Can I Be of Further Assistance? Using Unstructured Knowledge Access to Improve Task-oriented Conversational Modeling

Di Jin
Amazon Alexa AI
djinamzn@amazon.com

Seokhwan Kim
Amazon Alexa AI
seokhwk@amazon.com

Dilek Hakkani-Tur
Amazon Alexa AI
hakkanit@amazon.com

Abstract

Most prior work on task-oriented dialogue systems are restricted to limited coverage of domain APIs. However, users oftentimes have requests that are out of the scope of these APIs. This work focuses on responding to these beyond-API-coverage user turns by incorporating external, unstructured knowledge sources. Our approach works in a pipelined manner with knowledge-seeking turn detection, knowledge selection, and response generation in sequence. We introduce novel data augmentation methods for the first two steps and demonstrate that the use of information extracted from dialogue context improves the knowledge selection and end-to-end performances. Through experiments, we achieve state-of-the-art performance for both automatic and human evaluation metrics on the DSTC9 Track 1 benchmark dataset, validating the effectiveness of our contributions.

1 Introduction

Driven by the fast progress of natural language processing techniques, we are now witnessing a variety of task-orientated dialogue systems being used in daily life. These agents traditionally rely on pre-defined APIs to complete the tasks that users request (Williams et al., 2017; Eric et al., 2017); however, some user requests are related to the task domain but beyond these APIs' coverage (Kim et al., 2020a). For example, while task-oriented agents can help users book a hotel, they fall short of answering potential follow-up questions users may have, such as "whether they can bring their pets to the hotel". These beyond-API-coverage user requests frequently refer to the task or entities that were discussed in the prior conversation and can be addressed by interpreting them in context and retrieving relevant domain knowledge from web pages, for example, from textual descriptions and frequently asked questions (FAQs). Most task-oriented dialogue systems do not incorporate these external knowledge sources into dialogue modeling, making conversational interactions inefficient.

To address this problem, Kim et al. (2020a) recently introduced a new challenge on task-oriented conversational modeling with unstructured knowledge access, and provided datasets that are annotated for three related sub-tasks: (1) knowledge-seeking turn detection, (2) knowledge selection, and (3) knowledge-grounded response generation (one data sample is in Section B.1 of Supplementary Material). This problem was intensively studied as the main focus of the DSTC9 Track 1 (Kim et al., 2020b), where a total of 105 systems developed by 24 participating teams were benchmarked.

In this work, we also follow a pipelined approach and present novel contributions for the three sub-tasks: (1) For knowledge related turn detection, we propose a data augmentation strategy that makes use of available knowledge snippets. (2) For knowledge selection, we propose an approach that makes use of information extracted from the dialogue context via domain classification and entity tracking before knowledge ranking. (3) For the final response generation, we leverage powerful pre-trained models for knowledge grounded response generation in order to obtain coherent and accurate responses. Using the challenge test set as a benchmark, our pipelined approach achieves state-of-the art performance for all three sub-tasks, in both automated and manual evaluation.

2 Approach

Our approach to task-oriented conversation modeling with unstructured knowledge access (Kim et al., 2020a) includes three successive sub-tasks, as illustrated in Figure 1. First, knowledge-seeking turn detection aims to identify user requests that

Proceedings of the 1st Workshop on Document-grounded Dialogue and Conversational Question Answering, pages 119–127

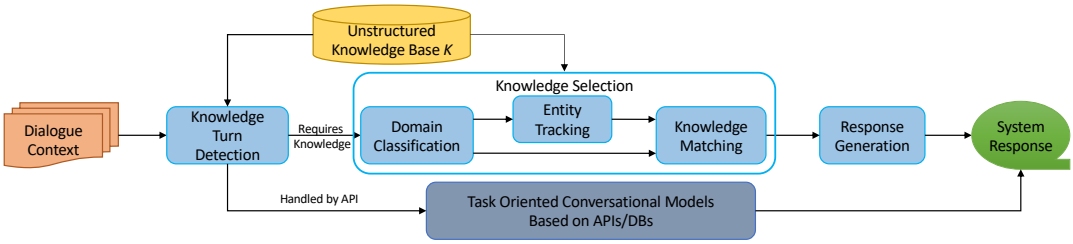

Figure 1: Task formulation and architecture of our knowledge-grounded dialog system.

are beyond the coverage of the task API. Then, for detected queries, knowledge selection aims to find the most appropriate knowledge that can address the user queries from a provided knowledge base. Finally, knowledge-grounded response generation produces a response given the dialogue history and selected knowledge.

