A Document Driven Dialogue Generation Model

Ke Li, Ziwei Bai, Xiaojie Wang, and Caixia Yuan

Beijing University of Posts and Telecommunications
{cocolike,bestbzw,xjwang,yuancx}@bupt.edu.cn

Abstract. Most of the current man-machine dialogues are at the two end-points of a spectrum of dialogues, i.e. goal-driven dialogues and non goal-driven chit-chats. Document-driven dialogues provide a bridge between them with the change of documents from structured data to unstructured free texts. This paper proposes a Document Driven Dialogue Generation model (D3G) which generates dialogues centering a given document, as well as answering user's questions. A Doc-Reader mechanism is designed to locate the content related to user's questions in documents. A Multi-Copy mechanism is employed to generate document-related responses. And the dialogue history is used in both mechanisms. Experimental results on the CMU_DOG dataset show that our D3G model can not only generate informative responses that are more relevant to the document, but also answer user's questions better than the baseline models.

Keywords: Document-driven dialogue · Doc-Reader · Multi-Copy.

1 Introduction

Most of the current man-machine dialogues are at the two end-points of a spectrum of dialogues. The goal-driven dialog [6,9,16] is on one end and chit-chat [13,14] is on the other end. Goal-driven dialogue systems communicate with users based on pre-defined task goals. While chit-chat aims to generate suitable responses without pre-defined task goals.

However, there are various dialogues between the two ends in daily life. For the goal-driven dialogue, it is often difficult to obtain the structured goal for a dialogue. For example, lots of services are illustrated by unstructured or semi-structured documents instead of structured frames, and operators are asked to talk with customers directly according to the information in documents. For the chit-chat, conversations are not completely unconstrained. A typical example is that a reasonable response might be relevant to the news when people chat centering on it. The dialogues centering on free texts are called document-driven.

Some work has been done to make a bridge between the goal-driven dialogue and chit-chat. For example, the goal-driven dialogue drew on the experience of the sequence to sequence framework which is widely used in chit-chat [5, 17]. While the chi-chat attempted to incorporate more structural information [8]. In recent years, people tried to add some external knowledge and information into dialogues, such as the knowledge base [1],the knowledge graph [18], and the image [15]. But there are few studies on the document-driven dialogue. In 2018, Zhou et al. [19] presented a document grounded dataset for conversations and proposed two baseline models on it. They demonstrated the model using the documents outperforms the model without documents. The study on document-driven dialogue bridges the gap between the goal-driven dialogue and chit-chat.

This paper aims to extend the document-driven dialogue, enabling it to answer specific questions as well as chatting with users based on the document. To achieve this goal, we propose a Document Driven Dialog Generation (D3G) model, which has two mechanisms: a Doc-Reader mechanism is designed to locate answers in the document regarding to user's question and a Multi-Copy mechanism is used to help generate document-related responses. Experimental results show that D3G model significantly outperforms baseline methods in several ways.

In general, our contributions are as follows:

- We propose a document-driven dialogue generation model(D3G). Our model can not only chat with users centering on a given document, but also answer user's questions related to the document.
- In the D3G model, we use the Doc-Reader mechanism to help locate the content related to user's questions. We also propose the Multi-Copy mechanism to generate document-related responses. And the dialogue history is used in both mechanisms.
- Experimental results on automatic evaluation show our model can generate much more document-related responses compared to the baseline models.

2 Related Work

There are few studies focused on document-driven dialog generation models. Most of the chat models aim to use the document to increase the diversity of responses, but pay less attention to the relevance between generated responses and documents they are referring, and let alone answer document-related questions.

Zhou et al. [19] is one of the few jobs that focus on the relevance of generated responses to documents. As far as we know, they present the first document grounded dataset for conversations, which called CMU_DOG, along with two baseline models. The conversations in the dataset are about the contents of a specified document. In order to verify the validity of this dataset, Zhou et al. proposes NW score and use BLEU [10] to measure the relevance between documents and the conversations generated by annotators. The two baseline models have similar structures, except for the input of decoder, where SEQS utilizes both current utterance and the document while SEQ only concerns the current utterance. Moreover, Zhou et al. uses human evaluation "Engagement" to measure the response generated by SEQ and SEQS, causing subjective errors.

