Granular Change Accuracy: A more accurate performance metric for Dialogue State Tracking

Anonymous ACL submission

Abstract

Current community-accepted metrics used to evaluate Dialogue State Tracking (DST) have key weaknesses: they do not assign partial scores and over-penalize for mistakes that occur in earlier turns. Their assumptions about error uniformity leads to inaccurate DST evaluation. We propose a new metric to address this challenge — Granular Change Accuracy (GCA) — that evaluates for predicted changes in dialogue state over the entire dialogue history. Our benchmarking shows that GCA mitigates irrelevant traits in predictions; *i.e.* distribution uniformity and position of mistakes over turns, leading to more accurate evaluation.

1 Introduction

Dialogue State Tracking (DST) is the task of extracting user preferences from a Task-Oriented Dialogue (TOD) to accomplish a task such as booking a hotel room (Henderson et al., 2014). How to appropriately evaluate the performances of these models is still an area of ongoing research.

While the community has adopted a set of metrics to report results (Ye et al., 2022; Feng et al., 2022; Zhu et al., 2022; Hung et al., 2022), we argue that they can result in imbalanced assessment, such that strong systems receive poor scores and vice versa.

Table 1 presents a sample TOD with two sets of DST predictions, P^1 and P^2. P^1 predicts five of seven slots correctly whereas P^2 only predicts one correctly. However, the metrics of Joint Goal Accuracy (JGA; Henderson et al. 2014), Flexible Goal Accuracy (FGA; Dey et al. 2022) and Average Goal Accuracy (AGA; Rastogi et al. 2020) evaluate the latter P^2 as the better prediction. Just as problematic, Slot Accuracy (SA; Wu et al. 2021) gives inflated and similar scores to both predictions.

This is due to several weaknesses that current metrics employ. Firstly these metrics account for the same predictions multiple times throughout turns. Secondly, they weigh each turn equally, averaging over the turn accuracies. Finally, most existing metrics do not assign partial scores to turns. These weaknesses make existing metrics under/over-estimate performance in two scenarios: (1) when mistakes occur early in the dialogue, or (2) are uniformly distributed among turns.

To address these weaknesses we propose Granular Change Accuracy (GCA). GCA evaluates the performance by scoring the changes in the prediction and ground truth belief states at each turn. This ensures that the same prediction is not multiply-accounted. Moreover, it avoids under/over-estimation by averaging over state changes.

We evaluate GCA on MultiWOZ 2.1 dataset (Eric et al., 2020), conducting benchmarking experiments with popular baselines and show that GCA positions in the middle of the spectrum, more optimistically than JGA and FGA's strict penalizing

Turn	Conversation Details	
0	U_0	I want to book a hotel with free internet.
	G_0	{hotel: {internet:yes} }
	P_0^1	{hotel: { internet:no } }
	P_0^2	{hotel: { internet:yes } }
1	S_1	Great, how about free parking?
	U_1	Yes, that would be great!
	G_1	{hotel: {internet:yes, parking:yes} }
	P_1^1	{hotel: { internet:no , parking:no } }
	P_1^2	{hotel: { internet:yes , parking:no } }
2	S_2	There is a cheap guesthouse near center.
	U_2	Okay, please book for 6 people 4 days starting this Sunday.
	G_2	{hotel: {internet: yes, parking: yes, day: Sunday, people: 6, stay: 4, price: cheap, type: guesthouse} }
	P_2^1	{hotel: { internet: no, parking: no , day: Sunday , people: 6 , stay: 4 , price: cheap , type: guesthouse }}
	P_2^2	{hotel: { internet: yes, parking: no , day: Monday , people: 3 , stay: 2 , price: expensive , type: hotel }}

Table 1: Sample dialogue with ground truth turn belief state G_t and two belief state predictions.

scheme but not as inflated as SA and AGA. We further conduct a qualitative analysis proving that GCA is 0.1 less correlated with the position of mistakes and 0.29 less correlated with the distribution uniformity of mistakes compared against the recent FGA metric with a significant difference.

