GKS: Graph-based Knowledge Selector for Task-oriented Dialog System

Jen-Chieh Yang*[1], Jia-Yan Wu*[2], Sung-Ping Chang[3], Ya-Chieh Huang[2]

[1] Academic Sinica
[2] National Taiwan University
[3] Columbia University
richardyangms3@gmail.com, b06701216@ntu.edu.tw, sksk29292929@gmail.com, b06701235@ntu.edu.tw

Abstract

In previous research, knowledge-selection tasks mostly rely on language model-based methods or knowledge ranking. However, while approaches that rely on the language models take all knowledge as sequential input, knowledge does not contain sequential information in most circumstances. On the other hand, the knowledge-ranking methods leverage dialog history and each given knowledge snippet separately, but they do not consider information between knowledge snippets. In the Tenth Dialog System Technology Challenges (DSTC10), we participated in the second Knowledge-grounded Task-oriented Dialogue Modeling on Spoken Conversations. To deal with the problems mentioned above, we modified training methods based on state-of-the-art (SOTA) models for the first and third sub-tasks. As for the second sub-task of knowledge selection, we proposed Graph-Knowledge Selector (GKS), utilizing a graph-attention base model incorporated with the language model. GKS makes knowledge-selection decisions in the dialog by simultaneously considering each knowledge embedding generated from the language model without sequential features. Moreover, GKS leverages considerable knowledge in decision-making and takes relations across knowledge as part of the selection process. As a result, GKS outperforms several SOTA models proposed in the data-set on knowledge selection from the Ninth Dialog System Technology Challenges (DSTC9).

1 Introduction

Nowadays, task-oriented dialog systems have been widely used, providing specific services in different industries. Therefore, building dialog systems to deal with heterogeneous data becomes significant to make robust and general frameworks for wide-range applied scenarios. The problem amplifies when input data is considered to be spoken language, making input data with even more variance. In this paper, we proposed our approach to constructing a dialog system in each sub-task to alleviate differentiation for data sets. We mainly focus on the Knowledge-selection task, aiming to solve potential obstacles past knowledge selection might encounter.

In DSTC10 track 2, the goal is to perform a dialog system on spoken language, which is not straightforward via a pre-trained language model. Regarding task 1, we design our knowledge-seeking turn detection model with a denoise language model. The model is pre-trained with text data, aiming to transfer spoken language to text-style data.

Knowledge selection approach in the past mainly utilized language models. These approaches often have minor or no change on model structure but training objective and some training techniques, such as data augmentation. Here, we assume information within knowledge for dialog should not be taken simply in a sequential style language model (see. Table 1).To better capture knowledge information, we form our knowledge selection model with a graph-based design for task 2. The proposed Graph-Knowledge Selector (GKS) takes knowledge-embedding as input to make selection tasks. GKS does not simply concatenate knowledge and question together into a language model as input and output prediction. Instead, GKS leverages information between each knowledge and question with a graph-attention model.

To assure our proposed frameworks in previous tasks, we select a model proposed in DSTC9 without much modification. Therefore, we choose (Bao et al. 2020) with modification as our generation model, which takes the output from knowledge selection task prediction as input.

The main contribution in this paper is a new knowledge selection model for the dialog system. However, since our approach does not solve the problem of spoken language quite well, we further test our framework on DSTC9 data set, which consists of text data. The experiment result shows that our proposed framework outperforms models proposed in last year's challenge. Our implementation will be released upon acceptance.

2 Related Work

The recent development of task-oriented dialogue systems benefits from pre-trained language models like GPT-2 (Radford et al. 2018; Budzianowski and Vulic 2019; Ham et al. 2020). Furthermore, BERT-like (Devlin et al. 2019) architectures, such as RoBERTa (Liu et al. 2019), achieved state-or-the-art performance on natural language understanding tasks like GLUE dataset (Wang et al. 2018). In subtasks 1 and 3, we leveraged this architecture to attain better performance.

*These authors contributed equally.

