{10}@2$	$R_{10}@5$
TADAM	66.0	83.4	97.5
separated segment	36.0	49.2	72.5
concatenated utterance	62.9	81.5	97.2
max-pool (match)	63.1	83.4	97.8
sum (match)	63.6	83.0	96.8
sum (aggregation)	62.8	84.3	97.6

Table 5: Results of different input modes & fusion methods.

find that our algorithm does split out topic segments, whose boundaries are very near the true ones. On the other hand, it is not ideal to encode segments and response separately. Encoding concatenated segments allows for relevant information supplement, while separated encoding tends to suffer from mixed topics caused by segmentation deviation.

Turns	Dialogue Text
Turn-1	A: *Hello.*
Turn-2	B: *Excuse me, has my order been sent out?*
Turn-3	A: *Please let me check.*
Turn-4	B: *I found I didn't buy the cotton one.*
Turn-5	A: *Your order has been sent out.*
Turn-6	B: *It's non-woven fabric.*
Turn-7	A: *Yes.*
Turn-8	B: *I'd like to switch to the plant fiber.*
Turn-9	A: *Ok.*
Turn-10	B: *Please change it for me.*
Turn-11	A: *Sorry, your order has been taken by the courier.*
Turn-12	B: *Can you get it back?*
Turn-13	A: *I'll try to intercept for you*
Turn-14	B: *I'm sorry*
Turn-15	A: *It doesn't matter*
Turn-16	B: *What is the natural plant fiber?*

Table 6: A topic-aware segmentation case from E-commerce Corpus. Solid lines are right boundaries, and dotted lines are labelled by our segmentation algorithm.

Second, we remove the segmentation part and just concatenate utterances as well as response (noted as UttDAM). Its results in line 4 of Table 5 also decrease much, which implies that taking the segment as a unit is more robust to the irrelevant contextual information than the utterance. we find that among all right justified dialogues in TADAM, 24.4% are segmented. Figure 2 shows the distribution of dialogue length for both TADAM and UttDAM. We can observe that topic segment does work in dialogues of different lengths.

Exploring Other Fusion Methods

In this section, we discuss other pooling or fusion methods in our model. Results are shown in Table 5 (line 5,6,7). We try max-pooling to get the representation for the whole segment with token vectors (in Equation 3, 4). Besides, we test the element-wise summation fusion method in both matching (in Equation 3, 4) and aggregation stage (in Equation 5).
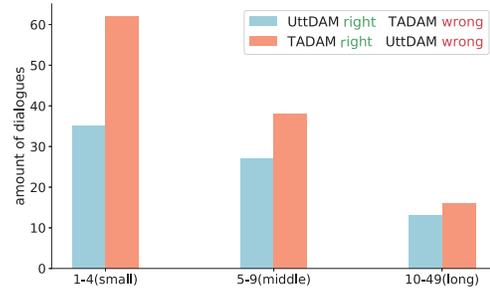

Figure 2: Distribution of dialogue length

Other fusion methods do not perform so well as TADAM, but still surpass the BERT baseline.

Effects of Topic-aware Segmentation

In order to investigate the effectiveness of our segmentation algorithm in the response selection task, we use a simple method: just segment using fixed ranges 2, 4, 6, 8, 10. Besides, we also apply our segmentation algorithm with the corresponding range R. Results are shown in Figure 3.
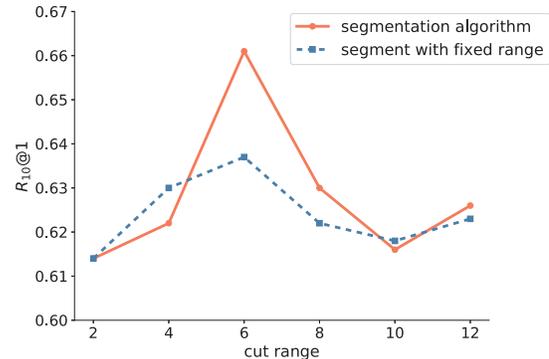

Figure 3: $R_{10}@1$ for different ranges on E-commerce.

For both methods, with the increase of cut range, range of 6 performs best. Both too small and large intervals hurt performance. Besides, applying our segmentation algorithm performs better than just using fixed ranges in most ranges especially in range of 6, which shows the effectiveness of our proposed topic-aware segmentation algorithm.

Conclusion

This paper presents the first topic-aware multi-turn dialogue modeling design in terms of explicitly segmenting and extracting topic-aware context utterances for retrieval-based dialogue systems. To fulfill our research purpose, we build two new datasets with topic boundary annotation and propose an effective topic-aware segmentation algorithm as prerequisites of this work. Then, we propose topic-aware dual-attention matching network to further improve the matching of response and segment contexts. Our models are evaluated on three benchmark datasets, showing new state-of-the-art, which verifies that the topic-aware clues and the related model design are effective.

