Recent Advances on Automatic Dialogue Evaluation

Luis Fernando D’Haro - @lfdharo

Chen Zhang – NUS (Singapore)

Fall School RTTH – Jaca, Nov 14-17, 2023

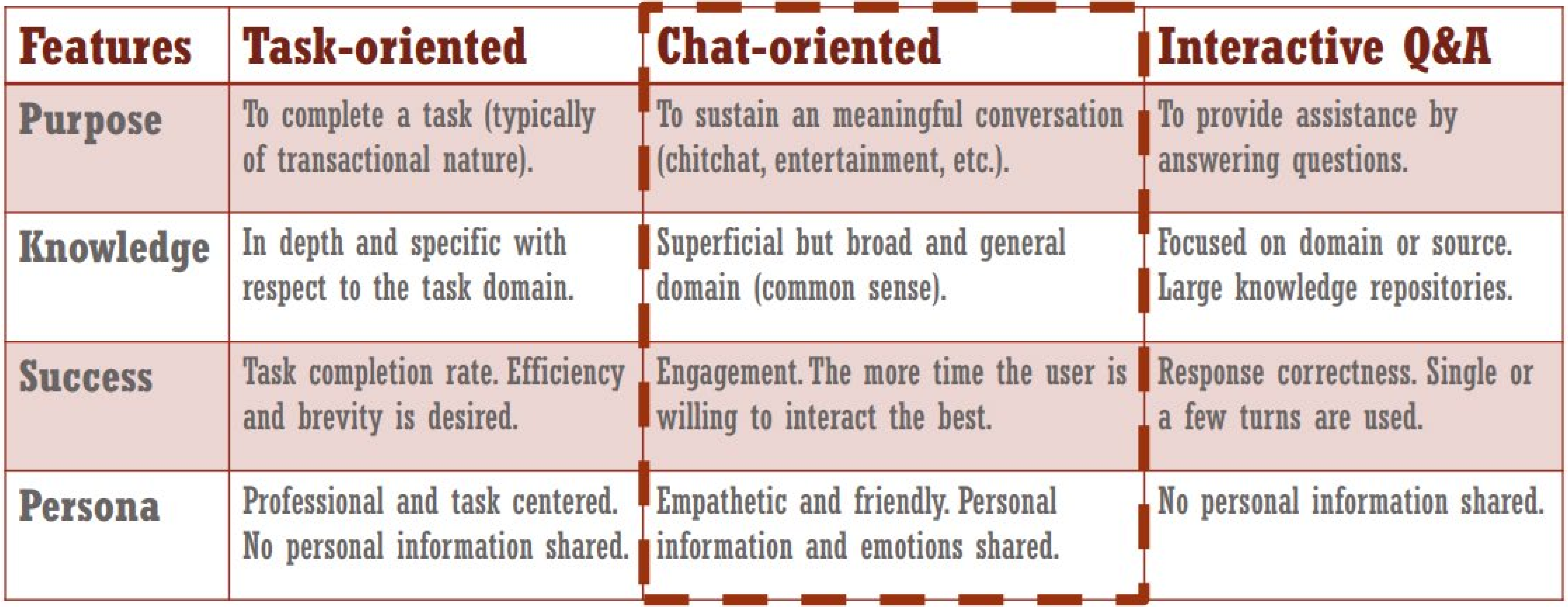
Grupo de Tecnología del Habla y Aprendizaje Automático - ETSI de Telecomunicación – Universidad Politécnica de Madrid

Tutorial Overview

- Introduction (15 min)
  - Overview on dialogue systems
  - Measuring progress: Human & Automatic Evaluation
- Reference-based Metrics (30 min)
  - Untrained Metrics
  - Trained Metrics
- Reference-free Metrics (50 min)
  - Untrained Metrics
  - Trained Metrics
- Challenges & Future Directions (10 min)
  - Challenges and needs
- Conclusion
- Hands-on (Google Colab)

Introduction

Dialogue System Overview



Features	Task-oriented	Chat-oriented	Interactive Q&A
Purpose	To complete a task (typically of transactional nature).	To sustain an meaningful conversation (chitchat, entertainment, etc.).	To provide assistance by answering questions.
Knowledge	In depth and specific with respect to the task domain.	Superficial but broad and general domain (common sense).	Focused on domain or source. Large knowledge repositories.
Success	Task completion rate. Efficiency and brevity is desired.	Engagement. The more time the user is willing to interact the best.	Response correctness. Single or a few turns are used.
Persona	Professional and task centered. No personal information shared.	Empathetic and friendly. Personal information and emotions shared.	No personal information shared.

- Deriu et al. "Survey on evaluation methods for dialogue systems." Artificial Intelligence Review 54.1 (2021): 755-810.

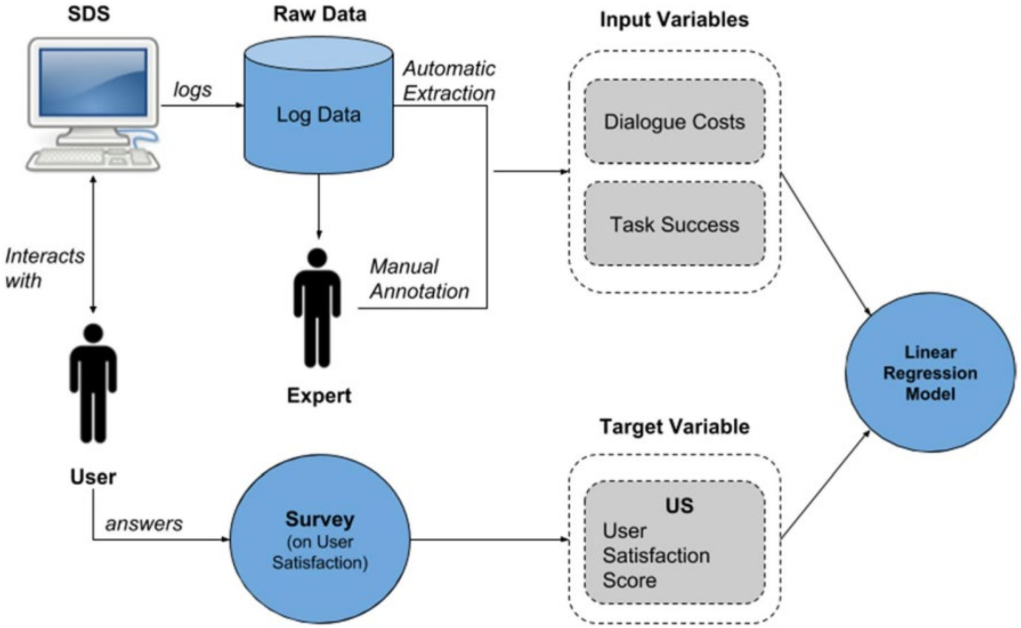
Dialogue System Overview - Task-Oriented Dialogue (TOD)

- Methodology
  - Combination of <u>rules</u> and <u>statistical</u> components (Young et al., 2013)
  - End-to-end approaches
    - End-to-end trainable task-oriented dialogue system (Wen et al., 2017)
    - End-to-end reinforcement learning dialogue system (Li et al., 2017; Zhao and Eskenazi, 2016)
    - Leveraging pre-trained language models (Ham et al., 2020)

- Young et al. "Pomdp-based statistical spoken dialog systems: A review." Proceedings of the IEEE 101.5 (2013).
- Wen et al. "A Network-based End-to-End Trainable Task-oriented Dialogue System." EACL (2017).
- Li et al. "End-to-End Task-Completion Neural Dialogue Systems." IJCNLP (2017).
- Zhao and Eskenazi. "Towards End-to-End Learning for Dialog State Tracking and Management using Deep Reinforcement Learning." SIGDial (2016).
- Ham, Donghoon, et al. "End-to-end neural pipeline for goal-oriented dialogue systems using GPT-2." ACL (2020).

Dialogue System Overview - Task-Oriented Dialogue (TOD)

- Evaluation
  - Two main aspects are measured - task-success and dialogue efficiency
  - User satisfaction modeling - the PARADISE framework (Walker et al., 1997)
    - Domain-independent, based on user ratings on the dialogue-level
    - Predict user satisfaction score based on linear regression of different input variables: ASR results, time, dialogue length, goal completion, user’s feedback, etc.



- Walker et al. "PARADISE: A Framework for Evaluating Spoken Dialogue Agents." ACL (1997).

Dialogue System Overview - Task-Oriented Dialogue (TOD)

- Other Evaluation Metrics
  - NLU evaluation - sentence level semantic accuracy (SLSA); slot error rate (SER); F-measures
  - DST evaluation - joint goal accuracy (JGA)
  - NLG evaluation
    - Correctness - F1 score
    - Quality of surface realization - BLEU (Papineni et al., 2002), ROUGE (Lin, 2004)

- Papineni, Kishore, et al. "BLEU: a method for automatic evaluation of machine translation." ACL (2002).
- Lin, Chin-Yew. "ROUGE: A package for automatic evaluation of summaries." Text summarization branches out (2004).

Dialogue System Overview - Open-Domain Dialogue (ODD)

- Retrieval-based methodology
  - Dual-encoder: LSTM (Lowe et al., 2015), ConveRT (Henderson et al., 2020)
  - Cross-encoder: BERT-based (Han et al., 2021)
  - Poly-encoder (Humeau et al., 2020)
- Evaluation of retrieval-based approaches
  - F1-score, Recall@k, Mean reciprocal rank

$$\mathrm{MRR} = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{\mathrm{rank}_i}.$$

- Lowe et al. "The Ubuntu Dialogue Corpus: A Large Dataset for Research in Unstructured Multi-Turn Dialogue Systems." SIGDial (2015).
- Henderson et al. "ConveRT: Efficient and Accurate Conversational Representations from Transformers." Findings of EMNLP (2020).
- Han et al. "Fine-grained Post-training for Improving Retrieval-based Dialogue Systems." NAACL (2021).
- Humeau et al. “Poly-encoders: Architectures and Pre-training Strategies for Fast and Accurate Multi-sentence Scoring.” ICLR (2020)

Dialogue System Overview - Open-Domain Dialogue (ODD)

- End-to-end generative approaches
  - RNN-based Seq2Seq models
    - HRED (Serban et al., 2016); VHRED (Serban et al., 2017)
  - Transformer-based decoder-only models
    - Transfer learning with GPT (Golovanov et al., 2016); DialoGPT (Zhang et al., 2020); LaMDA (Thoppilan et al., 2022)
  - Transformer-based encoder-decoder models
    - PLATO (Bao et al., 2020); Meena (Adiwardana et al., 2020); Blender (Roller et al., 2021), ChatGPT

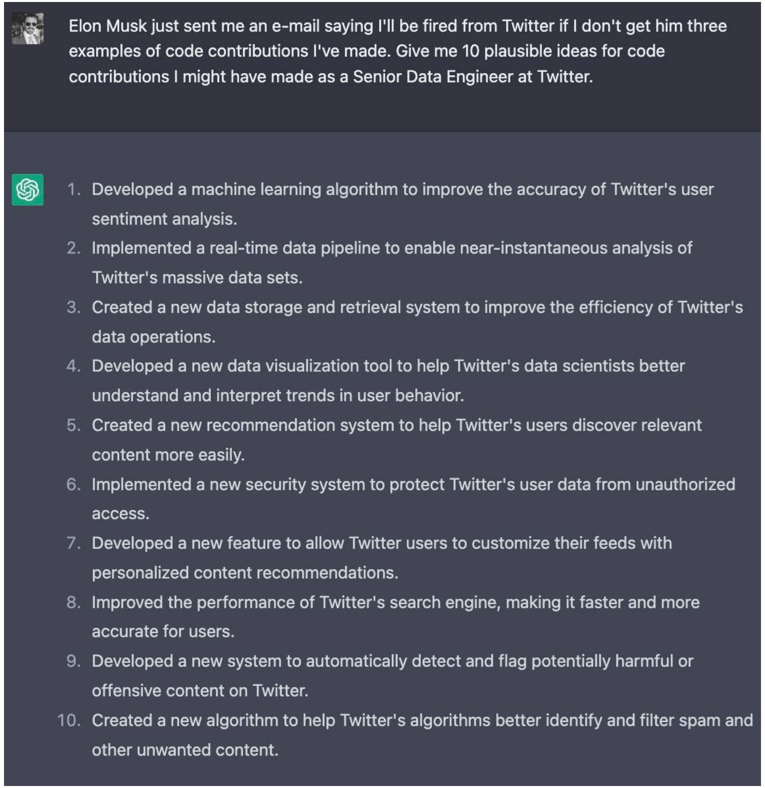
- Serban et al. "Building end-to-end dialogue systems using generative hierarchical neural network models." AAAI (2016).
- Serban et al. "A hierarchical latent variable encoder-decoder model for generating dialogues." AAAI (2017).
- Golovanov et al. "Large-scale transfer learning for natural language generation." ACL (2019).
- Zhang et al. "DIALOGPT: Large-Scale Generative Pre-training for Conversational Response Generation." ACL System Demonstrations. (2020).
- Thoppilan et al. "Lamda: Language models for dialog applications." arXiv preprint arXiv:2201.08239 (2022).
- Bao et al. "PLATO: Pre-trained Dialogue Generation Model with Discrete Latent Variable." ACL (2020).
- Adiwardana et al. "Towards a human-like open-domain chatbot." arXiv preprint arXiv:2001.09977 (2020).
- Roller et al. "Recipes for Building an Open-Domain Chatbot." EACL (2021).

Dialogue System Overview - Open-Domain Dialogue (ODD)

- Common evaluation method of end-to-end generative approaches
  - Human evaluation
    - Likert rating at both turn and dialogue level
    - Pairwise comparison at both turn and dialogue level
  - Automatic evaluation
    - Reference-based metrics, such as BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), etc
    - Reference-free metrics, such as perplexity, USR (Mehri and Eskenazi, 2020), etc

- Papineni et al. "BLEU: a method for automatic evaluation of machine translation." ACL (2002).
- Lin, "Rouge: A package for automatic evaluation of summaries." Text summarization branches out (2004).
- Mehri and Eskenazi. "USR: An Unsupervised and Reference Free Evaluation Metric for Dialog Generation." ACL (2020).

ChatGPT - The power of large-scale pre-training



Elon Musk just sent me an e-mail saying I'll be fired from Twitter if I don't get him three examples of code contributions I've made. Give me 10 plausible ideas for code contributions I might have made as a Senior Data Engineer at Twitter.

1. Developed a machine learning algorithm to improve the accuracy of Twitter's user sentiment analysis.
2. Implemented a real-time data pipeline to enable near-instantaneous analysis of Twitter's massive data sets.
3. Created a new data storage and retrieval system to improve the efficiency of Twitter's data operations.
4. Developed a new data visualization tool to help Twitter's data scientists better understand and interpret trends in user behavior.
5. Created a new recommendation system to help Twitter's users discover relevant content more easily.
6. Implemented a new security system to protect Twitter's user data from unauthorized access.
7. Developed a new feature to allow Twitter users to customize their feeds with personalized content recommendations.
8. Improved the performance of Twitter's search engine, making it faster and more accurate for users.
9. Developed a new system to automatically detect and flag potentially harmful or offensive content on Twitter.
10. Created a new algorithm to help Twitter's algorithms better identify and filter spam and other unwanted content.

- Existing dialogue models become more and more human-like and engaging
- Are the problem of dialogues really solved from the evaluation point of view?
  - Faithfulness (whether the information conveyed is true?)
  - Dialogue safety (how do we measure whether the chatbot is safe?)
  - Long-term consistency (how do we ensure the chatbot is consistent in a long conversation?)

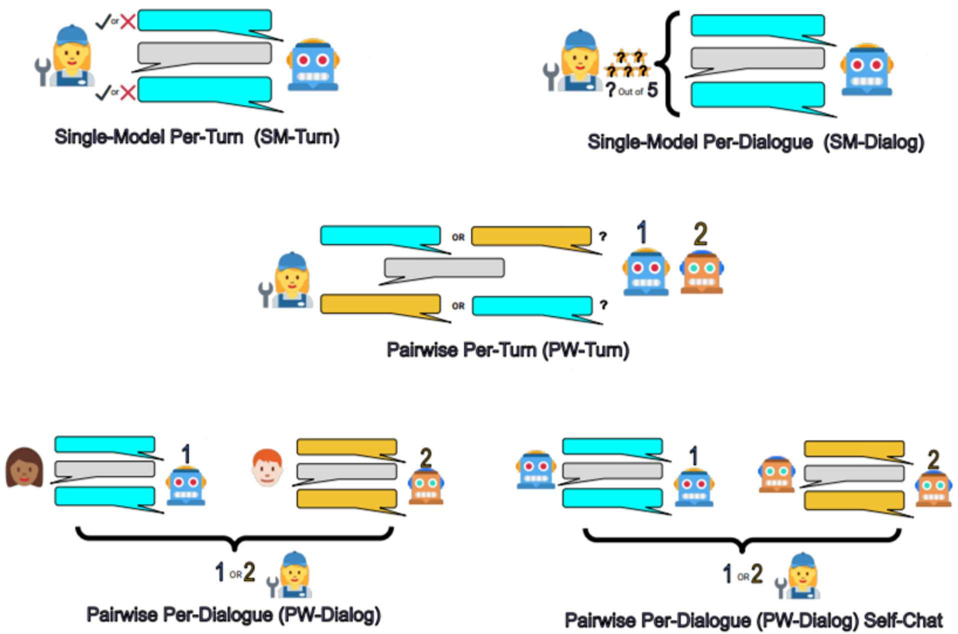
Measuring Progress

Human Evaluation

- The default option of quantifying progress in dialogue generation
  - **Levels** – Turn and Dialog
  - **Accurate** - humans possess a holistic understanding of natural language
  - **Multidimensionality** - humans are capable of judging dialogues from different perspectives
  - **Agreement** - we can rely on majority opinion vote from multiple annotators

Different settings of Human Evaluation

- Single-Model Per-Turn
  - Human annotators provide Likert ratings at turn-level
- Single-Model Per-Dialogue
  - Human annotators provide Likert ratings at dialogue-level
- Pairwise Per-Turn
  - Pairwise comparison at turn-level
- Pairwise Per-Dialogue
  - Pairwise comparison at dialogue-level (human-chatbot conversations)
- Pairwise Per-Dialogue Self-Chat
  - Pairwise comparison at dialogue-level (self-chat conversations)



Single-Model Per-Turn (SM-Turn)

Single-Model Per-Dialogue (SM-Dialog)

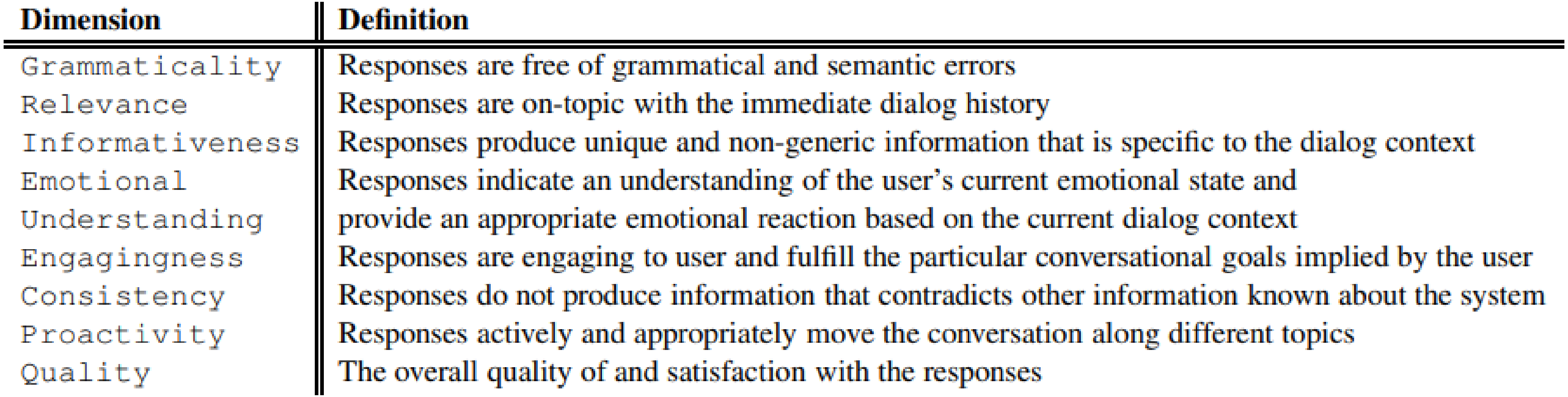
Pairwise Per-Turn (PW-Turn)

Pairwise Per-Dialogue (PW-Dialog)

Pairwise Per-Dialogue (PW-Dialog) Self-Chat

- Smith et al. "Human Evaluation of Conversations is an Open Problem: comparing the sensitivity of various methods for evaluating dialogue agents." NLP4ConvAI (2022).

Human Evaluation - Turn-Level Evaluation Criteria

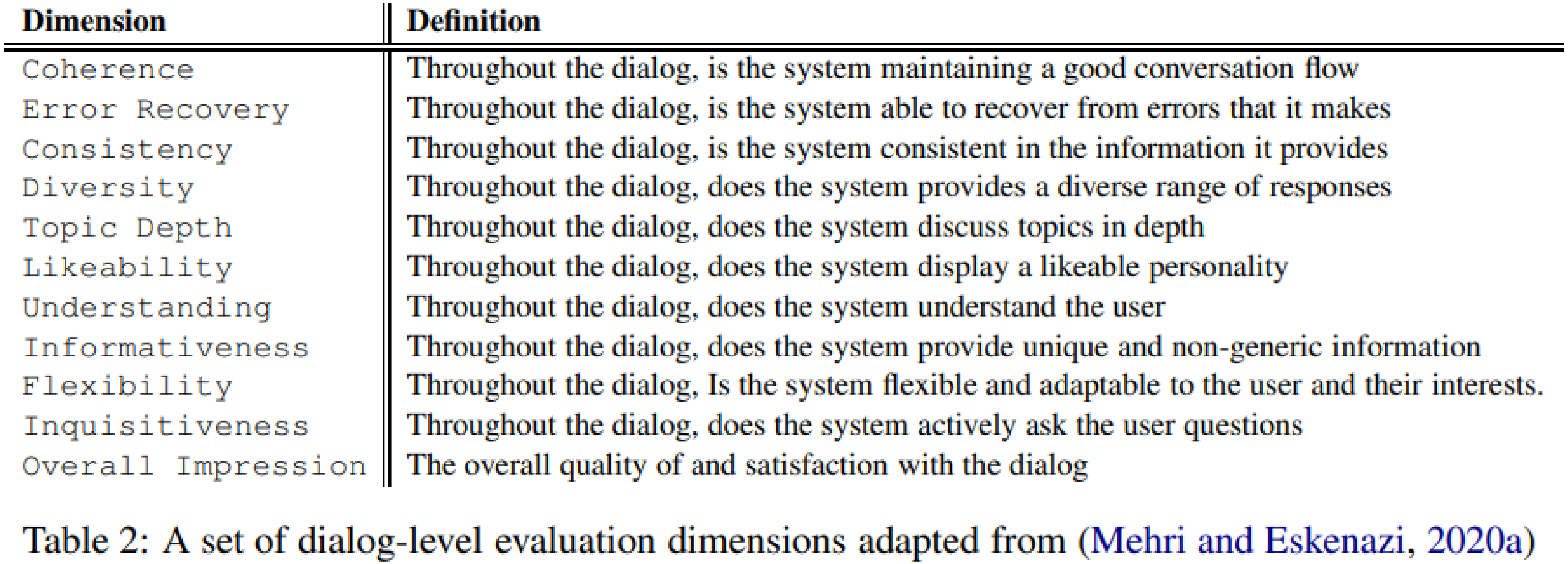


Dimension	Definition
Grammaticality	Responses are free of grammatical and semantic errors
Relevance	Responses are on-topic with the immediate dialog history
Informativeness	Responses produce unique and non-generic information that is specific to the dialog context
Emotional Understanding	Responses indicate an understanding of the user's current emotional state and provide an appropriate emotional reaction based on the current dialog context
Engagingness	Responses are engaging to user and fulfill the particular conversational goals implied by the user
Consistency	Responses do not produce information that contradicts other information known about the system
Proactivity	Responses actively and appropriately move the conversation along different topics
Quality	The overall quality of and satisfaction with the responses

Table 1: A set of turn-level evaluation dimensions adapted from (Finch and Choi, 2020)

- Finch and Choi. "Towards Unified Dialogue System Evaluation: A Comprehensive Analysis of Current Evaluation Protocols." SIGDial (2020).

Human Evaluation - Dialogue-Level Evaluation Criteria



Dimension	Definition
Coherence	Throughout the dialog, is the system maintaining a good conversation flow
Error Recovery	Throughout the dialog, is the system able to recover from errors that it makes
Consistency	Throughout the dialog, is the system consistent in the information it provides
Diversity	Throughout the dialog, does the system provides a diverse range of responses
Topic Depth	Throughout the dialog, does the system discuss topics in depth
Likeability	Throughout the dialog, does the system display a likeable personality
Understanding	Throughout the dialog, does the system understand the user
Informativeness	Throughout the dialog, does the system provide unique and non-generic information
Flexibility	Throughout the dialog, Is the system flexible and adaptable to the user and their interests.
Inquisitiveness	Throughout the dialog, does the system actively ask the user questions
Overall Impression	The overall quality of and satisfaction with the dialog

Table 2: A set of dialog-level evaluation dimensions adapted from (Mehri and Eskenazi, 2020a)

- Mehri and Eskenazi. "Unsupervised Evaluation of Interactive Dialog with DialoGPT." SIGDial (2020).

