A Compare Aggregate Transformer for Understanding Document-grounded Dialogue

Longxuan Ma, Weinan Zhang, Runxin Sun, Ting Liu
Research Center for Social Computing and Information Retrieval
Harbin Institute of Technology, Harbin, Heilongjiang, China
{lxma,wnzhang,rxsun,tliu}@ir.hit.edu.cn

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 CMU_DoG dataset show that the proposed CAT model outperforms the state-of-the-art approach and strong baselines.

1 Introduction

Dialogue system (DS) attracts great attention from industry and academia because of its wide application prospects. Sequence-to-sequence models (Seq2Seq) (Sutskever et al., 2014; Serban et al., 2016) 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 provids deficient information Li et al. (2016); Ghazvininejad et al. (2018). 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. (2016) introduced Maximum Mutual Information (MMI) as the objective function for generating diverse response. Serban et al. (2017) proposed a latent variable model to capture posterior information of golden response. Zhao et al. (2017) used conditional variational autoencoders to learn discourse-level diversity for neural dialogue models. The other way is introducing external knowledge, either unstructured knowledge texts Ghazvininejad et al. (2018); Ye et al. (2019); Dinan et al. (2019) or structured knowledge triples (Liu et al., 2018; Young et al., 2018; Zhou et al., 2018a) to help open-domain conversation generation by producing responses conditioned on selected knowledge.

Document: Movie Name: The Shape of Water. Year: 2017. Director: Guillermo del Toro. Genre: Fantasy, Drama.Cast: Sally Hawkins as Elisa Esposito, a mute cleaner who works at a secret government laboratory. ... Critical Response: one of del Toro's most stunningly successful works ...
Dialogue: **S1:** I thought The Shape of Water was one of Del Toro's best works. What about you? **S2:** *Yes, his style really extended the story.* **S1:** I agree. He has a way with fantasy elements that really helped this story be truly beautiful. It has a very high rating on rotten tomatoes, too. **S2: Sally Hawkins acting was phenomenally expressive. Didn't feel her character was mentally handicapped.** **S1:** The characterization of her as such was definitely off the mark.

Figure 1: One DGD example in the CMUDoG dataset. S1/S2 means Speaker-1/Speaker-2, respectively.

The Document-grounded Dialogue (DGD) (Zhou et al., 2018b; Zhao et al., 2019; Li et al., 2019) is a new way to use external knowledge. It establishes a conversation mode in which relevant information can be obtained from the given document. One example of DGD is presented in Figure 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 gener-

Findings of the Association for Computational Linguistics: EMNLP 2020, pages 1358–1367
November 16 - 20, 2020.

ate an informative response. Previous work Arora et al. (2019); Zhao et al. (2019); Qin et al. (2019); Tian et al. (2020); Ren et al. (2019) 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 Figure 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 (Vaswani et al., 2017) model for understanding the dialogues and generate informative responses in the DGD, named Compare Aggregate Transformer (CAT). Previous research (Sankar et al., 2019) has shown that the last utterance is the most important guidance for the response generation in the multi-turn setting. Hence we divide the dialogue into the last utterance and the dialogue history, then measure the effectiveness of the dialogue history. If the last utterance and the dialogue history are related, 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:

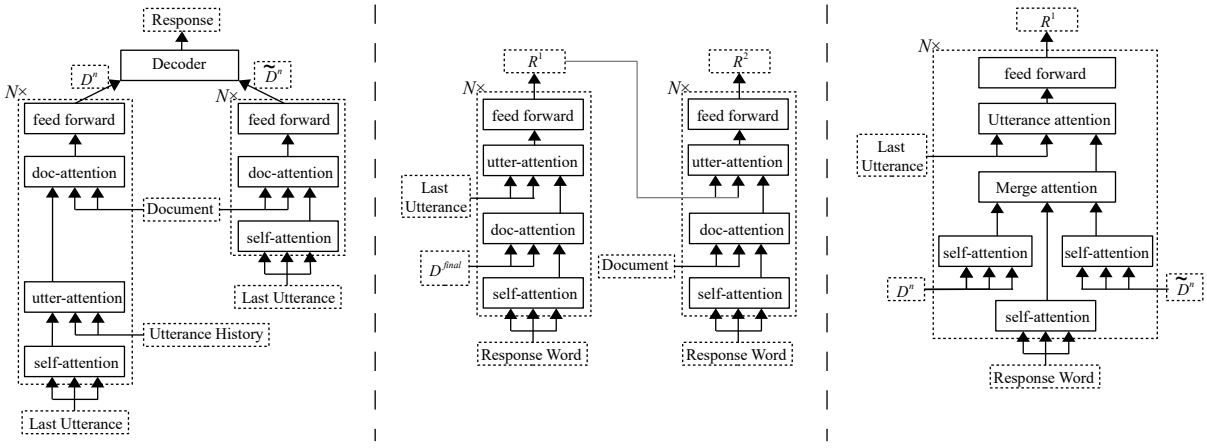
(1) We propose a compare aggregate method to determine the relationship between the historical dialogues and the last utterance. Experiments showed that our model outperformed strong baselines on the CMU_DoG dataset.

(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 Zhou et al. (2018b); Moghe et al. (2018); Qin et al. (2019); Gopalakrishnan et al. (2019) 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. (2018) 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. (2019) 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, (Zhang et al., 2019) 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 (Seo et al., 2017) to point out the most relevant token positions of the background sequence. Meng et al. (2019) used a decoding switcher to predict the probabilities of executing the reference decoding or generation decoding. Some other researchers (Zhao et al., 2019; Arora et al., 2019; Qin et al., 2019; Meng et al., 2019; Ren et al., 2019) also followed this parallel encoding method. For the second category, Kim et al. (2020) proposed a sequential latent knowledge selection model for Knowledge-Grounded Dialogue. Li et al. (2019) designed an incremental transformer to encode multi-turn utterances along with knowledge in the related document. Meanwhile, a two-way deliberation decoder (Xia et al., 2017) 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



(a) CAT Model (b) Deliberation Decoder (c) Transformer Enhanced Decoder

Figure 2: The architecture of the CAT model. "utter" is short for utterance. "doc" is short for document.

initiate a new topic (Akasaki and Kaji, 2019). 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:

$$P(\mathbf{R}|\mathbf{H},\mathbf{L},\mathbf{D};\Theta) = \prod_{i=1}^{r} P(R_i|\mathbf{H},\mathbf{L},\mathbf{D},\mathbf{R}_{<i};\Theta), \quad (1)$$

where $\mathbf{R}_{<i}$ = (R_1, R_2, ..., R_{i-1}), Θ is the model's parameters.

3.2 Encoder

The structure of the CAT model is shown in Figure 2. The hidden dimension of the CAT model is $\widehat{h}$. We use the Transformer structure (Vaswani et al., 2017). The self-attention is calculated as follow:

$$\text{Attention}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \text{softmax}(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}})\mathbf{V}, \quad (2)$$

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):

$$\text{MAtt}(\mathbf{Q},\mathbf{K},\mathbf{V}) = [\mathbf{A}_1, ..., \mathbf{A}_n]\mathbf{W}^O, \quad (3)$$

$$\mathbf{A}_i = \text{Attention}(\mathbf{Q}\mathbf{W}_i^Q, \mathbf{K}\mathbf{W}_i^K, \mathbf{V}\mathbf{W}_i^V), \quad (4)$$

where $\mathbf{W}_i^Q, \mathbf{W}_i^K, \mathbf{W}_i^V$ $(i = 1, ..., n)$ and $\mathbf{W}^O$ are learnable parameters.