Detected turn	**Does the hotel offer accessible parking?**
Knowledge0	Does the hotel offer accessible parking?
Knowledge1	Is there on-site private parking at the Bridge Guest House?
Knowledge2	Do I have to pay for parking?
Knowledge3	Is there a cost for parking?
Knowledge4	Can I make a reservation for parking?
Knowledge5	Do they have help for disabled parking?
Knowledge6	Do you provide room service daily?
Knowledge7	Are there any fitness center or gym?

Table 1: Example from dialog turn and knowledge. The desired knowledge for the user should be Knowledge0. However, there are no sequential relations between knowledge, but many approaches simply connect knowledge as input. Moreover, if we train the model simply pairing Detected Turn with each knowledge and sampling training pairs, the overlapping information between knowledge might be wasted. In the example, Knowledge0-5 refers to parking issues; we try to let the model distinguish Knowledge0 from all the others according to detected turn but not with sampling approaches.

2.1 Knowledge-Seeking Turn Detection

The knowledge-seeking turn detection was first introduced in DSTC9 track1 by (Kim et al. 2020). A binary classifier was proposed to solve this task, while Tang et al. (2021) used a supervised method with an ELECTRA-based model followed by a fully connected layer (Clark et al. 2020) to determine whether to trigger knowledge. In another way, Mi et al. (2021) employed an ensemble algorithm of four pre-trained models, such as RoBERTa and UniLM (Dong et al. 2019), to solve it as a classification task as well. Bao et al. (2020) proposed to consider API and external knowledge by schema description method. (Shah et al. 2019; Eric et al. 2019; Rastogi et al. 2020)

2.2 Knowledge Selection

The knowledge selection task is to retrieve candidate snippets from the knowledge base for response generation. Traditionally, TF-IDF technique (Ramos 2003) and language model are applied on similar tasks. As mentioned above, the limitations of previous works lie in the model structure. On the other hand, we are inspired by Kernel Graph Attention Network proposed by (Liu et al. 2021), which performs fine-grained fact verification with kernel-based attention. We believe such a graph-based model can better capture information and select a more plausible set of knowledge snippets.

2.3 Knowledge Grounded Generation

The third component of our system is to generate responses given select knowledge snippets. Recently, pre-trained language models such as BERT propel progress in natural language generation, but also demonstrate limitation while being directly fine-tuned on small conversation datasets (Rashkin et al. 2019; Wolf et al. 2019). PLATO (Bao et al. 2020) was proposed to address this problem by using uni- and bi-directional processing with further pre-training on large-scale Reddit and Twitter conversations. In addition, they introduced a discrete latent variable to grasp one-to-many utterances relationship information between conversations. In DSTC9 track1, Tan et al. (2020) further incorporated knowledge copy method to calculate the probability of response by combining generation distribution with the knowledge-attention distribution. Tan et al. (2020) provides an efficient way to generate sentences under the given knowledge, reducing the pressure added on the decoder and is easier for models to generalize to unseen knowledge.

3 Methodology

In this section, we first define our problem by dividing it into three separated sub-tasks, then discuss how we address each part of it with different models. Fig 1 shows our overall framework.

3.1 Knowledge-Seeking Turn Detection

In the first phase of our system, we deploy a binary classifier to decide whether to trigger the knowledge access branch for a given utterance and dialogue history.

Data Representations To capture whether the current dialog turn will need to trigger the knowledge, we decide to concatenate the current dialog with dialog history. We hope the model can consider richer information than only using one dialog turn. Besides, to make the model understand the speaker of dialog, we added a speaker token ([User] or [Sys]) before every dialog. The speaker tokens represent this dialog turn is spoken by the user or system. We believe that different speakers will provide implicit information to the dialog history and will make our system perform better. Below is the representation of our input data:

$$[User]U_i[Sys]S_0[User]U_0...[Sys]S_{i-1} \quad (1)$$

where U_i equals to the i_{th} turn of user, S_i equals to the i_{th} turn of system.

