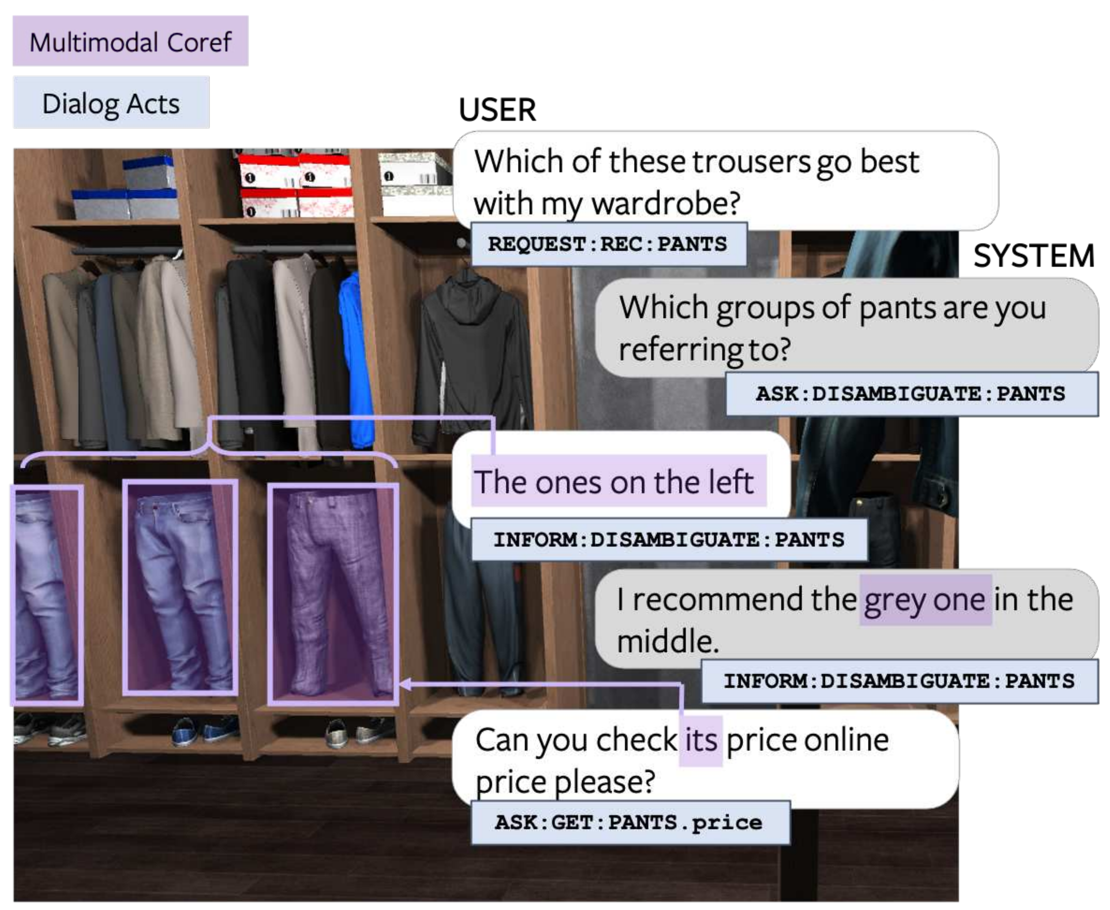
11777 Project: Multimodal Coreference Resolution in Task-Oriented Dialogue System

Yudong Liu, Zihao Deng, Hao-Ming Fu, Haoyang Wen
{yudongl, zihaoden, hfu2, hwen3}@andrew.cmu.edu
Carnegie Mellon University

Introduction

MultiModal Coreference Resolution (MM-Coref):



