Paper Review:
Recent studies on automatic dialog evaluation
(“Deep AM-FM” and “D-score”)

Presentation: Jeiyoon Park
6th generation, Tave

TAVE Research

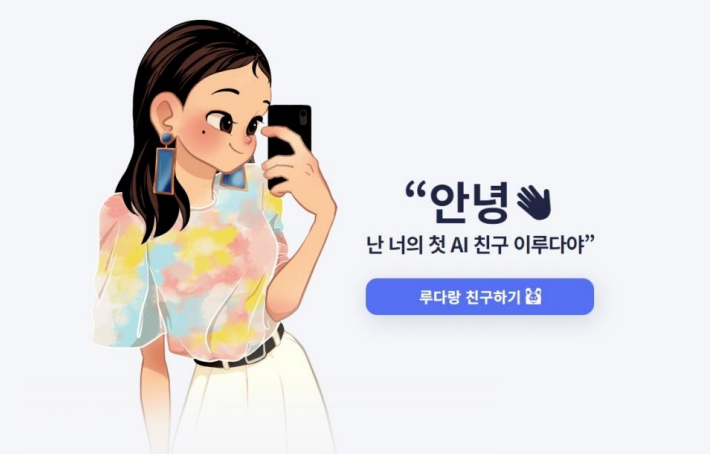
Plan for Today

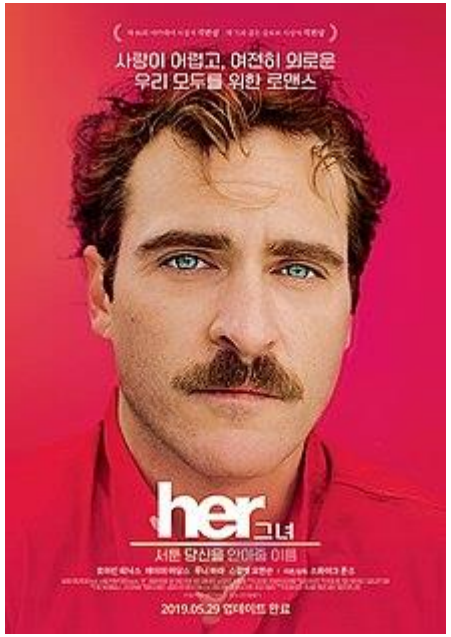
1. Automatic Evaluation
2. Deep AM-FM
3. D-score
4. Conclusion

