Part 1: Evaluation

Duration: 20 min

Presenter: Faegheh Hasibi

Synthetic Conversation Evaluation

Intrinsic Evaluation

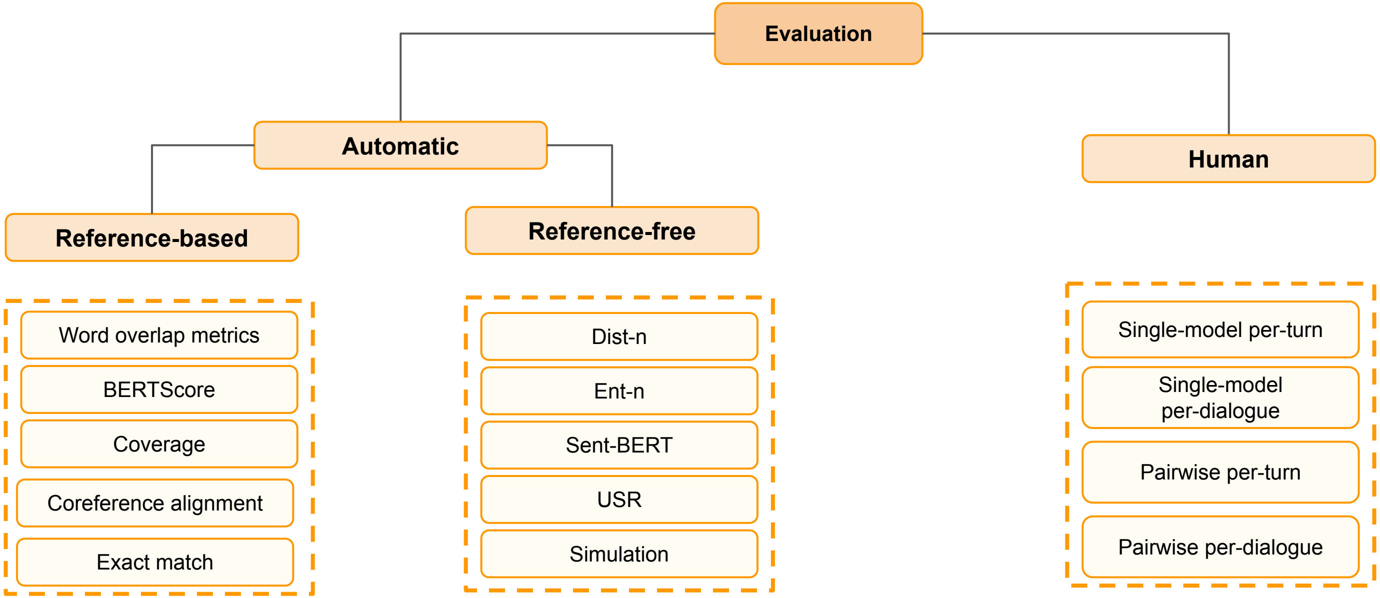
Evaluate directly the quality of generated dialogue