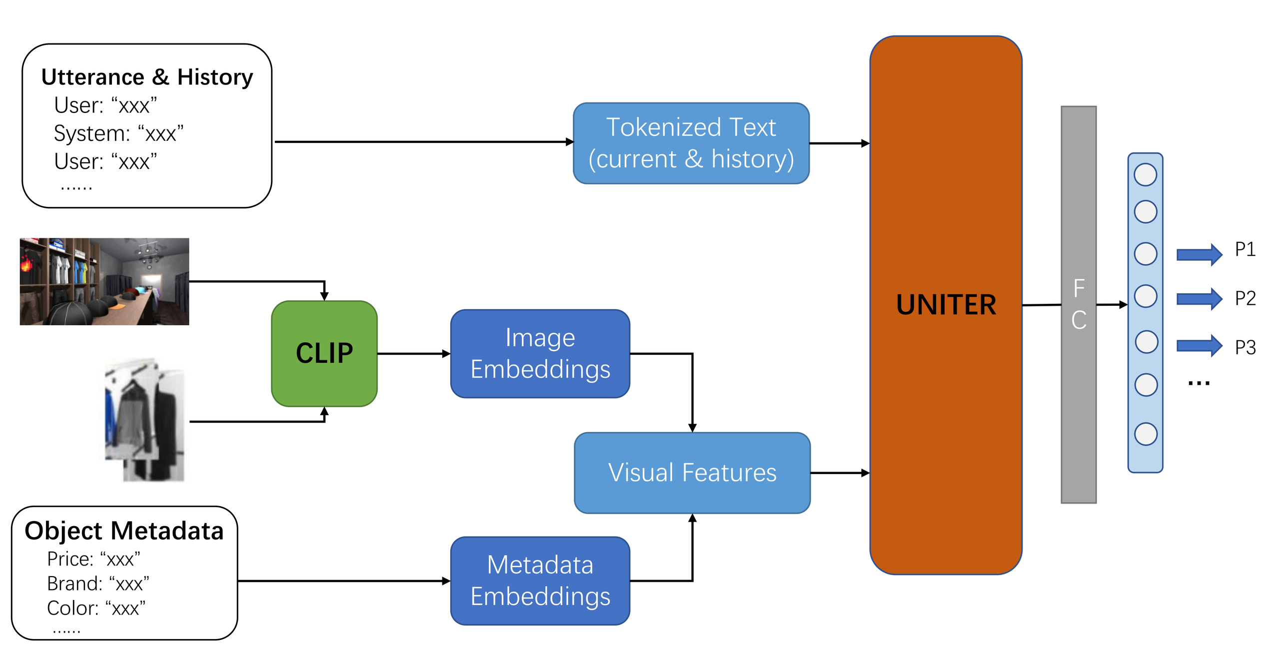
- *"What do you think of the grey pair on the left ?"*
- *"Add the one I mentioned to the cart."*

Background

Situated Interactive Multimodal Conversation 2.0 (SIMMC 2.0):

