A Study on Dialogue Evalaution Metrics Utilizing Large Language Models

Seungjun Lee, Dahyun Jung, Yuna Hur

Presenter: Seungjun Lee

Contents

01. Background

Our Methods

Research goals

- Dialogue Evaluation Metric research with LLM and Prompting Engineering→ is Feasible?
- Evaluate existing Dialog Evaluation Metrics with LLM

→ W/O Reference

→ No Labeled example

→ Consider contextual meaning and topic diversity

Task Definition

Score of Automatic Dialog Evaluation Metric

$$f : (c, r, g) \longrightarrow s$$

→ Dialog context c, Model response r

→ Evaluate a response generated based on a human written reference response g (reference-free)

Human quality annotation

$$(c, r) \longrightarrow q$$

→ A human annotator looks at the dialog context c and scores the quality of a given response r.

"Measure the performance of a metric by calculating the correlation between the **human rating** (Q) and **the metric score** (S)"

Experimental Settings

Quality-annotated Datasets

- DSTC6: Uses conversations from Twitter, which are noisier and more realistic than other conversational data
- DSTC9: Collected through direct interactions between real users and an open-domain chit-chat system
- PredictiveEngage: evaluates the overall quality of responses from DailyDialog, a daily conversation data set
- FED: Evaluates human-human and human-system conversations based on DialoGPT trained on Reddit conversation data.
- GRADE: Measures the consistency of conversation topics by graphing conversation transcripts
- HolisticEval: Evaluates conversation quality such as context coherence, language fluency, response diversity, and logical self-consistency.
- USR: Annotated for TopicalChat and PersonaChat datasets by considering qualities such as Understandable, Natural, Maintains Context, and Interesting.

Experimental Settings

Dialogue Evaluation Metrics

Metric	Pretrained Model	Training Dataset	Reference-Free?	Objective
BLEU	X	X	X	X
METEOR	X	X	X	X
ROUGE	X	X	X	X
BERTScore	X	X	X	X
BLEURT	BERT	WMT Metrics Shared Task	X	MSE
QuestEval	BERT	SQuAD-v2 / NewsQA	√	QA / QG
RUBER	T5	DailyDialog / PersonaChat	X	Triplet
BERT-RUBER	X	DailyDialog / PersonaChat	X	Triplet
PONE	BERT	DailyDialog	X	Triplet
MAUDE	BERT	PersonaChat	√	NCE
GRADE	BERT	DailyDialog	√	Triplet
DynaEval	RoBERTa	ED / ConvAI2 / DailyDialog	√	Triplet
USR	RoBERTa	TopicalChat / PersonaChat	√	MLM / CrossEntropy
USL-H	BERT	DailyDialog	√	VUP/NSP/MLM
DialogRPT	GPT-2	Reddit	√	CrossEntropy
Deep AM-FM	Multilingual BERT	Twitter	X	MLM
HolisticEval	BERT	DailyDialog	√	LM
PredictiveEngage	BERT	ConvAI	X	CrossEntropy
FED	DialoGPT	X	√	X

Experimental Settings

Prompt Engineering

- *Understandable (0 - 1):* Is the response understandable given the previous context?
- *Natural (1 - 3):* Does the response seem to be something that a person would naturally say?
- *Maintains Context (1 - 3):* Does the response serve as a valid continuation of the preceding conversation?
- *Interesting (1 - 3):* Is the response dull or interesting?
- *Uses Knowledge (0 - 1):* Given the fact that the response is conditioned on, how well does the response use that fact?
- *Overall Quality (1 - 5):* Given your answers above, what is your overall impression of the quality of this utterance?

```
    template = f"""You will be given a conversation between
two individuals. You will then be given several potential
responses for the next turn in the conversation. These
responses all concern an interesting fact, which will be
provided as well. Your task is to rate each of the
responses.  Evaluate the response the given context and
reference. Output only scores. Do not output any comments or
feedback.

[Annotation Instructions]
- Understandable (0 - 1): Is the response understandable
given the previous context?
- Natural (1 - 3): Does the response seem to be something
that a person would naturally say?
- Maintains Context (1 - 3): Does the response serve as a
valid continuation of the preceding conversation?
- Interesting (1 - 3): Is the response dull or interesting?
- Uses Knowledge (0 - 1): Given the fact that the response
is conditioned on, how well does the response use that fact?
- Overall Quality (1 - 5): Given your answers above, what is
your overall impression of the quality of this utterance?
    """
```

