Competitions overview

Summer Tracks

Track 1

Ambiguous Candidate Identification and Coreference Resolution for Immersive Multimodal Conversations

Task Proposal

Track 2

Intent Induction from Conversations for Task-Oriented Dialogue

Task Proposal

Track 3

Speech-Aware Dialog Systems Technology Challenge

Task Proposal

Winter Tracks

Track 4

Robust and Multilingual Automatic Evaluation Metrics for Open-Domain Dialogue Systems

Task Proposal

Track 5

Task-oriented Conversational Modeling with Subjective Knowledge

Task Proposal

Timelines

Summer:

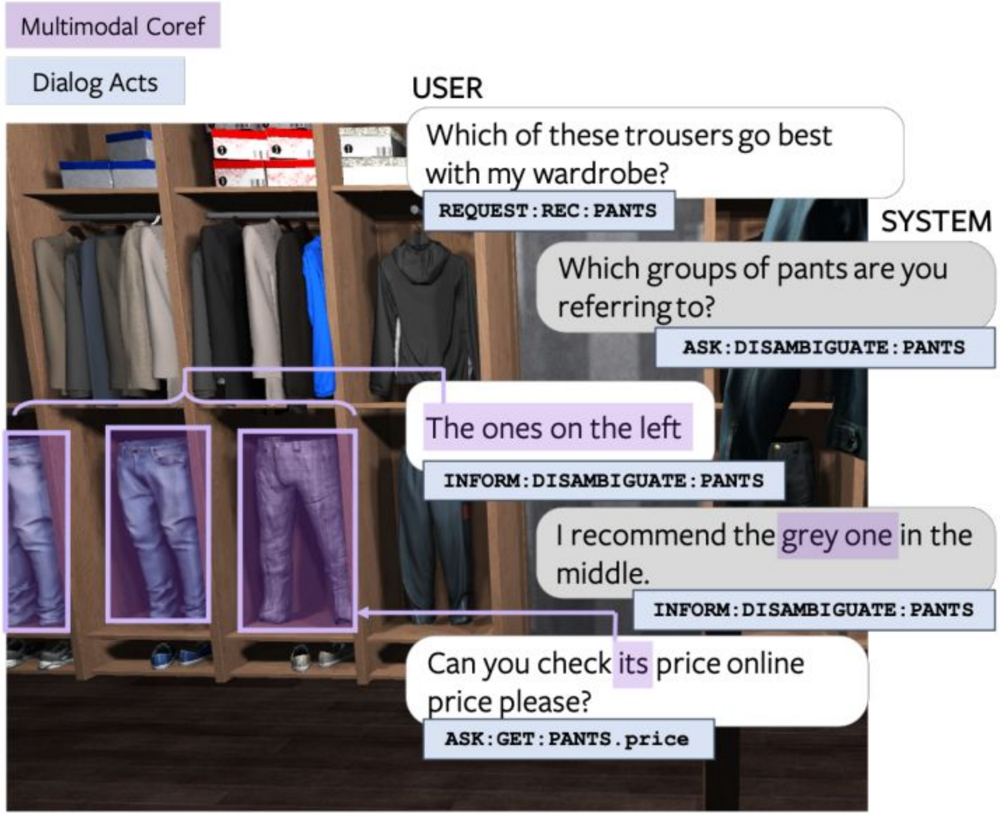
- **Training data release:** From end of May to middle of June
- **Test data release:** Middle of September
- **Submission of final results:** End of September
- **Final result announcement:** Early of October
- **Paper submission:** Middle of November
- **Workshop:** February or March in 2023

Winter:

- **Training data release:** From November to December
- **Test data release:** Middle of March in 2023
- **Submission of final results:** End of March in 2023
- **Final result announcement:** Early of April in 2023
- **Paper submission:** From March to May in 2023
- **Workshop:** July, August or September in 2023

SUMMER

RETRO: SIMMC II. ABOUT



(a) **SIMMC 2.0:** Cluttered, closer-to-real-world multimodal contexts

(b) **SIMMC 1.0:** Controlled and sanitized multimodal contexts

Figure 1: Illustration of a Situated Interactive Multimodal Conversation (SIMMC), which presents a task-oriented user↔assistant dialog grounded in a co-observed multimodal context. The newly collected SIMMC 2.0 dataset includes complex and photorealistic multimodal contexts, which poses more challenges for the Multimodal Coreference Resolution task (MM-Coref) and the Multimodal Dialog State Tracking task.

RETRO: SIMMC II. ABOUT

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 1: Proposed tasks and descriptions on SIMMC 2.0 dataset. Please see text for more details. Table: Kottur et al. (2021b)

<u>Disambiguation</u> ("grey jacket to the right of the one I mentioned") - is the task when we are trying to understand that words have different meanings based on the context of its usage in the sentence.

<u>Coreferences</u> ("directly behind it", "the one I mentioned") - is the task of finding all expressions that refer to the same entity in a text

RETRO: SIMMC II. ABOUT

We elaborate the multimodal context Mi in turn i as

$$M_i = (S_i, O_i, I_i)$$

S is a set of scene text descriptions, including the name, ID, bounding box and position of the objects appearing in the scene and their relationships

O is a set of object metadata, including name, brand, color and other attributes

I is a set of scene images

<u>Disambiguation </u>(“grey jacket to the right of the one I mentioned”) - is the task when we are trying to understand that words have different meanings based on the context of its usage in the sentence.

<u>Coreferences </u>(“directly behind it”, “the one I mentioned”) - is the task of finding all expressions that refer to the same entity in a text

