A Compare Aggregate Transformer for Understanding Document-grounded Dialogue

Longxuan Ma, Weinan Zhang, Runxin Sun, Ting Liu

{lxma,wnzhang,rxsun,tliu}@ir.hit.edu.cn
Research Center for Social Computing and Information Retrieval, Harbin Institute of Technology
Harbin, Heilongjiang, China

ABSTRACT
Unstructured documents serving as external knowledge of the dialogues help to generate more informative responses. Previous research focused on knowledge selection (KS) in the document with dialogue. However, dialogue history that is not related to the current dialogue may introduce noise in the KS processing. In this paper, we propose a Compare Aggregate Transformer (CAT) to jointly denoise the dialogue context and aggregate the document information for response generation. We designed two different comparison mechanisms to reduce noise (before and during decoding). In addition, we propose two metrics for evaluating document utilization efficiency based on word overlap. Experimental results on the CMUDoG dataset show that the proposed CAT model outperforms the state-of-the-art approach and strong baselines.

CCS CONCEPTS
- Computing methodologies → Natural language generation; Discourse, dialogue and pragmatics.

KEYWORDS
Dialogue system, Natural language generation, knowledge selection

1 INTRODUCTION
Dialogue system (DS) attracts great attention from industry and academia because of its wide application prospects. Sequence-to-sequence models (Seq2Seq) [24, 26] are verified to be an effective framework for the DS task. However, one problem of Seq2Seq models is that they tended to generate generic responses that provides deficient information Ghazvininejad et al. [5], Li et al. [10]. Previous researchers proposed different methods to alleviate this issue. One way is to focus on models’ ability to extract information from conversations. Li et al. [10] introduced Maximum Mutual Information (UMI) as the objective function for generating diverse response. Serban et al. [25] proposed a latent variable model to capture posterior information of golden response. Zhao et al. [33] used conditional variational autoencoders to learn discourse-level diversity for neural dialogue models. The other way is introducing external knowledge, either unstructured knowledge texts Dinan et al. [4], Ghazvininejad et al. [5], Ye et al. [30] or structured knowledge triples [13, 31, 36] to help open-domain conversation generation by producing responses conditioned on selected knowledge.

The Document-grounded Dialogue (DGD) [11, 34, 37] is a new way to use external knowledge. It establishes a conversation mode in which relevant information can be obtained from the given document. The DGD systems can be used in scenarios such as talking over merchandise against the product manual, commenting on news reports, etc. One example of DGD is presented in Table 1. Two interlocutors talk about the given document and freely reference the text segment during the conversation.

To address this task, two main challenges need to be considered in a DGD model: 1) Determining which of the historical conversations are related to the current conversation, 2) Using current conversation and the related conversation history to select proper document information and to generate an informative response. Previous work Arora et al. [2], Qin et al. [20], Ren et al. [21], Tian et al. [27], Zhao et al. [34] generally focused on selecting knowledge with all the conversations. However, the relationship between historical conversations and the current conversation has not been studied enough. For example, in Table 1, the italics utterance from user1, "Yes, his style really extended the story.", is related to dialogue history. While the black fold utterance from user1, "Sally Hawkins acting was phenomenally expressive. Didn’t feel her character was mentally handicapped.", has no direct relationship with the historical utterances. When employing this sentence as the last utterance, the dialogue history is not conducive to generate a response.

In this paper, we propose a novel Transformer-based [28] model for understanding the dialogues and generate informative responses in the DGD, named Compare Aggregate Transformer (CAT). Previous research [22] has shown that the last utterance is the most important guidance for the response generation in the multi-turn
When the dialogue history is less relevant to the last utterance, we need to consider all the conversations to filter the document information. Otherwise, the existence of dialogue history is equal to the introduction of noise, and its impact should be eliminated conditionally. For this purpose, on one side, the CAT filters the document information with the last utterance; on the other side, the CAT uses the last utterance to guide the dialogue history and employs the guiding result to filter the given document. We judge the importance of the dialogue history by comparing the two parts, then aggregate the filtered document information to generate the response. Experimental results show that our model can generate more relevant and informative responses than competitive baselines.