- Automatic evaluation
- Human evaluation

Extrinsic Evaluation

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

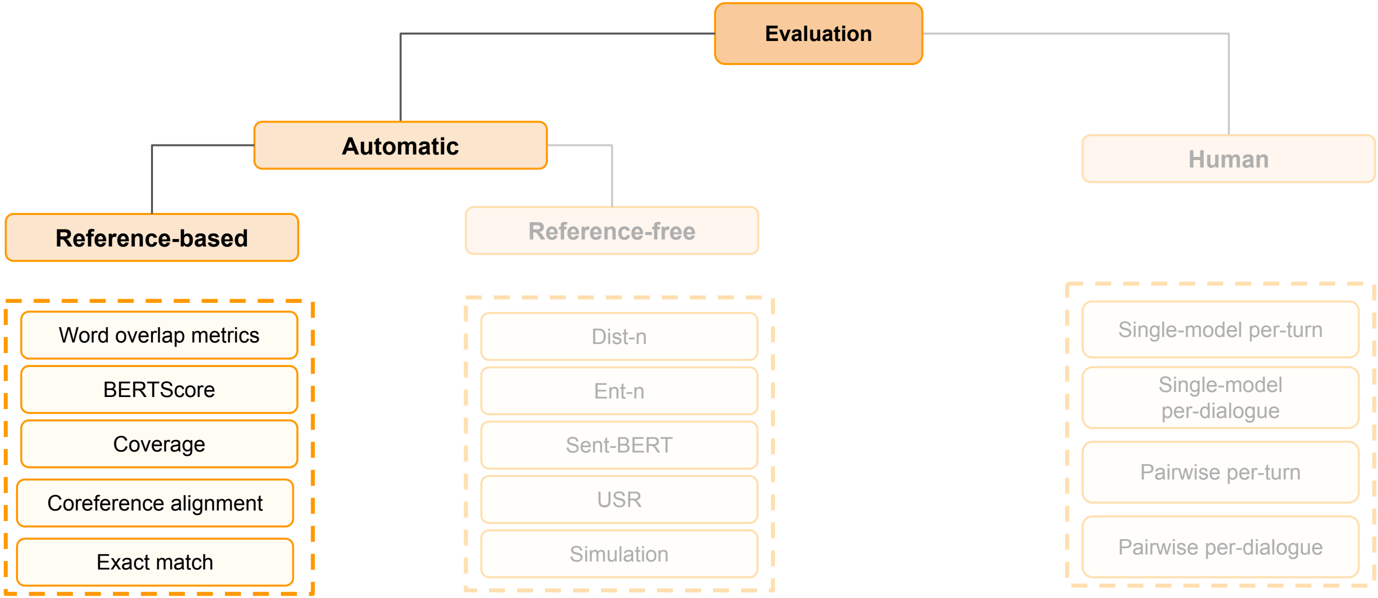
Overview



DISCLAIMER

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

Overview



Evaluation
Automatic
Human
Reference-based
Reference-free
Word overlap metrics
BERTScore
Coverage
Coreference alignment
Exact match
Dist-n
Ent-n
Sent-BERT
USR
Simulation
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:**
- ◆ BERTScore: Similarity between the generated and reference text using contextual embeddings

➔ **Subtask evaluation metrics:**
- ◆ E.g., Span coverage, Coreference alignment, Exact match, etc.

BERTScore

Computes a soft measure of similarity of using BERT.

