Data Augmentation for Conversational AI

The Web Conference 2024

Tutorial website

Presenters

Heydar Soudani
PhD Candidate
Radboud University
heydar.soudani@ru.nl

Evangelos Kanoulas
Full Professor
University of Amsterdam
e.kanoulas@uva.nl

Roxana Petcu
PhD Candidate
University of Amsterdam
r.m.petcu@uva.nl

Faegheh Hasibi
Assistant Professor
Radboud University
f.hasibi@cs.ru.nl

Part 1: Evaluation

Duration: 20 min

Presenter: Faegheh Hasibi

Synthetic Conversation Evaluation

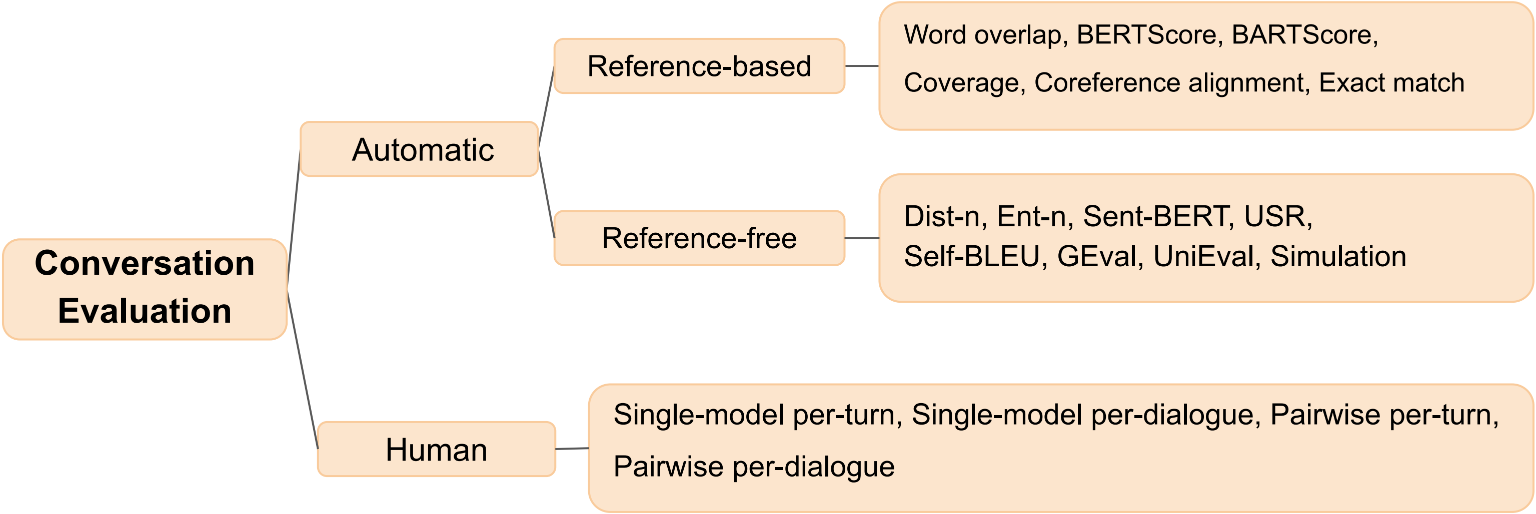
Extrinsic Evaluation

Train the dialogue model with synthetically generated data and evaluate the performance on downstream tasks