RETRO: SIMMC II. ABOUT

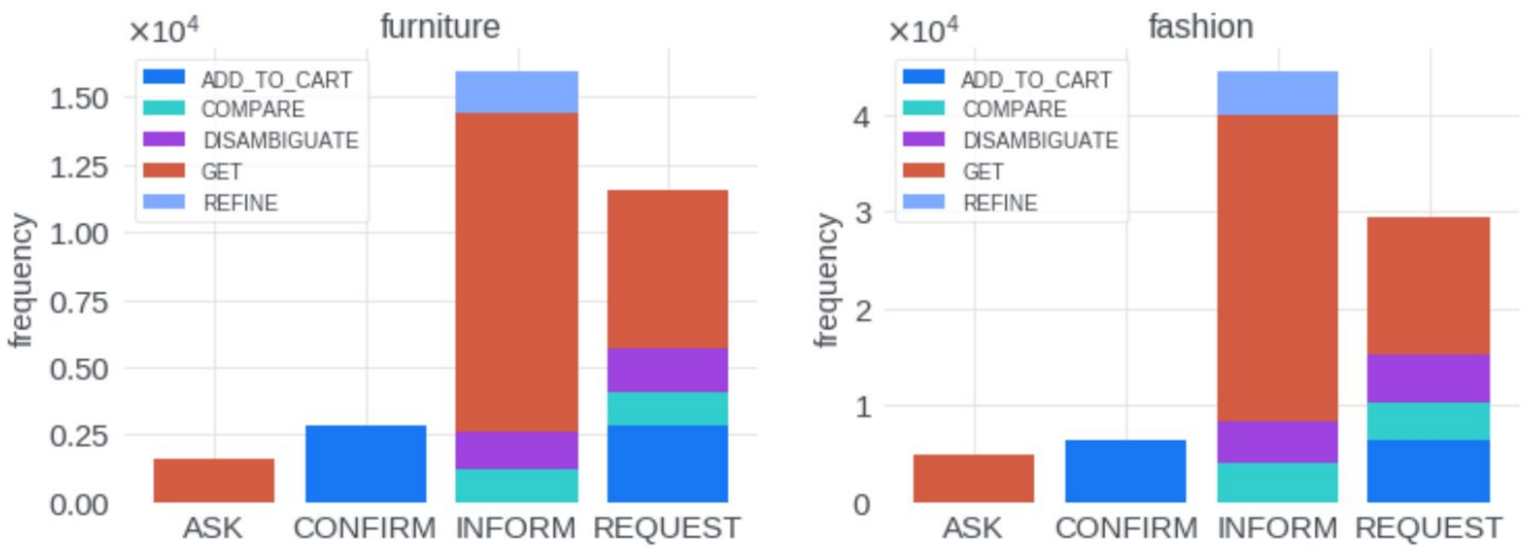


Figure 2: Distribution of dialog acts and activities for SIMMC 2.0. Figure from (Kottur et al. 2021b).

RETRO: SIMMC II. ABOUT

	UTTERANCES	ANNOTATIONS
USER:	Hey, could you show me some of your more expensive jackets?	`REQUEST:GET, slots: {type: jacket, price: expensive}`
ASSISTANT:	Sure, the grey and white jacket in the bottom left corner of the wall, the light blue jacket in the middle on the front table, and the grey jacket directly behind it are all in the expensive price range.	`INFORM:GET, objects: [11, 8, 0]`
USER:	What about something with good customer reviews in a size M?	`INFORM:REFINE, slots: {type: jacket, price: expensive, size: M, customerReview: good}`
ASSISTANT:	What do you think of the grey jacket to the right of the light blue one I mentioned? It is size M and has good customer reviews.	`INFORM:GET, objects: [9]`
USER:	I might do that one, but let's see some shirts first. Which of those on the wall do you think would best match my wardrobe?	`REQUEST:GET, slots: {type: shirt}`
ASSISTANT:	How do you like the grey shirt in the middle on the top row?	`INFORM:GET, objects: [3]`

RETRO: SIMMC II. RESULTS

Team ID	Org
0	Facebook
1	Sogang University
2*	UCLA
3	I2R A-STAR
4	KAIST, ETRI, Samsung Research
5	QS Goal Diggers
6	SKKU
7	Anonymous Submission
8	Kakao Enterprise
9	NYU
7	Heriot-Watt University

RETRO: SIMMC II. RESULTS

Subtask	Winner	Runner-up
#1: Disambiguation	Team 6 (SKKU)	Team 4 (KAIST, ETRI, Samsung Research), Team 5 (QS Goal Diggers), Team 10 (Heriot-Watt University)
#2: Multimodal Co-ref Resolution	Team 4 (KAIST, ETRI, Samsung Research)	Team 9 (NYU)
#3: Dialog State Tracking	Team 6 (SKKU)	Team 4 (KAIST, ETRI, Samsung Research)
#4-1: Assistant Response Generation	Team 5 (QS Goal Diggers), Team 10 (Heriot-Watt University)	Team 4 (KAIST, ETRI, Samsung Research), Team 8 (Kakao Enterprise)
#4-2: Assistant Response Retrieval	Team 4 (KAIST, ETRI, Samsung Research)	Team 3 (I2R A-STAR)

https://github.com/facebookresearch/simmc2/blob/main/CHALLENGE_RESULTS.md - REPO LINK 🔥🔥🔥

RETRO: SIMMC II. RESULTS

Team	Models	Joint (Pre)Train subtasks	Ens.	Language Model	MM Rep.	Subtask Ranks 1	2	3	4a	4b
Team 1	GPT-2 + Beam Search	2, 3, 4a, 4b	no	GPT-2 (large)	stringified	·	9	4	3	3
	GPT-2 + Beam Search	2, 3, 4a, 4b	no	GPT-2 (large)	stringified	·	9	4	4	3
Team 2	did not open-source	·	·	·	·	·	(1)	·	·	·
Team 3	BART+Poly-Encoder	1, 2, 3, 4b	no	BART	stringified	4	11	4	2	4
Team 4	BART	1, 2, 3, 4b	no	BART	object token	2	1	2	1	2
Team 5	BERT + ELECTRA	2, 3, 4a, 4b	no	BERT	stringified	2	8	3	5	1
	BERT + ELECTRA	2, 3, 4a, 4b	yes	BERT	stringified	·	·	·	·	1
Team 6	BART+ResNet (Ensemble)	2, 3, 4b	yes	BART	stringified (ResNet)	1	6	1	·	(2)
	BART+ResNet	2, 3, 4b	no	BART	stringified (ResNet)	1	·	·	·	·
Team 7	TOD-BERT	·	no	TOD-BERT	-	3	·	·	·	·
	LXMERT	·	no	LXMERT	LXMERT	·	7	·	·	·
Team 8	1: (UC+SM+MVM)	1, 2, 3, 4b	no	RoBERTa/GPT-2	DeIT	3	5	5	·	2
	2: (SM)	1, 2, 3, 4b	no	RoBERTa/GPT-2	DeIT	·	5	·	·	·
	3: (UC+SM+MVM w/o MI)	1, 2, 3, 4b	no	RoBERTa/GPT-2	DeIT	·	4	·	·	·
	1+3	1, 2, 3, 4b	no	RoBERTa/GPT-2	DeIT	·	3	·	·	·
Team 9	UNITER + Scene Graph	2b	no	UNITER	UNITER	·	2	·	·	·
Team 10	GLIMMeR	1, 2, 3, 4b	yes	GPT-2	stringified	·	10	·	·	·
	GLIMMeR (Ensemble)	1, 2, 3, 4b	yes	GPT-2	stringified	2	3	4	·	1

Table 4: **Summary of the developed models** for Subtasks 1, 2, 3, 4a and 4b.