Unlike the work of Zhou [19], our proposed D3G model employs the Doc-Reader mechanism and the Multi-Copy mechanism to make full use of the document information. The dialogue history is also used in both mechanisms. Our model can not only generate document-related responses, but also answer user's questions relevant to documents. In addition, we evaluate generated responses automatically.

There are some studies that introduce documents to the chit-chat. Those studies focus on solving safe response problems, such as "emmm..." or "I don't know". In 2017, Ghazvininejad et al. proposes the MTASK-R model to introduce external text

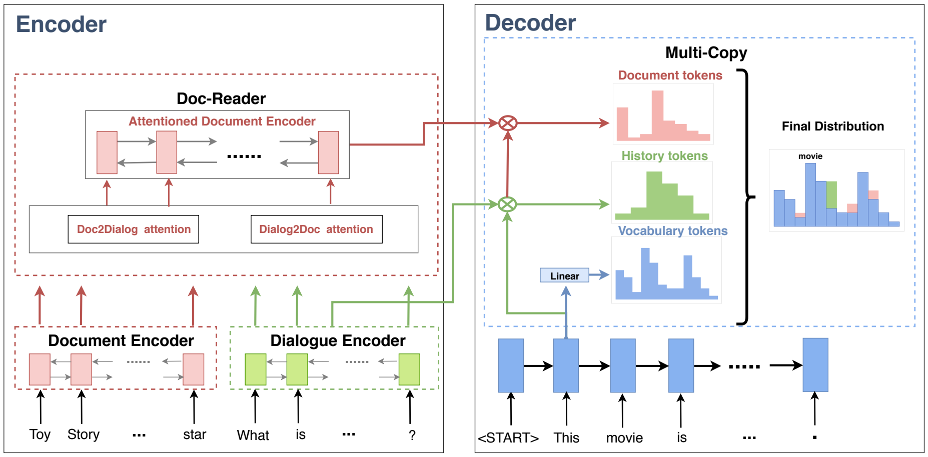

Fig. 1: The Document Driven Dialogue Generation Model

information into the chit-chat [2]. Experiments show that the model with external text information can generate more diverse responses. However, such a system can only chat with the user, without the ability to answer user's questions. The evaluation of these tasks also focuses on the fluency and diversity of the responses but pays less attention to the relevance to the document.

Other work such as conversational machine reading comprehension, focuses on whether the machine accurately locates the answers in the document. In 2018, Zhu et al. proposes an attention-based conversational deep neural network SDNet model [20], which helps the system determine the answer by understanding the document and dialogue history. Such research only requires the model to answer user's questions, but cannot chat with the user.

3 Model

Let $\{u_1, u_2, \ldots, u_t\}$ be the dialogue up to t-th utterance and $E^s = \{w_1, w_2, \ldots, w_m\}$ denotes a sequence of m tokens in the document. The response produced by D3G model is defined as $Y = \{y_1, y_2, \ldots, y_T\}$, where T is the number of tokens in the response. As shown in Fig.1, D3G model mainly consists of two modules: Encoder and Decoder. The Encoder Module first encodes the document and the dialogue, then uses a Doc-Reader mechanism to obtain dialogue-aware representations of document tokens. The Decoder Module proposes a Multi-Copy mechanism to generate document-related responses.

3.1 Encoder Module

Document Encoder: Given the document $E^s = \{w_1, w_2, \ldots, w_m\}$, a bidirectional LSTM is used to get the document contextual representation S,

$$S = BiLSTM(E^s), \tag{1}$$

where $S = [s_1, s_2, \dots, s_m] \in R^{2d \times m}$, d is the hidden size.

Dialogue Encoder: Given the dialogue history $\{u_1, u_2, \dots, u_{t-1}\}$ and the current user's utterance u_t, we concatenate the utterances to get $E^h = \{w_1, w_2, \dots, w_k\}$, where k is the number of tokens in dialogue. Then we use the same bidirectional LSTM to get the dialogue contextual representation H,

$$H = BiLSTM(E^h), \tag{2}$$

where $H = [h_1, h_2, \dots, h_k] \in R^{2d \times k}$.

Doc-Reader mechanism: The Doc-Reader mechanism is the core of the Encoder Module. It enables the D3G model with the ability to locate contents which are related to the dialogue. Inspired by BiDAF [12], a classical model for machine reading comprehension, we introduce the Doc-Reader mechanism to indicate the "importance" of tokens in the document. Moreover, we consider the attention from two directions: document to dialogue attention (Doc2Dialog Attention) and dialogue to document attention (Dialog2Doc Attention) to obtain the dialogue-aware representations of document tokens.

