DST

DSTC

Background

The DSTC shared tasks have provided common testbeds for the dialog research community since 2013.

From its sixth edition, it has been rebranded as "Dialog System Technology Challenge" to cover a wider variety of dialog related problems.

DSTC 9

DSTC 9

- Beyond Domain APIs: Task-oriented Conversational Modeling with Unstructured Knowledge Access (Amazon Alexa AI): This track aims to allow users to have requests that are out of the scope of APIs/DB but potentially available in external knowledge sources. Track participants will develop task-oriented dialogue systems to understand relevant domain knowledge, and generate system responses with the relevant selected knowledge. In addition, the track includes evaluation on generalization over unseen domains and modalities (i.e. moving from written to spoken conversations).

- Multi-domain Task-oriented Dialog Challenge II (Microsoft Research AI & Tsinghua University): This track follows its success in DSTC-8 continuing with the effort of building dialog systems under a multi-domain setting. This time extending the task by incorporating new datasets, creating new sub-tasks, and providing a new development platform. The new task specifically focuses on two aspects of dialog systems: language portability and end-to-end system complexity.

DSTC 9. Track 1

DSTC 9. Beyond Domain APIs: Task-oriented Conversational Modeling with Unstructured Knowledge Access

Beyond Domain APIs: Task-oriented Conversational Modeling with Unstructured Knowledge Access

Seokhwan Kim, Mihail Eric, Behnam Hedayatnia,
Karthik Gopalakrishnan, Yang Liu, Dilek Hakkani-tur
Amazon Alexa AI
1120 Enterprise Way, Sunnyvale, CA 94089, USA
{seokhwk,behnam,mihaeric,karthgop,yangliud,hakkanit}@amazon.com

Motivation

Traditionally, task-oriented dialogue systems have focused on providing information and performing actions according to users' requests which can be handled only by given databases or APIs. However, in addition to task-focused requests, users also have needs that require more than what is provided by backend resources. For example, while most virtual assistants can help users to book a hotel, a restaurant or movie tickets, they fall short of answering potential follow-up questions users may have, for example: how to park vehicles; whether they are allowed to bring pets or children to the reserved place; or what the cancellation policy is. No API/DB entry is usually available to handle such requests. On the other hand, relevant domain knowledge is already available on web pages in the form of descriptions, FAQs and customer reviews for many of these out-of-scope scenarios. Since current dialogue systems don't incorporate these external knowledge sources into task-oriented conversational modeling, users need to visit the websites by themselves to find out any additional information beyond API/DB coverage, making conversational interactions inefficient.

In this challenge, we propose the 'Beyond Domain APIs: Task-oriented Conversational Modeling with Unstructured Knowledge Access' track. This track aims to support frictionless task-oriented scenarios, where the flow of the conversation does not break when users have requests that are out of the scope of APIs/DB but potentially are already available in external knowledge sources. Inspired by recent studies on knowledge-grounded conversational modeling (Moghe et al. 2018; Zhou, Prabhumoye, and Black 2018; Dinan et al. 2018; Galley et al. 2019; Gopalakrishnan et al. 2019), track participants will develop dialogue systems to understand relevant domain knowledge, and generate system responses with the relevant selected knowledge. Different from previous work on social conversations, our proposed challenge track addresses task-oriented conversations grounded on fine-grained domain-/entity-level knowledge related to given dialogue contexts. In addition, the track includes evaluation of participant model submissions on generalization to new, unseen domains in the training dataset, as well as generalization to unseen modalities (i.e., moving from written to spoken conversations).

Proposed Tasks

This challenge track is designed based on a simple baseline architecture which decouples between turns that could be handled by the existing task-oriented conversational models with no extra knowledge and turns that require external knowledge resources to be answered by the dialogue system. We focus only on the turns that require knowledge access as the evaluation target in this track (Figure 1). More specifically, our proposed challenge track will include the following three sub tasks: 1) *Knowledge-seeking Turn Detection*, 2) *Knowledge Selection*, and 3) *Knowledge-grounded Response Generation* (Table 1).

Participants will develop systems to generate responses for the branch with knowledge access. They can leverage the annotations and the ground-

DSTC 9. Beyond Domain APIs: Task-oriented Conversational Modeling with Unstructured Knowledge Access

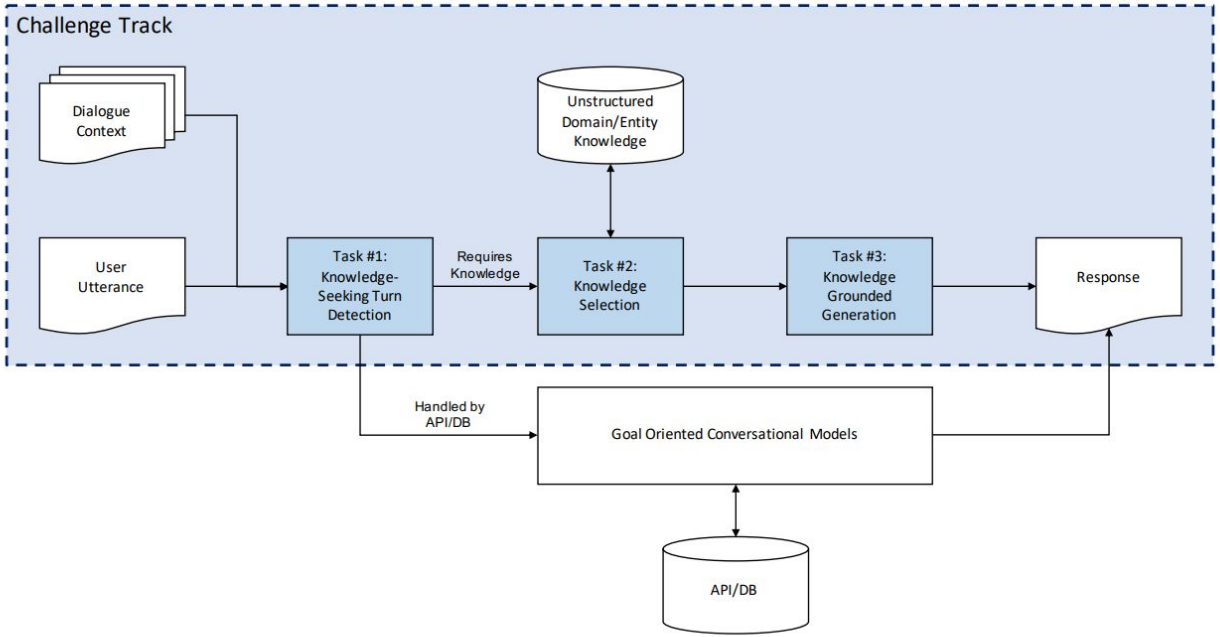


Figure 1: An overview of the challenge track that focuses on the knowledge access branch in the shaded box

DSTC 9. Beyond Domain APIs: Task-oriented Conversational Modeling with Unstructured Knowledge Access

Task #1	**Knowledge-seeking Turn Detection**
Goal	To decide whether to continue the existing scenario or trigger the knowledge access branch for a given utterance and dialogue history
Input	Current user utterance, Dialogue context, and API/Knowledge sources
Output	Binary class (requires knowledge access or not)
Metrics	Accuracy, Precision/Recall/F-measure
Task #2	**Knowledge Selection**
Goal	To select proper knowledge sources from the domain knowledge-base given a dialogue state at each turn with knowledge access
Input	Current user utterance, Dialogue context, and the entire set of domain-/entity-level knowledge snippets
Output	Ranking of top-k knowledge candidates
Metrics	R@k, MRR, MAP, NDCG, Avg rank
Task #3	**Knowledge-grounded Response Generation**
Goal	To take a triple of input utterance, dialog context, and the selected knowledge sources and generate a system response either in an extractive or an abstractive method
Input	Current user utterance, Dialogue context, and Selected knowledge sources
Output	Generated system response
Metrics	*Automated evaluation*: Perplexity, F1, BLEU, ROUGE, METEOR, CIDEr *Human evaluation*: grammatical/semantical correctness, naturalness, appropriateness, informativeness and relevance to given knowledge

Table 1: Summary of the proposed tasks

Task/ Locale	Data set	Number of dialogues	Number of original turns	Number of augmented turns	Total number of turns	Total number of knowledge snippets
MultiWoz	Train	8,438	56,778	~ 20k	~75k	~1k
	Valid	1,000	7,374	~2.5k	~10k	
	Test (seen)	1,000	7,372	~2.5k	~10k	
TBD	Test (unseen)	~1k	-	-	~10k	TBD

Table 2: Statistics of the data sets

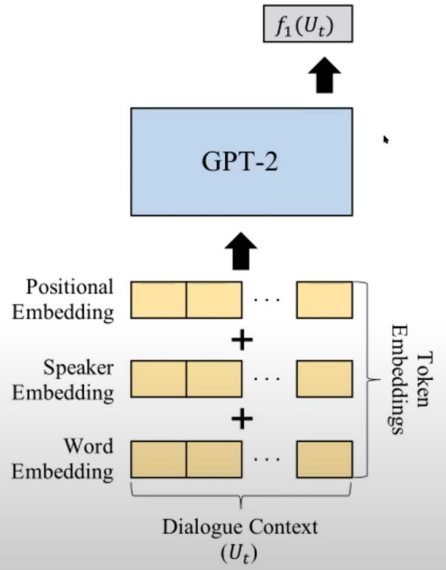
DSTC 9. Beyond Domain APIs: Task-oriented Conversational Modeling with Unstructured Knowledge Access

- Accuracy
  - Your task is to use the reference knowledge to score how accurate each agent's response is on a scale of 1 - 5. A score of 1 means that the response is completely wrong, and a score of 5 means that the response is completely accurate.
- Appropriateness
  - Appropriateness means how well the response is naturally connected to the conversation. A score of 1 means that the response is very inappropriate and it is *not* naturally connected to the conversation. A score of 5 means that the response is very appropriate, and it is very naturally connected to the conversation.
- Asked 3 turkers to score each of up to 14 responses
  - up to 12 finalists + Baseline + GT

DSTC 9. Beyond Domain APIs: Task-oriented Conversational Modeling with Unstructured Knowledge Access

Baseline: Knowledge-Seeking Turn Detection

- Neural Utterance Classification
  - Supervised binary classifier
  - Fine-tuned GPT-2
    - Taking a given dialogue context only
    - With binary cross-entropy loss

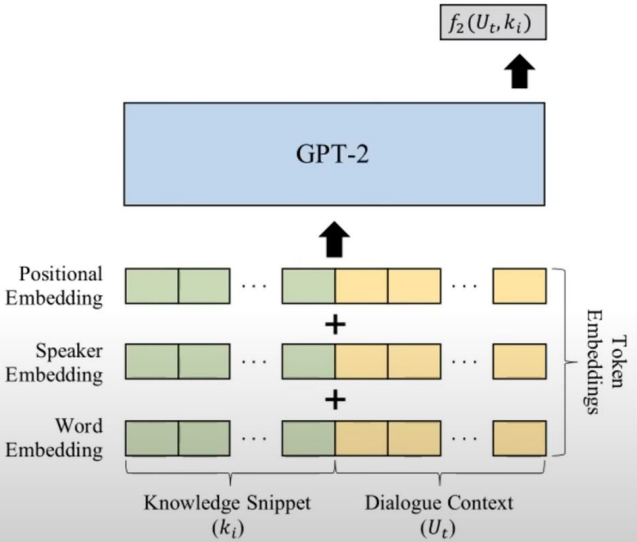


DSTC 9. Beyond Domain APIs: Task-oriented Conversational Modeling with Unstructured Knowledge Access

Baseline: Knowledge Selection

- Neural Relevance Classification
  - Supervised binary classifier
  - Fine-tuned GPT-2
    - Taking a given dialogue context and each knowledge candidate as a single concatenated sequence
    - With binary cross-entropy loss
    - 5 negative samples for each instance

Shows relevance of given pair of knowledge candidate and dialogue context

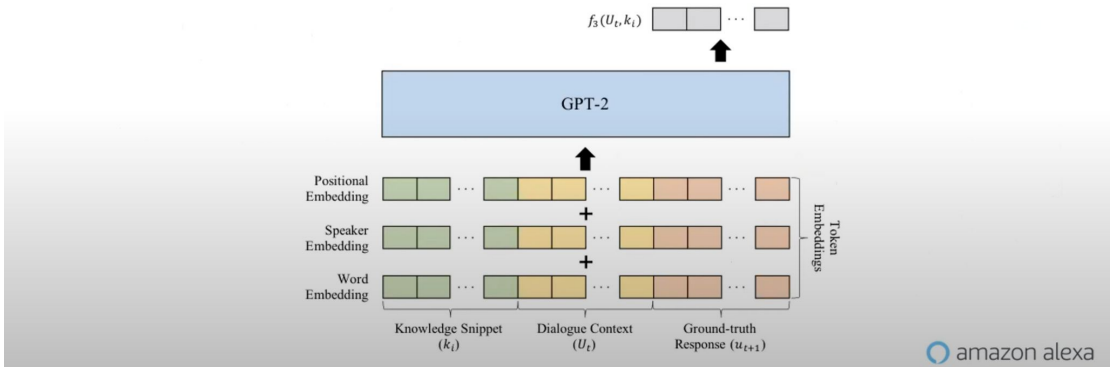


5 randomly selected negative samples per 1 target knowledge candidate

DSTC 9. Beyond Domain APIs: Task-oriented Conversational Modeling with Unstructured Knowledge Access

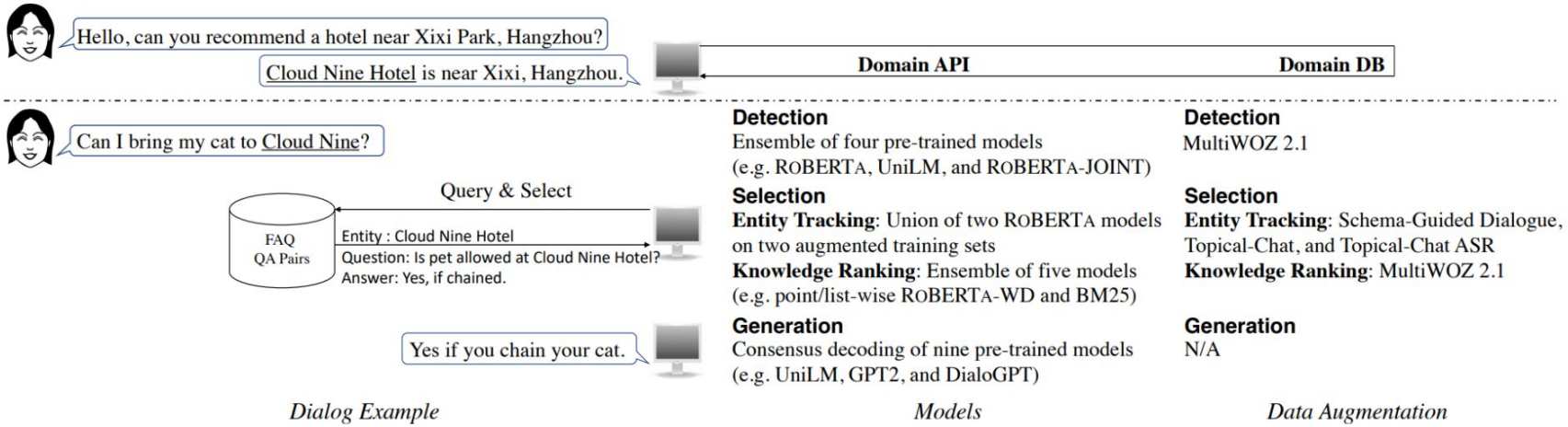
Baseline:
Knowledge Grounded Response Generation

- Neural Response Generation
  - Fine-tuned GPT-2
    - With ground-truth knowledge snippets concatenated



context + knowledge snippet + ground truth response

DSTC 9. Towards Generalized Models for Beyond Domain API Task-oriented Dialogue.



- $\langle user \rangle$: start of a user utterance;
- $\langle sys \rangle$: start of a system utterance;
- $\langle kng \rangle$: start of a knowledge;
- $\langle kng_k \rangle$: start of the k-th best knowledge;
- $\langle ent \rangle$: start of an entity or a domain name;
- $\langle ans \rangle$: start of a knowledge answer;
- $\langle resp \rangle$: start of a response.

2nd place

DSTC 9. Towards Generalized Models for Beyond Domain API Task-oriented Dialogue.