2 Related Work

The two most commonly reported DST metrics are JGA (Henderson et al., 2014) and SA (Wu et al., 2019). Both metrics take an arithmetic average of accuracy over turns assuming every turn is equally important — even when some turns may incorporate more slots compared to others (*c.f.* Turns 0 and 2 in Figure 1). Because dialogue states are accumulated across turns, these metrics account for the same prediction several times, leading to over/under-estimation of the DST model's performance. Moreover, JGA tends to under-estimate results since it denies partial credit from turns; whereas SA tends to over-estimate, as it rewards models for slots without an active value.

Flexible Goal Accuracy (Dey et al., 2022) (FGA) and Average Goal Accuracy (Rastogi et al., 2020) (AGA) are modified versions of JGA and SA, respectively. FGA redesigns JGA in order to diminish the repeated scoring of the same predictions by adding a decay parameter whereas AGA calculates recall over slots that have an active value in the turn. Although both of these metrics improve DST evaluation, they still do not completely address these identified problems (averaging accuracy over turns, multiply-accounting a prediction in different turns.

3 Background

Task Definition. DST is the task of extracting/generating the slot values for predefined slot labels specific to each domain, such as *restaurant-food: Indian* in the restaurant domain. We refer to a slot label/value as simply *slot* and *value* in this paper. A task-oriented dialogue is represented as $D = \{(S_0, U_0, BS_0), ..., (S_{n-1}, U_{n-1}, BS_{n-1})\}$ where S_i and U_i form the i_{th} turn pair and are system and user utterances, respectively; BS_i is the belief state of the i_{th} turn pair; and n is the number of turn pairs. Each turn pair can incorporate zero or more slot–value pairs, and these are summarized in the dialogue state, i.e. $BS = \{(S_0 : V_0), ..., (S_m : V_m)\}$ where $(S_j : V_j)$ is the j_{th} active slot–value pair and m is the number of slots predicted to have an active value in the current turn. Thus $m \leq M$ where M is the number of defined slots in the dataset (*e.g.* 30 for MultiWOZ). The rest $M - m$ slots acquire a "none" value indicating they are not active in the turn *i.e.* they do not have an actual value. Note that the dialogue state is formed cumulatively through the dialogue. Stated differently, any prediction made in an earlier turn will stick to the dialogue state unless a new value is predicted. (including "none" values).

Joint Goal Accuracy is the ratio of correctly predicted turn–pair slots over the number of turn pairs in the dialogue. A correct prediction requires every slot–value set within the turn–pair to match in prediction and ground truth belief states. $JGA = \frac{\sum_{t=0}^{n}(\mathbb{1} \mid G_t = P_t)}{n}$, where G_t and P_t are the ground truth and predicted belief states, respectively.

Slot Accuracy is calculated over all possible slot values regardless of which slots are predicted to have an active value in the turn. Thus it takes into consideration "none" valued slots, unlike JGA. $SA = \frac{\sum_{t=0}^{n} TA}{n}$ where TA, turn accuracy, is the ratio of correctly predicted slot values to M, *i.e.* the number of total slots defined in the dataset.

Average Goal Accuracy differs from earlier metrics because it evaluates only the performance of turns with active slots; i.e., if a turn does not have any ground truth values, it will be discarded during the evaluation. It calculates a recall value for all turns with non-empty ground truth belief states and returns the average.

$$AGA = \frac{\sum_{t=0}^{n}(\frac{G_t \cap P_t}{|G_t|} \mid |G_t| \geq 1)}{\sum_{t=0}^{n}(\mathbb{1} \mid |G_t| \geq 1)} \quad (1)$$

Flexible Goal Accuracy modifies the JGA metric to account for mistakes done in the current and earlier turns differently. Specifically, it copies JGA behavior for mistakes done in the current turn, completely ignoring the rest of the prediction and scoring the turn zero, however, unlike JGA when all slot values of the current turn are predicted correctly with a carried-over mistake from an earlier turn it penalizes the score rather than just scoring zero. This penalty decays by the number of turns passed since the mistake was made. They also provide a parameter, λ, to control this decay ratio.

4 Preliminary Analysis

We now categorize the weaknesses specific to each of the metrics reviewed above.