Intrinsic Evaluation

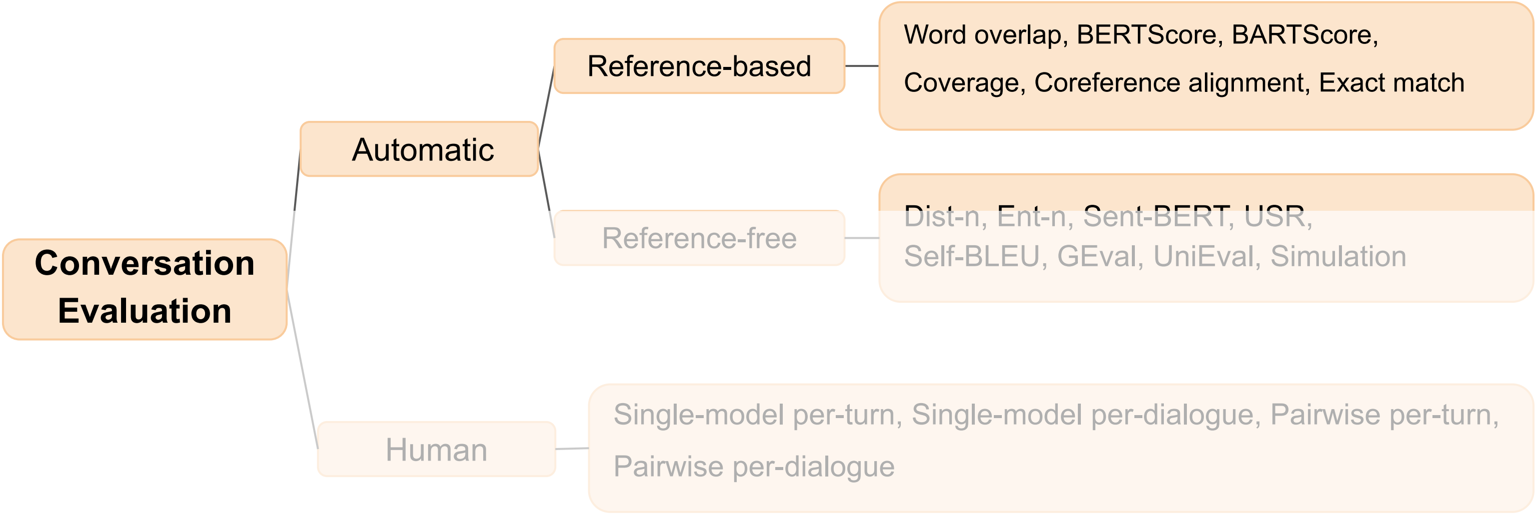
Evaluate directly the quality of generated dialogue

- Human evaluation
- Automatic evaluation



DISCLAIMER

The list is non-exhaustive and each paper uses some of these metrics.


Conversation Evaluation
Automatic
Reference-based
Word overlap, BERTScore, BARTScore, Coverage, Coreference alignment, Exact match
Reference-free
Dist-n, Ent-n, Sent-BERT, USR, Self-BLEU, GEval, UniEval, Simulation
Human
Single-model per-turn, Single-model per-dialogue, Pairwise per-turn, Pairwise per-dialogue

Automatic Reference-based Evaluation

- **Word overlap metrics:**
  - E.g., BLEU (1-3), ROUGE-L (R-L), METEOR, etc.

- **Embedding-based metrics:**
  - E.g., BERTScore and BARTScore (Zhang et al., 2020, (Yuan, et al., 2021)
  - Similarity between the generated and reference text using contextual embeddings

- **Subtask evaluation metrics:**
  - E.g., Coverage, Coreference alignment, Exact match

    (Wu et al., 2022, Kim et al., 2021, Gao et al., 2019)