When the dialogue history is less relevant to the last utterance, our model is verified to be even more effective. The main contributions of this paper are:

(1) We propose a compare aggregate method to determine the relationship between the historical dialogues and the last utterance. Experiments show that our method outperforms strong baselines on the CMUDoG dataset.

(2) We propose two new metrics to evaluate the document knowledge utilization in the DGD. They are both based on N-gram overlap among generated response, the dialogue, and the document.

2 RELATED WORK

The DGD maintains a dialogue pattern where external knowledge can be obtained from the given document. Most recently, some DGD datasets Gopalakrishnan et al. [6], Moghe et al. [18], Qin et al. [20], Zhou et al. [37] have been released to exploiting unstructured document information in conversations.

Models trying to address the DGD task can be classified into two categories based on their encoding process with dialogues: one is parallel modeling and the other is incremental modeling. For the first category, Moghe et al. [18] used a generation-based model that learns to copy information from the background knowledge and a span prediction model that predicts the appropriate response span in the background knowledge. Liu et al. [14] claimed the first to unify knowledge triples and long texts as a graph. Then employed a reinforce learning process in the flexible multi-hop knowledge graph reasoning process. To improve the process of using background knowledge, [32] firstly adopted the encoder state of the utterance history context as a query to select the most relevant knowledge, then employed a modified version of BiDAF [23] to point out the most relevant token positions of the background sequence. Meng et al. [16] used a decoding switcher to predict the probabilities of executing the reference decoding or generation decoding. Some other researchers [2, 16, 20, 21, 34] also followed this parallel encoding method. For the second category, Kim et al. [7] proposed a sequential latent knowledge selection model for Knowledge-Grounded Dialogue. Li et al. [11] designed an incremental transformer to encode multi-turn utterances along with knowledge in the related document. Meanwhile, a two-way deliberation decoder [29] was used for response generation. However, the relationship between the dialogue history and the last utterance is not well studied. In this paper, we propose a compare aggregate method to investigate this problem. It should be pointed out that when the target response changes the topic, the task is to detect whether the topic is ended and to initiate a new topic [1]. We do not study the conversation initiation problem in this paper, although we may take it as future work.

3 THE PROPOSED CAT MODEL

3.1 Problem Statement

The inputs of the CAT model are the given document \( D = (D_1, D_2, ..., D_d) \) with \( d \) words, dialogue history \( H = (H_1, H_2, ..., H_h) \) with \( h \) words and the last utterance \( L = (L_1, L_2, ..., L_l) \) with \( l \) words. The task is to generate the response \( R = (R_1, R_2, ..., R_r) \) with \( r \) tokens with probability:

\[
P(R|H, L, D; \Theta) = \prod_{i=1}^{r} P(R_i|H, L, D, R_{<i}; \Theta),
\]

where \( R_{<i} = (R_1, R_2, ..., R_{i-1}) \), \( \Theta \) is the model’s parameters.

3.2 Encoder

The structure of the CAT model is shown in Figure 1. The hidden dimension of the CAT model is \( h \). We use the Transformer structure [28]. The self-attention is calculated as follow:

\[
\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V,
\]

where \( Q, K, \) and \( V \) are the query, the key, and the value, respectively; \( d_k \) is the dimension of \( Q \) and \( K \). The encoder and the decoder stack \( N \) (\( N = 3 \) in our work) identical layers of multihead attention (MAtt):

\[
\text{MAtt}(Q, K, V) = [A_1, ..., A_n]W_Q^O.
\]

\[
A_i = \text{Attention}(QW_Q^O, KW_K^K, VW_V^K),
\]

where \( W_Q^O, W_K^K, W_V^K \) (\( i = 1, ..., n \)) and \( W_Q^O \) are learnable parameters.