1. 0/1 Scores: Both JGA and FGA have a strict scoring scheme that assigns either full or no credit for each turn, disallowing partial credit. FGA only partly addresses this by adding the flexibility to diminish the penalty for earlier mistakes. Under this scheme, predictions that correctly predict the majority or minority of the ground truth slot values are deemed equivalent.

2. Turn-centric Scores: All four metrics average over turns. This results in under/over-estimation of DST performance, as some turns have more slots compared to others (*c.f.* Table 1, Turns 0 and 2).

3. Multiple-counting score: All four metrics account for the same predictions multiple times across turns. Thus a prediction made in the earlier turns of the dialogue results in a large effect. This also results in over/under-estimation of performance. FGA's decay parameter only partially addresses this concern, as it still penalizes earlier mistakes more harshly.

5 Granular Change Accuracy

We design GCA to address these weaknesses. The first weakness is a direct result of using the belief state rather than individual slot–value pairs for evaluation. The second is the result of averaging over the number of turns. The third is caused by evaluating the whole belief state at each turn, rather than just the changes. We design our metric to consider the slots (*0/1 scores*) whose value was modified (*multiple-counting score*) since the last turn and take the average over the total number of modifications (*turn-centric scores*). Granular Change Accuracy is thus named to suggest that it assesses accuracy over changes in the belief state.

The state changes in GCA are calculated by four metrics: 1) missed predictions where the slot had a value in ground truth BS but not in the predicted BS; 2) wrong predictions where the slot had a value in both ground truth and predicted BS but do not match; 3) over-predictions where the slot had a value in predicted BS but not in ground truth BS; and 4) correct predictions where the slot–value pairs in the ground truth and predicted BS match. Smith (2014) define a similar taxonomy but report these four directly instead of aggregating them into a final value, unlike GCA.

Algorithm 1 gives pseudocode to calculate these four metrics. These metrics are used to calculate four other intermediate products:

Algorithm 1 Calculating missed (M), wrong (W), over (O), and correct (C) predictions.

G_{-1} = [],P_{-1}= []
M, W, O, C = 0
for $t = 0, 1, \ldots$ **do**
  Get G_t and P_t for turn t.
  $G'_t = G_t \setminus G_{t-1}$, $P'_t = P_t \setminus P_{t-1}$
  Cset, Wset = 0
  for s, v pair in G'_t **do**
    if s not in P_t **then**
      M += 1
    else if $\{s, v\}$ not in P_t **then**
      W += 1
      add s to Wset
    else
      C += 1
      Add s to Cset
    end if
  end for
  for s, v pair in P'_t **do**
    if s not in G_t **then**
      O += 1
    else if $\{s, v\}$ not in P_t & s not in Wset **then**
      W += 1
    else if s not in Cset **then**
      C += 1
    else
      continue
    end if
  end for
end for
return M, W, O, C

Value Precision $VP = \frac{C}{P}$ where $P = C + W + O$ is the number of state change predictions.

Value Recall $VR = \frac{C}{G}$ where $G = C + W + M$ is the number of ground truth state changes.

Label Precision $LP = \frac{C+W}{P}$

Label Recall $LR = \frac{C+W}{G}$

The numerator in the last two values is composed of predictions where the slot was predicted correctly, but where the value prediction can be either correct or wrong.

Finally, we take a weighted harmonic mean of these four to calculate GCA:

$$GCA = \frac{(P+G)}{\frac{P*\alpha}{VP} + \frac{G*\alpha}{VR} + \frac{P*(1-\alpha)}{LP} + \frac{G*(1-\alpha)}{LR}} \quad (2)$$

Weights for precision- and recall-based metrics are scaled by the number of predictions and the ground truth values, respectively.

We use α to weigh value accuracies differently from label accuracies. Since value accuracy is an exact match whereas label accuracy is a partial match we believe the former should have a higher value. Specifically, in our experiments, we set α so

Model	JGA	FGA	SA	AGA	GCA
TRADE	48.86	61.19	96.96	88.79	80.15
SOM-DST	53.09	71.04	97.36	91.71	88.63
Trippy	35.82	54.09	95.4	80.67	78.60
T5	51.4	67.27	97.32	91.72	87.19

Table 2: Single-run benchmarking results over baseline models with four existing evaluation metrics and GCA.

that the ratio between value and label accuracies are 10:1, *i.e.* $\alpha \approx 0.9$.