〈USR: An Unsupervised and Reference Free Evaluation Metric for Dialog Generation〉

03. Experiment

Results

USR-TopicalChat

Metric	Pearson	Spearman
BLEU-4	0.216	0.296
METEOR	0.336	0.391
ROUGE-L	0.275	0.287
BERTScore	0.298	0.325
BLEURT	0.282	0.317
QuestEval	0.3	0.338
RUBER	0.247	0.259
BERT-RUBER	0.342	0.348
PONE	0.271	0.274
MAUDE	0.044	0.083
DEB	0.18	0.116

Metric	Pearson	Spearman
GRADE	0.2	0.217
DynaEval	-0.032	-0.022
USR	0.412	0.423
USL-H	0.322	0.34
DialogRPT	0.12	0.105
Deep AM-FM	0.285	0.268
HolisticEval	-0.147	-0.123
PredictiveEngage	0.222	0.31
FED	-0.124	-0.135

Results

USR-TopicalChat

Metric	Pearson	Spearman
BLEU-4	0.216	0.296
METEOR	0.336	0.391
ROUGE-L	0.275	0.287
BERTScore	0.298	0.325
BLEURT	0.282	0.317
QuestEval	0.3	0.338
RUBER	0.247	0.259
BERT-RUBER	0.342	0.348
PONE	0.271	0.274
MAUDE	0.044	0.083
DEB	0.18	0.116

Metric	Pearson	Spearman
GRADE	0.2	0.217
DynaEval	-0.032	-0.022
USR	0.412	0.423
USL-H	0.322	0.34
DialogRPT	0.12	0.105
Deep AM-FM	0.285	0.268
HolisticEval	-0.147	-0.123
PredictiveEngage	0.222	0.31
FED	-0.124	-0.135
GPT-4	0.511	0.490
GPT-4 (ref-free)	**0.568**	**0.538**

Results

USR-PersonaChat

Metric	Pearson	Spearman
BLEU-4	0.135	0.090
METEOR	0.253	0.271
ROUGE-L	0.066	0.038
BERTScore	0.152	0.122
BLEURT	0.165	0.154
QuestEval	0.176	0.236
RUBER	0.131	0.19
BERT-RUBER	0.266	0.248
PONE	0.373	0.375
MAUDE	0.345	0.298
DEB	0.291	0.373

Metric	Pearson	Spearman
GRADE	0.358	0.352
DynaEval	0.149	0.171
USR	0.44	0.418
USL-H	0.495	0.523
DialogRPT	-0.064	-0.083
Deep AM-FM	0.228	0.219
HolisticEval	0.087	0.113
PredictiveEngage	-0.003	0.033
FED	-0.028	-0.000
GPT-4	0.606	0.626
GPT-4 (ref-free)	**0.645**	**0.665**

Results

Qualitative assessment

```
context =
  ["so , i 'm reading the latest film from studio ghibli is out the tale of
princess kaguya . dunno if you 're familiar with them , but studio ghibli has
made a lot of great animated films , like spirited away , and princess
mononoke ",
  " i do n't think i have heard of them . i have heard that one of the
directors recently passed away , and his last film was nominated for an
academy award ",
  " yeah , sadly , disney ( which owns the american rights to the films )
does n't tend to promote them very much . i think they 're worried they 'll
cut into their \" home grown \" market . anyway , dunno if you even like
animated movies , but they 're worth checking out . ",
  " i do n't watch them very often . apparently there was a showing of the
recent film in a park in d.c . that 's one u.s . city i have n't been to ",
  " sadly , i have n't been to dc either , although i 've always wanted to
visit there . apparently there 's a lot of interesting going down this summer
. they 're having a crab feast at the navy - marine corps stadium . they 'll
have 100 gallons of crab soup ! can you imagine that much soup ?"
]


responses = "i think it 's interesting that peter gabriel has been in the us
, he is a great performer ."
```

- Understandable (0 - 1): 0.1
- Natural (1 - 3): 1.0
- Maintains Context (1 - 3): 1.0
- Interesting (1 - 3): 1.0
- Uses Knowledge (0 - 1): 0.0
- Overall Quality (1 - 5): 1.0

The response doesn't really make sense in the context of the conversation. It suddenly brings up Peter Gabriel, a figure not previously mentioned, without any clear connection to the previous discussion about Studio Ghibli films or a trip to DC. It doesn't maintain the context of the conversation well and seems quite random. It's not particularly interesting as it doesn't add anything substantive to the conversation or use knowledge related to the topic at hand. Overall, the quality of the response is low.