Human Evaluation

- However, human annotations
  - Lack of consistency (e.g., experts vs non-experts vs crowd-workers)
  - Depend on age, mood, culture, topic knowledge, previous experience, expectations,...
  - Are costly and time-consuming, specially if number of evaluated dimensions is increased
  - Are highly conditioned on evaluation setup (i.e., scale, description of the task, selection of annotators, quality check, …)
- Hence, we need automatic evaluation metrics !!!

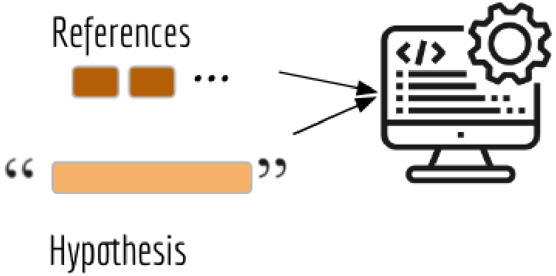
Automatic Evaluation

- The goal of automatic evaluation is not to replace human evaluation, but to supplement it with evaluation, that is **consistent**, **reproducible**, **efficient** and **cheap**.
- Favourable attributes of automatic metrics
  - Strong correlation with human judgment
  - Interpretability and multidimensionality
  - Generalizable across different domains
  - Robustness against adversarial attacks
  - Compatible with Human Evaluation

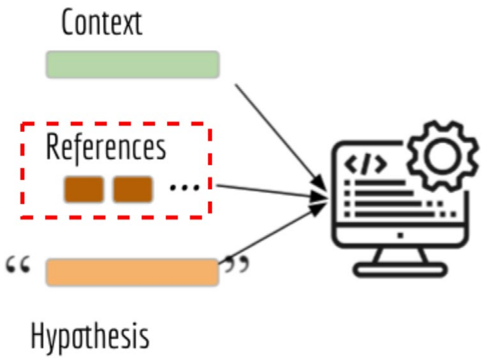
Categorization of Metrics

Different taxonomies for automatic metrics, in this course we follow:

- Context-based/Reference-free or Context-free/Reference-based metrics
- Untrained-based or Trained-based metrics



Reference-based but Context-free Metrics



Context-based Metrics with or without references

- Khapra and Sai. "A tutorial on evaluation metrics used in natural language generation." NAACL (2021).

Q&A

Reference-based Metrics

Untrained metrics

- **Character-based metrics:**
  - Work at character or phoneme level
  - Only consider lexical consistency, i.e., no fluency, syntactic and semantic integrity is considered.
  - Examples:
    - Edit distance, Jaccard, Hamming, characTER (Wang et al, 2016), Extended Edit Distance (Stanchev et al., 2019), etc.
  - Reduced usage in Dialogue Systems.

- Wang, W, et al. "Character: Translation edit rate on character level." *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*. 2016.
- Stanchev, Peter, Weiyue Wang, and Hermann Ney. "EED: Extended edit distance measure for machine translation." Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1). 2019.

Untrained metrics

- **N-gram Based Metrics**
  - Work at word-level or sequences of words (n-grams)
  - Examples:
    - BLEU (Bilingual Evaluation Understudy) (Papineni et al., 2002)
    - ROUGE (Recall Oriented Understudy for Gisting Evaluation)(Lin, C-Y, 2004)
    - METEOR (Metric for Evaluation for Translation with Explicit Ordering)(Banerjee and Lavie, 2005)
    - CIDEr (Vedantam et al., 2015)
  - High usage even in dialogue system due to simplicity and tradition

- Papineni, Kishore, et al. "BLEU: a method for automatic evaluation of machine translation." ACL (2002).
- Lin, Chin-Yew. "ROUGE: A package for automatic evaluation of summaries." Text summarization branches out (2004).
- Banerjee, Satanjeev, and Alon Lavie. "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments." Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization. 2005.
- Vedantam, R., Lawrence Zitnick, C., & Parikh, D. (2015). Cider: Consensus-based image description evaluation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4566-4575).

BLEU (Bilingual Evaluation Understudy)

Most popular metric proposed for machine translation. Intended for:

- adequacy
- fidelity
- fluency

Precision is approximated by modified n-gram precision:

- Fraction of n-grams in the candidate text which are present in any of the reference texts
- Match candidate's n-grams only as many times as they are present in any of the reference texts (to avoid repetitive words or arbitrary long texts)
- Include different n-gram orders by using geometric mean (precision decreases exponentially with *n)*

$$Precision = \exp(\sum_{n=1}^{N} w_n \log p_n), \quad \text{where} \quad w_n = 1/n$$

BLEU (Bilingual Evaluation Understudy)

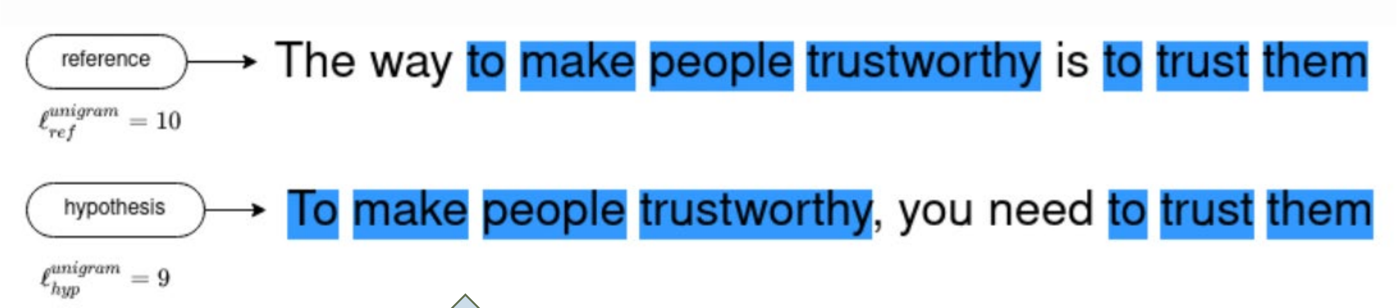
Recall is **approximated** by best match length:

- I.e., sentences with the same length or longer have more options to be correct (more n-grams matching will occur)
- c is the total length of candidate translation corpus, and r is the effective reference length of corpus, i.e., average length of all references.

$$BP = \begin{cases} 1, & \text{if } c > r \\ \exp(1 - \frac{r}{c}), & \text{otherwise} \end{cases}$$

Example with BLEU

- Human reference: The way to make people trustworthy is to trust them.
- Machine hypothesis: To make people trustworthy, you need to trust them.

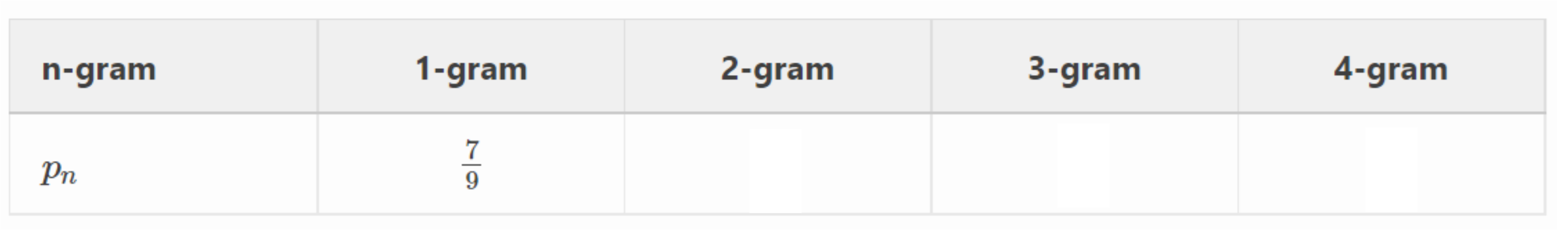


The way to make people trustworthy is to trust them

$\ell_{ref}^{unigram} = 10$

To make people trustworthy, you need to trust them

$\ell_{hyp}^{unigram} = 9$

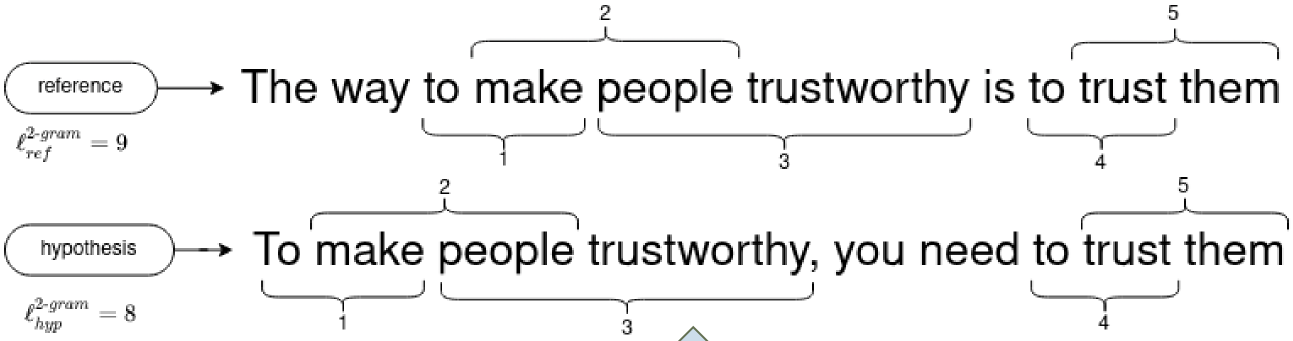


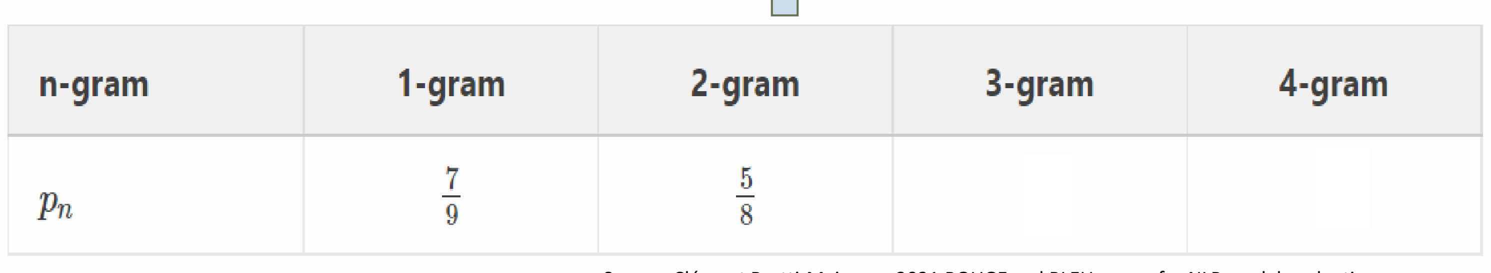
n-gram	1-gram	2-gram	3-gram	4-gram
p_n	$\frac{7}{9}$			

- Source: Clément Brutti-Mairesse, 2021.ROUGE and BLEU scores for NLP model evaluation

Example with BLEU

- Human reference: The way to make people trustworthy is to trust them.
- Machine hypothesis: To make people trustworthy, you need to trust them.


reference
The way to make people trustworthy is to trust them
$\ell_{ref}^{2\text{-gram}} = 9$
1
2
3
4
5
hypothesis
To make people trustworthy, you need to trust them
$\ell_{hyp}^{2\text{-gram}} = 8$
1
2
3
4
5

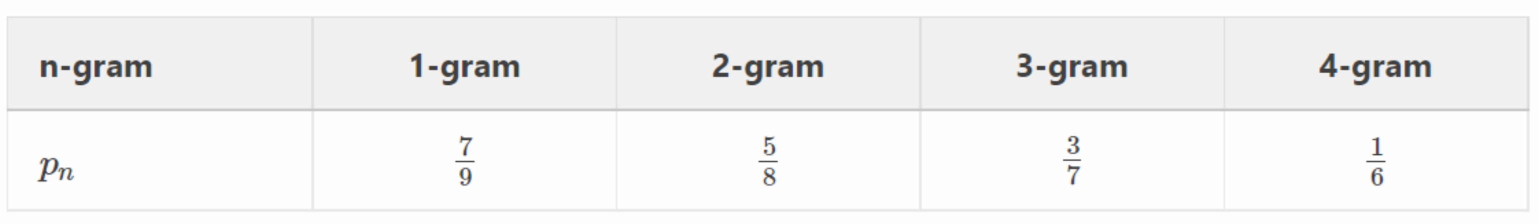


n-gram	1-gram	2-gram	3-gram	4-gram
p_n	$\frac{7}{9}$	$\frac{5}{8}$		

- Source: Clément Brutti-Mairesse, 2021.ROUGE and BLEU scores for NLP model evaluation

Example with BLEU

- Human reference: The way to make people trustworthy is to trust them. (L=10)
- Machine hypothesis: To make people trustworthy, you need to trust them. (L=9)



n-gram	1-gram	2-gram	3-gram	4-gram
p_n	$\frac{7}{9}$	$\frac{5}{8}$	$\frac{3}{7}$	$\frac{1}{6}$

$$BLEU_{N=4} = BP \cdot \exp\left(\sum_{n=1}^{N=4} \frac{1}{4} \log p_n\right)$$

$$BP = \begin{cases} 1 & \text{if } \ell_{hyp} > \ell_{ref} \\ e^{1-\frac{\ell_{ref}}{\ell_{hyp}}} & \text{if } \ell_{hyp} \leq \ell_{ref} \end{cases}$$

$$BP = e^{1-\frac{\ell_{ref}}{\ell_{hyp}}} = e^{-\frac{1}{9}}$$

$$BLEU_{N=4} \approx 0.33933$$

- Source: Clément Brutti-Mairesse, 2021.ROUGE and BLEU scores for NLP model evaluation

Problems with BLEU

Main issue:

- Require multiple references for better handling syntactic and semantic differences

Precision:

- Oly n-grams up-to order 4 are considered

Recall:

- Difficult to calculate the sensitivity of the candidate with respect to a general reference, therefore recall is not really calculated
- Average length is calculated over the entire corpus to avoid harshly punishing the length deviations on short sentences

ROUGE (Recall Oriented Understudy for Gisting Evaluation)

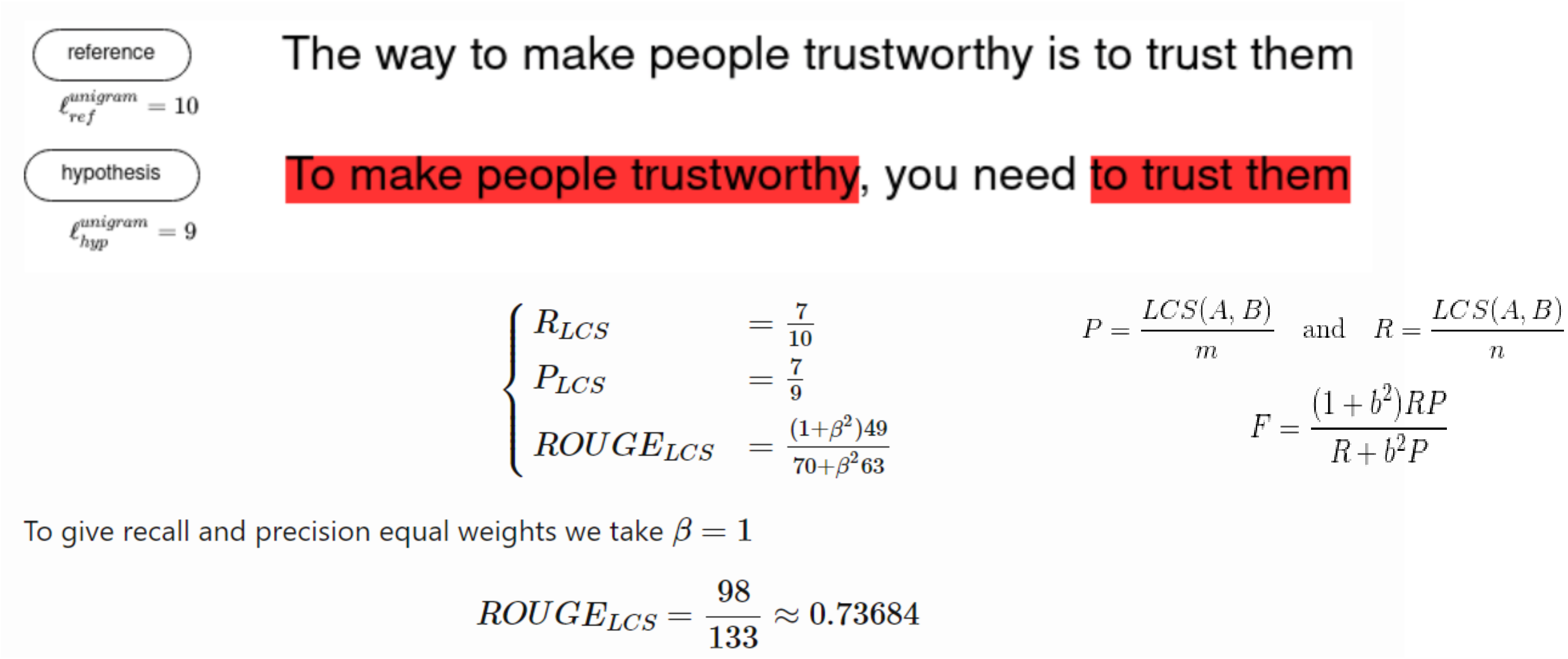
- It is based mainly on recall, and it is mostly used for **summary evaluation**. Intended for evaluating:
    - coherence
    - conciseness
    - grammaticality
    - readability
    - content
- Up to four different types to measure matching of n-grams with priority for longest matching
    - Most relevant ones are ROUGE-N and ROUGE-L

ROUGE (Recall Oriented Understudy for Gisting Evaluation)

- **ROUGE-N:** For any *n*, count the total number of n-grams across all the references, and find out how many of them are present in the candidate. This fraction is the required metric value.
- **ROUGE-L/W/S**: based on longest common subsequence (LCS), weighted LCS, and skip-bigram co-occurrence statistics, respectively. Use an F-score which is the harmonic mean of precision and recall values; *m* and *n* are lengths of candidate and reference.

$$P = \frac{LCS(A,B)}{m} \quad \text{and} \quad R = \frac{LCS(A,B)}{n} \qquad F = \frac{(1+b^2)RP}{R+b^2P}$$

Example ROUGE



reference

$\ell_{ref}^{unigram} = 10$

The way to make people trustworthy is to trust them

hypothesis

$\ell_{hyp}^{unigram} = 9$

To make people trustworthy, you need to trust them

$$\begin{cases} R_{LCS} & = \frac{7}{10} \\ P_{LCS} & = \frac{7}{9} \\ ROUGE_{LCS} & = \frac{(1+\beta^2)49}{70+\beta^2 63} \end{cases}$$

$$P = \frac{LCS(A,B)}{m} \quad \text{and} \quad R = \frac{LCS(A,B)}{n}$$

$$F = \frac{(1+b^2)RP}{R+b^2P}$$

To give recall and precision equal weights we take $\beta = 1$

$$ROUGE_{LCS} = \frac{98}{133} \approx 0.73684$$

- Source: Clément Brutti-Mairesse, 2021.ROUGE and BLEU scores for NLP model evaluation

Untrained metrics

- **Embedding-based metrics:**
  - Work using static (e.g., Wor2Vec or Glove) or contextual (BERT, ELMO) vector embeddings
  - Embeddings are trained on large corpora and capture distributional similarity between words
  - Examples:
    - Greedy Matching (Rus & Lintean, 2012)
    - Embedding Average metric (Landauer & Dumais, 1997)
    - Vector Extrema (Forgues et al., 2014)
    - BERTscore (Zhang et al., 2019)

- Rus, V., & Lintean, M. (2012, June). An optimal assessment of natural language student input using word-to-word similarity metrics. In International Conference on Intelligent Tutoring Systems (pp. 675-676). Springer, Berlin, Heidelberg.
- Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychological review, 104(2), 211.
- Forgues, G., Pineau, J., Larchevêque, J. M., & Tremblay, R. (2014, December). Bootstrapping dialog systems with word embeddings. In Nips, modern machine learning and natural language processing workshop (Vol. 2, p. 168).
- Zhang, Tianyi, et al. "Bertscore: Evaluating text generation with bert." arXiv preprint arXiv:1904.09675 (2019).

Embedding metrics

- Use different types of vector embeddings (static or contextual, e.g., Word2Vec, Glove, SentenceTransformers)

- **Greedy Matching:** evaluates each token in the reference to the closest token in the hypothesis using cosine similarity between the embeddings of the tokens. Then, averaging across all the tokens in the reference.
  - The greedy approach is direction-dependent, then the process is repeated in the reverse direction and averaged.

$$G(p,r) = \frac{\sum_{w \in r} \max_{\hat{w} \in p} cosine(\overrightarrow{w}, \overrightarrow{\hat{w}})}{|r|}$$

$$GM = \frac{G(p,r) + G(r,p)}{2}$$

- **Average embedding:** computes a sentence-level embedding by averaging the word embeddings of all the tokens in each sentence. The score is the cosine similarity between the embedding of the reference nd the embedding of the hypothesis.

$$\overrightarrow{s} = \frac{\sum_{w \in s} \overrightarrow{w}}{|s|} \Rightarrow EA = cosine(\overrightarrow{p}, \overrightarrow{r})$$

Embedding metrics

- **BERTscore** (Zhang et al., 2019)**:** Compute cosine similarity of each hypothesis token j with each token i in the reference sentence using contextualized embeddings.
  - Follow the greedy matching approach instead of a time-consuming best-case matching approach, and then compute the F1 measure:

$$R_{BERT} = \frac{1}{|r|} \sum_{i \in r} \max_{j \in p} \vec{i}^{\,T} \vec{j} \;, P_{BERT} = \frac{1}{|p|} \sum_{j \in p} \max_{i \in r} \vec{i}^{\,T} \vec{j}$$

$$\text{BERTscore} = F_{BERT} = 2 \frac{P_{BERT} . R_{BERT}}{P_{BERT} + R_{BERT}}$$

- Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2019). Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

AMFM: Adequacy and Fluency Metric

- Initially proposed by (Banchs et al, 2014) for MT, later modified for ODD (D'Haro et al., 2019) and finally adapted to DNN in (Zhang et al., 2020) using Transformers (BERT)
- Evaluates two dimensions: adequacy (coherence w.r.t. Context or Reference) and fluency (syntactic/semantic w.r.t. response)
  - Metric can be adapted to multiple references, changes in length (relative scores), and multilingual data (changing encoder)

$$AM_i = max_k \frac{S_i^T \cdot S_k^T}{\|S_i^T\| \cdot \|S_k^T\|}, k \equiv References, i \equiv turn\ pair \qquad FM_i = \frac{\min\left(PPL(H_i), PPL(R_k)\right)}{\max\left(PPL(H_i), PPL(R_k)\right)}, H_i \equiv Context, R_k \equiv Reference\ k$$

- Banchs, R. E., D'Haro, L. F., & Li, H. (2015). Adequacy–fluency metrics: Evaluating mt in the continuous space model framework. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 23(3), 472-482.
- D'Haro, L. F., Banchs, R. E., Hori, C., & Li, H. (2019). Automatic evaluation of end-to-end dialog systems with adequacy-fluency metrics. Computer Speech & Language, 55, 200-215.
- Zhang, C., D'Haro, L. F., Banchs, R. E., Friedrichs, T., & Li, H. (2021). Deep AM-FM: Toolkit for automatic dialogue evaluation. In Conversational Dialogue Systems for the Next Decade (pp. 53-69). Springer, Singapore.

Trained metrics

Context-free and context-based metrics that contain learnable components trained for automatic evaluation.

Feature-based trained metrics: combine various heuristic-based features using a learnable model. These features are obtained from the hypothesis and reference sentences; features such as n-gram precision, recall, BLEU or METEOR scores.

- **BLEND** (Ma et al., 2017): A SVM regression model combining various existing lexical, syntactic and semantic based metrics
- **Q-Metrics** (Nema & Khapra, 2018): categorize words into four categories: function words, question words, named entities and content words and then average the precision and recall for each one. Finally, interpolated with other metrics such as BLEU.

- Ma et al. "Blend: a novel combined MT metric based on direct assessment—CASICT-DCU submission to WMT17 metrics task." Proceedings of the second conference on machine translation. 2017.
- Nema, P., & Khapra, M. M. (2018). Towards a better metric for evaluating question generation systems. arXiv preprint arXiv:1808.10192.

Trained metrics

End-to-End Trained metrics: directly trained using the hypothesis and reference sentences. Most of them employ feed-forward neural networks, RNN or Transformer based models with static/contextualized word embeddings.

