

A Correction ratio

Table 1. Percentage of slots’ values changed in MultiWOZ 2.3 and MultiWOZ2.1, respectively, for “metadata” annotations. “Value Filled” stands for a value-filled from null, “none” or “not mentioned”. “Value Removed” means a slot value is changed to “not mentioned” or null. “Value dontcare” stands for slot values filled with “dontcare”.

Fixing Type	Count	Ratio
No Change	2,476,666	98.68%
Value Filled	20,639	0.82%
Value Changed	11,649	0.46%
Value Removed	221	0.01%
Value dontcare	563	0.02%

Table 1 shows statistics on the type of corrections we have made on the “metadata” annotations. Note that “dontcare” value is singled out during re-pairing since it is a significant factor (Table 8) on slot gate classifications in the TRADE model [21].

B Inconsistency

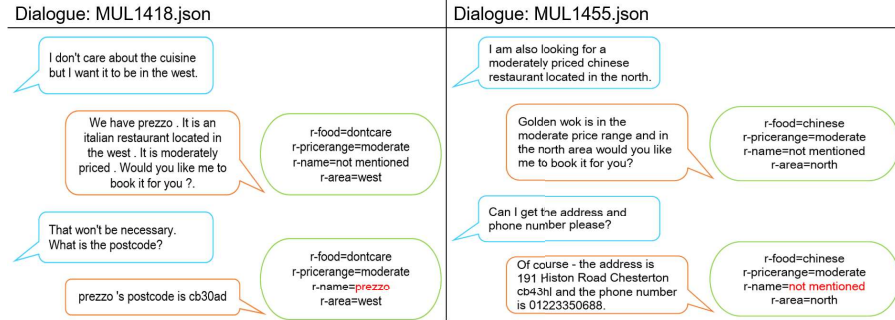


Fig. 1. Examples of inconsistent tracking on dialogue states of two different dialogues in similar scenarios from MultiWOZ 2.1. In the left column, dialogue *MUL1418.json* updates slot *r-name* with “prezzo” recommended by the system. However, for dialogue *MUL1455.json* in the right column, the value of slot *r-name* is remained as “not mentioned” even though “golden wok” is recommended by the system. “r” in the light green rectangle is an abbreviation for “restaurant”.

C Co-reference ratio

Table 2. Statistics of co-reference annotations. H/R/A/T represent “Hotel”, “Restaurant”, “Attraction” and “Train”, respectively.

Slot	Count	Ratio
Taxi.Depart	844	24.82%
H/R/A.Area	786	23.12%
Taxi.Dest	706	20.76%
H/R/A/T.Day	409	12.03%
H/R.Price	354	10.41%
H/R/T.People	201	5.91%
Taxi.Arrive	92	2.71%

Table 2 shows the statistics of the amount of “coreference” annotations for each slot type. We can see the most common co-referencing relationship is from “Taxi-Dest/Depart” and “xxx-Area”, followed by “Day”, “Price”, “People” and “Arrive”.

D Coreference sample

PMUL4852.json
<pre> 10: text: "That sounds wonderful ! Is it in the same area as the hotel ?" metadata: {} dialog_act: span_info: [] 1 item coreference: Hotel-Inform: [] 1 item 0: [] 5 items 0: "Area" 1: "same area" 2: "center" 3: 4 4: "12-12" turn_id: 10 </pre>

Fig. 2. Example of a co-referencing annotation. If the current turn involves more than one co-referencing relationships, all annotations will be gathered under the “coreference” key. The number “10” at the top left corner indicates the “turn_id” of dialogue *PMUL4852.json*.

The annotation takes a “Domain-Intent” format, including five parts: slot name, slot value in the current turn, referred value, referred turn id, and spans of referred value in the referred turn. Figure 2 depicts an example of “coreference” annotation and the corresponding values for the five parts are “Area”, “same area”, “center”, “4”, “12-12” under “Hotel-Inform”

E Dialogue State Tracking benchmarks

Table 3. Joint goal accuracies for different dialogue state tracking models on the MultiWOZ 2.1 and MultiWOZ-coref. We notice our work is cocurrent with MultiWOZ 2.2. However, we mainly base our refinement on MultiWOZ 2.1 and many models do not report joint goal accuracies on MultiWOZ 2.2. Therefore, MultiWOZ 2.2 is excluded from comparison.

Models	MultiWOZ 2.1	MultiWOZ 2.3
TRADE [21]	45.6%	49.2%
SUMBT [22]	49.2%	52.9%
COMER [5]	48.8%	50.2%
DSTQA [23]	51.2%	51.8%
SOM-DST [4]	53.1%	55.5%
TripPy [24]	55.3%	63.0%
ConvBERT-DG-Multi [2]	58.7%	67.9%
SAVN [3]	54.5%	58.0%

Upon code availability, we experiment MultiWOZ 2.3 on various dialogue state tracking models and Table 3 shows the corresponding joint goal accuracies.