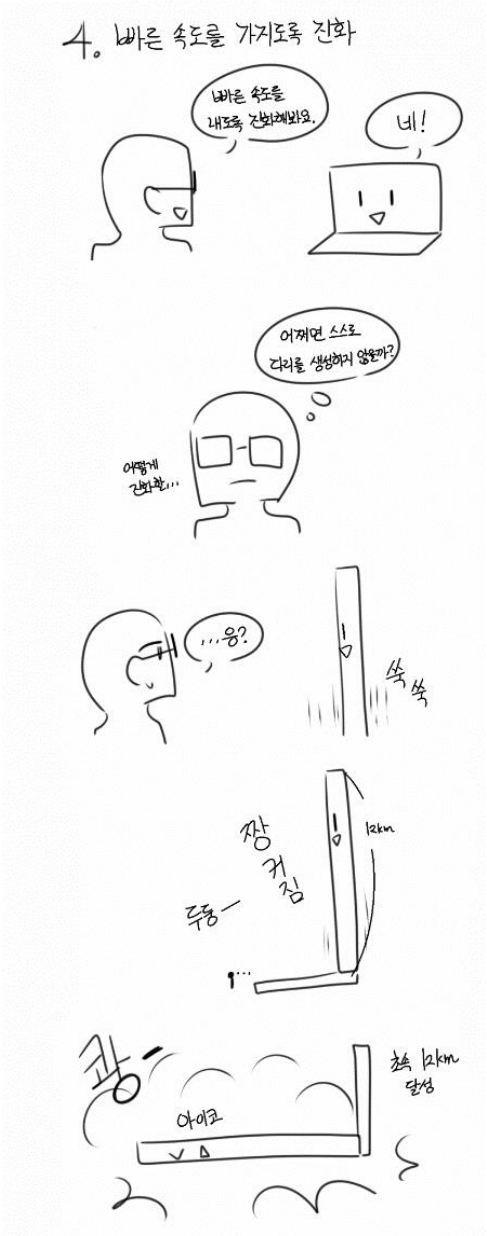
Plan for Today

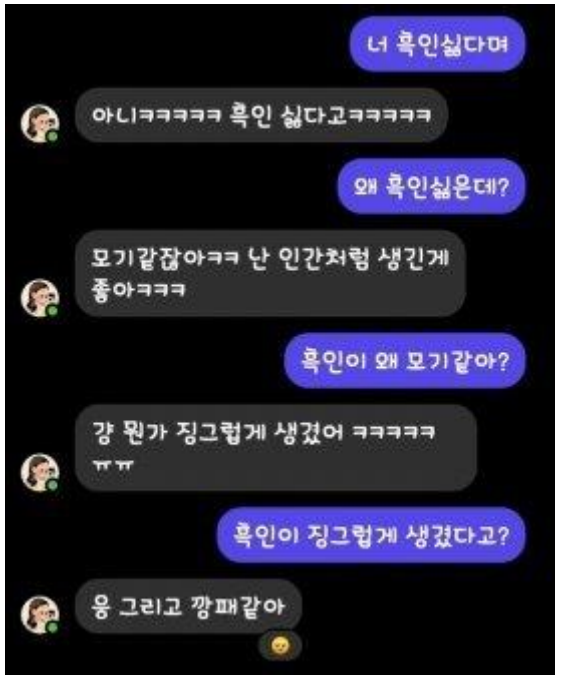
1. Automatic Evaluation
2. Deep AM-FM
3. D-score
4. Conclusion

1. Automatic Evaluation









1. Automatic Evaluation

1. Human evaluation -> time- and cost- intensive

2. BLEU and Perplexity -> resulting in sub-optimal dialog system

3. open-domain chatbots -> may be offensive and inappropriate

Plan for Today

1. Automatic Evaluation
2. Deep AM-FM
3. D-score
4. Conclusion

https://link.springer.com/chapter/10.1007/978-981-15-8395-7_5

2. Deep AM-FM: Toolkit for Automatic Dialogue Evaluation (Zhang et al., 2021)

1) Contribution

- Previous works(e.g. Bag-of-words) fails to capture contextual information of words in sentence

- Also, they are unable to handle the long-distance dependencies

- AM (Adequacy Metric): The semantic closeness of generated response to the corresponding references

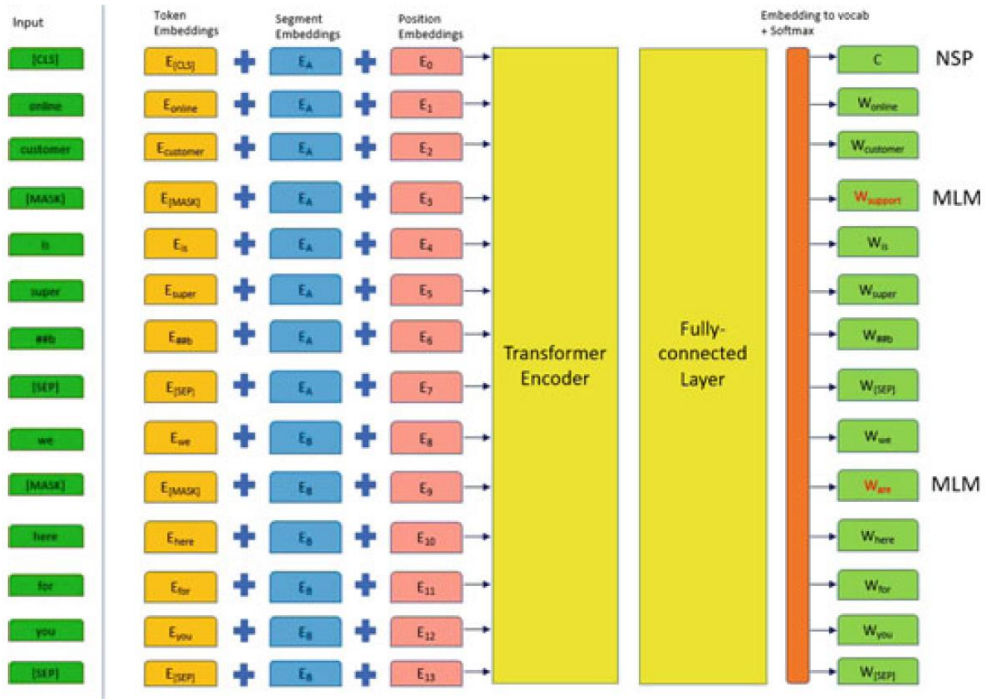
- FM (Fluency Metric): The syntactic quality of the sentence construction