- **BLEURT** (Sellam et al., 2020): Pretrained BERT pre-training scheme that uses millions of synthetic example for generalization and multiple subtasks

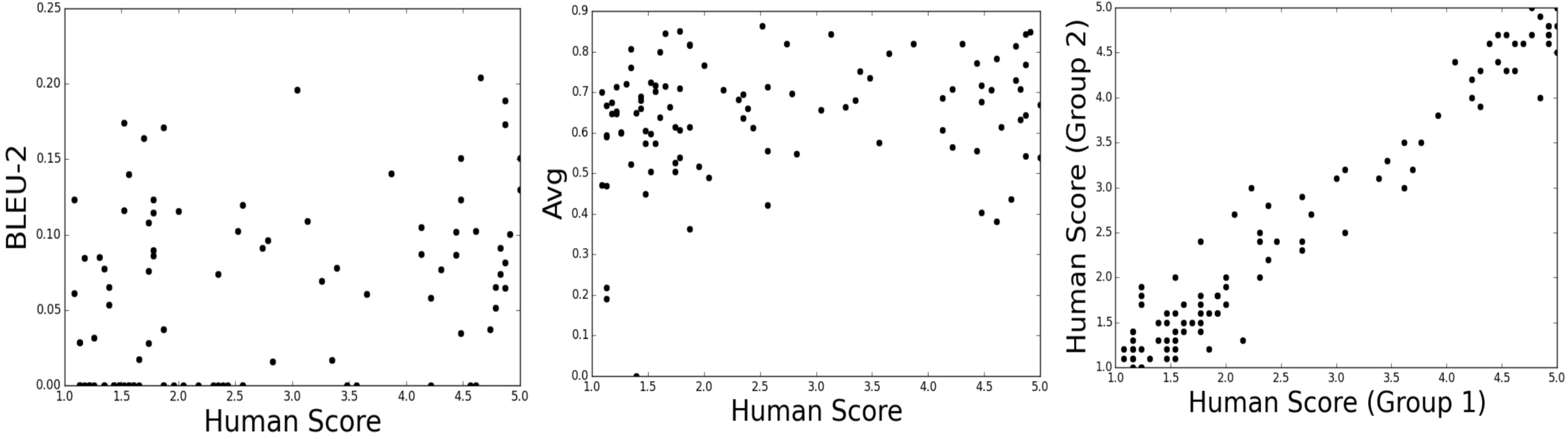
- Sellam, T., Das, D., & Parikh, A. P. (2020). BLEURT: Learning robust metrics for text generation. arXiv preprint arXiv:2004.04696.

Q
&
A
?

Reference-free Metrics

Problems with Reference-based Metrics

- Poor correlation with human evaluation (Liu et al., 2016)



- Liu et al. "How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation." EMNLP (2016).

Problems with Reference-based Metrics

- Reliance on multiple human-written references, which can be costly to obtain
- Poor explainability (Mehri et al., 2022)
  - A single score is hard to interpret
  - No one-size-fit-all solution to open-domain dialogue evaluation
- **Hence, it is necessary to have reference-free/context-dependent metrics**

- Mehri et al. "Report from the NSF future directions workshop on automatic evaluation of dialog: Research directions and challenges." arXiv preprint arXiv:2203.10012 (2022).

Untrained Reference-free Metrics

Perplexity

- Measures how well a probability distribution or probability model predicts a sample
- A common measure used to evaluation language model
- The lower the perplexity, the higher conditional probability of the word sequence

$$\mathrm{PP}(W) = \sqrt[N]{\prod_{i=1}^{N} \frac{1}{P(w_i|w_1 \ldots w_{i-1})}}$$

Semantic Coherence

- Measures similarity between dialogue context and the corresponding response
- A crude way of evaluating context-response entailment

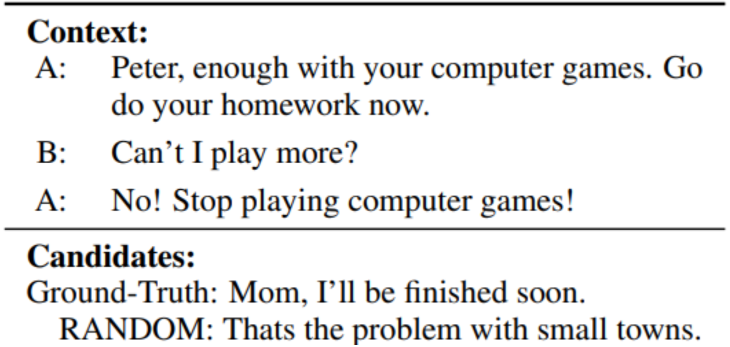
$$sim(\mathbf{p}, \mathbf{h}) = \frac{\mathbf{p}^T \mathbf{h}}{||\mathbf{p}|| \cdot ||\mathbf{h}||}$$

- Performance largely rely on the encoder
- Alternative solution is to use a pre-trained NLI model to predict the context-response entailment score

End-to-End Trained Reference-free Metrics

Key Attributes of End-to-End Metrics

- Self-supervision
  - Learn from unlabeled human-human dialogue data
  - Automatically generate supervision signals to train a classifier or a regressor
- Negative Sampling
  - Syntactic Perturbation - word shuffle, word drop, etc.
  - Semantic Perturbation - random response, generative model output, etc.
- Pre-trained Language Model
  - Direct application - response fluency, local coherence, etc.
  - End-to-end training - relevance, global coherence, etc.



Context:

A: Peter, enough with your computer games. Go do your homework now.

B: Can't I play more?

A: No! Stop playing computer games!

Candidates:

Ground-Truth: Mom, I'll be finished soon.

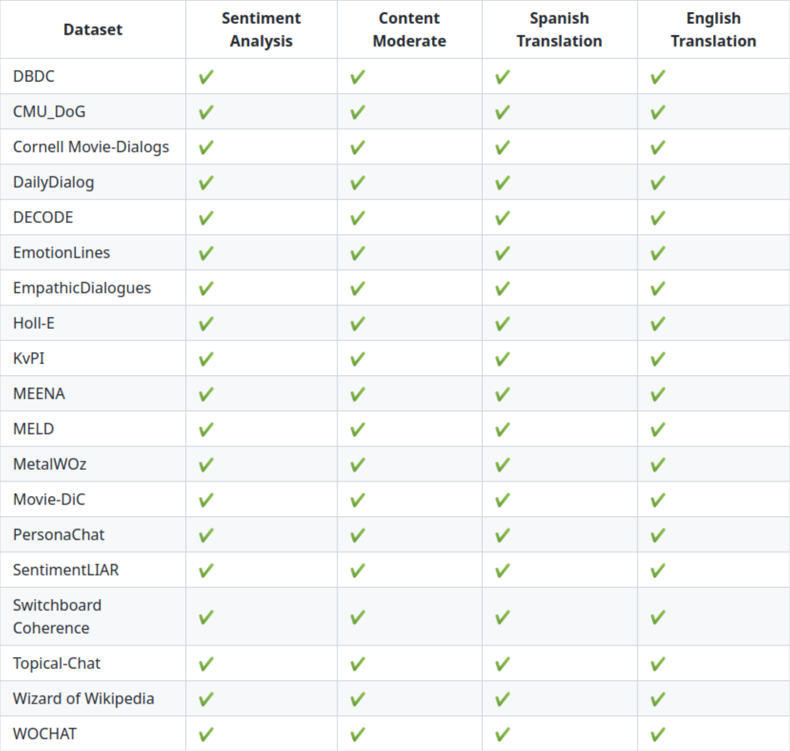
RANDOM: Thats the problem with small towns.

Commonly-Used Dialogue Datasets For Training

- There are different types of dialogue corpora
    - Daily chit-chat; Knowledge-grounded conversations;
    - Persona-guided conversations; Emotion dialogues;
- Dataset Statistics
    - DailyDialog (Li et al., 2017) - 13K dialogues, 110K utterances
    - PersonaChat (Zhang et al., 2018) - 110K dialogues, 162K utterances, 1155 persona profiles
    - TopicalChat (Gopalakrishnan et al., 2019) - 10.7K dialogues, 235K utterances, 300 entities across 8 topics, wikipedia lead section of each entity, 8-10 crowdsourced fun facts per entity, 3088 washington post articles
    - EmpatheticDialogues (Rashkin et al., 2019) - 25K dialogues, 110K utterances, each dialogue is accompanied by a situational context, 32 emotion labels

- Li et al. "Dailydialog: A manually labelled multi-turn dialogue dataset." IJCNLP (2017).
- Rashkin et al. “Towards Empathetic Open-domain Conversation Models: A New Benchmark and Dataset.” ACL (2019).
- Zhang et al. "Personalizing Dialogue Agents: I have a dog, do you have pets too?." ACL (2018).
- Gopalakrishnan et al. "Topical-Chat: Towards Knowledge-Grounded Open-Domain Conversations." INTERSPEECH (2019).

Open-domain Dialogue Datasets



Dataset	Sentiment Analysis	Content Moderate	Spanish Translation	English Translation
DBDC	✓	✓	✓	✓
CMU_DoG	✓	✓	✓	✓
Cornell Movie-Dialogs	✓	✓	✓	✓
DailyDialog	✓	✓	✓	✓
DECODE	✓	✓	✓	✓
EmotionLines	✓	✓	✓	✓
EmpathicDialogues	✓	✓	✓	✓
Holl-E	✓	✓	✓	✓
KvPI	✓	✓	✓	✓
MEENA	✓	✓	✓	✓
MELD	✓	✓	✓	✓
MetalWOz	✓	✓	✓	✓
Movie-DiC	✓	✓	✓	✓
PersonaChat	✓	✓	✓	✓
SentimentLIAR	✓	✓	✓	✓
Switchboard Coherence	✓	✓	✓	✓
Topical-Chat	✓	✓	✓	✓
Wizard of Wikipedia	✓	✓	✓	✓
WOCHAT	✓	✓	✓	✓

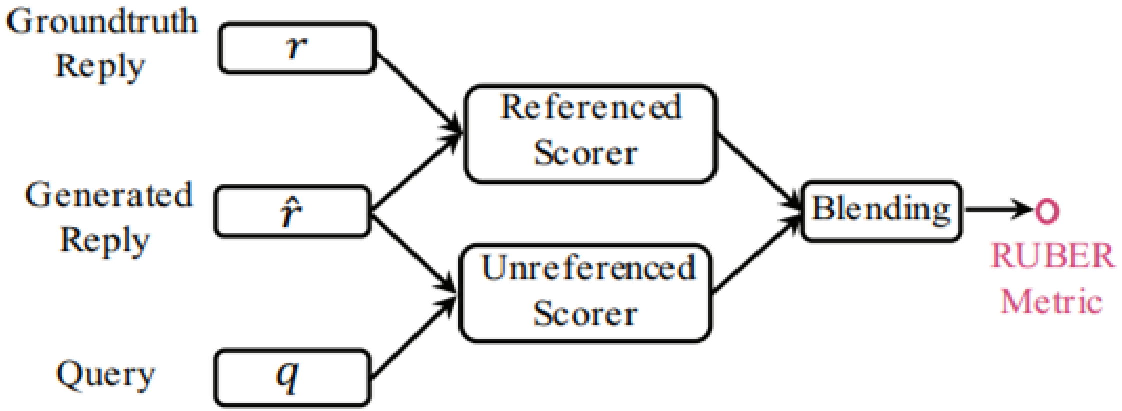
+18 dialogue datasets … and increasing

- Available at https://github.com/CHANEL-JSALT-2020/datasets

Turn-Level Metrics

Representative Examples - RUBER

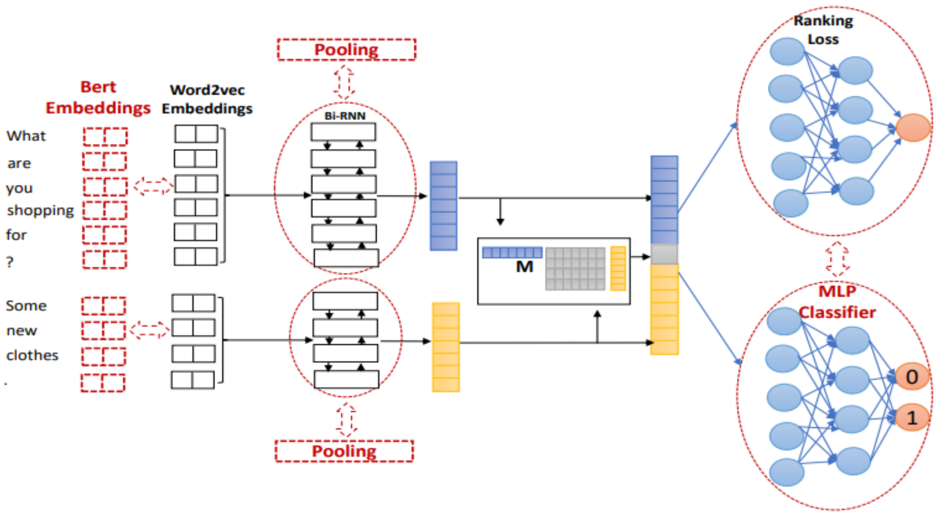
- Referenced and Unreferenced Metric Blended Evaluation Routine (Tao et al., 2018)
    - RUBER is introduced to improve correlation with human evaluation by combining
        - Referenced component - measures similarity between candidate and reference response embeddings
        - Unreferenced component - measures relevance between a candidate response and the corresponding query

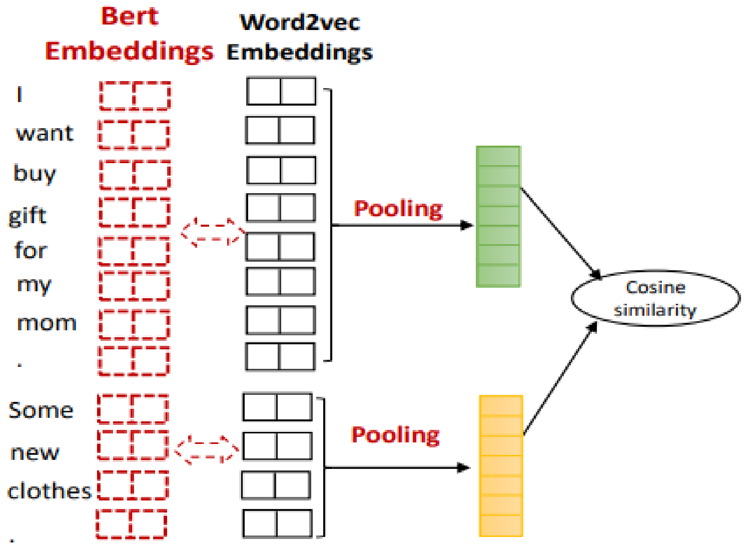


- Tao et al. "RUBER: An unsupervised method for automatic evaluation of open-domain dialog systems." AAAI (2018).

Representative Examples - BERT-RUBER

- Contextualized embeddings provided by pre-trained language model better represent the semantics of utterances than static word embeddings.
- Improving RUBER with contextualized representation (Ghazarian et al., 2019)





- Ghazarian et al. "Better Automatic Evaluation of Open-Domain Dialogue Systems with Contextualized Embeddings." Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation (2019).

Representative Examples - MAUDE

- RUBER and BERT-RUBER focus on static evaluation
- MAUDE (Sinha et al., 2020) targets online evaluation, taking into consideration of the dialogue structure
- RUBER and BERT-RUBER only adopt the random utterance strategy, for MAUDE, A variety of negative sampling strategies are considered

- Sinha et al. "Learning an Unreferenced Metric for Online Dialogue Evaluation." ACL (2020).

Representative Examples - MAUDE

- MAUDE Details
  - Designed to output a scalar

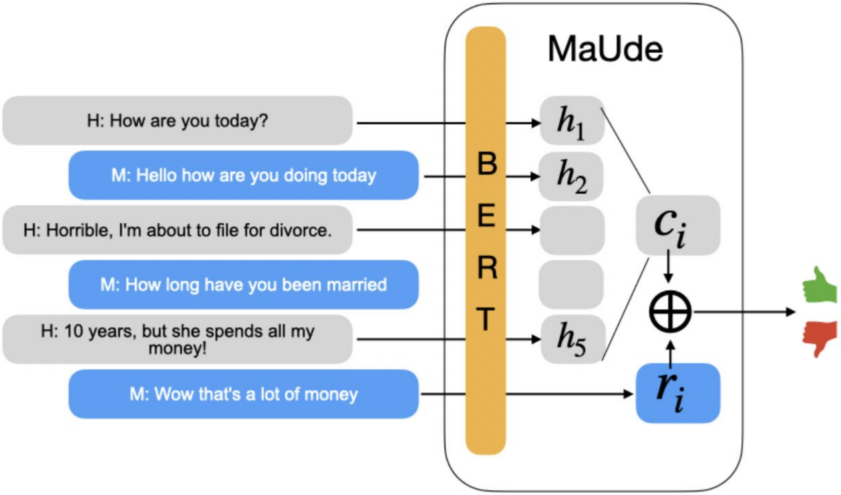
$$\text{score}(c_i, r_i) = \mathcal{R} \in (0,1)$$

  - Employ a dialogue-structure aware encoder

$$\begin{aligned}
\mathbf{h}_{u_i} &= \mathbf{D}_g f_e^{\text{BERT}}(u_i), \\
\mathbf{h}'_{u_{i+1}} &= f_R(\mathbf{h}_{u_i}, \mathbf{h}'_{u_i}), \\
\mathbf{c}_i &= \mathbf{W}.\text{pool}_{\forall t \in \{u_1, \ldots, u_{n-1}\}}(\mathbf{h}'_t) \\
\text{score}(c_i, r_i) &= \sigma(f_t([\mathbf{h}_{r_i}, \mathbf{c}_i, \mathbf{h}_{r_i} * \mathbf{c}_i, \mathbf{h}_{r_i} - \mathbf{c}_i])),
\end{aligned}$$

  - Trained with the Noise Contrastive Estimation (NCE) loss

$$\mathcal{L} = -\log(\text{score}(c, r)) - \mathbb{E}_{\hat{r} \sim P(\hat{r})} \log(-\text{score}(c, \hat{r})).$$



H: How are you today?

M: Hello how are you doing today

H: Horrible, I'm about to file for divorce.

M: How long have you been married

H: 10 years, but she spends all my money!

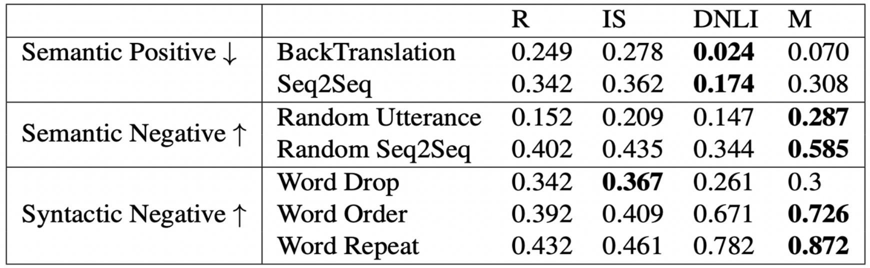
M: Wow that's a lot of money

Representative Examples - MAUDE

- Negative sampling strategies
  - Syntactic perturbations
    - Word-order (shuffling the ordering of the words of r)
    - Word-drop (dropping x% of words in r)
    - Word-repeat (randomly repeating words in r)
  - Semantic perturbations
    - Random utterance from another dialogue (random utterance)
    - Utterance produced by a seq2seq model conditioned on another dialogue context (random seq2seq)
    - Back-translated random utterance (random back-translation)

Representative Examples - MAUDE

- Performance
  - Compare to baseline metrics, MAUDE shows relatively small score differences for semantic positive cases
  - Maximum score differences for both semantic and syntactic negative cases
  - This showcase that MAUDE is able to discriminate negative samples from positive samples



		R	IS	DNLI	M
Semantic Positive ↓	BackTranslation	0.249	0.278	**0.024**	0.070
	Seq2Seq	0.342	0.362	**0.174**	0.308
Semantic Negative ↑	Random Utterance	0.152	0.209	0.147	**0.287**
	Random Seq2Seq	0.402	0.435	0.344	**0.585**
Syntactic Negative ↑	Word Drop	0.342	**0.367**	0.261	0.3
	Word Order	0.392	0.409	0.671	**0.726**
	Word Repeat	0.432	0.461	0.782	**0.872**

Table 1: Metric score evaluation ($\Delta = \text{score}(c, r_{\text{ground-truth}}) - \text{score}(c, r)$) between RUBER (R), InferSent (IS), DistilBERT-NLI (DNI) and MAUDE (M) on PersonaChat dataset's public validation set. For Semantic Positive tests, lower Δ is better; for all Negative tests higher Δ is better.

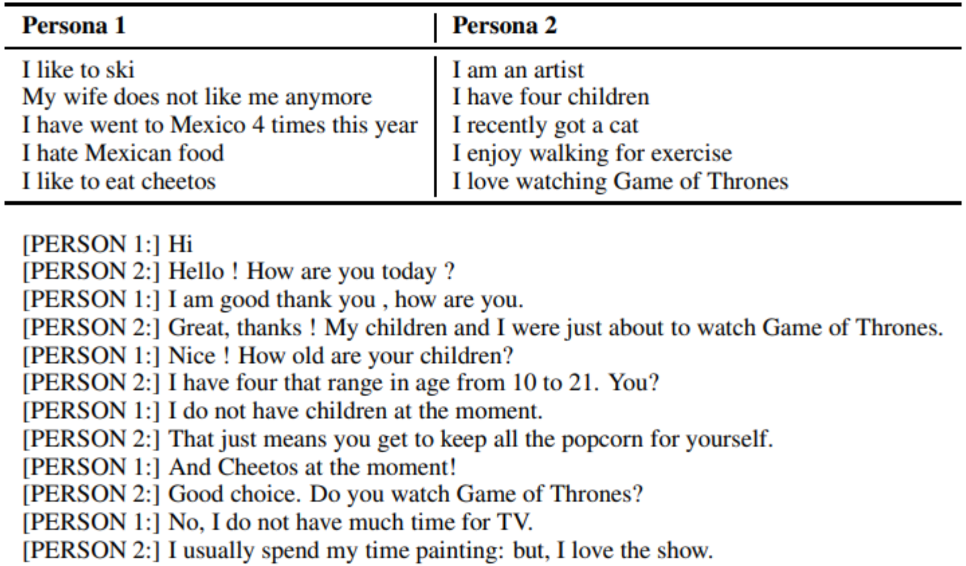
Representative Examples - USR

- MAUDE & RUBER focus mainly on contextual relevance of the response
- Dialogue evaluation is multi-facted in nature
- USR (Mehri & Eskenazi, 2020) Produces interpretable measures for desirable properties of dialogue
  - Understandable - Is the response understandable in the context of the history?
  - Natural - Is the response naturally written?
  - Maintains Context - Does the response serve as a valid continuation of the conversation history?
  - Uses Knowledge - Given the interesting fact that the response is conditioned on, how well does the response use the fact?
  - Interesting - Is the response dull/interesting?
  - Overall Quality - what is your overall impression of this utterance?

- Mehri and Eskenazi. "USR: An Unsupervised and Reference Free Evaluation Metric for Dialog Generation." ACL (2020).

Representative Examples - USR

- The USR Dataset
  - 120 dialogue contexts are sampled from PersonaChat (Zhang et al., 2018) & TopicalChat (Gopalakrishnan et al., 2019), 60 each
  - For each context, 6 responses of varying quality are created
  - Six dialogue researchers annotate each context-response pair along
    - Understandable (0-1); Natural (1-3); Maintains Context (1-3); Interesting (1-3); Uses Knowledge (0-1); Overall Quality (1-5)
  - Moderate to high inter-annotator agreement



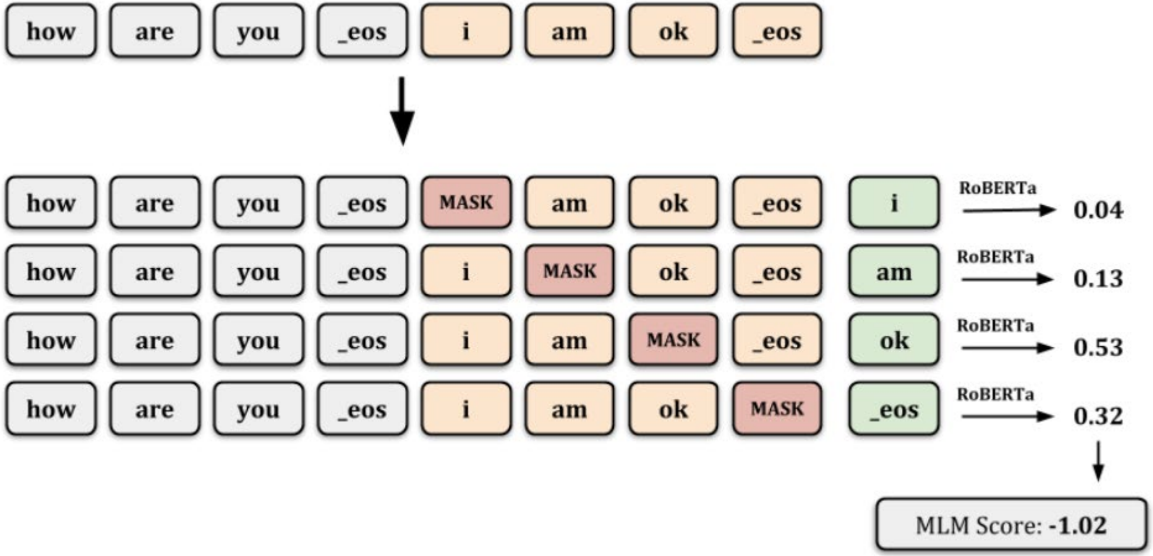
Persona 1	**Persona 2**
I like to ski My wife does not like me anymore I have went to Mexico 4 times this year I hate Mexican food I like to eat cheetos	I am an artist I have four children I recently got a cat I enjoy walking for exercise I love watching Game of Thrones

[PERSON 1:] Hi
[PERSON 2:] Hello ! How are you today ?
[PERSON 1:] I am good thank you , how are you.
[PERSON 2:] Great, thanks ! My children and I were just about to watch Game of Thrones.
[PERSON 1:] Nice ! How old are your children?
[PERSON 2:] I have four that range in age from 10 to 21. You?
[PERSON 1:] I do not have children at the moment.
[PERSON 2:] That just means you get to keep all the popcorn for yourself.
[PERSON 1:] And Cheetos at the moment!
[PERSON 2:] Good choice. Do you watch Game of Thrones?
[PERSON 1:] No, I do not have much time for TV.
[PERSON 2:] I usually spend my time painting: but, I love the show.