The encoder of CAT consists of two branches as figure 2 (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 $\mathbf{H}_s = \text{MAtt}(\mathbf{H},\mathbf{H},\mathbf{H})$ and $\mathbf{L}_s = \text{MAtt}(\mathbf{L},\mathbf{L},\mathbf{L})$. Then we employ $\mathbf{L}_s$ to guide the **H**. $\mathbf{H}^1 = \text{MAtt}(\mathbf{L}_s,\mathbf{H},\mathbf{H})$, where $\mathbf{H}^1$ is the hidden state at the first layer. Then we adopt $\mathbf{H}^1$ to select knowledge from the document **D**, $\mathbf{D}^1 = \text{FF}(\text{MAtt}(\mathbf{H}^1,\mathbf{D},\mathbf{D}))$. FF is the feed-forward process. In the second layer, $\mathbf{D}^1$ is the input, $\mathbf{D}_s^1 = \text{MAtt}(\mathbf{D}^1,\mathbf{D}^1,\mathbf{D}^1))$, $\mathbf{H}^2 = \text{MAtt}(\mathbf{D}_s^1,\mathbf{H},\mathbf{H})$, $\mathbf{D}^2 = \text{FF}(\text{MAtt}(\mathbf{H}^2,\mathbf{D},\mathbf{D}))$. After N layers, we obtain the information $\mathbf{D}^n$ selected by **H**. In the right branch, we use $\mathbf{L}_s$ to filter the **D**. $\widetilde{\mathbf{D}^n}$ is the information selected by **L**.

3.3 Comparison Aggregate

As demonstrated by (Sankar et al., 2019), the last utterance played an fundamental role in response generation. We need to preserve the document information filtered by **L**, and determine how much information selected by **H** is needed. We propose 2 different compare aggregate methods: one is concatenation before decoding and the other is attended comparison in the decoder.

3.3.1 Concatenation

We use average pooling to $\mathbf{H}_s$ and $\mathbf{L}_s$ to get their vector representations $\mathbf{H}_{sa}$ and $\mathbf{L}_{sa} \in \mathbb{R}^{\widehat{h}*1}$, respectively. The concatenation method calculates relevance score α to determine the importance of $\mathbf{D}^n$ as follow:

$$\alpha = \tanh(\mathbf{H}_{sa}\mathbf{W}^H + \mathbf{L}_{sa}\mathbf{W}^L), \quad (5)$$

$$\mathbf{D}_{final} = [\text{sigmoid}(\mathbf{W}^\alpha \alpha) * \mathbf{D}^n; \widetilde{\mathbf{D}}^n], \quad (6)$$

where $\mathbf{W}^H$, $\mathbf{W}^L \in \mathbb{R}^{\widehat{h}*\widehat{h}}$, $\mathbf{W}^\alpha \in \mathbb{R}^{1*\widehat{h}}$ are learnable parameters. $[\mathbf{X}; \mathbf{Y}]$ is the concatenation of $\mathbf{X}$ and $\mathbf{Y}$ in sentence dimension. $*$ is the element-wise multiplication. Note that the $\mathbf{D}^n$ is guided by $\mathbf{H}$, the concatenation method performs a second level comparison with $\mathbf{H}$ and $\mathbf{L}$ and then transfers the topic-aware $\mathbf{D}_{final}$ to the two-pass Deliberation Decoder (DD) (Xia et al., 2017). The structure of the DD is shown in Figure 2 (b). The first-pass takes $\mathbf{L}$ and $\mathbf{D}_{final}$ as inputs and learns to generate a contextual coherently response $\mathbf{R}^1$. The second-pass takes $\mathbf{R}^1$ and the document $\mathbf{D}$ as inputs and learns to inject document knowledge. The DD aggregates document, conversation, and topic information to generate the final response $\mathbf{R}^2$. Loss is from both the first and the second layers:

$$L = -\sum_{m=1}^{M}\sum_{i=1}^{r}(\log P(R_i^1) + \log P(R_i^2)), \quad (7)$$

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

3.3.2 Attended Comparison

We employ an Enhanced Decoder (Zheng and Zhou, 2019) to perform the attended comparing. The structure of our Enhanced Decoder is illustrated in Figure 2 (c). It accepts $\mathbf{D}^n$, $\widetilde{\mathbf{D}}^n$ and the response $\mathbf{R}$ as inputs, applying a different way to compare and aggregate. The merge attention computes weight across all inputs:

$$\mathbf{P} = [\mathbf{R}; \mathbf{D}^n; \widetilde{\mathbf{D}}^n]\mathbf{W}_P, \quad (8)$$

$$\mathbf{V}_{merge} = P_R\mathbf{R} + P_D\mathbf{D}^n + P_{\widetilde{D}}\widetilde{\mathbf{D}}^n, \quad (9)$$