The encoder of CAT consists of two branches as figure 1 (a). The left branch learns the information selected by dialogue history \( H \), the right part learns the information chosen by the last utterance \( L \). After self-attention process, we get \( H_s = \text{MAtt}(H, H, H) \) and \( L_s = \text{MAtt}(L, L, L) \). Then we employ \( L_s \) to guide the \( H, H^1 = \text{MAtt}(H_s, H, H), \) where \( H^1 \) is the hidden state at the first layer. Then we adopt \( H^1 \) to select knowledge from the document \( D \), \( D^1 = \text{FF}(\text{MAtt}(H^1, D, D)) \). FF is the feed-forward process in the second layer, \( D^1 \) is the input, \( D^1 = \text{MAtt}(D, D^1, D^1), H^2 = \text{MAtt}(D^1, H, H), \) \( D^2 = \text{FF}(\text{MAtt}(H^2, D, D)) \). After \( N \) layers, we obtain the information \( D^n \) selected by \( H \). In the right branch, we use \( L_s \) to filter the \( D, D^n \) is the information selected by \( L \).

3.3 Comparison Aggregate

As demonstrated by [22], the last utterance played a fundamental role in response generation. We need to preserve the document information filtered by \( L \), and determine how much information

---

1The code and data will be released in Github.
3.3.1 Concatenation. We use average pooling to \( H_s \) and \( L_s \) to get their vector representations \( H_s \alpha \) and \( L_s \alpha \in \mathbb{R}^{h \times 1} \), respectively. The concatenation method calculates relevance score \( \alpha \) to determine the importance of \( D^n \) as follow:

\[
\alpha = \tanh(H_s \alpha W^H + L_s \alpha W^L), \tag{5}
\]

\[
D_{final} = [\text{sigmoid}(W^s \alpha) \ast D^n; \widetilde{D}^n], \tag{6}
\]

where \( W^H, W^L \in \mathbb{R}^{h \times r}, W^s \in \mathbb{R}^{1 \times h} \) are learnable parameters. \([X, Y]\) is the concatenation of \( X \) and \( Y \) in sentence dimension, \( \ast \) is the element-wise multiplication. Note that the \( D^n \) is guided by \( H \), the concatenation method performs a second level comparison with \( H \) and \( L \) and then transfers the topic-aware \( D_{final} \) to the two-pass Deliberation Decoder (DD) [29]. The structure of the DD is shown in Figure 1 (b). The first-pass takes \( L \) and \( D_{final} \) as inputs and learns to generate a contextual coherently response \( R^1 \). The second-pass takes \( R^1 \) and the document \( D \) as inputs and learns to inject document knowledge. The DD aggregates document, conversation, and topic information to generate the final response \( R^2 \). Loss is from both the first and the second layers:

\[
L = - \sum_{m=1}^{M} \sum_{i=1}^{r} (\log P(R^1_i) + \log P(R^2_i)), \tag{7}
\]

where \( M \) is the total training example; \( R^1_i \) and \( R^2_i \) are the \( i \)-th word generated by the first and second decoder layer, respectively.

3.3.2 Attended Comparison. We employ an Enhanced Decoder [35] to perform the attended comparing. The structure of our Enhanced Decoder is illustrated in Figure 1 (c). It accepts \( D^n \), \( \widetilde{D}^n \) and the response \( R \) as inputs, applying a different way to compare and aggregate. The merge attention computes weight across all inputs:

\[
P = [R, D^n; \widetilde{D}^n]W_p, \tag{8}
\]

\[
V_{merge} = P_R R + P_D D^n + P_{\widetilde{D}} \widetilde{D}^n, \tag{9}
\]

where \( W_p \) is learnable parameters. The dimension of \( P \) is \( 3, P_R, P_D \) and \( P_{\widetilde{D}} \) are the Softmax results of \( P \), \( V_{merge} \) and \( L \) are used for next utterance attention as shown in Figure 1 (c). The output of the Enhanced Decoder is connected to the second layer of DD and we define this new structure as Enhanced Deliberation Decoder (EDD). The loss is the same as Eq. (7).