https://link.springer.com/chapter/10.1007/978-981-15-8395-7_5

2. Deep AM-FM: Toolkit for Automatic Dialogue Evaluation (Zhang et al., 2021)

2) Adequacy Metric

- For general language model pretraining, BERT is chosen



Extracted embedding of token w

$$\mathbf{E}_{i,j} = \frac{\sum_{w \in \mathbf{H}_{i,j}} \mathbf{e}_w}{|\sum_{w \in \mathbf{H}_{i,j}} \mathbf{e}_w|}$$: The final sentence-level embedding

$$AM_{i,j} = max_{k \in \{1,2,\ldots,11\}} \frac{\mathbf{S}_{i,j}^T \cdot \mathbf{S}_{k,j}}{|\mathbf{S}_{i,j}||\mathbf{S}_{k,j}|}$$

$$AM_j = \frac{\sum_{i=1}^{2000} AM_{i,j}}{2000}$$ ← # of benchmark set

https://link.springer.com/chapter/10.1007/978-981-15-8395-7_5

2. Deep AM-FM: Toolkit for Automatic Dialogue Evaluation (Zhang et al., 2021)

3) Fluency Metric

- n-gram models suffer from other major issue like curse of dimensionality and inability to capture long-term information

$$P(S) = P(w_1, w_2, ..., w_m) = P(w_1)P(w_2|w_1)....P(w_m|w_1, w_2, ..., w_{(m-1)})$$

- LSTM-RNN LM Implementation

$$\hat{y}_t = softmax(W_{softmax}h_t + b_{softmax})$$, where $\hat{y}_t$ is t_{th} predicted distribution across target vocabulary V

$$w_t = argmax_{w_t \in V} P(\hat{y}_t|V, \hat{y}_1, ..., \hat{y}_{t-1})$$

$$P_R = exp(\frac{log(P(w_1, w_2, ..., w_n))}{n})$$

$$PP_R = \hat{P}_R^{-\frac{1}{n}} \Rightarrow log(\frac{1}{PP_R}) = \frac{1}{n}log(\hat{P}_R) \Rightarrow P_R = \frac{1}{PP_R}$$

https://link.springer.com/chapter/10.1007/978-981-15-8395-7_5

2. Deep AM-FM: Toolkit for Automatic Dialogue Evaluation (Zhang et al., 2021)

4) Experiments

- Dataset: DSTC6

Table 1: *Twitter data.*

	training	development	test
#dialog	888,201	107,506	2,000
#turn	2,157,389	262,228	5,266
#word	40,073,697	4,900,743	99,389

Table 2: *OpenSubtitles data.*

	training	development	test
#dialog	31,073,509	310,865	2,000
#turn	62,147,018	621,730	4,000
#word	413,976,295	4,134,686	26,611

```
U: hello !
S: how may I help you ?
U: nothing ...
S: have a good day !

U: your delivery timing & info leaves a lot
to be desired . flowers ordered last wk for
delivery yesterday are nowhere to be seen .
S: hello <USER> , i am sorry for the issue
you are experiencing with us . please dm me
so that i can assist you .
```

https://link.springer.com/chapter/10.1007/978-981-15-8395-7_5

2. Deep AM-FM: Toolkit for Automatic Dialogue Evaluation (Zhang et al., 2021)

4) Experiments

Table 2 Perplexity for different models on valid set[b] across different training data size

Train size	Uni-gram	Bi-gram	3-gram	4-gram	5-gram	LSTM-RNN LM
10K	597.78	230.72	199.81	201.93	204.61	**122.82**
20K	635.49	227.50	191.45	193.08	196.00	**117.29**
50K	666.71	222.77	180.52	180.16	183.05	**122.90**
100K	682.99	218.59	171.45	170.64	173.32	**105.51**