BERTScore

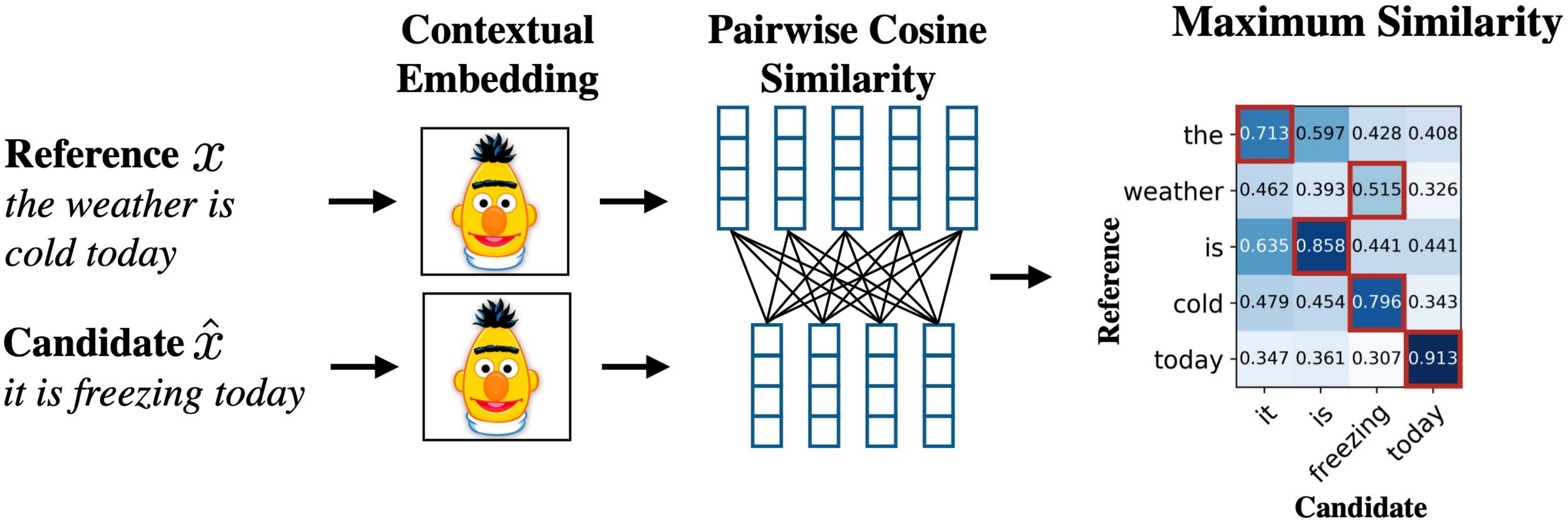
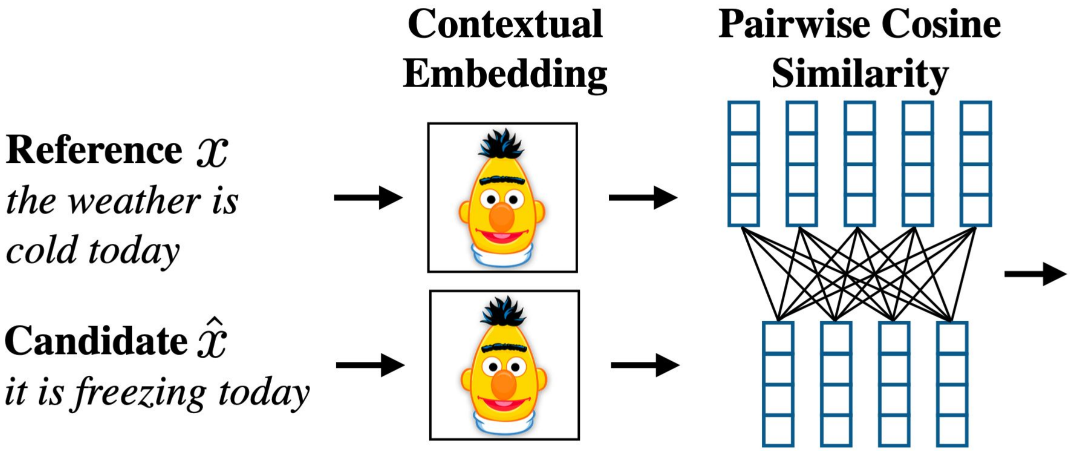


Image: (Zhang et al., 2020)