Task 1

- Data representation

Context	Target
$\langle user \rangle U_1 \ \ldots \ \langle sys \rangle S_{i-1} \langle user \rangle U_i$	True/False

- Models
  - RoBERTa (Liu et al., 2019)
  - RoBERTa-Hidden States
  - UniLM (Dong et al., 2019)

- Data representation

Context		Target
Sentent 1	Sentent 2	
$\langle user \rangle U_1 \ \ldots \ \langle sys \rangle S_{i-1} \langle user \rangle U_i$	$\langle kng \rangle q_i^1 \langle ans \rangle a_i^1$	1
		0
	⋮	0
		0
	$\langle kng \rangle q_i^5 \langle ans \rangle a_i^5$	0
	False	0

- Model
  - RoBERTa-Joint, a similarity-based model

However, the above models might not be well extended to unseen domains because they only look at dialogue utterances, and the prediction becomes harder on utterances from unseen domains. This leads us to propose a similarity based ROBERTA-JOINT model which jointly learns detection and selection tasks and directly models the similarity between the dialogue history and knowledge. ROBERTA JOINT takes a dialog history and a top-n knowledge list as inputs and calculates their similarity.

2nd place

DSTC 9. Towards Generalized Models for Beyond Domain API Task-oriented Dialogue.

Data Augmentation Since the training data only covers a portion of entities and knowledge in the training $\mathbf{K}$, we augment the training samples through revising MultiWOZ 2.1 (Eric et al. 2019) data as follows:

- *Entity Substitution:* In order to reduce the distributional bias of entities, we leverage the MultiWOZ labels in terms of span information, and replace those entities with new ones in the utterances so as to ensure all entities appear evenly in the augmented data.
- *Knowledge Insertion:* Given the training $\mathbf{K}$ and an entity e_i mentioned in the current dialogue, we randomly choose one knowledge k_j, whose e_j equals to e_i, and insert the question part q_j as a new knowledge seeking turn.
- *Sentence Rewriting.* We also rewrite knowledge questions with an off-the-shelf paraphrase toolkit[1].

2nd place

DSTC 9. Towards Generalized Models for Beyond Domain API Task-oriented Dialogue.

Task 2

Sub-task	Context		Target	Type
Entity Tracking	Sentence 1	Sentence 2	True/False	Binary
	$\langle user \rangle U_1 \ldots \langle sys \rangle S_{i-1} \langle user \rangle U_i$	$\langle ent \rangle e_j$	True	
Knowledge Ranking	Sentence 1	Sentence 2	True/False	Binary
	$\langle user \rangle U_1 \ldots \langle sys \rangle S_{i-1} \langle user \rangle U_i$	$\langle kng \rangle q_j \langle ans \rangle a_j$	True	
	Sentence 1	Sentence 2	One hot	Multi-class
	$\langle user \rangle U_1 \ldots \langle sys \rangle S_{i-1} \langle user \rangle U_i$	$\langle kng \rangle q_j^1 \langle ans \rangle a_j^1, \ldots, \langle kng \rangle q_j^5 \langle ans \rangle a_j^5$	[0, 1, 0, 0, 0]	

Table 1: The data representations for entity tracking and knowledge ranking tasks. In the knowledge ranking task, we first use a point-wise ROBERTA to select top-5 related knowledge (rows 4 and 5), then, we develop a list-wise ROBERTA (the last two rows) to rank top-5 knowledge again, $\langle kng \rangle q_j^1 \langle ans \rangle a_j^1, \ldots, \langle kng \rangle q_j^5 \langle ans \rangle a_j^5$ means a batch of top-5 knowledge, and the objective function is to minimize the cross-entropy loss between the true distribution and the system prediction of five classes.

For the list-wise model, we first run a point-wise model, and get top-5 knowledge for each example. Then we apply list-wise models on those top-5 results (see the last two rows in Table 1 for an example). Our ensemble re-ranks all knowledge by using the sum of probabilities of each knowledge from all single systems.

2nd place

DSTC 9. Towards Generalized Models for Beyond Domain API Task-oriented Dialogue.

Task 3

Type	Context			Target	
All	History	Top-5 Knowledge	Last Turn	Response	
	$\langle user \rangle U_1 \ldots \langle sys \rangle S_{i-1}$	$\langle kng_5 \rangle \langle ent \rangle e_i^5 \langle ans \rangle a_i^5 \ldots \langle kng_1 \rangle \langle ent \rangle e_i^1 \langle ans \rangle a_i^1$	$\langle user \rangle U_i$	$\langle resp \rangle S_i$	
Each	History	nth-best Knowledge	Last Turn	True/False	Response
	$\langle user \rangle U_1 \ldots \langle sys \rangle S_{i-1}$	$\langle kng_1 \rangle \langle ent \rangle e_i^1 \langle ans \rangle a_i^1$	$\langle user \rangle U_i$	$\langle resp \rangle$ True	# S_i
	$\langle user \rangle U_1 \ldots \langle sys \rangle S_{i-1}$	$\langle kng_2 \rangle \langle ent \rangle e_i^2 \langle ans \rangle a_i^2$	$\langle user \rangle U_i$	$\langle resp \rangle$ False	# Null

Table 2: The data representations for generation task. *All* means all top-5 knowledge are given, while *Each* only gives a knowledge each time, and our generation models generate response for each of them independently, and pick the best 'True' response based on the combing score of selection and generation models.

decoding model (log-linear)

$$\underset{S'}{\text{argmax}} \left\{ \sum_i \psi_i(\mathbf{S}, S') w_i + r(S') w_r + \sum_j t_j(S') w_{t_j} \right\}. \quad (2)$$

where S' is one system response from the ensemble of all n-best responses of all single systems, S is a set of 1-best responses of each single system. ψ(S, S') is a function that measures the similarity between S' and S. wi is a weight of ψ(S, S'). Those similarity functions include 9 metrics. We also add features related to each response S 0 : reciprocal rank r(S') in a n-best list, and some system type indicators t(S')

2nd place

DSTC 9. Towards Generalized Models for Beyond Domain API Task-oriented Dialogue.

Type	System	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-1	ROUGE-2	ROUGE-L
Single	UniLM Large WS (A)	0.4222	**0.2891**	**0.1952**	**0.1426**	0.4291	0.4616	**0.2374**	**0.4134**
	UniLM Uncased (A*)	0.4167	0.2818	0.1923	0.1410	0.4252	0.4548	0.2319	0.4087
	UniLM Large WS (E)	**0.4274**	0.2872	0.1918	0.1380	0.4271	**0.4644**	0.2308	0.4129
	GPT2 pretune (E)	0.4227	0.2828	0.1872	0.1367	**0.4338**	0.4587	0.2274	0.4129
	GPT2 (A)	0.4220	0.2817	0.1829	0.1357	0.4268	0.4558	0.2254	0.4074
	DialoGPT (A)	0.4129	0.2744	0.1841	0.1339	0.4201	0.4468	0.2205	0.3985
Ensemble	10 basic feats	**0.4442**	0.3049	0.2088	0.1566	0.4508	0.4779	0.2474	0.4289
	+ 7 sys type feats	0.4430	**0.3061**	**0.2133**	**0.1616**	**0.4535**	**0.4795**	**0.2520**	**0.4304**

Table 8: Validation results of top six single systems (ranked by BLEU-4) and consensus decoding with different feature sets. (A) shows *All* top-5 knowledge in context, while (E) only gives *Each* knowledge in context. (A*) uses top-2 knowledge. 10 basic features include 9 similarity features, and 1 reciprocal rank feature. For the system type features, each system has its own type except for GPT2 and DialoGPT models (GPT2 (A) and DialoGPT (A) share a same system type, same for GPT2 (E) and DialoGPT (E)). Thus, we have 7 additional system type features in total.

2nd place

DSTC 9. Learning to Select External Knowledge with Multi-Scale Negative Sampling (1st place)

Figure 1: Some schema descriptions from MultiWOZ 2.2.

To determine whether to seek external knowledge or not, the most straightforward way is to rely on the dialogue context.

In practice, the system is supposed to know the functions of APIs and external knowledge before making a choice between them. For the sake of a comprehensive decision, we take the dialogue context as well as the API/knowledge functions into consideration.

The schema descriptions are denoted as S = {s1, s2, · · · ,si}, where si is one slot/intent description from MultiWOZ 2.2. The external knowledge snippets are represented as K = {k1, k2, · · · , kn}, where ki represent the i-th knowledge snippet. The dialogue context is referred as Ct = {u1, u2, · · · , ut}, where ui is the i-th utterance in a multiturn conversation and t is the current time step. In the schema guided knowledge decision, we will estimate the following probability pdecision(lx = 1|Ct, x), where x can be one schema description si or knowledge snippet ki . lx stands for the label to choose x or not given the dialogue context. The input is fed into transformer network in the following format: [CLS] Ct [SEP] x [SEP], and the hidden embedding of [CLS] in the last layer is used to estimate the above probability.

1st place

DSTC 9. Learning to Select External Knowledge with Multi-Scale Negative Sampling

Negatives Enhanced Knowledge Selection

Baidu 百度 | NLP

- Training with multi-scale negative samples $k_{i,j}^{-}$
  - Random: randomly selected from the whole knowledge snippet set
  - In-Domain: those within the same domain as k_i
  - In-Entity: those belonging to the same entity as k_i
  - Cross-Entity: those belonging to aforementioned entities in the dialogue context
  - $\mathcal{L}_{selection} = -\log p_{selection}(l_{k_i} = 1 | C_t, k_i) - \sum_j \log p_{selection}\left(l_{k_{i,j}^{-}} = 0 \middle| C_t, k_{i,j}^{-}\right)$
  - The ratio of positive to negative training samples is 1:4
- Inference process
  - optimal knowledge snippet $k^* = \max_{k_i \in K} p_{selection}(l_{k_i} = 1 | C_t, k_i)$

1st place

DSTC 9. Learning to Select External Knowledge with Multi-Scale Negative Sampling

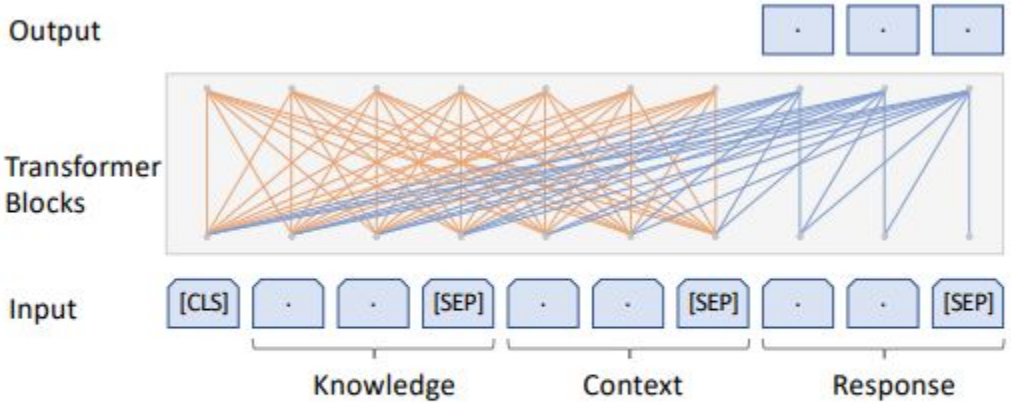


Figure 2: Knowledge grounded response generation. Orange lines denote bi-directional attention, and blue lines denote uni-directional attention

In this paper, we leverage powerful pre-trained models for knowledge grounded response generation. The network infrastructure is sketched in Figure 2. The backbone of the pre-training network consists of transformer blocks. The input to the network is the sum of the following four representations (Devlin et al. 2019; Bao et al. 2020a).

- Token Embedding. Following the conventional pre-processing, the input text is tokenized into byte-pair-encoding (BPE) tokens (Sennrich, Haddow, and Birch 2016).
- Segment Embedding. To better differentiate the input information, distinct segment embeddings are assigned to the knowledge snippet, dialogue context and response.
- Role Embedding. As the multi-turn conversation is interactive, role embeddings are employed to distinguish the utterances from different characters.
- Position Embedding. To obtain better extensibility on the input length, relative position embeddings are embraced in this generation network.

$$\mathcal{L}_{\text{generation}} = -\mathbb{E} \log p_{\text{generation}}(r|C_t, k) \quad (4)$$

where r refers to the target response. During training, we utilize the golden knowledge snippet $\tilde{k}$ for response generation. During inference, we rely on the knowledge snippet k^* retrieved with Equation (3) for response generation.

1st place

DSTC 9. Learning to Select External Knowledge with Multi-Scale Negative Sampling

Method		Task1: Knowledge-seeking Turn Detection			Task2: Knowledge Selection			Task3: Response Generation							
Team ID	Entry ID	Precision	Recall	F1	MRR@5	Recall@1	Recall@5	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-1	ROUGE-2	ROUGE-L
Baseline		0.9933	0.9021	0.9455	0.7263	0.6201	0.8772	0.3031	0.1732	0.1005	0.0655	0.2983	0.3386	0.1364	0.3039
3	1	**0.9964**	**0.9859**	**0.9911**	0.9395	0.9013	**0.9840**	**0.3864**	**0.2539**	0.1692	**0.1190**	0.3914	**0.4332**	**0.2115**	**0.3885**
15	3	0.9933	0.9677	0.9803	0.9195	0.8975	0.9460	0.3779	0.2532	**0.1731**	0.1175	**0.3931**	0.4204	0.2113	0.3765
19 (**Ours**)	0	0.9941	0.9430	0.9679	0.9181	0.8870	0.9554	0.3726	0.2402	0.1556	0.1064	0.3802	0.4103	0.1936	0.3665
	1	0.9911	0.9566	0.9735	0.9214	0.8883	0.9612	0.3780	0.2449	0.1594	0.1088	0.3853	0.4167	0.1978	0.3727
	2	0.9954	0.9818	0.9886	**0.9504**	**0.9235**	0.9814	0.3803	0.2449	0.1590	0.1081	0.3869	0.4192	0.1976	0.3738
	3	0.9954	0.9818	0.9886	**0.9504**	**0.9235**	0.9814	0.3735	0.2402	0.1550	0.1045	0.3815	0.4132	0.1940	0.3685
	4	0.9954	0.9818	0.9886	**0.9504**	**0.9235**	0.9814	0.2243	0.1537	0.0998	0.0663	0.2887	0.3530	0.1855	0.3150

Table 3: Experimental results on the test set, with the highest value written in bold.

1st place

DSTC 9. Learning to Select External Knowledge with Multi-Scale Negative Sampling

Rank	Team ID	Entry ID	Accuracy	Appropriateness	Average
Ground-truth			4.5930	4.4513	4.5221
1	19 (**Ours**)	2	**4.3917**	**4.3922**	**4.3920**
2	3	1	4.3480	4.3634	4.3557
3	10	0	4.3544	4.3201	4.3373
4	15	3	4.3793	4.2755	4.3274
5	17	0	4.3360	4.3076	4.3218
Baseline			3.7155	3.9386	3.8271

Table 4: Final human evaluation on the test set, with the highest value written in bold.

1st place

DSTC 9. RADGE: Relevance Learning and Generation Evaluating Method for Task-Oriented Conversational Systems

Input Representation: As Figure 2 shows, the model takes the concatenation of the domain name, API keywords, and the dialog context as input instances.

Loss Function and Output Representation: The ELECTRA-based model is fine-tuned with binary cross-entropy loss (Rubinstein, 1999) as follows:

$$L_{domain} = -\frac{1}{N}\sum_{i=1}^{N}[d_i \log \hat{d}_i + (1-d_i)\log(1-\hat{d}_i)], \quad (1)$$

$$L_{api} = -\frac{1}{N}\sum_{i=1}^{N}[a_i \log \hat{a}_i + (1-a_i)\log(1-\hat{a}_i)], \quad (2)$$

$$L_{total} = \alpha L_{domain} + \beta L_{api}, \quad (3)$$

where $\hat{d}_i$ indicates the probability whether the last utterance domain is the same as the given domain, whereas $\hat{a}_i$ refers to the correlation between the API keywords, domains, and dialog contexts. The final loss L_{total} is the weighted sum of two losses.