We build M, a matrix donates the token-level similarity between the dialogue representation H and document representation S,

$$M = S^{\mathrm{T}} H, \tag{3}$$

where $M \in R^{m \times k}$, and M_{ij} indicates the similarity between i-th document token and j-th dialogue token.

Doc2Dialog Attention: We use the Doc2Dialog Attention to signify which dialogue tokens are more relevant to each document token. First of all, we use a column softmax to get the attention weights α on dialogue tokens,

$$\alpha = softmax_{col}(M), \tag{4}$$

where $\alpha = [\alpha_1, \alpha_2, \dots, \alpha_m]^{\mathrm{T}} \in R^{m \times k}$, $\alpha_i \in R^k$ indicates the attention weights on dialogue tokens for the i-th document token.

Then, we obtain the attended dialogue context $\widetilde{H}$ for the entire document,

$$\widetilde{H} = H\alpha^{\mathrm{T}}, \tag{5}$$

where $\widetilde{H} = [\widetilde{h}_1, \widetilde{h}_2, \dots, \widetilde{h}_m] \in R^{2d \times m}$, $\widetilde{h}_i \in R^{2d}$ indicates the attended dialogue context for the i-th document token.

Dialog2Doc Attention: Dialog2Doc Attention signifies which document tokens have more closer similarity to the dialogue context.

First of all, we obtain the attention weights on the document tokens $\beta \in R^m$,

$$\beta = softmax(max_{col}(M)), \tag{6}$$

Then we obtain the attended document vectors $\widetilde{s} \in R^{2d}$.

$$\widetilde{s} = S\beta. \tag{7}$$

To maintain the contextual relevance, we fuse the results of Doc2Dialog Attention and Dialog2Doc Attention, then feed it into a bidirectional LSTM layer. Finally we obtain the dialogue-aware representations of document tokens $D \in R^{2d \times m}$,

$$G = [S \circ \widetilde{H}; S \circ (\widetilde{s} \otimes e_m)] \in R^{4d \times m}, \tag{8}$$

$$D = BiLSTM(G), \tag{9}$$

where the outer product $(\cdot \otimes e_m)$ produces a matrix or row vector by repeating the vector or scalar on the left for m times.

3.2 Decoder Module

In this section, we give a detailed introduction to the proposed Multi-Copy mechanism. Our model is primarily motivated by the Pointer Generator [11], which aims to handle the OOV problems by copying tokens from the dynamic dialogue context and generating tokens from the external vocabulary at the same time. In our task, the response tokens may come from the external vocabulary, the dialogue, and the document. To generate a document related response, we propose the Multi-Copy mechanism which allows generating tokens from vocabulary and copying tokens from both dialogue and document.

Given the dialogue-aware document representation D and dialogue representation H, we use a single layer LSTM as decoder, which receives the word embedding of the previous token, the document attention representation, dialogue attention representation, and the decoder hidden state.

Generate from Vocab: At each time step t, we use a two-layer fully connected network and a softmax function to map the decoder hidden state $d_t \in R^d$ into the probability distribution of each token in the external vocabulary. The probability of token w generated through vocabulary is $p^v(w)$:

$$p^v(w) = \begin{cases} softmax(W_2^v(W_1^v d_t + b_1^v) + b_2^v) \cdot e^w & \text{if } w \in vocabulary, \\ 0 & \text{otherwise.} \end{cases} \tag{10}$$

where $e^w \in R^v$ is a one-hot vector used to distinguish each token in the vocabulary, v is the external vocabulary size. $W_1^v \in R^{d \times d}$, $b_1^v \in R^d$, $W_2^v \in R^{v \times d}$, and $b_2^v \in R^v$ are parameters to be learned.

Copy from Dialogue: We first obtain the attention weights $\gamma^h \in R^k$ over the contextual dialogue tokens and the dialogue context vector $C^h \in R^{2d}$.

$$F^h = tanh(W^f H + W^b d_t \otimes e_k), \tag{11}$$

$$\gamma^h = softmax({F^h}^{\mathrm{T}} W^{\gamma} + b^f \otimes e_k), \tag{12}$$

$$C^h = H\gamma^h. \tag{13}$$

where $W^{\gamma} \in R^{2d}$, $b^f \in R$, $W^f \in R^{2d\times 2d}$, and $W^b \in R^{2d\times d}$ are parameters to be learned. $(\cdot \otimes e_k)$ follows the same definition as before.