BERTScore



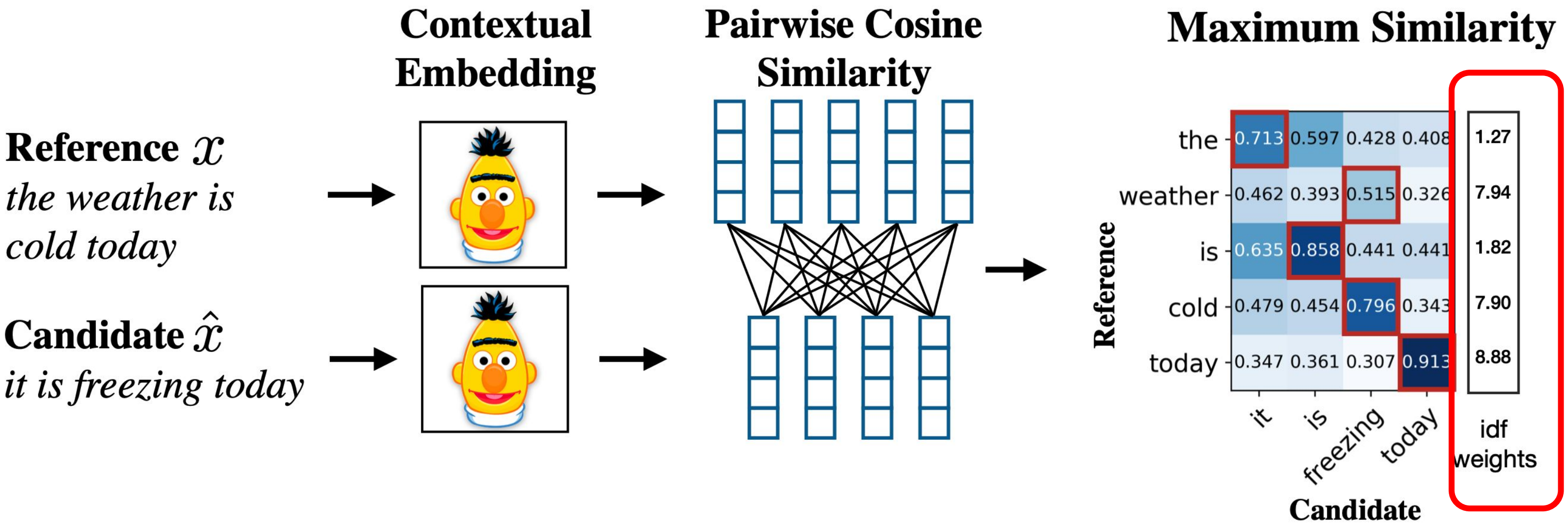
$$R_{\mathrm{BERT}} = \frac{1}{|x|} \sum_{x_i \in x} \max_{\hat{x}_j \in \hat{x}} \mathbf{x}_i^\top \hat{\mathbf{x}}_j$$

$$P_{\mathrm{BERT}} = \frac{1}{|\hat{x}|} \sum_{\hat{x}_j \in \hat{x}} \max_{x_i \in x} \mathbf{x}_i^\top \hat{\mathbf{x}}_j$$

$$F_{\mathrm{BERT}} = 2 \frac{P_{\mathrm{BERT}} \cdot R_{\mathrm{BERT}}}{P_{\mathrm{BERT}} + R_{\mathrm{BERT}}}$$