- Zhang et al. "Personalizing Dialogue Agents: I have a dog, do you have pets too?." ACL (2018).
- Gopalakrishnan et al. "Topical-Chat: Towards Knowledge-Grounded Open-Domain Conversations." INTERSPEECH (2019).

Representative Examples - USR

- USR - Mask Language Modeling (MLM) Metric
  - Finetune RoBERTa-base (Liu et al., 2019) on PersonaChat or TopicalChat
  - Evaluate the understandability and naturalness of responses
  - Compute the log probability of the masked word



- Liu et al. "Roberta: A robustly optimized bert pretraining approach." arXiv preprint arXiv:1907.11692 (2019).

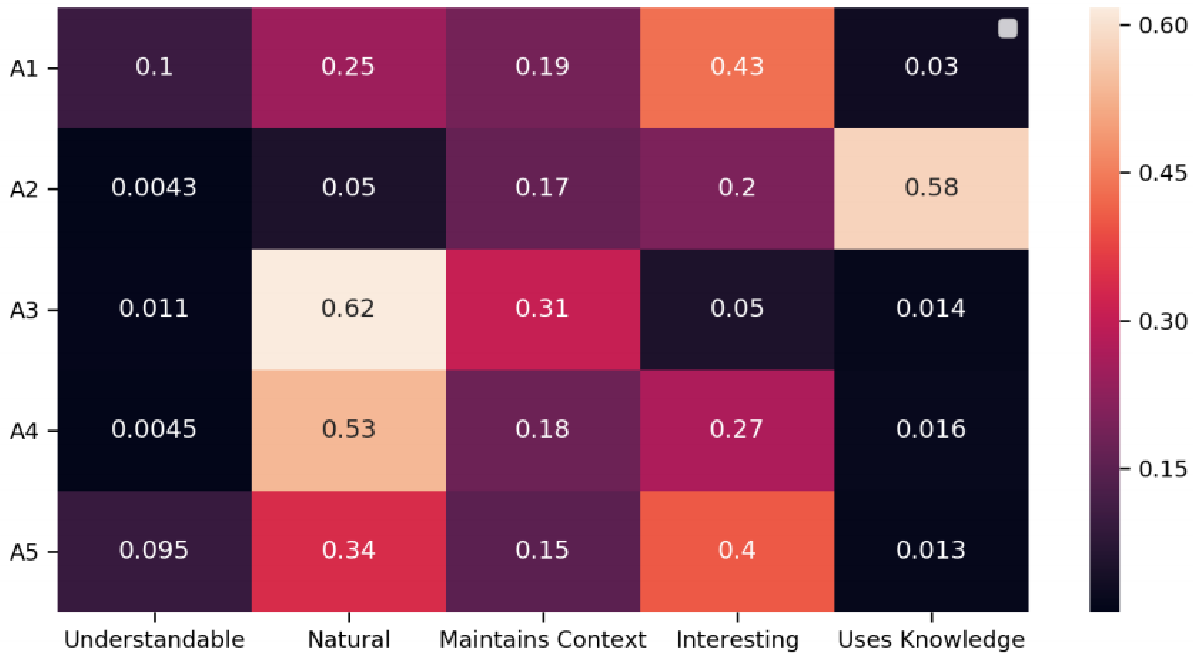
Representative Examples - USR

- USR - Dialogue Retrieval (DR) Metric
  - Evaluate dialogue qualities: “maintaining contexts”, “interesting”, and “using knowledge”
  - Rely on context information to discriminate original responses from random ones, two different contexts are used:
    - Context consists of both the dialogue history and the fact
    - Context is just the fact associated with the dialogue

- Lowe et al. "The Ubuntu Dialogue Corpus: A Large Dataset for Research in Unstructured Multi-Turn Dialogue Systems." SIGDial. (2015).

Representative Examples - USR

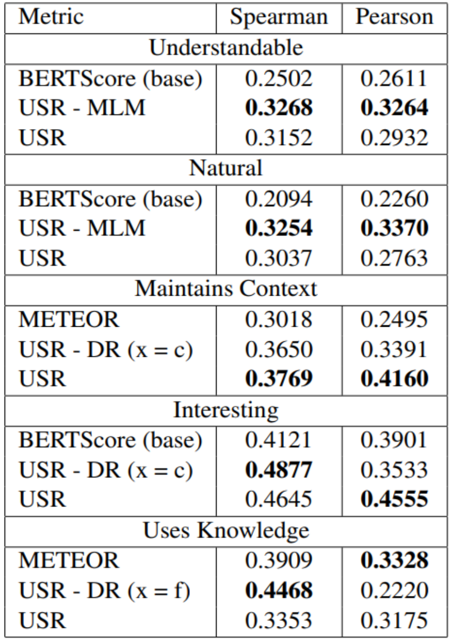
- USR - Combining the sub-metrics
  - Configurable weights assigned to scores with respect to the sub-metrics
  - Evaluate the “Overall Quality”



	Understandable	Natural	Maintains Context	Interesting	Uses Knowledge
A1	0.1	0.25	0.19	0.43	0.03
A2	0.0043	0.05	0.17	0.2	0.58
A3	0.011	0.62	0.31	0.05	0.014
A4	0.0045	0.53	0.18	0.27	0.016
A5	0.095	0.34	0.15	0.4	0.013

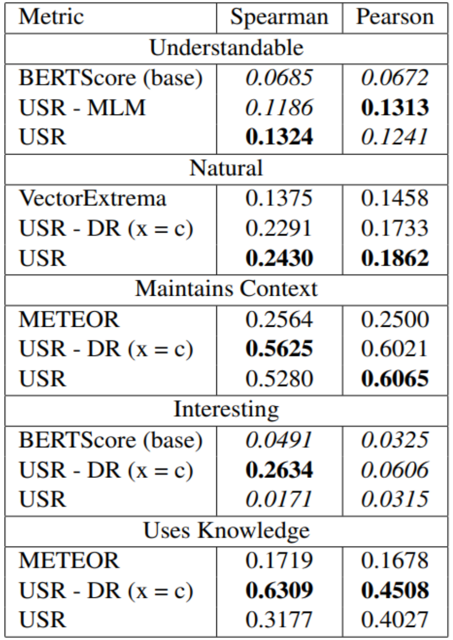
Representative Examples - USR

- Performance of USR
  - USR or its sub-metrics outperforms the best reference-based metrics across all the dimensions
  - USR-MLM performs well on Natural & Understandable dimensions
  - USR-DR performs well on Maintaining context, Interestingness and Usage of Knowledge
  - The performance is still far from perfect



Metric	Spearman	Pearson
Understandable		
BERTScore (base)	0.2502	0.2611
USR - MLM	**0.3268**	**0.3264**
USR	0.3152	0.2932
Natural		
BERTScore (base)	0.2094	0.2260
USR - MLM	**0.3254**	**0.3370**
USR	0.3037	0.2763
Maintains Context		
METEOR	0.3018	0.2495
USR - DR (x = c)	0.3650	0.3391
USR	**0.3769**	**0.4160**
Interesting		
BERTScore (base)	0.4121	0.3901
USR - DR (x = c)	**0.4877**	0.3533
USR	0.4645	**0.4555**
Uses Knowledge		
METEOR	0.3909	**0.3328**
USR - DR (x = f)	**0.4468**	0.2220
USR	0.3353	0.3175

Table 3: Turn-level correlations on Topical-Chat. We show: (1) best non-USR metric, (2) best USR sub-metric and (3) USR metric. All measures in this table are statistically significant to $p < 0.01$.

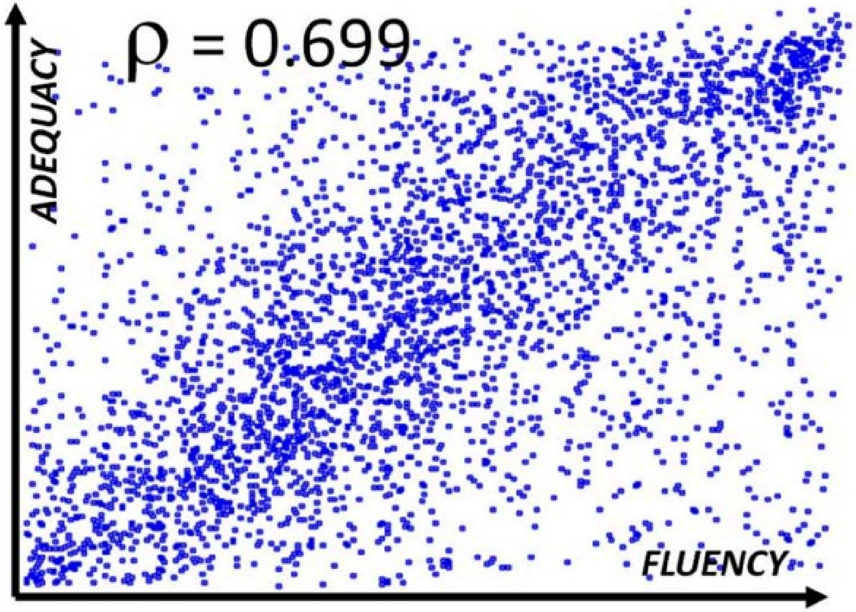


Metric	Spearman	Pearson
Understandable		
BERTScore (base)	*0.0685*	*0.0672*
USR - MLM	*0.1186*	**0.1313**
USR	**0.1324**	*0.1241*
Natural		
VectorExtrema	0.1375	0.1458
USR - DR (x = c)	0.2291	0.1733
USR	**0.2430**	**0.1862**
Maintains Context		
METEOR	0.2564	0.2500
USR - DR (x = c)	**0.5625**	0.6021
USR	0.5280	**0.6065**
Interesting		
BERTScore (base)	*0.0491*	*0.0325*
USR - DR (x = c)	**0.2634**	*0.0606*
USR	*0.0171*	*0.0315*
Uses Knowledge		
METEOR	0.1719	0.1678
USR - DR (x = c)	**0.6309**	**0.4508**
USR	0.3177	0.4027

Table 4: Turn-level correlations on Persona-Chat. We show: (1) best non-USR metric, (2) best USR sub-metric and (3) USR metric. All values with $p > 0.05$ are italicized.

Representative Examples - D-score

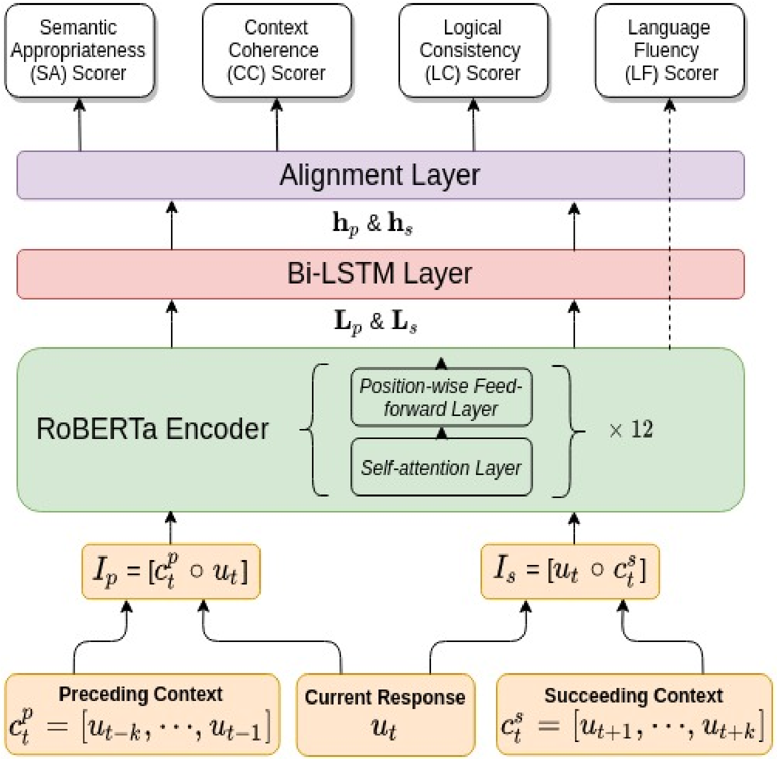
- USR treats the evaluation of individual dimensions as independent
- Human judges do not evaluate different aspects in a completely independent manner
- Dimension-independent dialogue features may potentially benefit the evaluation of different dimensions
- D-score (Zhang et al., 2021) adopts multi-task learning to learn a holistic metric that evaluate
  - Naturalness; Response appropriateness; Coherence; Consistency



- Zhang et al. "D-score: Holistic dialogue evaluation without reference." IEEE/ACM TASLP (2021).

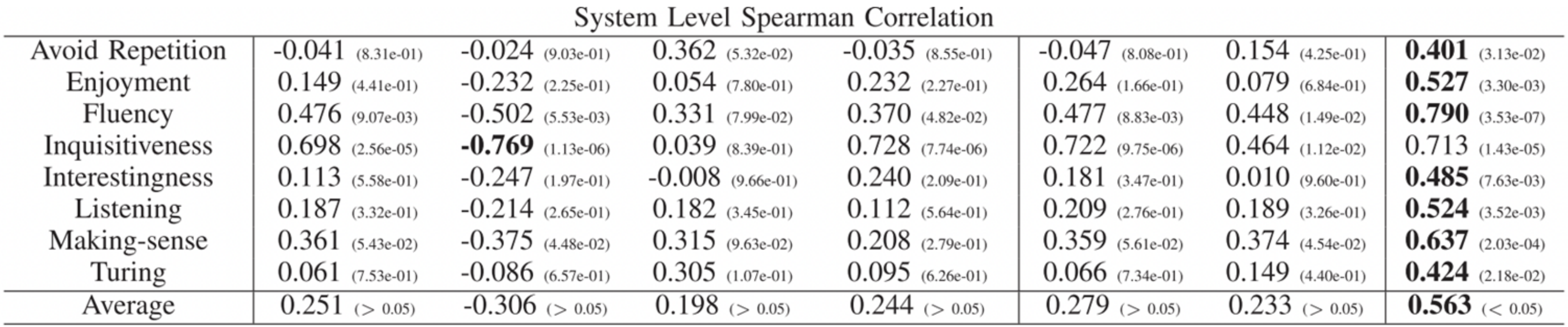
Representative Examples - D-score

- Evaluate different dialogue aspects while keeping a holistic perspective
  - Different pre-text tasks to handle different dialogue aspects
    - Discriminate original response against random utterance
    - Shuffle the ordering of the utterances
    - Natural Language Inference
    - Language modeling
  - Rely on multi-task learning to simultaneously learn the pre-text tasks
  - A shared encoder to encode regularities in dialogues


Semantic Appropriateness (SA) Scorer
Context Coherence (CC) Scorer
Logical Consistency (LC) Scorer
Language Fluency (LF) Scorer
Alignment Layer
$\mathbf{h}_p$ & $\mathbf{h}_s$
Bi-LSTM Layer
$\mathbf{L}_p$ & $\mathbf{L}_s$
RoBERTa Encoder
Position-wise Feed-forward Layer
Self-attention Layer
× 12
$I_p = [c_t^p \circ u_t]$
$I_s = [u_t \circ c_t^s]$
Preceding Context
$c_t^p = [u_{t-k}, \cdots, u_{t-1}]$
Current Response
u_t
Succeeding Context
$c_t^s = [u_{t+1}, \cdots, u_{t+k}]$

Representative Examples - D-score

- Performance
  - D-score achieves highest correlation across almost all the dimensions
  - Outperforms both USR and metric fusion, demonstrating the advantage of multi-task learning

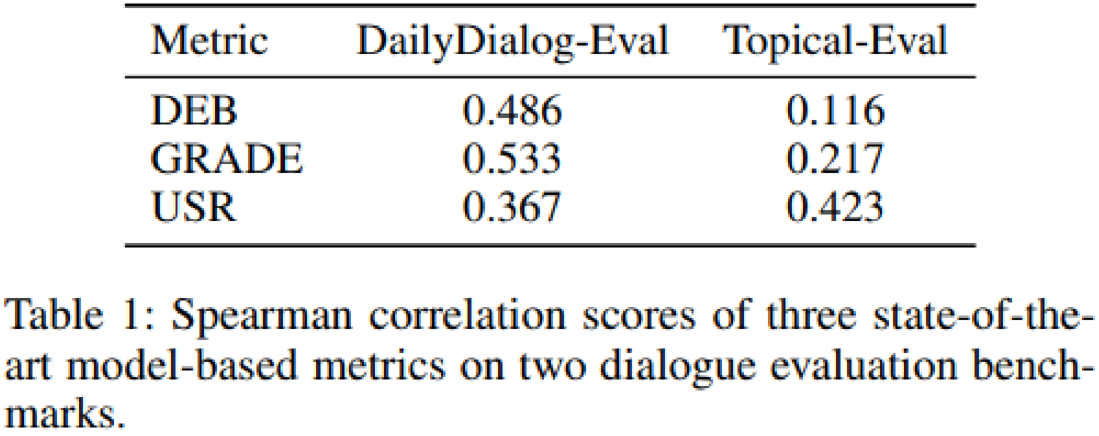


System Level Spearman Correlation

Avoid Repetition	-0.041 (8.31e-01)	-0.024 (9.03e-01)	0.362 (5.32e-02)	-0.035 (8.55e-01)	-0.047 (8.08e-01)	0.154 (4.25e-01)	**0.401** (3.13e-02)
Enjoyment	0.149 (4.41e-01)	-0.232 (2.25e-01)	0.054 (7.80e-01)	0.232 (2.27e-01)	0.264 (1.66e-01)	0.079 (6.84e-01)	**0.527** (3.30e-03)
Fluency	0.476 (9.07e-03)	-0.502 (5.53e-03)	0.331 (7.99e-02)	0.370 (4.82e-02)	0.477 (8.83e-03)	0.448 (1.49e-02)	**0.790** (3.53e-07)
Inquisitiveness	0.698 (2.56e-05)	**-0.769** (1.13e-06)	0.039 (8.39e-01)	0.728 (7.74e-06)	0.722 (9.75e-06)	0.464 (1.12e-02)	0.713 (1.43e-05)
Interestingness	0.113 (5.58e-01)	-0.247 (1.97e-01)	-0.008 (9.66e-01)	0.240 (2.09e-01)	0.181 (3.47e-01)	0.010 (9.60e-01)	**0.485** (7.63e-03)
Listening	0.187 (3.32e-01)	-0.214 (2.65e-01)	0.182 (3.45e-01)	0.112 (5.64e-01)	0.209 (2.76e-01)	0.189 (3.26e-01)	**0.524** (3.52e-03)
Making-sense	0.361 (5.43e-02)	-0.375 (4.48e-02)	0.315 (9.63e-02)	0.208 (2.79e-01)	0.359 (5.61e-02)	0.374 (4.54e-02)	**0.637** (2.03e-04)
Turing	0.061 (7.53e-01)	-0.086 (6.57e-01)	0.305 (1.07e-01)	0.095 (6.26e-01)	0.066 (7.34e-01)	0.149 (4.40e-01)	**0.424** (2.18e-02)
Average	0.251 (> 0.05)	-0.306 (> 0.05)	0.198 (> 0.05)	0.244 (> 0.05)	0.279 (> 0.05)	0.233 (> 0.05)	**0.563** (< 0.05)

Representative Examples - MDD-Eval

- Existing metrics lack a generalized skill to evaluate dialogues across multiple domains
- Lack of high-quality multi-domain training data
  - Existing negative sampling strategies are too easy to learn



Metric	DailyDialog-Eval	Topical-Eval
DEB	0.486	0.116
GRADE	0.533	0.217
USR	0.367	0.423

Table 1: Spearman correlation scores of three state-of-the-art model-based metrics on two dialogue evaluation benchmarks.

- Zhang et al. "MDD-Eval: self-training on augmented data for multi-domain dialogue evaluation." AAAI (2022).

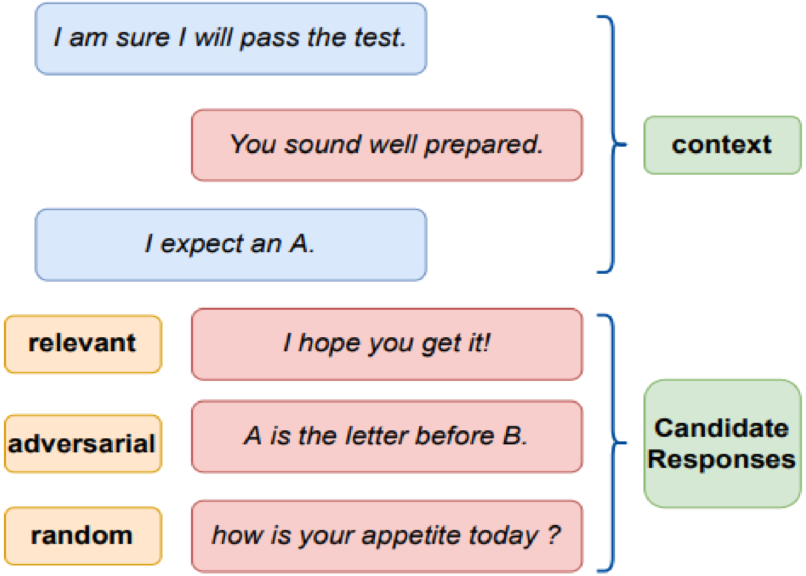
Representative Examples - MDD-Eval

- Possible solutions
  - Recruit humans to write multiple relevant and hard negative responses given a dialogue context (Sai et al. 2020)
    - Too expensive and difficult to scale
  - Rely on data augmentation techniques (Gupta et al. 2021)
    - Lack of quality control, such as introduction of false negatives
- Combine advantages of both worlds
  - Semi-supervised learning - Self Training
    - Apply a teacher model, trained using labeled data, to create synthetic labels for unlabeled examples (Scudder. 1965)
    - Combine the pseudo-labeled data and labeled data, to train a student model

- Sai et al. "Improving Dialog Evaluation with a Multi-reference Adversarial Dataset and Large Scale Pretraining" TACL (2020).
- Gupta et al. 2021. "Synthesizing Adversarial Negative Responses for Robust Response Ranking and Evaluation" Findings of ACL-IJCNLP (2021).
- H Scudder. Probability of error of some adaptive pattern-recognition machines. IEEE Transactions on Information Theory, 11(3):363–371 (1965).

Representative Examples - MDD-Eval

- Step 1. Train a strong teacher classifier
  - Fine-tune RoBERTa (Liu et al., 2019) on DailyDialog++ dataset (Sai et al., 2020)



I am sure I will pass the test.

You sound well prepared.

I expect an A.

context

relevant — *I hope you get it!*

adversarial — *A is the letter before B.*

random — *how is your appetite today ?*

Candidate Responses

- Sai et al. "Improving Dialog Evaluation with a Multi-reference Adversarial Dataset and Large Scale Pretraining" TACL (2020).
- Liu et al. “RoBERTa: A robustly optimized bert pretraining approach”. arXiv preprint arXiv:1907.11692 (2019).

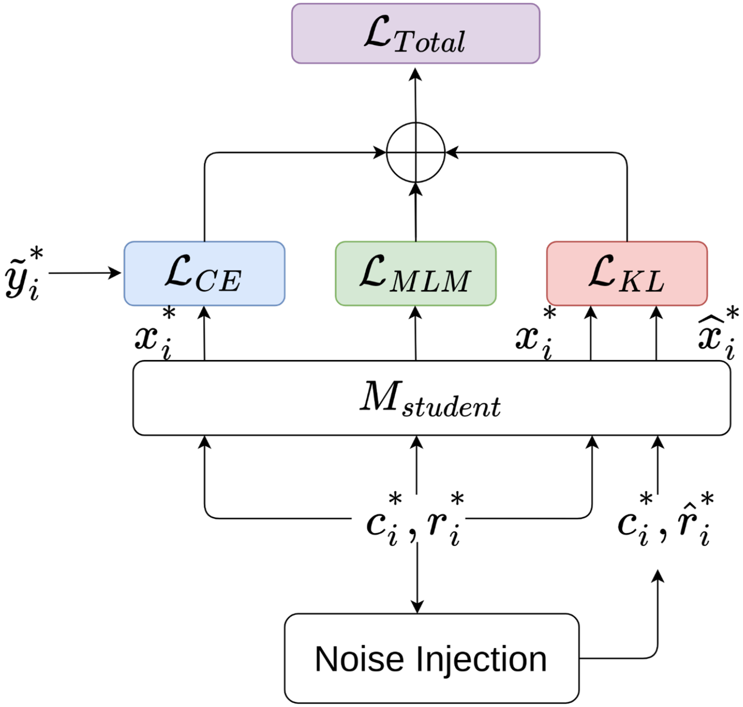
Representative Examples - MDD-Eval

- Step 2. Perform data augmentation to obtain large-scale multi-domain data
  - Original human-human dialogues from 4 different datasets, such as DailyDialog, PersonaChat, EmpatheticDialogue, and TopicalChat
  - Back-translation of dialogue responses
  - Mask-and-fill (Gupta et al. 2021)
  - Random sampling
  - Generate responses with open-domain chatbots, such as DialoGPT (Zhang et al. 2020) and BlenderBot (Roller et al. 2021)
  - Syntactic perturbations, such as word drop, word shuffle, and word repeat

- Li et al. "DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset" IJCNLP (2017).
- Zhang et al. “Personalizing Dialogue Agents: I have a dog, do you have pets too?” ACL (2018).
- Zhang et al. “DIALOGPT : Large-Scale Generative Pre-training for Conversational Response Generation” ACL (2020).
- Roller et al. “Recipes for Building an Open-Domain Chatbot”. EACL (2021).