RETRO: SIMMC II. EVALUATION

Overall, the track received 16 model entries across the world. Below are the full results as of **Oct 25, 2021** then edited to Dec 16, 2021.

Participant Team			#1. Disambiguation	#2. Coref	#3. DST		#4-1. Response Retrieval					#4-2. Response Generation
			Test-std	Test-std	Test-std		Test-std					Test-std
Team ID	Org	Label	Accuracy	Object F1	Slot F1	Intent F1	MRR	R@1	R@5	R@10	M. Rank	BLEU-4
0	Facebook	GPT-2	73.5 +- 1.2 %	44.1%	83.8%	94.1%						0.202
		MTN			76.7%	92.8%						0.211
1	Sogang University	Model 1		52.1%	88.3%	96.3%	53.5%	42.8%	65.4%	74.9%	11.0	0.285
		Model 2		51.9%	88.4%	96.3%	51.7%	41.2%	62.8%	72.5%	11.9	0.279
2	UCLA			78.3%								
3	I2R Singapore		89.5%	42.2%	87.8%	96.2%	61.2%	49.6%	74.7%	84.5%	6.6	0.256
4	KAIST, ETRI, and Samsung Research		93.9%	75.8%	90.3%	95.9%	81.5%	71.2%	95.0%	98.2%	1.9	0.295
5	QS Goal Diggers	Single	93.8%	56.4%	89.3%	96.4%	32.0%	19.9%	41.8%	61.2%	12.9	0.322
		Ensemble										0.322
6	SKKU	Single	94.7%	59.5%	91.5%	96.0%						
		Ensemble	94.5%									0.309*
7	Anonymous		93.1%	57.3%								
8	Kakao Enterprise	Sub 2-1	93.1%	63.4%	4.0%	41.4%						0.297
		Sub 2-2		63.0%								
		Sub 2-3		66.7%								
		Sub 2-4		68.2%								
9	New York University Shanghai			73.3%								
10	Heriot-Watt University	Baseline		50.6%								
		Ensemble	93.6%	68.2%	87.7%	95.8%						0.327

RETRO: SIMMC II. Sub-task 1 - Multimodal Disambiguation

Subtask 1: Fig. 4 shows the distribution of the disambiguation accuracy as turns of the dialog progress.

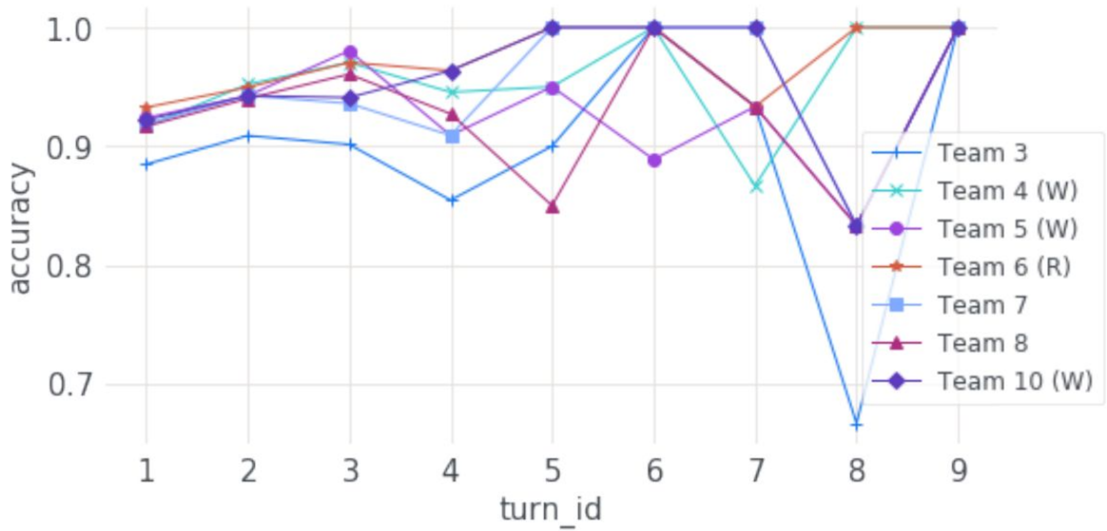


Figure 4: Distribution of disambiguation accuracy as dialog progresses.

RETRO: SIMMC II. Sub-task 1 - Multimodal Disambiguation. [1st place SKKU 🌿]

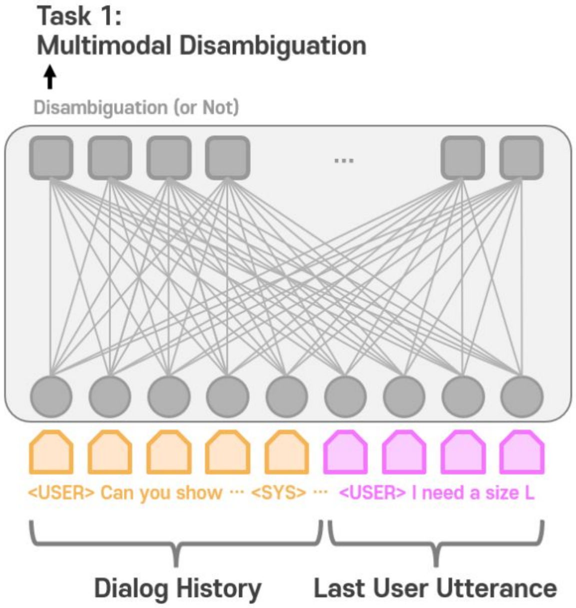


Figure 2: Overview of the proposed model for subtask 1. The dialogue context including the last user utterance is inputted in the encoder-based model, and the output [CLS] token embedding is used to classify the disambiguation from the dialog context.

Since SIMMC 2.0 consists of virtual shopping scenarios for fashion and furniture domains, it is a domain-specific dataset. However, in existing studies on SIMMC, no studies have focused on learning domain information. Therefore, we here leverage post-learning to learn domain-specific information of SIMMC and improve performance.