F Value Normalization

Table 4. Value normalization rules when updating values from dialogue acts to dialogue states.

Type	Content
Number	zero': '0', 'one': '1', 'two': '2', 'three': '3', 'four': '4', 'five': '5', 'six': '6', 'seven': '7', 'eight': '8', 'nine': '9', 'ten': '10', 'eleven': '11', 'twelve': '12'
Pricerange	high end': 'expensive', 'expensively': 'expensive', 'upscale': 'expensive', 'inexpensive': 'cheap', 'cheaply': 'cheap', 'cheaper': 'cheap', 'cheapest': 'cheap', 'moderately priced': 'moderate', 'moderately': 'moderate'
dontcare	do n't have a preference': 'dontcare', 'do not have a preference': 'dontcare', 'no particular': 'dontcare', 'not particular': 'dontcare', 'do not care': 'dontcare', 'do n't care': 'dontcare', 'any': 'dontcare', 'does not matter': 'dontcare', 'does n't matter': 'dontcare', 'not really': 'dontcare', 'do nt care': 'dontcare', 'does n really matter': 'dontcare', 'do n't really care': 'dontcare'
Area	center': 'centre', 'northern': 'north', 'northside': 'north', 'eastern': 'east', 'eastside': 'east', 'westside': 'west', 'western': 'west', 'southside': 'south', 'southern': 'south'
Time	Remove words as 'after', 'before' and etc., and sort to the 'hh:mm' time format. 'X pm' format is remained as the original.
Stars	[0-9]-stars, converted to [0-9]
Parking & Internet	Free' value for parking and internet slot is converted to 'yes'
Plural	hotels': 'hotel', 'guesthouses': 'guesthouse', 'churches': 'church', 'museums': 'museum', 'entertainments': 'entertainment', 'colleges': 'college', 'nightclubs': 'nightclub', 'swimming pools': 'swimming pool', 'architectures': 'architecture', 'cinemas': 'cinema', 'boats': 'boat', 'boating': 'boat', 'theatres': 'theatre', 'concert halls': 'concert hall', 'parks': 'park', 'local sites': 'local site', 'hotspots': 'hotspot'

G SUMBT Slot Accuracy

Table 5. Slot accuracies among MultiWOZ 2.1, MultiWOZ 2.2, MultiWOZ 2.3 and MultiWOZ-coref in terms of different slot types. The bold number indicates the highest accuracy across all three datasets for each slot. The red bold number indicates higher accuracy between MultiWOZ 2.3 and MultiWOZ-coref for each slot.

Slot type	MultiWOZ 2.1	MultiWOZ 2.2	MultiWOZ 2.3	MultiWOZ-coref
attraction-area	95.94	95.97	96.28	96.80
attraction-name	93.64	93.92	95.28	94.59
attraction-type	96.76	97.12	96.53	96.91
hotel-area	94.33	94.44	94.65	95.02
hotel-book day	98.87	99.06	99.04	99.32
hotel-book people	98.66	98.72	98.93	99.17
hotel-book stay	99.23	99.50	99.70	99.70
hotel-internet	97.02	97.02	97.45	97.56
hotel-name	94.67	93.76	94.71	94.71
hotel-parking	97.04	97.19	97.90	98.34
hotel-pricerange	96.00	96.23	95.90	96.40
hotel-stars	97.88	97.95	97.99	98.09
hotel-type	94.67	94.22	95.92	95.65
restaurant-area	96.30	95.47	95.52	96.05
restaurant-book day	98.90	98.91	98.83	99.66
restaurant-book people	98.91	98.98	99.17	99.21
restaurant-book time	99.43	99.24	99.31	99.46
restaurant-food	97.69	97.61	97.49	97.64
restaurant-name	92.71	93.18	95.10	94.91
restaurant-pricerange	95.36	95.65	95.75	96.26
taxi-arriveBy	98.36	98.03	98.18	98.45
taxi-departure	96.13	96.35	96.15	97.49
taxi-destination	95.70	95.50	95.56	97.59
taxi-leaveAt	98.91	98.96	99.04	99.02
train-arriveBy	96.40	96.40	96.54	96.76
train-book people	97.26	97.04	97.29	97.67
train-day	98.63	98.60	99.04	99.38
train-departure	98.43	98.40	97.56	97.50
train-destination	98.55	98.30	97.96	97.86
train-leaveAt	93.64	94.14	93.98	93.96