Pre-normalized vectors to reduce the calculation to the inner product

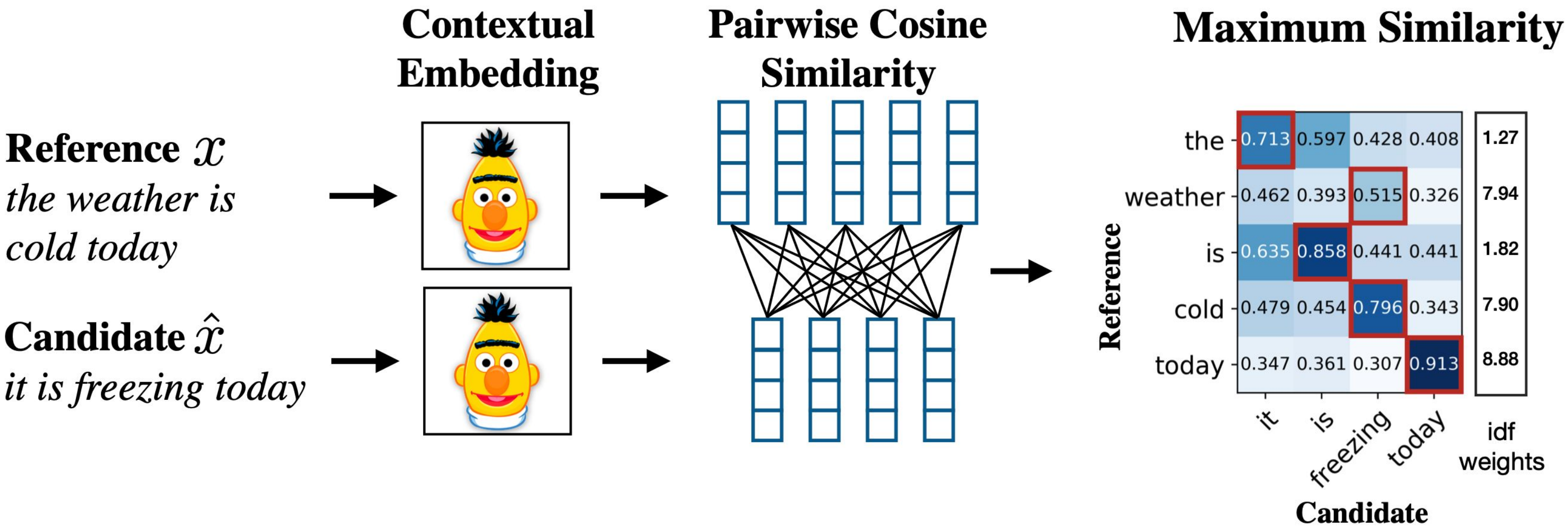
BERTScore - Optional IDF Weighting



The effect is marginal and dependent on the domain and test data

BERTScore

- Strong segment-level correlation with human
- Ineffective at dealing with conversations



Subtask Evaluation Metrics

Span Coverage

- How much the extracted spans cover the original documents
- Dialogue generation models trained on spans with higher span coverage perform better

$$\text{Coverage} = \frac{\sum_{\text{span}} |\bigcup_{d \in \text{doc}_i} \bigcup_{s \in d} s|}{|\text{document}_i|}$$

S: span within document

(Wu et al., 2022)

Span Match

- Exact Match: the predicted span exactly matches the reference span
- F1 of span n-grams

(Kim et al., 2022)

Correference alignment

- Precision, Recall, and F1 of pronouns

(Gao et al., 2019)