Results

Qualitative assessment

```
context =
  ["so , i 'm reading the latest film from studio ghibli is out the tale of
princess kaguya . dunno if you 're familiar with them , but studio ghibli has
made a lot of great animated films , like spirited away , and princess
mononoke ",
  " i do n't think i have heard of them . i have heard that one of the
directors recently passed away , and his last film was nominated for an
academy award ",
  " yeah , sadly , disney ( which owns the american rights to the films )
does n't tend to promote them very much . i think they 're worried they 'll
cut into their \" home grown \" market . anyway , dunno if you even like
animated movies , but they 're worth checking out . ",
  " i do n't watch them very often . apparently there was a showing of the
recent film in a park in d.c . that 's one u.s . city i have n't been to ",
  " sadly , i have n't been to dc either , although i 've always wanted to
visit there . apparently there 's a lot of interesting going down this summer
. they 're having a crab feast at the navy - marine corps stadium . they 'll
have 100 gallons of crab soup ! can you imagine that much soup ?"
]


responses = "i think it 's interesting that peter gabriel has been in the us
, he is a great performer ."
```

- Understandable (0 - 1): 0.1
- Natural (1 - 3): 1.0
- Maintains Context (1 - 3): 1.0
- Interesting (1 - 3): 1.0
- Uses Knowledge (0 - 1): 0.0
- Overall Quality (1 - 5): 1.0

The response doesn't really make sense in the context of the conversation. It suddenly brings up Peter Gabriel, a figure not previously mentioned, without any clear connection to the previous discussion about Studio Ghibli films or a trip to DC. It doesn't maintain the context of the conversation well and seems quite random. It's not particularly interesting as it doesn't add anything substantive to the conversation or use knowledge related to the topic at hand. Overall, the quality of the response is low.

ToyTest

Human Evaluation Measures

- **Fluency**: We ask the evaluators to judge the fluency of the dialogue on a scale from 1 to 5, where 1 is “not fluent at all”, 5 is “extremely fluent”, and 3 is “fine”.
- **Situation Consistency**: We ask the evaluators to judge the situation consistency on a scale of from 1 to 5, where 1 is “not maintained at all”, 5 is “well maintained”, and 3 is “fine”.
- **Engagingness**: We ask the evaluators to judge the engagingness disregarding fluency from 1 to 5, where 1 is “not engaging at all”, 5 is “extremely engaging”, and 3 is “fine”.

〈PEEP-Talk: A Situational Dialogue-based Chatbot for English Education〉

ChatGPT’s response

Fluency: We ask the evaluators to judge the fluency of the dialogue on a scale from 1 to 5, where 1 is “not fluent at all”, 5 is “extremely fluent”, and 3 is “fine”.

Situation Consistency: We ask the evaluators to judge the situation consistency on a scale of from 1 to 5, where 1 is “not maintained at all”, 5 is “well maintained”, and 3 is “fine”.

Engagingness:Weasktheevaluatorstojudge the engagingness disregarding fluency from 1 to 5, where 1 is “not engaging at all”, 5 is
“extremely engaging”, and 3 is “fine”.

[context]
A: how is the market's reaction to the newly released product ?
B: the sales increase is faster than the previous product .

[response]
A: It's really nice

Fluency: 5 - The response is perfectly fluent English.

Situation Consistency: 4 - The response is consistent with the situation but it's somewhat vague and doesn't further the conversation in a meaningful way.

Engagingness: 2 - The response is not engaging. It's an extremely brief and non-specific affirmation of the information provided by the other participant, but it doesn't ask a new question or add any new information to keep the conversation going.

03. Conclusion

Conclusion

- Evaluation with Consistent Criteria: Evaluated multiple Dialogue Evaluation Metrics on the same evaluation dataset.
- Feasibility of LLM: Proposed a Dialogue Evaluation Metric using LLM without labeled datasets and recorded high performance.
- Experiments on different datasets: Different datasets have different features that we want to measure in the dialog
- Various Prompt Engineering: Experimenting with different prompts to get the maximum performance of the Dialog Evaluation Metric.