6 Experiments and Analysis

We conduct experiments on MultiWOZ 2.1 dataset spanning 7 distinct domains with over 10,000 dialogues and report results with four DST models: TRADE (Wu et al., 2021), SOM-DST (Kim et al., 2020), Trippy (Heck et al., 2020), and T5 based model by Lin et al. (2021). For TRADE and SOM-DST we re-use the predictions reported in Dey et al. (2022). We trained Trippy and T5 from scratch on an NVIDIA-V100 using the best hyperparameter settings reported by the authors.

6.1 Benchmarking Results

Table 2 shows the benchmarking results. We set $\lambda = 0.5$ for FGA following Dey et al. (2022). JGA and FGA are at the lower side of the spectrum due to 0/1 scoring whereas SA and AGA present the highest scores with small differences across models. These metrics can be very deceptive. One could claim that there is a very big gap for industry-ready models judging from JGA or claim the models are not far from ideal judging from SA. Unlike these two community-accepted metrics, GCA gives a more accurate standing avoiding both under and over-estimation.

6.2 Fine-Grained Analysis

To analyze edge cases, we filter out 20 predictions of TRADE and SOM-DST models where FGA > GCA and GCA > FGA with the largest disagreement. FGA over-estimates the performance when errors are accumulated in a few turns, *i.e.* the mistakes are **not uniformly** distributed. Especially if these accumulations occur in the later part of the dialogue, *i.e.* when the mistakes show a **tail-oriented** distribution (c.f. samples in Appendix A.1).

6.3 Effect of Spurious Traits

Tail-Oriented Mistake Distribution. To further explore how tail-oriented mistakes affect its FGA and GCA evaluation, we define a new measure: $TO = \frac{\sum_i^K d_i}{n}$ where d_i is the distance of mistake i's turn from the middle turn of the dialogue (the distance is negative if the turn is in the first half of the dialogue and vice versa), and K is the number of mistakes. The nominator is the average turn index of mistakes whereas the denominator acts as a normalization factor.

Non-Uniform Mistake Distribution. Similarly, we define a non-uniformity measure inspired by the chi-square metric: $NU = \sum_i^T (m_i - \frac{\sum_j^T m_j}{T})^2$ where m_i is number of mistakes done in turn i and T is the number of turns.

Results. The Pearson Correlation Coefficients between TO and FGA/GCA are $0.08/-0.02$, whereas between NU and FGA/GCA are $0.12/-0.17$ respectively. The differences between these correlations are significant according to Zou (2008)'s confidence interval tests. FGA's correlation with both features is significantly stronger with a 95% confidence level (Further analysis in Appendix A.2).

7 Limitations

Though GCA is more exhaustive than existing metrics, there is still room for improvement by partial credit of slot values; *i.e.* by calculating the similarity of ground-truth and predicted values. Our experiments could also be generalized to other datasets than just MultiWOZ 2.1, to validate the empirical credibility of the metric.

8 Conclusion

We highlight the critical weaknesses of existing DST evaluation metrics and how they over/under-estimate performance. To address these, we propose Granular Change Accuracy (GCA) which evaluates accuracy over the belief state changes. We show through analysis that GCA, avoids both over and under-estimation in existing metrics. Moreover it has significantly less correlation with the insignificant traits of the dialogue, such as non-uniformity or tail-skewness of mistakes, compared to recent FGA metric. We claim better DST evaluation through GCA or alike metrics would open doors to much fairer and accurate performance results and thus enabling more trustworthy research in this field. Future work may check the similarity between the ground truth and predicted values to enable partial evaluation at the slot level rather than on the turn level as in GCA.

References

Suvodip Dey, Ramamohan Kummara, and Maunendra Desarkar. 2022. Towards fair evaluation of dialogue state tracking by flexible incorporation of turn-level performances. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 318–324, Dublin, Ireland. Association for Computational Linguistics.

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

Yue Feng, Aldo Lipani, Fanghua Ye, Qiang Zhang, and Emine Yilmaz. 2022. Dynamic schema graph fusion network for multi-domain dialogue state tracking. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 115–126, Dublin, Ireland. Association for Computational Linguistics.

