SUMBot: Summarizing Context in Open-Domain Dialogue Systems

Anonymous ACL submission

Abstract

In this paper, we investigate the problem of including relevant information as context in a dialogue system. Most models struggle to identify and incorporate important knowledge from dialogues and simply use the entire turns as context, which increases the size of the input fed to the model with unnecessary information. In order to surpass this problem, we substitute part of the context with a summary and increase the ability of models to keep track of all the previous utterances. We show that including a summary as input to a dialogue model increases the overall quality of generated responses and enhances the ability to capture information from the context long ago.

1 Introduction

Chit-chat systems have become more and more prominent with the emergence of large pre-trained models and the increased access to public libraries (Wolf et al., 2020; Gardner et al., 2017; Miller et al., 2017) that allow to easily train and deploy these models. However, these models tend to generate meaningless responses and fail to capture long-term language dependencies, particularly in the dialogue setting where conversations can attain lots of interactions and contain long turns.

Recent approaches have studied the ability of deep generative models to capture relevant information from the dialogue context (Sankar et al., 2019; Dušek and Jurcicek, 2016). They have found that these models do not efficiently make use of all parts from the dialogue history and tend to ignore relevant turn information. Other approaches (Mehri et al., 2019; Ortega and Vu, 2017; Kale and Rastogi, 2020; Henderson et al., 2020) have attempted to represent the context and leverage the resulting representations to various dialogue tasks. However, none of these approaches has studied the substitution of the context with a summary.

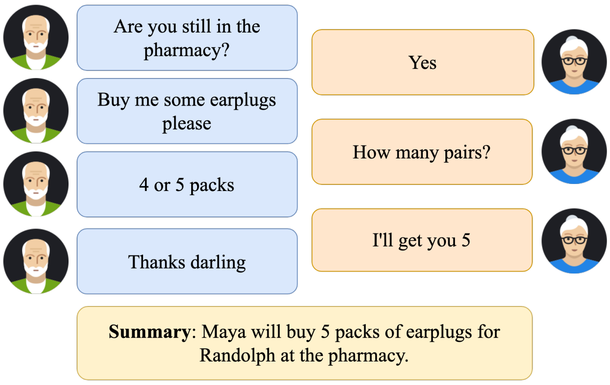
Are you still in the pharmacy?
Yes
Buy me some earplugs please
How many pairs?
4 or 5 packs
I'll get you 5
Thanks darling
Summary: Maya will buy 5 packs of earplugs for Randolph at the pharmacy.

Figure 1: Example of a dialogue between two speakers and the respective summary on the SAMSum dataset.

In this paper, we investigate the importance of reducing the context size in the open-domain dialogue task and attempt to answer the following question: can a summary of the previous context include all the important information and also decrease the input size fed to a model? To answer this question, we propose a simple yet effective method that incorporates summaries of the previous turns that are not included as input. To the best of our knowledge, we are the first to introduce and substitute parts of context with summaries as input in a decoder framework. Furthermore, we train different versions of the model by varying the amount of context fed as input and also by including (or not) a summary, which allows to directly compare the impact of adding the summary in the quality of the responses.

The training is divided in two independent stages: first, we fine-tune BART (Lewis et al., 2020) in the SAMSum corpus (Gliwa et al., 2019) and use it to generate summaries for the dialogue context. Then, we fine-tune GPT-2 with the summaries from the previous stage by incorporating them with the dialogue between both speakers.

We evaluate our model on the open-domain Persona-Chat dataset (Zhang et al., 2018) and show that it is possible to increase the overall performance of the models by substituting part of the