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 $\mathbf{P}$. $\mathbf{V}_{merge}$ and $\mathbf{L}$ are used for next utterance attention as shown in Figure 2 (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 CMU_DoG (Zhou et al., 2018b) 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 1. Please refer to Zhou et al. (2018b) for more details. It is worth noting that the sampled version does not represent the proportion of all conversation topic transfers, but it demonstrates this problem better than the Reduced version. We also test our method on the Holl-E Moghe et al. (2018) dataset. Since the processing of the dataset and the experimental conclusions obtained are similar to CMU_DoG, we did not present in this article.

4.2 Baselines

We evaluated several competitive baselines.

Dataset	U.Num(train / dev / test)	W/Utter
Original	72922 / 3626 / 11577	18.6
Reduced	66332 / 3269 / 10502	19.7
Sampled	66332 / 3269 / 1317	19.6

Table 1: Statistics of the CMU_DoG dataset. "U.Num" means Utterances Numbers, "W/Utter" means average words per utterance.

4.2.1 RNN-based models

VHRED: A Hierarchical Latent Variable Encoder-Decoder Model (Serban et al., 2017), which introduces a global (semantic level) latent variable Z for the problem that HRED (Serban et al., 2016) 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. (2017) 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 (Luong et al., 2015) 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 (Li et al., 2019): 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's encoding process. ITDD incrementally models dialogues without detecting topic transitions.

4.3 Evaluation Metrics

Automatic Evaluation: We employ perplexity (PPL) (Bengio et al., 2000), BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004). 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.

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 $\mathbf{G}_d^N$, $\mathbf{G}_c^N$ and $\mathbf{G}_r^N$, respectively. The number of overlapped N-grams of $\mathbf{G}_d^N$ and $\mathbf{G}_r^N$ is recorded as $\mathbf{G}_{dr}^N$. Tuples which are in $\mathbf{G}_{dr}^N$ but not in $\mathbf{G}_c^N$ is named $\mathbf{G}_{dr-c}^N$. Then $\mathbf{KU} = len(\mathbf{G}_{dr-c}^N)/len(\mathbf{G}_{dr}^N)$ reflects how many N-grams in the document are used in the generated replies, $len(\mathbf{G})$ is the tuple number in $\mathbf{G}$. 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 $\mathbf{G}$ as $\mathbf{R}_k^G$, which represents the importance of a tuple. Then the **Quality of Knowledge Utilization (QKU)** is calculated as:

$$\mathbf{QKU} = \sum_{(\mathbf{D},\mathbf{C},\mathbf{R})} \frac{\sum_k \mathbf{R}_k^{G_r}}{\sum_k \mathbf{R}_k^{G_d}}, \quad k \in \mathbf{G}_{dr-c}. \quad (10)$$

If $\mathbf{R}_k^{G_r}$ is more important in response and $\mathbf{R}_k^{G_d}$ 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

Model	PPL	BLEU (%)	ROUGE-L	KU-2/3 (%)	QKU-2/3
VHRED(-k)	97.3◇ (99.3)*	0.49* (0.49)*	7.80* (7.82)*	–/– (–/–)	–/– (–/–)
VHRED(c)	80.2◇ (85.4)*	0.79* (0.77)*	8.64* (8.63)*	12.0/27.0◇ (12.1/27.6)◇	3.36/2.82◇ (3.35/2.80)◇
VHRED(a)	77.2◇ (78.5)*	0.84* (0.80)*	8.98* (8.99)*	**13.7/31.7**◇ (**13.1/31.3**)*	3.23/2.72* (3.23/2.72)*
T-DD	18.2* (20.5)*	0.90* (0.89)*	9.23* (9.24)*	8.0/23.1* (8.0/23.0)*	2.55/1.94* (2.55/1.95)*
T-EDD	18.2* (20.3)*	0.91* (0.90)*	9.35* (9.36)*	8.3/23.5* (8.1/23.4)*	2.45/1.91* (2.45/1.92)*
ITDD	16.2* (18.7)*	1.01* (0.99)*	10.12◇ (10.10)*	9.0/24.5* (9.1/24.4)*	2.18/1.84* (2.15/1.82)*
CAT-EDD	16.0* (18.2)*	1.14* (1.14)*	11.10* (11.12)*	9.5/24.8* (9.7/24.9)*	2.12/1.77* (2.11/1.76)*
CAT-DD	**15.2** (**16.1**)	**1.22** (**1.21**)	**11.22** (**11.22**)	11.0/26.5 (11.1/26.4)	**2.08/1.64** (**2.05/1.62**)