Representative Examples - MDD-Eval

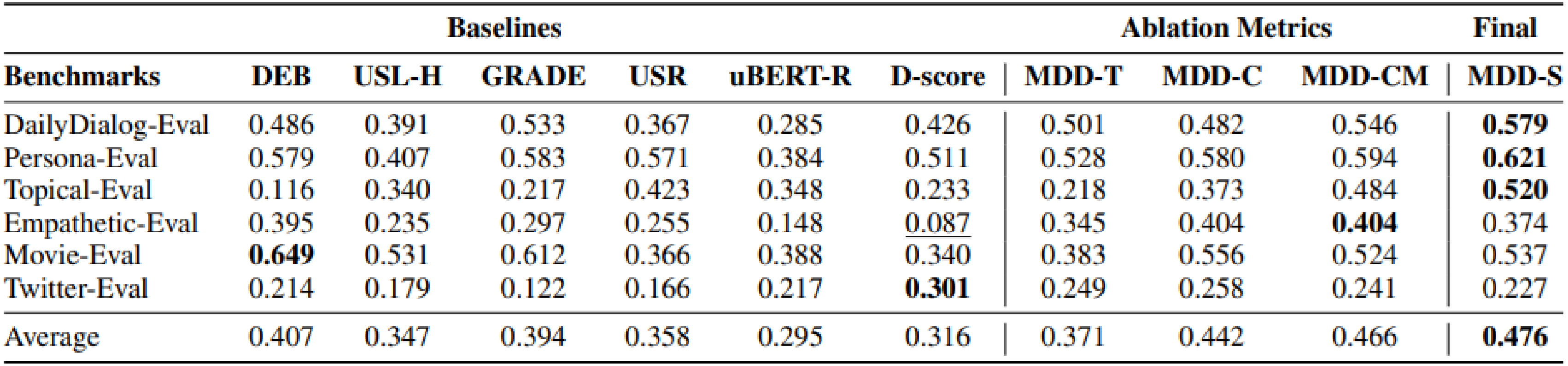
- Step 3. Pseudo-labeling
  - The teacher model assigns soft labels to the augmented unlabeled data
  - The soft pseudo label is a probability distribution over the three classes (random, adversarial and relevant)
  - Keep data that the teacher is confident about (class probability >= 0.7)
- Step 4. Student model training
  - Optimized with three losses: (1) Standard classification cross-entropy loss; (2) Masked language model loss; (3) KL loss for consistency regularization
  - Consistency regularization (Bachman et al. 2014)



- Bachman et al. "Learning with pseudo-ensembles." NIPS (2014).

Representative Examples - MDD-Eval

- Performance
  - MDD-S (the student model) works the best across different datasets
  - The teacher model (MDD-T) and DEB perform better than USL-H, GRADE, and USR, showcasing the importance of using adversarial response for training



	Baselines						Ablation Metrics			Final
Benchmarks	**DEB**	**USL-H**	**GRADE**	**USR**	**uBERT-R**	**D-score**	**MDD-T**	**MDD-C**	**MDD-CM**	**MDD-S**
DailyDialog-Eval	0.486	0.391	0.533	0.367	0.285	0.426	0.501	0.482	0.546	**0.579**
Persona-Eval	0.579	0.407	0.583	0.571	0.384	0.511	0.528	0.580	0.594	**0.621**
Topical-Eval	0.116	0.340	0.217	0.423	0.348	0.233	0.218	0.373	0.484	**0.520**
Empathetic-Eval	0.395	0.235	0.297	0.255	0.148	0.087	0.345	0.404	**0.404**	0.374
Movie-Eval	**0.649**	0.531	0.612	0.366	0.388	0.340	0.383	0.556	0.524	0.537
Twitter-Eval	0.214	0.179	0.122	0.166	0.217	**0.301**	0.249	0.258	0.241	0.227
Average	0.407	0.347	0.394	0.358	0.295	0.316	0.371	0.442	0.466	**0.476**

Q&A

Dialogue-Level Metrics

Representative Examples - FED

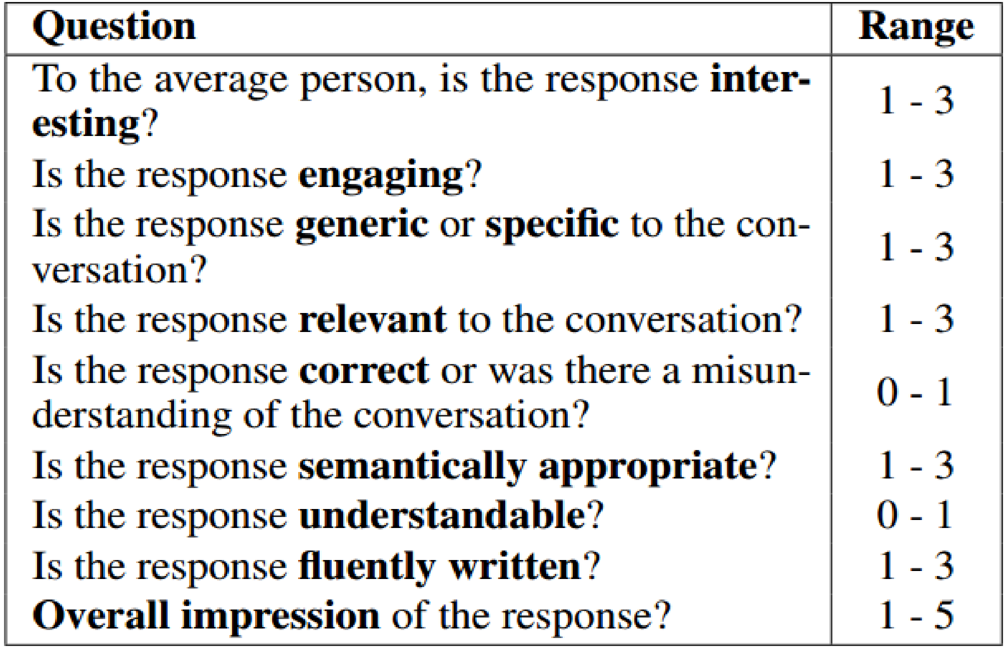
- Turn-level evaluation and dialogue-level evaluation are very different
  - Static evaluation vs interactive evaluation
  - Some erroneous behaviors can only be captured after observing the entire dialogue
  - Turn-level and dialogue-level focus on different set of dialogue quality dimensions

- Most existing metrics and evaluation datasets are about turn-level evaluation
  - FED (Fine-grained Evaluation of Dialogue) (Mehri & Eskenazi, 2020) targets multi-dimensional evaluation at both the turn-level and the dialogue-level
  - A high-quality dataset, named the "FED dataset", is created to facilitate research on dialogue-level evaluation

- Mehri and Eskenazi. "Unsupervised Evaluation of Interactive Dialog with DialoGPT." SIGDial (2020).

Representative Examples - FED

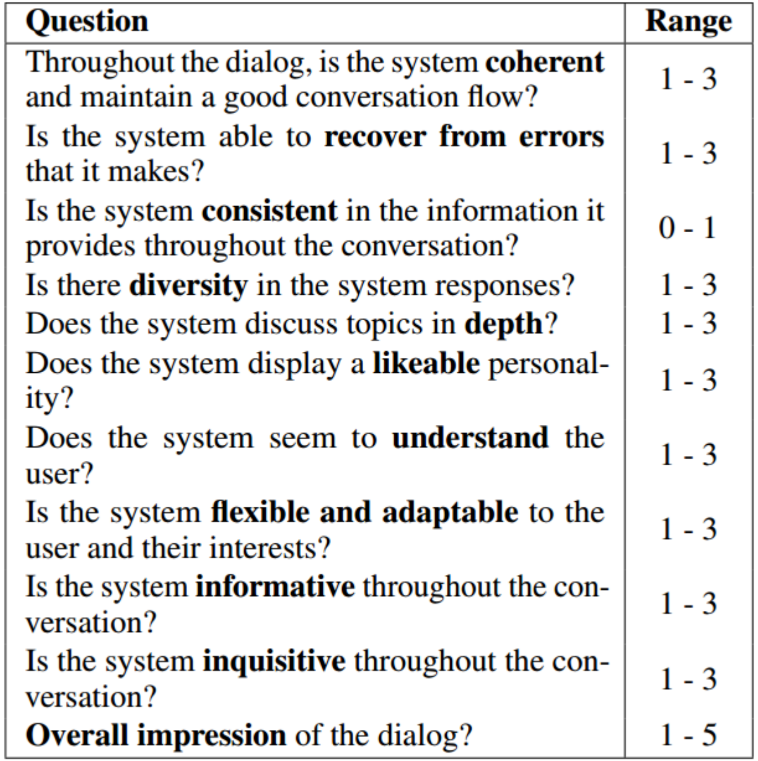
- Turn-level annotations in the FED dataset



Question	Range
To the average person, is the response **interesting**?	1 - 3
Is the response **engaging**?	1 - 3
Is the response **generic** or **specific** to the conversation?	1 - 3
Is the response **relevant** to the conversation?	1 - 3
Is the response **correct** or was there a misunderstanding of the conversation?	0 - 1
Is the response **semantically appropriate**?	1 - 3
Is the response **understandable**?	0 - 1
Is the response **fluently written**?	1 - 3
Overall impression of the response?	1 - 5

Representative Examples - FED

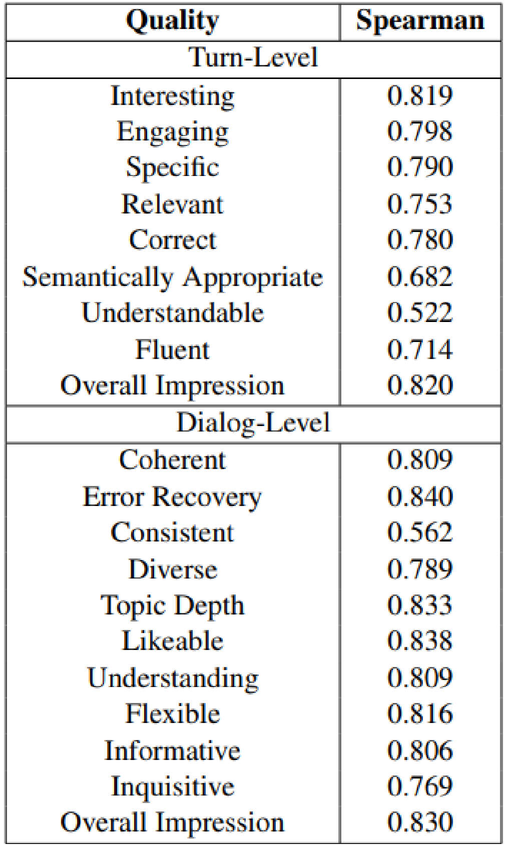
- Dialogue-level annotations in the FED dataset



Question	Range
Throughout the dialog, is the system **coherent** and maintain a good conversation flow?	1 - 3
Is the system able to **recover from errors** that it makes?	1 - 3
Is the system **consistent** in the information it provides throughout the conversation?	0 - 1
Is there **diversity** in the system responses?	1 - 3
Does the system discuss topics in **depth**?	1 - 3
Does the system display a **likeable** personality?	1 - 3
Does the system seem to **understand** the user?	1 - 3
Is the system **flexible and adaptable** to the user and their interests?	1 - 3
Is the system **informative** throughout the conversation?	1 - 3
Is the system **inquisitive** throughout the conversation?	1 - 3
Overall impression of the dialog?	1 - 5

Representative Examples - FED

- The FED dataset details
    - ○ 125 dialogue are annotated (41 Human-Meena, 44 Human-Mitsuku, 40 Human-Human)
    - ○ For each conversation, three system responses were hand-selected to be annotated at the turn level (375 annotated responses)
    - ○ Each data instance is annotated by five crowdsource workers
    - ○ The inter-annotator agreements are high for all the dimensions



Quality	**Spearman**
Turn-Level	
Interesting	0.819
Engaging	0.798
Specific	0.790
Relevant	0.753
Correct	0.780
Semantically Appropriate	0.682
Understandable	0.522
Fluent	0.714
Overall Impression	0.820
Dialog-Level	
Coherent	0.809
Error Recovery	0.840
Consistent	0.562
Diverse	0.789
Topic Depth	0.833
Likeable	0.838
Understanding	0.809
Flexible	0.816
Informative	0.806
Inquisitive	0.769
Overall Impression	0.830

Representative Examples - FED

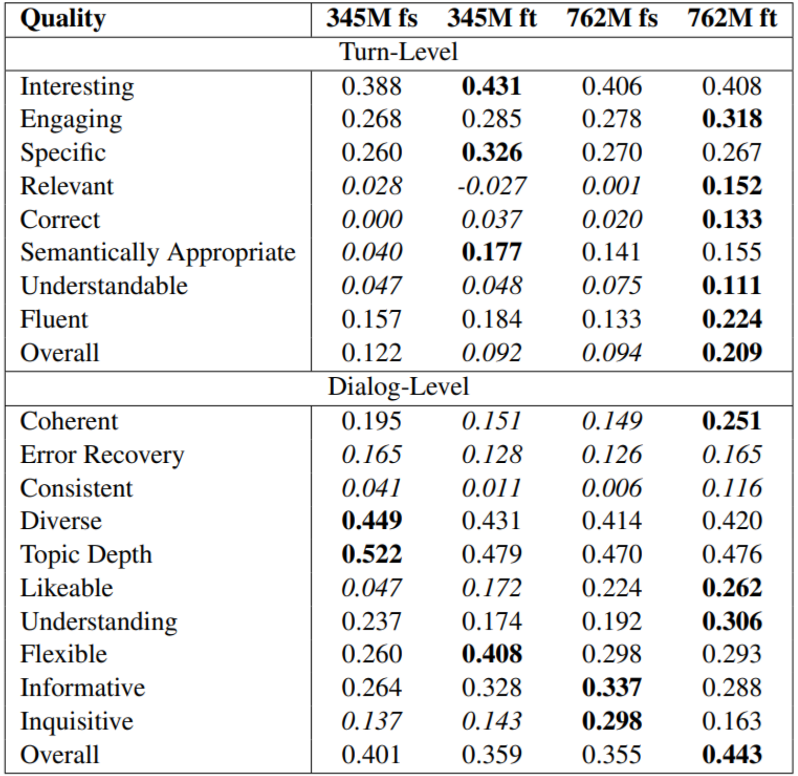
- The FED metric
  - Follow-Up utterance for evaluation - compute the likelihood of the model generating various follow-up utterances

$$\sum_{i=1}^{|p|} \mathcal{D}(c+r, p_i) - \sum_{i=1}^{|n|} \mathcal{D}(c+r, n_i)$$

  - Positive follow-up utterances for “interestingness”: ["Wow that is really interesting.", "That's really interesting!", "Cool! That sounds super interesting."]
  - Negative follow-up utterances for “interestingness”: ["That's not very interesting.", "That's really boring.", "That was a really boring response."]

Representative Examples - FED

- Performance
  - The table shows correlations of different FED metric variants using different versions of DialoGPT respectively
  - FED works fairly good across different dimensions at both turn-level and dialogue-level



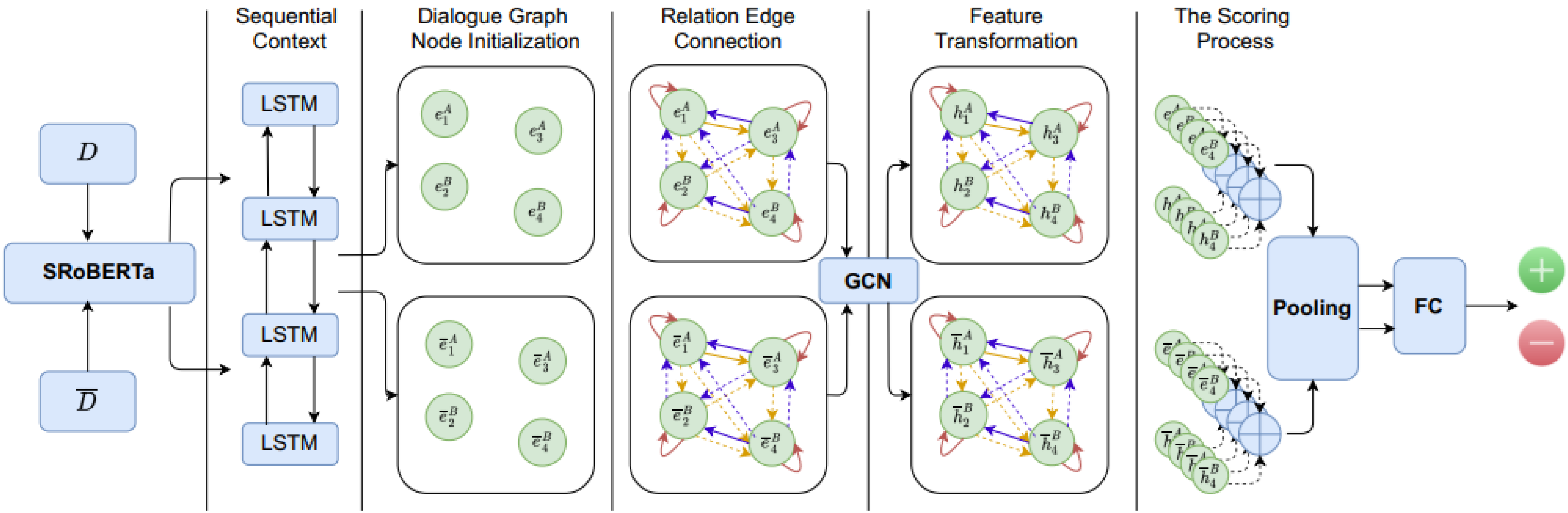
Quality	**345M fs**	**345M ft**	**762M fs**	**762M ft**
		Turn-Level		
Interesting	0.388	**0.431**	0.406	0.408
Engaging	0.268	0.285	0.278	**0.318**
Specific	0.260	**0.326**	0.270	0.267
Relevant	*0.028*	*-0.027*	*0.001*	**0.152**
Correct	*0.000*	*0.037*	*0.020*	**0.133**
Semantically Appropriate	*0.040*	**0.177**	0.141	0.155
Understandable	*0.047*	*0.048*	*0.075*	**0.111**
Fluent	0.157	0.184	0.133	**0.224**
Overall	0.122	*0.092*	*0.094*	**0.209**
		Dialog-Level		
Coherent	0.195	*0.151*	*0.149*	**0.251**
Error Recovery	*0.165*	*0.128*	*0.126*	*0.165*
Consistent	*0.041*	*0.011*	*0.006*	*0.116*
Diverse	**0.449**	0.431	0.414	0.420
Topic Depth	**0.522**	0.479	0.470	0.476
Likeable	*0.047*	*0.172*	0.224	**0.262**
Understanding	0.237	0.174	0.192	**0.306**
Flexible	0.260	**0.408**	0.298	0.293
Informative	0.264	0.328	**0.337**	0.288
Inquisitive	*0.137*	*0.143*	**0.298**	0.163
Overall	0.401	0.359	0.355	**0.443**

Representative Examples - DynaEval

- Static evaluation cannot capture dialogue-level errors
- Dialogue is essentially a multi-turn, dynamic, and interactive process between the interlocutors
- Two types of dependency in the interactive process (Ghosal et al., 2019)
    - Speaker Level Dependency
    - Utterance Level Dependency
- DynaEval (Zhang et al., 2021) adopts the graph structure to model the interactive process

- Ghosal et al. "DialogueGCN: A graph convolutional neural network for emotion recognition in conversation." EMNLP (2019).
- Zhang et al. "DynaEval: Unifying Turn and Dialogue Level Evaluation." ACL-IJCNLP (2021).

Representative Examples - DynaEval



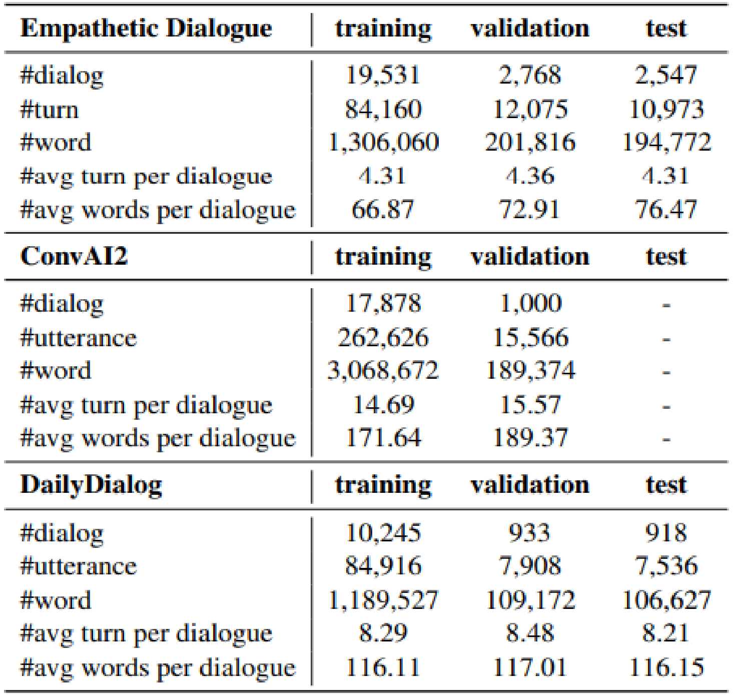
Representative Examples - DynaEval

- Margin ranking loss

$$y = \begin{cases} 1 & \text{if } D \text{ is preferred over } \bar{D} \\ -1 & \text{if } \bar{D} \text{ is preferred over } D \end{cases}$$

$$\mathcal{L} = \max(0, -y * (s_{dial} - s_{d\bar{i}al}) + 1)$$

- Negative Sampling Strategies
  - Utterance Replacement (UR)
  - Speaker Level Utterance Shuffling (SS)
- Training Datasets
  - EmpatheticDialogue (Rashkin et al., 2019)
  - DailyDialog (Li et al., 2017)
  - ConvAI2 (Dinan et al., 2020)

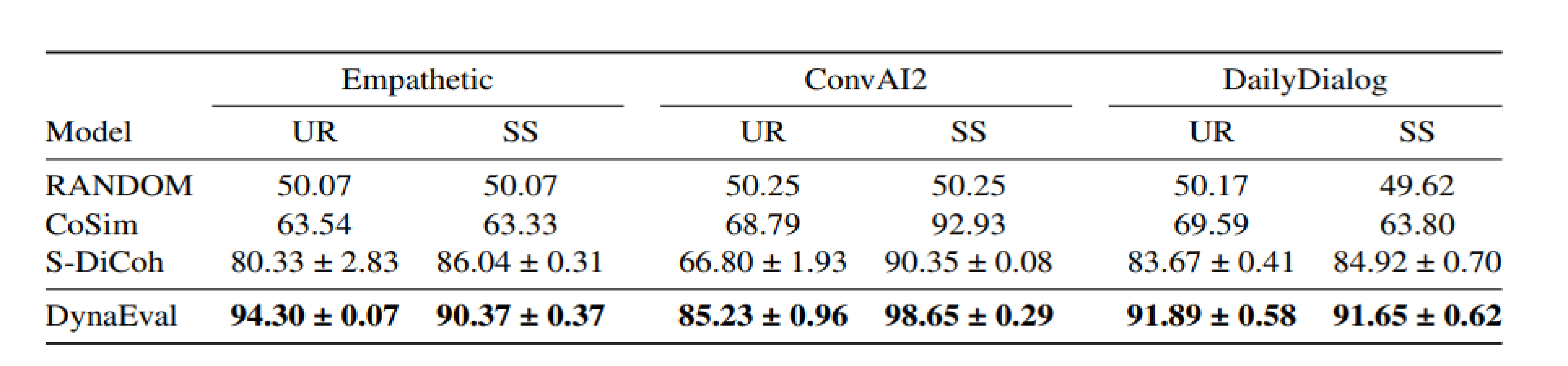


Empathetic Dialogue	**training**	**validation**	**test**
#dialog	19,531	2,768	2,547
#turn	84,160	12,075	10,973
#word	1,306,060	201,816	194,772
#avg turn per dialogue	4.31	4.36	4.31
#avg words per dialogue	66.87	72.91	76.47
ConvAI2	**training**	**validation**	**test**
#dialog	17,878	1,000	-
#utterance	262,626	15,566	-
#word	3,068,672	189,374	-
#avg turn per dialogue	14.69	15.57	-
#avg words per dialogue	171.64	189.37	-
DailyDialog	**training**	**validation**	**test**
#dialog	10,245	933	918
#utterance	84,916	7,908	7,536
#word	1,189,527	109,172	106,627
#avg turn per dialogue	8.29	8.48	8.21
#avg words per dialogue	116.11	117.01	116.15

- Rashkin et al. “Towards Empathetic Open-domain Conversation Models: A New Benchmark and Dataset.” ACL (2019).
- Li et al. "Dailydialog: A manually labelled multi-turn dialogue dataset." IJCNLP (2017).
- Dinan et al. “The second conversational intelligence challenge (ConvAI2).” The NeurIPS’18 Competition, Springer (2020).

