A Dataset for Document Grounded Conversations

Kangyan Zhou, Shrimai Prabhumoye, Alan W Black
Carnegie Mellon University, Pittsburgh, PA, USA
{kangyanz, sprabhum, awb}@cs.cmu.edu

Abstract

This paper introduces a document grounded dataset for conversations. We define "Document Grounded Conversations" as conversations that are about the contents of a specified document. In this dataset the specified documents were Wikipedia articles about popular movies. The dataset contains 4112 conversations with an average of 21.43 turns per conversation. This positions this dataset to not only provide a relevant chat history while generating responses but also provide a source of information that the models could use. We describe two neural architectures that provide benchmark performance on the task of generating the next response. We also evaluate our models for engagement and fluency, and find that the information from the document helps in generating more engaging and fluent responses.

1 Introduction

At present, dialog systems are considered to be either task-oriented, where a specific task is the goal of the conversation (e.g. getting bus information or weather for a particular location); or non-task oriented where conversations are more for the sake of themselves, be it entertainment or passing the time. Ultimately, we want our agents to smoothly interleave between task-related information flow and casual chat for the given situation. There is a dire need of a dataset which caters to both these objectives.

Serban et al. (2015) provide a comprehensive list of available datasets for building end-to-end conversational agents. Datasets based on movie scripts (Lison and Tiedemann, 2016; Danescu-Niculescu-Mizil and Lee, 2011a) contain artificial conversations. The Ubuntu Dialogue Corpus (Lowe et al., 2015) is based on technical support logs from the Ubuntu forum. The Frames dataset (Asri et al., 2017) was collected to solve the problem of frame tracking. These datasets do not provide grounding of the information presented in the conversations. Zhang et al. (2018) focuses on personas in dialogues: each worker has a set of predefined facts about the persona that they can talk about. Most of these datasets lack conversations with large number of on-topic turns.

We introduce a new dataset which addresses the concerns of grounding in conversation responses, context and coherence in responses. We present a dataset which has real human conversations with grounding in a document. Although our examples use Wikipedia articles about movies, we see the same techniques being valid for other external documents such as manuals, instruction booklets, and other informational documents. We build a generative model with and without the document information and find that the responses generated by the model with the document information is more engaging (+7.5% preference) and more fluent (+0.96 MOS). The perplexity also shows a 11.69 point improvement.

2 The Document Grounded Dataset

To create a dataset for document grounded conversations, we seek the following things: (1) A set of documents (2) Two humans chatting about the content of the document for more than 12 turns. We collected conversations about the documents through Amazon Mechanical Turk (AMT). We restrict the topic of the documents to be movie-related articles to facilitate the conversations. We initially experimented with different potential domains. Since movies are engaging and widely known, people actually stay on task when discussing them. In fact in order to make the task interesting, we offered a choice of movies to the participants so that they are invested in the task.

Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 708–713
Brussels, Belgium, October 31 - November 4, 2018.

2.1 Document Set Creation

We choose Wikipedia (Wiki) [1] articles to create a set of documents $\boldsymbol{D} = \{d_1, \ldots, d_{30}\}$ for grounding of conversations. We randomly select 30 movies, covering various genres like thriller, super-hero, animation, romantic, biopic etc. We extract the key information provided in the Wiki article and divide it into four separate sections. This was done to reduce the load of the users to read, absorb and discuss the information in the document. Hence, each movie document d_i consists of four sections $\{s_1, s_2, s_3, s_4\}$ corresponding to basic information and three key scenes of the movie. The basic information section s_1 contains data from the Wikipedia article in a standard form such as year, genre, director. It also includes a short introduction about the movie, ratings from major review websites, and some critical responses. Each of the key scene sections $\{s_2, s_3, s_4\}$ contains one short paragraph from the plot of the movie. Each paragraph contains on an average 7 sentences and 143 words. These paragraphs were extracted automatically from the original articles, and were then lightly edited by hand to make them of consistent size and detail. An example of the document is attached in Appendix.

2.2 Dataset Creation

To create a dataset of conversations which uses the information from the document, involves the participation of two workers. Hence, we explore two scenarios: (1) Only one worker has access to the document and the other worker does not and (2) Both the workers have access to the document. In both settings, they are given the common instructions of chatting for at least 12 turns.