γ^h_j also respresents the probability of the j-th dialogue token. In many cases, a token may appear more than one times in dialogue context. $p^h(w)$ is the sum of all probabilities of the token "w".

$$p^h(w) = \begin{cases} \sum_{j:w_j=w} \gamma^h_j & \text{if } w \in dialogue, \\ 0 & \text{otherwise.} \end{cases} \tag{14}$$

Copy from Document: Similar to Copy from Dialogue context, we obtain the attention weights $\gamma^s \in R^m$, the token distribution in the document p^s and the attention weighted document $C^s \in R^{2d}$.

We feed the decoder state d_t, the attention weighted dialogue context C^h, and the attention weighted document C^s into a nonlinear neural network to obtain $\delta \in R^3$, which indicates the probabilities of choosing tokens from the external vocabulary, dialogue context or document.

$$\delta = softmax(W^d d_t + W^s C^s + W^h C^h + b). \tag{15}$$

where $W^d \in R^{3\times d}$. $W^s, W^h \in R^{3\times 2d}$, and $b \in R^3$.

Then we obtain the final token distribution p,

$$p(w) = \delta_1 p^v(w) + \delta_2 p^h(w) + \delta_3 p^s(w), \tag{16}$$

where $w \in V \cup S \cup H$. $\delta_1 p^v(w)$, $\delta_2 p^h(w)$, $\delta_3 p^s(w)$ represent probabilities of the token w from vocabulary, dialogue and document respectively. We have a detailed description in section 4.4.

3.3 Training

In the training stage, the loss for timestep t is the negative log likelihood of the target word ${w_t}^*$ for that timestep:

$$loss_t = -log\, p({w_t}^*), \tag{17}$$

and the overall loss for the whole sequence is:

$$loss = \frac{1}{T}\sum_{t=1}^{T} loss_t, \tag{18}$$

4 Experiment

4.1 Experimental settings

Dataset: We conduct experiments on the Document Grounded Conversations dataset (CMU_DOG). It is built by Carnegie Mellon University in 2018. The conversations in CMU_DOG are about the contents of a specified document(Wikipedia articles about popular movies). The dataset contains 4112 conversations with an average of 21.43 turns per conversation. According to the visibility of the documents by two interlocutors, CMU_DOG is divided into Scenario 1 and Scenario 2[1]. In Scenario 1, only one interlocutor has access to the document. The interlocutor who has no access to the document asks the other one to get information. In Scenario 2, both the interlocutors have access to the same Wiki document. They discuss the content in the document. In this paper, we experiment in Scenario1 and Scenario 2 respectively.

In order to evaluate our model's ability to answer question correctly, we automatically generate 98 questions about movies in Scenario 1. The answers of all questions can be found in the documents.

Baselines: We use the SEQ and SEQS proposed in the Zhou et al. [19] as our baseline models, which are only document-driven dialogue models as far as we know.

Implement Details: In all the models explored in this paper, we keep previous two utterances as dialogue history[2]. We use a two-layer bidirectional LSTM as an encoder. The dropout rate is set to be 0.3. The batch size is 32. The size of hidden units for both LSTMs is 300. The size of the vocabulary is 10000. We cut off the first 100 words in the documents during training. We use the pre-trained 100-dimensional glove embedding[3] and fine-tune it during the training. The models are trained with Adam optimizer [3] with learning rate 0.001 until they converge on the validation set for the valid loss. During the test, we use beam search with size 5.

Evaluation Metrics: Doc_BLEU[4] and NW [18] are used to measure the relevance between generated responses and documents. Target_BLEU [10] and METEOR [4] scores are employed to measure the similarity between generated responses and standard outputs. We also report the diversity [7] and the average length of the generated responses. We use the accurancy to evaluate the question answering.

4.2 Experimental Results

Table 1 shows the NW and Doc_BLEU scores in Scenario1 and Scenario2. As we can see, our D3G model significantly outperforms the baseline models on both datasets. It

[1] Scenario1 contains 2128 conversations and Scenario 2 contains 1984 conversations.