MLM is a learning method that randomly masks the tokens in the input data and predicts the masked token based on the context, and it is effective for learning the language representation. In this study, we used MLM based post-learning to reduce the learning bias of the existing pre-trained model that was trained only with articles and dictionary data; at the same time, we can add domain-specific information of the target data to our encoder-based model. We set the random masking probability of the MLM to 15%, and the loss of post-training was computed using the same method as the standard BERT.

RETRO: SIMMC II. Sub-task 2 - Multimodal Coreference Resolution [11th place A-STAR⭐]

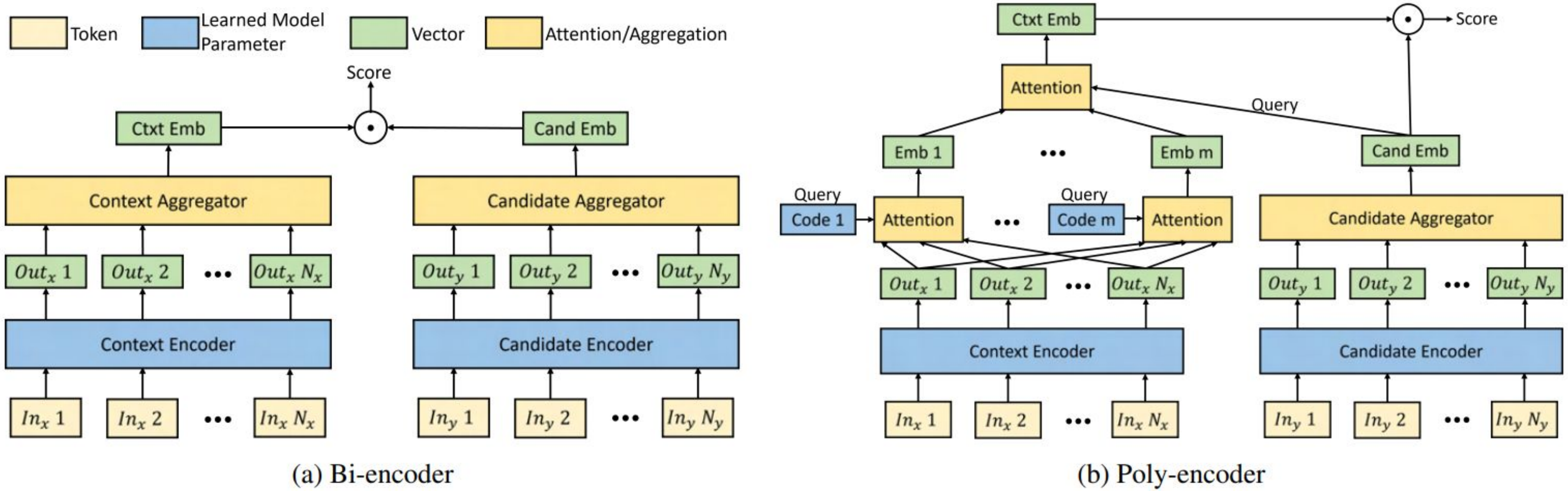
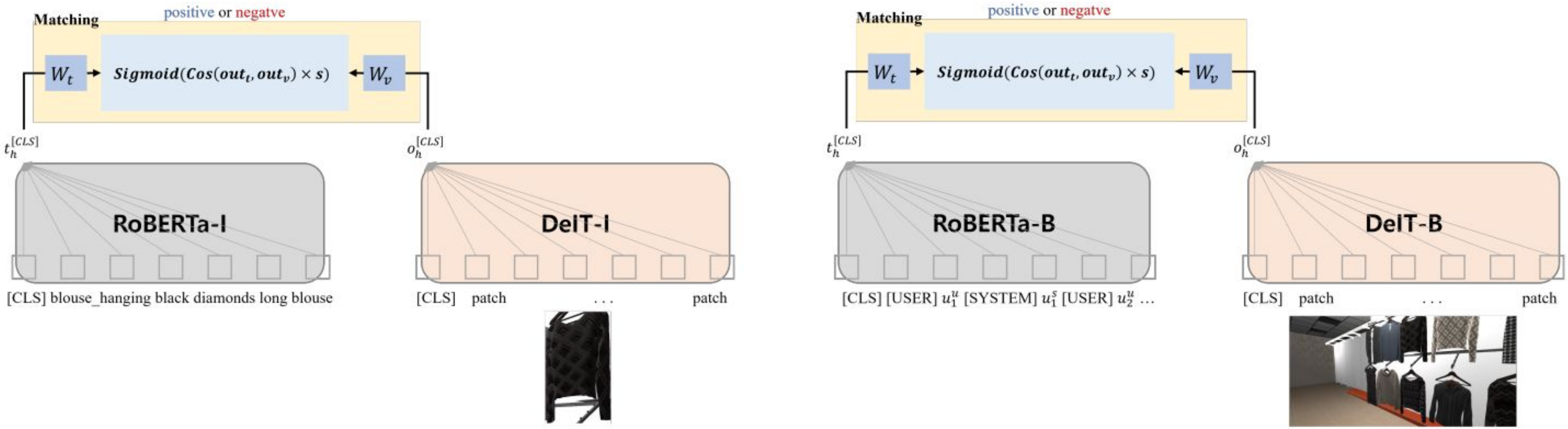


Figure 3: Bi-encoder & Poly-encoder Illustration. (a) Bi-encoder encodes context and candidate separately. (b) Poly-encoder also encodes context and candidate separately, but adds an attention between the global features of the input context and a given candidate to catch the interactions. Adapt from (Humeau et al. 2020)

RETRO: SIMMC II. Sub-task 2 - Multimodal Coreference Resolution [3rd place KAKAO☕Pretrain]



(a) The structure of Image-to-text matching (ITM).

(b) The structure of Background-to-text matching (BTM).

Figure 1: Multimodal pretraining for mutual understanding of representations between text and image.