A dialog context, API keywords, and domains are defined as $U_t = \{u_{t-w+1}, \dots, u_{t-1}, u_t\}$, $A = \{a_1, \dots, a_n\}$, $D = \{d_1, \dots, d_n\}$, respectively, where u_i is the i-th utterance in a given dialog, t is the time-step of the current user utterance to be processed. To a given user turn at t, each pair of U_t and $a_i \in A$, $d_i \in D$ is taken into account to predict whether they are relevant as follows:

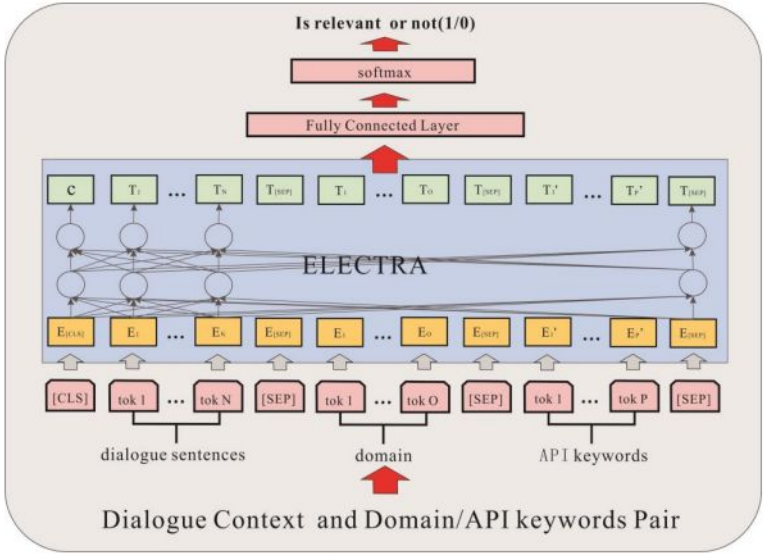


Figure 1: The model overview

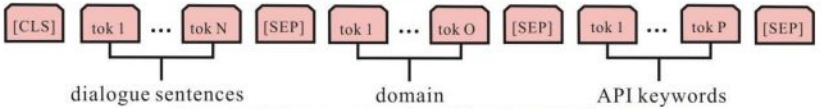


Figure 2: The input representation of the proposed model in Task 1

3d place

DSTC 9. RADGE: Relevance Learning and Generation Evaluating Method for Task-Oriented Conversational Systems

$$L_{domain} = -\frac{1}{N}\sum_{i=1}^{N}[d_i \log \hat{d}_i + (1-d_i)\log(1-\hat{d}_i)], \quad (1)$$

$$L_{api} = -\frac{1}{N}\sum_{i=1}^{N}[a_i \log \hat{a}_i + (1-a_i)\log(1-\hat{a}_i)], \quad (2)$$

$$L_{total} = \alpha L_{domain} + \beta L_{api}, \quad (3)$$

where $\hat{d}_i$ indicates the probability whether the last utterance domain is the same as the given domain, whereas $\hat{a}_i$ refers to the correlation between the API keywords, domains, and dialog contexts. The final loss L_{total} is the weighted sum of two losses.

Positive Sample: For each dialog context handled by the API-based branch, a positive sample is constructed with the corresponding domain and DB-API keywords.

Negative sample:

(1) For each dialog handled by the API-based branch, a negative sample is constructed with the other domain name but the corresponding DB-API keywords;

(2) For each dialog handled by the API-based branch, a negative sample is constructed with the corresponding domain name and the KL-API keywords;

(3) For each dialog handled by the knowledge access branch, a negative sample is constructed with the corresponding domain name and the DB-API keywords.

3d place

DSTC 9. RADGE: Relevance Learning and Generation Evaluating Method for Task-Oriented Conversational Systems

Dataset	System	Precision	Recall	F1
all	baseline(official)	0.9933	0.9021	0.9455
	baseline(ELECTRA)	0.9829	0.9263	0.9537
	proposed model w/o optimized with KL-API keywords	0.9817	0.9465	0.9638
	proposed model (single)	0.9799	0.9591	0.9694
seen domains	baseline(official)	--	--	--
	baseline(ELECTRA)	0.9818	0.9536	0.9675
	proposed model w/o optimized with KL-API keywords	0.9821	0.9691	0.9755
	proposed model (single)	0.9826	0.9680	0.9752
unseen domain(attraction)	baseline(official)	--	--	--
	baseline(ELECTRA)	1.0000	0.6374	0.7786
	proposed model w/o optimized with KL-API keywords	0.9758	0.7076	0.8203
	proposed model (single)	0.9487	0.8655	0.9052

Table 1. Comparison of four models in Task 1 performance

3d place

DSTC 9. RADGE: Relevance Learning and Generation Evaluating Method for Task-Oriented Conversational Systems

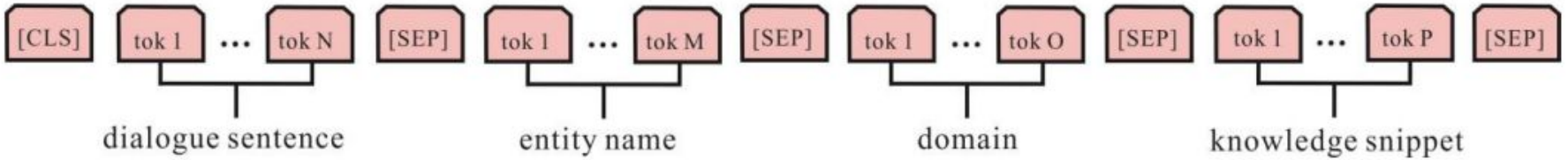


Figure 3: The input tokens of the proposed model in Task 2

System	MRR@5	R@1	R@5	Average
baseline(official)	0.7263	0.6201	0.8772	0.7412
baseline(ELECTRA)	0.7856	0.6871	0.9225	0.7984
baseline optimized with Entity candidates	0.8889	0.8227	0.9660	0.8925
proposed model(single)	0.9372	0.9117	0.9665	0.9385

Table 2. Comparison of five systems in Task 2 performance

3d place

DSTC 9. RADGE: Relevance Learning and Generation Evaluating Method for Task-Oriented Conversational Systems

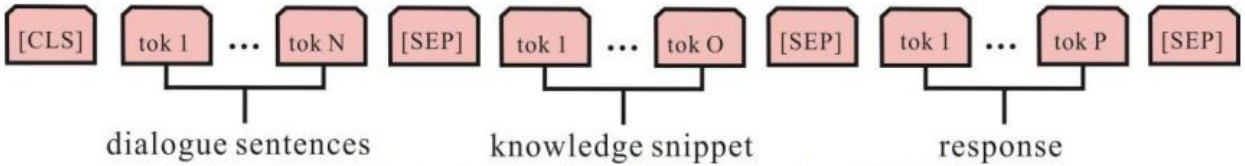


Figure 4: The input representation of the proposed model in Task 3

Loss Function and Output Representation: The ELECTRA-based model is fine-tuned with binary cross-entropy loss as below:

$$L = -\frac{1}{N}\sum_{i=1}^{N}[s_i \log \hat{s}_i + (1 - s_i)\log(1 - \hat{s}_i)], \quad (10)$$

where $\hat{s}_i$ shows the probability whether the response is the same as the ground truth.

3d place

DSTC 9. RADGE: Relevance Learning and Generation Evaluating Method for Task-Oriented Conversational Systems

System	Automated Metrics								
	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-1	ROUGE-2	ROUGE-L	Average
baseline (official)	0.3031	0.1732	0.1005	0.0655	0.2983	0.3386	0.1364	0.3039	0.2149
proposed model (single)	0.3505	0.2171	0.1359	0.0893	0.3491	0.3938	0.1759	0.3551	0.2583
proposed model (generation-evaluating)	0.3684	0.2374	0.1531	0.1030	0.3719	0.4113	0.1938	0.3692	0.2760

Table3. Comparison of three models in Task 3 performance in automated metrics

System	Human Evaluation		
	Accuracy	Appropriateness	Average
ground-truth	4.5930	4.4513	4.5221
baseline(official)	3.7155	3.9386	3.8271
proposed system	4.3544	4.3201	4.3373

Table4. Comparison of two systems in Task 3 performance in official human evaluation

3d place

DSTC 9. Beyond Domain APIs: Task-oriented Conversational Modeling with Unstructured Knowledge Access

Outcomes:

1. Задача хорошо решается ансамблями с учетом мета инфы из api скиллов
2. Крайне важно креативно подойти к Negative sampling для ваших классификаторов
3. Важно делать как можно больше осмысленной аугментации (backtrans, paraphrase, аугментация через сущности итд)

DSTC 9. Track 2

DSTC 9. Multi-domain Task-oriented Dialog Challenge II

1 Introduction

There has been an increasing interest in building a dialog system crossing multiple domains to accomplish a complex goal [1, 2, 3, 4, 5, 6]. With the success of Multi-Domain Task Completion Dialog Challenge in DSTC-8 Track 1, we continue with the effort of building dialog systems under the multi-domain setting in this proposal. Compared with the previous challenge, in this track, we extend the tasks by incorporating new datasets, creating new sub-tasks, and providing a new development platform. We specifically focus on two aspects of dialog systems: language portability and end-to-end system complexity.

First, with the rapid globalization process, the need for building dialog systems that supports multiple languages in the same or a similar scenario is ever increasing. However, building a task-oriented dialog system requires a considerable amount of annotated training data, and for some languages, the training corpus is very limited. Therefore, cross-lingual transfer learning becomes a popular topic during the years and helps reduce the cost of developing dialog systems for resource-poor target languages. To advance the state-of-the-art cross-lingual technology in building dialog systems, we introduce the task of cross-lingual dialog state tracking in this track.

Second, with the advancement of end-to-end learning, building a dialog system, and evaluating its performance in an end-to-end fashion has attracted increasing interest from the community. In this challenge, by offering the end-to-end dialog system task, we also provide the latest development platform that enables users to build, evaluate, and diagnose a full dialog system with ease.

2 Task Description

This track consists of two tasks:

- Participants will build a cross-lingual multi-domain dialog state tracker.
- Participants will develop multi-domain end-to-end dialog systems with the provided development platform.

DSTC 9. Multi-domain Task-oriented Dialog Challenge II

2.2 Multi-domain End-to-end Dialog Challenge

Most prior works focus on improving individual components in a dialog system, without evaluating the performance of the entire system. However, the modular performance improvement does not necessarily contribute to the end-to-end performance increment. To foster the development of end-to-end dialog systems, DSTC-8 Track 1 introduced an end-to-end multi-domain task (Task 1) and received reasonable submissions from participants.

In this task, we continue with the same setting as DSTC-8 Track 1 Task 1 with two changes:

- Dataset. Instead of using MultiWOZ 2.0 [4] as in DSTC-8, we will build dialog systems based on MultiWOZ 2.1 [12] in this task. Compared with the previous version, MultiWOZ 2.1 re-annotated states and utterances based on the original utterance to fix the original noisy annotation. It also contains user dialog act annotation, which is missing in MultiWOZ 2.0. We believe that with MultiWOZ 2.1, participants are empowered to build more effective dialog systems.
- Development Platform. ConvLab [14] is employed as the development platform in the previous challenge. In this task, we are providing the next generation of ConvLab (ConvLab-2), which integrates more powerful models and supports more datasets on top of ConvLab. Also, it includes an interactive visual toolkit that enables searchers to inspect the components and diagnose the dialog systems with an enhanced evaluator.

The participants are encouraged to experiment with various approaches based on ConvLab-2 to build a dialog system that takes natural language utterance as input, tracks dialog states during the conversation, interacts with a task-specific dataset, and generates a system response at each turn. Both automatic evaluation and human evaluation results will be reported in the challenge.

- Automatic Evaluation. We will provide an automatic evaluation script equipped with an end-to-end user simulator and evaluator, and report a range of metrics including task success rate, return (reward), number of turns for dialog policy, book rate, and precision/recall/F1 score for intent/slot detection.
- Human Evaluation. For the human evaluation, we will crowdsource the work on Amazon Mechanic Turk so human judges can communicate with the agent via natural language, and make a judgment of the system based on the dialog success/failure, language understanding score, and response appropriateness score.

The final ranking only considers the success rate in human evaluation results.

DSTC 9. Multi-domain Task-oriented Dialog Challenge II

In this task, following a similar scheme as in DSTC-5, our goal is to build a cross-lingual dialog state tracker with a training set in the rich resource language and a small development set in the low resource language. The performance of each dialog state tracker will be evaluated on an unlabelled test set in the low resource language and compared with reference annotation. In particular, we will offer two sub-tasks based on MultiWOZ 2.1 [12] and CrossWOZ [13], respectively.

- MultiWOZ 2.1. MultiWOZ is a multi-domain dialog dataset spanning 7 distinct domains and containing over 10,000 dialogs under the travel planning setting. We augment the dataset with Chinese translations as the development/test set. English is considered as the rich resource language and Chinese as the low resource dataset.
- CrossWOZ. CrossWOZ is the first large scale Chinese multi-domain task-oriented dialog dataset spanning 5 distinct domains and containing over 6,000 dialogs under the setting of travel planning to Beijing. English translation is provided along with the original Chinese corpus. In this sub-task, Chinese is considered as the rich resource language and English as the low resource dataset.

We evaluate the performance of the dialog state tracker using two metrics:

- Joint Goal Accuracy. This metric evaluates whether the predicted dialog state is exactly equal to the ground truth.
- Slot Precision/Recall/F1. These metrics evaluate whether the predicted labels for individual slots in dialog state are equal to the ground truth, microaveraged over all slots.

DSTC 9. Multi-domain Task-oriented Dialog Challenge II

Table 1: An example of partial dialog for CrossWOZ dataset.

Speaker	Utterance	Dialog States
User	你好，可以帮我找一个人均消费是100-150元，评分是5分的餐馆吗? Hello, can you help me find a restaurant with a per capita consumption of 100-150 yuan and a score of 5?	餐馆: {人均消费: 100-150元, 评分: 5分, 名称: “”} Restaurant: {per capita consumption: 100-150 yuan, score: 5 points, name: “”}
System	全聚德烤鸭(吕营店)和咸亨酒店(和平店)都是比较火的餐馆，你感兴趣不? Quanjude Roast Duck (Luying) and Xianheng Hotel (Heping) are hot restaurants. Are you interested?	餐馆: {人均消费: 100-150元, 评分: 5分, selectedResults: [全聚德烤鸭(吕营店),咸亨酒店(和平店)] } Restaurant: {per capita consumption: 100-150 yuan, score: 5 points, selectedResults: [Quanjude Roast Duck (Luying), Xianheng Hotel (Heping Branch)] }
User	我选择去全聚德烤鸭(吕营店)吃饭，它家都有什么推荐菜啊? I choose to eat at Quanjude Roast Duck (Luying). What recommendations does it have?	餐馆: { 名称: 全聚德烤鸭(吕营店), 推荐菜: “”, 人均消费:100-150元, 评分: 5分} Restaurant: { name: Quanjude Roast Duck (Luying), recommended dishes: “””, per capita consumption: 100-150 yuan, score: 5 points }
System	推荐菜有清蒸鲈鱼, 盐水鸭干。 Recommended dishes are steamed sea bass and salted duck.	餐馆: { 名称: 全聚德烤鸭(吕营店), 推荐菜: ””, 人均消费: 100-150元, 评分: 5分, selectedResults: [全聚德烤鸭(吕营店)] } Restaurant: { name: Quanjude Roast Duck (Luying), recommended dishes: ””, per capita consumption: 100-150 yuan, score: 5 points, selectedResults: [Quanjude Roast Duck (Luying)] }
User	好的，那它家周边有景点吗? OK, are there any attractions around it?	餐馆: { 名称: 全聚德烤鸭(吕营店), 推荐菜:[清蒸鲈鱼, 盐水鸭干], 人均消费: 100-150元, 评分: 5分, 周边景点: [] } Restaurant: { name: Quanjude Roast Duck (Luying), recommended dishes: [steamed sea bass, dried salted duck], per capita consumption: 100-150 yuan, score: 5 points, attractions nearby: [] }

DSTC 9. A Hybrid Task-Oriented Dialog System with Domain and Task Adaptive Pretraining

- Data delexicalization

```
i show a train leaving birmingham new street at 17:40 and arriving at 20:23 on wednesday .
i show a train leaving [value_departure] at [value_leave] and arriving at [value_arrive] on [value_day] .

it 's a hotel . there are 5 guesthouses in the area . do you prefer cheap or moderate for the price range ?
it 's a [value_type] . there are 5 [value_type] in the area . do you prefer [value_pricerange] or
[value_pricerange] for the price range ?
```

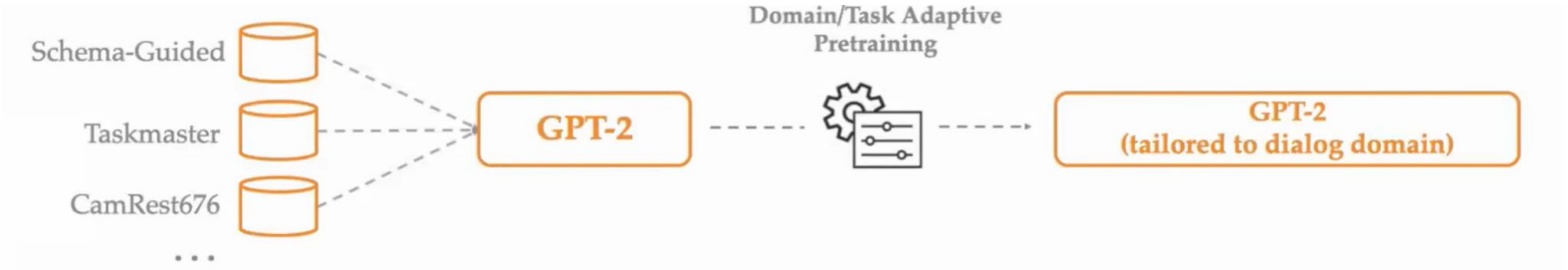
- Data cleaning & normalization

```
4pm -> 16:00
we're -> we are
...
```

- Recover delexicalization
  - Neural model predicts delexicalized response, to generate the final response, we replace placeholders with the real value from the dialog context or database.

