Controllable Generation of Dialogue Acts for Dialogue Systems via Few-Shot Response Generation and Ranking

Angela Ramirez, Kartik Aggarwal , Juraj Juraska , Utkarsh Garg and Marilyn A. Walker
University of California Santa Cruz
Contact: aramir62@ucsc.edu, kartik@ucsc.edu, mawalker@ucsc.edu

Motivation: NLG for Spoken Dialogue Systems

- **Dialogue Acts:** You need to control the dialogue act.
- **Semantic Attributes:** You need to control the expression of the Semantic Attributes. System has a specific thing to convey, either because dialogue is task oriented, or it is based on some kind of knowledge.
- **Crowdsourced Datasets:** Getting large datasets can be expensive. May not exhibit the properties you want. Crowdsourced training data can have quality issues.
- Could Prompt-Based Learning be a solution to these challenges?
- Can we control both Dialogue Acts and Semantic Attributes with few-shot learning?

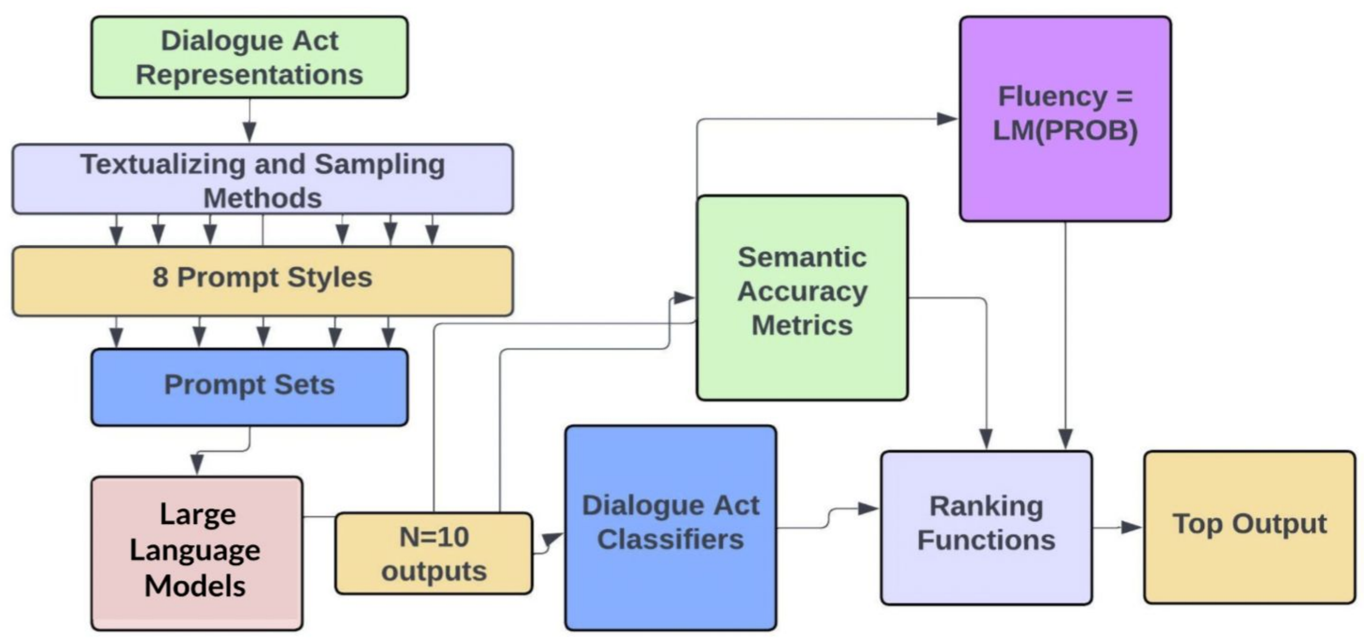
Datasets and Models

1. Viggo, 9 dialogue acts associated with the **video game domain**
2. RNNLG **Laptop and TV domain,** where they consist of 13 dialogue acts
3. Models: Jurassic Jumbo-1, GPT-3, GPT-NEO 1B,

Examples of Viggo Dialogue Acts

D-Act	Example/ MR
Suggest	Alright! Have you played any **adventure** games by **Remedy Entertainment**, like **Alan Wake?**
MR	D-Act = **suggest**, name = Alan Wake, developer = Remedy Entertainment, genres=adventure
Confirm	Oh, do you mean the **2017** game from **Ninja Theory**, **Hellblade: Senua's Sacrifice?**
MR	D-Act = **confirm**, name = Hellblade: Senua's Sacrifice, release_year = 2017, developer = Ninja Theory
Give Opinion	I think that **SpellForce 3** is one of the **worst** games I've ever played. Trying to combine the **real-time strategy and role-playing** genres just doesn't work, and the **bird's eye view** makes it near impossible to play.
MR	D-Act = **give_opinion**, name = SpellForce 3, rating =poor. genres = (real-time strategy, role-playing), player_perspective =bird view