RETRO: SIMMC II. Sub-task 2 - Multimodal Coreference Resolution [3rd place KAKAO☕ Fine-tune]

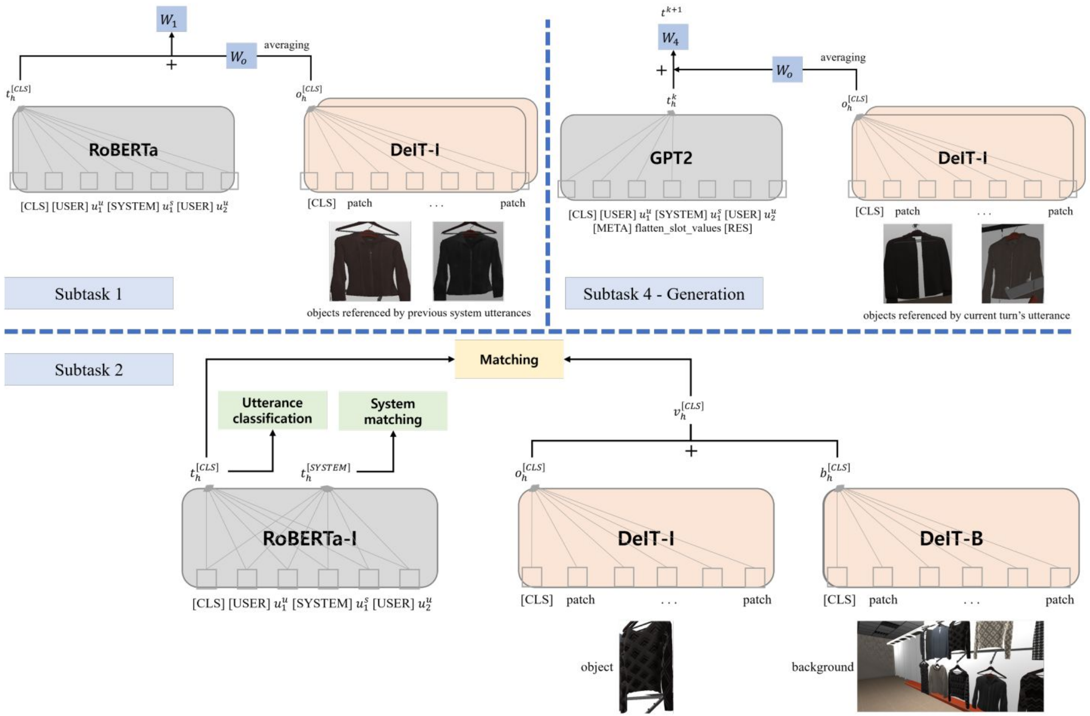


Figure 2: The proposed model architecture. The top left is subtask #1, the top right is subtask #4, and the bottom is subtask #2 architecture. Subtask #2 includes two multi-task learning.

RETRO: SIMMC II. Sub-task 2 - Multimodal Coreference Resolution [7th place HERIOT-WATT 🌄]

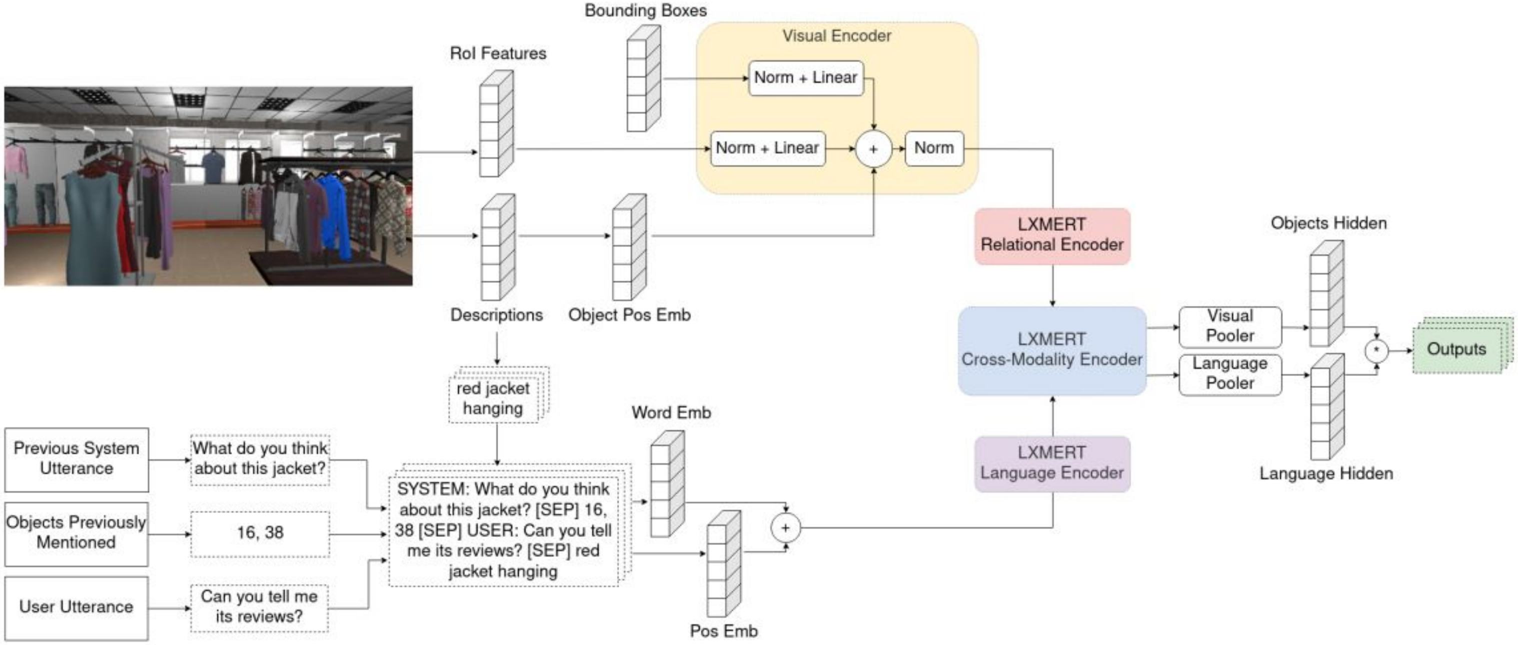


Figure 1: The architecture of the model submitted for Sub-Task #2. It uses vision to derive textual descriptions, and it combines the RoI features with language to obtain the final outputs. 'Norm' and 'Linear' stand for normalisation and linear layers.

