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 by a large margin, especially by 3.4% on E-commerce dataset that has an obvious topic shift.

1 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; Tao et al. 2019a; Gu, Ling, and Liu 2019).

The retrieval-based response selection task is aimed to select a most suitable response from a collection of candidate answers according to a dialogue history. Early studies concatenate context together and then match with each candidate 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 by a large margin, especially by 3.4% on E-commerce dataset that has an obvious topic shift.

| 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 Corpus. The topic has changed after Turn-6.

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

Corresponding author. This paper was partially supported by National Key Research and Development Program of China (No. 2017YFB0304100) and Key Projects of National Natural Science Foundation of China (U1836222 and 61733011). Copyright © 2021, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.
et al. (2019) select semantic-relevant information or like Wu et al. (2018a) use topic information at word-level for better response 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.

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.

2 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, Carennini, 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 such as BERT (Devlin et al. 2018), ALBERT (Lan et al. 2020), RoBERTa (Liu et al. 2019), and then split out each utterance. For matching networks, there are various forms. For example, DUA (Zhang et al. 2018) presents a self-attention based deep utterance aggregation model to form a fine-grained context representation. 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.

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. Also, 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, the ‘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.

3 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 Section 5.4.

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
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: \hspace{1em} $l = u_{i-d} \oplus \cdots \oplus u_{i-1}$
4: \hspace{1em} $j = 1$, $c_0 = u^w$
5: \hspace{1em} while $i + j \leq n + 1$ and $j \leq R$ do
6: \hspace{2em} $c_j = c_{j-1} \oplus u_{i+j-1}$
7: \hspace{2em} if $j \mod K = 0$ then
8: \hspace{3em} $r_j = u_{i+j} \oplus \cdots \oplus u_{i+j+d-1}$
9: \hspace{3em} $\text{cost}_{c_j} = \max( \text{sim}(E(c_j), E(l)), \text{sim}(E(c_j), E(r_j)))$
10: \hspace{2em} $j + = 1$
11: \hspace{2em} $c_j^* = \min_{c_j} \text{cost}_{c_j}$
12: \hspace{1em} $S.append(c_j^*$), $i = |c_j^*| + 1$
return $S$

4 TADAM Network

4.1 Problem Formalization

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 $D = \{(C, R, Y)\}^n_{i=1}$, 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. Then, the response is used to match each weighted segment in a dual cross-attention way, 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 (we also add a linear layer here), and pass the final fused vector to a linear layer for a matching score.

4.2 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 $\{S_i\}^n_{i=1}$ and response $R$ with special tokens: $X = [CLS] S_1 [SEP] \cdots [SEP] S_n [SEP] R [SEP]$, 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 $\{S_i \in \mathbb{R}^{L \times d}\}^T_{i=1}$. More encoding modes are discussed in Section 6.2.

4.3 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:

$$M^1 = \text{transfer}(C, W, R),$$

$$M = \frac{1}{\sqrt{d}} \tanh(M^1)V,$$

$$M_{\text{pool}} = \left[ \max_{\text{dim}=2} \frac{M}{\text{dim}=3} \right],$$

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

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

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.

**Combination:** In order to fuse both two levels, we set a hyper-parameter $\alpha$ and sum up $s_1, s_2$. Then we multiply $s$...
Topic segments & Response  Weighting  Dual Cross-attention Matching  Aggregation

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

4.4 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 processing.

