Document-Grounded Goal-Oriented Dialogue Systems on Pre-Trained Language Model with Diverse Input Representation

Boeun Kim[*], **Dohaeng Lee**[*], **Yejin Lee** and **Harksoo Kim**
Konkuk University / Seoul, South Korea
{boeun, dsdhlee, jinjin096, nlpdrkim}@konkuk.ac.kr

Sihyung Kim
Kangwon National University / Chuncheon, South Korea
sureear@kangwon.ac.kr

Jin-Xia Huang, Oh-Woog Kwon
Electronics and Telecommunications Research Institute / Daejeon, South Korea
{hgh, ohwoog}@etri.re.kr

Abstract

Document-grounded goal-oriented dialog system understands users' utterances, and generates proper responses by using information obtained from documents. The Dialdoc21 shared task consists of two subtasks; subtask1, finding text spans associated with users' utterances from documents, and subtask2, generating responses based on information obtained from subtask1. In this paper, we propose two models (i.e., a knowledge span prediction model and a response generation model) for the subtask1 and the subtask2. In the subtask1, dialogue act losses are used with RoBERTa, and title embeddings are added to input representation of RoBERTa. In the subtask2, various special tokens and embeddings are added to input representation of BART's encoder. Then, we propose a method to assign different difficulty scores to leverage curriculum learning. In the subtask1, our span prediction model achieved F1-scores of 74.81 (ranked at top 7) and 73.41 (ranked at top 5) in test-dev phase and test phase, respectively. In the subtask2, our response generation model achieved sacreBLEUs of 37.50 (ranked at top 3) and 41.06 (ranked at top 1) in in test-dev phase and test phase, respectively.

1 Introduction

The Dialdoc21 shared task is a task that generates a proper response by finding a knowledge span from a document associated with a dialogue history. It consists of two subtasks; subtask1 for finding useful knowledge spans from a document and subtask2 for generating proper responses based on the knowledge spans. The doc2dial dataset, the dataset for the Dialdoc21 shared task, contains conversations between users and agents in real-world situations. The user and the agent engage in a conversation associated with a document, and the agent should provide the user with document-grounded information in order to guide the user. In this paper, we propose two models to perform the Dialdoc21 shared task using a pre-trained language model. In particular, we show that in the process of fine-tuning the pre-trained model, the proposed input representations significantly contribute to improving performances.

2 Related Work

The baseline models for the subtask1 and the subtask2 were proposed by Feng et al. (2020), the composers of doc2dial datasets. They formulated the subtask1 as a span selection, inspired by extractive question answering tasks such as SQuAD task (Rajpurkar et al., 2016, 2018). Zheng et al. (2020) proposed a method to reflect the differences between knowledge spans used for each turn and current knowledge span candidates. The differential information is fused with or disentangled from the contextual information to facilitate final knowledge selection. Wolf et al. (2019) constructed input presentation using word, dialog state and positional embedding.

3 Task Description

In the subtask1, our goal is to find a relevant knowledge span required for agent's response in a conversation composed of multi-turns from a given document. Inspired by Feng et al. (2020), we propose a joint model to perform both dialogue act prediction and knowledge span prediction. In the subtask2, our goal is to generate agent's response

[*] equal contribution

Proceedings of the 1st Workshop on Document-grounded Dialogue and Conversational Question Answering, pages 98–102

Title	Section ID	Span ID	Text
For Your Surviving Divorced Spouse	8	31	For Your Surviving Divorced Spouse
	9	32	If you have a surviving divorced spouse
		33	they could get the same benefits as your widow or widower provided that your marriage lasted 10 years or more.
	10	34	Benefits paid to a surviving divorced spouse won't affect the benefit amounts your other survivors will receive based on your earnings record.
	11	35	If your former spouse is caring for your child who is under age 16 or disabled and gets benefits on your record ,
		36	they will not have to meet the length - of - marriage rule.
		37	The child must be your natural or legally adopted child.

Table 1: Example of span extensions from a sentence to a title. The red cell denotes an answer span predicted by the subtask1 model. The green cell denotes a section span containing the answer span. The blue cell denotes a title span containing the predicted span.

in natural language based on a dialogue history and a document associated with the dialogue history. The dialogue history consists of speakers and utterances. Then, the document consists of sentences, tags, titles, and so on. Based on these structural information of the dialogue history and the document, we define special tokens and embeddings. Then, we propose a method to reflect these special tokens and embeddings to the well-known BART model (Lewis et al., 2020). The doc2dial dataset contains goal-oriented dialogues and knowledge documents. For developing models, three sub-datasets in four domains (DMV, VA, SSA, and studentaid) were deployed; a train dataset, a validation dataset, and a test-dev dataset. For evaluating the models, a test dataset in five domains (i.e., four seen domains (DMV, VA, SSA, studentaid) and an unseen domain (COVID-19)) was used. The test-dev dataset embodied 30% of the test dataset except for the unseen domain.

