Building Goal-oriented Document-grounded Dialogue Systems

Xi Chen[¶ *], **Faner Lin**[¶*], **Yeju Zhou**[¶*], **Kaixin Ma**[¶],
Jonathan Francis[¶§], **Eric Nyberg**[¶], **Alessandro Oltramari**[§]
[¶]Language Technologies Institute, Carnegie Mellon University
[§]Human-Machine Collaboration, Bosch Research Pittsburgh
{xc3, fanerl, yejuz, kaixinm, jmf1, ehn}@cs.cmu.edu,
alessandro.oltramari@us.bosch.com

Abstract

In this paper, we describe our systems for solving the two Doc2Dial shared task: knowledge identification and response generation. We proposed several pre-processing and post-processing methods, and we experimented with data augmentation by pre-training the models on other relevant datasets. Our best model for knowledge identification outperformed the baseline by 10.5+ *f1*-score on the test-dev split, and our best model for response generation outperformed the baseline by 11+ *SacreBleu* score on the test-dev split.

1 Introduction

There has been a recent surge of interest in building domain-specific question answering (QA) systems, in both academia and industry. Compelling real-world applications include customer services and decision-support, wherein there is strong reliance on such QA systems to be of high quality. A significant challenges for building QA systems is that domain-specific data is relatively sparse and much noisier, compared to samples from well-studied public benchmark datasets. Also, when the answer is not explicitly present in the context, models must generate new answers instead of extracting from document, adding complexity to the problem.

In this paper, we make efforts toward building domain-oriented question answering systems, by tackling the two Doc2Dial shared-tasks[1]: *knowledge identification* and *response generation*. For knowledge identification (Subtask1), the main goal is to identify the grounding knowledge, in form of a document span, for the next-agent conversational turn. For response generation (Subtask2), the main objective is to generate the next-agent response, in natural language. We experiment with various baseline models, and we developed and evaluated our proposed solutions. Some improvement strategies we tried include post-processing, hyper-parameter tuning, and pre-training on other well-known datasets such as SQuAD (Rajpurkar et al., 2016). We found that with carefully-selected hyperparameters, and with pre-processing and post-processing heuristics, the baseline model's performance can be significantly improved: our best model is able to out-perform the provided baseline by 10.5+ *f1*-score on the test-dev split for Subtask1, and our best model for Subtask2 out-performed the baseline by 11+ *SacreBleu* score on test-dev.

2 Related Work

There are many previous works that study the problem of dialogue-based question answering. Some of them only focused on answering the questions based on dialogue history alone (Ma et al., 2018; Li et al., 2020), while, for others, the dialogue and question-answer pairs are based on a document (Choi et al., 2018). Most of these tasks are extractive in nature, meaning that the exact answer can be located in the document or dialogue. Among them, CoQA (Reddy et al., 2019) is the most similar task to Doc2Dial dataset. The main objective of the CoQA challenge is to measure machine learning models' ability to comprehend text and answer related questions that appear in a conversation; also, because some answers may not appear explicitly in the document, the model may be required to synthesize the answer based on evidence. The two sub-tasks we study in this paper differ from those described above—mainly in terms of dataset attributes. The Doc2Dial dataset mostly contains long documents and dialogues that inter-connect with each other. Moreover, the ground-truth answers in Doc2Dial are usually long as well, which makes the associated prediction tasks harder to

* Equal contribution; alphabetized by surname

[1] https://doc2dial.github.io/workshop2021/shared.html

Proceedings of the 1st Workshop on Document-grounded Dialogue and Conversational Question Answering, pages 109–112

tackle. Thus, the models and heuristics we have developed are mainly targeted towards handling these specific scenarios and problems.

3 Experiments

3.1 Dataset

The Doc2Dial dataset (Feng et al., 2020) contains two tasks: knowledge identification (Subtask1) and response generation (Subtask2). For knowledge identification: given a long document as the context, and a dialogue history between a user and an agent, the task is to identify a span of text in the document that serves as the knowledge which grounds for the next dialogue turn from agent. For response generation: given a full document and the dialogue history, the task is to directly generate an agent response for the next turn in natural language. We tackle both tasks in this paper and describe our approaches below.

3.2 Baselines

For Subtask1, the baseline model is the BERT-large-uncased-whole-word-masking model (Devlin et al., 2019). A span-extraction head is added on top of BERT, and the model is fine-tuned on the Doc2Dial knowledge identification dataset. For each example, an entire document is used as the context and the reverse concatenated dialogue history is used as the question.