Michael Heck, Carel van Niekerk, Nurul Lubis, Christian Geishauser, Hsien-Chin Lin, Marco Moresi, and Milica Gasic. 2020. TripPy: A triple copy strategy for value independent neural dialog state tracking. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 35–44, 1st virtual meeting. Association for Computational Linguistics.

Matthew Henderson, Blaise Thomson, and Jason D. Williams. 2014. The second dialog state tracking challenge. In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 263–272, Philadelphia, PA, U.S.A. Association for Computational Linguistics.

Chia-Chien Hung, Anne Lauscher, Simone Ponzetto, and Goran Glavaš. 2022. DS-TOD: Efficient domain specialization for task-oriented dialog. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 891–904, Dublin, Ireland. Association for Computational Linguistics.

Sungdong Kim, Sohee Yang, Gyuwan Kim, and Sang-Woo Lee. 2020. Efficient dialogue state tracking by selectively overwriting memory. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 567–582, Online. Association for Computational Linguistics.

Zhaojiang Lin, Bing Liu, Seungwhan Moon, Paul Crook, Zhenpeng Zhou, Zhiguang Wang, Zhou Yu, Andrea Madotto, Eunjoon Cho, and Rajen Subba. 2021. Leveraging slot descriptions for zero-shot cross-domain dialogue StateTracking. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5640–5648, Online. Association for Computational Linguistics.

Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. 2020. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34:8689–8696.

Ronnie Smith. 2014. Comparative error analysis of dialog state tracking. In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 300–309, Philadelphia, PA, U.S.A. Association for Computational Linguistics.

Chen Henry Wu, Yinhe Zheng, Xiaoxi Mao, and Minlie Huang. 2021. Transferable persona-grounded dialogues via grounded minimal edits. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2368–2382, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Chien-Sheng Wu, Andrea Madotto, Ehsan Hosseini-Asl, Caiming Xiong, Richard Socher, and Pascale Fung. 2019. Transferable multi-domain state generator for task-oriented dialogue systems. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 808–819, Florence, Italy. Association for Computational Linguistics.

Fanghua Ye, Yue Feng, and Emine Yilmaz. 2022. ASSIST: Towards label noise-robust dialogue state tracking. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2719–2731, Dublin, Ireland. Association for Computational Linguistics.

Qi Zhu, Bing Li, Fei Mi, Xiaoyan Zhu, and Minlie Huang. 2022. Continual prompt tuning for dialog state tracking. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1124–1137, Dublin, Ireland. Association for Computational Linguistics.

Guang Zou. 2008. Toward using confidence intervals to compare correlations. *Psychological methods*, 12:399–413.

Turn	Conversation Details	
0	G_0	{hotel: {name: el shaddai} }
	P_0	{hotel: { name: el shaddai } }
1	G_1	{hotel: {name: el shaddai} }
	P_1	{hotel: { name: el shaddai } }
2	G_2	{hotel: {name: el shaddai} attraction: {type: museum } }
	P_2	{hotel: { name: el shaddai } }
3	G_3	{hotel: {name: el shaddai} attraction: {type: museum } }
	P_3	{hotel: { name: el shaddai } }
4	G_4	{hotel: {name: el shaddai} attraction: {type: museum } }
	P_4	{hotel: { name: el shaddai } }
5	G_5	{hotel: {name: el shaddai} attraction: {type: museum , area: dontcare , name: dontcare } }
	P_5	{hotel: { name: el shaddai },attraction: { name: Cambridge artworks } }
6	G_6	{hotel: {name: el shaddai} attraction: {type: museum , area: dontcare , name: dontcare } }
	P_6	{hotel: { name: el shaddai },attraction: { name: Cambridge artworks } }
7	G_7	{hotel: {name: el shaddai} attraction: {type: museum , area: dontcare , name: dontcare } }
	P_7	{hotel: { name: el shaddai },attraction: { name: Cambridge artworks } }

Table 3: Sample dialouge from MultiWOZ 2.1 dataset, (MUL1110) with ground truth and predicted belief states. GCA: 31.43, FGA: 54.74

A Appendix

A.1 Sample GCA and FGA Scores

This section presents two sample dialogues from the MultiWOZ 2.1 dataset along with DST model predictions, and FGA/GCA evaluations.