RETRO: SIMMC II. Sub-task 2 - Multimodal Coreference Resolution [2nd place NYU 🗽 UNITER]

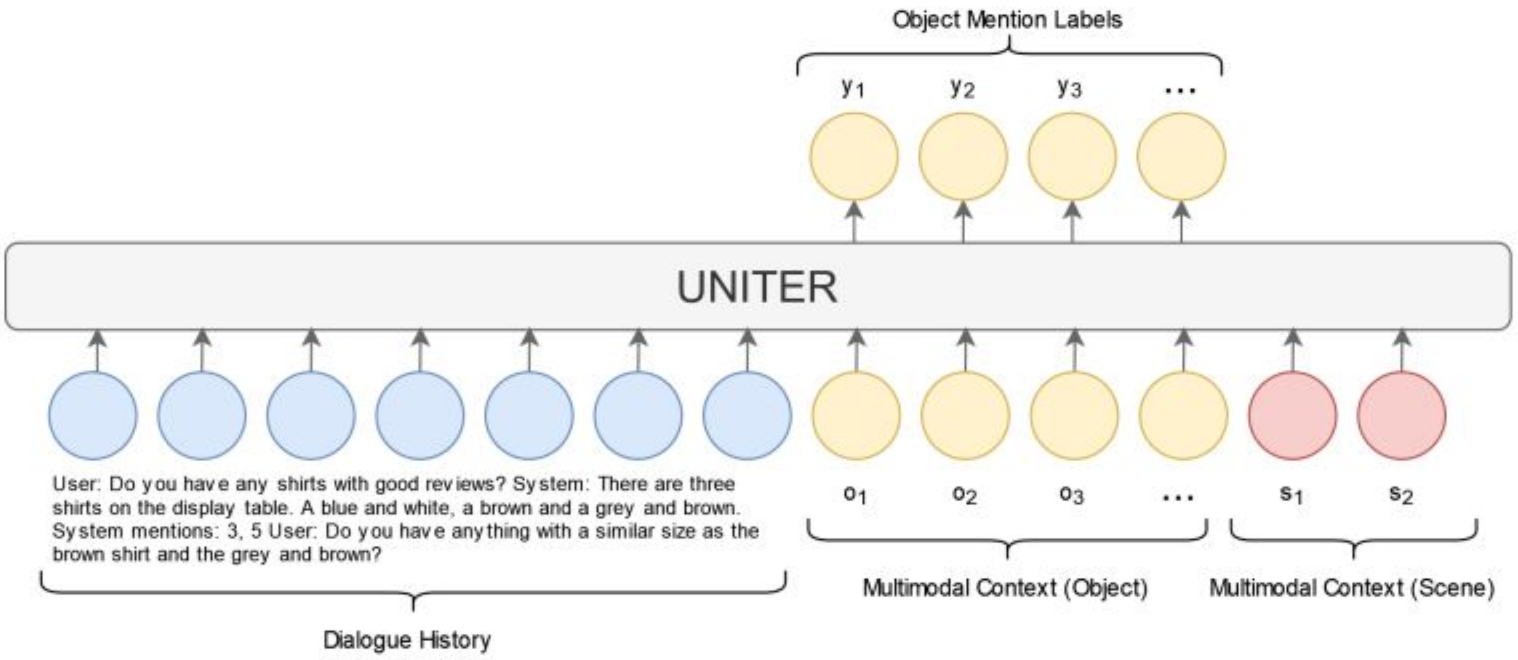


Figure 1: An overview of the proposed model.

RETRO: SIMMC II. Sub-task 2 - Multimodal Coreference Resolution [2nd place NYU 🗽 UNITER]

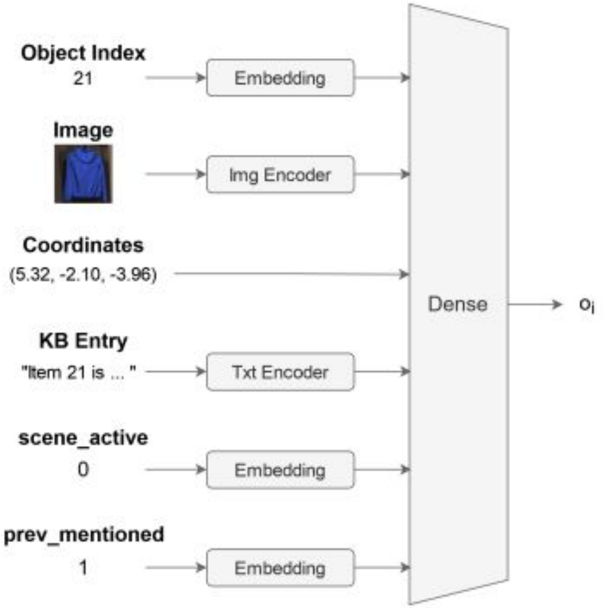


Figure 2: The separate embeddings of multimodal object features are concatenated and aggregated through a dense layer.

RETRO: SIMMC II. Sub-task 3 - Multimodal Dialog State Tracking [1st place SKKU 🌿]

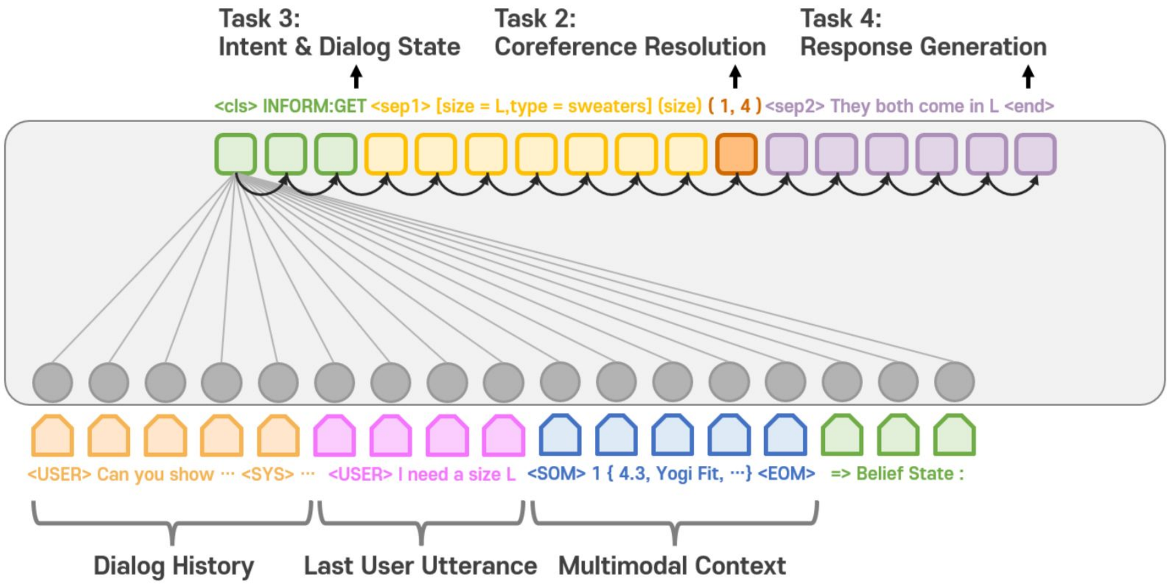


Figure 3: Overview of the encoder decoder-based model for subtasks 2-4. The model takes the dialog and multimodal context input and it should generate the output containing the user intention and states, object ID(s) that the user wants, and natural response. The special tokens for input are used to distinguish a user utterance and system utterance in the dialog context. And the special tokens for output are used to distinguish the different types of the generated output.