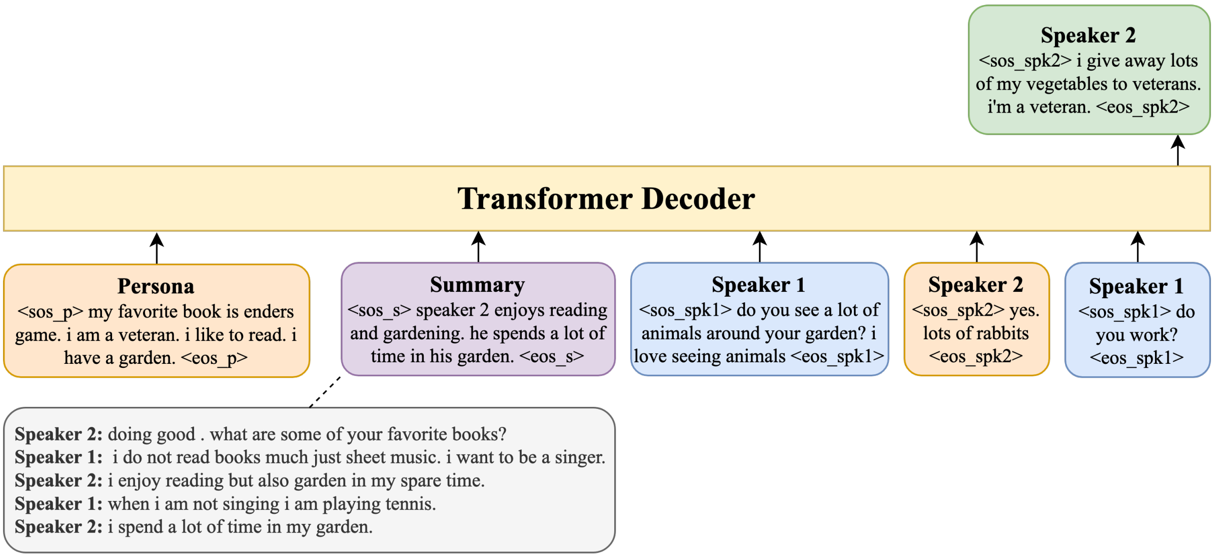
Speaker 2
<sos_spk2> i give away lots of my vegetables to veterans. i'm a veteran. <eos_spk2>
Transformer Decoder
Persona
<sos_p> my favorite book is enders game. i am a veteran. i like to read. i have a garden. <eos_p>
Summary
<sos_s> speaker 2 enjoys reading and gardening. he spends a lot of time in his garden. <eos_s>
Speaker 1
<sos_spk1> do you see a lot of animals around your garden? i love seeing animals <eos_spk1>
Speaker 2
<sos_spk2> yes. lots of rabbits <eos_spk2>
Speaker 1
<sos_spk1> do you work? <eos_spk1>
Speaker 2: doing good . what are some of your favorite books?
Speaker 1: i do not read books much just sheet music. i want to be a singer.
Speaker 2: i enjoy reading but also garden in my spare time.
Speaker 1: when i am not singing i am playing tennis.
Speaker 2: i spend a lot of time in my garden.

Figure 2: Example of an input fed to our model and the corresponding generated answer. Here, the summary represents the whole history that was not included as context.

context with a simple summary, and thus significantly reduce the size of information fed to the model, also allowing to decrease the power consumption and memory usage in the training phase.

2 Related Work

Since the introduction of sequence-to-sequence models (Sutskever et al., 2014; Bahdanau et al., 2015), chit-chat dialogue systems have been in constant evolution and are more capable of generating fluent and human-like sentences. In these systems, the encoder extracts important features from the utterances and passes that information to a decoder that generates a response.

Considering that our approach attempts to provide a proper substitute for the dialogue history, the related work that becomes more relevant is the work that focuses on studying and representing the context in the dialogue task. Sankar et al. (2019) study the aptitude of RNNs and Transformers models to interpret and understand the dialogue context by introducing synthetic perturbations to the history. They show that Transformers are less sensitive to structure perturbations and seem to fail in capturing the dialogue dynamics between turns. Henderson et al. (2020) introduce ConveRT, a lightweight framework to represent multi-turn context where it is possible to transfer learned encodings at different layers to other dialogue tasks. Liu et al. (2021) encode the dialogue context using ConveRT and merge that representations with the user request to generate appropriate and context-aware responses.

Although recent advances have introduced different techniques to represent and embody dialogue context into generative models, none of these approaches has studied the impact of using summaries as replacements for the dialogue history.

3 Method

3.1 Summary Generation

In order to summarize the dialogue context, we use BART-large (Lewis et al., 2020), a transformer architecture with a bidirectional encoder similar to BERT (Devlin et al., 2019) and a decoder similar to GPT (Radford et al., 2019), pre-trained on the English Wikipedia and BookCorpus dataset. In a preliminary phase, we fine-tune the model on the SAMSum corpus (Gliwa et al., 2019), a dataset collected for the abstractive summarization task where the goal is to summarize a dialogue between different speakers. Figure 1 shows an example of a dialogue from this dataset. After that, we use this fine-tuned model to generate the summaries for every turn of the Persona-Chat dataset.