[2] After many experiments, the results obtained by using the previous two utterances as dialogue history are the best.

[3] https://nlp.stanford.edu/projects/glove/

[4] We only use the Bleu-1 and ignore the brevity penalty.

Table 1: NW and Doc_BLEU scores in Scenario 1 & 2

Models	**Scenario 1**		**Scenario 2**	
	NW	Doc_BLEU	NW	Doc_BLEU
SEQ	0.245	0.153	0.160	0.181
SEQS	0.467	0.218	0.186	0.237
D3G	**0.712**	**0.345**	**0.209**	**0.262**

Table 2: Targe_BLEU and METEOR scores in Scenario 1 & 2

Models	**Scenario 1**					**Scenario 2**				
	B-1	B-2	B-3	B-4	METEOR	B-1	B-2	B-3	B-4	METEOR
SEQ	0.086	0.061	0.054	0.052	0.016	0.031	0.021	0.019	0.019	0.016
SEQS	0.097	0.068	0.060	0.057	0.025	0.029	0.020	0.019	0.018	0.019
D3G	**0.142**	**0.084**	**0.069**	**0.065**	**0.044**	**0.059**	**0.042**	**0.037**	**0.036**	**0.024**

Table 3: Diversity scores and Average lengths in Scenario 1 & 2

Models	**Scenario 1**			**Scenario 2**		
	Dist-1	Dist-2	Avg_len	Dist-1	Dist-2	Avg_len
SEQ	0.017	0.037	3.85	0.008	0.024	2.83
SEQS	0.024	0.061	4.99	0.016	0.056	2.91
D3G	**0.071**	**0.312**	**7.27**	**0.046**	**0.205**	**4.96**

Table 4: Accuracy Rate

Models	Acc(%)
SEQ	8.16
SEQS	11.22
D3G	**16.32**

demonstrates that our D3G model can generate more document-related responses. The improvements of NW and Doc_BLUE in Scenario1 are more than 24.5% and 12.7% compared to SEQS. Compared to Scenario 1, the conversations in Scenario 2 are freer and much lower related to the document, so the improvements are relatively small in Scenario2.

Table 2 shows the Target_BLEU and the METEOR scores in Scenario1 and Scenario2. Compared to the baseline models, our D3G model slightly improves the BLEU score and METEOR score. In combination with Table 1, it shows that responses generated by our model are not only document-related, but also with high quality.

Table 3 shows the Dist and Avg_len scores in Scenario 1 and Scenario 2. It demonstrates that D3G model can generate longer and more diverse responses than that generated by SEQ and SEQS.

Table 4 shows the accuracy rates on our 98 automatically generated questions. It also validates the superiority of our D3G model. There is an improvement of 5% in the accuracy rate. Our D3G model can generate better responses relative to the baseline models.

Table 5: Ablation study in Scenario1

Models	NW	Doc_BLEU	Target_BLEU	METEOR	AVG_LEN	Dist-1	Dist-2	Acc
D3G	**0.712**	**0.345**	**0.065**	**0.044**	**7.27**	**0.071**	**0.312**	**16.32**
-Multi-Copy	0.550	0.312	0.059	0.034	6.89	0.061	0.231	13.27
-Doc-Reader	0.651	0.299	0.065	0.039	6.44	0.069	0.305	12.24
-History	0.476	0.293	0.059	0.034	5.65	0.064	0.253	13.27

Table 6: Ablation study in Scenario2

Models	NW	Doc_BLEU	Target_BLEU	METEOR	AVG_LEN	Dist-1	Dist-2
D3G	**0.209**	**0.262**	**0.036**	**0.024**	**4.96**	**0.046**	**0.205**
-Multi-Copy	0.189	0.255	0.031	0.021	4.49	0.043	0.196
-Doc-Reader	0.174	0.261	0.029	0.022	4.91	0.044	0.193
-History	0.181	0.234	0.024	0.021	4.55	0.037	0.148

In addition, our model performs better in Scenario 1 than in Scenario 2. One possible reason is that both the interlocutors in Scenario 2 can see the content of the document, which has high degree of freedom and relatively complex sentences.

4.3 Ablation study

To validate the effectiveness of our D3G model, we conduct three ablation experiments in Scenario1 and 2 separately since two datasets have different characteristics.