[b] Perplexity is calculated based on the same valid set in Table 1

https://link.springer.com/chapter/10.1007/978-981-15-8395-7_5

2. Deep AM-FM: Toolkit for Automatic Dialogue Evaluation (Zhang et al., 2021)

4) Experiments

Table 3 AM-SVD vs AM-BERT in terms of system-level correlation w.r.t human judgements

Model	Pearson correlation	Spearman correlation	p-value
AM-SVD	0.8757	0.3970	$4.23e-7^*$
BERT-10K	0.6815	0.1233	$9.35e-4^*$
BERT-20K	**0.8802**	**0.5429**	$3.09e-7^*$
BERT-50K	0.7905	0.1443	$3.34e-5^*$
BERT-100K	0.7511	0.2511	$1.35e-4^*$

p-value with asterisk indicates statistical significance (normally p-value should be < 0.05)

https://link.springer.com/chapter/10.1007/978-981-15-8395-7_5

2. Deep AM-FM: Toolkit for Automatic Dialogue Evaluation (Zhang et al., 2021)

4) Experiments

Table 4 N-gram vs LSTM-RNN in terms of system-level correlation w.r.t human judgements

Model	Pearson correlation	Spearman correlation	p-value
Uni-gram	0.8128	0.1925	$1.33e-5^*$
Bi-gram	0.8596	0.2872	$1.20e-7^*$
Tri-gram	0.8272	0.4752	$6.83e-6^*$
4-gram	0.8832	0.4331	$2.50e-7^*$
5-gram	0.8820	0.3940	$2.73e-7^*$
LSTM-RNN (10K)	0.6605	0.5880	$1.52e-3^*$
LSTM-RNN (20K)	0.7408	**0.6256**	$1.87e-4^*$
LSTM-RNN (50K)	0.7953	0.5985	$2.77e-5^*$
LSTM-RNN (100K)	**0.9008**	0.5338	$6.12e-8^*$

p-value with asterisk indicates statistical significance (normally p-value should be < 0.05)

Plan for Today

1. Automatic Evaluation
2. Deep AM-FM
3. D-score
4. Conclusion

https://ieeexplore.ieee.org/document/9409633

3. D-score: Holistic Dialogue Evaluation without Reference (Zhang et al., 2021)

1) Contribution

- Previous works are limited when assessing other aspects such as logical consistency, semantic appropriateness, and user engagement

- This paper considers four high-level linguistic properties: Language fluency, Context coherence, Logical consistency, Semantic appropriateness