The dialogue in Table 3 is an example where FGA over-estimates the performance of a dialogue scoring it 55% even though it only predicts one out of four slots correctly. This is because the majority of mistakes in the dialogue occur closer to the tail of the dialogue.

Table 4 on the other hand presents an example where FGA under-estimates the performance scoring a prediction 26% even though it predicts three out of four slots correctly. This is because contrary to the previous example the majority of mistakes in this scenario are head-oriented *i.e.* they are closer to the beginning of the dialogue.

In line with these observations the TO measure (*c.f.* section 6.3) of dialogues in Table 3 and 4 are 0.55 and 0.36 respectively. This also suggests that the tail-orientedness of dialogues spuriously increases the FGA performance.

Turn	Conversation Details	
0	G_0	{hotel: {pricerange: expensive}, {parking: yes} }
	P_0	{hotel: { pricerange: expensive } , {parking: yes } }
1	G_1	{hotel: {pricerange: expensive}, {parking: yes} , {type: hotel}, { area: dontcare } }
	P_1	{hotel: { pricerange: expensive } , {parking: yes },{type: hotel } }
2	G_2	{hotel: {pricerange: expensive}, {parking: yes} , {type: hotel}, { area: dontcare } }
	P_2	{hotel: { pricerange: expensive } , {parking: yes },{type: hotel } }
3	G_3	{hotel: {pricerange: expensive}, {parking: yes} , {type: hotel}, { area: dontcare } }
	P_3	{hotel: { pricerange: expensive } , {parking: yes },{type: hotel } }

Table 4: Sample dialouge from MultiWOZ 2.1 dataset, (SNG0779) with ground truth and predicted belief states. GCA: 75, FGA: 26.38

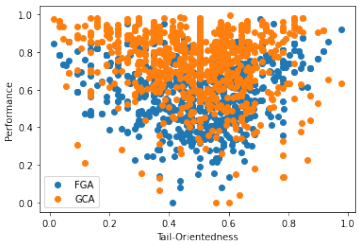
A.2 Analysis of Spurious Traits

We further analyze the distribution of FGA and GCA across the two spurious traits: tail-orientedness and distribution uniformity of mistakes.

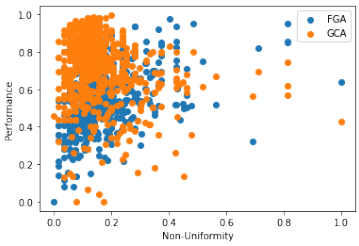
Figure 1a shows how performance distribution changes according to the TO measure of predictions. Although GCA results are generally higher, FGA shows higher results between 0.8-1.0 TO values. This suggests that as dialogues' tail-orientedness increases FGA tends to evaluate performance higher even though this should not have any effect on an ideal evaluation metric.

Figure 1b shows the effect of the NU measure on performance distribution. On the left-hand side of the plot, FGA tends to have lower values compared to GCA whereas on the rightmost edge it shows higher values. This suggests that FGA is affected by positions of the mistakes in the prediction which again should be an insignificant trait when it comes to evaluating DST models.

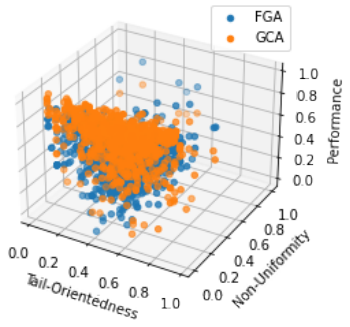
Finally, Figure 1c shows the effect of both traits together. Again moving higher on both traits' axes FGA's evaluation gets affected significantly. The dialogues that are evaluated highest by FGA are clustered on the higher ends of both axes supporting the hypotheses that they have a significant effect.



(a)



(b)



(c)

Figure 1: Figures showing how spurious dialogue traits effect GCA and FGA scoring. 1a - tail-orientedness of mistakes vs performance, 1b - non-uniformity of mistakes vs performance, 1c - tail-orientedness and non uniformity of mistakes vs performance.