Representative Examples - DynaEval

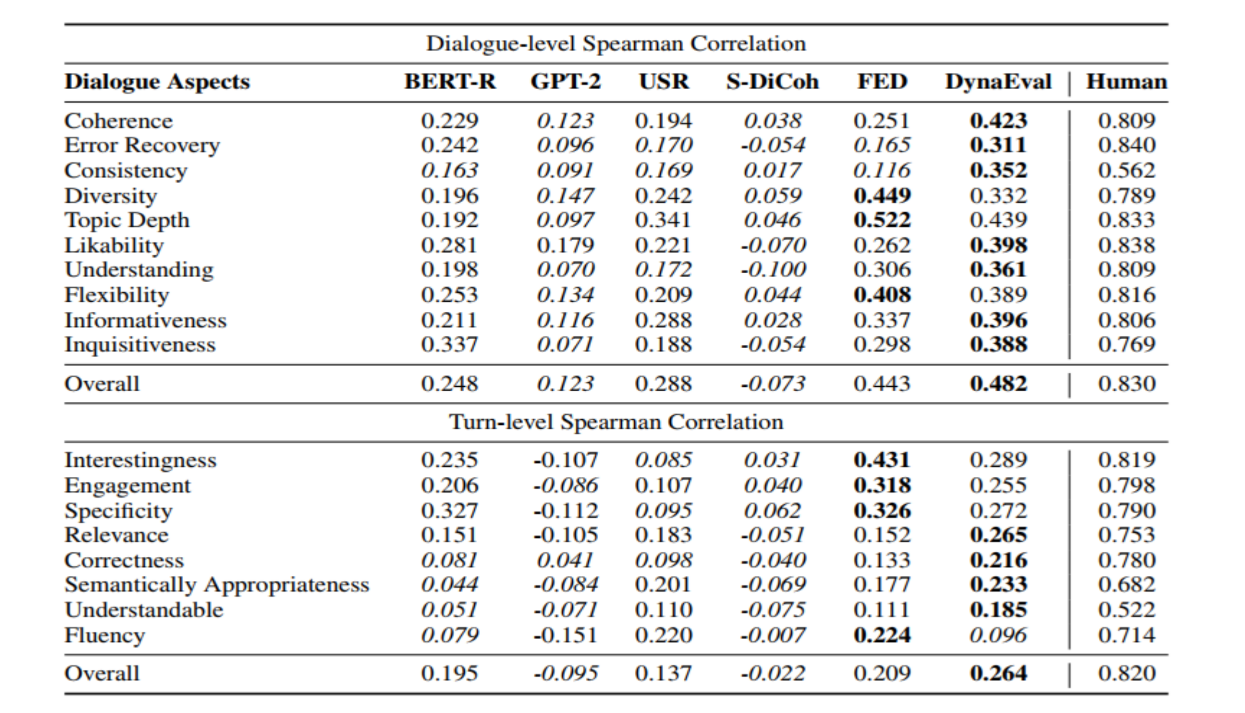
- Performance
  - Coherence evaluation - binary classification task
  - DynaEval is capable of discriminating positive dialogue samples from negative ones, outperforming baselines by a significant margin



	Empathetic		ConvAI2		DailyDialog	
Model	UR	SS	UR	SS	UR	SS
RANDOM	50.07	50.07	50.25	50.25	50.17	49.62
CoSim	63.54	63.33	68.79	92.93	69.59	63.80
S-DiCoh	80.33 ± 2.83	86.04 ± 0.31	66.80 ± 1.93	90.35 ± 0.08	83.67 ± 0.41	84.92 ± 0.70
DynaEval	**94.30 ± 0.07**	**90.37 ± 0.37**	**85.23 ± 0.96**	**98.65 ± 0.29**	**91.89 ± 0.58**	**91.65 ± 0.62**

Representative Examples - DynaEval

- Performance
  - Correlation analysis on the FED dataset
  - DynaEval significantly outperforms turn-level metrics across all dimensions
  - DynaEval and FED are complementary
  - Still far from human upper bound



Dialogue-level Spearman Correlation							
Dialogue Aspects	**BERT-R**	**GPT-2**	**USR**	**S-DiCoh**	**FED**	**DynaEval**	**Human**
Coherence	0.229	*0.123*	0.194	*0.038*	0.251	**0.423**	0.809
Error Recovery	0.242	*0.096*	*0.170*	*-0.054*	*0.165*	**0.311**	0.840
Consistency	*0.163*	*0.091*	0.169	*0.017*	*0.116*	**0.352**	0.562
Diversity	0.196	*0.147*	0.242	*0.059*	**0.449**	0.332	0.789
Topic Depth	0.192	*0.097*	0.341	*0.046*	**0.522**	0.439	0.833
Likability	0.281	0.179	0.221	*-0.070*	0.262	**0.398**	0.838
Understanding	0.198	*0.070*	*0.172*	*-0.100*	0.306	**0.361**	0.809
Flexibility	0.253	*0.134*	0.209	*0.044*	**0.408**	0.389	0.816
Informativeness	0.211	*0.116*	0.288	*0.028*	0.337	**0.396**	0.806
Inquisitiveness	0.337	*0.071*	0.188	*-0.054*	0.298	**0.388**	0.769
Overall	0.248	*0.123*	0.288	*-0.073*	0.443	**0.482**	0.830
Turn-level Spearman Correlation							
Interestingness	0.235	-0.107	*0.085*	*0.031*	**0.431**	0.289	0.819
Engagement	0.206	*-0.086*	0.107	*0.040*	**0.318**	0.255	0.798
Specificity	0.327	-0.112	*0.095*	*0.062*	**0.326**	0.272	0.790
Relevance	0.151	-0.105	0.183	*-0.051*	0.152	**0.265**	0.753
Correctness	*0.081*	*0.041*	*0.098*	*-0.040*	0.133	**0.216**	0.780
Semantically Appropriateness	*0.044*	*-0.084*	0.201	*-0.069*	0.177	**0.233**	0.682
Understandable	*0.051*	*-0.071*	0.110	*-0.075*	0.111	**0.185**	0.522
Fluency	*0.079*	-0.151	0.220	*-0.007*	**0.224**	*0.096*	0.714
Overall	0.195	*-0.095*	0.137	*-0.022*	0.209	**0.264**	0.820

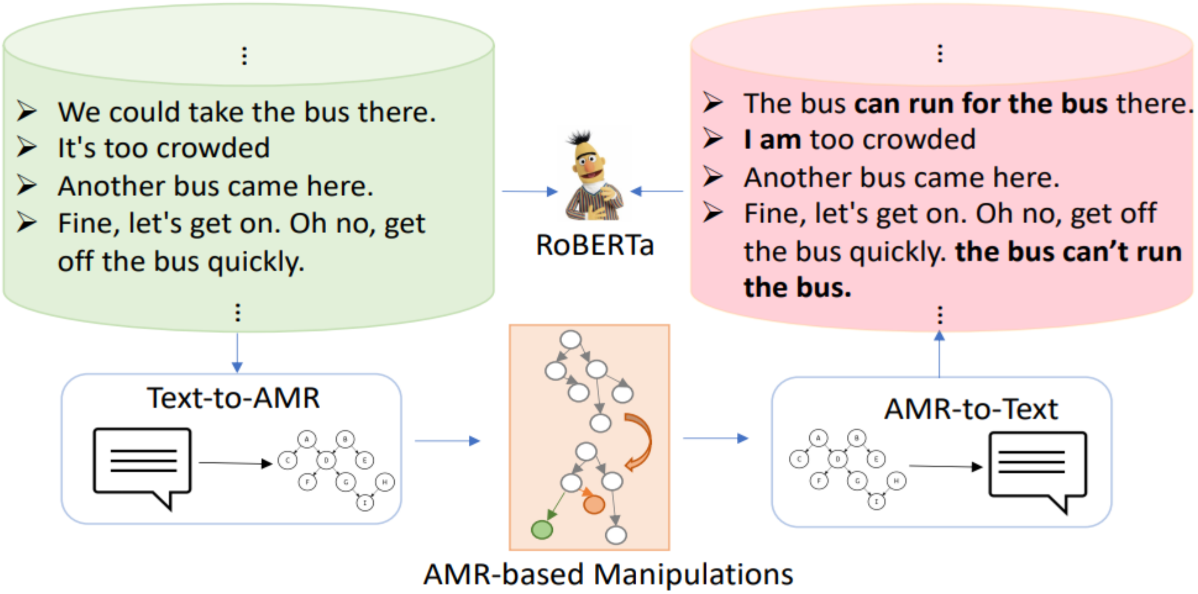
Representative Examples - DEAM

- Coherence - measures how well the utterances in a conversation are unified leading to a consistent interaction
- Existing methods, such as DynaEval, rely on discriminating original H-H dialogues and the heuristically generated negative samples
- Heuristic text-level manipulations are insufficient to reflect errors in advanced dialogue systems
- DEAM (Ghazarian et al., 2022) Apply Abstract Meaning Representation (AMR) for semantic perturbation
    - Coreference inconsistency
    - Irrelevancy
    - Contradictions
    - Decrease engagement

- Ghazarian et al. "DEAM: Dialogue Coherence Evaluation using AMR-based Semantic Manipulations." ACL (2022).

Representative Examples - DEAM

- Metric Details - Overview



⋮

- ➢ We could take the bus there.
- ➢ It's too crowded
- ➢ Another bus came here.
- ➢ Fine, let's get on. Oh no, get off the bus quickly.

⋮

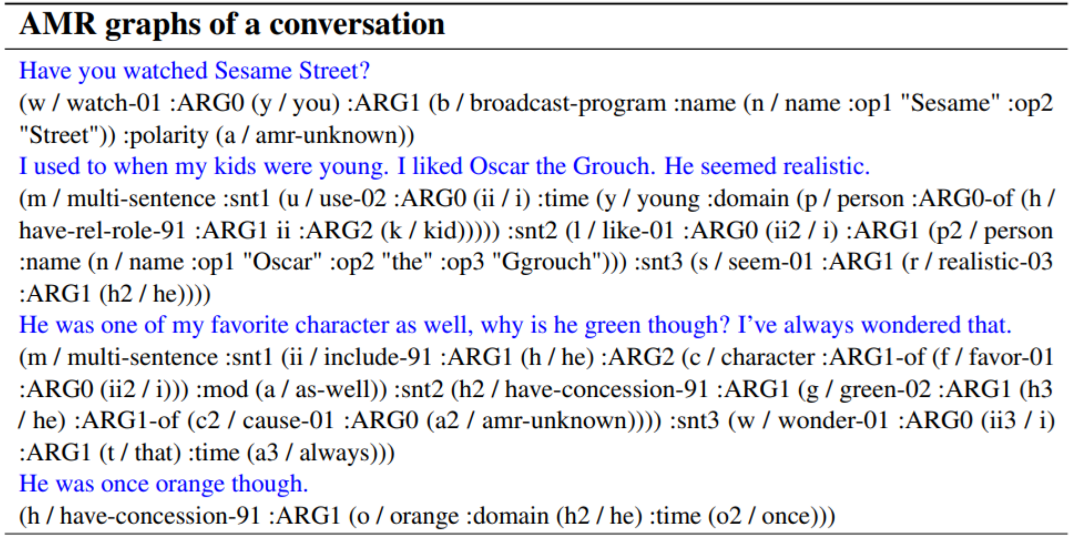
⋮

- ➢ The bus **can run for the bus** there.
- ➢ **I am** too crowded
- ➢ Another bus came here.
- ➢ Fine, let's get on. Oh no, get off the bus quickly. **the bus can't run the bus.**

⋮

Representative Examples - DEAM

- Metric Details - Text-to-AMR & AMR-to-Text
  - Pre-trained AMR parsing model translates conversation texts to directed and acyclic AMR graphs
  - The graph contains relation edges between concept nodes
  - Perform specific manipulations on the AMR graph
  - Adopt pre-trained AMR-to-Text generation model to convert the perturbed graph back to conversation texts



AMR graphs of a conversation

Have you watched Sesame Street?

(w / watch-01 :ARG0 (y / you) :ARG1 (b / broadcast-program :name (n / name :op1 "Sesame" :op2 "Street")) :polarity (a / amr-unknown))

I used to when my kids were young. I liked Oscar the Grouch. He seemed realistic.

(m / multi-sentence :snt1 (u / use-02 :ARG0 (ii / i) :time (y / young :domain (p / person :ARG0-of (h / have-rel-role-91 :ARG1 ii :ARG2 (k / kid))))) :snt2 (l / like-01 :ARG0 (ii2 / i) :ARG1 (p2 / person :name (n / name :op1 "Oscar" :op2 "the" :op3 "Ggrouch"))) :snt3 (s / seem-01 :ARG1 (r / realistic-03 :ARG1 (h2 / he))))

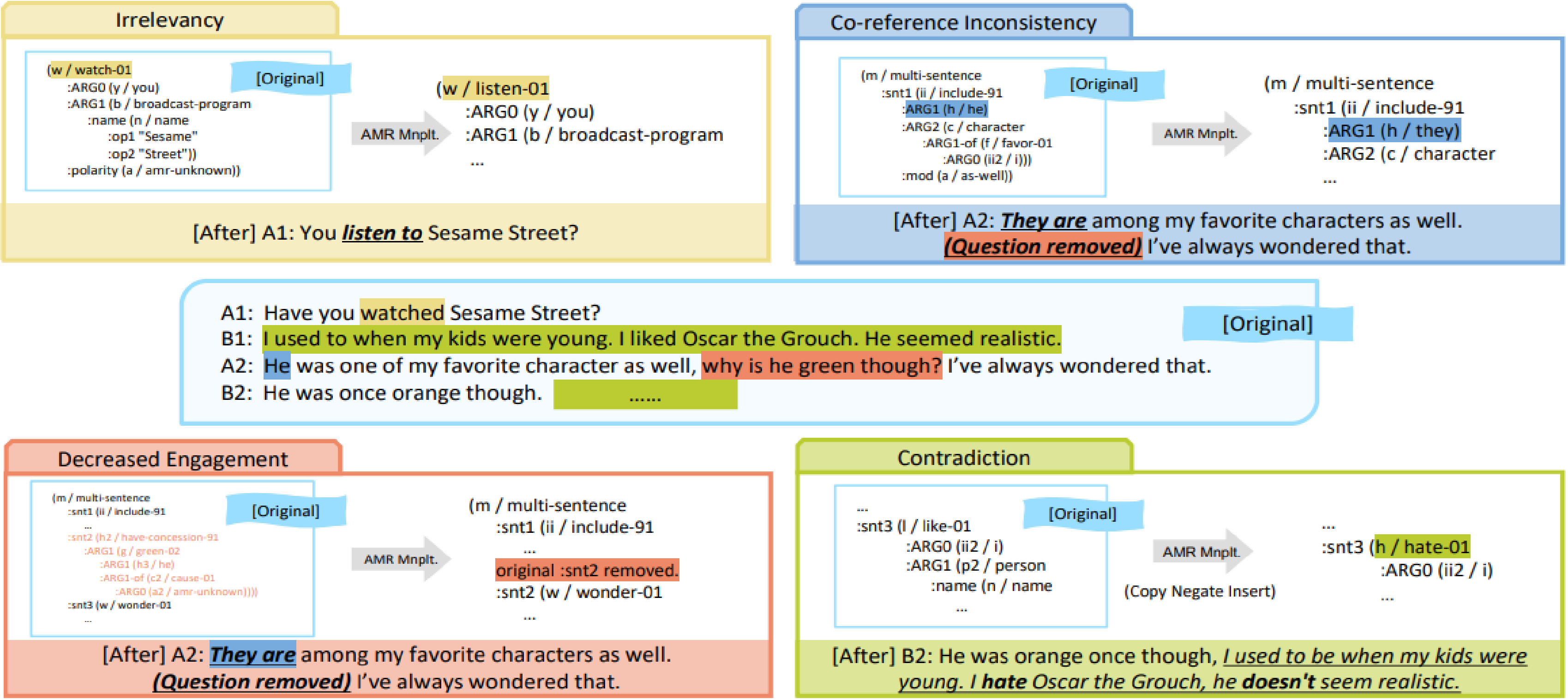
He was one of my favorite character as well, why is he green though? I've always wondered that.

(m / multi-sentence :snt1 (ii / include-91 :ARG1 (h / he) :ARG2 (c / character :ARG1-of (f / favor-01 :ARG0 (ii2 / i))) :mod (a / as-well)) :snt2 (h2 / have-concession-91 :ARG1 (g / green-02 :ARG1 (h3 / he) :ARG1-of (c2 / cause-01 :ARG0 (a2 / amr-unknown)))) :snt3 (w / wonder-01 :ARG0 (ii3 / i) :ARG1 (t / that) :time (a3 / always)))

He was once orange though.

(h / have-concession-91 :ARG1 (o / orange :domain (h2 / he) :time (o2 / once)))

Representative Examples - DEAM



Irrelevancy

[Original]

```
(w / watch-01
    :ARG0 (y / you)
    :ARG1 (b / broadcast-program
        :name (n / name
            :op1 "Sesame"
            :op2 "Street"))
    :polarity (a / amr-unknown))
```

```
(w / listen-01
    :ARG0 (y / you)
    :ARG1 (b / broadcast-program
    ...
```

[After] A1: You ***listen to*** Sesame Street?

Co-reference Inconsistency

[Original]

```
(m / multi-sentence
    :snt1 (ii / include-91
        :ARG1 (h / he)
        :ARG2 (c / character
            :ARG1-of (f / favor-01
                :ARG0 (ii2 / i)))
        :mod (a / as-well))
```

```
(m / multi-sentence
    :snt1 (ii / include-91
        :ARG1 (h / they)
        :ARG2 (c / character
        ...
```

[After] A2: ***They are*** among my favorite characters as well.
(Question removed) I've always wondered that.

A1: Have you watched Sesame Street?
B1: I used to when my kids were young. I liked Oscar the Grouch. He seemed realistic.
A2: He was one of my favorite character as well, why is he green though? I've always wondered that.
B2: He was once orange though. ......

[Original]

Decreased Engagement

[Original]

```
(m / multi-sentence
    :snt1 (ii / include-91
        ...
    :snt2 (h2 / have-concession-91
        :ARG1 (g / green-02
            :ARG1 (h3 / he)
            :ARG1-of (c2 / cause-01
                :ARG0 (a2 / amr-unknown))))
    :snt3 (w / wonder-01
        ...
```

```
(m / multi-sentence
    :snt1 (ii / include-91
        ...
    original :snt2 removed.
    :snt2 (w / wonder-01
        ...
```

[After] A2: ***They are*** among my favorite characters as well.
(Question removed) I've always wondered that.

Contradiction

[Original]

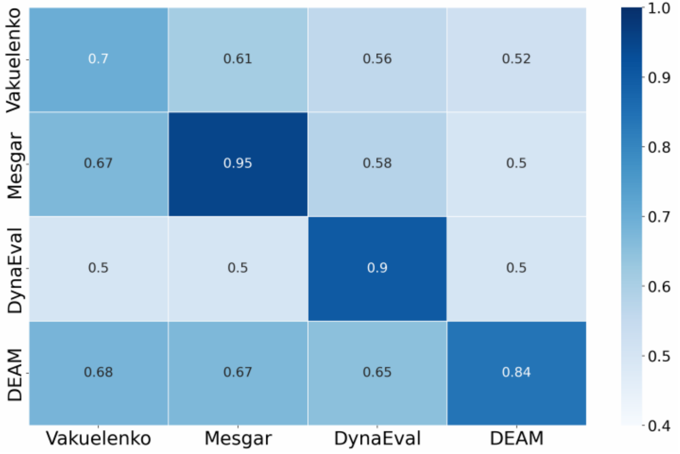
```
...
:snt3 (l / like-01
        :ARG0 (ii2 / i)
        :ARG1 (p2 / person
            :name (n / name
        ...
```

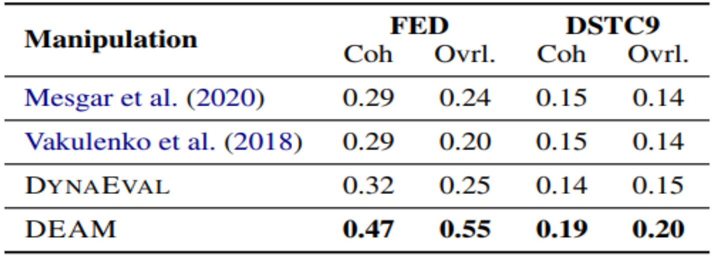
```
...
:snt3 (h / hate-01
        :ARG0 (ii2 / i)
        ...
```

[After] B2: He was orange once though, *I used to be when my kids were young. I **hate** Oscar the Grouch, he **doesn't** seem realistic.*

Representative Examples - DEAM

- Performance
  - DEAM is able to distinguish incoherent dialogues generated by both baseline perturbations and AMR-based perturbations
  - The baselines only performs well with their respective perturbation strategies
  - DEAM significantly outperforms the baselines on both FED and DSTC9 along Coherence and Overall





Manipulation	FED		DSTC9	
	Coh	Ovrl.	Coh	Ovrl.
Mesgar et al. (2020)	0.29	0.24	0.15	0.14
Vakulenko et al. (2018)	0.29	0.20	0.15	0.14
DYNAEVAL	0.32	0.25	0.14	0.15
DEAM	**0.47**	**0.55**	**0.19**	**0.20**

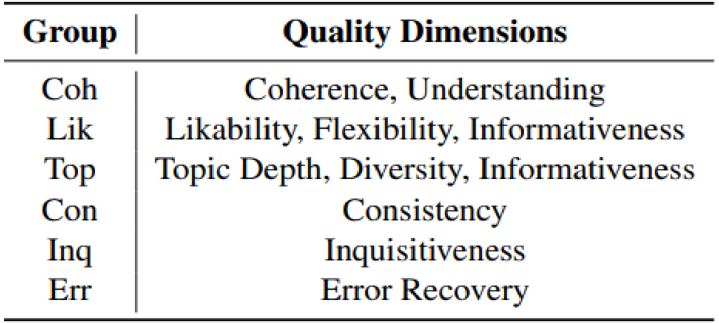
Representative Examples - FineD-Eval

- The correlation of DEAM and DynaEval with human evaluation is still not strong.
- A major reason is that their perturbation strategy only targets dialogue coherence.
- We should consider more fine-grained dimensions when designing dialogue-level metrics.
- Fine-grained Automatic Dialogue-Level Evaluation (Zhang et al., 2022) Target multi-dimensional evaluation at the dialogue level.

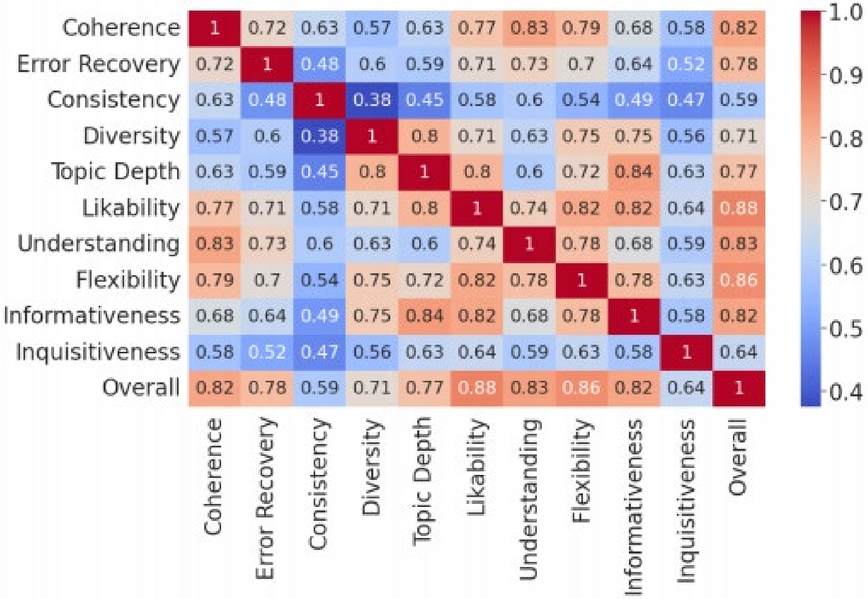
- Zhang et al. "FineD-Eval: Fine-grained Automatic Dialogue-Level Evaluation." EMNLP (2022).

Representative Examples - FineD-Eval

- Categorization of dimensions based on correlation analysis of human scores
  - Six different groups for 10 different fine-grained dimensions.
  - FineD-Eval targets three of them, Coh, Lik, and Top.