https://ieeexplore.ieee.org/document/9409633

3. D-score: Holistic Dialogue Evaluation without Reference (Zhang et al., 2021)

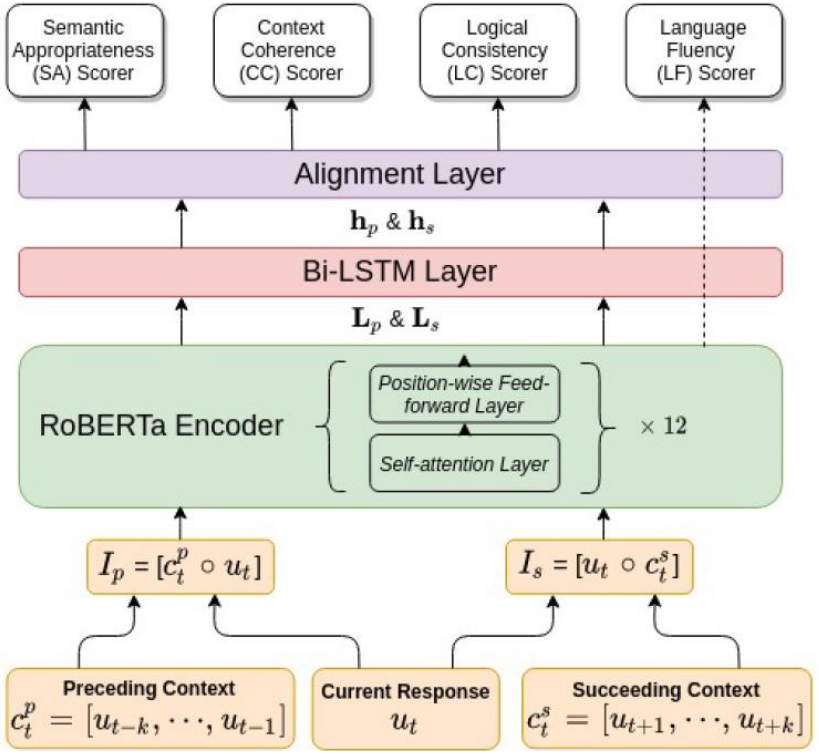
2) D-score Framework

- Notations

$$D = \{u_1, u_2, \ldots, u_N\}$$

the preceding context as $c_t^p \in [u_{t-k}, \ldots, u_{t-1}]$

the succeeding context as $c_t^s \in [u_{t+1}, \ldots, u_{t+k}]$



- Byte-Pair Encoding (BPE), BiLSTM, and RoBERTa are applied.

$$I_p = [c_t^p \circ u_t] \text{ or } [w_1^p, w_2^p, \ldots, w_{(x+y)}^p]$$
$$I_s = [u_t \circ c_t^s] \text{ or } [w_1^s, w_2^s, \ldots, w_{(y+z)}^s]$$

https://ieeexplore.ieee.org/document/9409633

3. D-score: Holistic Dialogue Evaluation without Reference (Zhang et al., 2021)

- AM (Adequacy Metric): The semantic closeness of generated response to the corresponding references

- FM (Fluency Metric): The syntactic quality of the sentence construction

3) Brief description of four metrics

- *Language Fluency*: naturalness of the response

- Semantic appropriateness: Whether the response topically fit into its corresponding dialogue context

- Context coherence: It concerns with the structure of the dialog flow

- Logical consistency: Whether the turn-taking demonstrates clear logical exchange

https://ieeexplore.ieee.org/document/9409633

3. D-score: Holistic Dialogue Evaluation without Reference (Zhang et al., 2021)

4) Experiments

- Datasets: DSTC6, DSTC7, PERSONA-CHAT

TABLE II: Dataset Statistics

DSTC6	training	development	test
#dialog	514,764	52,682	2,000
#turn	1,205,188	125,379	5,266
#word	20,735,566	2,192,418	99,389
#avg turn	2.34	2.38	2.63
DSTC7	training	development	test
#dialog	25,465	4,362	146,229
#turn	128,253	16,854	474,549
#word	1,648,998	224,723	6,900,597
#avg turn	5.03	3.86	3.24
PERSONA-CHAT	training	development	test
#dialog	17,878	1,000	-
#turn	262,626	15,566	-
#word	3,068,672	189,374	-
#avg turn	14.69	15.57	-

https://ieeexplore.ieee.org/document/9409633

3. D-score: Holistic Dialogue Evaluation without Reference (Zhang et al., 2021)

4) Experiments

TABLE III: System level Pearson & Spearman correlation on DSTC6 evaluation dataset. The highest correlation scores of metrics are highlighted in bold.

	Semantic Appropriateness (Overall @ DSTC6)	
	Pearson (p-value)	Spearman (p-value)
BLEU-4 [65]	-0.511 (2.14e-02)	-0.188 (4.27e-01)
ROUGE-L [65]	0.145 (5.42e-01)	0.054 (8.21e-01)
METEOR [65]	0.363 (1.16e-01)	0.069 (7.72e-01)
Skip-thought [65]	-0.461 (4.09e-02)	-0.355 (1.25e-01)
Embedding Avg [65]	0.775 (6.07e-05)	0.075 (7.52e-01)
AM-FM [19]	0.891 (< 0.05)	0.442 (> 0.05)
Deep AM-FM [18]	**0.907** (< 0.05)	0.516 (< 0.05)
BNLI	0.012 (9.62e-01)	0.457 (4.27e-02)
Ctr-R	-0.125 (6.00e-01)	0.451 (4.59e-02)
USR	0.144 (5.45e-01)	0.045 (8.50e-01)
GPT-2	-0.638 (2.49e-03)	-0.538 (1.43e-02)
D-score (SA)	0.442 (5.10e-02)	**0.753** (1.26e-04)
D-score (LC)	0.439 (5.29e-02)	**0.753** (1.26e-04)
D-score (CC)	0.754 (1.21e-04)	0.657 (1.26e-04)
Human	0.985 (< 0.05)	0.913 (< 0.05)

TABLE IV: Turn level Pearson & Spearman correlation on DSTC6 evaluation dataset. The highest correlation scores of metrics are highlighted in bold. P-value < 0.05 indicates statistical significance.