DSTC9 Track 1 (Kim et al., 2020b) organizers provided a baseline system that adopted the fine-tuned GPT2-small (Radford et al., 2019) for all three sub-tasks. The winning teams (Team 19 and Team 3) extensively utilized ensembling strategies to boost the performance of their submissions (He et al., 2021; Tang et al., 2021; Mi et al., 2021). We follow the pipelined architecture of the baseline system, but made innovations and improvements for each sub-task, outlined in detail below.

2.1 Knowledge-seeking Turn Detection

We treat knowledge-seeking turn detection as a binary classification task, given the dialogue context as the input, and fine-tuned a pre-trained language model for this purpose. The knowledge provided in the knowledge base constitutes a set of FAQs. We augmented the available training sets by treating all questions in the knowledge base as new potential user queries. Furthermore, for all questions in this augmentation that contain an entity name, we created a new question by replacing this entity name with "it". In this way, we obtained 13,668 additional data samples. In contrast to the baseline, we found that replacing GPT2-small with RoBERTa-Large (Liu et al., 2019) improved the performance. The other changes we made include feeding only the last user utterance instead of the whole dialogue context into the model and fine-tuning the decision threshold t_{ktd} (when the inferred probability score $p > t_{ktd}$, the prediction is positive, otherwise negative) to optimize the F1 score on the validation set, both of which helped achieve better performance.

2.2 Knowledge Selection

For knowledge selection, the baseline system predicts the relevance between a given dialogue context and every candidate in the whole knowledge base, which is very time-consuming especially when the size of knowledge base is substantially expanded. Instead, we propose a hierarchical filtering method to narrow down the candidate search space. Our proposed knowledge selection pipeline includes the following three modules: domain classification, entity tracking, and knowledge matching, as illustrated in Figure 1. Specifications of each module are detailed below.

2.2.1 Domain Classification

In multi-domain conversations, if the system knows what domain a given turn belongs to, the search space for knowledge selection can be greatly reduced by taking the domain-specific knowledge only. The DSTC9 Track 1 data includes the augmented turns for "Train", "Taxi", "Hotel", and "Restaurant" domains in its training set, where the first two domains have domain-level knowledge only, while the others can be further subdivided for each entity-specific knowledge. To improve the generalizability of our filtering mechanism for unseen domains, we merged the domains which require further entity-level analysis into an "Others" class and defined this task as a three-way classification: {"Train", "Taxi", and "Others"}.

We implemented a domain classifier by fine-tuning the RoBERTa-Large model which takes the whole dialogue context and outputs a domain label. Considering that a new domain (i.e., "Attraction") is introduced in the test set, we augmented the training data with 3,350 additional samples of the "Attraction" domain, which were obtained from the MultiWOZ 2.1 (Eric et al., 2020), the source of the DSTC9 Track 1 data (all augmented samples are labeled as "Others"). More specifically, we first find out those dialogues for "Attraction" in the train-

ing set of the MultiWOZ 2.1 dataset (this dataset contains seven domains including "Attraction") by selecting dialogue turns that contain "Attraction" related slots. We then replace the original "Attraction" related slots with entities of the "Attraction" domain in the knowledge base K. Meanwhile we replace the last user utterances in the dialogues with the knowledge questions that belong to the replaced new entities. Table 1 gives one example for explanation. In this example, we replace the original entity of "funky fun house" with a new entity of "California Academy of Science" randomly selected from the "Attraction" domain of the knowledge base. Besides, we replace the original last user utterance with a knowledge question randomly selected from the FAQs of this new entity "California Academy of Science".