Scenario 1: One worker has document. In this scenario, only one worker has access to the document. The other worker cannot see the document. The instruction to the worker with the document is: *Tell the other user what the movie is, and try to persuade the other user to watch/not to watch the movie using the information in the document*; and the instruction to the worker without the document is: *After you are told the name of the movie, pretend you are interested in watching the movie, and try to gather all the information you need to make a decision whether to watch the movie in the end*. An example of part of the dialogue for this scenario is shown in Table 1.

user2:	Hey have you seen the inception?
user1:	No, I have not but have heard of it. What is it about
user2:	It's about extractors that perform experiments using military technology on people to retrieve info about their targets.

Table 1: An example conversation for scenario 1. User 1 does not have access to the document, while User 2 does. The full dialogue is attached in the Appendix.

User 2:	I thought The Shape of Water was one of Del Toro's best works. What about you?
User 1:	Did you like the movie?
User 1:	Yes, his style really extended the story.
User 2:	I agree. He has a way with fantasy elements that really helped this story be truly beautiful.

Table 2: An example conversation for scenario 2. Both User 1 and User 2 have access to the Wiki document. The full dialogue is attached in the Appendix.

Scenario 2: Both workers have document. In this scenario, both the workers have access to the same Wiki document. The instruction given to the workers are: *Discuss the content in the document with the other user, and show whether you like/dislike the movie*. An example of the dialogue for this scenario is shown in Table 2.

Workflow: When the two workers enter the chat-room, they are initially given only the first section on basic information s_1 of the document d_i. After the two workers complete 3 turns (for the first section 6 turns is needed due to initial greetings), the users will be shown the next section. The users are encouraged to discuss information in the new section, but are not constrained to do so.

2.3 Dataset Statistics

The dataset consists of total 4112 conversations with an average of 21.43 turns. The number of conversations for scenario 1 is 2128 and for scenario 2 it is 1984. We consider a turn to be an exchange between two workers (say $w1$ and w_2). Hence an exchange of w_1, w_2, w_1 has 2 turns (w_1, w_2) and (w_2, w_1). We show the comparison of our dataset as **CMU Do**cument **G**rounded Conversations (CMU_DoG) with other datasets in Table 3.

[1] https://en.wikipedia.org

Dataset	# Utterances	Avg. # of Turns
CMU_DoG	130000	31
Persona-chat (Zhang et al., 2018)	164,356	14
Cornell Movie (Danescu-Niculescu-Mizil and Lee, 2011b)	304,713	1.38
Frames dataset (Asri et al., 2017)	19,986	15

Table 3: Comparison with other datasets. The average number of turns are calculated as the number of utterances divided by the number of conversations for each of the datasets.

	Rating 1	Rating 2	Rating 3	Rating 2& 3
Total # of conversations	1443	2142	527	2669
Total # of utterances	28536	80104	21360	101464
Average # utterances/conversation	19.77(13.68)	35.39(8.48)	40.53(12.92)	38.01(9.607)
Average length of utterance	7.51(50.19)	10.56(8.51)	16.57(15.23)	11.83(10.58)

Table 4: The statistics of the dataset. Standard deviation in parenthesis.

One of the salient features of CMUDoG dataset is that it has mapping of the conversation turns to each section of the document, which can then be used to model conversation responses. Another useful aspect is that we report the quality of the conversations in terms of how much the conversation adheres to the information in the document.

Split Criteria: We automatically measure the quality of the conversations using BLEU (Papineni et al., 2002) score. We use BLEU because we want to measure the overlap of the turns of the conversation with the sections of the document. Hence, a good quality conversation should use more information from the document than a low quality conversation. We divide our dataset into three ratings based on this measure. The BLEU score is calculated between all the utterances $\{x_1, \ldots, x_n\}$ of a conversation C_i and the document d_i corresponding to C_i. We eliminate incomplete conversations that have less than 10 turns. The percentiles for the remaining conversations are shown in Table 5. We split the dataset into three ratings based on BLEU score.

Percentile	20	40	60	80	99
BLEU	0.09	0.20	0.34	0.53	0.82

Table 5: The distribution of BLEU score for conversations with more than 10 turns.

Rating 1: Conversations are given a rating of 1 if their BLEU score is less than or equal to 0.1. We consider these conversations to be of low-quality.

Rating 2: All the conversations that do not fit in rating 1 and 3 are marked with a rating of 2.