4 Key Components of Our Model

4.1 Subtask1

We adopt pre-trained RoBERTa-large model (Liu et al., 2019) as a backbone. Each dialogue turn in the train dataset and the validation dataset has a dialogue act label. We assume that agent's dialogue act aids to find a proper knowledge span. For dialogue act prediction, we use a fully connected layer added on the [CLS] output vector of the RoBERTa-large model. Then, we pointwise add special embeddings called title embeddings to conventional input representation of the RoBERTa-large model. As shown in Table 1, each span in a knowledge document has its own title. By adding the title embedding, we expect that spans sharing the same title will be tied together to help find a knowledge span. For knowledge span prediction, we use the well-known machine reading comprehension (MRC) architecture proposed by (Devlin et al., 2019). In the MRC model, each output vector of the RoBERTa-large model is fed into a bi-directional gated recurrent unit (Bi-GRU) (Cho et al., 2014) layer. Then, each output of the Bi-GRU layer is again fed into a fully connected layer for predicting a starting position and an ending position of a knowledge span. Finally, the knowledge span prediction model expands predicted spans (a sequence of words) into span segments predefined with span IDs. In this paper, these predefined span segments are called answer spans. The final loss function of the proposed span prediction model, L_{total}, is the weighted sum of the dialogue act prediction loss, $L_{dialogueact}$, and the span prediction loss, L_{span}, as follows:

$$L_{total} = \alpha * L_{dialogueact} + \beta * L_{span}$$

where α and β are weighting parameters that are set to 0.3 and 0.7, respectively. Then, the dialogue act prediction loss and the span prediction loss are calculated by minimizing cross-entropies between predicted values and gold values, respectively.

4.2 Subtask2

Token	Meaning
<user>	Beginning of user's utterance
<agent>	Beginning of agent's utterance
<doc>	Beginning of a knowledge document
<title>	Beginning of a knowledge document's title
<rank>	Bordering between answer spans
<u>	Ending of underline markup that is existed in a knowledge document
<h>	Ending of heading markup that is existed in a knowledge document

Table 2: Special tokens and their meanings.

We adopt pre-trained BART-base model (Lewis

et al., 2020) as a backbone. An input of BART's encoder consists of a dialogue history and a knowledge document. We use a current utterance and 7 previous utterances, $u_i, u_{i-1}, \ldots, u_{i-7}$, as a dialogue history. Then, we use answer spans that are constructed from 100 span candidates predicted by the knowledge span prediction model, $\hat{s}_0, \hat{s}_1, \ldots, \hat{s}_{100}$, as a knowledge document. For enriching input representation of BART's encoder, we use special tokens and additional embeddings. We first add some special tokens to BART's input, as shown in Table 2. Then, we pointwise add the following special embeddings to input representation of BART's encoder:

Type-of-Input embedding: Embedding to distinguish between a dialogue history and a knowledge document.

Rank Embedding: Embedding for representing rankings of title spans containing answer spans that are returned by the knowledge span prediction model.

Rank-in-Section Embedding: Embedding for representing rankings of answer spans in each title.

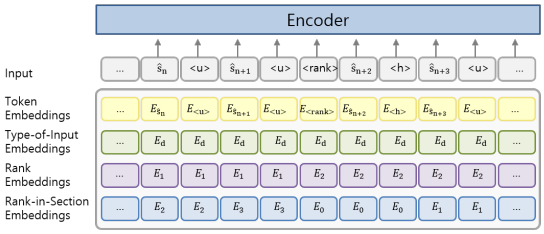

Figure 1: Special tokens and embeddings.

Figure 1 illustrates the proposed special tokens and embeddings.

5 Curriculum Learning

To improve performances, we train the proposed models through curriculum learning (Xu et al., 2020). Figure 2 illustrates the training process by curriculum learning. We first divide the training dataset into *N* buckets and train *N* teacher model (i.e., a teacher model per bucket). In this paper, *N* is set to four. Then, we measure performances of each teacher model by using *N-1* dataset except for those used for training each teacher model. Based on the performances of the teacher models, we assign each data to difficulty levels.

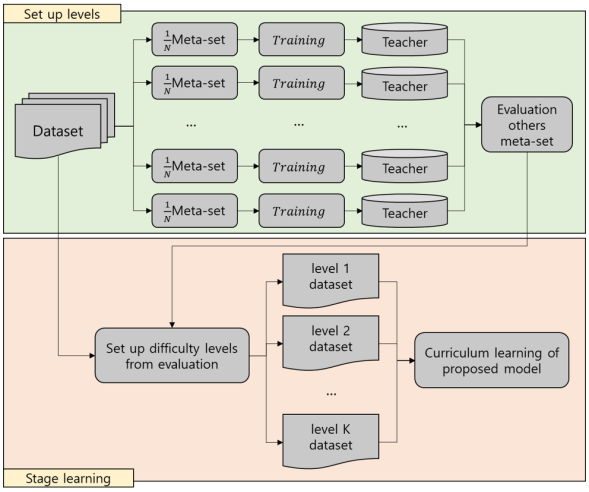

Figure 2: Curriculum learning process. *K* denotes the number of difficulty levels.