2.2.2 Entity Tracking

Once the domain classifier predicts the 'Others' label for a given turn, the entity tracking module is executed to detect the entities mentioned in the dialogue context and align them to the entity-level candidates in the knowledge base. We adopt an unsupervised approach based on fuzzy n-gram matching whose details can be referred to Section A.2 of the Supplementary Material. After extracting these entities, we determined the character-level start position of each entity in the dialogue context and selected the last three mentioned entities as the output of this module.

2.2.3 Knowledge Matching

The knowledge matching module receives a list of knowledge candidates and ranks them in terms of relevance to the input dialogue context. We concatenated the dialogue context, domain/entity name, and each knowledge snippet into a long sequence, which is then sent to the fine-tuned RoBERTa-Large model to get a relevance score.

To train the model, we adopted Hinge loss, which was reported to perform better for the ranking problems (Wang et al., 2014; Elsayed et al., 2018) than Cross-entropy loss used in the baseline system. For each positive instance, we drew four negative samples, each of which is randomly selected from one of four sources: 1) the whole knowledge base, 2) the knowledge snippets in the ground truth domain, 3) the knowledge snippets of the ground truth entity, and 4) the knowledge snippets of other entities mentioned in the same dialogue. In the execution time, we fed the knowledge candidates filtered by the predicted domain and entity from Section 2.2.1 and 2.2.2, repectively. Then, the module outputs a list of the candidates ranked by relevance score.

2.3 Response Generation

For response generation, we compared the following three pre-trained sequence-to-sequence (seq2seq) models: T5-Base (Raffel et al., 2020), BART-Large (Lewis et al., 2020), and Pegasus-Large (Zhang et al., 2020). Each model inputs a concatenated sequence of the whole dialogue context and the knowledge answer and then outputs a response. The ground-truth knowledge answer is used in the training phase, while the top-1 candidate from the knowledge selection result is used in the test phase.

3 Experiments and Results

We used the same data split and evaluation metrics as the official DSTC9 Track 1 challenge. All model training and dataset details are summarized in the Section B of the Supplementary Material.

3.1 Knowledge Seeking Turn Detection

Table 2 compares the knowledge seeking turn detection performance between our proposed models and the best single model and ensemble-based systems from the DSTC9 Track 1 official results.[1] The results show that our proposed data augmentation method helped to improve the recall of our detection model and led to the highest F1 score among all the single models in the challenge.

3.2 Knowledge Selection

Our domain classification and entity tracking modules achieved 99.5% in accuracy and 97.5% in recall, respectively. The data augmentation method helped to improve the domain classification accuracy from 97.1% to 99.5%.

Table 3 summarizes the knowledge selection performance of our system based on the proposed hierarchical filtering mechanism using the results from both domain classification and entity tracking modules. Our proposed system outperformed the challenge baseline in all three metrics with a largely reduced execution time from more than 20 hours by the baseline to less than half an hour to process the whole test set with a single V100 GPU.

[1] There are up to five entries submitted by each team in the competition and we report only the best entries by a single model and ensemble-based systems.

Speaker	Original Dialogue	New Dialogue
User	I was hoping to see local places while in Cambridge. Some entertainment would be great.	I was hoping to see local places while in Cambridge. Some entertainment would be great.
Agent	I got 5 options. which side is okay for you?	I got 5 options. which side is okay for you?
User	It doesn't matter. Can I have the address of a good one?	It doesn't matter. Can I have the address of a good one?
Agent	How about funky fun house, they are located at 8 mercers row, mercers row industrial estate.	How about California Academy of Sciences, they are located at 8 mercers row, mercers row industrial estate.
User	Could I also get the phone number and postcode?	Is WiFi available?

Table 1: An example of data augmentation for domain classification. The left dialogue is the original dialogue from the MultiWOZ 2.1 dataset while the right one is synthesized by replacing the original entity and last user utterance highlighted by red with a new entity and knowledge question from the knowledge base highlighted by blue.