Binary Classification Model In this part, we defined Knowledge-Seeking Turn Detection as a binary classification task. To extract informative features in the dialog context, we chose to use RoBERTa as our encoder since it currently outperformed most pre-trained language models. Besides, we applied a new dialogue turn embedding, which represents the number of turns in the whole dialogue, in our training procedure. We hope the model can learn more from this embedding and regard turn number as important information. After fine-tuning the RoBERTa model, the probability of x_0 being the correct answer is calculated as:

$$e_0 = \sum W_h h_0, \quad (2)$$

$$p(x_0) = \texttt{sigmoid}(e_0), \quad (3)$$

where $h_0 \in \mathbb{R}^{d_{h_j}}$ is the output hidden states of [CLS] token, $W_h \in \mathbb{R}^{d_{h_j}}$ are trainable parameters, $p(x_0)$ is the probability that input dialog will need to trigger the knowledge branch access.

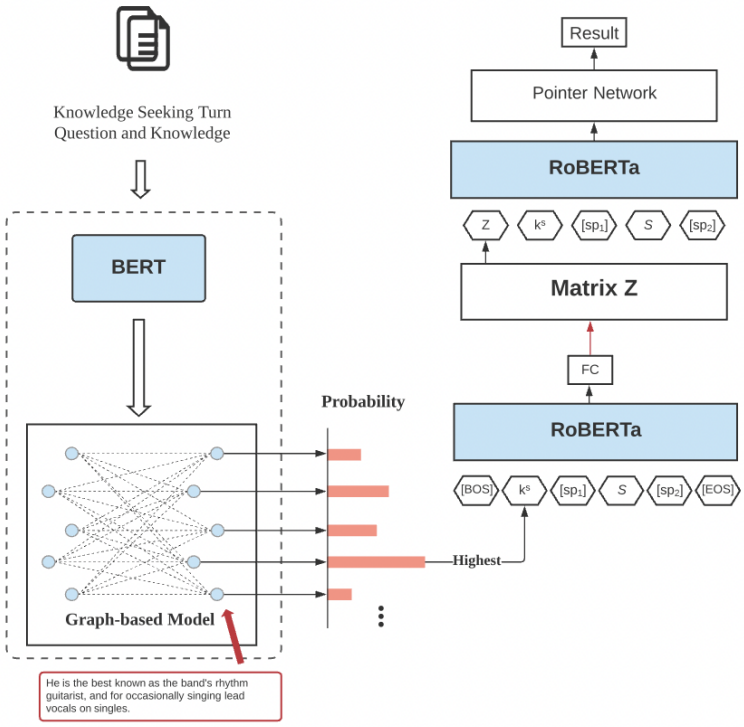

Figure 1: The figure shows the overall framework for knowledge selection and knowledge-grounded generation. The left part describes the workflow of knowledge selection, generating the probability for each node. The generation model takes the highest probability among all nodes as knowledge snippets as input for generation.

3.2 Knowledge Selection

This section describes how we develop our graph-based knowledge selection model GKS.

Knowledge Embedding To construct node embedding for GKS, we first build knowledge embedding with the BERT model. For the pre-training BERT model, we first concatenate detected knowledge-needed questions (refer to detected knowledge-seeking turns) and each knowledge with [SEP] token as input, and add [CLS] at the first of every question-knowledge pair. We train the BERT model combined with a linear layer as a binary classification task, referring to "Related" and "Unrelated." We then reference this pre-trained BERT model to make node embedding.

$$e^n = BERT(k^n) \tag{4}$$

k^n is the n^{th} knowledge in the set K connected with question with [SEP], where K is the knowledge set consists of $k^1...k^n...k^l$ with l knowledge pieces of the current entity. e^n is the hidden state of the n^{th} knowledge-question pair. To elaborate, e^n_0 represent [CLS]. $e^n_{1:m}$ and $e^n_{1:m+p}$ represent question and knowledge where question contains m words (tokens) and knowledge p words (tokens).