5.1 Difficulty level for subtask1

In the subtask1, we implement four teacher model based on the RoBERTa-base model (Liu et al., 2019). Each teacher model calculates an average of F1-score and EM-score (i.e., F1-score + EM score / 2) per input data. Then, the average scores of three teacher models are summed. According to the summed average scores, we divide the training dataset into an easiest level (the summed average score of 300), an easy level (the summed average score of (200,300)), a median level (the summed average score of (100,200]), and a difficult level (the summed average score of (0,100]). The numbers of data in each level are 5,390, 3,215, 6,538 and 9,260, respectively.

5.2 Difficulty level for subtask2

In the subtask2, we implement four teacher models based on the BART-base model (Lewis et al., 2020). We compute an average sum of sacreBLEUs evaluated by each teacher model. Then, we perform human evaluations on the computed average sums. Based on the human evaluations, we divide the training dataset into an easy level (sacreBLEU of [30,100]), a median level (sacreBLEU of [15,30)), and a difficult level (sacreBLEU of [0,15)). The numbers of data in each level are 8,165, 3,976, and 12,262, respectively.

5.3 Training detail

Based on the measured difficulty scoring, the total training stage consists of *K+1* phases. For instance, if *K* is set to two, the difficulty level comprises of two levels, i.e., "easy" and "difficult", and the

training stage is composed of three phases. Concisely, we sort training datasets through difficulty levels. In the first stage, we train the model by using *1/K* dataset of "easy" level. In the second stage, we train the model by using *1/K* dataset of "easy" level and *1/K* dataset of "difficult" level excluding data used for the previous training stage. In the last stage, we train the model by using the entire training dataset until convergence. Since we use K as 4 in subtask1 and K as 3 in subtask2, each stage is composed of five phases and four phases.

6 Experiments

Models	F1	EM
BERT-large	67.96	52.02
+ DA	69.29	54.04
RoBERTa-large	-	-
+ DA	72.23	56.06
+ DA + T	72.91	57.07
+ DA + T + CL	74.81	59.59

Table 3: Subtask1 test-dev phase results. DA denotes the dialogue act prediction, T denotes the title embedding, and CL denotes the curriculum learning.

As shown in Table 3, the span prediction model based on RoBERTa-large showed better performances than that based on BERT-large (Devlin et al., 2019). The dialogue act contributed to improving performances: "BERT-large+DA" showed F1-score of 1.33%p higher and EM score of 2.02%p higher than "BERT-large". The title embedding contributed to improving performances: "RoBERTa-large+DA+T" showed F1-score of 0.68%p higher and EM score of 1.01%p higher than "RoBERTa-large+DA". Moreover, the curriculum learning significantly contributed to improving performances: "RoBERTa-large+DA+T+CL" showed F1-score of 1.9%p higher and EM score of 2.52%p higher than "RoBERTa-large+DA+T". Table 4 lists results of the subtask2 in the test-dev phase.

As shown in Table 4, the Type-of-Input embedding contributed to improving the sacreBLEU of 2.74%p compared to BART-base. Adding the Rank embedding improved the score by 5.39%p, and adding the Rank-in-Section embedding boosts the performance by another 4.47%p. Finally, the curriculum learning improved the sacreBLEU of 1.81%p.

Models	SacreBLEU
BART-base	23.09
+ TI	25.83
+ TI + R	31.22
+ TI + R + RS	35.69
+ TI + R + RS + CL	37.50

Table 4: Subtask2 test-dev phase results. TI denotes the Type-of-Input embedding, R denotes the Rank embedding, RS denotes the Rank-in-Section embedding, and CL denotes the curriculum learning.

7 Conclusion

We proposed a document-grounded goal-oriented dialogue system for the Dialdoc21 shared task. The proposed model used various special tags and embeddings for enriching input representation of pre-trained language models, RoBERTa-large for knowledge span prediction and BART for response generation. In addition, curriculum learning was adopted to achieve performance improvements. In the subtask1, our span prediction model achieved F1-scores of 74.81 (ranked at top 7) and 73.41 (ranked at top 5) in test-dev phase and test phase, respectively. In the subtask2, our response generation model achieved sacreBLEUs of 37.50 (ranked at top 3) and 41.06 (ranked at top 1) in test-dev phase and test phase, respectively.

Acknowledgments

This work was supported by Institute of Information & Communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (2019-0-00004, Development of semi-supervised learning language intelligence technology and Korean tutoring service for foreigners). Also, this work was supported by Institute of Information & communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (No.2020-0-00368, A Neural-Symbolic Model for Knowledge Acquisition and Inference Techniques)

References

Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Hol-

ger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. In *EMNLP*.

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.

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)*, pages 8118–8128.

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.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392.

Thomas Wolf, Victor Sanh, Julien Chaumond, and Clement Delangue. 2019. Transfertransfo: A transfer learning approach for neural network based conversational agents. *arXiv preprint arXiv:1901.08149*.

Benfeng Xu, Licheng Zhang, Zhendong Mao, Quan Wang, Hongtao Xie, and Yongdong Zhang. 2020. Curriculum learning for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6095–6104.

Chujie Zheng, Yunbo Cao, Daxin Jiang, and Minlie Huang. 2020. Difference-aware knowledge selection for knowledge-grounded conversation generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 115–125.