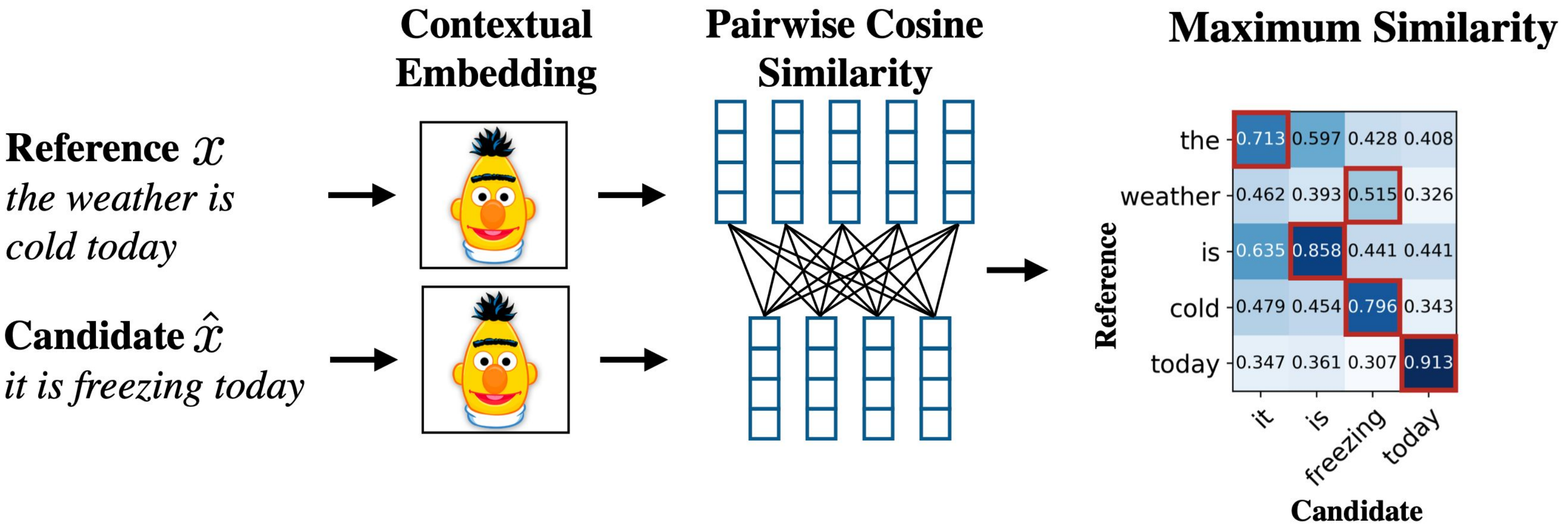
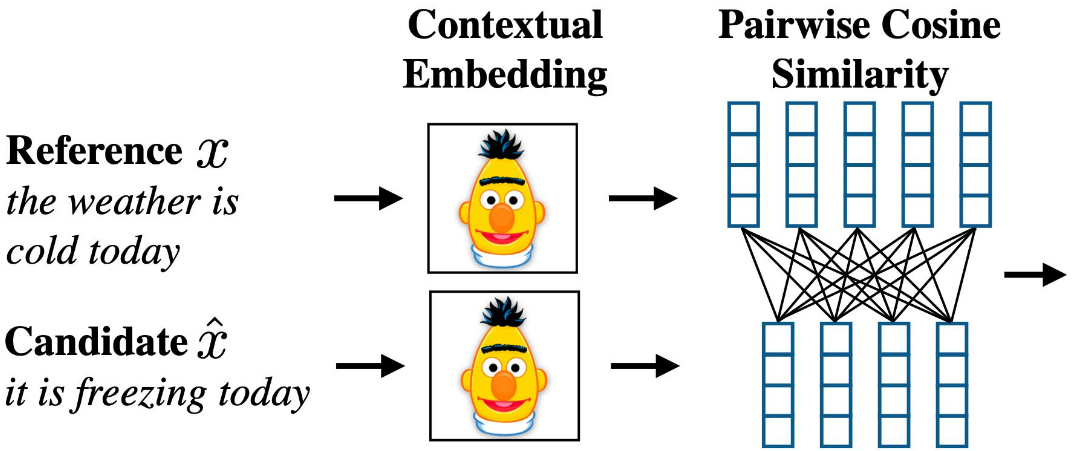


Image: Zhang et al., 2020

BERTScore



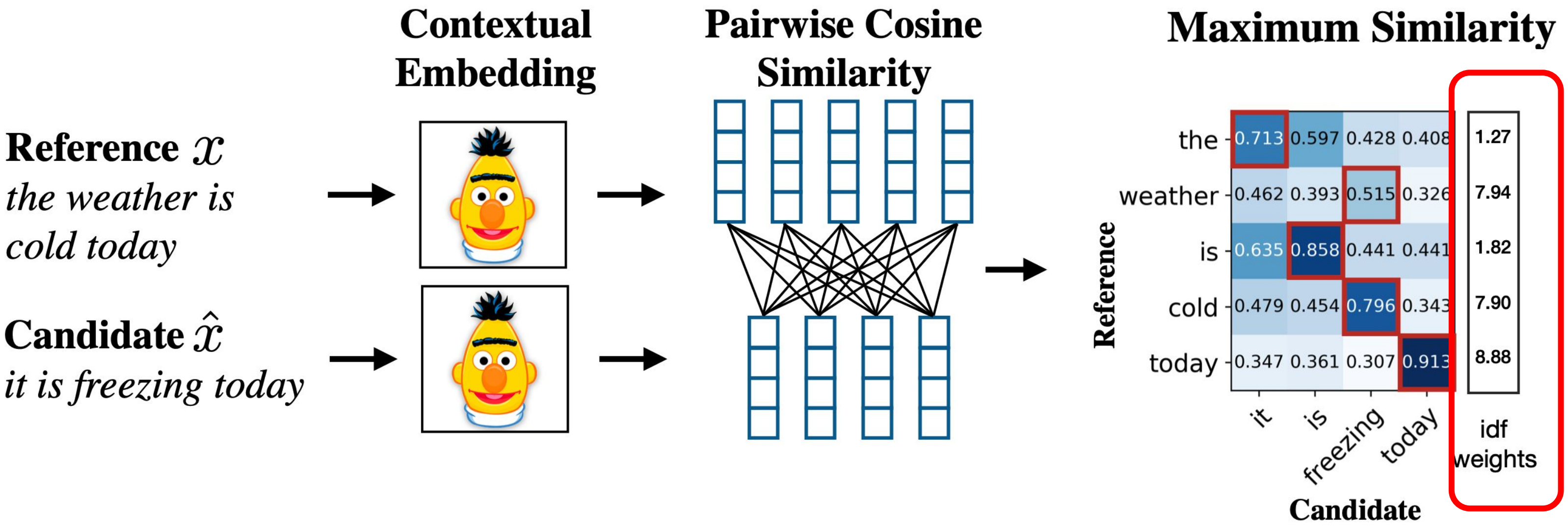
$$R_{\mathrm{BERT}} = \frac{1}{|x|} \sum_{x_i \in x} \max_{\hat{x}_j \in \hat{x}} \mathbf{x}_i^\top \hat{\mathbf{x}}_j$$