RETRO: SIMMC II. Sub-task 4.1 - Response Generation

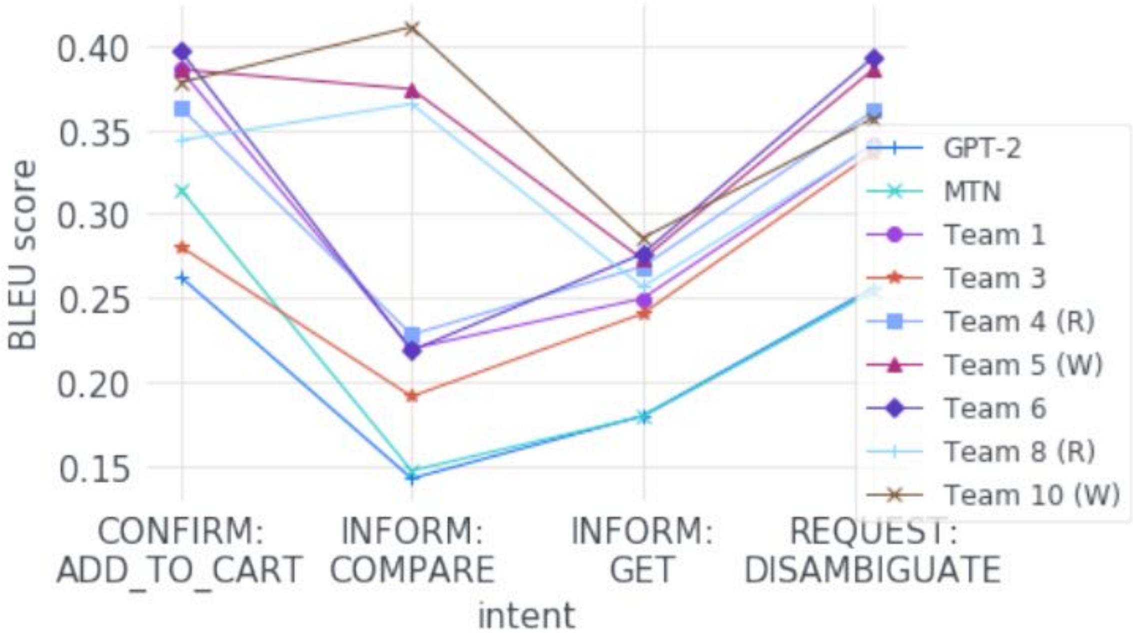


Figure 5: Distribution of BLEU score based on the natural language generation act.

RETRO: SIMMC II. Sub-task 4.1 - Response Generation [2nd place KAKAO☕]

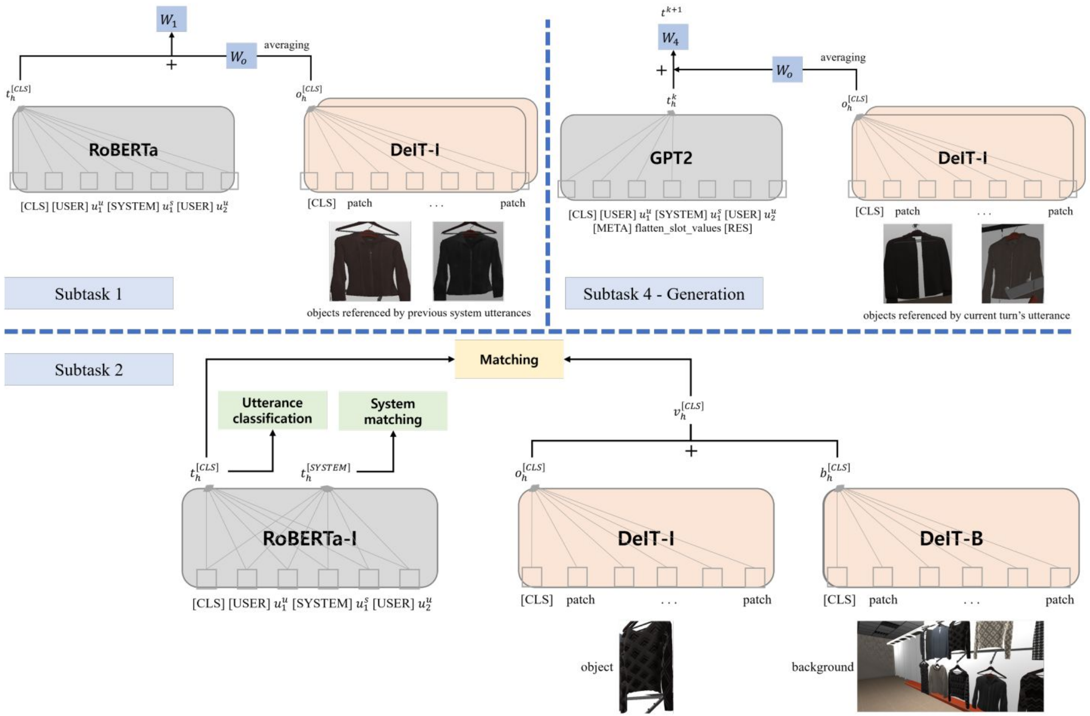


Figure 2: The proposed model architecture. The top left is subtask #1, the top right is subtask #4, and the bottom is subtask #2 architecture. Subtask #2 includes two multi-task learning.