- The full D3G model with all of the components.
- D3G model without Multi-Copy mechanism.[5]
- D3G model without Doc-Reader mechanism.
- D3G model without dialog history.

The results are shown in Table 5 and Table 6 respectively. Table 6 removes the accuracy since the document is accessible for both interlocutors in Scenario 2, there is no need to ask questions.

We can draw that the methods proposed by D3G are indispensable for all evaluation metrics on both Scenarios. Firstly, Multi-Copy mechanism is very important to D3G model. All scores decrease when Multi-Copy is moved from D3G model. Especially, the NW score decreases significantly from 0.712 to 0.550 in Scenario 1. It shows that Multi-Copy not only has a great impact on generating document-related responses, but also has some influences on diversity and fluency of responses.

Secondly, the Doc-Reader mechanism leads to a significant improvement on accuracy as shown in Table 5. The accuracy of questions answering drops significantly from 16.32 to 12.24. This shows that the Doc-Reader mechanism can effectually help locate the content related to user's questions.

Finally, as shown in Table 5 and Table 6, removing the dialogue history causes drop on all scores. Because the dialogue history is used in both of the above mechanisms, removing history will have an impact on them.

[5] We follow SEQ, only copy tokens from dialogue.

Table 7: Examples of Question Answering

ID	Document	Inputs	Responses
1	**Name** Toy Story **Year** 1995 **Genre** Comedy, Animation,Adventure **Director** John Lasseter **Introduction** Toy Story is a 1995 American... comedy adventure film ... as they evolve from rivals ... Catmull.	What is the movie name?	**SEQ:** the wolf of wall street **SEQS:** the movie is the shape of water. **Ours:** toy story
2	**Name** The Wolf of Wall Street **Year** 2013 **Genre** black comedy **Director** Martin Scorsese **Introduction** The Wolf of Wall Street is a 2013 American black comedy[3] ... Leonardo DiCaprio (who was also a producer) stars as Belfort... also star.	What is the movie name?	**SEQ:** have you seen it? **SEQS:** the movie is based on a true story. **Ours:** the wolf of wall street
3	**Name** Zootopia **Year** 2016 **Genre** Comedy, Adventure, Animation **Director** Byron Howard, Rich Moore **Introduction** In a world populated ... her childhood dream of becoming ... Nick Wilde and Finnick.	awesome! I bet my kids would like it, is it old?	**SEQ:** nod. **SEQS:** i do n't know. **Ours:** it came out in 2016.

4.4 Case study

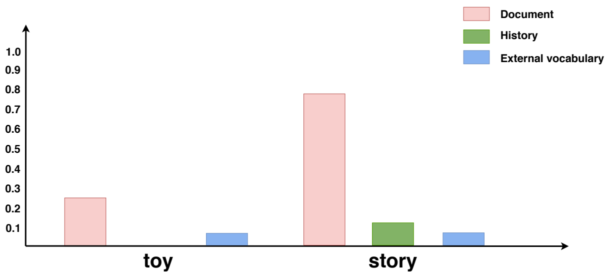

Fig. 2: Probabilities of the token in different sources

In this part, we show typical examples of generated responses of answering question (as shown in 7) and chatting with users (as shown in Table 8).

Table 7 shows the responses to user's questions. For example, in Item1, our proposed D3G model can generate correct movie name "Toy Story". The SEQ model can just generate the high frequency movie name in the train set. The SEQS model that uses the document also doesn't generate the correct answer.

We visually analyze the reasons why the D3G model accurately answers user's questions. As shown in the Fig.2, D3G allows copying tokens from multi-sources (dialogue, document, external vocabulary). The token "Toy" has a low probability in the external vocabulary, but highest in the document. Combining the above probabilities, the final output token is "Toy".

Table 8: Examples of chatting with users

ID	document	Inputs	Responses
1	**Name** The Avengers **Year** 2012 **Genre** Drama, Superhero **Director** Joss Whedon **Introduction** Marvel's The Avengers is a 2012 American ... Robert Downey Jr., Chris Evans, ... form a team that must stop Thor's brother Loki from subjugating Earth.	No way. Robert Downey Jr makes the character so funny, but so cool.	**SEQ:** good **SEQS:** it no on then not together . **Ours:** it do a very cool film really best.
2	**Name** Imitation Game **Year** 2014 **Genre** Drama, biopic **Director** Morten Tyldum **Introduction** The Imitation Game is a ... Benedict Cumberbatch decrypted German intelligence codes for the British government during the Second World War ... also star.	Actually, what I enjoyed most was Benedict Cumberbatch's performance. He was excellent!	**SEQ:** en **SEQS:** yeah, it is good . **Ours:** He made a great supporting role.