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

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

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

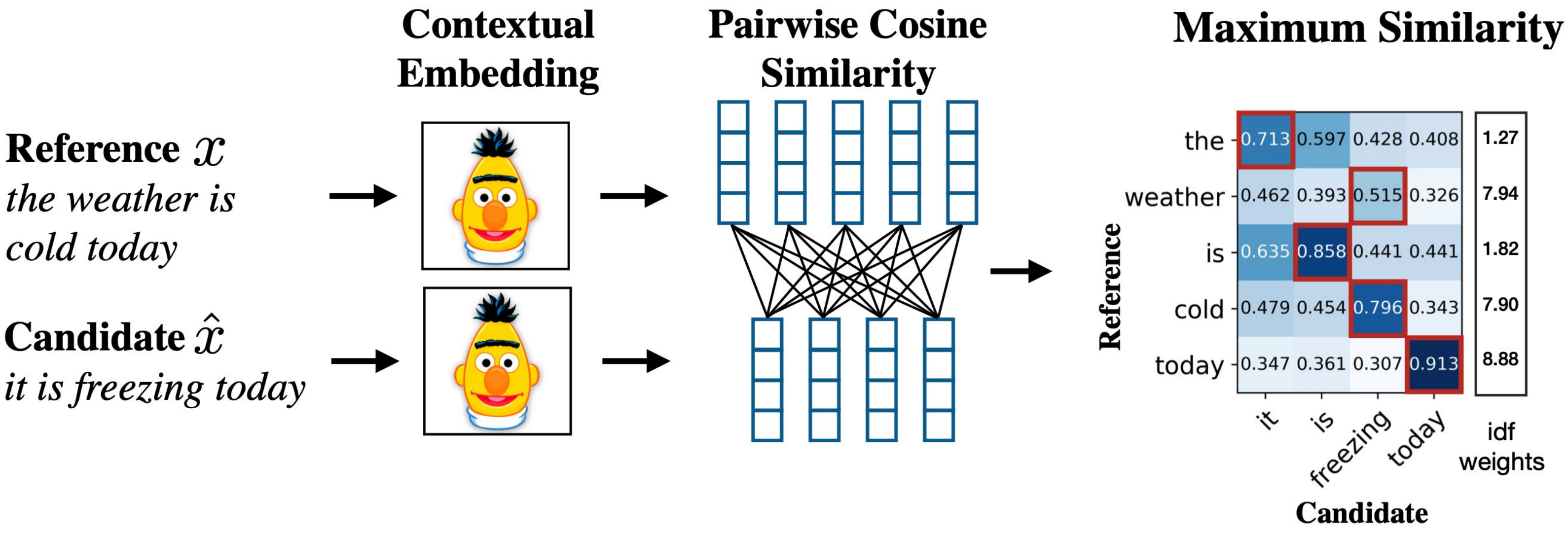
BERTScore - Optional IDF Weighting



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., 2021)