4 EXPERIMENTS
4.1 Dataset
We evaluate our model with the CMUDoG [37] dataset. There are 4112 dialogs based on 120 documents in the dataset. One document contains 4 sections, such as movie introduction and scenes. A related section is given for every several consequent utterances. However, the conversations are not constrained to the given section. In our setting, we use the full document (with 4 section) as external knowledge. The average length of documents is around 800 words. We concatenate consequent utterances of the same person as one utterance. When training, we remove the first two or three rounds of greeting sentences. Each sample contains one document, two or more historical utterances, one last utterance, and one golden response. When testing, we use two different versions of the test set. The first follows the process of training data, we name it Reduced version. The second is constructed by comparing the original document section of the conversation based, we preserve the examples that the dialogue history and the last utterance are based on different document sections. For example, dialogue history is based on section 2, the last utterance and response are based on section 3. We name it Sampled version and it is used for testing our models’ comprehending ability of the topic transfer in conversations. The
data statistics are shown in Table 2. Please refer to Zhou et al. [37] for more details. It is worth noting that the sampled version does not represent the proportion of all conversation topic transitions, but it demonstrates this problem better than the Reduced version.

### 4.2 Baselines

We evaluated several competitive baselines.

#### 4.2.1 RNN-based models.

VHRED: A Hierarchical Latent Variable Encoder-Decoder Model [25], which introduces a global (semantic level) latent variable \( Z \) for the problem that HRED [24] is difficult to generate meaningful and high-quality replies. \( Z \) is calculated with the encoder RNN outputs and the context RNN outputs. The latent variable \( Z \) contains some high-level semantic information, which encourages the model to extract abstract semantic concepts. Please refer to Serban et al. [25] for more details. We use \( Z \) to capture the topic transfer in conversations and test three different settings. For the first setting, we do not employ the document knowledge, only use dialogue as input to generate the response. It is recorded as VHRED-(k). For the second one, we use the same encoder RNN with shared parameters to learn the representation of the document and the utterance, then concatenate the final hidden state of them as the input of the context RNN. It is denoted by VHRED(c). For the third one, we use word-level dot-attention [15] to get the document-aware utterance representation and use it as the input of context RNN. It is termed as VHRED(a).

#### 4.2.2 Transformer-based models.

T-DD/T-EDD: They both use the Transformer as the encoder. The inputs are the concatenation of dialogues and the document. These two models parallel encode the dialogue without detecting topic transfer. The T-DD uses a Deliberation Decoder (DD) as the decoder. The T-EDD uses an Enhanced Deliberation Decoder (EDD) as the decoder.

ITDD [11]: It uses Incremental Transformer Encoder (ITE) and two-pass Deliberation Decoder (DD). Incremental Transformer uses multi-head attention to incorporate document sections and context into each utterance. Encodings encoding process. ITDD incrementally models dialogues without detecting topic transitions.

### 4.3 Evaluation Metrics

#### Automatic Evaluation

We employ perplexity (PPL) [3], BLEU [19] and ROUGE [12]. The PPL of the gold response is measured, lower perplexity indicates better performance. BLEU measures the n-gram overlap between a generated response and a gold response. Since there is only one reference for each response, BLEU scores are extremely low. ROUGE measures the n-gram overlap based on the recall rate. Since the conversations are constrained by the background material, ROUGE is reliable.