1st place

DSTC 9. A Hybrid Task-Oriented Dialog System with Domain and Task Adaptive Pretraining



- Continue to pretrain GPT-2 on the plain text of dialog domain datasets
- Domain adaptive pretraining
  - Pretrain GPT-2 on external dialog related datasets, such as Taskmaster and Schema-guided Dialog, etc.
- Task adaptive pretraining
  - Pretrain GPT-2 on the plain text of MultiWoz dataset

DiDi

1st place

DSTC 9. A Hybrid Task-Oriented Dialog System with Domain and Task Adaptive Pretraining

During training, we adopt the multi-task fine-tuning strategy of Peng et al. (2020):

- **Belief Prediction**: We use $\mathbf{c}_1$ (dialog history) to predict $\mathbf{c}_2$ (turn domain) and $\mathbf{c}_3$ (belief state), and define the objective as:

$$\mathcal{L}_B = \log p(\mathbf{c}_2, \mathbf{c}_3 | \mathbf{c}_1) = \sum_{t=1}^{T_{\mathbf{c}_{2,3}}} \log p_\theta(c_t | c_{<t}, \mathbf{c}_1)$$

, where $c_{<t}$ indicates all tokens before t.

- **Response Prediction**: We use $\mathbf{c}_3$ (Belief State) to query the database and create $\mathbf{c}_4$ (DB Match). Then we predict $\mathbf{c}_5$ (Dialog Response) which is conditioned on $\mathbf{c}_1$, $\mathbf{c}_2$, $\mathbf{c}_3$, and $\mathbf{c}_4$. The response prediction objective is defined as:

$$\mathcal{L}_R = \log p(\mathbf{c}_5 | \mathbf{c}_{1-4}) = \sum_{t=1}^{T_{\mathbf{c}_5}} \log p_\theta(c_t | c_{<t}, \mathbf{c}_{1-4})$$

, where $c_{<t}$ indicates all tokens before t.

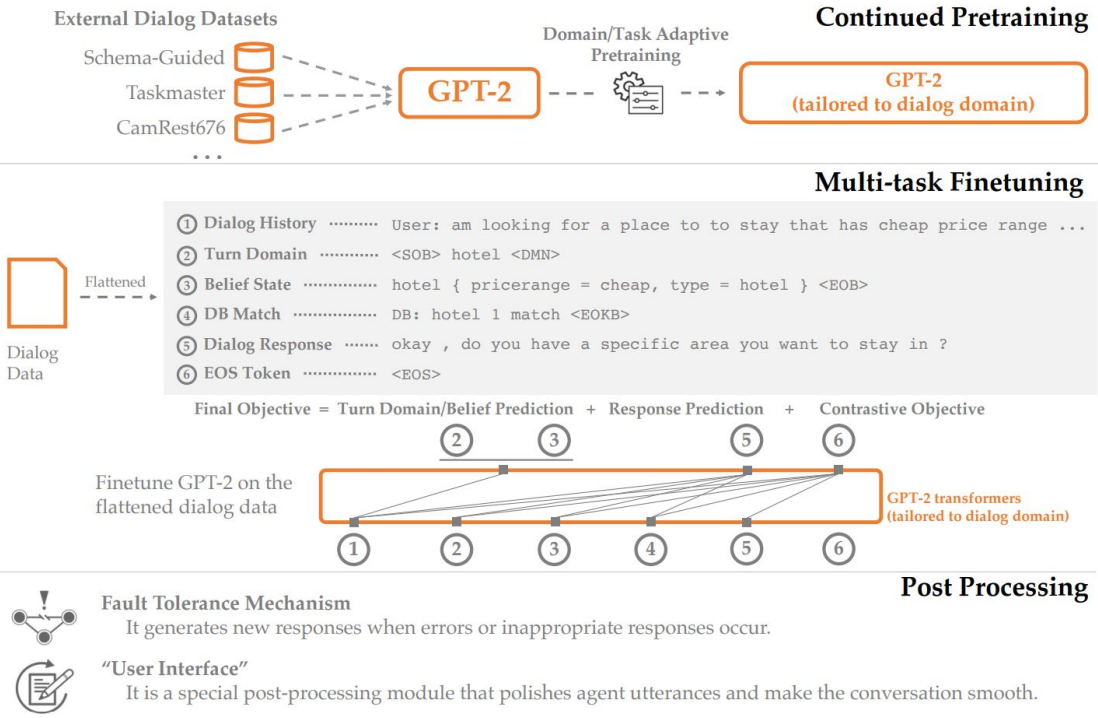
- **Contrastive Ojbective**: It is a common practice to introduce constrastive samples for training machine learning models (Peng et al. 2020; Ham et al. 2020a). We create the negative sample by modifying the positive sample, through one of the following three ways: 1) replace $\mathbf{c}_2$ (belief state) with another random $\mathbf{c}_2$, 2) replace $\mathbf{c}_5$ (dialog response) with another random $\mathbf{c}_5$, or 3) replace both $\mathbf{c}_2$ and $\mathbf{c}_5$ with another $\mathbf{c}_2$ and $\mathbf{c}_5$. We apply a binary classifier on "<EOS>" token of the positive and negative samples. The objective function is:

$$\mathcal{L}_C = y \log(p_\theta(\textbf{positive})) + (1-y) \log(1 - p_\theta(\textbf{negative}))$$

Thus, the full fine-tuning objective is:

$$\mathcal{L} = \mathcal{L}_B + \mathcal{L}_R + \mathcal{L}_C.$$

During inference, there are two stages of predictions: 1) given $\mathbf{c}_1$ (dialog history), the model predicts $\mathbf{c}_2$ (turn domain) and $\mathbf{c}_3$ (belief state), and queries the database to generate $\mathbf{c}_4$ (DB match), and 2) the model predicts $\mathbf{c}_5$ (dialog response) based on $\mathbf{c}_{1-4}$.



1st place

DSTC 9. A Hybrid Task-Oriented Dialog System with Domain and Task Adaptive Pretraining

Rank	Team	Success Rate			Language Understanding	Response Appropriateness	Turns
		Avg.†	w/ DB	w/o DB			
10	Team 10	19.5	6.0	33.0	3.23	2.93	18.8
9	Team 8	35.0	26.0	44.0	3.27	3.15	18.5
8	Team 9	55.2	43.2	67.2	4.15	3.98	19.2
7	Team 5	58.4	50.4	66.4	4.15	4.06	19.7
6	Team 4	60.3	51.4	69.2	4.49	4.22	17.7
5	Team 3	67.8	60.0	75.6	4.56	4.42	21.0
4	Team 6	70.6	60.8	80.4	4.41	4.41	20.1
3	Team 7	72.3	62.0	82.6	4.53	4.41	17.1
1	Team 1	74.8	70.2	79.4	4.54	4.47	18.5
1	**Ours**	**74.8**	**68.8**	**80.8**	**4.51**	**4.45**	**19.4**

† Ranking of the teams is based on the average success rate.

Table 3: Official results of the human evaluation. We tie for first place with Team 1. The rank is based on the average success rate. Please refer to the Human Evaluation section for details of the evaluation metric.

Rank	Team	Success Rate†	Book Rate	Inform Rate			Turns	
				P	R	F	succ.	all
10	Team 10	21.4	0.0	55.4	60.0	54.1	11.0	25.9
9	Team 9	44.4	26.5	57.9	64.5	58.9	12.2	14.6
8	Team 8	52.6	66.7	57.5	80.7	64.8	13.2	22.5
7	Team 7	57.8	85.0	68.7	81.6	72.6	13.7	16.4
6	Team 6	67.7	90.8	70.4	85.6	75.2	12.8	14.2
5	Team 5	83.3	89.1	81.1	90.3	83.5	13.5	13.8
4	Team 4	89.8	96.3	72.4	96.0	80.1	15.1	15.8
3	Team 3	90.8	96.7	81.0	95.4	85.9	13.4	13.6
1	Team 1	93.0	94.6	84.1	96.2	88.1	12.5	12.7
2	**Ours**	**91.4**	**96.9**	**80.2**	**97.3**	**86.0**	**15.3**	**15.7**

† Ranking of the teams is based on the success rate.

Table 4: Official results of the user simulator based automatic evaluation. The rank is based on the success rate. Please refer to the Automatic Evaluation section for details of the evaluation metric.

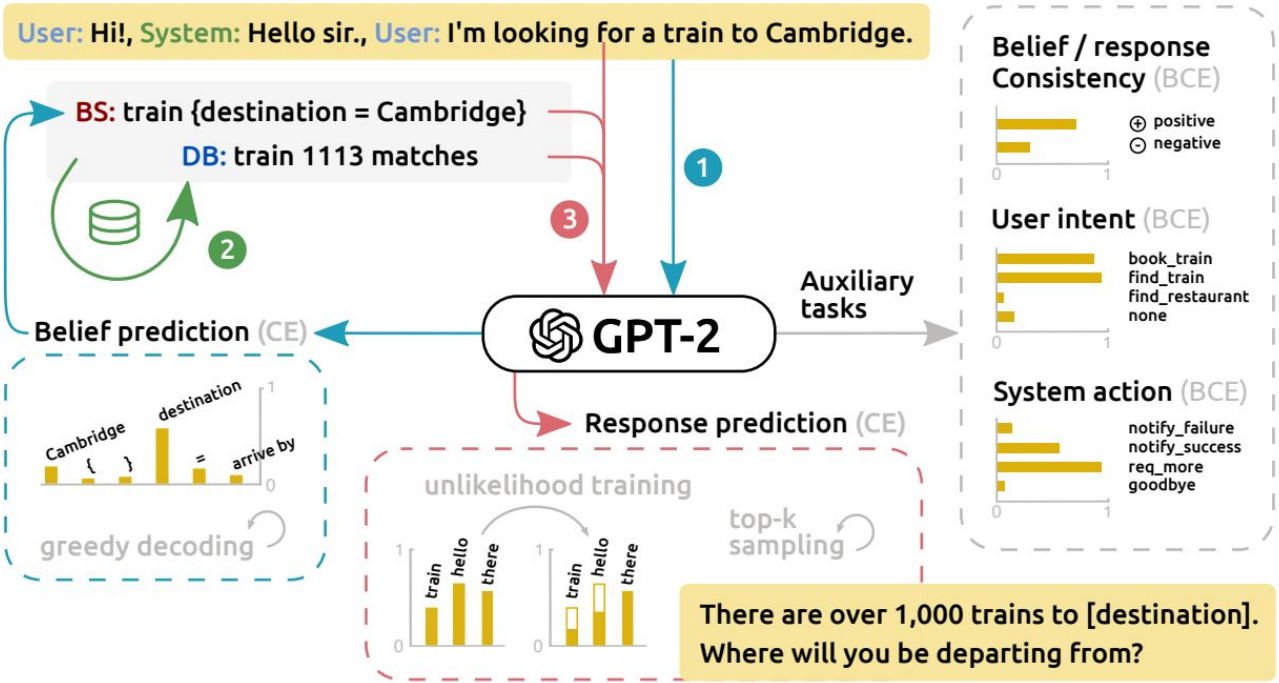
1st place

DSTC 9. AuGPT: Dialogue with Pre-trained Language Models and Data Augmentation

- model based on SOLOIST (Peng et al., 2020)
- auto-regressive language model used for decoding of (1) belief state and (2) response
- belief state also used as the query for the database
- finetuned GPT2 used as the language model

2nd place

DSTC 9. AuGPT: Dialogue with Pre-trained Language Models and Data Augmentation



2nd place

DSTC 9. AuGPT: Dialogue with Pre-trained Language Models and Data Augmentation

- we corrupt 50% of the samples and train a binary classifier to detect them, this is called **consistency detection** auxiliary task
- the binary classifier is attached to the last hidden representation of GPT2
- when constructing a negative sample, one of more of the following changes are applied with the same probability:
  - a different belief state is randomly sampled from the dataset
  - a different database counts are randomly sampled from the dataset
  - values in the belief state are uniformly randomly resampled **(new)**
- we also tried user intent and system action classifiers, but that did not help performance

2nd place

DSTC 9. AuGPT: Dialogue with Pre-trained Language Models and Data Augmentation

- massive **data augmentation** via backtranslation
- improved consistency detection **auxiliary training task**
- using **unlikelihood loss** for response to increase output diversity
- different decoding strategies for belief state and response
- training on **'clean'** samples only

2nd place

DSTC 9. AuGPT: Dialogue with Pre-trained Language Models and Data Augmentation

Method	Average Success	Success w/ DB	Success w/o DB	NLU score	Response appropriatenes	Turns
Baseline	69.6	56.8	82.4	4.34	4.18	18.5
Team1 (winner)	**74.8**	**70.2**	79.4	**4.54**	**4.47**	18.5
Team7 (ours)	72.3	62.0	**82.6**	4.53	4.41	**17.1**

Table 3: Human evaluation results on the DSTC9 Track 2. We show a shortened version of the table, the full version is published online at https://convlab.github.io. Note that only 4 teams outperformed the Baseline according to the average success metric.

	MultiWOZ 2.1			ConvLab 2							
							inform			turn	
method	inform	success	BLEU	complete	success	book	P	R	F1	succ	all
AuGPT*	91.4	**72.9**	17.2	**89.4**	**60.1**	85.7	64.5	**82.1**	70.3	12.7	14.6
w/o. unlikelihood*	90.8	70.4	16.9	89.2	59.3	**90.8**	63.9	81.6	69.5	12.8	14.6
w/o. clean	**91.6**	70.7	15.8	85.0	57.7	85.6	65.6	79.1	69.6	12.7	14.5
w/o. unlikelihood, clean*	90.4	72.7	**17.5**	85.9	58.4	81.3	62.2	79.8	67.5	12.6	**14.1**
w. all auxiliary*	91.1	71.4	16.8	88.7	59.2	86.0	64.6	81.1	69.9	**12.6**	14.4
w/o. pre-training*†	90.7	67.9	15.1	88.1	59.8	83.7	**68.1**	80.9	72.1	13.5	15.6
w/o. back-translations	89.1	67.9	15.2	88.9	58.2	87.4	68.0	81.6	**72.2**	12.9	14.9
w. old consistency	90.7	71.8	17.0	85.5	57.8	86.0	65.2	80.0	69.8	12.7	14.6
w/o. consistency	90.4	68.7	16.8	86.4	57.1	84.1	66.3	81.2	70.9	13.1	14.6

Table 4: AuGPT ablation study. The model version with the best ConvLab 2 success rate is chosen as our best model and dubbed AuGPT. Variants are denoted with their respective modifications compared to AuGPT: "w/o. unlikelihood" = unlikelihood loss was not used for training; "w/o. clean" uses all training samples as opposed to using only the ones consistent with the database; "w/o. pre-training" = the additional Taskmaster-1 and Schema-Guided datasets were not used for training; "all auxiliary" = using two additional auxiliary tasks (see the Method section for details); "w/o. consistency" = dialogue consistency task is not used; "old consistency" refers to the consistency task as defined by Peng et al. (2020) (see the Method section for details). All variants submitted to the DSTC9 shared task are denoted "*", the model chosen for DSTC9 human evaluation is denoted "†".