	Precision	Recall	F1
Our proposed model	0.9920	0.9344	0.9623
+ data augmentation	0.9903	0.9833	0.9868
DSTC9 Track 1 Systems:			
Baseline	0.9933	0.9021	0.9455
Team 17†	0.9933	0.9748	0.9839
Team 3‡	**0.9964**	**0.9859**	**0.9911**

Table 2: Test results on task 1: knowledge-seeking turn detection. † and ‡ denote the best DSTC9 Track 1 systems with a single model and model ensemble, respectively. Overall highest scores are made bold while highest scores for single models are underlined.

	MRR@5	Recall@1	Recall@5
Our proposed model	0.9461	**0.9251**	0.9702
DSTC9 Track 1 Systems:			
Baseline	0.7263	0.6201	0.8772
Team 7†	0.9309	0.8988	0.9666
Team 19‡	**0.9504**	0.9235	**0.9840**

Table 3: Test results on task 2: knowledge selection.

Compared with the best knowledge selection results from the challenge, our model achieved higher performances than the best single model-based system in all metrics, and even surpassed the best ensemble model in recall@1. To be noted, recall@1 is the most important metric, since the response generation is grounded on only the top-1 result from knowledge selection.

3.2.1 Ablation Study

First of all, Table 5 summarizes the ablation results by imposing two kinds of changes based on our full knowledge matching model: instead of concatenating the dialogue context, domain name, entity name, and knowledge question and answer pair as the input to the model, we only concatenate the dialogue context and knowledge question and answer pair (w/o entity names); we replace the Hinge loss with Cross-entropy loss (w/o Hinge Loss). To be noted, we should pay more attention to the Recall@1 score in the Table 5, which is the most important metric. And we can see that adding the domain and entity names are beneficial and the use of Hinge loss for optimization is better than Cross-entropy for this ranking problem.

As above-mentioned, for training the knowledge matching module, we need to sample several negative samples for each position sample and instead of using only one negative sampling strategy, we used a mixed strategy. More specifically, for sampling each negative sample, we randomly adopted one of the following four strategies:

1. Randomly select from all knowledge snippets;
2. Randomly select from the knowledge snippets of entities that are the in the same domain as the ground truth one (i.e., the entity of the positive sample);
3. Randomly select from the knowledge snippets of the ground truth entity;
4. Randomly select from the knowledge snippets of entities that are mentioned in the same dialogue as the ground truth one.

Each strategy $i \in \{1, 2, 3, 4\}$ is sampled at a certain sampling ratio p^i_{ns}. We tuned this sampling ratio by trying several combinations, and the results are summarized in Table 6. From it, we can see that: (1) Strategy 4 is the most effective among all four ones; (2) Mixing four strategies is better than using only one of them; (3) Allocating higher ratio to strategy 4 is better than uniform ratios for every strategy.

3.3 Response Generation

Table 4 summarizes the automated evaluation results for the generated responses with different seq2seq models. Our fine-tuned T5-Base model achieved lower BLEU scores than BART-Large and Pegasus-Large, while its METEOR score is

	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-1	ROUGE-2	ROUGE-L
Our Systems:								
BART-Large	0.3743	0.2428	0.1620	0.1098	0.3869	0.4163	0.1992	0.3639
T5-Base	0.3575	0.2432	0.1685	0.1155	**0.4379**	0.4139	0.2103	0.3536
Pegasus-Large	0.3808	0.2531	0.1727	**0.1192**	0.4013	0.4237	0.2099	0.3656
DSTC9 Track 1 Systems:								
Baseline	0.3031	0.1732	0.1005	0.0655	0.2983	0.3386	0.1364	0.3039
Team 15†	0.3779	0.2532	**0.1731**	0.1175	0.3931	0.4204	0.2113	0.3765
Team 3‡	**0.3864**	**0.2539**	0.1692	0.1190	0.3914	**0.4332**	**0.2115**	**0.3885**

Table 4: Test results on task 3: knowledge grounded response generation.

Settings	MRR@5	Recall@1	Recall@5
Original model	**0.9811**	**0.9693**	**0.9936**
w/o entity names	0.9788	0.9656	0.9933
w/o Hinge Loss	0.9734	0.9613	0.9905

Table 5: Ablation study of the knowledge matching module for knowledge selection by removing entities and hinge loss. Scores are reported on the validation set.