Rating 3: Conversations are labeled with a rating of 3, only if the conversation has more than 12 turns and has a BLEU score larger than 0.587. This threshold was calculated by summing the mean (0.385) and the standard deviation (0.202) of BLEU scores of the conversations that do not belong rating 1.

The average BLEU score for workers who have access to the document is 0.22 whereas the average BLEU score for the workers without access to the document is 0.03. This suggests that even if the workers had external knowledge about the movie, they have not extensively used it in the conversation. It also suggests that the workers with the document have not used the information from the document verbatim in the conversation. Table 4 shows the statistics on the total number of conversations, utterances, and average number of utterances per conversation and average length of utterances for all the three ratings.

3 Models

In this section we discuss models which can leverage the information from the document for generating responses. We explore generative models for this purpose. Given a dataset $\boldsymbol{X} = \{x_0, \ldots, x_n\}$ of utterances in a conversation C_i, we consider two settings: (1) to generate a response x_{i+1} when given only the current utterance x_i and (2) to generate a response x_{i+1} when given the corresponding section s_i and the previous utterance x_i.

Without section: We use the sequence-to-sequence model (Sutskever et al., 2014) to build our baseline model. Formally, let $\boldsymbol{\theta}_E$ represent the

parameters of the encoder. Then the representation $\boldsymbol{h}_{x_i}$ of the current utterance x_i is given by:

$$\boldsymbol{h}_{x_i} = \text{Encoder}(\boldsymbol{x}_i; \boldsymbol{\theta}_E) \quad (1)$$

Samples of x_{i+1} are generated as follows:

$$p(\hat{\boldsymbol{x}}|\boldsymbol{h}_{x_i}) = \prod_t p(\hat{x}_t|\hat{\boldsymbol{x}}^{<t}, \boldsymbol{h}_{x_i}) \quad (2)$$

where, $\hat{\boldsymbol{x}}^{<t}$ are the tokens generated before $\hat{x}_t$. We also use global attention (Luong et al., 2015) with copy mechanism (See et al., 2017) to guide our generators to replace the unknown (UNK) tokens. We call this model **SEQ**.

With section: We extend the sequence-to-sequence framework to include the section s_i corresponding the current turn. We use the same encoder to encode both the utterance and the section. We get the representation $\boldsymbol{h}_{x_i}$ of the current utterance x_i using Eq. 1. The representation of the section is given by:

$$\boldsymbol{h}_{s_i} = \text{Encoder}(\boldsymbol{s}_i; \boldsymbol{\theta}_E) \quad (3)$$

The input at each time step t to the generative model is given by $h_t = [x_{t-1}; h_s]$, where x_{t-1} is the embedding of the word at the previous time step. We call this model **SEQS**.

Experimental Setup: For both SEQ and SEQS model, we use a two-layer bidirectional LSTM as the encoder and a LSTM as the decoder. The dropout rate of the LSTM output is set to be 0.3. The size of hidden units for both LSTMs is 300. We set the word embedding size to be 100, since the size of vocabulary is relatively small[2]. The models are trained with adam (Kingma and Ba, 2014) optimizer with learning rate 0.001 until they converge on the validation set for the perplexity criteria. We use beam search with size 5 for response generation. We use all the data (i.e all the conversations regardless of the rating and scenario) for training and testing. The proportion of train/validation/test split is 0.8/0.05/0.15.

4 Evaluation

In what follows, we first present an analysis of the dataset, then provide an automatic metric for evaluation of our models–perplexity and finally present the results of human evaluation of the generated responses for engagement and fluency.

scenario	NW	LT
1	0.78	12.85
2	5.84	117.12

Table 6: The results of data analysis. LT refers to the average length of x_i in scenario 1 and $x_i, \ldots, x_{i+k}$ in scenario 2.

Dataset analysis: We perform two kinds of automated evaluation to investigate the usefulness of the document in the conversation. The first one is to investigate if the workers use the information from the document in the conversation. The second analysis is to show that the document adds value to the conversation. Let the set of tokens in the current utterance x_i be N, the set of tokens in the current section s_i be M, the set of tokens in the previous three utterances be H, and the set of stop words be S. In scenario 1, we calculate the set operation (NW) as $|((N \cap M) \setminus H) \setminus S|$. Let the tokens that appear in all the utterances $(x_i, \ldots, x_{i+k})$ corresponding to the current section s_i be K and the tokens that appear in all the utterances $(x_i, \ldots, x_{i+p})$ corresponding to the previous section s_{i-1} be P. In scenario 2, we calculate the set operation (NW) as $|((K \cap M) \setminus P) \setminus S|$. The results in Table 6 show that people use the information in the new sections and are not fixated on old sections. It also shows that they use the information to construct the responses.