2nd place

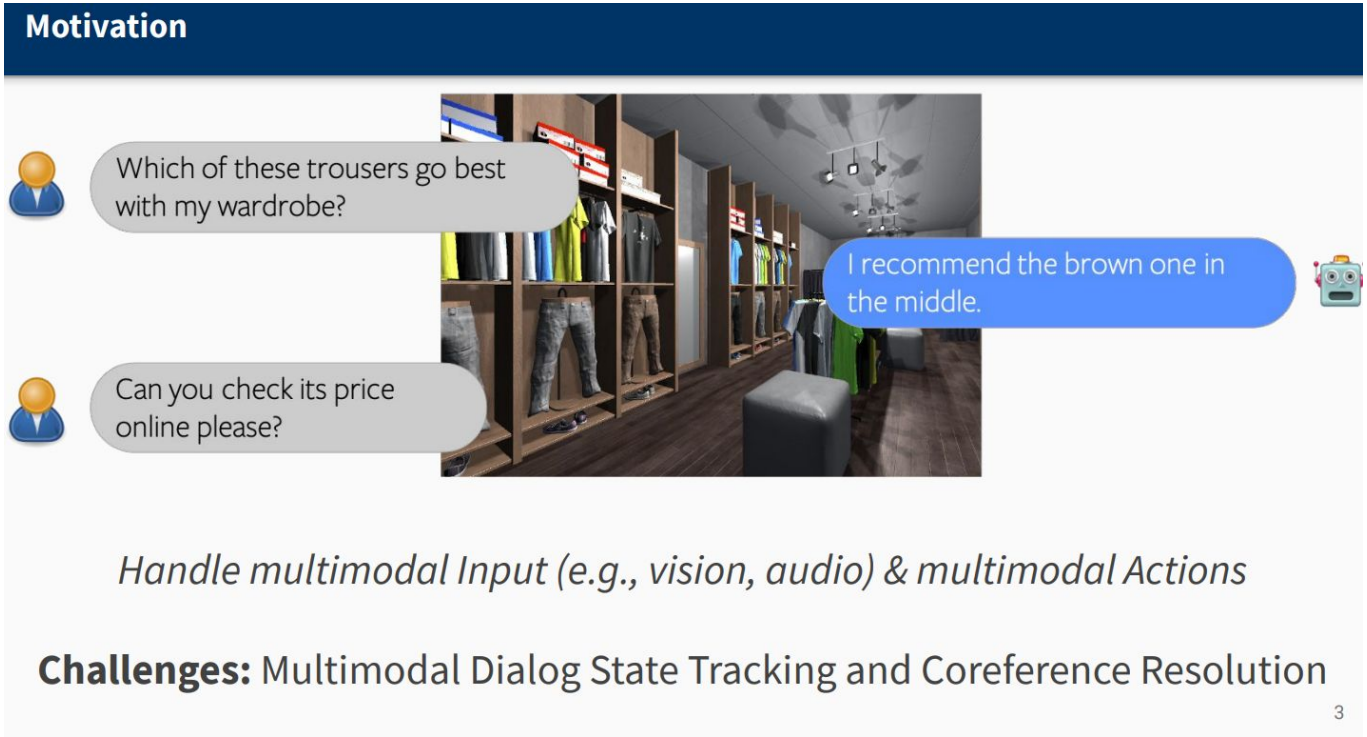
DSTC 9. Multi-domain Task-oriented Dialog Challenge II

Outcomes:

1. Важно делать pretraining, причем pretraining domain adaptive, здесь был полезней чем task adaptive
2. Хорошая эвристика с делексикализацией, единственное надо будет потратить на обратный декодинг

DSTC 9. Track 4

DSTC 9. Situated Interactive Multimodal Conversations (SIMMC)



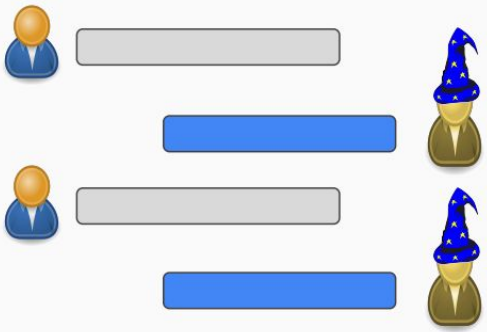
DSTC 9. Situated Interactive Multimodal Conversations (SIMMC)

Dialogs grounded in **co-observed**, **evolving** multimodal context; provide shopping experience

Role of the User
Browse through a catalog (and purchase items)

Co-observed multimodal context

Manipulated by the assistant



Role of the Assistant
Facilitate a shopping experience for the user

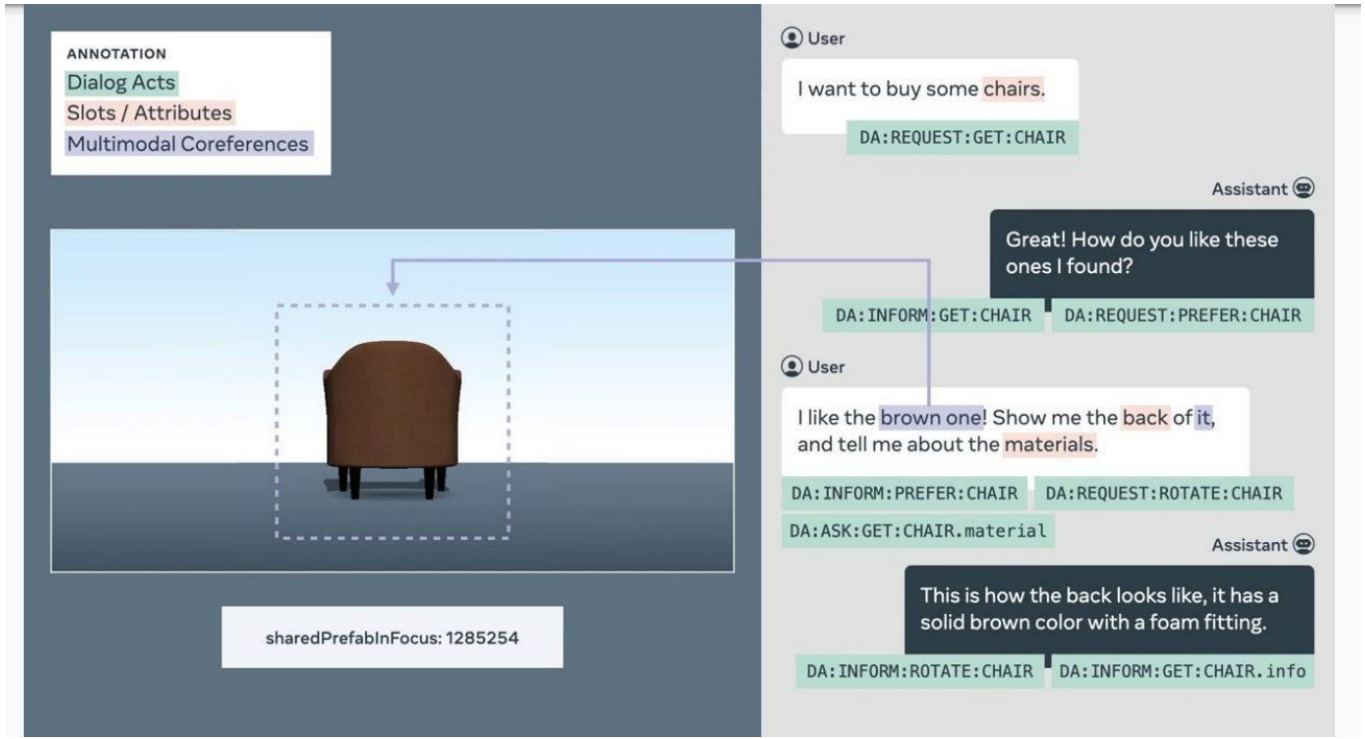
Populate the carousel
'Show me some red couches'

Interact with items (focus, rotate)
'How does the back of this chair look?'

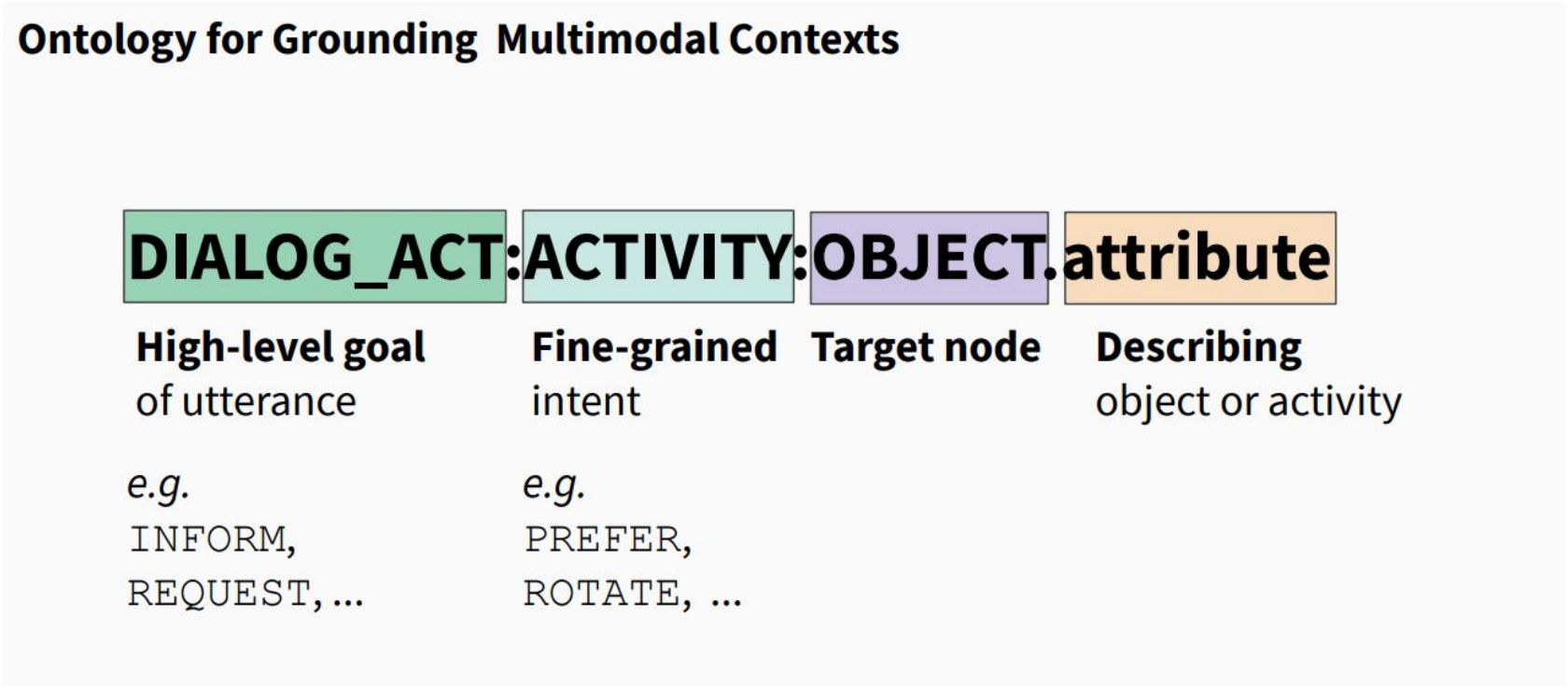
Provide additional information
'Tell me more about this table'

Data Collection Interface : Crook and Podder et al., ASRU (2019)

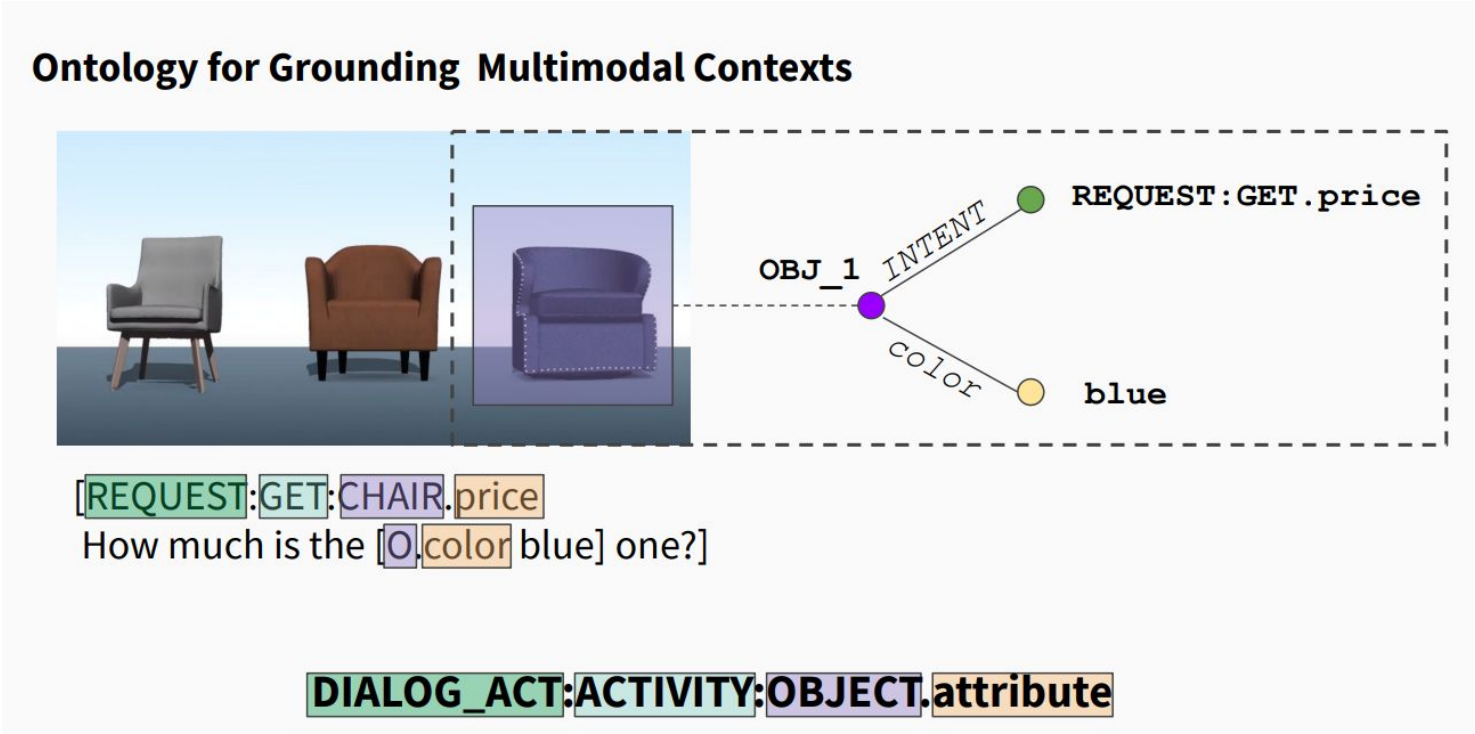
DSTC 9. Situated Interactive Multimodal Conversations (SIMMC)



DSTC 9. Situated Interactive Multimodal Conversations (SIMMC)



DSTC 9. Situated Interactive Multimodal Conversations (SIMMC)



DSTC 9. Situated Interactive Multimodal Conversations (SIMMC)

Statistics	Furniture (VR)		Fashion (Image)
	Text	Audio†	
Total # dialogs	6.4k	1.3k	6.6k
Total # utterances	97.6k	15.8k	71.2k
Avg # rounds / dialog	7.62	7.16	5.39
Avg # tokens (user)	11.0	N/A	11.10
Avg # tokens (assistant)	12.2	N/A	10.87

DSTC 9. Situated Interactive Multimodal Conversations (SIMMC)

Hierarchical Dialog Acts
e.g.