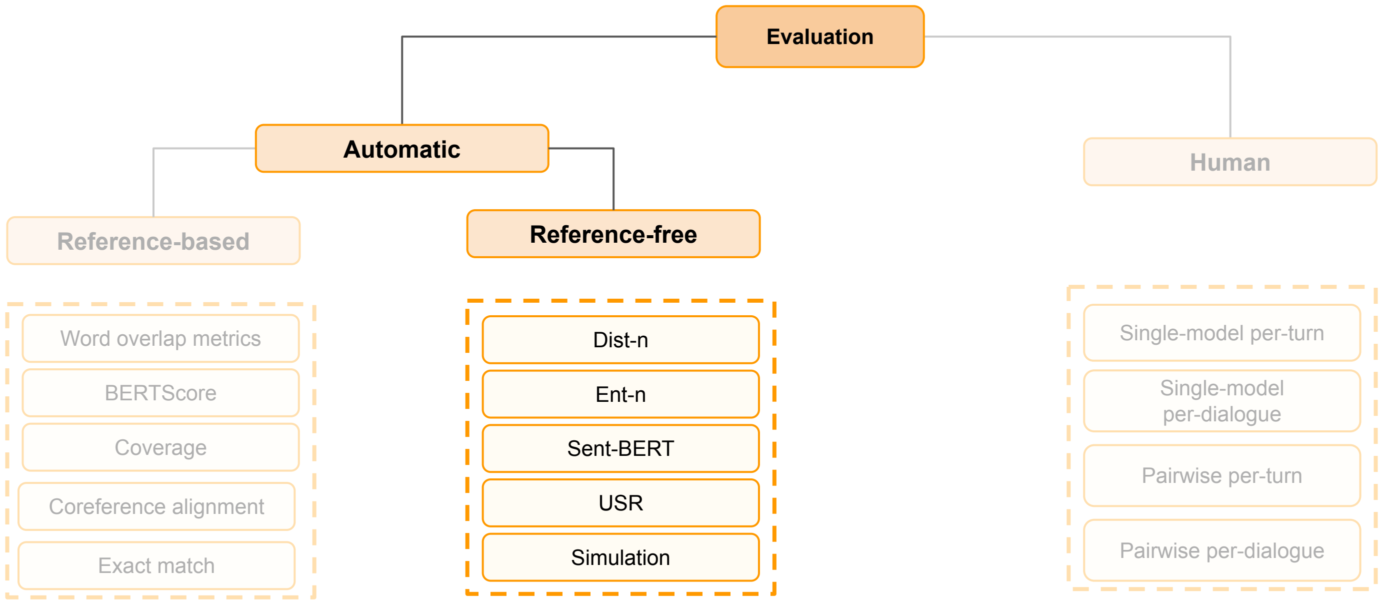
Span Match

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

Correference alignment

- Precision, Recall, and F1 of pronouns

Overview


Evaluation
Automatic
Human
Reference-based
Reference-free
Word overlap metrics
BERTScore
Coverage
Coreference alignment
Exact match
Dist-n
Ent-n
Sent-BERT
USR
Simulation
Single-model per-turn
Single-model per-dialogue
Pairwise per-turn
Pairwise per-dialogue

Automatic Reference-free Evaluation

- **Diversity metrics:**
  - Dist-n: number of distinct unigrams and bigrams / total number of generated words.
  - Ent-n: how evenly the n-gram distribution is over all generated questions
  - Sent-BERT: the average negative cosine similarity between SentenceBERT embedding for each pair of responses
  - etc.
- **Dialogue quality metrics:**
  - Learned metrics such as USR

USR: UnSupervised and Reference-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 and Eskenazi., 2020)

USR: UnSupervised and Reference-free metric for dialog

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

<table>
<tr><td>Understandable</td><td rowspan="2">r: response
i: i-th word of response
l_i: mask log likelihood of word i
$$-\sum_{i}^{|r|} l_i$$</td></tr>
<tr><td>Natural</td></tr>
<tr><td>Maintains Context</td><td rowspan="3">RoBERTa further fine tuned to predict $P(y=1|x, r)$