#### Knowledge Utilization (KU)

We also introduce two metrics to automatically evaluate the Knowledge Utilization (KU), they are both based on N-grams overlaps. We define one document, conversations and generated response in Test set as \((D,C,R)\). The N-grams set of each \((D,C,R)\) are termed as \(G_d^N, G_c^N, \) and \(G_r^N\), respectively. The number of the overlapping N-grams of \(G_d^N\) and \(G_r^N\) is recorded as \(G_{dr}^N\). Tuples which are in \(G_d^N\) but not in \(G_c^N\) is named \(G_{dr-c}^N\). Then \(KU = \frac{\text{len}(G_{dr-c}^N)}{\text{len}(G_{dr}^N)}\) reflects how many \(N\)-grams in the document are used in the generated replies, \(\text{len}(G)\) is the tuple number in G. The larger the KU is, the more \(N\)-grams of the document is utilized. Since low-frequency tuples may be more representative of text features, we define the reciprocal of the frequency of each tuple \(k\) in \(G\) as \(R_k^G\), which represents the importance of a tuple. Then the Quality of Knowledge Utilization (QKU) is calculated as:

$$QKU = \frac{\sum_{k \in G_{dr-c}^N} R_k^G}{\sum_k R_k^G}, \quad k \in G_{dr-c}^N.$$  \hfill (10)

If \(R_k^G\) is more important in response and \(R_k^{G_{dr-c}}\) is less important in document, the QKU will become even larger. So the smaller QKU means the higher quality of the used document knowledge.

#### Human Evaluation

We randomly sampled 100 conversations from the Sampled test set and obtained 800 responses from eight models. We have 5 graduate students as judges. They score each response with access to previous dialogues and the document. We use three metrics: Fluency, Coherence, and Informativeness. Fluency measures whether the response is a human-like utterance. Coherence measures if the response is coherent with the dialogue context. Informativeness measures if the response contains relevant and correct information from the document. They are scored from 1 to 5 (1:very bad, 2:bad, 3:acceptable, 4:good, 5:very good). Overall inter-rater agreement measured by Fliess’s Kappa is 0.32 ("fair").

### 4.4 Experimental Setup

We use OpenNMT-py [9] as the code framework. For all models, the pre-trained 300 dimension word embedding [17] is shared by dialogue, document, and generated responses, the dimension of the hidden size is 300. For the RNN-based models, 3-layer bidirectional GRU and 3-layer GRU are applied for encoder and decoder, respectively. For the Transformer-based models, the layers of both encoder and decoder are set to 3, the number of heads in multi-head attention is 8 and the filter size is 2048. We use Adam (\(\alpha = 0.001, \beta_1 = 0.9, \beta_2 = 0.999, \) and \(\epsilon = 10^{-8}\)) [8] for optimization. The beam size is set to 5 in the decoder. We truncate the words of the document to 800 and the dialogue utterance to 40. All models are trained on a Titan X (Pascal) GPU. The average training time per epoch is around 40 minutes for the Transformer-based models and around 20 minutes for the RNN-based models.

### 5 ANALYSIS

#### 5.1 Experimental Results study

Table 3 shows the automatic evaluations for all models on the Reduced (Sampled) dataset. The dialogue history is 2 rounds. We only present ROUGE-L as ROUGE-1/2 show the same trend as ROUGE-L. Through experiments, we can see that the change range

| Dataset   | Utterances (train / dev / test) | Word/Utterance |
|-----------|---------------------------------|----------------|
| Original  | 72922 / 3626 / 11577            | 18.6           |
| Reduced   | 66332 / 3269 / 10502            | 19.7           |
| Sampled   | 66332 / 3269 / 1317             | 19.6           |

Table 2: Statistics of the CMUDoG dataset.
of KU-2 (8.0–13.7) is less than KU-3 (23.1–31.7) on the Reduced data, indicating that the KU-3 can better reflect the amount of knowledge used than KU-2.

In the RNN-based models, the VHRED(-k) gets the worst PPL/BLEU/ROUGE, which reveals the importance of injecting document knowledge in the DGD task. We did not calculate the KU/QKU of the VHRED(-k) since the model did not use document knowledge. The VHRED(a) gets better PPL/BLEU/ROUGE/QKU than the VHRED(c) model, which means the smaller granular extraction of document information benefits more in generating responses.