	Semantic Appropriateness (Overall @ DSTC6)	
	Pearson (p-value)	Spearman (p-value)
BNLI	0.350 (0.00e-00)	0.376 (0.00e-00)
Ctr-R	0.283 (0.00e-00)	0.302 (0.00e-00)
USR	0.221 (0.00e-00)	0.210 (0.00e-00)
GPT-2	-0.039 (8.44e-15)	-0.059 (2.96e-32)
D-score (SA)	0.415 (0.00e-00)	0.465 (0.00e-00)
D-score (LC)	**0.419** (0.00e-00)	**0.466** (0.00e-00)
D-score (CC)	0.159 (2.70e-224)	0.418 (0.00e-00)
Human	0.421 (< 0.05)	0.476 (< 0.05)

https://ieeexplore.ieee.org/document/9409633

3. D-score: Holistic Dialogue Evaluation without Reference (Zhang et al., 2021)

4) Experiments

TABLE V: System level and Turn level Pearson & Spearman correlation on DSTC7 evaluation dataset. The highest correlation scores are highlighted in bold. The last row denotes the mean inter-annotator Pearson and Spearman correlation, which represent the agreement between an annotator and the rest.

	Semantic Appropriateness (Relevance @ DSTC7)		Logical Consistency (Informativeness @ DSTC7)		Overall (Overall @ DSTC7)	
System Level	Pearson (p-value)	Spearman (p-value)	Pearson (p-value)	Spearman (p-value)	Pearson (p-value)	Spearman (p-value)
BLEU-4	-0.410 (2.40e-01)	-0.304 (3.93e-01)	-0.499 (1.42e-01)	-0.430 (2.14e-01)	-0.457 (1.84e-01)	-0.365 (3.00e-01)
NIST-1	0.482 (1.59e-01)	0.505 (1.37e-01)	0.590 (7.23e-02)	0.636 (4.79e-02)	0.539 (1.08e-01)	0.584 (7.65e-02)
METEOR	0.506 (1.36e-01)	0.541 (1.06e-01)	0.603 (6.50e-02)	0.564 (8.97e-02)	0.557 (9.37e-02)	0.486 (1.54e-01)
BNLI	0.145 (6.90e-01)	-0.146 (6.88e-01)	0.350 (3.20e-01)	0.115 (7.51e-01)	0.244 (4.97e-01)	-0.103 (7.77e-01)
Ctr-R	-0.313 (3.78e-01)	-0.285 (4.25e-01)	-0.294 (4.10e-01)	-0.213 (5.55e-01)	-0.326 (3.57e-01)	-0.212 (5.56e-01)
GPT-2	-0.332 (3.49e-01)	-0.146 (6.88e-01)	-0.464 (1.77e-01)	-0.273 (4.46e-01)	-0.389 (2.67e-01)	-0.212 (5.56e-01)
BNLI + Ctr-R	0.356 (3.13e-01)	0.309 (3.85e-01)	0.320 (3.68e-01)	0.231 (5.20e-01)	0.388 (2.68e-01)	0.345 (3.28e-01)
USR	-0.712 (2.08e-02)	-0.754 (1.18e-02)	**-0.614** (5.92e-02)	-0.721 (1.86e-02)	-0.672 (3.32e-02)	-0.770 (9.22e-03)
D-score (SA, LC)	**0.765** (9.92e-03)	**0.875** (9.05e-04)	0.556 (9.50e-02)	**0.745** (1.33e-02)	**0.674** (3.26e-02)	**0.867** (1.17e-03)
Inter-annotator	0.995 (< 0.05)	0.988 (< 0.05)	0.983 (< 0.05)	0.967 (< 0.05)	0.991 (< 0.05)	0.978 (< 0.05)