For Subtask2, the baseline model is the BART-large-CNN (Lewis et al., 2020) model: a pre-trained BART model is first fine-tuned on the CNN summarization task, then fine-tuned on Doc2Dial response generation dataset. The entire document and full dialogue history are used as the context and the model is trained to generate the next dialogue response.

3.3 Approaches: Knowledge Identification

Based on error analysis of baseline results, we found that the model is making a lot of empty predictions. This is mainly because the documents in Doc2Dial are very long, necessitating a sliding-window approach. Consequently, if a text chunk does not contain any relevant information to the question, the model would predict *no-answer* with a very high confidence, preventing the model from choosing answers from other chunks. To alleviate this issue, we developed heuristics to post-process the prediction at inference time, to ensure that the model produces a valid answer. Specifically, we skip the empty prediction and select the candidate with the second highest probability at inference time. Also, prediction with the highest probability is extended to a longer span if another prediction candidate contains the prediction with the highest probability as sub-string and also has a higher start or end position probability. Besides post-processing, we also increase the sliding-window overlap size to 256 and max answer length to 80 during training, so as to get more positive instances. Moreover, since the Doc2dial dataset size is relatively small, we pre-trained the model on other QA datasets and then fine-tuned on Doc2dial. To this end, we selected SQuAD 1.1, because it is a widely used span-extraction dataset, and CoQA, because of its similar task structure, where models must answer questions based on both dialogue history and document-based context.

3.3.1 Approaches: Response Generation

For Subtask2, we start with error analysis of the baseline model and found that the model often generate responses based on the irrelevant content in the supporting documents. We hypothesize that this is because the document and the dialogue history are too long, thus it is hard for models to locate the relevant information and generate a response at the same time. If we keep only relevant knowledge grounding as input, the model will be able to generate better responses.

To test this hypothesis, we used the model trained on Subtask1 to select a chunk of document to feed in as Subtask2 input, instead of the full document. Since the span selection model is not perfect, it can select a completely wrong span, which would prevent the Subtask2 model from producing a valid response. Thus to increase the recall, we start with the best-selected span and iterate over the top-20 span predictions, in order to expand the selected span boundary and cover the next best prediction, if the the next best span is near the current selection boundary. Here, we set the threshold to be less than 500 characters away. For example, given the current start and end indices of (400, 520), if the next span prediction is (580, 650), we will change the boundary to (400, 650). However, if the next span prediction is (1200, 1300), we will stop iteration and return (400, 520). We also experimented with the ground-truth response grounding span, in order to find an upper bound of this approach.

Additionally, we only append the past two dialogue

turns to the supporting document in the input, instead of using the whole dialogue history as in (Reddy et al., 2019); it is found that most questions in a dialogue only have limited dependency, and including the past two dialogue turns may give comparable performance as including the complete dialogue history.

Another adjustment we make is to feed the past two turns of the dialogue to the decoder as input and the response will be generated following the past two dialogue turns. The intuition is that the decoder will also have more context to look at when generating its response, and we think this will make the task easier to learn.

Finally, we are interested in studying the effect of adding data. Thus, we re-formulated the CoQA dataset into a dialogue response task, and we pre-trained the BART model on CoQA before fine-tuning on Doc2Dial. Since the documents in CoQA are much shorter, we did not perform span selection as is proposed for Doc2Dial.

4 Result and Analysis

For Subtask1, we report *f1*-score and exact-match score on the dev set for our proposed method. For Subtask2, we report *SacreBleu* (Post, 2018) on the dev set. Finally, we report the test set results achieved with our best model, for both tasks.

4.1 Sub-task1: Knowledge Identification

The results for Subtask1 are shown in Table 1. We see that applying the post-processing heuristics improved the results by a significant margin. For pre-training the model on SQuAD and CoQA datasets, we see that the model achieves a small performance gain in both cases, suggesting that more data is helping the model learning more effectively and that the selection of these pre-training tasks does not conflict with the downstream task at hand. The advantage of CoQA over SQuAD also suggests that tasks with similar structure may transfer better. Finally, with the increased size of the overlap between each sliding-window, we see a decent improvement over the baseline, indicating the usefulness of the carefully chosen hyper-parameters. However, when we combined the larger overlap stride with pre-training on CoQA or SQuAD, we did not see further improvement; we leave the further investigation of this issue to future work.