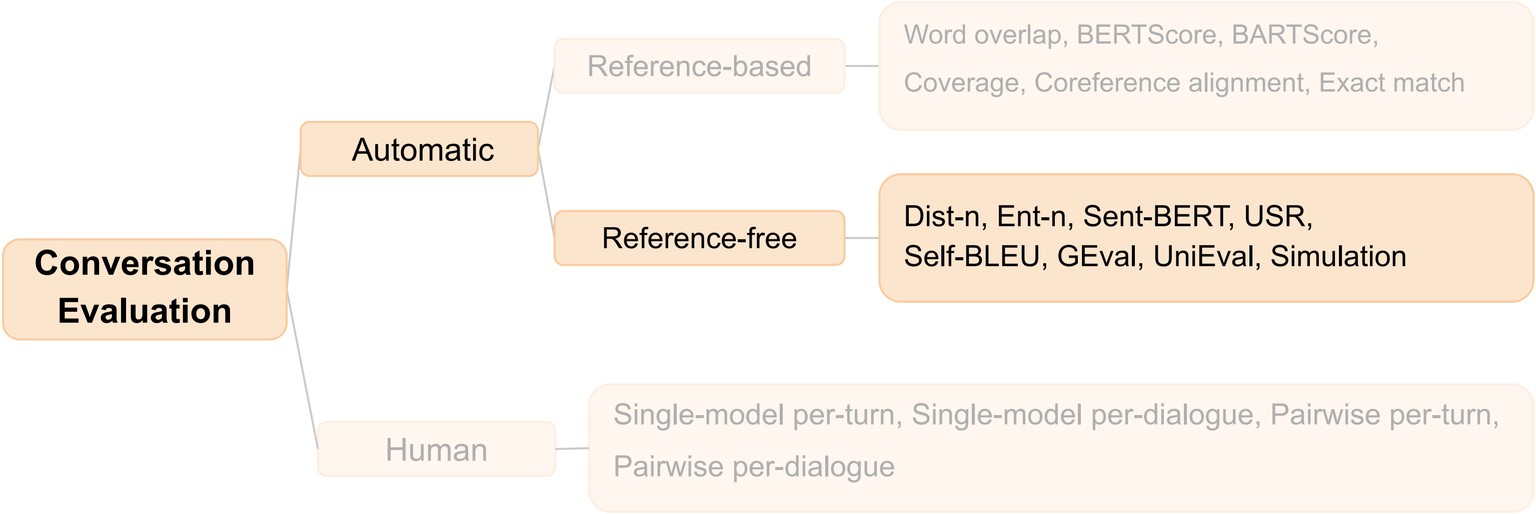
Subtask Evaluation Metrics - TOD

Turn-based evaluation:

- On intent-level: Active Intent Accuracy
- On slot-level: Requested slot F1
- Zero-shot Coverage: Measures the accuracy ratio between zero-shot learning outcomes and a fully trained model (Kim et al., 2021)

Conversation evaluation:

- On goal-level: Success Rate, Completion Rate, Book Rate, Inform Prec/Rec/F1


Conversation Evaluation
Automatic
Reference-based
Word overlap, BERTScore, BARTScore, Coverage, Coreference alignment, Exact match
Reference-free
Dist-n, Ent-n, Sent-BERT, USR, Self-BLEU, GEval, UniEval, Simulation
Human
Single-model per-turn, Single-model per-dialogue, Pairwise per-turn, Pairwise per-dialogue

Automatic Reference-free Evaluation

Diversity Metrics:

- Dist-n (Li et al., 2016)
  - Number of distinct unigrams and bigrams / total number of generated words.
- Ent-n (Zhang et al., 2018)
  - How evenly the n-gram distribution is over all generated questions
- Sent-BERT (Reimers et al., 2019)
  - The average negative cosine similarity between SentenceBERT embedding for each pair of responses
- Self-BLEU (Zhu et al., 2018)
  - Uses one sentence from a set as a hypothesis and the rest as references, calculating a BLEU score for each sentence. The average of these scores is termed Self-BLEU

Mind length normalization in Diversity metrics!

USR: **U**n**S**upervised and **R**eference-free metric for dialog

Consists of five sub-metrics, combined to measure the ***Overall Quality*** metric.

Understandable	Response being understandable given the previous context
Natural	Response being similar to what a person would naturally say
Maintains Context	Response being a valid continuation of the conversation
Interesting	Dull or interesting response
Uses Knowledge	Response using a given fact

(Mehri et al., 2020)