Table 8 shows the responses to chat with users. For example, in Item 2, the SEQ model can only generate short and general responses, such as "en". Our D3G model can generate "He made a great supporting role", which is meaningful, diverse and natural.

5 Conclusion

In this paper, we propose a Document Driven Dialogue Generation model, called D3G. Experimental results on CMU_DOG dataset show that D3G model outperforms state-of-art approaches. Ablation study shows that Doc-Reader mechanism effectively helps to locate the user's question related to document contents. Multi-Copy mechanism plays an important role to generate document-related responses. Finally, we visually analyze the reasons why our model works.

References

1. Cui, W., Xiao, Y., Wang, H., Song, Y., Wei, W.: Kbqa: Learning question answering over qa corpora and knowledge bases. Proceedings of the Vldb Endowment **10**(5), 565–576 (2017)
2. Ghazvininejad, M., Brockett, C., Chang, M.W., Dolan, B., Gao, J., Yih, W.t., Galley, M.: A knowledge-grounded neural conversation model. In: Thirty-Second AAAI Conference on Artificial Intelligence (2018)
3. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. Computer Science (2014)
4. Lavie, A., Agarwal, A.: Meteor: An automatic metric for mt evaluation with high levels of correlation with human judgments. In: Proceedings of the Second Workshop on Statistical Machine Translation. pp. 228–231. Association for Computational Linguistics (2007)
5. Lei, W., Jin, X., Kan, M.Y., Ren, Z., He, X., Yin, D.: Sequicity: Simplifying task-oriented dialogue systems with single sequence-to-sequence architectures. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 1437–1447 (2018)
6. Levin, E., Narayanan, S.S., Pieraccini, R., Zeljkovic, I.: Method of using a natural language interface to retrieve information from one or more data resources. At & T (1999)

7. Li, J., Galley, M., Brockett, C., Gao, J., Dolan, B.: A diversity-promoting objective function for neural conversation models. arXiv preprint arXiv:1510.03055 (2015)
8. Lian, R., Xie, M., Wang, F., Peng, J., Wu, H.: Learning to select knowledge for response generation in dialog systems (2019)
9. McTear, M.F.: Modelling spoken dialogues with state transition diagrams: experiences with the cslu toolkit. In: Fifth International Conference on Spoken Language Processing (1998)
10. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proc Meeting of the Association for Computational Linguistics (2002)
11. See, A., Liu, P.J., Manning, C.D.: Get to the point: Summarization with pointer-generator networks (2017)
12. Seo, M., Kembhavi, A., Farhadi, A., Hajishirzi, H.: Bidirectional attention flow for machine comprehension (2016)
13. Serban, I.V., Sordoni, A., Bengio, Y., Courville, A., Pineau, J.: Building end-to-end dialogue systems using generative hierarchical neural network models. In: Thirtieth Aaai Conference on Artificial Intelligence (2016)
14. Vinyals, O., Le, Q.: A neural conversational model. Computer Science (2015)
15. Yang, Z., He, X., Gao, J., Li, D., Smola, A.: Stacked attention networks for image question answering (2016)
16. Young, S., Gasic, M., Thomson, B., Williams, J.D.: Pomdp-based statistical spoken dialog systems: A review. Proceedings of the IEEE **101**(5), 1160–1179 (2013)
17. Zhao, T., Lu, A., Lee, K., Eskenazi, M.: Generative encoder-decoder models for task-oriented spoken dialog systems with chatting capability (2017)
18. Zhou, H., Young, T., Huang, M., Zhao, H., Xu, J., Zhu, X.: Commonsense knowledge aware conversation generation with graph attention. In: IJCAI. pp. 4623–4629 (2018)
19. Zhou, K., Prabhumoye, S., Black, A.W.: A dataset for document grounded conversations (2018)
20. Zhu, C., Zeng, M., Huang, X.: Sdnet: Contextualized attention-based deep network for conversational question answering (2018)