Sampling ratios	**MRR@5**	**Recall@1**	**Recall@5**
Original model			
[0.1,0.1,0.1,0.7]	**0.9811**	**0.9693**	**0.9936**
[0.25,0.25,0.25,0.25]	0.9761	0.9615	0.9929
[1.0,0.0,0.0,0.0]	0.9712	0.9514	0.9933
[0.0,1.0,0.0,0.0]	0.9559	0.9248	0.9906
[0.0,0.0,1.0,0.0]	0.9728	0.9540	0.9933
[0.0,0.0,0.0,1.0]	0.9751	0.9596	0.9929

Table 6: Ablation study of the knowledge matching module for knowledge selection by tuning the mixed negative sampling ratio. Scores are reported on the validation set. The sampling ratio is represented in the format of $[p_{ns}^1, p_{ns}^2, p_{ns}^3, p_{ns}^4]$.

substantially higher than the others. Note that our generation system does not perform any model ensemble, and it surpasses the best single system in the DSTC9 Track 1 for half of the metrics.

Following the official evaluation protocol in the challenge, we performed human evaluation to compare our system with the top systems from the challenge[2], as shown in Table 7. Specifically, we hired three crowd-workers for each instance, asked them to score each system output in terms of its "accuracy" and "appropriateness" in five point Likert scale, and reported the averaged scores. We have three findings: (1) T5 achieves higher accuracy, while Pegasus is slightly better for appropriateness; (2) our systems generates more accurate responses than the top DSTC9 systems, while the appropriateness scores is comparable (confirmed by significance testing in Section C.2 of Supplementary Material); (3) the final average scores of our systems rank the highest. We present several examples of the generated responses by our system compared against the baseline and top 2 systems in Section C.3 of Supplementary Material.

Systems	Accuracy	Appropriateness	Average
Our Systems:			
T5-Base	**4.5994***	4.4572†	**4.5283***
Pegasus-Large	4.5451†	4.4591†	4.5021†
DSTC9 Track 1 Systems (Top-2):			
Team 19	4.4979	**4.4698**	4.4838
	(4.3917)	(4.3922)	(4.3920)
Team 3	4.4524	4.4064	4.4294
	(4.3480)	(4.3634)	(4.3557)

Table 7: Human evaluation results of the test set for response generation. Numbers within the parentheses are official scores from DSCT9 (Kim et al., 2020b). The symbol * means our score is significantly higher than the best previous system while † means our score is not significantly different from the best previous system, according to paired t-test with $p < 0.05$.

4 Conclusions

In this work, we propose a comprehensive system to enable the task-orientated dialogue models to answer user queries that are out of the scope of APIs. We significantly improved the system's capability of finding the most relevant knowledge snippets, consequently providing excellent responses by introducing a novel data augmentation method, incorporating domain and entity identification modules for knowledge selection, and utilizing mixed negative sampling. To demonstrate the efficacy of our approach, we benchmark our system on the DSTC9 Track 1 challenge dataset and report the state-of-the-art performance.

[2]https://github.com/alexa/alexa-with-dstc9-track1-dataset/tree/master/results

References

Gamaleldin Elsayed, Dilip Krishnan, Hossein Mobahi, Kevin Regan, and Samy Bengio. 2018. Large margin deep networks for classification. In *Advances in Neural Information Processing Systems*, volume 31, pages 842–852. Curran Associates, Inc.

Mihail Eric, Rahul Goel, Shachi Paul, Abhishek Sethi, Sanchit Agarwal, Shuyang Gao, Adarsh Kumar, Anuj Goyal, Peter Ku, and Dilek Hakkani-Tur. 2020. MultiWOZ 2.1: A consolidated multi-domain dialogue dataset with state corrections and state tracking baselines. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 422–428, Marseille, France. European Language Resources Association.

Mihail Eric, Lakshmi Krishnan, Francois Charette, and Christopher D. Manning. 2017. Key-value retrieval networks for task-oriented dialogue. In *Proceedings of the SIGDIAL 2017 Conference*, pages 37–49.