Graph-Attention Knowledge Selection Inspired by (Liu et al. 2021), which successfully captures text information with a graph attention model for factual verification, we develop our selection model with the graph-based model. Based on our aforementioned assumption, we believe the graph-based neural networks can better capture similar traits in knowledge snippets, which resembles clustering for unsupervised learning. The prediction is made per-node, which the node with the highest probability indicates the predicted knowledge. We follow (Zhou et al. 2019), utilizing node kernel to determine relevance between dialog turn and each knowledge with "readout" function. First, Graph-Attention Knowledge Selector (GKS) applies the translation matrix M^{q,e^n} on the n^{th} knowledge hidden state k^n and q, where q is the question of the user, by the hidden state $e^n_{1:m}$ and $e^n_{1:m+p}$. GKS then applies the kernel $Kernel$ match feature on the translation matrix to construct node representation $S(k^n)$ for knowledge selection

$$S(k^n) = \frac{1}{m}\sum_{i}^{m} Kernel(M^{q,k^n}) \tag{5}$$

and

$$P(k^n|E) = softmax_p(Linear(S(k^n))) \tag{6}$$

function as the readout to calculate n^{th} knowledge selection probability $P(k^n|E)$. The whole model is trained end-to-end by minimizing the cross entropy loss:

$$L = CrossEntropy(k^*, P(k^p|E)) \tag{7}$$

where k^* is the golden knowledge to the detect knowledge turn.

Model	Recall	Precision	F1
Baseline	0.9992	0.9719	0.9853
ROBERTA-HS	0.9981	0.9963	0.9972
ROBERTA-WD (DA)	**0.9996**	**0.9985**	**0.9990**
(He et al. 2021)	0.99102	0.9969	0.9939
(Tang et al. 2021)	0.9817	0.9465	0.9638
Ours	0.9918	0.9921	0.9920

Table 2: Our result of Knowledge-Seeking Turn Detection. Where ROBERTA-HS and ROBERTA-WD (DA) are from Mi et al. (2021).

3.3 Knowledge Grounded Generation

After candidate knowledge snippets are given, we then select the one snippet with the highest probability as an input for knowledge grounded generation. Inspired by Bao et al. (2020) and Tan et al. (2020), we then leverage RoBERTa-based architecture with the consideration of latent variable. Inspired by Tan et al. (2020), we concatenate knowledge snippet and dialogue history with special tokens as input. Unlike Tan et al. (2020), we consider both questions and answer part of the knowledge snippet:

$$< BOS > k^s < sp1 > s_1 < sp2 > s_2 ... < sp2 > r < EOS > \quad (8)$$

where k^s represents selected knowledge snippet, $< sp_1 >$ and $< sp_2 >$ represent two speakers in the dialogue respectively, and s_n denotes n^{th} term in dialog history . r represents the response. Following Tan et al. (2020), we encode response into the Z matrix, of which each row represents a special z corresponding to given examples. To select a specific z as our latent variable, we estimate posterior probability $q_\phi(z|S, k^s, r)$, where S denotes dialogue history. The rest of the architecture and calculation process, such as knowledge copy mechanism, segmented response generation, and modified beam search, are essentially identical to the ones in Tan et al. (2020). We illustrate it with the subtask2 model in Figure 1.

4 Experiments

This section demonstrates our experiment results. For the Baseline and chosen baselines, we report from their paper presented in DSTC9 last year.

4.1 Knowledge Seeking-Turn Detection

Table 2 shows our result on the DSTC9 data-set. As the proposed baseline model of DSTC9 is already performing very well, other proposed models have only a slight difference in their results. Our experiment results show that our model outperformed several baselines on the F1 score, which indicates that our approach on the training model with [User] and [Sys] tokens gives the language model more ability to learn user utterances patterns. Mi et al. (2021) perform better from selected models. We assume their proposed training strategy with data augmentation is the key reason to gain performance under the condition that most language models could gain excellent performance on the original data-set.

Model	Acc@5	Acc@1	MRR@5
Baseline	0.8772	0.6201	0.7263
ROBERTA-WD	0.9745	0.9145	0.9428
ROBERTA-WD (IS)	0.9741	0.9456	0.9589
ROBERTA-WD-listwise	0.9752	0.9394	0.9566
(He et al. 2021)	0.9892	0.9465	**0.9665**
(Tang et al. 2021)	0.9665	0.9117	0.9372
KGS	**0.9899**	**0.95435**	-

Table 3: Result of our presented Knowledge Selection model KGS. ROBERTA-WD, ROBERTA-WD (IS), and ROBERTA-WD-listwise are proposed in Mi et al. (2021).