Group	Quality Dimensions
Coh	Coherence, Understanding
Lik	Likability, Flexibility, Informativeness
Top	Topic Depth, Diversity, Informativeness
Con	Consistency
Inq	Inquisitiveness
Err	Error Recovery



Representative Examples - FineD-Eval

- Metric Details - Overview
  - Adopt preference learning to train different sub-metrics (similar to DynaEval)

$$\mathcal{L}_q = max(0, y * (x_1^q - x_2^q) + 0.1)$$

  - Each sub-metric measures one dimension - Coherence, Likability, and Topic Depth
  - The sub-metrics are combined through ensemble or multitask learning

Representative Examples - FineD-Eval

- Metric Details - Coherence
  - Utterance order shuffling
    - Randomly permute the order of utterances in human-human dialogues
  - Question-answer (QA) relevance scoring
    - Select dialogues in existing dialogue corpora that are more than 4 utterances and contain at least one question-answer pair
    - Use a pretrained BERT-based QA evaluator to score each QA pair within a dialogue.
    - Average the relevance scores of all QA pairs within the dialogue to derive the dialogue-level QA relevance score.
    - Those with low QA relevance score are treated as incoherent dialogues

Representative Examples - FineD-Eval

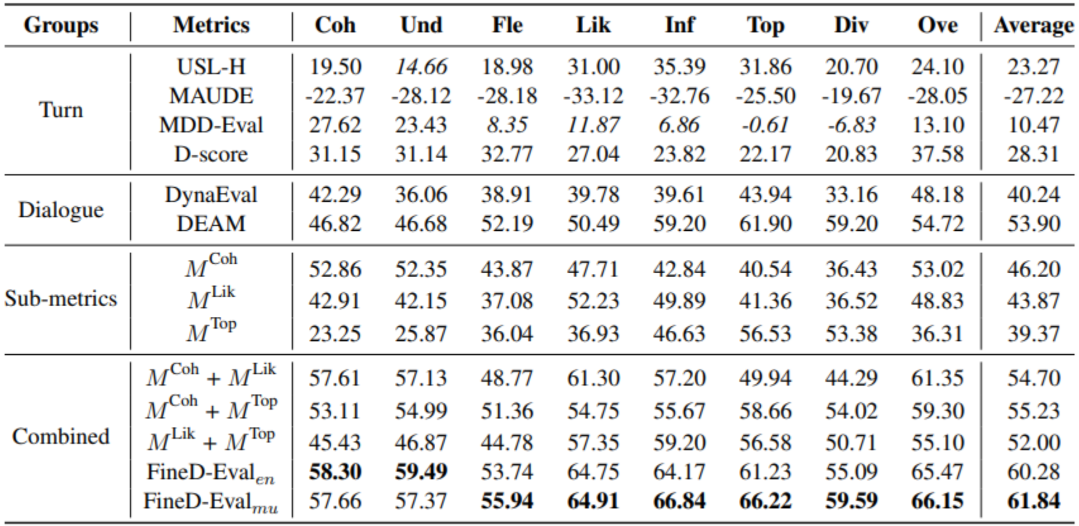
- Metric Details - Likability
  - Contradiction scoring
    - People tend to favour others who share similar opinions or preferences with them
    - Adopt a pre-trained NLI model to provide contradiction score to adjacent utterance pairs
    - For a dialogue containing k utterances, we have k −1 adjacency pairs, thus k −1 contradiction scores.
    - The dialogue-level contradiction score is derived by computing the average of the k − 1 scores
  - Number of utterances that carry positive sentiment
    - A pre-trained sentiment classification model is used to classify the emotion of utterances

Representative Examples - FineD-Eval

- Metric Details - Topic Depth
  - Entailment scoring
    - A dialogue with good topic depth rating should carry much more information than a dull dialogue
    - Entailment is a way to measure semantic similarity between a pair of utterances
    - A content-rich dialogue should contain less similar utterances, hence, low entailment score
    - A pre-trained NLI model is adopted to score each utterance pair in a dialogue
    - The dialogue-level entailment score is obtained by averaging the entailment scores of all utterance pairs within the dialogue

Representative Examples - FineD-Eval

- Performance
  - FineD-Eval yields significantly better correlation across multiple dimensions as well as the overall dimension than both turn-level and dialogue-level baselines
  - The designed sub-metrics work as expected - they perform well along their respective target dimension

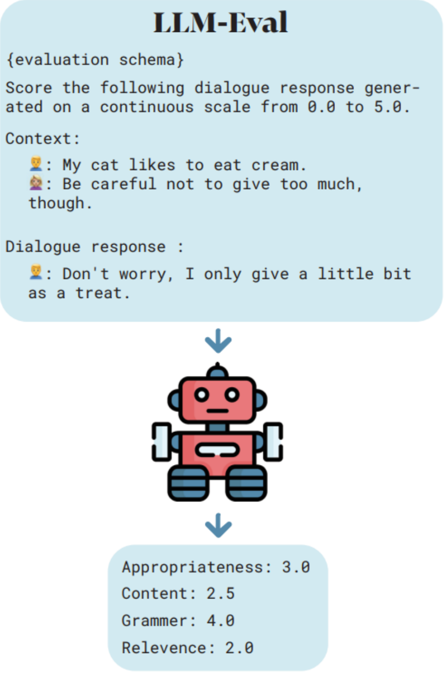


Groups	Metrics	Coh	Und	Fle	Lik	Inf	Top	Div	Ove	Average
Turn	USL-H	19.50	*14.66*	18.98	31.00	35.39	31.86	20.70	24.10	23.27
	MAUDE	-22.37	-28.12	-28.18	-33.12	-32.76	-25.50	-19.67	-28.05	-27.22
	MDD-Eval	27.62	23.43	*8.35*	*11.87*	*6.86*	*-0.61*	*-6.83*	13.10	10.47
	D-score	31.15	31.14	32.77	27.04	23.82	22.17	20.83	37.58	28.31
Dialogue	DynaEval	42.29	36.06	38.91	39.78	39.61	43.94	33.16	48.18	40.24
	DEAM	46.82	46.68	52.19	50.49	59.20	61.90	59.20	54.72	53.90
Sub-metrics	M^{Coh}	52.86	52.35	43.87	47.71	42.84	40.54	36.43	53.02	46.20
	M^{Lik}	42.91	42.15	37.08	52.23	49.89	41.36	36.52	48.83	43.87
	M^{Top}	23.25	25.87	36.04	36.93	46.63	56.53	53.38	36.31	39.37
Combined	$M^{\text{Coh}} + M^{\text{Lik}}$	57.61	57.13	48.77	61.30	57.20	49.94	44.29	61.35	54.70
	$M^{\text{Coh}} + M^{\text{Top}}$	53.11	54.99	51.36	54.75	55.67	58.66	54.02	59.30	55.23
	$M^{\text{Lik}} + M^{\text{Top}}$	45.43	46.87	44.78	57.35	59.20	56.58	50.71	55.10	52.00
	FineD-Eval$_{en}$	**58.30**	**59.49**	53.74	64.75	64.17	61.23	55.09	65.47	60.28
	FineD-Eval$_{mu}$	57.66	57.37	**55.94**	**64.91**	**66.84**	**66.22**	**59.59**	**66.15**	**61.84**

The Era of LLMs

Representative Examples – LLM-Eval

- LLM-Eval: a unified multi-dimensional automatic evaluation method with LLMs
  - Provide natural language instructions to LLMs
  - Prompt the LLMs to generate multi-dimensional scores in a JSON format
  - Right figure is the output format instruction (now this can be easily achieved with OpenAI's JSON mode)



LLM-Eval

{evaluation schema}

Score the following dialogue response generated on a continuous scale from 0.0 to 5.0.

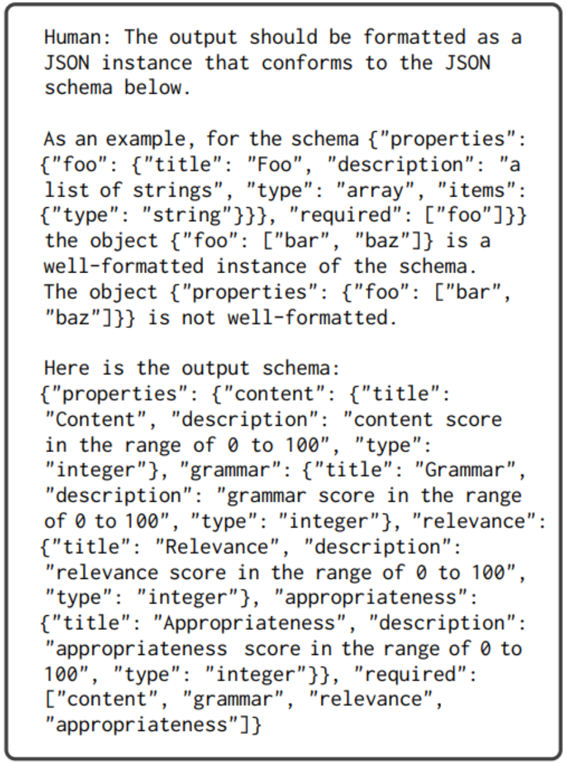
Context:

👱: My cat likes to eat cream.
💁: Be careful not to give too much, though.

Dialogue response :

👱: Don't worry, I only give a little bit as a treat.

Appropriateness: 3.0
Content: 2.5
Grammer: 4.0
Relevence: 2.0



Human: The output should be formatted as a JSON instance that conforms to the JSON schema below.

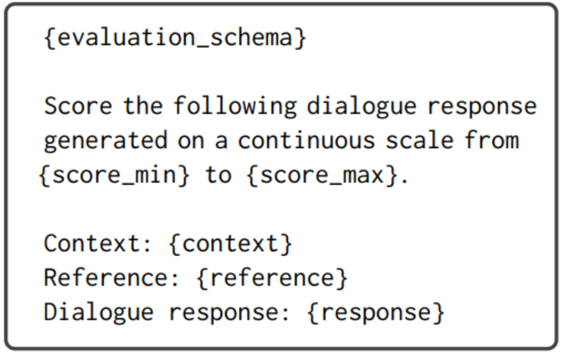
As an example, for the schema {"properties": {"foo": {"title": "Foo", "description": "a list of strings", "type": "array", "items": {"type": "string"}}}, "required": ["foo"]}} the object {"foo": ["bar", "baz"]} is a well-formatted instance of the schema. The object {"properties": {"foo": ["bar", "baz"]}} is not well-formatted.

Here is the output schema:
{"properties": {"content": {"title": "Content", "description": "content score in the range of 0 to 100", "type": "integer"}, "grammar": {"title": "Grammar", "description": "grammar score in the range of 0 to 100", "type": "integer"}, "relevance": {"title": "Relevance", "description": "relevance score in the range of 0 to 100", "type": "integer"}, "appropriateness": {"title": "Appropriateness", "description": "appropriateness score in the range of 0 to 100", "type": "integer"}}, "required": ["content", "grammar", "relevance", "appropriateness"]}

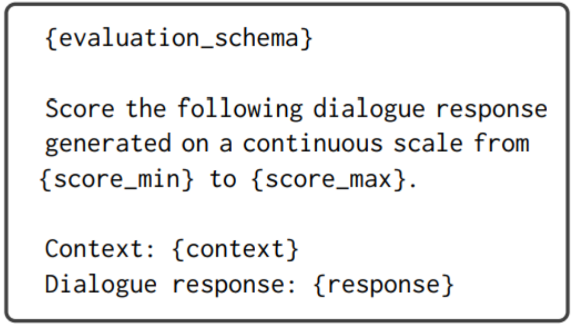
- Yen-Ting Lin and Yun-Nung Chen. LLM-Eval: Unified Multi-Dimensional Automatic Evaluation for Open-Domain Conversations with Large Language Models. NLP4ConvAI 2023,

Representative Examples – LLM-Eval

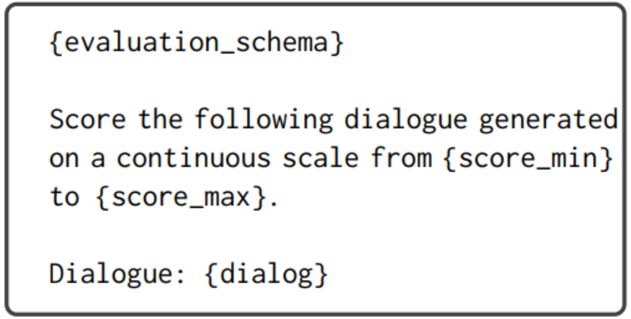
- LLM-Eval: a unified multi-dimensional automatic evaluation method with LLMs
  - Instructions to three different settings:
    - (1) Reference-based; (2) reference-free turn-level; (3) reference-free dialogue-level
  - The schema and output format are used together to prompt LLMs
  - The evaluation_schema defines the dimension to evaluate



```
{evaluation_schema}

Score the following dialogue response
generated on a continuous scale from
{score_min} to {score_max}.

Context: {context}
Reference: {reference}
Dialogue response: {response}
```



```
{evaluation_schema}

Score the following dialogue response
generated on a continuous scale from
{score_min} to {score_max}.

Context: {context}
Dialogue response: {response}
```

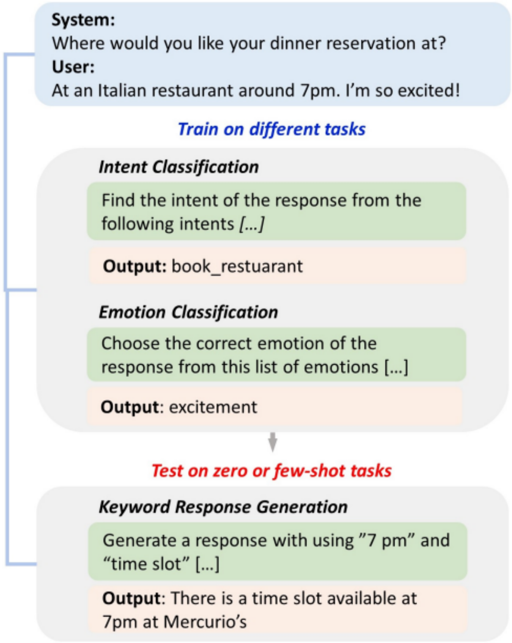


```
{evaluation_schema}

Score the following dialogue generated
on a continuous scale from {score_min}
to {score_max}.

Dialogue: {dialog}
```

- Yen-Ting Lin and Yun-Nung Chen. LLM-Eval: Unified Multi-Dimensional Automatic Evaluation for Open-Domain Conversations with Large Language Models. NLP4ConvAI 2023),

Representative Examples – InstructDial

- Format 52 diverse dialogue tasks into unified instruction-based text-to-text format
  - Incorporate both open-domain and task-oriented tasks
  - Classification, generation, dialogue generation, etc.
- Perform instruction-tuning with language models, such as BART and T0
- Train and evaluate on disjoint set of tasks



System:
Where would you like your dinner reservation at?
User:
At an Italian restaurant around 7pm. I'm so excited!

Train on different tasks

Intent Classification

Find the intent of the response from the following intents *[...]*

Output: book_restuarant

Emotion Classification

Choose the correct emotion of the response from this list of emotions [...]

Output: excitement

Test on zero or few-shot tasks

Keyword Response Generation

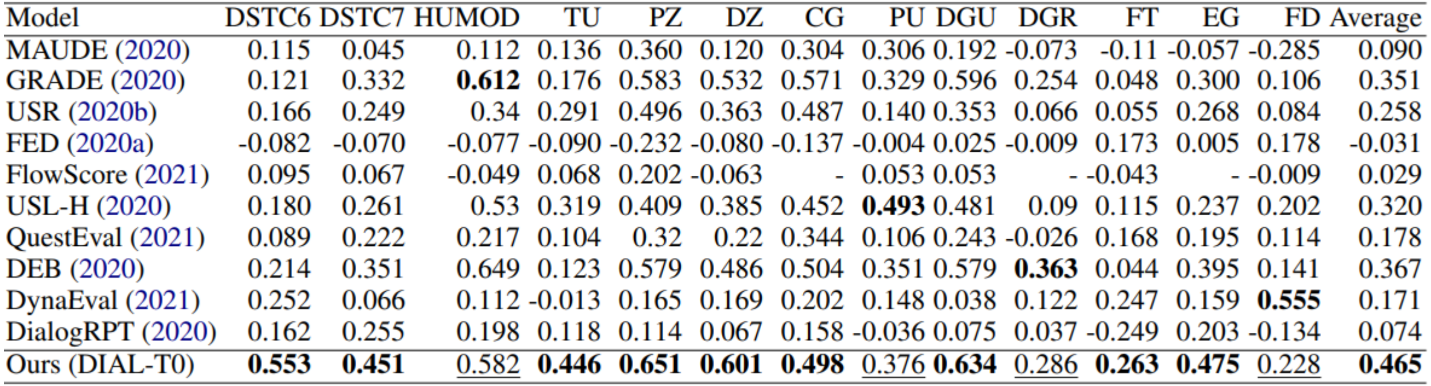
Generate a response with using "7 pm" and "time slot" [...]

Output: There is a time slot available at 7pm at Mercurio's

- Prakhar Gupta, et al. 2022. InstructDial: Improving Zero and Few-shot Generalization in Dialogue through Instruction Tuning. EMNLP-2022.

Representative Examples – InstructDial

- Prompting Instruction-Tuned Model for Automatic Dialogue Evaluation
  - Instruct the model to predict “yes” if the response is relevant to the context, otherwise predict “no”
  - calculate the probability of “yes” as p(yes) = p(yes)/(p(yes) + p(no))
  - Compute Spearman correlation of the model’s prediction with human ratings for relevance

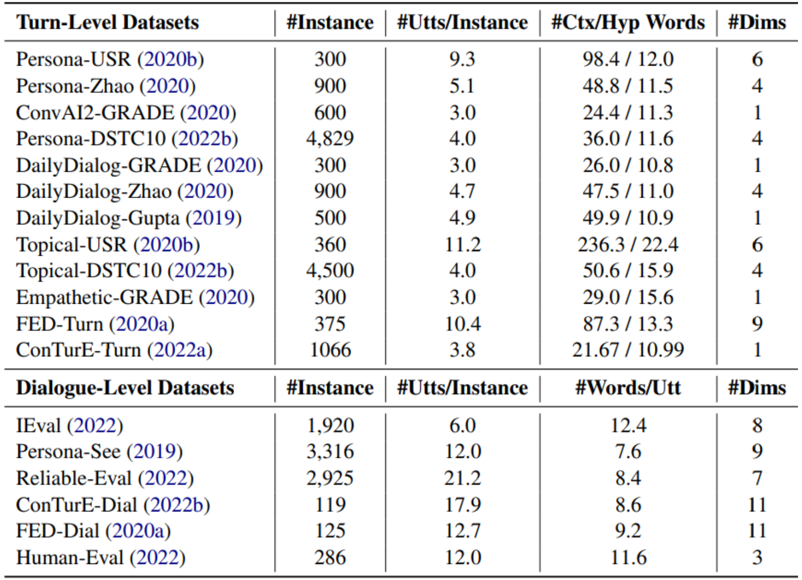


Model	DSTC6	DSTC7	HUMOD	TU	PZ	DZ	CG	PU	DGU	DGR	FT	EG	FD	Average
MAUDE (2020)	0.115	0.045	0.112	0.136	0.360	0.120	0.304	0.306	0.192	-0.073	-0.11	-0.057	-0.285	0.090
GRADE (2020)	0.121	0.332	**0.612**	0.176	0.583	0.532	0.571	0.329	0.596	0.254	0.048	0.300	0.106	0.351
USR (2020b)	0.166	0.249	0.34	0.291	0.496	0.363	0.487	0.140	0.353	0.066	0.055	0.268	0.084	0.258
FED (2020a)	-0.082	-0.070	-0.077	-0.090	-0.232	-0.080	-0.137	-0.004	0.025	-0.009	0.173	0.005	0.178	-0.031
FlowScore (2021)	0.095	0.067	-0.049	0.068	0.202	-0.063	-	0.053	0.053	-	-0.043	-	-0.009	0.029
USL-H (2020)	0.180	0.261	0.53	0.319	0.409	0.385	0.452	**0.493**	0.481	0.09	0.115	0.237	0.202	0.320
QuestEval (2021)	0.089	0.222	0.217	0.104	0.32	0.22	0.344	0.106	0.243	-0.026	0.168	0.195	0.114	0.178
DEB (2020)	0.214	0.351	0.649	0.123	0.579	0.486	0.504	0.351	0.579	**0.363**	0.044	0.395	0.141	0.367
DynaEval (2021)	0.252	0.066	0.112	-0.013	0.165	0.169	0.202	0.148	0.038	0.122	0.247	0.159	**0.555**	0.171
DialogRPT (2020)	0.162	0.255	0.198	0.118	0.114	0.067	0.158	-0.036	0.075	0.037	-0.249	0.203	-0.134	0.074
Ours (DIAL-T0)	**0.553**	**0.451**	0.582	**0.446**	**0.651**	**0.601**	**0.498**	0.376	**0.634**	0.286	**0.263**	**0.475**	0.228	**0.465**

- Prakhar Gupta, et al. 2022. InstructDial: Improving Zero and Few-shot Generalization in Dialogue through Instruction Tuning. EMNLP-2022.

Representative Examples – xDial-Eval

- The current research on dialogue evaluation primarily focuses on English dialogues
  - A primary reason is the lack of a multilingual dialogue evaluation benchmark
- xDial-Eval includes 12 turn-level and 6 dialogue-level English datasets, comprising 14930 annotated turns and 8691 annotated dialogues respectively.
- The English dialogue data are extended to 9 other languages with commercial machine translation systems.
  - Chinese (ZH), Spanish (ES), German (DE), French (FR), Japanese (JA), Korean (KO), Hindi (HI), Arabic (AR), and Russian (RU)



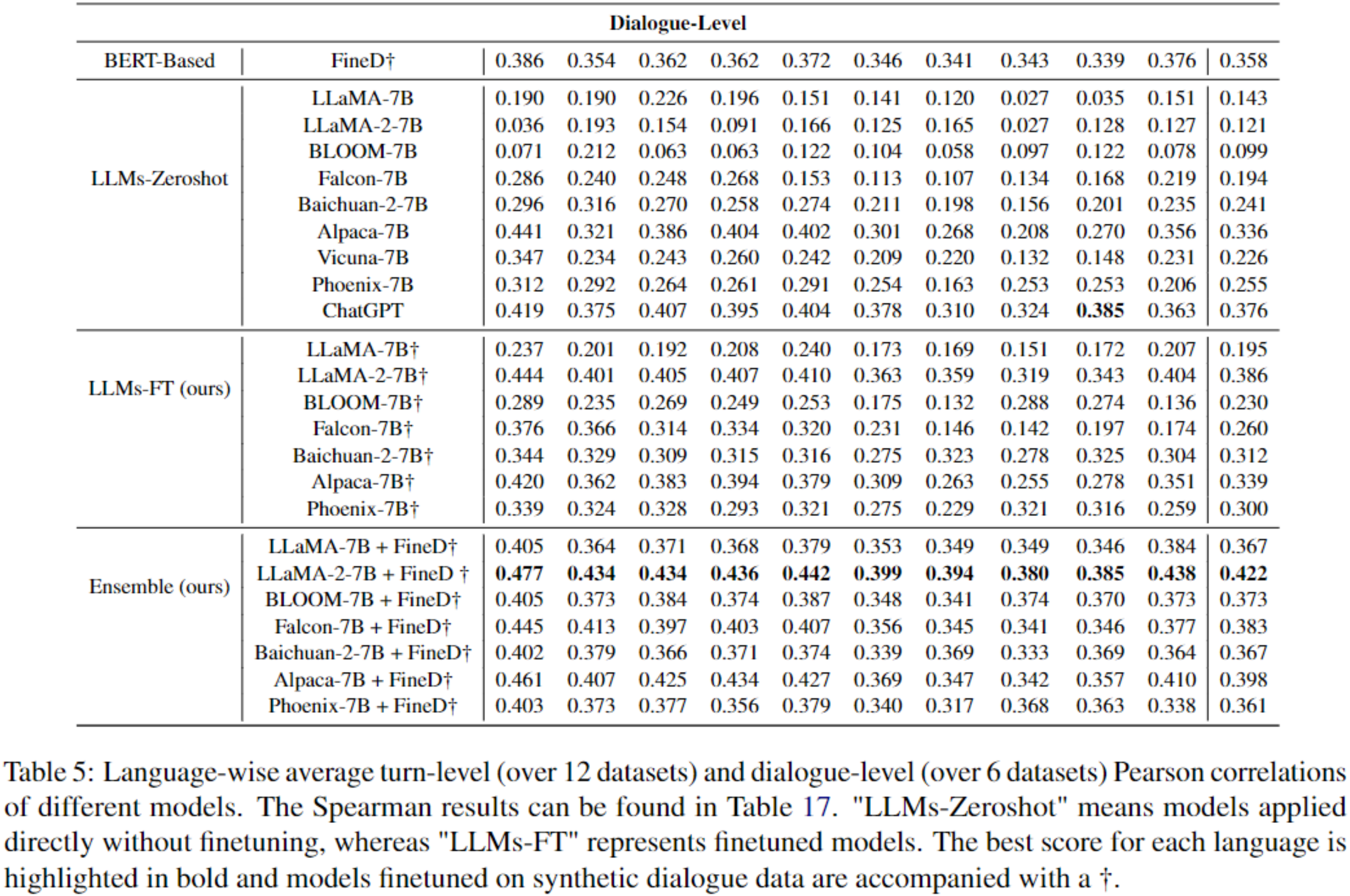
Turn-Level Datasets	#Instance	#Utts/Instance	#Ctx/Hyp Words	#Dims
Persona-USR (2020b)	300	9.3	98.4 / 12.0	6
Persona-Zhao (2020)	900	5.1	48.8 / 11.5	4
ConvAI2-GRADE (2020)	600	3.0	24.4 / 11.3	1
Persona-DSTC10 (2022b)	4,829	4.0	36.0 / 11.6	4
DailyDialog-GRADE (2020)	300	3.0	26.0 / 10.8	1
DailyDialog-Zhao (2020)	900	4.7	47.5 / 11.0	4
DailyDialog-Gupta (2019)	500	4.9	49.9 / 10.9	1
Topical-USR (2020b)	360	11.2	236.3 / 22.4	6
Topical-DSTC10 (2022b)	4,500	4.0	50.6 / 15.9	4
Empathetic-GRADE (2020)	300	3.0	29.0 / 15.6	1
FED-Turn (2020a)	375	10.4	87.3 / 13.3	9
ConTurE-Turn (2022a)	1066	3.8	21.67 / 10.99	1
Dialogue-Level Datasets	**#Instance**	**#Utts/Instance**	**#Words/Utt**	**#Dims**
IEval (2022)	1,920	6.0	12.4	8
Persona-See (2019)	3,316	12.0	7.6	9
Reliable-Eval (2022)	2,925	21.2	8.4	7
ConTurE-Dial (2022b)	119	17.9	8.6	11
FED-Dial (2020a)	125	12.7	9.2	11
Human-Eval (2022)	286	12.0	11.6	3