Table 2: Automatic evaluations on the CMU_DoG Dataset. · (·) means Reduced (Sampled) test data. We take the CAT-DD as the base model to do the significant test, ◇ and * stands p<0.05 and p<0.01, respectively.

5 (1:very bad, 2:bad, 3:acceptable, 4:good, 5:very good). Overall inter-rater agreement measured by Fliess' Kappa is 0.32 ("fair").

4.4 Experimental Setup

We use OpenNMT-py (Klein et al., 2017) as the code framework. For all models, the pre-trained 300 dimension word embedding (Mikolov et al., 2013) 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 (α = 0.001, β_1 = 0.9, β_2 = 0.999, and $\epsilon = 10^{-8}$) (Kingma and Ba, 2015) 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 2 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 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/KU/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-L/KU/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 we expected. The CAT-EDD 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 3 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

Model	Flu.	Coh.	Inf.
VHRED(-k)	3.71 (3.72)	2.82 (2.72)	3.01 (2.82)
VHRED(c)	3.73 (3.82)	3.04 (3.11)	3.03 (3.05)
VHRED(a)	3.84 (3.77)	3.11 (3.14)	3.22 (3.06)
T-DD	3.84 (3.82)	3.03 (3.06)	3.03 (3.06)
T-EDD	3.84 (3.83)	3.02 (3.08)	3.05 (3.05)
ITEDD	3.90 (3.91)	3.11 (3.12)	3.43 (3.42)
CAT-EDD	4.02 (3.93)	3.12 (3.33)	3.33 (3.41)
CAT-DD	**4.09** (**4.09**)	**3.39** (**3.43**)	**3.44** (**3.61**)

Table 3: Manual evaluations on the CMU_DoG Dataset. Flu. /Coh. /Inf. /· (·) mean Fluency /Coherence /Informativeness /Reduced (Sampled) test data, respectively.

Models	PPL	BLEU	KU-2(%)/QKU-2
CAT-DD	16.1	1.21	11.1 / 2.05
w/o-left	19.8*	0.90*	8.2* / 2.56*
w/o-(5,6)	18.7*	0.93*	9.1* / 2.48⋄
w/o-(G)	18.2*	0.96*	9.2⋄ / 2.46*

Table 4: Ablation Study on the Sampled test set. 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. w/o means without.

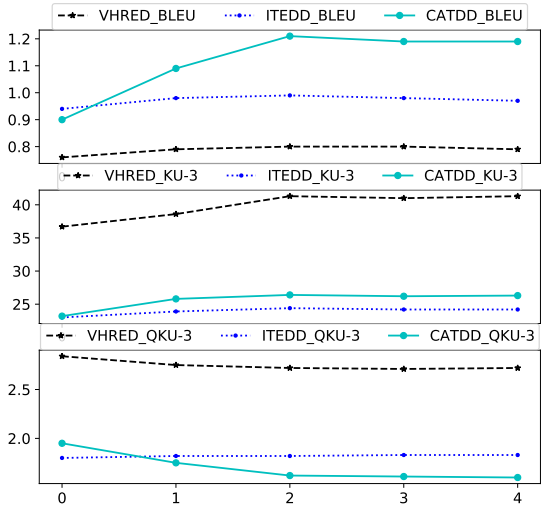


Figure 3: The effect of dialogue history rounds on VHRED(a)/ITDD/CAT-DD models. The abscissa represents the historical dialogue rounds. The ordinate represents the BLEU/KU-3/QKU-3 values.

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 4 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/KU-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 3 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 QKU-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.