4.2 Knowledge Selection

Since our motivation is aiming to develop a better solution for the knowledge section without noise in spoken language translation and preventing potential defects mentioned in earlier sections that previous approaches don't cope with, we further test our proposed model on the DSTC9 dataset. Table 3 shows the final result of KGS model performance on the DSTC9 track one dataset. We selected several models proposed in last year's competition and SOTA models as the baseline. ROBERTA-WD (IS) in (Mi et al. 2021) used sampling technique and k-fold cross-validation during the training process. (He et al. 2021) acquired multi-scale negatives to replace random sampling, which might lead to coarse-grained class separation. (Tang et al. 2021) is an ELECTRA-based model with proposed aggregated loss, which contains the correlation between the domains, entity names, knowledge snippets, and dialog contexts. The result shows that our model, which applies a graph-based model in the selection process, outperforms past approaches that only rely on language models, even without data augmentation. The results suggested that our proposed graph-based architecture did enhance the performance as our settings on knowledge embedding generation was simpler than other SOTA models.

4.3 Knowledge Grounded Generation

The generated results are demonstrated in Table 4, it is commensurate with others in DSTC9. Following Tan et al. (2020), our RoBERTa-based model has the same hyperparameters as the baseline model in Kim et al. (2021). The learning rate is $6.25e-5$, the batch size is 4, and the number of gradient accumulation steps is 32. The number of hidden variable z is set to 5. Our model is trained in 20 epochs, and we use a copy mechanism followed by vanilla beam search to get our final generated result.

5 Conclusions

In this paper, we proposed a framework for DSTC10 and DSTC9. Our main goal is to develop a better solution for knowledge selection tasks, which only rely on language models to perform selection in the past. The results showed that our proposed knowledge selection model with a graph-based model performed better than the proposed models last year. For our future goal, we are interested in replacing knowledge turn question embedding, which is constructed

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-1	ROUGE-2	ROUGE-L
Baseline	0.3601	0.2202	0.1389	0.0956	0.3600	0.3939	0.1749	0.3501
(Mi et al. 2021)	0.4330	**0.3061**	**0.2133**	**0.1616**	**0.4535**	**0.4795**	**0.2520**	**0.4304**
(Tang et al. 2021)	0.3684	0.2374	0.1531	0.1030	0.3719	0.4113	0.1938	0.3692
(He et al. 2021)	0.4267	0.2789	0.1858	0.1357	0.4324	0.4587	0.2249	0.4093
Ours	**0.4356**	0.2978	0.1993	0.1378	0.4400	0.4711	0.2415	0.4262

Table 4: Automatic metric of our generation model against other baselines in subtask3.

with text sentence embedding in our original setting, with wave embedding. We assume this could better obtain spoken without hurting the overall system.

References

Bao, S.; He, H.; Wang, F.; Wu, H.; and Wang, H. 2020. PLATO: Pre-trained Dialogue Generation Model with Discrete Latent Variable. arXiv:1910.07931.

Budzianowski, P.; and Vulic, I. 2019. Hello, It's GPT-2 - How Can I Help You? Towards the Use of Pretrained Language Models for Task-Oriented Dialogue Systems. *CoRR*, abs/1907.05774.

Clark, K.; Luong, M.-T.; Le, Q. V.; and Manning, C. D. 2020. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. arXiv:2003.10555.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 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)*, 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics.

Dong, L.; Yang, N.; Wang, W.; Wei, F.; Liu, X.; Wang, Y.; Gao, J.; Zhou, M.; and Hon, H.-W. 2019. Unified Language Model Pre-training for Natural Language Understanding and Generation. arXiv:1905.03197.