14182

References

Ba, J. L.; Kiros, J. R.; and Hinton, G. E. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450* .

Budzianowski, P.; Wen, T.-H.; Tseng, B.-H.; Casanueva, I.; Ultes, S.; Ramadan, O.; and Gašić, M. 2018. MultiWOZ - A Large-Scale Multi-Domain Wizard-of-Oz Dataset for Task-Oriented Dialogue Modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP 2018)*, 5016–5026.

Chen, C.-Y.; Yu, D.; Wen, W.; Yang, Y. M.; Zhang, J.; Zhou, M.; Jesse, K.; Chau, A.; Bhowmick, A.; Iyer, S.; et al. 2018. Gunrock: Building A Human-Like Social Bot By Leveraging Large Scale Real User Data. *Alexa Prize Proceedings* .

Cho, K.; Van Merriënboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; and Bengio, Y. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078* .

Choi, F. Y. Y.; Wiemer-Hastings, P.; and Moore, J. 2001. Latent Semantic Analysis for Text Segmentation. In *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing (EMNLP 2001)*.

Clark, K.; Luong, M.-T.; Le, Q. V.; and Manning, C. D. 2020. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. In *ICLR*. URL https://openreview.net/pdf?id=r1xMH1BtvB.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* .

Eric, M.; Krishnan, L.; Charette, F.; and Manning, C. D. 2017. Key-Value Retrieval Networks for Task-Oriented Dialogue. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, 37–49.

Galley, M.; McKeown, K.; Fosler-Lussier, E.; and Jing, H. 2003. Discourse segmentation of multi-party conversation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics (ACL 2003)*, 562–569.

Gu, J.-C.; Li, T.; Liu, Q.; Ling, Z.-H.; Su, Z.; Wei, S.; and Zhu, X. 2020. Speaker-Aware BERT for Multi-Turn Response Selection in Retrieval-Based Chatbots. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management*, 2041–2044.

Gu, J.-C.; Ling, Z.-H.; and Liu, Q. 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*, 2321–2324.

Hearst, M. A. 1997. TextTiling: Segmenting text into multi-paragraph subtopic passages. *Computational linguistics* 23(1): 33–64.

Joty, S.; Carenini, G.; and Ng, R. T. 2013. Topic segmentation and labeling in asynchronous conversations. *Journal of Artificial Intelligence Research* 47: 521–573.

Kadlec, R.; Schmid, M.; and Kleindienst, J. 2015. Improved deep learning baselines for ubuntu corpus dialogs. *arXiv preprint arXiv:1510.03753* .

Lan, Z.; Chen, M.; Goodman, S.; Gimpel, K.; Sharma, P.; and Soricut, R. 2020. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. In *International Conference on Learning Representations (ICLR 2020)*.

LeCun, Y.; Bengio, Y.; and Hinton, G. 2015. Deep learning. *nature* 521(7553): 436–444.

Li, L.; Li, C.; and Ji, D. 2020. Deep context modeling for multi-turn response selection in dialogue systems. *Information Processing & Management* 102415.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* .

Lowe, R.; Pow, N.; Serban, I.; and Pineau, J. 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*, 285–294.

Morris, J.; and Hirst, G. 1991. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational linguistics* 17(1): 21–48.

Pevzner, L.; and Hearst, M. A. 2002. A critique and improvement of an evaluation metric for text segmentation. *Computational Linguistics* 28(1): 19–36.

Serban, I.; Klinger, T.; Tesauro, G.; Talamadupula, K.; Zhou, B.; Bengio, Y.; and Courville, A. 2017a. Multiresolution Recurrent Neural Networks: An Application to Dialogue Response Generation. In *Thirty-First AAAI Conference on Artificial Intelligence (AAAI 2017)*, 3288–3295.

Serban, I. V.; Sordoni, A.; Lowe, R.; Charlin, L.; Pineau, J.; Courville, A.; and Bengio, Y. 2017b. A hierarchical latent variable encoder-decoder model for generating dialogues. In *Thirty-First AAAI Conference on Artificial Intelligence (AAAI 2017)*, 32953302.

Song, Y.; Mou, L.; Yan, R.; Yi, L.; Zhu, Z.; Hu, X.; and Zhang, M. 2016. Dialogue Session Segmentation by Embedding-Enhanced TextTiling. In *INTERSPEECH*, 2706–2710.

Takanobu, R.; Huang, M.; Zhao, Z.; Li, F.; Chen, H.; Zhu, X.; and Nie, L. 2018. A Weakly Supervised Method for Topic Segmentation and Labeling in Goal-oriented Dialogues via Reinforcement Learning. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018*, 4403–4410.

Tan, M.; Santos, C. d.; Xiang, B.; and Zhou, B. 2015. Lstm-based deep learning models for non-factoid answer selection. *arXiv preprint arXiv:1511.04108* .