Experimental Design



NLG Conditioned on DAs and Semantic Attributes

- We want to generate an utterance y, conditioned on an input x, composed of DA D and attribute values a.

$$p(y|d,a) = p(d|y,a) * p(a|y) * p(y)$$

- p(d | y, a) is the **DA probability** given the generated utterance y and the semantic attributes a
- p(a|y) is the **semantic accuracy,** which can be measured in different ways
- p(y) is the unconditional probability of y, which is normally taken to be a measure of **fluency** of the output y.

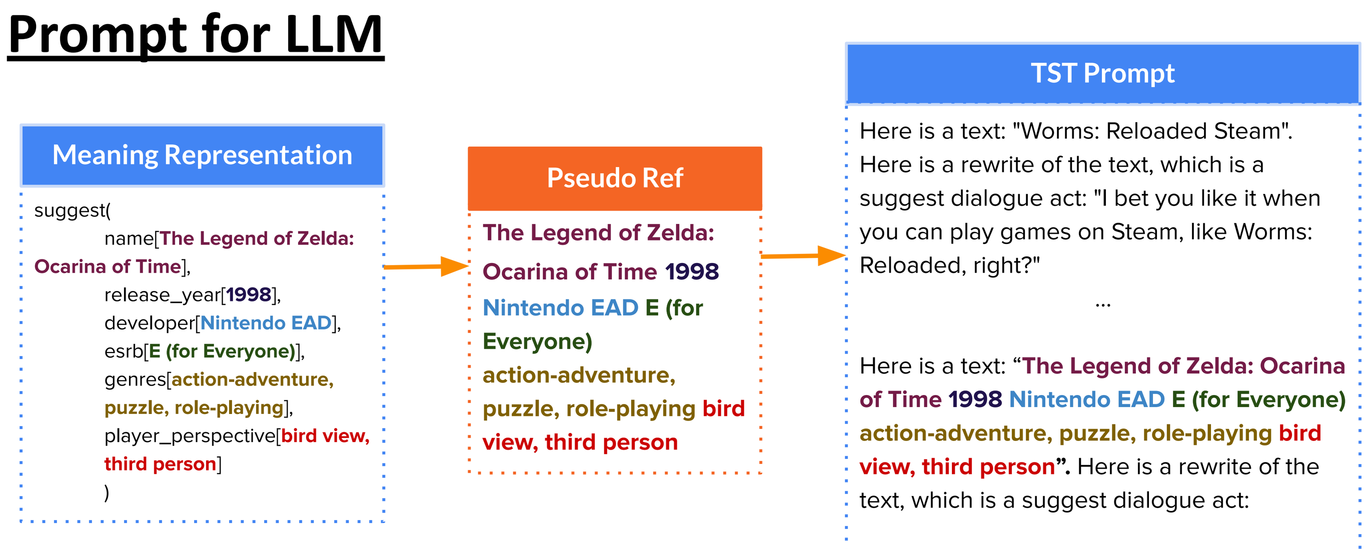
Ranking Functions

Below are the 6 ranking functions we utilize where DAC is dialogue act accuracy, SACC is semantic accuracy, P(S) is fluency where we used language model probability, pBBLEU is pseudo beyond BLEU, and pBLEU is pseudo BLEU.

- RF1 combines dialogue act accuracy, semantic accuracy, and fluency
- Noticed that pBLEU may pick up hallucinations. Added RF2
- RF2DA: FILTER at each step
- RF3, replace SACC with Beyond BLEU
- RF4: use pBBLEU as a baseline

RF1	DAC*SACC*P(S)
RF2	DAC*SACC*pBBLEU*P(S)
RF2da	DAC | SACC | pBLEU | P(S)
RF3	DAC*pBBLEU*P(S)
RF4	pBBLEU
RF5	pBLEU

Process of Converting Meaning Representations to use in Prompt for LLM



Results

- Can we achieve a high level of dialogue act control with PBL?
  - DAC column
- Can we achieve high semantic accuracy with PBL?
  - SACC column
  - Not as high as SOTA, but surprisingly good given only 10 examples
- Which prompt format works best?
- How many samples are optimal?
  - All the TST formats, giving 10 samples slightly better than 5
  - Definitional has promise, maybe needs tweaking
- Should samples always be of the same dialogue act?
  - yes
- How does it compare to few-shot fine tuning?
  - Same data 5-per much worse
  - With 100 per DA, fine-tuning has significantly lower DAC but better SACC