Huang He, Hua Lu, Siqi Bao, Fan Wang, Hua Wu, Zhengyu Niu, and Haifeng Wang. 2021. Learning to select external knowledge with multi-scale negative sampling. *arXiv preprint arXiv:2102.02096*.

Seokhwan Kim, Mihail Eric, Karthik Gopalakrishnan, Behnam Hedayatnia, Yang Liu, and Dilek Hakkani-Tur. 2020a. Beyond domain APIs: Task-oriented conversational modeling with unstructured knowledge access. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 278–289, 1st virtual meeting. Association for Computational Linguistics.

Seokhwan Kim, Mihail Eric, Karthik Gopalakrishnan, Behnam Hedayatnia, Yang Liu, and Dilek Hakkani-Tur. 2020b. Beyond domain apis: Task-oriented conversational modeling with unstructured knowledge access. *arXiv preprint arXiv:2006.03533*.

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.

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. *CoRR*, abs/1907.11692.

Haitao Mi, Qiyu Ren, Yinpei Dai, Yifan He, Jian Sun, Yongbin Li, Jing Zheng, and Peng Xu. 2021. Towards generalized models for beyond domain api task-oriented dialogue. *AAAI-21 DSTC9 Workshop*.

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

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Liang Tang, Qinghua Shang, Kaokao Lv, Zixi Fu, Shijiang Zhang, Chuanming Huang, and Zhuo Zhang. 2021. Radge relevance learning and generation evaluating method for task-oriented conversational system-anonymous version. *AAAI-21 DSTC9 Workshop*.

Jiang Wang, Yang Song, Thomas Leung, Chuck Rosenberg, Jingbin Wang, James Philbin, Bo Chen, and Ying Wu. 2014. Learning fine-grained image similarity with deep ranking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1386–1393.

Jason D. Williams, Kavosh Asadi, and Geoffrey Zweig. 2017. Hybrid code networks: practical and efficient end-to-end dialog control with supervised and reinforcement learning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL 2017)*.

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 von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 11328–11339. PMLR.

A Methods

A.1 Entity Extraction

Specifically, we first normalize the entity names in the knowledge base using a set of heuristic rules, such as replacing the punctuation "&" with "and". Table A.1 summarizes the full list of normalization rules and we give an example for each rule as illustration. Then we perform the fuzzy n-gram matching between an entity and a certain piece of dialogue context. For example of an entity of "Alexander Bed and Breakfast", it is a four-gram, therefore we extract all four-grams from the dialogue context and match each of them against it. And the process of matching is to first find out the longest contiguous matching sub-sequence and then calculate the matching ratio by the equation of $2M/T$, where M is the length of the matched sub-sequence while T is the total length of the two n-grams to be matched.[3] If this ratio is higher than 0.95, we deem this pair of n-grams as matched. In this way, we can find out which entities in the knowledge base are mentioned in a certain dialogue.

B Experiments

B.1 Data Samples & Statistics

Table B.2 shows an example conversation with unstructured knowledge access. The user utterance at turn $t = 5$ requests the information about the gym facility, which is out of the coverage of the structured domain APIs. However, the relevant knowledge contents can be found from the external sources as in the rightmost column which includes the sampled QA snippets from the FAQ lists for each corresponding entity within domains such as train, hotel, or restaurant. With access to these unstructured external knowledge sources, the agent manages to continue the conversation with no friction by selecting the most appropriate knowledge.

The data statistics are summarized in Table B.3.[4] The main data is an augmented version of MultiWOZ 2.1 that includes newly introduced knowledge-seeking turns in the MultiWOZ conversations. A total of 22,834 utterance pairs were newly collected based on 2,900 knowledge candidates from the FAQ webpages about the domains and the entities in MultiWOZ databases. To be noted, for the test set, other conversations collected from scratch about touristic information for San Francisco are added. To evaluate the generalizability of models, the new conversations cover knowledge, locale and domains that are unseen from the train and validation data sets. In addition, this test set includes not only written conversations, but also spoken dialogues to evaluate system performance across different modalities.