5.4 History Importance Study

Figure 4 shows the average sigmoid$(W^{\alpha}\alpha)$ 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

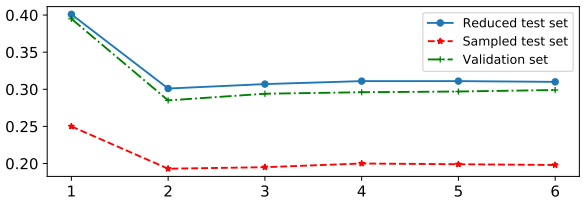


Figure 4: The rating of dialogue history in the CAT-DD model with Reduced and Sampled test set. The abscissa represents the dialogue rounds and the ordinate represents the correlation score in the model.

Document:
... sally hawkins as elisa esposito, a mute cleaner who works at a secret government laboratory. michael shannon as colonel richard strickland ... rating rotten tomatoes: 92% The shape of water is a 2017 american fantasy film ... it stars sally hawkins, michael shannon, richard jenkins, Doug jones, michael stuhlbarg, and octavia spencer ...

Dialogue history:
S1: I wonder if it's a government creation or something captured from the wild. i would assume the wild.
S2: It was captured for governmental experiments.
The last Utterance:
S1: Is it a big name cast?

Groud truth:
S2: Sally hawkins played the role of the mute cleaner, michael shannon played the role of colonel richard strickland.

Generated response:
VHRED(a): it has rating rotten tomatoes: 92%.
TDD: i am not sure about it.
ITDD: yes, sally hawkins as elisa esposito.
CAT-DD: sally hawkins, michael shannon, richard jenkins, doug jones, michael stuhlbarg, and octavia spencer.
(w/o-(5,6)): yes, sally hawkins works at a secret government laboratory.
(w/o-(G)): it is a 2017 american fantasy film.

Figure 5: Case study in the CMU_DoG Sampled Dataset. S1/S2 means Speaker-1/Speaker-2, respectively. (w/o-(5,6)) and (w/o-(G)) are models in the ablation study.

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 Figure 5, 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.

Acknowledgments

This paper is supported by the National Natural Science Foundation of China under Grant No. 62076081, No.61772153 and No.61936010.

References

Satoshi Akasaki and Nobuhiro Kaji. 2019. Conversation initiation by diverse news contents introduction. In *NAACL-HLT (1)*, pages 3988–3998. Association for Computational Linguistics.

Siddhartha Arora, Mitesh M. Khapra, and Harish G. Ramaswamy. 2019. On knowledge distillation from complex networks for response prediction. In *NAACL-HLT (1)*, pages 3813–3822. Association for Computational Linguistics.

Yoshua Bengio, Réjean Ducharme, and Pascal Vincent. 2000. A neural probabilistic language model. In *NIPS*, pages 932–938. MIT Press.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of wikipedia: Knowledge-powered conversational agents. In *ICLR (Poster)*. OpenReview.net.

Marjan Ghazvininejad, Chris Brockett, Ming-Wei Chang, Bill Dolan, Jianfeng Gao, Wen-tau Yih, and Michel Galley. 2018. A knowledge-grounded neural conversation model. In *AAAI*, pages 5110–5117. AAAI Press.

Karthik Gopalakrishnan, Behnam Hedayatnia, Qinlang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, Dilek Hakkani-Tür, and Amazon Alexa AI. 2019. Topical-chat: Towards knowledge-grounded open-domain conversations. *Proc. Interspeech 2019*, pages 1891–1895.

Byeongchang Kim, Jaewoo Ahn, and Gunhee Kim. 2020. Sequential latent knowledge selection for knowledge-grounded dialogue. *CoRR*, abs/2002.07510.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada. Association for Computational Linguistics.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *HLT-NAACL*, pages 110–119. The Association for Computational Linguistics.

Zekang Li, Cheng Niu, Fandong Meng, Yang Feng, Qian Li, and Jie Zhou. 2019. Incremental transformer with deliberation decoder for document grounded conversations. In *ACL (1)*, pages 12–21. Association for Computational Linguistics.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Shuman Liu, Hongshen Chen, Zhaochun Ren, Yang Feng, Qun Liu, and Dawei Yin. 2018. Knowledge diffusion for neural dialogue generation. In *ACL (1)*, pages 1489–1498. Association for Computational Linguistics.