Initially, we did some preliminary experiences where we summarized the dialogue between both speakers' utterances, which resulted in a summary that embodied relevant information from both speakers. However, by examining a few examples, we observed that the resulting summaries sometimes focused on the information from Speaker 1 and omitted relevant information from Speaker 2, which is the target speaker that functions as the answering bot. Additionally, the Personas that occur

Context	Summary	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L	Avg. Length	Max. Length
0	N	3.70	0.184	0.0423	0.176	71	115
1	N	3.94	0.192	0.0455	0.183	96	291
2	N	3.86	0.194	0.0462	0.185	118	309
3	N	4.03	**0.196**	0.0430	**0.186**	136	366
4	N	3.32	0.193	0.0450	0.184	150	274
5	N	3.89	0.180	0.0366	0.172	160	434
0	Y	3.76	0.187	0.0423	0.179	86	115
1	Y	3.95	0.195	**0.0472**	0.185	107	305
2	Y	3.95	0.191	0.0419	0.182	127	349
3	Y	3.73	0.189	0.0428	0.180	140	376
4	Y	**4.11**	0.195	0.0444	**0.186**	153	380
5	Y	4.05	0.193	0.0413	0.183	162	386

Table 1: Results for the experiments with and without summaries on Persona-Chat. Table shows that with exception of the version with context 3, the addition of a summary increases the overall BLEU score of the model. We also observe improvements in ROUGE score for all versions except with context 2 and 3.

in each dialogue are related only to Speaker 2, so we chose to only summarize the turns from Speaker 2 as we want to keep track of the context related to that speaker.

3.2 Decoder Fine-Tuning

In this stage, we fine-tune a GPT-2 transformer decoder in the Persona-Chat dataset. We use a pre-trained version trained on a large corpora of dialogues, DialoGPT (Zhang et al., 2020), which produces more relevant and context-consistent answers in comparison to the original pre-trained version, and thus becomes more suitable for our application. We build the input as seen in Figure 2, where the second speaker corresponds to the agent that will answer to the other speaker's request utterance.

Consider a dialogue d with n turns and a persona p. Then, the input t_n at the n-th turn can be described as:

$$t_n = \{p, s_{0..i-1}, c_{i..n-1}, x_n\},$$

where s is the summary of the dialogue until the context $i - 1$ (inclusive), c corresponds to the pairs of Speaker 1 and Speaker 2 full sentences represented as context, and x is the current request from Speaker 1. The model then generates an appropriate response for the n-th turn of Speaker 2 r_t according to the distribution $p(r_n|t_n)$. For instance, if the context is only the last pair of sentences, then the summary will be from the beginning of the dialogue until the last but one pair of sentences.

We create special tokens for each part of the input in order to help the model distinguish between the different segments.

4 Experiments

4.1 Experimental Setup

In order to compare and evaluate the impact of adding summaries as input, we train different versions of the model where we vary the size of context that is given. As discussed in Section 3.2, we use a version of GPT-2 trained specially for the dialogue generation task, which contains 12 layers of decoder Transformer blocks. The maximum input size is 1024 and we generate an answer with a maximum size of 200. We use Adam as the optimizer with a learning rate of $6.25e^{-5}$, and train the model for 5 epochs with patience 1. We use HuggingFace's library (Wolf et al., 2020) which provides an implementation with a language modelling head on top of the GPT-2 decoder, and generate the answer using a greedy search approach, where the next word selected is the one with the highest probability. We report BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004), both automatic metrics that measure fluency by comparing the word occurrences between the generated and the ground truth responses.

4.2 Results and Discussion

In Table 1, we present the results of the experiments with and without the inclusion of summaries, and observe that the version that achieves the higher BLEU score is the model with context 4. When we fix the context and directly compare the versions with and without context, we observe that in all with exception of the version with context 3, the inclusion of summaries improves the overall BLEU score of the results. We also observe improvements on ROUGE score in all versions except with con-

		BLEU per Turn Size						
Context	**Summary**	0	1	2	3	4	5	6
0	Y	6.33	**4.71**	3.65	**4.18**	**3.06**	**2.46**	**2.44**
0	N	**7.20**	4.63	**4.04**	3.73	2.76	2.39	2.31
1	Y	6.81	4.77	3.87	**4.05**	**3.12**	2.79	**2.99**
1	N	**7.10**	**5.09**	**4.01**	3.86	2.90	**2.80**	2.61
2	Y	**3.95**	**6.88**	4.87	3.77	3.99	**3.45**	**2.86**
2	N	3.86	5.55	**4.99**	**4.21**	**4.07**	3.25	2.75
3	Y	**6.71**	4.35	3.52	3.66	2.94	2.85	2.37
3	N	6.37	**4.60**	**4.15**	**4.60**	**3.39**	**3.16**	**2.97**
4	Y	**6.70**	**4.61**	**4.31**	**4.63**	**3.51**	2.78	**3.06**
4	N	5.74	4.17	3.48	3.56	2.67	**2.88**	1.95
5	Y	**7.43**	**4.97**	4.12	**4.29**	2.99	**2.97**	2.57
5	N	5.89	4.71	**4.25**	4.02	**3.25**	2.93	**2.64**