ID	N	PERF	SACC	DAC
Few-Shot Fine-Tuning Experiments				
FTune 5-per	45	38.88	85.71	54.44
FTune 25-per	225	62.22	92.19	79.72
FTune 50-per	450	71.94	96.43	79.44
FTune 100-per	900	78.61	97.74	80.56
Prompt-Based In-Context Per-DA Experiments				
5-per DA	45	60.0	83.6	92.0
2-per DA	18	61.0	83.2	91.4
2-per DA + Def	18	64.0	84.0	98.6
Prompt-Based In-Context Specific-DA Experiments				
TST Vanilla	10	**85.6**	**94.7**	**100**
TST Dialogue	10	83.9	94.2	100
TST Paraphrase	10	83.9	94.2	100
Pseudo	10	75.8	94.2	100
S2S	10	70.6	86.5	100
Definition (each)	10	76.9	91.2	100
Definition (top)	10	82.2	93.5	100
Paraphrase	10	77.8	92.1	100
Dialogic	10	77.2	91.5	100
TST Vanilla	5	80.6	92.6	98.7
TST Dialogue	5	**83.9**	**93.9**	100
TST Paraphrase	5	80.2	92.6	99.7
Pseudo	5	52.2	82.6	88.6
S2S	5	66.7	83.5	98.7
Definition (each)	5	80.0	92.7	99.4
Definition (top)	5	77.2	91.3	**100**
Paraphrase	5	70.8	89.7	100
Dialogic	5	66.9	88.3	99.1
TST Vanilla	1	69.2	88.2	92.0
TST Dialogue	1	69.2	88.2	93.3
TST Paraphrase	1	**72.2**	**89.8**	93.6
Definition	1	63.9	85.3	**98.3**
Paraphrase	1	41.9	75.1	83.8
Dialogic	1	38.8	71.8	82.3

- Which ranking function works best?
  - Weighting by DAC works better not just for DA but also for SACC
- Do we need domain specific ranking functions? yes
- Does it generalize across domains?
  - Yes
    - Better results for TV
    - Comparable for Laptop

RF	Terms	PERF	SACC	DAC	BLEU
ViGGO					
RF1	DAC, SACC, P(S)	79.17	91.82	99.72	38.41
RF2	DAC, SACC, pBLEU, P(S)	78.33	91.72	99.00	38.67
$RF2_{DA}$	DAC, SACC, pBLEU, P(S)	**85.56**	**94.73**	**100.00**	40.08
RF3	DAC, pBBLEU, P(S)	62.78	84.38	100.00	**49.87**
RF4	pBBLEU	60.55	91.63	77.78	42.82
RF5	pBLEU	44.22	81.66	75.28	40.08
TV					
RF1	DAC, SACC, P(S)	85.40	96.86	100.00	72.55
RF2	DAC, SACC, pBLEU, P(S)	88.19	97.43	100.00	72.55
$RF2_{DA}$	DAC, SACC, pBLEU, P(S)	**98.85**	**99.76**	**100.00**	60.51
RF3	DAC, pBBLEU, P(S)	73.96	93.87	100.00	**72.89**
RF4	pBBLEU	90.14	97.88	99.71	60.51
RF5	pBLEU	63.45	91.50	99.57	66.71
Laptop					
RF1	DAC, SACC, P(S)	49.25	86.70	100.00	61.24
RF2	DAC, SACC, pBLEU, P(S)	57.29	89.47	100.00	59.39
$RF2_{DA}$	DAC, SACC, pBLEU, P(S)	**80.95**	**95.90**	**100.00**	**61.36**
RF3	DAC, pBBLEU, P(S)	35.55	80.41	100.00	45.03
RF4	pBBLEU	61.79	90.97	98.88	36.32
RF5	pBLEU	42.38	84.25	97.77	**61.36**

Conclusion

1. Few Shot prompt-based tuning with ranking performs better than a fine-tuned model
2. Achieved perfect dialogue act accuracy (DAC) and near perfect semantic accuracy (SACC)
3. Method is generalizable to different domains
4. Future work may be needed to evaluate if certain prompts such as the definitional style can be readjusted to achieve higher scores, and whether different fine-tuned models can outperform prompt-based learning