Perplexity: To automatically evaluate the fluency of the models, we use perplexity measure. We build a language model on the train set of responses using ngrams up to an order of 3[3]. The generated test responses achieve a perplexity of 21.8 for the SEQ model and **10.11** for the SEQS model. This indicates that including the sections of document helps in the generation process.

4.1 Human Evaluation

We also perform two kinds of human evaluations to evaluate the quality of predicted utterances – engagement and fluency. These experiments are performed on Amazon Mechanical Turk.

Engagement: We set up a pairwise comparison following Bennett (2005) to evaluate the engagement of the generated responses. The test presents the chat history (1 utterance) and then, in random

[2]The total number of tokens is 46000, and we limit the vocabulary to be 10000 tokens.

[3]We use the SRILM toolkit (Stolcke, 2002)

order, its corresponding response produced by the SEQ and SEQS models. A third option "No Preference" was given to participants to mark no preference for either of the generated responses. The instruction given to the participants is "Given the above chat history as context, you have to pick the one which can be best used as the response based on the engagingness." We randomly sample 90 responses from each model. Each response was annotated by 3 unique workers and we take majority vote as the final label. The result of the test is that SEQ generated responses were chosen only 36.4% times as opposed to SEQS generated responses which were chosen **43.9%** and the "No Preference" option was chosen 19.6% of times. This result shows the information from the sections improves the engagement of the generated responses.

Fluency: The workers were asked to evaluate the fluency of the generated response on a scale of 1 to 4, where 1 is unreadable and 4 is perfectly readable. We randomly select 120 generated responses from each model and each response was annotated by 3 unique workers. The SEQ model got a low score of 2.88, contrast to the SEQS score of **3.84**. This outcome demonstrates that the information in the section also helps in guiding the generator to produce fluent responses.

5 Conclusion

In this paper we introduce a crowd-sourced conversations dataset that is grounded in a predefined set of documents which is available for download [4]. We perform multiple automatic and human judgment based analysis to understand the value the information from the document provides to the generation of responses. The SEQS model which uses the information from the section to generate responses outperforms the SEQ model in the evaluation tasks of engagement, fluency and perplexity.

Acknowledgments

This work was funded by a fellowship from Robert Bosch, and in part by Facebook Inc. and Microsoft Corporation. This work was performed as a part of The Conversational Intelligence Challenge (ConvAI, NIPS 2017) and we would like to thank the ConvAI team. We are also grateful to the anonymous reviewers for their constructive feedback and to Carolyn Penstein Rose, Shivani Poddar, Sreecharan Sankaranarayanan, Samridhi Shree Choudhary and Zhou Yu for valuable discussions at earlier stages of this work.

[4] https://www.github.com/festvox/datasets/CMU_DoG

References

Layla El Asri, Hannes Schulz, Shikhar Sharma, Jeremie Zumer, Justin Harris, Emery Fine, Rahul Mehrotra, and Kaheer Suleman. 2017. Frames: A corpus for adding memory to goal-oriented dialogue systems. *arXiv preprint arXiv:1704.00057*.

Christina L Bennett. 2005. Large scale evaluation of corpus-based synthesizers: Results and lessons from the blizzard challenge 2005. In *Ninth European Conference on Speech Communication and Technology*.

Cristian Danescu-Niculescu-Mizil and Lillian Lee. 2011a. Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. In *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics*, pages 76–87. Association for Computational Linguistics.

Cristian Danescu-Niculescu-Mizil and Lillian Lee. 2011b. Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics, ACL 2011*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Pierre Lison and Jörg Tiedemann. 2016. Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles.

Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. 2015. The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. *arXiv preprint arXiv:1506.08909*.

Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Abigail See, Peter J Liu, and Christopher D Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proc. ACL*.

Iulian Vlad Serban, Ryan Lowe, Peter Henderson, Laurent Charlin, and Joelle Pineau. 2015. A survey of available corpora for building data-driven dialogue systems. *arXiv preprint arXiv:1512.05742*.

Andreas Stolcke. 2002. Srilm-an extensible language modeling toolkit. In *Seventh international conference on spoken language processing*.

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

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? *arXiv preprint arXiv:1801.07243*.