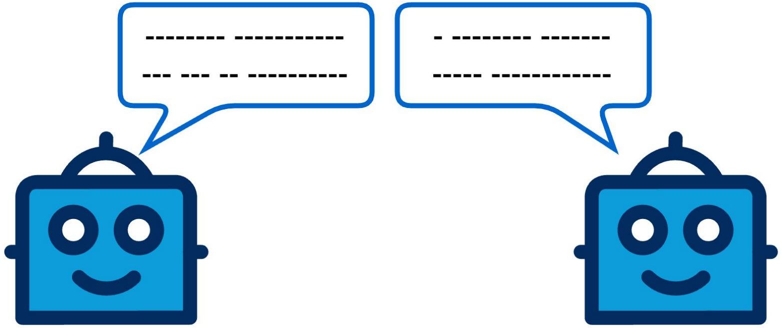
y: whether r is true response or randomly sampled
x: dialogue history and/or the fact</td></tr>
<tr><td>Interesting</td></tr>
<tr><td>Uses Knowledge</td></tr>
<tr><td>Overall Quality</td><td>Combines sub-metrics using a regression model trained on human annotation</td></tr>
</table>

(Mehri and Eskenazi., 2020)

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

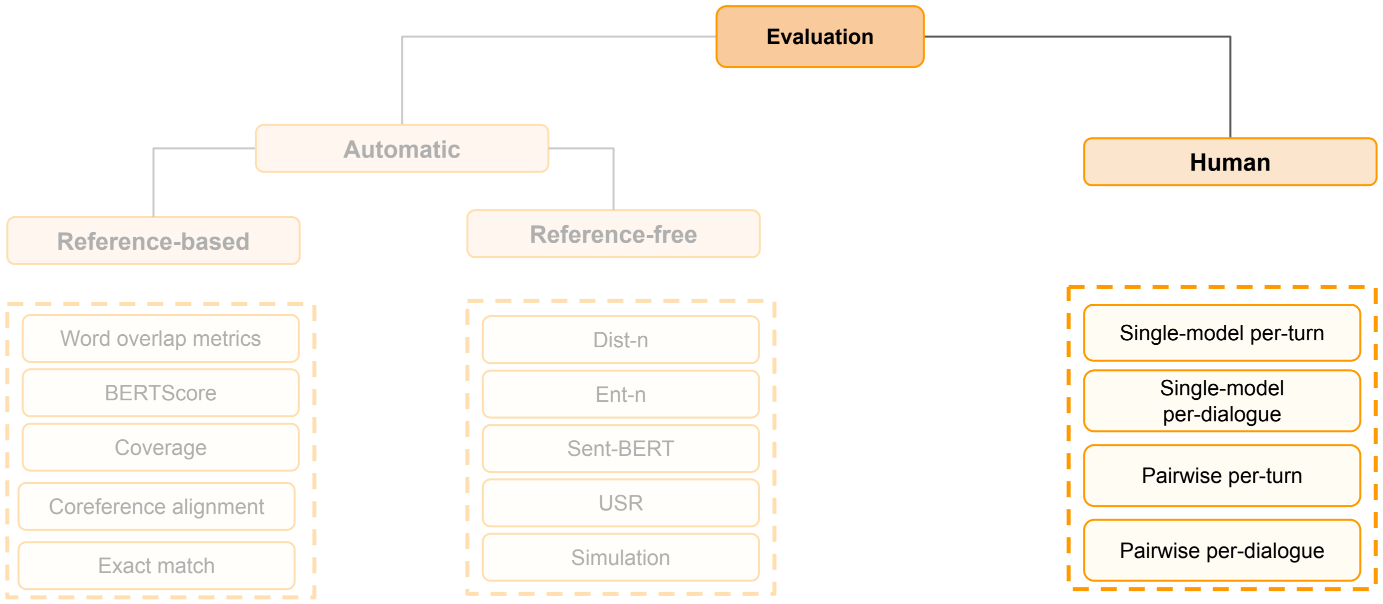
Agent role:
Randomly picks a target
and starting point



Human role:
converse with agent without
knowing the target

(Tang et al., 2023)