Among the Transformer-based models, the ITDD model gets better PPL/BLEU/ROUGE/QKU than the T-DD model, which means the incremental encoding method is stronger than parallel encoding. The CAT-EDD and the CAT-DD models achieve better performance than the T-DD and the T-EDD models, respectively. It indicates that our Compare-Aggregate method is helpful to understand the dialogue. The CAT-EDD model outperforms the ITDD model on all metrics, which indicates that our CAT module automatically learns the topic transfer between conversation history and the last utterance as expected. The CAT-DD does not perform as good as the CAT-DD, which shows that it is necessary to set up an independent mechanism to learn topic transfer, rather than automatic learning by attentions in the decoder.

Comparing with the RNN-based models, the Transformer-based models get better performance on PPL/BLEU/ROUGE. It proves that the latter is better in the ability of convergence to the ground truth. The VHRED(c) and the VHRED(a) get better KU and worse QKU than the Transformer-based models. It means that the latent variable models increase the diversity of replies and use more document tuples, but their ability to extract unique tuples is not as good as the Transformer-based ones.

Table 4 shows the manual evaluations for all models on the Reduced/Sampled data. We take the CAT-DD as the base model to do the significant test, ⋄ and * stand for p < 0.05 and p < 0.01, respectively.

Table 5 shows the manual evaluations for all models on the Reduced/Sampled dataset. The CAT-DD model gets the highest scores on Fluency/Coherence/Informativeness. When experimenting with the Sampled test set, we can see that the advantages of our models become greater than the results of the Reduced version in both automatic and manual evaluations. Our model shows more advantages in datasets with more topic transfer.

### 5.2 Ablation Study

Table 5 illustrates the ablation study of the CAT-DD model. w/o-left means the left branch is removed and the model degenerates to T-DD which takes the last utterance and document as inputs. We can see that all the automatic evaluation indexes significantly reduce, indicating the dialogue history can not be simply ignored. w/o-(5,6) is a model without Eq. (5) and (6), which is equivalent to simply connect the outputs of the left and the right encoder branches. The results showed that the ability of the model to distinguish the conversation topic transfer is weakened. w/o-(G) is a model removing the utter-attention in the left branch, which means we do not use L to guide the H, the structure of left branch changes to the right branch and the input is H. The performance is declining, which indicates that the guiding process is useful. The significant tests (two-tailed student t-test) on PPL/BLEU/QKU-2/QKU-2 reveal the effectiveness of each component.

### 5.3 History Round Study

We use the CAT-DD model and the Sampled test set to study the influence of the historical dialogue rounds. For example, setting dialogue history to 0 means we use only the last utterance, the CAT-DD becomes the w/o-left model in the ablation study. Setting
dialogue history to \( N \) means we use \( N \) rounds of dialogue history for the input of the left branch. We set the conversation history to 0/1/2/3/4 to test the response of VHRED(a)/ITDD/CAT-DD models. Figure 2 shows the trend of BLEU/KU-3/QKU-3. The top figure shows the BLEU trend, the CAT-DD reaches the maximum when the rounds are 2. The continuous increase of rounds does not significantly improve the generation effect. In the middle picture, with the increase of historical dialogue from 0 to 2, the VHRED(a) and the CAT-DD have a visible improvement on the KU-3, which shows that the information contained in the historical dialogue can be identified and affect the extraction of document information. The ITDD model is not as sensitive as the others on the KU-3, indicating that the incremental encoding structure pays more attention to the information of the last utterance. The bottom figure shows the trend of the KU-3. When the history dialogue increases, the ITDD model keeps stable and the VHRED(a) and the CAT-DD models have a declining trend, which again indicates that the VHRED(a) and the CAT-DD are more sensitive to the historical dialogue.