Table 1: Model performance on Doc2Dial sub-task1. Here "Post." means post-processing.

Model	F1	EM
BERT	63.80	51.79
BERT + Post.	69.73	54.91
BERT + Post. + SQuAD	70.89	56.31
BERT + Post. + CoQA	72.15	57.18
BERT + Post. + 256 stride + 80 len	**72.74**	**58.53**

Table 2: Model performance on Doc2Dial sub-task2. "SS" means span selection and "DI" means additional decoder input.

Model	SacreBleu
BART (CNN)	17.69
BART (CNN) + SS	18.82
BART (CNN) + Gold span	24.86
BART + SS + DI	**31.61**
BART + SS + DI + CoQA	27.87

4.2 Sub-task2: Response Generation

The results for Subtask2 are shown in Table 2. We see that when using the selected span of text, instead of the full document, we achieved a small improvement on Bleu score; when using the ground-truth grounding span, we got a large improvement. This verified our hypothesis that shorter input will help the model generate relevant responses. The gap between these two settings suggest that a stronger span-selection model would further help the Subtask2 model improve.

Regarding the strategy of adding the last two dialogue turns to the decoder input: we switched from BART model pre-trained on CNN to a plain BART model, since the task setup is less like summarization and more like sentence completion. We see that, by adding the last 2 dialogue turns, the model's performance is improved by a large margin, showing that providing more context to the decoder indeed helps the model learn better. On the other hand, we see that pre-training on CoQA dataset actually leads to worse performance. We hypothesize that this is because of document length, where questions and answers for most dialogue turns in the CoQA dataset are much shorter than those of Doc2Dial datasets: models pre-trained on CoQA may not glean useful training signals for Doc2Dial.

Table 3: Results on Doc2Dial sub-task1 test splits.

	Test-Dev		Test	
Model	**F1**	**EM**	**F1**	**EM**
Baseline	59.51	45.45	-	-
Schlussstein	**70.12**	**56.57**	**67.31**	**50.32**

Table 4: Results on Doc2Dial Subtask2 test splits

Model	**Test-dev**	**Test**
Baseline	16.73	-
Schlussstein	**27.93**	**30.68**

4.3 Leaderboard Submission

We submitted our best models to both subtask leaderboards, and the results are shown in tables 3 and 4. Overall, our models out-performed baselines by large margins, and we got 8th place for Subtask1 and 6th place for Subtask2.

5 Conclusion

In this paper, we proposed several pre/post-processing heuristics that improve the model performance, on both knowledge identification and response generation tasks in the Doc2Dial challenge. We also found that pre-training on other question answering datasets only slightly improves the performance on knowledge identification, but did not help for response generation task. For future work, we think it is worth looking into other directions for improvement, including incorporating external knowledge bases (Ma et al., 2019) or synthetic data generation (Ma et al., 2021).

References

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. QuAC: Question answering in context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2174–2184, Brussels, Belgium. Association for Computational Linguistics.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

Song Feng, Hui Wan, Chulaka Gunasekara, Siva Patel, Sachindra Joshi, and Luis Lastras. 2020. doc2dial: A goal-oriented document-grounded dialogue dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 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, Online. Association for Computational Linguistics.

Jiaqi Li, Ming Liu, Min-Yen Kan, Zihao Zheng, Zekun Wang, Wenqiang Lei, Ting Liu, and Bing Qin. 2020. Molweni: A challenge multiparty dialogues-based machine reading comprehension dataset with discourse structure. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2642–2652, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Kaixin Ma, Jonathan Francis, Quanyang Lu, Eric Nyberg, and Alessandro Oltramari. 2019. Towards generalizable neuro-symbolic systems for commonsense question answering. *arXiv preprint arXiv:1910.14087*.

Kaixin Ma, Filip Ilievski, Jonathan Francis, Yonatan Bisk, Eric Nyberg, and Alessandro Oltramari. 2021. Knowledge-driven data construction for zero-shot evaluation in commonsense question answering. *In Proceedings of the 35th AAAI Conference on Artificial Intelligence (AAAI-21)*.

Kaixin Ma, Tomasz Jurczyk, and Jinho D. Choi. 2018. Challenging reading comprehension on daily conversation: Passage completion on multiparty dialog. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2039–2048, New Orleans, Louisiana. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Siva Reddy, Danqi Chen, and Christopher D Manning. 2019. Coqa: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.