RETRO: SIMMC II. Sub-task 4.2 - Response Retrieval [2nd place A-STAR⭐]

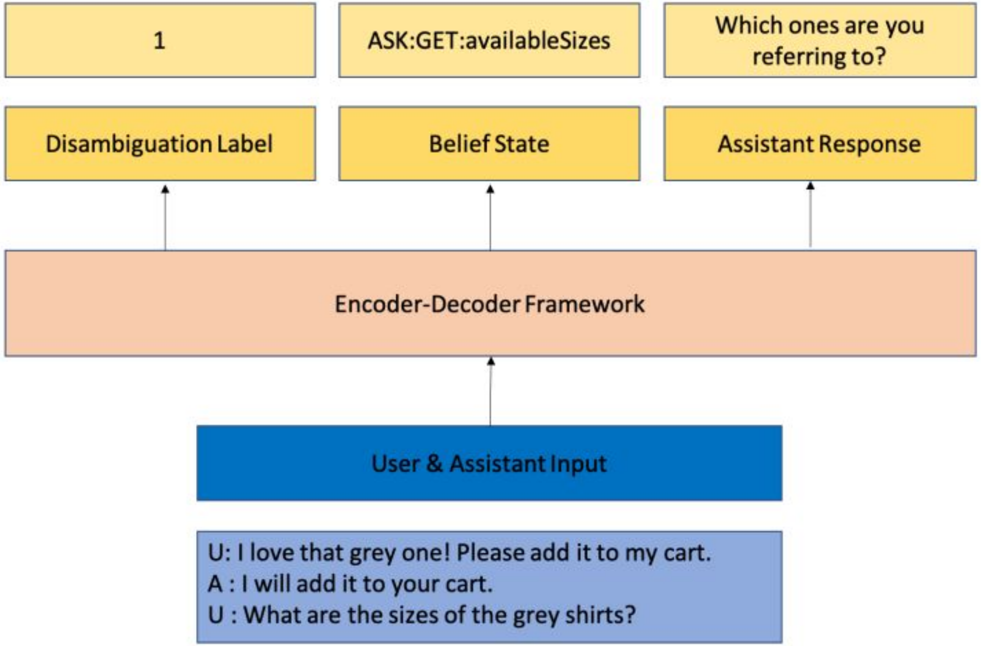


Figure 2: Illustration of generating the three sub-tasks' outputs in a single string.

RETRO: SIMMC II. INSIGHTS

1. There is a strong linguistic modeling vibe: almost all the performance comes from LLMs giving image processing techniques a room to add a pinch of cross-modal interaction
2. There are a lot of ways to interleave text and images, it could easily be a whiff if not to do it properly – the point is to use all available visual features
3. The thing common to all retros is multitask training pretends to be better than the independent one. Nevertheless, there could be raised an issue how to compose the tasks and what impact has one on the others
4. Multimodal context is heterogeneous, so it is necessary to set any relations and orders to rank entities by the criteria

DSTC 11. Ambiguous Candidate Identification and Coreference Resolution for Immersive Multimodal Conversations

Task Name	Goal	Evaluation
1. Ambiguous Candidate Identification	Given user utterances with ambiguous object mentions, resolve all referent candidate objects to their canonical ID(s) as defined by the catalog.	Object identification Precision / Recall / F1
2. Multimodal Coreference Resolution	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	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-4 score

Table 1: **Proposed tracks and descriptions.**

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**

WINTER

DSTC 11. Task-oriented Conversational Modeling with Subjective Knowledge

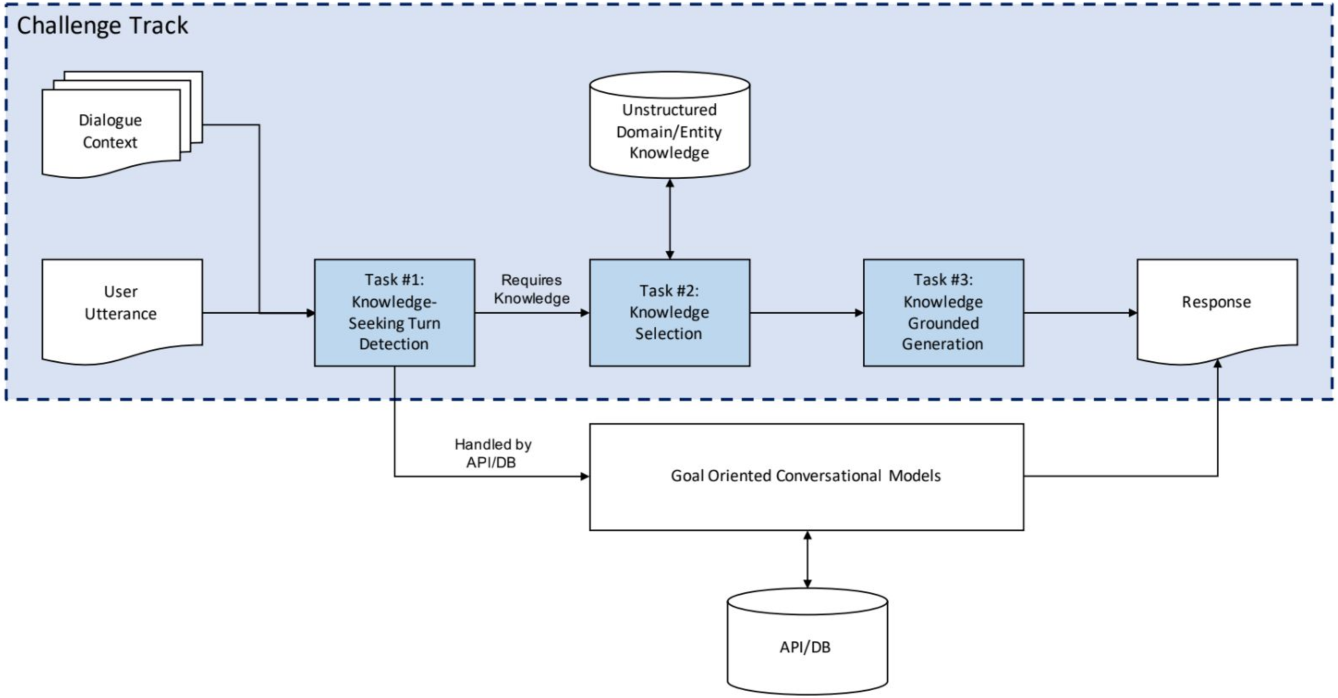


Figure 1: An overview of the challenge track that focuses on the knowledge access branch in the shaded box.

DSTC 11. Task-oriented Conversational Modeling with Subjective Knowledge

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 (including both FAQs and review posts)
Output	A ranked list of relavant knowledge candidates
Metrics	Precision/Recall/F-measure, 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*: BLEU, ROUGE, METEOR *Human evaluation*: appropriateness and accuracy to given knowledge

Table 1: Summary of the proposed tasks

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

RETRO 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)