Tao, C.; Wu, W.; Xu, C.; Hu, W.; Zhao, D.; and Yan, R. 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*, 267–275.

Tao, C.; Wu, W.; Xu, C.; Hu, W.; Zhao, D.; and Yan, R. 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 (ACL 2019)*, 1–11.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in neural information processing systems*, 5998–6008.

Wan, S.; Lan, Y.; Xu, J.; Guo, J.; Pang, L.; and Cheng, X. 2016. Match-SRNN: Modeling the Recursive Matching Structure with Spatial RNN. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI 2016)*, 29222928.

Wang, S.; and Jiang, J. 2016. Learning Natural Language Inference with LSTM. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL 2016)*, 1442–1451.

Whang, T.; Lee, D.; Lee, C.; Yang, K.; Oh, D.; and Lim, H. 2020. An Effective Domain Adaptive Post-Training Method for BERT in Response Selection. In *Proc. Interspeech 2020*.

Wu, Y.; Li, Z.; Wu, W.; and Zhou, M. 2018a. Response selection with topic clues for retrieval-based chatbots. *Neurocomputing* 316: 251–261.

Wu, Y.; Wu, W.; Xing, C.; Zhou, M.; and Li, Z. 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)*, 496–505.

Wu, Y.; Wu, W.; Yang, D.; Xu, C.; and Li, Z. 2018b. Neural response generation with dynamic vocabularies. In *Thirty-Second AAAI Conference on Artificial Intelligence (AAAI 2018)*.

Xing, C.; Wu, W.; Wu, Y.; Liu, J.; Huang, Y.; Zhou, M.; and Ma, W.-Y. 2017. Topic Aware Neural Response Generation. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, 33513357.

Yan, R.; Song, Y.; and Wu, H. 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*, 5564.

Yuan, C.; Zhou, W.; Li, M.; Lv, S.; Zhu, F.; Han, J.; and Hu, S. 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)*, 111–120.

Zhang, S.; Zhao, H.; Wu, Y.; Zhang, Z.; Zhou, X.; and Zhou, X. 2019. DCMN+: Dual co-matching network for multi-choice reading comprehension. *arXiv preprint arXiv:1908.11511* .

Zhang, Z.; Chen, K.; Wang, R.; Utiyama, M.; Sumita, E.; Li, Z.; and Zhao, H. 2020a. Neural Machine Translation with Universal Visual Representation. In *International Conference on Learning Representations*. URL https://openreview.net/forum?id=Byl8hhNYPS.

Zhang, Z.; Li, J.; Zhu, P.; Zhao, H.; and Liu, G. 2018. Modeling Multi-turn Conversation with Deep Utterance Aggregation. In *Proceedings of the 27th International Conference on Computational Linguistics*, 3740–3752.

Zhang, Z.; Wu, Y.; Zhao, H.; Li, Z.; Zhang, S.; Zhou, X.; and Zhou, X. 2020b. Semantics-aware BERT for language understanding. In *the Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-2020)*.

Zhang, Z.; Wu, Y.; Zhou, J.; Duan, S.; Zhao, H.; and Wang, R. 2020c. SG-Net: Syntax-Guided Machine Reading Comprehension. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence*.

Zhang, Z.; Yang, J.; and Zhao, H. 2021. Retrospective Reader for Machine Reading Comprehension. In *The Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI-21)*.

Zhang, Z.; and Zhao, H. 2018. One-shot Learning for Question-Answering in Gaokao History Challenge. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING 2018)*.

Zhou, G.; Luo, P.; Cao, R.; Lin, F.; Chen, B.; and He, Q. 2017. Mechanism-aware neural machine for dialogue response generation. In *Thirty-First AAAI Conference on Artificial Intelligence (AAAI 2017)*, 34003408.

Zhou, H.; Huang, M.; Zhang, T.; Zhu, X.; and Liu, B. 2018a. Emotional chatting machine: Emotional conversation generation with internal and external memory. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Zhou, X.; Dong, D.; Wu, H.; Zhao, S.; Yu, D.; Tian, H.; Liu, X.; and Yan, R. 2016. Multi-view response selection for human-computer conversation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP 2016)*, 372–381.

Zhou, X.; Li, L.; Dong, D.; Liu, Y.; Chen, Y.; Zhao, W. X.; Yu, D.; and Wu, H. 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*, 1118–1127.

Zhu, P.; Zhang, Z.; Li, J.; Huang, Y.; and Zhao, H. 2018. Lingke: a Fine-grained Multi-turn Chatbot for Customer Service. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, 108–112.

Zhu, P.; Zhao, H.; and Li, X. 2020. Dual Multi-head Co-attention for Multi-choice Reading Comprehension. *arXiv preprint arXiv:2001.09415* .