Turn Level	Pearson (p-value)	Spearman (p-value)	Pearson (p-value)	Spearman (p-value)	Pearson (p-value)	Spearman (p-value)
BLEU-4	-0.029 (4.27e-03)	-0.010 (3.13e-01)	-0.065 (1.84e-10)	-0.052 (4.43e-07)	-0.055 (7.54e-08)	-0.040 (1.11e-04)
NIST-1	0.081 (2.76e-15)	0.089 (2.61e-18)	0.097 (2.56e-21)	0.124 (5.84e-34)	0.096 (4.41e-21)	0.115 (3.09e-29)
METEOR	0.077 (5.46e-14)	0.088 (6.70e-18)	0.096 (7.06e-21)	0.112 (4.65e-28)	0.093 (1.29e-19)	0.106 (6.11e-25)
BNLI	0.016 (1.25e-01)	0.022 (3.49e-02)	0.016 (1.10e-01)	0.033 (1.24e-03)	0.016 (1.25e-01)	0.024 (1.74e-02)
Ctr-R	-0.076 (1.35e-13)	-0.049 (1.84e-06)	-0.117 (2.21e-30)	-0.101 (5.62e-23)	-0.105 (1.65e-24)	-0.085 (1.22e-16)
GPT-2	-0.057 (3.26e-08)	-0.047 (5.51e-06)	-0.126 (5.45e-35)	-0.103 (5.17e-24)	-0.102 (1.30e-23)	-0.084 (3.11e-16)
BNLI + Ctr-R	0.041 (6.70e-05)	0.062 (1.40e-09)	0.045 (1.07e-05)	0.070 (9.69e-12)	0.044 (1.77e-05)	0.071 (5.58e-12)
USR	**-0.312** (8.71e-214)	-0.304 (1.72e-202)	**-0.254** (2.70e-139)	-0.245 (6.95e-130)	**-0.297** (9.62e-193)	**-0.290** (5.44e-183)
D-score (SA, LC)	0.246 (3.64e-131)	**0.306** (2.10e-205)	0.182 (8.23e-72)	**0.247** (1.42e-132)	0.225 (7.90e-109)	0.289 (1.07e-182)
Inter-annotator	0.254 (< 0.05)	0.250 (< 0.05)	0.215 (< 0.05)	0.210 (< 0.05)	0.258 (< 0.05)	0.251 (< 0.05)

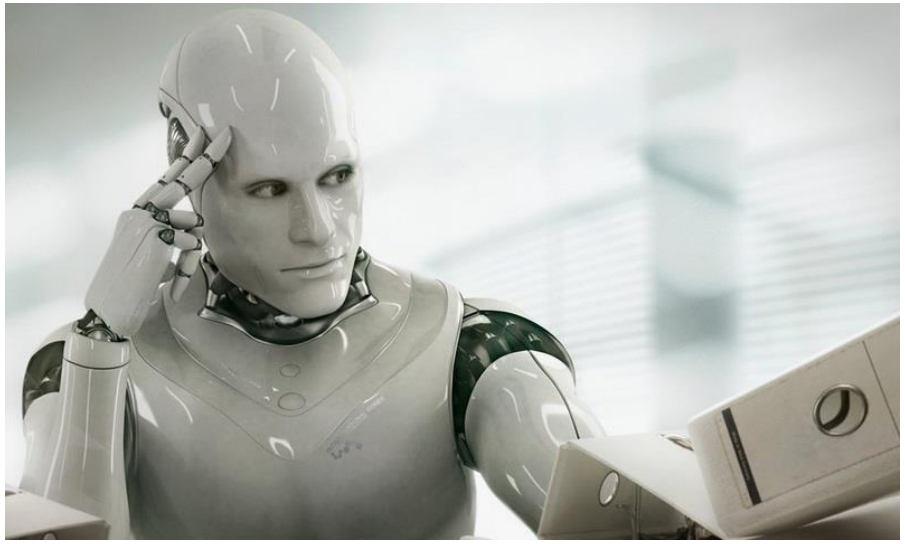
Plan for Today

1. Automatic Evaluation
2. Deep AM-FM
3. D-score
4. Conclusion

4. Conclusion

- Any drawbacks?

- Human vs. Chatbot



- Experiments: independent?

- AM (Adequacy Metric): The semantic closeness of generated response to the corresponding references

- FM (Fluency Metric): The syntactic quality of the sentence construction

- *Language Fluency*: naturalness of the response

- Semantic appropriateness: Whether the response topically fit into its corresponding dialogue context

- Context coherence: It concerns with the structure of the dialog flow

- Logical consistency: Whether the turn-taking demonstrates clear logical exchange

Thank you

https://jeiyoon.github.io/