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 sequence via Scaled Dot-Product Attention (Vaswani et al. 2017), then applies those attention results upon the value sentence, which is defined as:

\[ V_{att} = \text{softmax}(\frac{QK^T}{\sqrt{d}})V, \]

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 feedforward 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:

\[ \text{AttentiveModule}(Q, K, V), \]

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

\[ S^\text{cross}_i = \text{AttentiveModule}(\tilde{S}_i, R, R), \]
\[ R^\text{cross}_i = \text{AttentiveModule}(R, \tilde{S}_i, \tilde{S}_i), \]

where \( S^\text{cross}_i, R^\text{cross}_i \in \mathbb{R}^{L \times d} \). All the \( S^\text{cross}_i, R^\text{cross}_i \) construct \( C^\text{cross}_S \in \mathbb{R}^{T \times L \times d}, C^\text{cross}_R \in \mathbb{R}^{T \times L \times d} \), which are then mean-pooled and concatenated as \( C^\text{cross} \) (other pooling and fusion methods are discussed in Section 6.3):

\[ C^\text{cross}_S = \text{mean}(C^\text{cross}_S), \text{mean}(C^\text{cross}_R), \]

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

4.5 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 \( C^\text{cross}_T \), which is the dual matching between the last segment \( S^\text{cross}_T \) and the response \( R^\text{cross}_T \). Furthermore, we add a linear layer to \( C^\text{cross}_T \):

\[ \hat{H} = \text{GRU}(C^\text{cross}_T), \]
\[ C^\text{cross}_T = \text{mean}(S^\text{cross}_T), \text{mean}(R^\text{cross}_T), \]
\[ \hat{C}^\text{cross}_T = W_3C^\text{cross}_T + b_3, \]

where \( \hat{H}, C^\text{cross}_T, \hat{C}^\text{cross}_T \in \mathbb{R}^{d}, W_3 \in \mathbb{R}^{2d \times 2d}, b_3 \in \mathbb{R}^{d} \) are learnable parameters. Finally, we get the score for context \( C \) and its candidate response \( r \) using a linear layer, where \( W_4 \in \mathbb{R}^{d}, b_4 \in \mathbb{R} \) are learnable parameters (other fusion methods are discussed in Section 6.3):

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

5.1 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 and Stanford Dialog Dataset, where each dialogue is about one topic.

For response selection, TADAM is tested on three widely used public datasets. (1) Ubuntu Corpus: consists of English multi-turn conversations about technical support collected from chat logs of the Ubuntu forum. (2) Douban Corpus: consists of multi-turn conversations from the Douban group, which is a popular social networking service in China. (3) E-commerce Corpus: 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.

Statistics about all datasets can be found in Appendix.

5.2 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 $R_0\times k$ 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.

5.3 Experimental Settings
For both tasks, we use pre-trained BERT$^5$ (Devlin et al. 2018) (bert-base-uncased & bert-base-chinese) as encoder.

For topic-aware segmentation: In both datasets, we set $R = 8$, jump step $K = 2$, window size $d = 2$ and threshold $\theta_{cost} = 0.6$. For our baseline 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.

For response selection: We apply topic-aware segmentation algorithm to Ubuntu, Douban and E-commerce with range $R = 2, 6$. 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 $\alpha$ of word-level weights is set to 0.5. As for the baseline BERT, we have finetuned it. The epoch is set as $\{3, 2, 3\}$ for three datasets, and other settings are the same as our model.

5.4 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     | 2.00    | 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. Segmentation statistics on test sets are shown in Appendix. Because the average range $R$ of Ubuntu and Douban is small, most dialogues are segmented.

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 significantly outperforms all public works and especially gets much improvement (3.4% in $R_{10}\times 0.1$) over the strong

---

5 https://doi.org/10.17863/CAM.41572
6 https://www.douban.com/group
7 https://www.taobao.com
8 https://github.com/huggingface/transformers
we concatenate the multi-turn matching result (line 3-5). Second, in the aggregation part (in Section 4.5), by moving word or segment level weights, or both of them, the influence of the segment weighting part (in Section 4.3) is shown in Table 4. First, we explore the model, we conduct serious ablation experiments from three aspects and results are shown in Table 5 (line 3,4). Third, we do dual segment matching result (line 6-7). Third, we do dual cross-attention matching which includes cross-attentioned segments \(C^T_{\text{cross}}\) and cross-attentioned responses \(C^R_{\text{cross}}\) 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 effectiveness of our dual matching design. Results of Ubuntu are not so obvious as that in 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.