Eric, M.; Goel, R.; Paul, S.; Kumar, A.; Sethi, A.; Ku, P.; Goyal, A. K.; Agarwal, S.; Gao, S.; and Hakkani-Tur, D. 2019. MultiWOZ 2.1: A Consolidated Multi-Domain Dialogue Dataset with State Corrections and State Tracking Baselines. arXiv:1907.01669.

Ham, D.; Lee, J.-G.; Jang, Y.; and Kim, K.-E. 2020. End-to-End Neural Pipeline for Goal-Oriented Dialogue Systems using GPT-2. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 583–592. Online: Association for Computational Linguistics.

He, H.; Lu, H.; Bao, S.; Wang, F.; Wu, H.; Niu, Z.; and Wang, H. 2021. Learning to Select External Knowledge with Multi-Scale Negative Sampling. arXiv:2102.02096.

Kim, S.; Eric, M.; Gopalakrishnan, K.; Hedayatnia, B.; Liu, Y.; and Hakkani-Tur, D. 2020. Beyond Domain APIs: Task-oriented Conversational Modeling with Unstructured Knowledge Access. arXiv:2006.03533.

Kim, S.; Eric, M.; Hedayatnia, B.; Gopalakrishnan, K.; Liu, Y.; Huang, C.-W.; and Hakkani-Tur, D. 2021. Beyond Domain APIs: Task-oriented Conversational Modeling with Unstructured Knowledge Access Track in DSTC9. arXiv:2101.09276.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *CoRR*, abs/1907.11692.

Liu, Z.; Xiong, C.; Sun, M.; and Liu, Z. 2021. Fine-grained Fact Verification with Kernel Graph Attention Network. arXiv:1910.09796.

Mi, H.; Ren, Q.; Dai, Y.; He, Y.; Sun, J.; Li, Y.; Zheng, J.; and Xu, P. 2021. Towards Generalized Models for Beyond Domain API Task-oriented Dialogue. *DSTC 9*.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; and Sutskever, I. 2018. Language Models are Unsupervised Multitask Learners. *OpenAI blog*.

Ramos, J. E. 2003. Using TF-IDF to Determine Word Relevance in Document Queries. In *Proceedings of the first instructional conference on machine learning*.

Rashkin, H.; Smith, E. M.; Li, M.; and Boureau, Y.-L. 2019. Towards Empathetic Open-domain Conversation Models: a New Benchmark and Dataset. arXiv:1811.00207.

Rastogi, A.; Zang, X.; Sunkara, S.; Gupta, R.; and Khaitan, P. 2020. Schema-Guided Dialogue State Tracking Task at DSTC8. arXiv:2002.01359.

Shah, D. J.; Gupta, R.; Fayazi, A. A.; and Hakkani-Tur, D. 2019. Robust Zero-Shot Cross-Domain Slot Filling with Example Values. arXiv:1906.06870.

Tan, C.-H.; Yang, X.; Zheng, Z.; Li, T.; Feng, Y.; Gu, J.-C.; Liu, Q.; Liu, D.; Ling, Z.-H.; and Zhu, X. 2020. Learning to Retrieve Entity-Aware Knowledge and Generate Responses with Copy Mechanism for Task-Oriented Dialogue Systems. arXiv:2012.11937.

Tang, L.; Shang, Q.; Lv, K.; Fu, Z.; Zhang, S.; Huang, C.; and Zhang, Z. 2021. RADGE: Relevance Learning and Generation Evaluating Method for Task-Oriented Conversational Systems. *DSTC 9*.

Wang, A.; Singh, A.; Michael, J.; Hill, F.; Levy, O.; and Bowman, S. R. 2018. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. *CoRR*, abs/1804.07461.

Wolf, T.; Sanh, V.; Chaumond, J.; and Delangue, C. 2019. TransferTransfo: A Transfer Learning Approach for Neural Network Based Conversational Agents. arXiv:1901.08149.

Zhou, J.; Han, X.; Yang, C.; Liu, Z.; Wang, L.; Li, C.; and Sun, M. 2019. GEAR: Graph-based Evidence Aggregating and Reasoning for Fact Verification. arXiv:1908.01843.