- *INFORM*
  - *GET*
  - *REFINE*
  - *PREFER / DISPREFER*
  - *ADD_TO_CART*
  - *COMPARE*
  - *COUNT ...*
- *CONFIRM*
  - *GET*
  - *REFINE*
  - *ADD_TO_CART*
  - *…*
- *REQUEST …*
- *ASK …*
- *PROMPT ...*

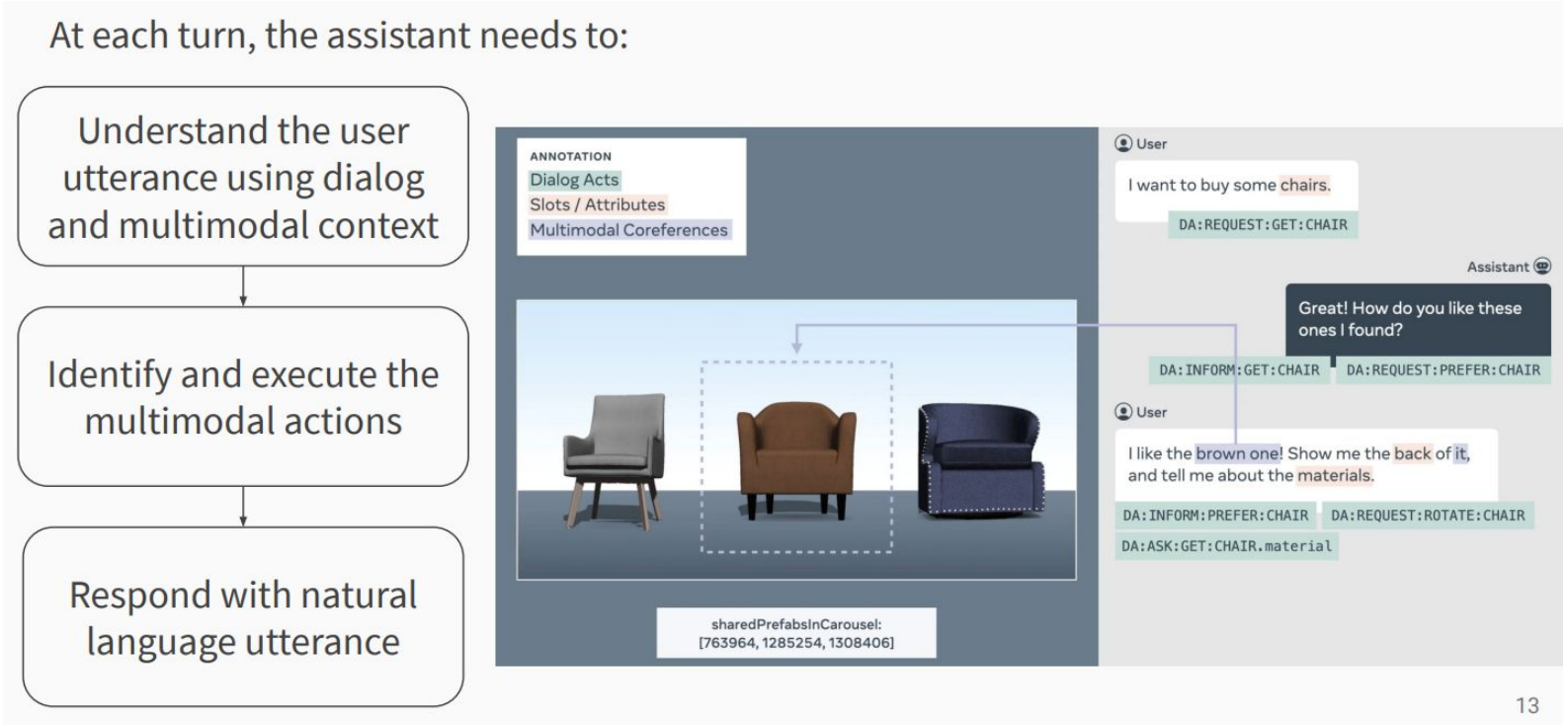
42 Unique Intent Nodes

~100 Unique Attribute Slots
e.g.

color, date_time, user_attention_on, price, size, pattern, necklineStyle, sleeveLength, ...

~1600 Unique Values
e.g.

DSTC 9. Situated Interactive Multimodal Conversations (SIMMC)



DSTC 9. Situated Interactive Multimodal Conversations (SIMMC)

U: I want to buy some chairs.

A: Great! How do you like these ones I found?

H_t

M_t

I like the brown one! Show me the back of it, and tell me about the materials.

U_t

ROTATE[behind]

a_t

Task 1: Assistant API Prediction

Action accuracy, attribute accuracy, action perplexity

This is how the back looks like, it has a solid brown color with a foam fitting.

A_t

Task 2: Assistant Response Generation

BLEU-4 (generation)
Mean reciprocal rank, r@k, mean rank (retrieval)

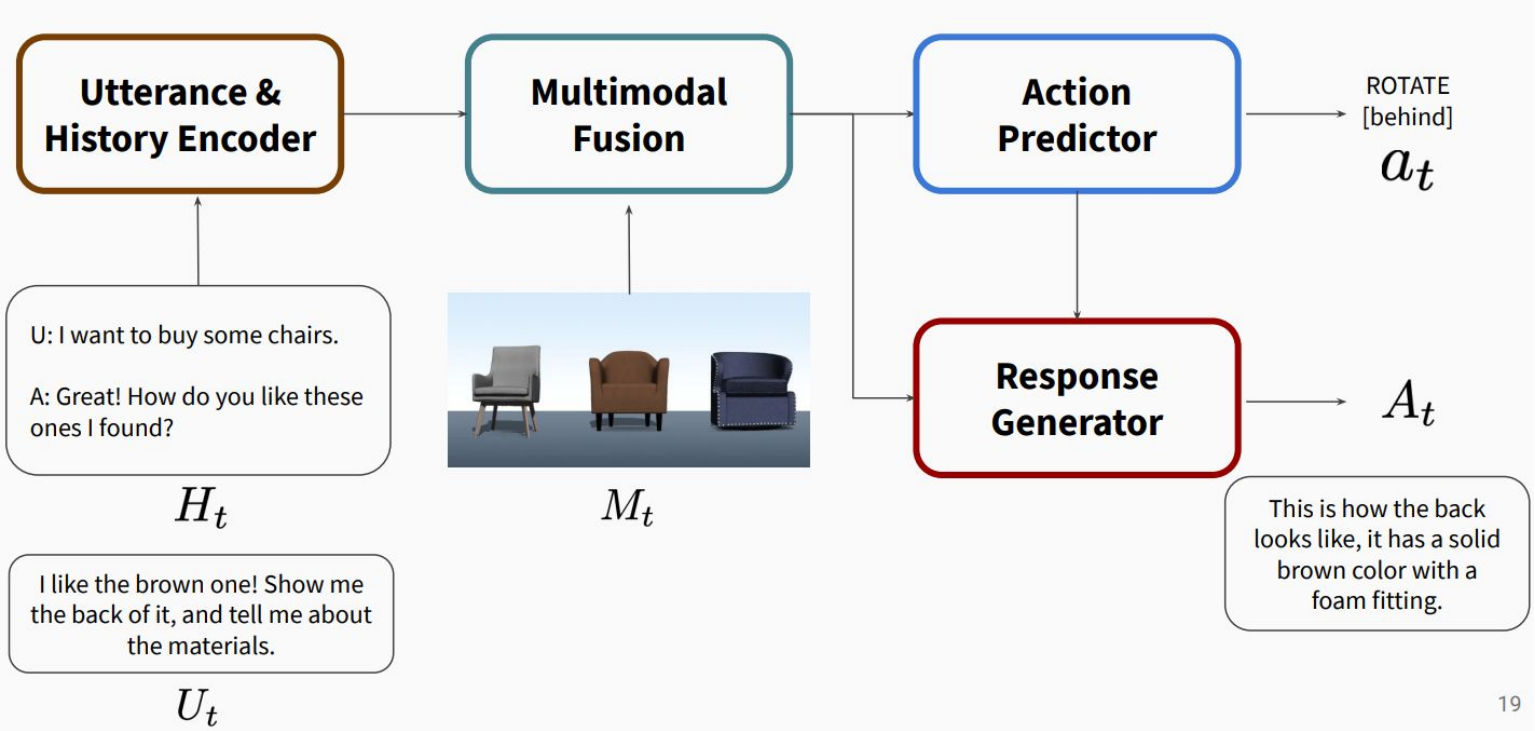
DA:CONFIRM:PREFER:FURNITURE.color=Brown
DA:REQUEST:ROTATE:FURNITURE.back
DA:ASK:GET:FURNITURE.material

D_t

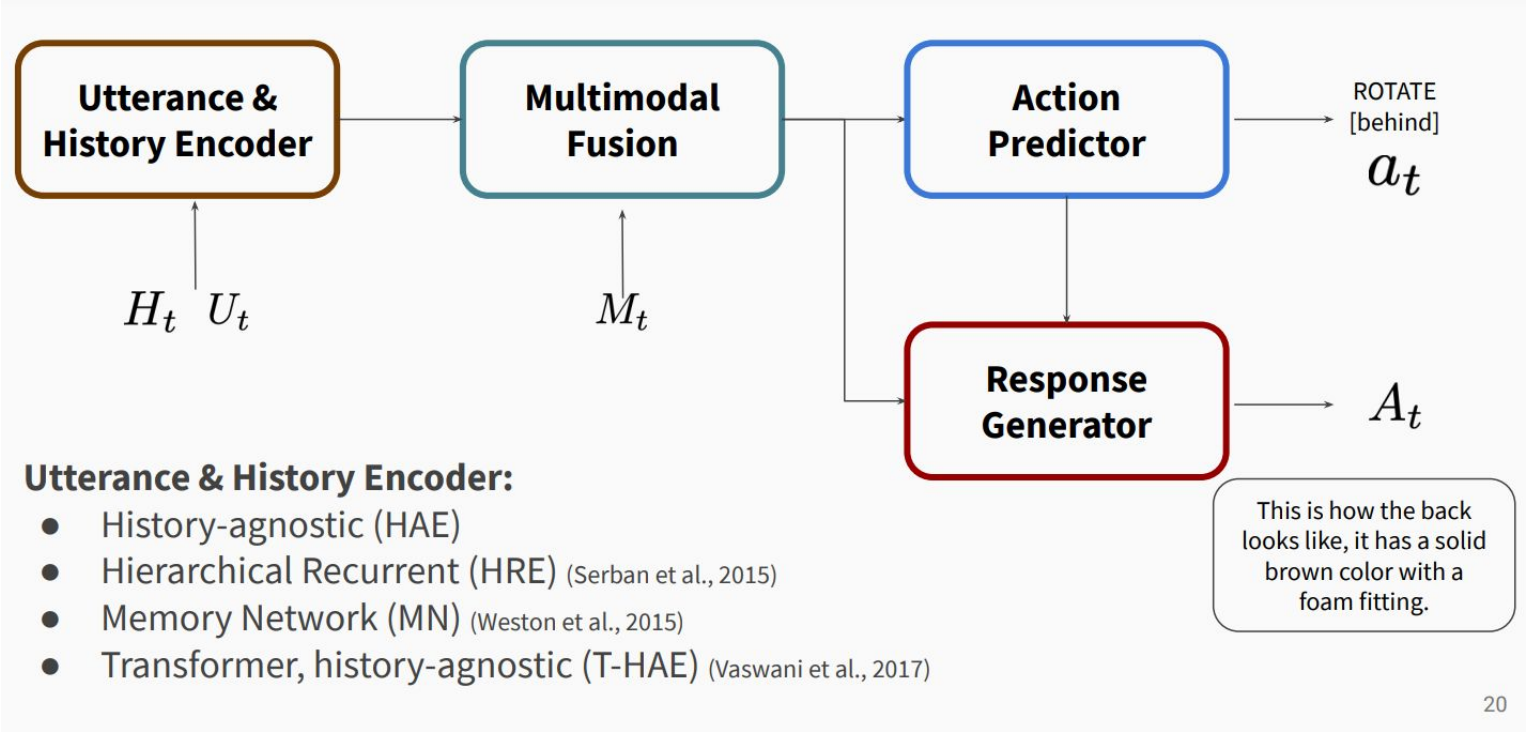
Task 3: MM-Dialog State Tracking

Intent and Slot F1

DSTC 9. Situated Interactive Multimodal Conversations (SIMMC)



DSTC 9. Situated Interactive Multimodal Conversations (SIMMC)



DSTC 9. Situated Interactive Multimodal Conversations (SIMMC)

1. History-Agnostic Encoder (HAE) that ignores dialogue context H_t and encodes only the user utterance through a LSTM for downstream components.
2. Hierarchical Recurrent Encoder (HRE) (Serban et al. 2016) that models dialogues at two hierarchical recurrence levels of utterance and turn.
3. Memory Network (MN) encoder (Sukhbaatar et al. 2015) that treats dialogue history H_t as a collection of memory units and then selectively attends to them with the current utterance u_t.
4. Transformer-based History-Agnostic Encoder (T-HAE) which is similar to the HAE except with Transformer units (Vaswani et al. 2017) instead of LSTMs.
5. Pretrained GPT2 model for action prediction and dialog state tracking.

DSTC 9. Joint Generation and Bi-Encoder for Situated Interactive MultiModal Conversations

- Task: Given dialog history H_t, multimodal context M_t, and current user utterance U_t, jointly predict the assistant API action Act and arguments Act_{arg} (Sub-Task #1), system response R (Sub-Task #2 Response Generation), a list of belief state actions $I^{(1)}, \ldots, I^{(N)}$ and associated belief state slots (Sub-Task #3).
- Source input string $X^{(i)}$ at i-th turn:

$$X^{(i)} = \left[U^{(0)}, \cdots, U^{(i-1)}, U^{(i)}, m_{ctx}^{(i)}\right]$$

- Target output string:

$$Act\ [Act_{arg}]\ [SEP1]\ I^{(1)}\ [I_{arg}^{(1)}], \ldots, I^{(i)}\ [I_{arg}^{(i)}]\ [SEP2]\ R$$

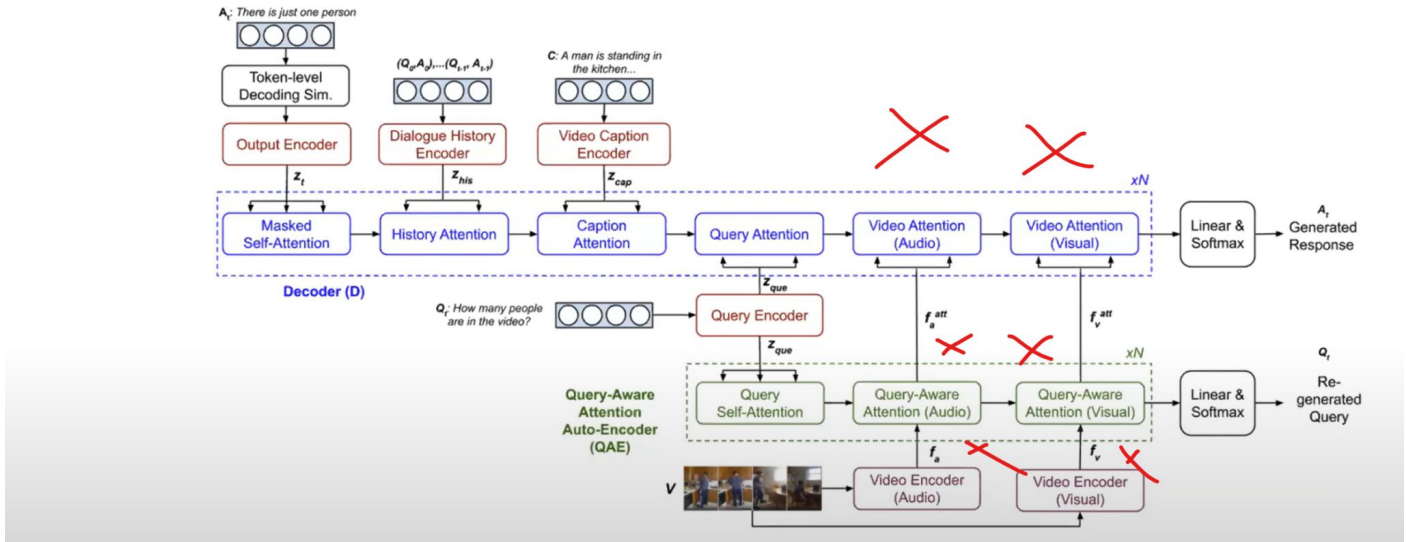
Where $[SEP1]$ and $[SEP2]$ are two special tokens for the target output string.