Zhibin Liu, Zheng-Yu Niu, Hua Wu, and Haifeng Wang. 2019. Knowledge aware conversation generation with reasoning on augmented graph. *CoRR*, abs/1903.10245.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 1412–1421. The Association for Computational Linguistics.

Chuan Meng, Pengjie Ren, Zhumin Chen, Christof Monz, Jun Ma, and Maarten de Rijke. 2019. Refnet: A reference-aware network for background based conversation. *CoRR*, abs/1908.06449.

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

Nikita Moghe, Siddhartha Arora, Suman Banerjee, and Mitesh M. Khapra. 2018. Towards exploiting background knowledge for building conversation systems. In *EMNLP*, pages 2322–2332. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*, pages 311–318. ACL.

Lianhui Qin, Michel Galley, Chris Brockett, Xiaodong Liu, Xiang Gao, Bill Dolan, Yejin Choi, and Jianfeng Gao. 2019. Conversing by reading: Contentful neural conversation with on-demand machine reading. In *ACL (1)*, pages 5427–5436. Association for Computational Linguistics.

Pengjie Ren, Zhumin Chen, Christof Monz, Jun Ma, and Maarten de Rijke. 2019. Thinking globally, acting locally: Distantly supervised global-to-local knowledge selection for background based conversation. *CoRR*, abs/1908.09528.

Chinnadhurai Sankar, Sandeep Subramanian, Chris Pal, Sarath Chandar, and Yoshua Bengio. 2019. Do neural dialog systems use the conversation history effectively? an empirical study. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 32–37.

Min Joon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2017. Bidirectional attention flow for machine comprehension. In *ICLR (Poster)*. OpenReview.net.

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 Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA*, pages 3776–3784.

Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron C. Courville, and Yoshua Bengio. 2017. A hierarchical latent 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*, pages 3295–3301.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *NIPS*, pages 3104–3112.

Zhiliang Tian, Wei Bi, Dongkyu Lee, Lanqing Xue, Yiping Song, Xiaojiang Liu, and Nevin L. Zhang. 2020. Response-anticipated memory for on-demand knowledge integration in response generation. *CoRR*, abs/2005.06128.

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*, pages 5998–6008.

Yingce Xia, Fei Tian, Lijun Wu, Jianxin Lin, Tao Qin, Nenghai Yu, and Tie-Yan Liu. 2017. Deliberation networks: Sequence generation beyond one-pass decoding. In *NIPS*, pages 1784–1794.

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.

Tom Young, Erik Cambria, Iti Chaturvedi, Hao Zhou, Subham Biswas, and Minlie Huang. 2018. Augmenting end-to-end dialogue systems with commonsense knowledge. In *AAAI*, pages 4970–4977. AAAI Press.

Yangjun Zhang, Pengjie Ren, and Maarten de Rijke. 2019. Improving background based conversation with context-aware knowledge pre-selection. *CoRR*, abs/1906.06685.

Tiancheng Zhao, Ran Zhao, and Maxine Eskénazi. 2017. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. In *ACL (1)*, pages 654–664. Association for Computational Linguistics.

Xueliang Zhao, Chongyang Tao, Wei Wu, Can Xu, Dongyan Zhao, and Rui Yan. 2019. A document-grounded matching network for response selection in retrieval-based chatbots. In *IJCAI*, pages 5443–5449. ijcai.org.

Wen Zheng and Ke Zhou. 2019. Enhancing conversational dialogue models with grounded knowledge. In *CIKM*, pages 709–718. ACM.

Hao Zhou, Tom Young, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. 2018a. Commonsense knowledge aware conversation generation with graph attention. In *IJCAI*, pages 4623–4629. ijcai.org.

Kangyan Zhou, Shrimai Prabhumoye, and Alan W. Black. 2018b. A dataset for document grounded conversations. In *EMNLP*, pages 708–713. Association for Computational Linguistics.