Table B.4 gives the statistics of the knowledge base, which is a collection of frequently asked questions (FAQs). To be noted, there are no entities for the "Train" and "Taxi" domains while for "Hotel", "Restaurant", and "Attraction" domains, each entity has its corresponding list of FAQ pairs. Besides, the knowledge base for the test set covers the train & validation sets and is further expanded by adding one more domain of "Attraction" and more entities.

B.2 Experimental Details

We implemented our proposed system based on the DSTC9 Track 1 baseline provided by Kim et al. (2020b) and the transformers library (Wolf et al., 2020). For all sub-tasks, the maximum sequence length for the dialogue context and the knowledge snippet is both 128. For the knowledge seeking turn detection sub-task, the model is fine-tuned for 5 epochs with the batch size of 16, while for other sub-tasks, 8 epochs and the batch size of 4 are used. A model checkpoint is saved after each epoch, and the best checkpoint is picked based on the validation results. For decoding process of the response generation model, we replaced the nucleus sampling in the baseline to beam search (beam width is 5), which achieved higher performances in the validation set.

C Results

C.1 Significance Testing for Human Evaluation

Since those scores of human evaluation for response generation are quite close to each other, we resort to significance testing to confirm our system's superior performance. Table C.5 summarizes the significance testing p-value between our systems and the top-2 submitted systems in the DSTC9 challenge for the accuracy, appropriateness, and average scores, respectively. From it, we can see that T5-Base is significantly higher than the competing systems in terms of accuracy ($p < 0.05$). Besides,

[3] https://towardsdatascience.com/sequencematcher-in-python-6b1e6f3915fc

[4] Data can be downloaded from: https://github.com/alexa/alexa-with-dstc9-track1-dataset

Normalization rules	Examples
Replace the punctuation "&" with "and"	Bay Subs & Deli → Bay Subs and Deli
If the entity contains any symbol of " - ", ", " or "/", split this entity by this symbol and remove the second part	Hard Knox Cafe - Potrero Hill → Hard Knox Cafe
Replace "guesthouse" with "guest house"	ARBURY LODGE GUESTHOUSE → ARBURY LODGE GUEST HOUSE
If the entity contains a place name such as "Fisherman's Wharf" and "San Francisco" in the end, remove it (since the entities in the knowledge base do not contain these place names)	Bay Bridge Inn San Francisco → Bay Bridge Inn
Replace numbers with corresponding English words	Pho Huynh Hiep 2 → Pho Huynh Hiep Two

Table A.1: Entity normalization rules and the corresponding examples.

Turn	Speaker	Utterance	Sampled Knowledge Snippets from FAQs
1	User	I'm looking for a place to stay in the south of town. It doesn't need to have free parking.	
2	Agent	There are 4 hotels that are in the area you are looking for. Would you prefer a 3 or 4 star rated hotel?	
3	User	I don't care about the star rating as long as it's expensive.	
4	Agent	The ***Lensfield Hotel*** is the only expensive hotel in the south area. Would you like any more information on this location?	
5	User	I'm interested in knowing, do they have a workout facility on the premises?	
6	Agent	There are both a fitness center and gym available on the premises. Does this sound ok?	Q1: Do you have room service for your guests? A1: Yes, the Lensfield Hotel provides room services. **Q2: Is there a gym available at your location?** **A2: There is both a fitness center and gym available on the premises.** Q3: Can I bring my dog? A3: Pets are not allowed at the Lensfield Hotel.
7	User	That is perfect can you book that for me please.	
8	Agent	The Lensfield Hotel is located in the South. It has a 3 star rating and is expensive. There is free parking and internet. I have booked it for you.	
9	User	Great, thank you!	

Table B.2: Examples of task-oriented conversations with unstructured knowledge access. Three sampled FAQ pairs for the entity "Lensfield Hotel" are listed in the rightmost column for turn 5 which is beyond the coverage of API and needs external knowledge support. The most appropriate FAQ pair to address turn 5 is highlighted in bold font.