![Figure 2: The effect of dialogue history rounds on VHRED(a)/ITDD/CAT-DD models. The abscissa represents the historical dialogue rounds. The ordinate represents the BLEU/KU-3/QKU-3 values.](image)

In Table 6, we randomly select an example in the Sampled test set for a case study. The document, the dialogue history, the last utterance, and the ground truth are presented. We can observe that the last utterance is irrelevant to the dialogue history. The generated responses of different models are listed below. The VHRED(a) and CAT-DD(w/o-(G)) models misunderstand the dialogue and use the wrong document knowledge. The TDD gives a generic reply. The ITDD model answers correctly but without enough document information. The CAT-DD(w/o-(5,6)) model gives a response that was influenced by the irrelevant historical dialogue which we want to eliminate. Only the CAT-DD model generates a reasonable reply.

In Figure 3, we show the average sigmoid(\( \frac{1}{\alpha} \)) value in the CAT-DD model over the Reduced/Sampled test set and the Validation set. A higher value means a stronger correlation between the last utterance and the historical dialogue. We can see that on the Reduced test set and the Validation set, the relevance score is higher than that of the Sampled data, which proves that the last utterance and the historical dialogue are more irrelevant in the latter. Our model captures this change and performs better on the Sampled data than the Reduced data. When the historical rounds increase from 1 to 2, the relevance score reduces obviously for all data sets, which means the increase of dialogue history introduces more unrelated information. When the historical conversations increases from 2 to 6, all data have no significant change, indicating that increasing the dialogue rounds does not improve the recognition ability of the model to the topic change.

![Figure 3: The rating of dialogue history in the CAT-DD model with Reduced and Sampled test set. The abscissa represents the dialogue rounds and the ordinate represents the correlation score in the model.](image)

### 5.4 History Importance Study

Figure 3 shows the average sigmoid(\( \frac{1}{\alpha} \)) value in the CAT-DD model over the Reduced/Sampled test set and the Validation set. A higher value means a stronger correlation between the last utterance and the historical dialogue. We can see that on the Reduced test set and the Validation set, the relevance score is higher than that of the Sampled data, which proves that the last utterance and the historical dialogue are more irrelevant in the latter. Our model captures this change and performs better on the Sampled data than the Reduced data. When the historical rounds increase from 1 to 2, the relevance score reduces obviously for all data sets, which means the increase of dialogue history introduces more unrelated information. When the historical conversations increases from 2 to 6, all data have no significant change, indicating that increasing the dialogue rounds does not improve the recognition ability of the model to the topic change.

### 5.5 Case Study

In Table 6, we randomly select an example in the Sampled test set for a case study. The document, the dialogue history, the last utterance, and the ground truth are presented. We can observe that the last utterance is irrelevant to the dialogue history. The generated responses of different models are listed below. The VHRED(a) and CAT-DD(w/o-(G)) models misunderstand the dialogue and use the wrong document knowledge. The TDD gives a generic reply. The ITDD model answers correctly but without enough document information. The CAT-DD(w/o-(5,6)) model gives a response that was influenced by the irrelevant historical dialogue which we want to eliminate. Only the CAT-DD model generates a reasonable reply.
and uses the correct document knowledge, which means it correctly understands the dialogues.

6 CONCLUSION

We propose the Compare Aggregate method to understand Document-grounded Dialogue (DGD). The dialogue is divided into the last utterance and the dialogue history. The relationship between the two parts is analyzed to denoise the dialogue context and aggregate the document information for response generation. Experiments show that our model outperforms previous work in both automatic and manual evaluations. Our model can better understand the dialogue context and select proper document information for response generation. We also propose Knowledge Utilization (KU) and Quality of Knowledge Utilization (QKU), which are used to measure the quantity and quality of the imported external knowledge, respectively. In the future, we will further study the topic transition problem and the knowledge injecting problem in the DGD.