1st place

DSTC 9. Joint Generation and Bi-Encoder for Situated Interactive MultiModal Conversations

Joint Generation of Sub-Task 1-3

Multimodal Transformer Networks (MTN):



1st place

DSTC 9. Joint Generation and Bi-Encoder for Situated Interactive MultiModal Conversations

- **Bert2Bert** (Rothe, Narayan, and Severyn 2020): The encoder is a pre-trained BERT model which consists of 12 layers and 12 attention heads, where hidden size is 768. The decoder is another pre-trained BERT model of the same size as the encoder, but its weights are not synchronized with those of the encoder and are fine-tuned independently from the encoder. In addition, the decoder includes a cross-attention sub-layer.
- **Bert2Share** (Rothe, Narayan, and Severyn 2020): The encoder and the decoder share the same pre-trained BERT model and its weights, while the decoder additionally has a cross-attention sub-layer.
- **Bert2GPT2** (Rothe, Narayan, and Severyn 2020): The encoder is a pre-trained BERT model. The decoder is a pre-trained GPT2 model which consists of 12 layers, 12
- **Bert2DistilGPT2** (Rothe, Narayan, and Severyn 2020): The encoder is a pre-trained BERT model. The decoder is a pre-trained distilled version of GPT2 model which consists of 6 layers, 12 attention heads and a cross-attention sub-layer, where hidden size is 768.
- **MTN** (Le et al. 2019): The encoder is a text sequence encoder of Multimodal Transformer Network (MTN) model (Le et al. 2019). The decoder is an auto-regressive decoder of the MTN model. They are not pre-trained, but trained from scratch with the SIMMC data. We do not use the video and caption encoders from the MTN model since the SIMMC datasets have only text data.
- **SimpleTOD** (Hosseini-Asl et al. 2020): SimpleTOD uses a single, causal language model jointly trained end-to-end on all generation sub-tasks as a single sequence prediction problem. The underlying pre-trained model used is GPT2.
- **BART** (Lewis et al. 2020): The encoder and the decoder are those of the BART model (Lewis et al. 2020), but the loss of fine-tuning the model for the SIMMC datasets is calculated only based on the errors of the next token prediction task, not involving the reconstruction loss of the denoising step of the original BART model. We use both the pre-trained BART-base and BART-large models for the experiments.

1st place

DSTC 9. Joint Generation and Bi-Encoder for Situated Interactive MultiModal Conversations

Source Text Encoder The sequence of concatenated source text $X^{(i)} = x_1, ..., x_i$ that combines a user utterance $U^{(i)}$ at the i-th turn, a dialogue history $U^{(0)}, ..., U^{(i-1)}$ at previous turns and multimodal contexts $m_{ctx}^{(i)}$ is fed into the *encoder* of a model $(Encoder)$ to get the last layer hidden representation of the input texts, $h_{src}^{(i)}$:

$$X^{(i)} = \left[U^{(0)}, \cdots, U^{(i-1)}, U^{(i)}, m_{ctx}^{(i)}\right] \quad (1)$$

$$h_{src}^{(i)} = Encoder\left(X^{(i)}\right) \quad (2)$$

Target Text Decoder Given the hidden representation output of the encoder $h_{src}^{(i)}$, the goal of the decoder $(Decoder)$ is to generate the next token (y_m) given previous target text sequence $Y_{m-1}^{(i)} = y_1, ..., y_{m-1}$ which includes the information of assistant API action Act, Act_{arg}, system response R, and a list of belief states $I^{(1)} I_{arg}^{(1)}, ..., I^{(N)} I_{arg}^{(N)}$. We assume that the $Decoder$ has an innate multi-head cross-attention layer $CrossAttn$, where the hidden representation of the previous target string $h_{tgt}^{(i)}$ attends to the hidden representation of the encoder's last layer $h_{src}^{(i)}$, as follows (Vaswani et al. 2017):

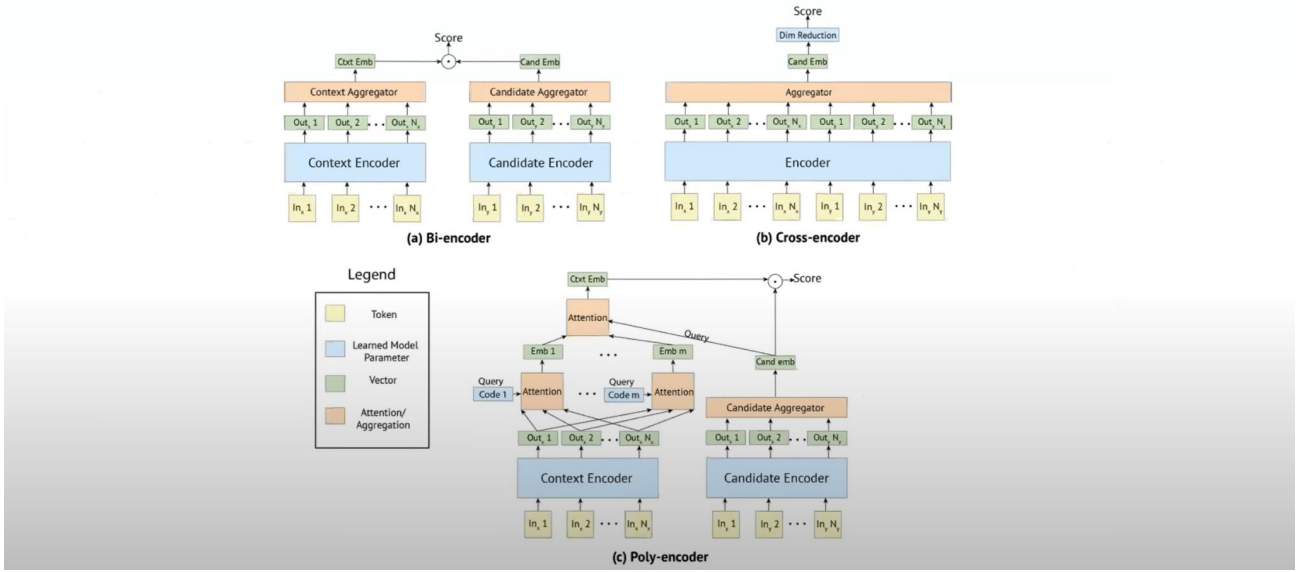
$$h_{tgt}^{(i)} = Decoder\left(Y_{m-1}^{(i)}\right) \quad (3)$$

$$h'^{(i)}_{tgt} = CrossAttn\left(h_{tgt}^{(i)}, h_{src}^{(i)}\right) \quad (4)$$

The output of the attention layer $h'^{(i)}_{tgt}$ is used to proceed the decoding process. We apply a linear layer for language modeling after the last layer of the decoder to get probabilities for the next token. A cross-entropy loss is used for fine-tuning the model.

Sub-Task 2 – Response Retrieval

Task: Given dialog history, multimodal context, and current utterance, select the best response from a list of 100 candidates



1st place

DSTC 9. Situated Interactive Multimodal Conversations (SIMMC)

Outcomes:

1. Ребята с 1-го места сразу перебежали бэйзлайн, тем что не стали тренировать кучу параллельных моделей для каждого сабтаска, мол в таком случае *model couldn’t capture semantics relationships between the different outputs*
2. Pretrained модели конечно лучше чем тренировать MTN с нуля
3. Cross-attention интересная штука, надо разобраться

DSTC 10

DSTC 10

- SIMMC 2.0: Situated Interactive Multimodal Conversational AI
  https://github.com/facebookresearch/simmc2

- Reasoning for Audio Visual Scene-Aware Dialog
  https://github.com/dialogtekgeek/AVSD-DSTC10_Official

- Automatic Evaluation and Moderation of Open-domain Dialogue Systems
  https://chateval.org/dstc10 (!!!)

DSTC 10. SIMMC 2.0: Situated Interactive Multimodal Conversational AI

Task Name	Goal	Evaluation
1. Multimodal Disambiguation	Given user utterances, classify if the assistant should disambiguate in the next turn.	Binary classification accuracy
2. Multimodal Coreference Resolution (MM-Coref)	Given user utterances with object mentions, resolve referent objects to their canonical ID(s) as defined by the catalog.	Coref Precision / Recall / F1
3. Multimodal Dialog State Tracking (MM-DST)	Given user utterances, track user belief states across multiple turns.	Intent Accuracy, Slot Precision / Recall / F1
4. Response Generation	Given user utterances, ground-truth APIs and ground-truth object IDs, generate Assistant responses or retrieve from a candidate pool.	Generation: BLEU; Retrieval: Accuracy@k, mean reciprocal rank, mean rank

Table 4: Proposed tasks and descriptions on our SIMMC 2.0 dataset. Please see Sec. 4 for more details.

Total # dialogs	11,244
Total # utterances	117,236
Total # scene snapshots	1566
Avg # words per user turns	12
Avg # words per assistant turns	13.7
Avg # utterances per dialog	10.4
Avg # objects mentioned per dialog	4.7
Avg # objects in scene per dialog	19.7

Table 1: **SIMMC 2.0 Dataset Statistics**

DSTC 10. SIMMC 2.0: Situated Interactive Multimodal Conversational AI

TASK 1. Multimodal Disambiguation.

It is important for the system to recognize ambiguous uses of such coreferences even before attempting to resolve them. For example:

‘A: The blue trousers are priced at $45. U: What about those?’, where the phrase those could be ambiguous in the following situations:

(a) The user refers to a group of trousers without specifying the exact one they have in mind,
(b) The user incorrectly uses a shorthand for a novel pair of trousers not mentioned in the dialog due to conversational brevity.

In either cases, identifying the need for disambiguation and responding with:
‘Which ones are you talking about? The red or the green pair?’ is a desirable trait for a robust assistant system. The multimodal disambiguation task tests this ability of the agent.

More concretely, given the dialog history and the current user utterance, multimodal disambiguation requires the agent to predict a binary label conditioned on the multimodal context, to indicate the ***presence of a referential ambiguity in the user utterance.***

DSTC 10. SIMMC 2.0: Situated Interactive Multimodal Conversational AI

TASK 2. Multimodal Coreference Resolution.

For this task, we aim to resolve referential mentions in user utterances to their canonical object IDs as defined for each scene. These mentions can be resolved through the dialog context

A: 'This shirt comes in XL and is $29.' → U: 'Please add it to cart.'

or the multimodal context:

U: 'How much is that red shirt?' or both ***U: 'How much is the one next to the one you mentioned?'***

The input for this task includes the ground-truth bounding boxes defining each object ID, to avoid the performance bottleneck by the object detection algorithms. The main evaluation metric includes F1, precision and recall performance.

DSTC 10. SIMMC 2.0: Situated Interactive Multimodal Conversational AI

TASK 3. Multimodal Dialog State Tracking.

We propose multimodal dialog state tracking (MM-DST), slots are grounded on the coexisting multimodal context, which requires handling of multimodal objects (as opposed to textual tokens) as part of dialog states.

The underlying reasoning behind this task is that the MM-DST labels will be able to provide sufficient information for a multimodal dialog system to carry out dialog policies and actions, given the detected and resolved items in each multimodal scene. Therefore, the MM-DST task measures the model’s holistic understanding of user requests throughout each dialog, including the disambiguation needs as well as the coreferences.

DSTC 10. SIMMC 2.0: Situated Interactive Multimodal Conversational AI

TASK 4. Assistant Response Generation.

The goal of this task is to generate assistant responses or retrieve from a candidate pool, given user utterances, ground-truth belief state, and object IDs. While we assume the assistant agent has the ground-truth meta information on each object, each response needs to naturally describe the referent objects as observed and understood by the user through the co-observed scene or the dialog context:

INFORM:RECOMMEND (OBJ_ID: 3) → A: "I recommend the blue shirt directly behind the brown jacket.".

We propose two ways to evaluate the performance of systems for response generation:

(a) As a generation task, where the agent is seen as conditional language model. Performance is measured using BLEU-4 score between the generated response and the ground truth response provided with the dataset.

(b) As a retrieval task, where the agent has to pick the ground truth response from a list of candidate responses (generated randomly; unique to each utterance). We use traditional information retrieval metrics like recall@k (k = {1, 5, 10}), mean rank, and mean reciprocal rank for comparing model performances.

DSTC 10. SIMMC 2.0: Situated Interactive Multimodal Conversational AI

Baselines

We benchmark the dataset by adopting:

(a) We train a ***multi-task GPT-2*** (Radford et al., 2019) based Transformer model using the joint supervision signals for the Disambiguation, MM-Coref, DST, and Response Generation tasks. Specifically, the model takes as input the dialog context and the flattened multimodal contexts (as structurally formatted strings) to predict the belief states and the responses, following the popular causal language model approach (Peng et al., 2020; Hosseini-Asl et al., 2020).
We use the 12-layer GPT-2 (117M parameters) as the pre-trained language model and fine-tune for ten epochs. Note that this baseline uses the ground-truth multimodal contexts provided from the scene generator, instead of consuming raw images as input, and thus serves as a soft oracle on the proposed dataset.

(b) ***Multimodal Transformer Network (MTN)*** (Le et al., 2019) for the DST and Response Generation tasks. In particular, MTN uses image features extracted from scene snapshots and attends to relevant parts as guided by the dialog. We use the same training setting and hyperparameters as Le et al. (2019)

DSTC 10. SIMMC 2.0: Situated Interactive Multimodal Conversational AI

1. Disamb.	2. MM-Coref	3. DST		4. Gen.
Acc↑	Coref F1↑	Slot F1↑	Intent F1↑	BLEU↑
73.9±1.2	36.64±0.58	81.72±0.51	94.53±0.36	0.192±0.002
-	-	74.75±0.42	93.40±0.26	0.217±0.002

Table 5: Baseline performances: Moon et al. (2020) (top), Le et al. (2019) (bottom). **(1) Multimodal Disambiguation (Disamb.)**, via classification accuracy, **(2) Multimodal Coreference Resolution (MM-Coref)**, via coref prediction F1, **(3) Dialog State Tracking (DST)**, via slot and intent F1, **(4) Response Generation** via BLEU. ↑: higher is better.

DSTC 9. Multimodal Interactions Using Pretrained Unimodal Models for SIMMC 2.0

3 Method

In this section, we describe an approach to solving each subtask. First, we pretrain the multimodal model by image-to-text matching to mutually understand the text and image representations before finetuning the subtasks. Then, we train a new model for each subtask based on the pretrained multimodal model. Table 1 shows the notations used in the paper.

Symbol	Description
i_t	input of the text model (RoBERTa or GPT2)
i_v	input of the image model (DeIT)
t_h^k	output of text model superscript: the kth token or special tokens ([CLS] or [SYSTEM])
o_h^k	output of the image model for the object image superscript: the kth patch or a special patch ([CLS])
b_h^k	output of the image model for the background image
u_k^t	kth utterance superscript (t): user or system
M	matching function
W_k	matrix of the output layer of subtask #k (k: 1 or 2 or 4)

Table 1: Notations used in papers

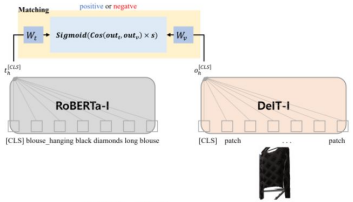
DSTC 9. Multimodal Interactions Using Pretrained Unimodal Models for SIMMC 2.0

Types of subtaks

1. Multimodal Disambiguation
2. Multimodal Conference Resolution (MM-Coref)
3. Multimodal Dialog State Tracking (MM-DST) (not participating)
4. Multimodal Dialog Response Generation & Retrieval (Participate only in generation)

Multimodal pre-training

1. ITM (Image-to-Text Matching)
- Input: object description ↔ cropped object
- Text Model: RoBERTa
- Image Model: DeIT
- Matching score: BCE Loss, s(scale size): 100



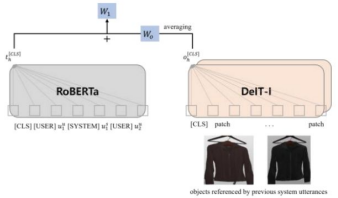
$$score = \mathrm{M}(t_h^{[CLS]}, o_h^{[CLS]}) = sigmoid(cos(W_t(t_h^{[CLS]}), W_v(o_h^{[CLS]})) \times s)$$

2. BTM (Background-to-Text Matching)
- Input: background Image ↔ dialog context
- Learning the same way as ITM

DSTC 9. Multimodal Interactions Using Pretrained Unimodal Models for SIMMC 2.0

Subtask #1, #2

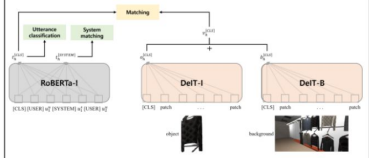
#1. Multimodal Disambiguation
- Input: dialog context + cropped object
- Model: RoBERTa (scratch), DeIT-I (ITM)



#2. Multimodal Conference Resolution
- Input: dialog context ↔ cropped object + background
- Model: RoBERTa-I, DeIT-I, DeIT-B
- Two multitask learning

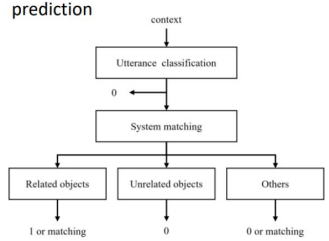
1) Utterance Classification
- Whether there is an object to match

2) System Matching
- Classification of previous system utterances that referred to the same object as the current utterance

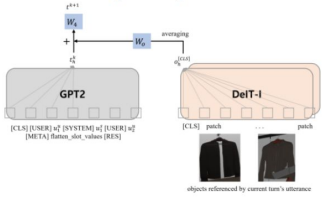


Subtask #2, #4-1

#2. Multimodal Conference Resolution
- Inference process using multitask prediction



#4-1. Multimodal Dialog Response Generation
- Input: dialog context + slot_values + cropped object
- Model: GPT2 (scratch) + DeIT-I



Results

- Subtask #1 (Disambiguation)

	Accuracy (devtest)	Accuracy (teststd)
Baseline (GPT2)	73.9	73.5
Ours	**92.28**	**93.1**

- Subtask #2 (MM-Coref)

	F1 Score (devtest)	F1 Score (teststd)
Baseline (GPT2)	36.6	44.1
Baseline (all S)	32.2	-
Ours+M	59.5	63.4
-ITM	57.9	-
-BTM	59.4	-
-D	58.3	-
-U	56.8	-
-SU	53.5	-
Ours (only S)	60.4	63
Ours-M	60.7	66.7
Ours (+M & -M)	**60.8**	**68.2**

- Subtask #4-1 (Generation)

	BLEU-4 (devtest)	BLEU-4 (teststd)
Baseline (GPT2)	0.192	0.202
Baseline (MTN)	0.217	0.211
Ours	**0.285**	**0.297**
-O	0.275	-
-M	0.219	-
-D	0.281	-

Conclusion

- We propose a multimodal model based on the unimodal models of RoBERTa and DeIT.
- Pretraining strategies called ITM,BTM.
- Two multitask learning in subtask #2
- 3rd best performance in subtask #1, #2, and runner-up in subtask #4-1.