Table 2: BLEU score per turn. We divide the dialogues by the size of the turn. As we can observe, all models with summary except the one with context 3 perform better. We also observe that the models with summary achieve greater results when the turn size is higher.

text 2 and 3. This shows that the addition of the summary may increase the ability of the model to generate responses more appropriate and closer to the golden ones.

Input Size. We report the average size and the maximum size of the input fed to the model in the evaluation setting. As we can see, the addition of a summary to the input only increases the size provided to the model a few points. Additionally, if we required to reduce the model's input size or if we consider a scenario where the dialogues were very extensive, the inclusion of the summary would allow to reduce the size but embody important information from all previous context. Here, the summaries are fundamental as they encapsulate the context that could not be included and also reduce the size of the input needed to generate an appropriate answer.

Scores by Turn Size. Finally, we performed an extensive analysis of the results by calculating the overall score for each turn size, and compared which version of the model obtained the higher score in every turn. We consider turn size as the number of pairs of sentences between Speaker 1 and 2 plus the additional last request from Speaker 1. In Table 2, we report BLEU score for each turn size for the versions with and without summary until turn size 5. As we can observe, when the context fed to the model is lower, the addition of a summary improves the model's score, especially when the turn size is higher. This shows the effectiveness of the summaries at capturing information from the context long ago.

Response Generation. Although achieving higher results when comparing to the versions without summaries, the models still obtain a low BLEU score, which indicates that these systems are not yet prepared to the real-world scenario. By examining a few generated examples, we observed that in some cases the summary generated included irrelevant information such as greetings "*Speaker 2 wants to know how are you doing*" or excluded from the summaries important information, as turns where it mentions some of the speaker's hobbies. By omitting this information, the model is not able to understand that this information was already mentioned in a previous turn and leads to the generation of similar and repetitive responses.

5 Conclusion and Future Work

In this paper, we present a simple yet effective method for representing dialogue context in the open-domain setting. We show that it is possible to reduce the size of the input and maintain the ability to keep track of the relevant information in the previous turns. This is useful especially when the dialogues and the turns are very extensive and carry out too much irrelevant knowledge. In future work, we would like to extend the use of dialogue summarization to the task-oriented setting.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Ondřej Dušek and Filip Jurcicek. 2016. A context-aware natural language generator for dialogue systems. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, page 185–190. Association for Computational Linguistics.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke S. Zettlemoyer. 2017. Allennlp: A deep semantic natural language processing platform.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 70–79, Hong Kong, China. Association for Computational Linguistics.

Matthew Henderson, Iñigo Casanueva, Nikola Mrkšić, Pei-Hao Su, Tsung-Hsien Wen, and Ivan Vulić. 2020. Convert: Efficient and accurate conversational representations from transformers. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, page 2161–2174. Association for Computational Linguistics.

Mihir Kale and Abhinav Rastogi. 2020. Template guided text generation for task-oriented dialogue. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, page 6505–6520. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *ACL 2004*.

Ye Liu, Wolfgang Maier, Wolfgang Minker, and Stefan Ultes. 2021. Context matters in semantically controlled language generation for task-oriented dialogue systems. *arXiv:2111.14119 [cs]*. ArXiv: 2111.14119.

Shikib Mehri, Evgeniia Razumovskaia, Tiancheng Zhao, and Maxine Eskénazi. 2019. Pretraining methods for dialog context representation learning. In *ACL*.

Alexander H. Miller, Will Feng, Dhruv Batra, Antoine Bordes, Adam Fisch, Jiasen Lu, Devi Parikh, and Jason Weston. 2017. Parlai: A dialog research software platform. In *EMNLP*.

Daniel Ortega and Ngoc Thang Vu. 2017. Neural-based context representation learning for dialog act classification. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, page 247–252. 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 *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

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 Annual Meeting of the Association for Computational Linguistics*, page 32–37.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Scao, Sylvain Gugger, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. pages 38–45.

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? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and William B. Dolan. 2020. Dialogpt : Large-scale generative pre-training for conversational response generation. In *ACL*.