REFERENCES

[1] Satoshi Akasaka and Nobuhiro Kaji. 2019. Conversation Initiation by Diverse News Contents Introduction. In NAACL-HLT (1). Association for Computational Linguistics, 3988–3998.
[2] Siddharta Arora, Mitesh M. Khapra, and Harish G. Ramaswamy. 2019. On Knowledge distillation from complex networks for response prediction. In NAACL-HLT (1). Association for Computational Linguistics, 3813–3822.
[3] Yoshua Bengio, Réjean Ducharme, and Pascal Vincent. 2000. A Neural Probabilistic Language Model. In NIPS MIT Press, 932–938.
[4] Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Kuester. 2020. A Knowledge-Grounded Neural Variable Encoder-Decoder Model for Generating Dialogues. In Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA. 3776–3784. http://www.aaai.org/ocs/index.php/AAAI/AAAI16/paper/view/11957
[5] Min Joon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2017. Bidirectional Attention Flow for Machine Comprehension. In ICLR (Poster) OpenReview.net
[6] Iulian Vlad Serban, Alessandro Sordoni, Yoshua Bengio, Aaron C. Courville, and Joelle Pineau. 2016. Building End-To-End Dialogue Systems Using Generative Hierarchical Neural Network Models. In Proceedings of the Thirteenth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA. 3776–3784. http://www.aaai.org/ocs/index.php/AAAI/AAAI16/paper/view/11957
[7] Tiancheng Zhao, Ran Zhao, and Maxine Eskénazi. 2017. Learning Discourse-level Semantic Comprehension for Neural Conversation Models. In Proceedings of the Twenty-Sixth Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Lisbon, Portugal, September 17-21, 2017, Lilis Márquez, Chris Callison-Burch, Juan Su, Daniele Pighin, and Yuval Marton (Eds.). The Association for Computational Linguistics, 1412–1421. https://doi.org/10.18653/v1/d17-1166
[8] Tom Young, Erik Cambria, Iti Chaturvedi, Hao Zhou, Subham Biswas, and Minlie Huang. 2019. Diversity for Neural Dialog Models using Conditional Variational Autoencoders. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA. 3295–3301. http://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14567
[9] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to Sequence Learning with Neural Networks. In NIPS 3104–3112.
[10] Zihang Tian, Wei Bi, Dongyue Lee, Lanqing Xie, Yiping Song, Xiaoying Liu, and Nevin L. Zhang. 2020. Response-Anticipated Memory for On-Demand Knowledge Integration in Response Generation. CoRR abs/2005.06128 (2020). arXiv:2005.06128 https://arxiv.org/abs/2005.06128
[11] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In NIPS. 5998–6008.
[12] Yangjun Zhang, Pengjie Ren, and Maarten de Rijke. 2019. Improving Background Based Conversation with Context-aware Knowledge Pre-selection. CoRR abs/1906.06685 (2019).
[13] Tiancheng Zhao, Ran Zhao, and Maxine Eskenazi. 2017. Learning Discourse-level Diversity for Neural Dialog Models using Conditional Variational Autoencoders. In ACL (1). Association for Computational Linguistics, 654–664.
[14] Hao-Tong Ye, Kai-Ling Lo, Shang-Yu Su, and Yun-Nung Chen. 2019. Knowledge-Grounded Response Generation with Deep Attentional Latent-Variable Model. CoRR abs/1903.09813 (2019).
[15] Tom Young, Iryk Cambria, Iti Chaturvedi, Hao Zhou, Subham Biswas, and Minlie Huang. 2018. Augmenting End-to-End Dialogue Systems With Commonsense Knowledge. In AAAI. AAAI Press, 4970–4977.
[16] Dongyan Zhao, Can Xu, Dongyan Zhao, and Rui You. 2019. A Knowledge-Grounded Neural Variable Encoder-Decoder Model for Generating Dialogues. In Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4–9, 2017, San Francisco, California, USA. 3295–3301. http://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14567
[17] Tiancheng Zhao, Ran Zhao, and Maxine Eskenazi. 2017. Learning Discourse-level Diversity for Neural Dialog Models using Conditional Variational Autoencoders. In ACL (1). Association for Computational Linguistics, 654–664.