DSTC 10. Reasoning for Audio Visual Scene-Aware Dialog

Reasoning for Audio Visual Scene-Aware Dialog Track in DSTC10

Shijie Geng[†], Peng Gao[‡], Anoop Cherian[◇] Chiori Hori[◇], Tim K. Marks[◇]

[†]Rutgers University, Piscataway, NJ, USA
[‡]The Chinese University of Hong Kong
[◇]Mitsubishi Electric Research Laboratories (MERL), Cambridge, MA, USA
{chori, cherian}@merl.com

Abstract

To build a dialog system that can enable an agent to discuss audio visual scenes with humans, we held two challenges on Audio Visual Scene-Aware Dialog (AVSD) in DSTC7 and DSTC8. We defined the task and prepared the AVSD dataset, which was the first attempt to promote the combination of conversation systems and multimodal reasoning algorithms into a single framework. The task setup for the previous challenges allowed the participants to use human-created video captions to generate answers for the dialog questions. However, such manual descriptions are not available in real-world applications of an AVSD system and in such applications, the system needs to learn to produce the answers without the captions. To encourage progress towards this end, we propose a third challenge in DSTC10 under the video-based scene-aware dialog track. Our proposed task is to generate or select a system response to a query, where this query is part of a multi-turn dialog about a video. Specifically, the participants will use the video, its associated audio, and the dialog text to train end-to-end deep learning models to produce the answers, however unlike previous challenges, participants cannot use the video captions. We will continue to use the AVSD dataset, however we may also include additional datasets that are more attuned towards our task.

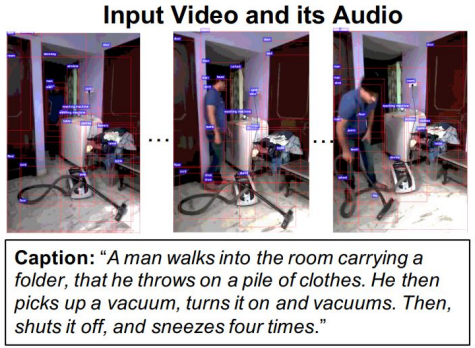


Figure 1: A sample dialog for the AVSD challenge data set. Given a video clip (including audio), its caption, dialog history, and a follow-up question, the AVSD generation task aims to generate an answer to the question in natural language form.

DSTC 10. Reasoning for Audio Visual Scene-Aware Dialog

3 Task definition

In this track, the system must generate responses to a user input in the context of a given dialog. The dialog context consists of a dialog history between the user and the system in addition to the video and audio information in the scene. Additionally, reasoning needs to be shown to support answers. There are two tasks, each with two versions (a and b):

Task 1: Video QAs Answer generation
(a) Using the provided video, QA and manual description training data, but no external data sources, other than publicly available pre-trained feature extraction models. However, manual descriptions for test set are not provided.
(b) Also using external data for training (e.g., How2 dataset [14])

Task 2: Grounding for QAs Grounding to support answers needs to shown without using manual descriptions. When a system answer "A dog is barking.", The *sound* of the dog's barking or the *image* of the dog should be provided/highlighted/attended as evidence.
(a) Using the provided video, QA and manual description training data and additional grounding data supporting answers. However, manual descriptions for test set are not provided. No external data sources, other than publicly available pre-trained feature extraction models
(b) Also using external data for training.

Challenge participants can select to submit entries in Task 1, Task 2, or both. The training data and a baseline system will be released to all participants of DSTC8.

DSTC 10. Automatic Evaluation and Moderation of Open-domain Dialogue Systems

2. Track Details

This track consists of two tasks:

1. Participants will develop effective automatic open-ended dialogue evaluation metrics that perform robustly across a range of dialogue evaluation tasks.
2. Participants will build generative models that can respond to toxic questions with appropriate and polite sentences.

DSTC 10. Automatic Evaluation and Moderation of Open-domain Dialogue Systems

2.1. Automatic Metrics Development

Effective automatic dialogue evaluation metrics possess the following two important properties as indicated in (Deriu et al., 2019):

- Correlated to human judgements - the metrics should produce evaluation scores that well correlate to human judgements (scores) across multiple dialogue evaluation aspects.
- Explainable - the metrics should provide constructive and explicit feedback to the generative models in terms of the quality of their generated responses. For instance, if a generative model is contradicting itself, the evaluation metrics should signal such behavior to the generative models.

In this task, our goal is to seek effective automatic dialogue evaluation metrics that exhibit the above properties. These metrics can serve as a proxy to human evaluation for fast prototyping of open-domain chatbots. We have identified the following already existing datasets to test the effectiveness of the proposed evaluation metrics[1]:

DSTC 10. Automatic Evaluation and Moderation of Open-domain Dialogue Systems

For this task a dataset of pairs of 100K messages (training and validation set) will be automatically collected with the following characteristics:

- A toxic user sends a Tweet message using one or several of the most common swear words found on the Internet. The Tweet message must be directed to one of the customer service channels.
- A toxic user writes a Tweet message using one or several swear words and the message is replied by another user.
- A toxic user posts a message in Reddit using one or several swear words and the message is replied by another user.

DSTC 10. MME-CRS: Multi-Metric Evaluation Based on Correlation Re-Scaling for Evaluating Open-Domain Dialogue

Methodology

Figure 1 shows the architecture of our proposed metric MME-CRS. In this section, we will first introduce 5 groups of sub-metrics in detail. Then score composition approach CRS is discussed to integrate sub-metric scores for diverse qualities.

Automatic Evaluation Metrics

The evaluation quality contains various aspects, such as fluency, relevancy, specificity, and user engagement. For example, sub-task1 of DSTC10 track5 contains 14 development datasets, of which 37 different qualities are included in the total. What's more, the evaluation of each aspect usually relies on several metrics, and the weight distribution over sub-metric varies from aspect to aspect. To better measure each evaluation aspect of dialogue, we design 5 groups of fundamental sub-metrics as follows.

Fluency Metric (FM) quantifies whether or not a response is fluency or understandable. A fluency utterance does not have to be grammatically correct because an open-domain response is usually the central part of a complete sentence. The auxiliary verb or stop words may be missing.

We use this characteristic to build a training set of fluent and non-fluent responses. First, we randomly determine if a response r is fluent. If it is, we assign response r with label one and randomly apply one of the following rules: (i) no modification, (ii) delete each stopword with a probability of 0.5. Otherwise, we label response r with zero and apply one of the following rules following Sinha et al. (2020) for negative sampling: (i) word reorder (shuffle the order of all words), (ii) word drop (randomly drop x% words), or (iii) words repeat (randomly select span(s) of words and randomly repeat them).

For a response r with $(w_1, w_2, ...w_n)$ words, we fine-tune SimCSE (Gao, Yao, and Chen 2021) to embed each word in r and apply Max-Pooling to get the utterance embedding. Then a Softmax layer is used to obtain the probability, and we use it as the fluency score S_F.

Relevance Metric (RM) measures coarse-grained relevance between context and response. We fine-tune another SimCSE model based on the next utterance prediction task to predict whether a context-response pair is relevant or not. Similar to the fluency metric, we first randomly determine a context-response pair from the Daily Dialog dataset (Li et al. 2017) is valid or not. For the valid case, we randomly apply one of the following changes to the response: (i) no modification, (ii) remove stop words.

Lan et al. (2020) observes that most random sampled negative responses are low-quality, and the decision boundary

DSTC 10. Automatic Evaluation and Moderation of Open-domain Dialogue Systems

learned is far from the actual decision boundary, which hurts the performance. Hence, for the invalid case, we propose a simple but effective negative response sampling method. First, we randomly choose ten responses from the response pool and compute the Word2Vec similarity (Mikolov et al. 2013) between reference and candidate responses. Then we sort candidate responses based on their similarity score and choose the middle one as a negative response.

To fine-tune the SimCSE model, we first concatenate a context-response pair to a single sentence. Then we compute the score S_R using the same approach as the fluency metric.

Topic Coherence Metric (TCM) measures fine-grained topic transition dynamics of dialogue flows. Huang et al. (2020) demonstrates the effectiveness of incorporation graph information into dialogue evaluation. Following Huang et al. (2020), topic-level dialogue graphs are firstly constructed based on ConceptNet (Speer, Chin, and Havasi 2017). The topic transition dynamics over topic-level dialogue graphs are modeled applying a graph neural network. Then the topic-level graph representation is fed into an MLP layer to predict topic coherence score S_T. Huang et al. (2020) also embeds the context-response pair and jointly predict coherence score together with topic-level graph representation. The former embedding is ignored in this part to focus on the topic coherence metric.

Engagement Metric (EM) measures whether the user is willing to participate in the dialogue. We build a training set based on the human engagement scores. User engagement score usually ranges from 0 to 5, and the user's enthusiasm is proportional to the engagement score. Ghazarian et al. (2020) propose to label response with engagement score less than two as zero, while we find that scaling the engagement score to between 0 and 1 yields more significant benefits.

We train an utterance-level engagement classifier to predict whether the user engagement is high or low. Specifically, for a response r with $(w_1, w_2, ..., w_n)$ words, we fine-tune SimCSE to get the contextual embedding h_i for each word w_i. We use average-pooling here to get the embedding of the whole response. Then an MLP layer followed by a Softmax layer is added to predict the engagement score S_E.

Ghazarian et al. (2020) aggregates the embedding of both context and response to predict the score of user engagement, while we observe that user engagement mainly relies on the model response. The relationship between dialogue context and response should be handled by relevancy metric or topic coherence metric.

Specificity Metric (SM) measures the model's ability to handle diverse words in complex open-domain talking context. We introduce the specificity metric here because some deep models tend to generate general or ambiguous answers.

Mehri and Eskenazi (2020b) uses a Roberta model (Liu et al. 2019) to compute the mask language model (MLM) task, while we use a more light SimCSE model following other proposed sub-metrics. Similar to (Phy, Zhao, and Aizawa 2020), we only use the response r with $(w_1, w_2, ..., w_n)$ words to compute specific score. In detail, we mask each response word w_i and predict negative log-likelihood (SM-LL) based on SimCSE-MLM. We also investigate negative cross-entropy (SM-NCE) and perplexity (SM-PPL) to further improve the effectiveness of specific metrics.

Correlation Re-Scaling Method

Instead of designing a score composition function for the overall aspect alone, we propose to compute weight distribution along designed sub-metrics for each evaluation aspect. The evaluation of each evaluation aspect usually relies on several designed sub-metric. For example, suppose an annotator thinks a response generated by the dialogue model is specific. In that case, he probably implies that the response is also fluent and relevant to the dialogue context. However, the designed specific metric is only trained to predict the specific score for a response. Hence, to better evaluate each dialogue aspect, we propose to model the relationship between designed sub-metrics and diverse evaluation qualities.

We propose a novel Correlation Re-Scaling (CRS) method to compute the weight distribution for each aspect. For a dialogue evaluation dataset D_i with $(q_1, q_2, ..., q_n)$ qualities, we first randomly sample 300 dialogues for Spearman correlation computation.

For each dialogue q_{ij} in dataset D_i, we compute fundamental sub-metric scores as S_{ijk}, where k is the number of sub-metrics. If Spearman correlation score S_{ijk} is less than 0, then the corresponding sub-metric is believed to have no contribution to dialogue quality q_{ij}; thus, S_{ijk} is simply set to 0. We treat correlation score S_{ijk} as the importance of the corresponding sub-metric to quality q_{ij}.

We believe that important sub-metrics should be given higher weight, it is significant for score composition over multiple scores. Hence we compute the normalized weight distribution w_{ijk} as follows:

$$w_{ijk} = \frac{S_{ijk}^{d_{ij}}}{\sum_k S_{ijk}^{d_{ij}}} \tag{1}$$

Where d_{ij} is the power number of S_{ijk}, and the assigned weight to S_{ijk} is $S_{ijk}^{d_{ij}-1}$. The larger d_{ij} is, the more weight is given to more important sub-metrics. According to our experiments, the effect of the score composition method works best when $max(S_{ijk})$ is between 1/3 and 1/2. It is a simple but effective way to determine the value of d_{ij}.

To further improve the generalization ability of CRS method, we calculate the average w_{ijk} of 14 development datasets as follows:

$$w_{jk} = \frac{1}{|D_{q_j}|} \sum_i w_{ijk} \tag{2}$$

Where $|D_{q_j}|$ is the number of development datasets that have q_j quality, and w_{jk} is the normalized weight distribution over each sub-metric for diverse qualities.

For each test dataset D_i, we first compute 7 kinds of sub-metric scores. Then the composition score for each evaluation quality q_{ij} can be computed as follows:

$$S_{ij} = \sum_k w_{jk} \cdot S_{ijk} \tag{3}$$

DSTC 11

DSTC 11

- Bridging the Gap between Databases and Dialogues: Task-oriented Dialogue Modeling with Interactive Databases, *Beijing University of Posts and Telecommunications & Meituan Group*
- Intent Induction from Conversations for Task-Oriented Dialogue, *AWS AI*
- Robust and Multilingual Automatic Evaluation Metrics for Open-Domain Dialogue Systems, *Universidad Politécnica de Madrid (UPM), National University of Singapore (NUS), New York University (NYU) & Carnegie Mellon University (CMU)*
- Speech-Aware Dialog Systems Technology Challenge, *Google Research*
- Cross-Lingual Task-Oriented Dialogue Agents, *Tianjin University, Stanford University, Hanyang University, Laboratory for Integration of Systems and Technology, Microsoft & International Institute of Information Technology*
- Ambiguous Candidate Identification and Coreference Resolution for Immersive Multimodal Conversations, *Meta AI & Meta Reality Labs*
- End-to-End Flowchart Grounded Task-Oriented Dialog Response Generation, *IBM Research, University of Massachusetts Amherst & Indian Institute of Technology*
- Task-oriented Conversational Modeling with Subjective Knowledge, *Amazon Alexa AI*

DSTC 10

Total # dialogs	11,244
Total # utterances	117,236
Total # scenes	3,133
Avg # words per user turns	12
Avg # words per assistant turns	13.7
Avg # utterances per dialog	10.4
Avg # objects mentioned per dialog	4.7
Avg # objects in scene per dialog	19.7
Avg # candidates per ambiguous turn	5.6

Table 2: **SIMMC 2.1 Dataset Statistics**

todos

1. Аугментации обозреть (text-attack, nlpaug)
2. Собрать все диалоговые мультимодальные датасеты (поискать)
3. модели mmdst
4. Какие таски доступны на 11 dstc
5. SemEval