Split	Source	# dialogues	# samples	# knowledge seeking turns
Train	MultiWOZ	7,190	71,348	19,184
Valid	MultiWOZ	1,000	9,663	2,673
Test	MultiWOZ	977	2,084	977
	SF Written	900	1,834	900
	SF Spoken	107	263	104

Table B.3: Statistics of the data divided into training, validation, and test sets. The test set contains three sources of samples: MultiWOZ, San Francisco tourism in written English, and San Francisco tourism in spoken English, which is different from train and validation sets.

Domain	Train & Val		Test	
	# Entities	# Snippets	# Entities	# Snippets
Train	–	26	–	26
Taxi	–	5	–	5
Hotel	33	1,219	178	4,346
Restaurant	110	1,650	391	7,155
Attraction	–	–	97	507
Total	143	2,900	666	12,309

Table B.4: Statistics of the knowledge base (the list of FAQs). "Train" and "Taxi" domains do not have any entities and there is no "Attraction" domain for the knowledge base in train and validation sets.

	Accuracy		Appropriateness		Average	
	T5-Base	Pegasus-Large	T5-Base	Pegasus-Large	T5-Base	Pegasus-Large
Team 19	1.86×10^{-6}	1.45×10^{-1}	6.07×10^{-1}	5.56×10^{-1}	1.96×10^{-3}	4.25×10^{-1}
Team 3	1.21×10^{-16}	8.01×10^{-7}	2.89×10^{-4}	5.68×10^{-5}	8.31×10^{-15}	2.48×10^{-8}

Table C.5: Significance Testing p-values between our systems and top-2 systems in the DSTC9 competition for the accuracy, appropriateness, and average score of response generation. Paired t-test is used.

	Example
Dialogues	**User:** I need a train to Broxbourne, that arrives by 20:30. **Agent:** Certainly. Can I assume you're leaving from Cambridge? **User:** Yes, that is correct. **Agent:** There is a train leaving at 5:01, would you like a seat? **User:** Is that leaving on Sunday? **Agent:** I have the TR8991 leaving at 19:01 and arriving by 20:01. Would you like me to book that for you? **User:** Not yet but I also need a restaurant that serve mediterranean food and should be in the centre of town. **Agent:** I have a few restaurants that fit your criteria. The Gardenia, La Mimosa, and Shiraz Restaurant. Which one are you interested in hearing about? **User:** Tell me about The Gardenia. **Agent:** It's a cheap restaurant located at 2 Rose Crescent City Centre. Phone number and postcode are 01223356354 and cb23ll. Do you need anything else? **User:** Does the restaurant offer live music on the weekend? **Agent:** Sorry, there is no live music at The Gardenia. What else can I help you with? **User:** Does the train have a data limit for wifi usage?
Knowledge Answer	Data is limited to 50MB per day with no option of additional data.
Our systems:	
T5-Base	Data is limited to 50MB per day with no option of additional data. Is there anything else I can do for you today or would you like to make a reservation?
Pegasus-Large	Data is limited to 50MB per day with no option of additional data. Is there anything else I can help you with?
Top-2 submitted systems:	
Team 3	No, there is no additional data available to purchase. Anything else I can do for you?
Team 19	No, the train does not have a data limit for wifi usage. Anything else I can do for you?

Table C.6: Qualitative comparison between our system with previous strong competitors. Knowledge answer is the answer part of the ground truth knowledge snippet. We are comparing against the top-2 systems submitted to the DSTC9 competition.

T5-Base and Pegasus-Large are comparable to the best previous system in terms of appropriateness. Finally, with regards to the average score, our T5-Base significantly rivals the previous best system.

C.2 Qualitative Examples of Responses

Table C.6 gives one qualitative example to compare our system's responses against those of the top-2 submitted systems in the DSTC9 competition (i.e., Team 3 and 19)[5]. Overall, we can see that our system's responses are more accurate. Taking the example in Table C.6, our responses can exactly answer the user query and it is strictly aligning with the ground truth knowledge, while the response from Team 19 is totally wrong and that from Team 3 does not address the user query at all.

[5]https://github.com/alexa/alexa-with-dstc9-track1-dataset/tree/master/results