Overview

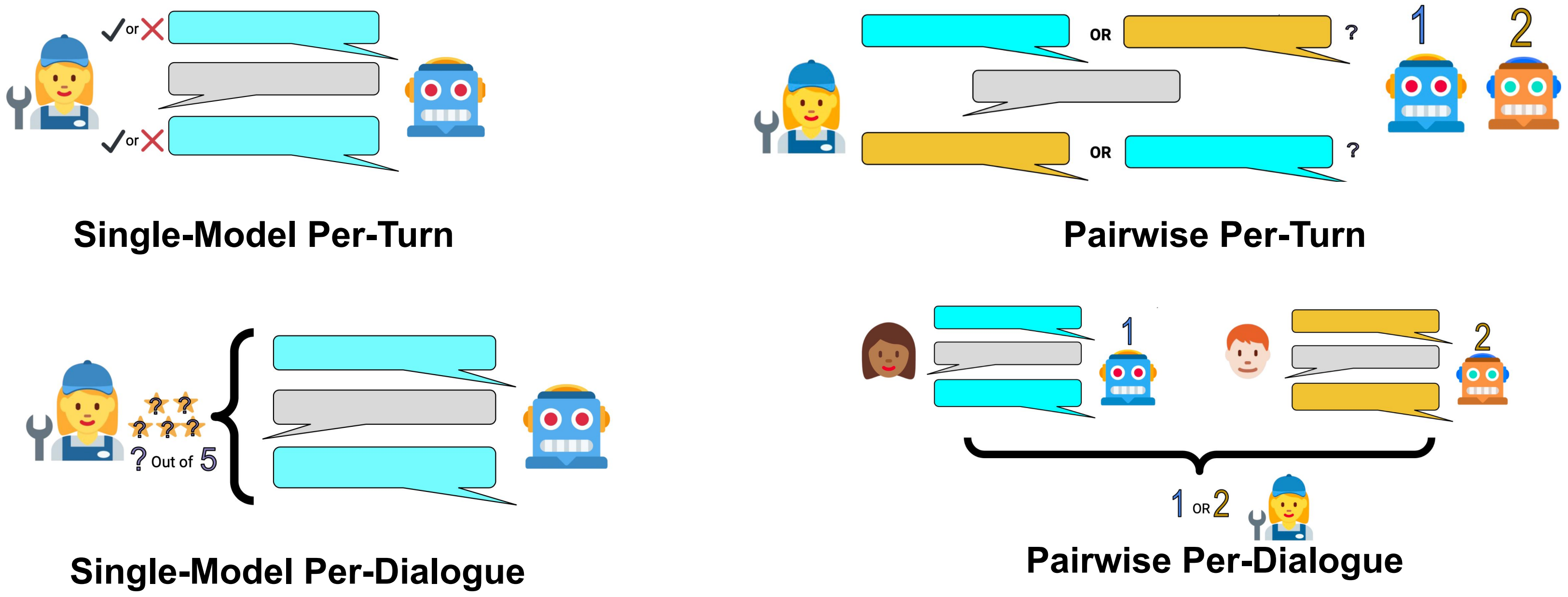

Evaluation
Automatic
Human
Reference-based
Reference-free
Word overlap metrics
BERTScore
Coverage
Coreference alignment
Exact match
Dist-n
Ent-n
Sent-BERT
USR
Simulation
Single-model per-turn
Single-model per-dialogue
Pairwise per-turn
Pairwise per-dialogue

Human Evaluation

- **Evaluation criteria**
  - Naturalness,Informativeness, context relevance, answer accuracy, etc.
  - Overall quality
- **Method of evaluation**
  - Single-model: Assign integer scores (e.g., 1-3) for a question/dialogue
  - Pair-wise: Compare two responses/dialogues and select the best one
  - Ranking: Provide a ranking of (>2) systems for a given evaluation criteria
- **Turn-level Evaluation vs. Dialogue-level Evaluation**

Human evaluations are not comparable across different experiments and papers.

Human Evaluation Methods - Comparison



Methods are compared on three aspects: Preference, Humanness, Interestingness

Image: Michael Smith et al., 2022

Human Evaluation Methods - Comparison

- Pairwise per-turn evaluation tends to work well when differences in models' replies are easily detectable
  - E.g., training models on different datasets
- Pairwise per-dialogue evaluation performs best when model differences appear after several conversation turns
  - E.g., a pattern in average length of conversation
- Single-model evaluation perform well when comparing models that are similar, only differ slightly in quality
  - E.g., models with different numbers of parameters

(Michael Smith et al., 2022)