USR: **U**n**S**upervised and **R**eference-free metric for dialog

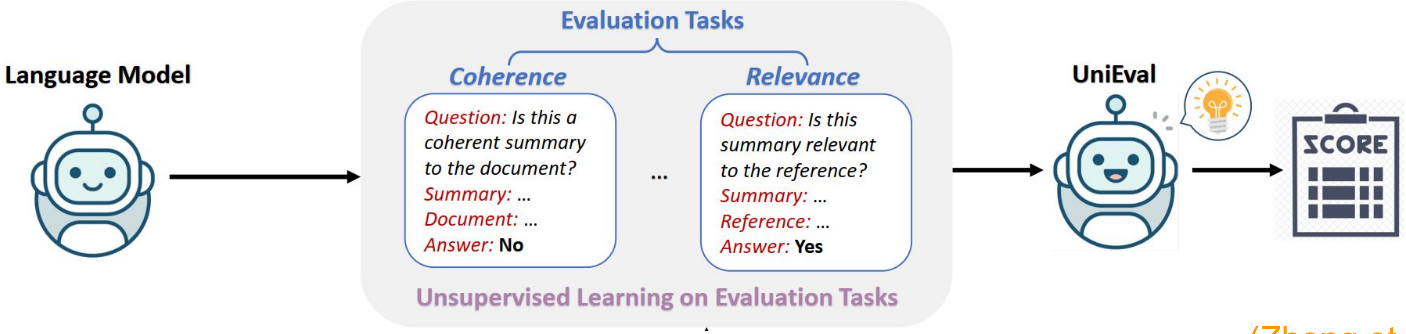
Uses RoBERTa, fine tuned on dialogue corpus used for evaluation.

Understandable	r: response i: i-th word of response l_i: mask log likelihood of word *i*	$$-\sum_{i}^{\lvert r \rvert} l_i$$
Natural		
Maintains Context	RoBERTa further fine tuned to predict *P(y=1|x, r)* *y:* whether r is true response or randomly sampled *x*: dialogue history and/or the fact	
Interesting		
Uses Knowledge		
Overall Quality	Combines sub-metrics using a regression model trained on human annotation	

(Mehri et al., 2020)

UniEval

- An aspect-based reference-free evaluator for NLG tasks
- Casts each evaluation aspect to a Boolean QA problem:
  - Coherence: "Is this a coherent summary of the document?"
- Intermediate training of T5 for each task (similar to USR aspects for conversations)

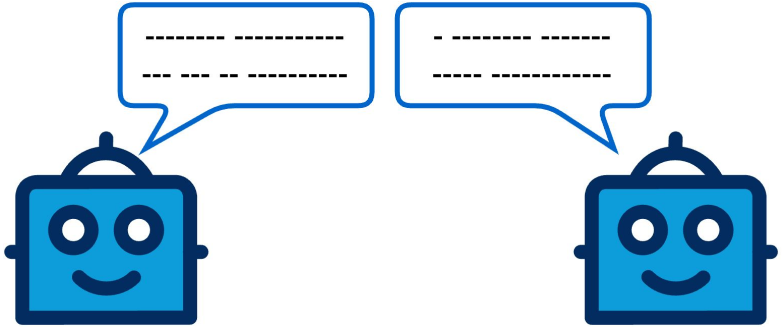


(Zhong et al., 2022)

Automatic Simulation-based Evaluation

- Used for evaluating (target-guided) open domain dialogue systems
- Two dialogue agents converse with each other
- Automatically measures the **success rate** of achieving the target
- Often a max. allowed number of turn is set

Agent role:
Randomly picks a target
and starting point



Human role:
converse with agent without
knowing the target

(Tang et al., 2019)

Conversation Evaluation

- Automatic
  - Reference-based
    - Word overlap, BERTScore, BARTScore, Coverage, Coreference alignment, Exact match
  - Reference-free
    - Dist-n, Ent-n, Sent-BERT, USR, Self-BLEU, GEval, UniEval, Simulation
- Human
  - Single-model per-turn, Single-model per-dialogue, Pairwise per-turn, Pairwise per-dialogue