Zhang, C., D'Haro, L. F., Tang, C., Shi, K., Tang, G., & Li, H. (2023). xDial-Eval: A Multilingual Open-Domain Dialogue Evaluation Benchmark. *arXiv preprint arXiv:2310.08958*. Accepted to **EMNLP2023**

Representative Examples – xDial-Eval

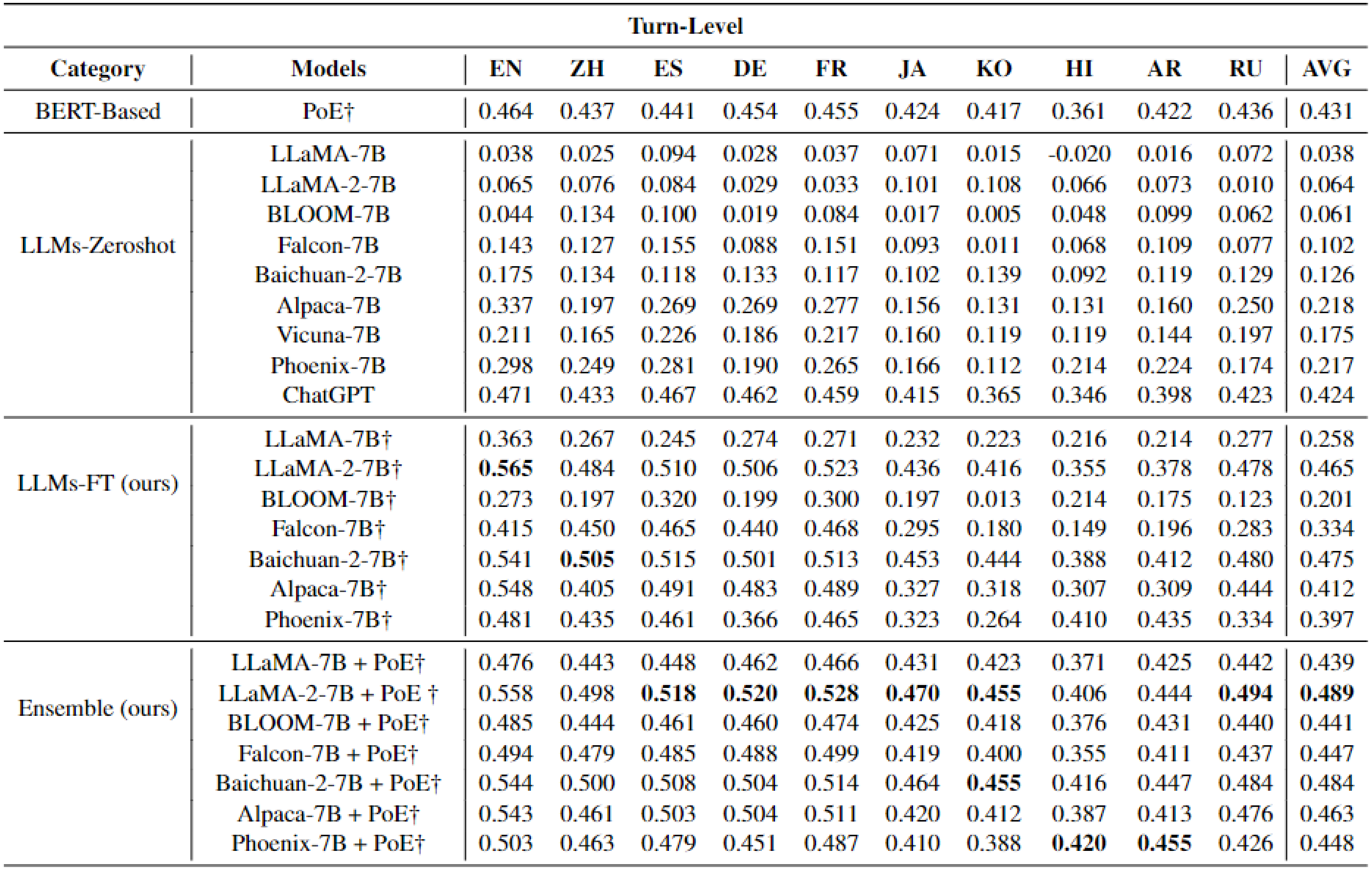
- Comprehensive analyses of previous BERT-based metrics and 9 LLMs
- Results: avg. Pearson correlations over all datasets and languages:
    - Best baseline outperforms OpenAI's ChatGPT by absolute improvements of 6.5\% and 4.6\% at the turn and dialogue levels respectively
    - Data and code available at https://github.com/e0397123/xDial-Eval

Zhang, C., D'Haro, L. F., Tang, C., Shi, K., Tang, G., & Li, H. (2023). xDial-Eval: A Multilingual Open-Domain Dialogue Evaluation Benchmark. *arXiv preprint arXiv:2310.08958*. Accepted to **EMNLP2023**



		Dialogue-Level										
BERT-Based	FineD†	0.386	0.354	0.362	0.362	0.372	0.346	0.341	0.343	0.339	0.376	0.358
LLMs-Zeroshot	LLaMA-7B	0.190	0.190	0.226	0.196	0.151	0.141	0.120	0.027	0.035	0.151	0.143
	LLaMA-2-7B	0.036	0.193	0.154	0.091	0.166	0.125	0.165	0.027	0.128	0.127	0.121
	BLOOM-7B	0.071	0.212	0.063	0.063	0.122	0.104	0.058	0.097	0.122	0.078	0.099
	Falcon-7B	0.286	0.240	0.248	0.268	0.153	0.113	0.107	0.134	0.168	0.219	0.194
	Baichuan-2-7B	0.296	0.316	0.270	0.258	0.274	0.211	0.198	0.156	0.201	0.235	0.241
	Alpaca-7B	0.441	0.321	0.386	0.404	0.402	0.301	0.268	0.208	0.270	0.356	0.336
	Vicuna-7B	0.347	0.234	0.243	0.260	0.242	0.209	0.220	0.132	0.148	0.231	0.226
	Phoenix-7B	0.312	0.292	0.264	0.261	0.291	0.254	0.163	0.253	0.253	0.206	0.255
	ChatGPT	0.419	0.375	0.407	0.395	0.404	0.378	0.310	0.324	**0.385**	0.363	0.376
LLMs-FT (ours)	LLaMA-7B†	0.237	0.201	0.192	0.208	0.240	0.173	0.169	0.151	0.172	0.207	0.195
	LLaMA-2-7B†	0.444	0.401	0.405	0.407	0.410	0.363	0.359	0.319	0.343	0.404	0.386
	BLOOM-7B†	0.289	0.235	0.269	0.249	0.253	0.175	0.132	0.288	0.274	0.136	0.230
	Falcon-7B†	0.376	0.366	0.314	0.334	0.320	0.231	0.146	0.142	0.197	0.174	0.260
	Baichuan-2-7B†	0.344	0.329	0.309	0.315	0.316	0.275	0.323	0.278	0.325	0.304	0.312
	Alpaca-7B†	0.420	0.362	0.383	0.394	0.379	0.309	0.263	0.255	0.278	0.351	0.339
	Phoenix-7B†	0.339	0.324	0.328	0.293	0.321	0.275	0.229	0.321	0.316	0.259	0.300
Ensemble (ours)	LLaMA-7B + FineD†	0.405	0.364	0.371	0.368	0.379	0.353	0.349	0.349	0.346	0.384	0.367
	LLaMA-2-7B + FineD †	**0.477**	**0.434**	**0.434**	**0.436**	**0.442**	**0.399**	**0.394**	**0.380**	**0.385**	**0.438**	**0.422**
	BLOOM-7B + FineD†	0.405	0.373	0.384	0.374	0.387	0.348	0.341	0.374	0.370	0.373	0.373
	Falcon-7B + FineD†	0.445	0.413	0.397	0.403	0.407	0.356	0.345	0.341	0.346	0.377	0.383
	Baichuan-2-7B + FineD†	0.402	0.379	0.366	0.371	0.374	0.339	0.369	0.333	0.369	0.364	0.367
	Alpaca-7B + FineD†	0.461	0.407	0.425	0.434	0.427	0.369	0.347	0.342	0.357	0.410	0.398
	Phoenix-7B + FineD†	0.403	0.373	0.377	0.356	0.379	0.340	0.317	0.368	0.363	0.338	0.361

Table 5: Language-wise average turn-level (over 12 datasets) and dialogue-level (over 6 datasets) Pearson correlations of different models. The Spearman results can be found in Table 17. "LLMs-Zeroshot" means models applied directly without finetuning, whereas "LLMs-FT" represents finetuned models. The best score for each language is highlighted in bold and models finetuned on synthetic dialogue data are accompanied with a †.



Turn-Level												
Category	**Models**	**EN**	**ZH**	**ES**	**DE**	**FR**	**JA**	**KO**	**HI**	**AR**	**RU**	**AVG**
BERT-Based	PoE†	0.464	0.437	0.441	0.454	0.455	0.424	0.417	0.361	0.422	0.436	0.431
LLMs-Zeroshot	LLaMA-7B	0.038	0.025	0.094	0.028	0.037	0.071	0.015	-0.020	0.016	0.072	0.038
	LLaMA-2-7B	0.065	0.076	0.084	0.029	0.033	0.101	0.108	0.066	0.073	0.010	0.064
	BLOOM-7B	0.044	0.134	0.100	0.019	0.084	0.017	0.005	0.048	0.099	0.062	0.061
	Falcon-7B	0.143	0.127	0.155	0.088	0.151	0.093	0.011	0.068	0.109	0.077	0.102
	Baichuan-2-7B	0.175	0.134	0.118	0.133	0.117	0.102	0.139	0.092	0.119	0.129	0.126
	Alpaca-7B	0.337	0.197	0.269	0.269	0.277	0.156	0.131	0.131	0.160	0.250	0.218
	Vicuna-7B	0.211	0.165	0.226	0.186	0.217	0.160	0.119	0.119	0.144	0.197	0.175
	Phoenix-7B	0.298	0.249	0.281	0.190	0.265	0.166	0.112	0.214	0.224	0.174	0.217
	ChatGPT	0.471	0.433	0.467	0.462	0.459	0.415	0.365	0.346	0.398	0.423	0.424
LLMs-FT (ours)	LLaMA-7B†	0.363	0.267	0.245	0.274	0.271	0.232	0.223	0.216	0.214	0.277	0.258
	LLaMA-2-7B†	**0.565**	0.484	0.510	0.506	0.523	0.436	0.416	0.355	0.378	0.478	0.465
	BLOOM-7B†	0.273	0.197	0.320	0.199	0.300	0.197	0.013	0.214	0.175	0.123	0.201
	Falcon-7B†	0.415	0.450	0.465	0.440	0.468	0.295	0.180	0.149	0.196	0.283	0.334
	Baichuan-2-7B†	0.541	**0.505**	0.515	0.501	0.513	0.453	0.444	0.388	0.412	0.480	0.475
	Alpaca-7B†	0.548	0.405	0.491	0.483	0.489	0.327	0.318	0.307	0.309	0.444	0.412
	Phoenix-7B†	0.481	0.435	0.461	0.366	0.465	0.323	0.264	0.410	0.435	0.334	0.397
Ensemble (ours)	LLaMA-7B + PoE†	0.476	0.443	0.448	0.462	0.466	0.431	0.423	0.371	0.425	0.442	0.439
	LLaMA-2-7B + PoE †	0.558	0.498	**0.518**	**0.520**	**0.528**	**0.470**	**0.455**	0.406	0.444	**0.494**	**0.489**
	BLOOM-7B + PoE†	0.485	0.444	0.461	0.460	0.474	0.425	0.418	0.376	0.431	0.440	0.441
	Falcon-7B + PoE†	0.494	0.479	0.485	0.488	0.499	0.419	0.400	0.355	0.411	0.437	0.447
	Baichuan-2-7B + PoE†	0.544	0.500	0.508	0.504	0.514	0.464	**0.455**	0.416	0.447	0.484	0.484
	Alpaca-7B + PoE†	0.543	0.461	0.503	0.504	0.511	0.420	0.412	0.387	0.413	0.476	0.463
	Phoenix-7B + PoE†	0.503	0.463	0.479	0.451	0.487	0.410	0.388	**0.420**	**0.455**	0.426	0.448

Q
&
A

Challenges & Future Directions

Remaining challenges (Mehri et al., 2022)

- Metrics are limited in scope
  - Measure a limited set of dialog qualities or dimensions, languages, cultures
- Metrics struggle to generalize
  - Lack of robustness
- Metrics are not strongly correlated with human judgment
  - Real users (e.g., Alexa Prize), inconsistencies between datasets

- Mehri, S., Choi, J., D'Haro, L. F., Deriu, J., Eskenazi, M., Gasic, M., ... & Zhang, C. (2022). Report from the NSF future directions workshop on automatic evaluation of dialog: Research directions and challenges. arXiv preprint arXiv:2203.10012.

Needs for datasets

Necessity for defining more standard annotation schemes for dialogue evaluation

- Including unified definition or terminology for some dimensions of evaluation
- Demographics of the annotators

New dimensions not fully covered yet

- E.g., toxicity, bias, coherence, hallucination, common-sense, engagement, cultural and language issues.

Fined-grained annotations:

- Dialogue level vs Turn level vs Long-term interaction

Design of progressively and more difficult benchmarks and availability of repositories

- Similar as for MT, Q&A, NLG fields
- Repository for benchmarking: https://github.com/exe1023/DialEvalMetrics (Yeh et al., 2021)
- Repository for datasets: (DSTC10-T5)(Zhang et al., 2021): https://github.com/e0397123/dstc10_metric_track

- Yeh, Y. T., Eskenazi, M., & Mehri, S. (2021). A comprehensive assessment of dialog evaluation metrics. *arXiv preprint arXiv:2106.03706*.
- Chen, Z., Sadoc, J., D'Haro, L. F., Banchs, R., & Rudnicky, A. (2021). Automatic evaluation and moderation of open-domain dialogue systems. *arXiv preprint arXiv:2111.02110*.

Desired properties for metrics

- Strong Correlation with Human Judgements
  - Improved models: beyond dialogue context & response, better data augmentation & training objectives, better training data,
- Interpretability
  - fine-grained explanations, type and severity of errors, natural language feedback
- Robustness against Adversarial Attacks
  - Avoid gaming the metric
- Generalizable across different domains and new dimensions
  - High-quality data + Meta-learning approaches, e.g., mixture of experts
- Forward- and Backward-looking
  - Possibility of considering whole dialogues and sessions, not just previous turns
- Compatible with Human Evaluation
  - Human-in-the-loop for error correction

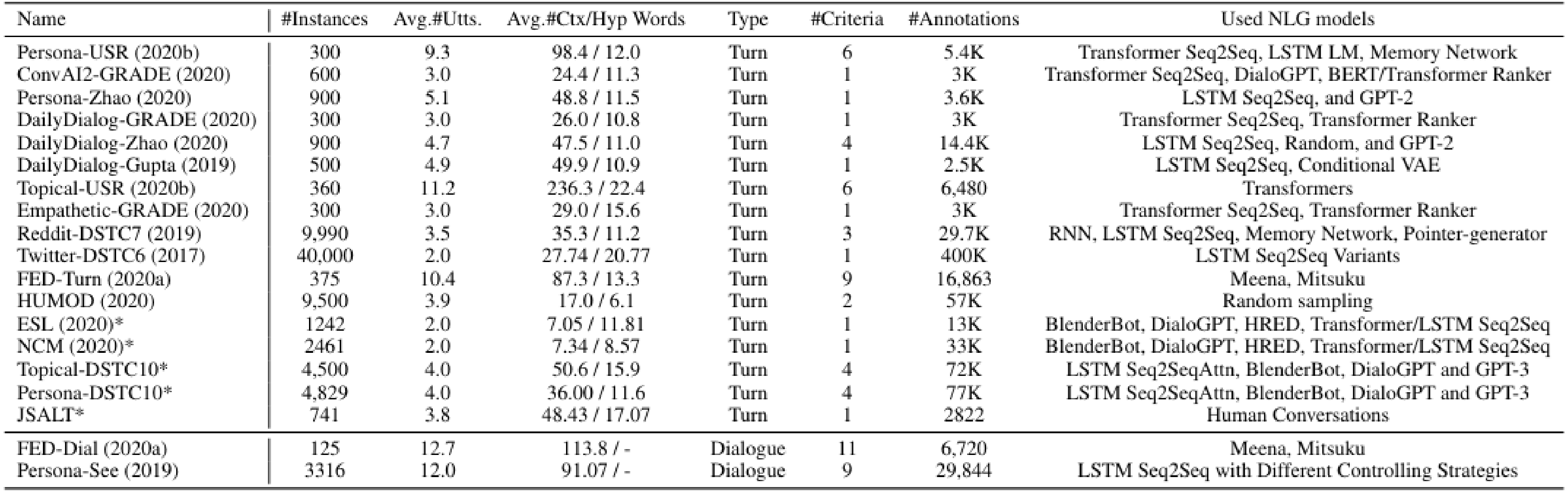
Dialogue Evaluation Datasets

- The DSTC9-interactive datasets (Mehri et al., 2022)
  - Consists of 3300 turn-level and 2200 dialogue-level annotated data
  - Annotated in the same manner as the FED dataset (Mehri et al., 2020)
  - Turn-level data: http://dialog.speech.cs.cmu.edu:9993/static_data.json
  - Dialogue-level data: http://dialog.speech.cs.cmu.edu:9993/interactive_data.json
  - The dialogues are significantly longer than those in FED
  - More advanced dialogue systems are incorporated, such as PLATO-2 (Bao et al., 2022), DialoGPT (Zhang et al., 2020), etc.

- Mehri et al. "Interactive Evaluation of Dialog Track at DSTC9." LREC (2022).
- Mehri and Eskenazi. "Unsupervised Evaluation of Interactive Dialog with DialoGPT." SIGDial (2020).
- Bao et al. "PLATO-2: Towards Building an Open-Domain Chatbot via Curriculum Learning." Findings of ACL-IJCNLP (2021).
- Zhang et al. "DIALOGPT: Large-Scale Generative Pre-training for Conversational Response Generation." ACL System Demonstrations (2020).

Dialogue Evaluation Datasets

- The DSTC10 Benchmark (Zhang et al., 2021)
  - Available at https://chateval.org/dstc10



Name	#Instances	Avg.#Utts.	Avg.#Ctx/Hyp Words	Type	#Criteria	#Annotations	Used NLG models
Persona-USR (2020b)	300	9.3	98.4 / 12.0	Turn	6	5.4K	Transformer Seq2Seq, LSTM LM, Memory Network
ConvAI2-GRADE (2020)	600	3.0	24.4 / 11.3	Turn	1	3K	Transformer Seq2Seq, DialoGPT, BERT/Transformer Ranker
Persona-Zhao (2020)	900	5.1	48.8 / 11.5	Turn	1	3.6K	LSTM Seq2Seq, and GPT-2
DailyDialog-GRADE (2020)	300	3.0	26.0 / 10.8	Turn	1	3K	Transformer Seq2Seq, Transformer Ranker
DailyDialog-Zhao (2020)	900	4.7	47.5 / 11.0	Turn	4	14.4K	LSTM Seq2Seq, Random, and GPT-2
DailyDialog-Gupta (2019)	500	4.9	49.9 / 10.9	Turn	1	2.5K	LSTM Seq2Seq, Conditional VAE
Topical-USR (2020b)	360	11.2	236.3 / 22.4	Turn	6	6,480	Transformers
Empathetic-GRADE (2020)	300	3.0	29.0 / 15.6	Turn	1	3K	Transformer Seq2Seq, Transformer Ranker
Reddit-DSTC7 (2019)	9,990	3.5	35.3 / 11.2	Turn	3	29.7K	RNN, LSTM Seq2Seq, Memory Network, Pointer-generator
Twitter-DSTC6 (2017)	40,000	2.0	27.74 / 20.77	Turn	1	400K	LSTM Seq2Seq Variants
FED-Turn (2020a)	375	10.4	87.3 / 13.3	Turn	9	16,863	Meena, Mitsuku
HUMOD (2020)	9,500	3.9	17.0 / 6.1	Turn	2	57K	Random sampling
ESL (2020)*	1242	2.0	7.05 / 11.81	Turn	1	13K	BlenderBot, DialoGPT, HRED, Transformer/LSTM Seq2Seq
NCM (2020)*	2461	2.0	7.34 / 8.57	Turn	1	33K	BlenderBot, DialoGPT, HRED, Transformer/LSTM Seq2Seq
Topical-DSTC10*	4,500	4.0	50.6 / 15.9	Turn	4	72K	LSTM Seq2SeqAttn, BlenderBot, DialoGPT and GPT-3
Persona-DSTC10*	4,829	4.0	36.00 / 11.6	Turn	4	77K	LSTM Seq2SeqAttn, BlenderBot, DialoGPT and GPT-3
JSALT*	741	3.8	48.43 / 17.07	Turn	1	2822	Human Conversations
FED-Dial (2020a)	125	12.7	113.8 / -	Dialogue	11	6,720	Meena, Mitsuku
Persona-See (2019)	3316	12.0	91.07 / -	Dialogue	9	29,844	LSTM Seq2Seq with Different Controlling Strategies

- Ni et al. "Recent advances in deep learning based dialogue systems: A systematic survey." Artificial Intelligence Review: 1-101 (2022).

DSTC11-Track 4

Benchmarks and challenges are required for progress in the field. This current challenge targets two main tasks:

- Task 1: Propose and develop effective Automatic Metrics for evaluation of open-domain multilingual dialogs.
- Task 2: Propose and develop Robust Metrics for dialogue systems trained with back translated and paraphrased dialogs in English.

Datasets:

- For training: Up to 18 Human-Human curated multilingual datasets (+3M turns), with turn/dialogue level automatic annotations including QE metrics or toxicity.
- Dev/Test: Up to 10 Human-Chatbot curated multilingual datasets (+150k turns), with turn/dialogue level human annotations.

Links for registration and baselines:

- ChatEval: https://chateval.org/dstc11
- GitHub: https://github.com/Mario-RC/dstc11_track4_robust_multilingual_metrics

Schedule:

- Training/Validation data release: From November to December in 2022
- Test data release: Middle of March in 2023
- Entry submission deadline: Middle of March in 2023
- Submission of final results: End of March in 2023
- Final result announcement: Early of April in 2023
- Paper submission: From March to May in 2023
- Workshop: July-September/2023 in a venue to be announced with DSTC11

Conclusions

Important points

- Automatic evaluation is needed to improve performance of dialogue systems
  - Reduce costs and speed up deployment
- Different types of metrics:
  - Turn-level and Dialogue-level
  - Reference-based or Reference-free, trained or untrained
  - Currently mostly measuring syntactic, semantic and coherence dimensions
- Improvements required in:
  - Evaluation of new dimensions, robustness, languages, explainability
  - New annotated datasets and schemas, benchmarks and languages

Useful Resources

- Repositories
  - https://github.com/ricsinaruto/dialog-eval
  - https://github.com/Maluuba/nlg-eval
  - https://github.com/exe1023/DialEvalMetrics
  - https://github.com/e0397123/dstc10_metric_track
- Overview Papers
  - A Comprehensive Assessment of Dialog Evaluation Metrics (Yeh et al., 2021)
  - Report from the NSF future directions workshop on automatic evaluation of dialog: Research directions and challenges (Mehri et al., 2022)
  - A Survey of Evaluation Metrics Used for NLG Systems (Sai et al., 2022)
  - Survey on evaluation methods for dialogue systems (Deriu et al., 2021)
  - Towards Unified Dialogue System Evaluation: A Comprehensive Analysis of Current Evaluation Protocols (Finch & Choi, 2020)
  - Human Evaluation of Conversations is an Open Problem: comparing the sensitivity of various methods for evaluating dialogue agents (Smith et al., 2022)
  - Achieving Reliable Human Assessment of Open-Domain Dialogue Systems (Ji et al., 2022)
  - Investigating the Impact of Pre-trained Language Models on Dialog Evaluation (Zhang et al., 2021)

Relevant bibliography for follow up

Current State-of-the-Art approaches:

- Dialogue-Level
  - Zhang, C., D'Haro, L. F., Zhang, Q., Friedrichs, T., & Li, H. (2022). FineD-Eval: Fine-grained Automatic Dialogue-Level Evaluation. *arXiv preprint arXiv:2210.13832*.
- Turn-level
  - Zhang, C., D'Haro, L. F., Friedrichs, T., & Li, H. (2022, June). MDD-Eval: self-training on augmented data for multi-domain dialogue evaluation. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 36, No. 10, pp. 11657-11666).

Complete survey on evaluating LLMs and hallucinations:

- Guo, Z., Jin, R., Liu, C., Huang, Y., Shi, D., Yu, L., ... & Xiong, D. (2023). Evaluating Large Language Models: A Comprehensive Survey. arXiv preprint arXiv:2310.19736.

Hands-on

Goals

- Understand the process of performing human annotations
    - Instructions for annotating different dimensions at turn & dialogue levels
    - Tools for analysis: correlations
    - Randomly initialized models
- Perform automatic evaluation using SotA models available on HuggingFace
    - Single-T, PoE: Turn-level
    - FineD-Eval: Dialogue-Level
- Evaluation using LLM
    - Prompt-based techniques
    - Two different OS models: Phi-1.5B and Vicuna-7B

Resources

- Colab notebook: https://short.upm.es/fb50t
  - Load Google drive (requires a Google account)
  - Select T4-GPU (for LLMs)

- Annotation files: https://short.upm.es/d280r