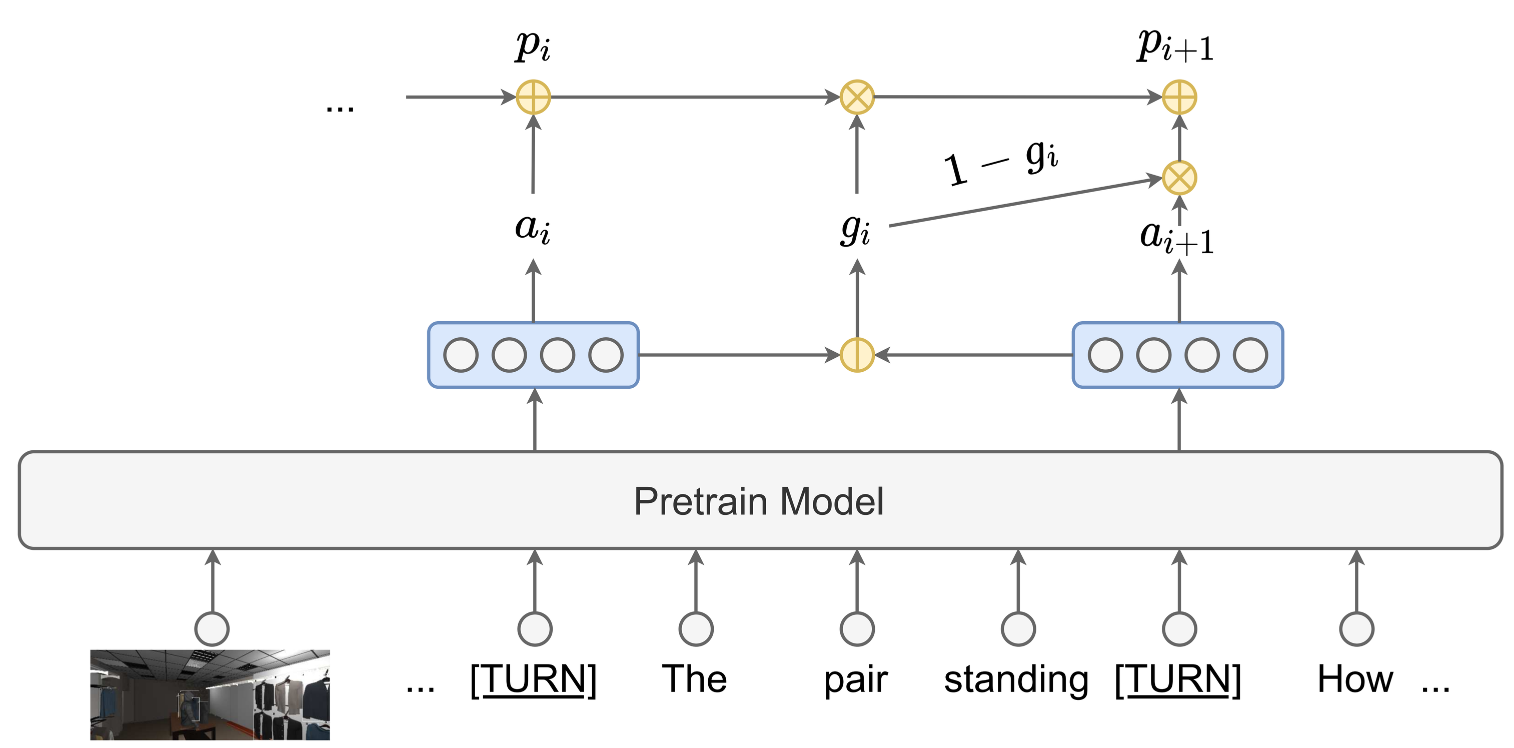
- Interactive shopping dataset
- User-assistant conversations about furniture/fashion
- 11K dialogs (117K utterances)
- 1566 snapshots from 160 3D scenes, with 19.7 items on average in a single scene.
- 10 types of referring expression.

Our baseline - NYU's system for MM-Coref:



Idea 1: Tracking Objects Over Turns

Typical Error: Inconsistent predictions on consecutive turns
Proposed Method: Explicitly model the transition over turns



- Local prediction at each user utterance for an object

$$a_{o,t} = \sigma(\mathrm{FFN}_2(\tanh(\mathrm{FFN}_1([\boldsymbol{h}_t; \boldsymbol{h}_o]))))$$

- Gate probability from consecutive utterances

$$g_t = \sigma\left(\boldsymbol{W}_g[\boldsymbol{h}_{t-1}; \boldsymbol{h}_t] + \boldsymbol{b}_g\right)$$

- Transition from previous turn

$$p_{o,t} = (1 - g_t) \times a_{o,t} + g_t \times p_{o,t-1}$$

Idea 2: Predicting the Count of Objects

Typical Error: The total number of predicted objects is incorrect
Proposed Method: First predict the count of objects, then filter the object predictions.

- Model: We propose to train an additional module that takes the corresponding [CLS] representations using NYU's architecture and predict the count of objects:

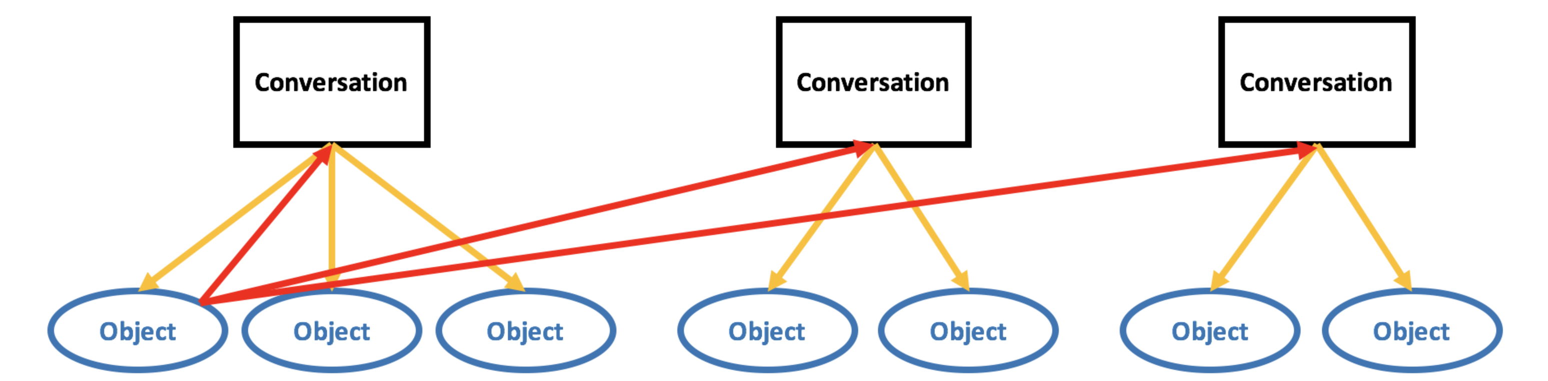
$$N = \mathrm{argmax}[\mathrm{softmax}(\boldsymbol{W}_{\mathrm{cls}}\boldsymbol{h}_{\mathrm{cls}} + \boldsymbol{b}_{\mathrm{cls}})]$$

- Prediction: We use the predicted N to select the top-N objects from NYU's prediction

$$O_{\mathrm{pred}} = \mathrm{topk}(p, N)$$

Idea 3: Contrasting Conversations

Goal: Improve alignment between conversations and objects
Proposed Method: For every object, add a contrastive learning object between positive/negative conversations



- Orange arrows: Original contrasts between positive/negative objects
- Red arrows: Add contrasts between positive/negative conversations

Experiments & Analysis

Models	Precision	Recall	F_1
GPT-2	40.0	40.5	40.3
Kakao	37.7	70.6	49.1
NYU	63.4	75.3	68.9
Idea 1	**63.8**	**75.8**	**69.3**
Idea 2	62.98	50.80	56.24
Idea 3	56.74	74.85	64.55

Table 1:The performance (%) of models on development-test set.

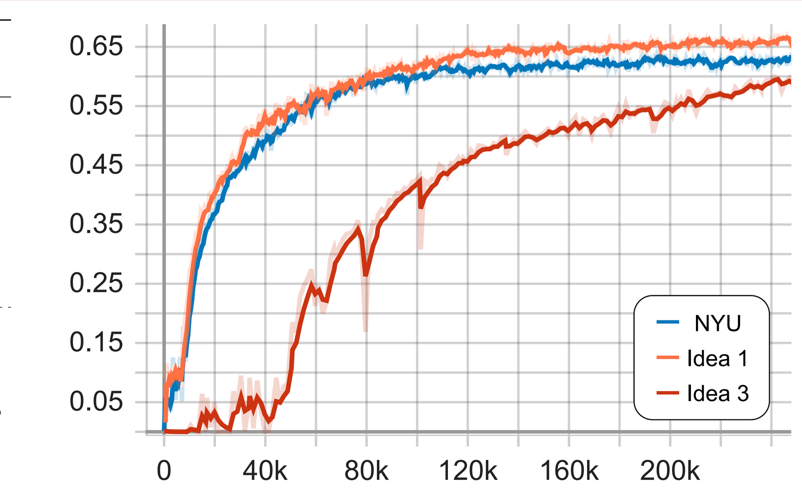


Figure 1:Learning curve of the models on the development set.

Analysis:

- The model of idea 1 outperforms the NYU baseline. We can also see steady improvement from the learning curve on development set.
- We can find cases that the model of idea 1 make consistent prediction using the probability from previous turn.
- Although our module from idea 2 itself can achieve high accuracy (98%), we see a drop on Recall, which indicates that there may be some high confidence wrong objects and we remove correct objects with low confidence.
- The model of idea 3 harms the performance. The reason may be the misalignment between the added training objective and the goal of the task.