### 6 Analysis

#### 6.1 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 Section 4.3) by moving word or segment level weights, or both of them (line 3-5). Second, in the aggregation part (in Section 4.5), we concatenate the multi-turn matching result \(\hat{H}\) and last segment matching result \(\hat{C}^T_{\text{cross}}\) to get a score. Hence we remove either of both time (line 6-7). Third, we do dual cross-attention matching which includes cross-attentioned segments \(C^T_{\text{cross}}\) and cross-attentioned responses \(C^R_{\text{cross}}\) 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 effectiveness 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.

#### 6.2 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
Table 5: Results of different input modes (line 3,4) and fusion methods (line 5,6,7).

| Model                      | $R_{10}@1$ | $R_{10}@2$ | $R_{10}@5$ |
|----------------------------|------------|------------|------------|
| TADAM                      | 66.1       | 83.4       | 97.5       |
| separated segment          | 36.0       | 49.2       | 72.5       |
| concatenated utterance     | 61.9       | 83.3       | 96.8       |
| 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 6 shows a topic-aware segmentation case from E-commerce Corpus. We can 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.

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 significantly, 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, where we can observe that topic segment does work in dialogues of different lengths.

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

6.4 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.

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.

7 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.
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).

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.; 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 dialogues. 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.

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.; Tesauo, 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), 3295–3302.

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, L.; 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), 2922–2928.

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.

Wu, H.; 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.

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.
A Appendix

A.1 Data Statistics
Statistics about our Chinese and English datasets for topic-aware segmentation task are shown in Table 7. For response selection task, TADAM is tested on three widely used public datasets, whose statistics are shown in Table 8. Note that all our experiments except for TADAM-Douban are conducted in one NVIDIA RTX 2080 24G GPU. TADAM-Douban is done on two GPUs for its batch size is 32.

| Dataset Statistics       | Chinese | English |
|--------------------------|---------|---------|
| #Dialogues               | 505     | 711     |
| #Segments                | 2019    | 3465    |
| Avg. Seg./Dial.          | 4.0     | 4.87    |
| Avg. Utter./Seg.         | 6.37    | 5.58    |
| Avg. Tokens/Utter.       | 19.56   | 12.43   |

Table 7: Topic-aware segmentation.

| Dataset Statistics       | Ubuntu       | Douban       | E-commerce   |
|--------------------------|--------------|--------------|--------------|
|                         | Train  | Valid | Test | Train  | Valid | Test | Train  | Valid | Test |
| #Context-response Pairs  | 1M     | 500K  | 500K | 1M     | 50K   | 50K  | 1M     | 10K   | 10K  |
| #Candidates/Context      | 2      | 10    | 10   | 2      | 2     | 10   | 2      | 2     | 10   |
| Avg. Turns/Context       | 10.13  | 10.11 | 10.11| 6.69   | 6.75  | 6.45 | 5.51   | 5.48  | 5.64 |
| Avg. Tokens/Utterance    | 11.35  | 11.34 | 11.37| 18.56  | 18.50 | 20.74| 7.02   | 6.99  | 7.11 |

Table 8: Response selection.

A.2 Segmentation statistics on response selection datasets
We apply the segmentation algorithm to dialogues on Ubuntu, Douban and E-commerce datasets to get inputs of TADAM. Segmentation statistics on test sets are shown in Table 9. Because the judge range $R$ of Ubuntu and Douban is small, most dialogues are segmented.

| Statistics                | Ubuntu | Douban | E-commerce |
|---------------------------|--------|--------|------------|
| #Dialogues                | 50K    | 1k     | 1K         |
| Avg. Seg./Dial.           | 4.76   | 3.02   | 1.52       |
| Avg. Utter./Seg.          | 2.13   | 2.15   | 3.71       |
| Avg. Utter./Dial.         | 10.11  | 6.48   | 5.64       |
| Segmented Dial.           | 84.4%  | 99.9%  | 22.1%      |

Table 9: Segmentation statistics on test sets.