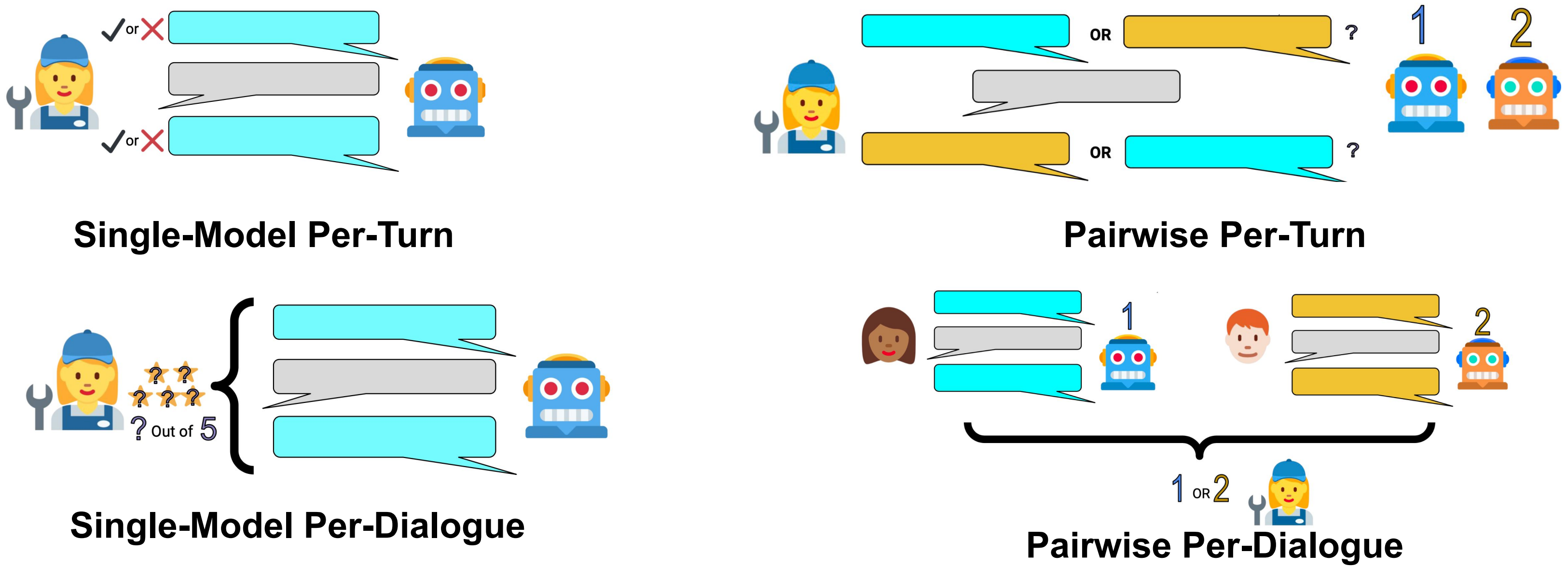
Human Evaluation

- **Evaluation criteria**
  - Naturalness, Informativeness, context relevance, answer accuracy, etc.
  - Overall quality

- **Method of evaluation**
  - **Single-model:** Assigning integer scores (e.g., 1-3) for a question/dialogue
  - **Pair-wise:** Comparing two responses/dialogues and select the best one
  - **Turn-level:** Human rating after every system response
  - **Dialogue-level:** Human rating at the end of conversation

Human evaluations are not comparable across different experiments and papers.

Human Evaluation Methods - Comparison



- Comparison on three aspects: Preference, Humanness, Interestingness
- Three model comparison types: Length, parameter size, Fine-tuning

Image: (Michael Smith et al., 2022)

Human Evaluation Methods - Comparison

- **Per-turn evaluation:** More fine-grained, can capture small differences
- **Pairwise per-turn evaluation:** Performs best on fine tuning comparison
  - Differences in models' replies are easily detectable
- **Pairwise per-dialogue evaluation:** Performs best on length comparison
  - Differences appear after several conversation turns
- **Single model evaluation:** Performs best on model size comparison (#params)
  - Slight differences in quality

(Michael Smith et al., 2022)