Applied Deep Learning

NLG Evaluation

September 25th, 2024 **http://adl.miulab.tw**

National Taiwan University
國立臺灣大學

2

Automatic Evaluation Metrics

- Word overlap metrics: BLEU, ROUGE, METEOR, etc.
  - Not ideal for machine translation
  - Much worse for summarization
  - Even worse for dialogue, storytelling

more open-ended

- Embedding metrics
  - Computing the similarity of word embeddings
  - Capturing semantics in a flexible way

Evaluating the outputted results instead of the generative model

3 BLEU

- N-Gram Precision

$$p_n = \frac{\sum_{ngram \in hyp} count_{clip}(ngram)}{\sum_{ngram \in hyp} count(ngram)}$$

highest count of n-gram in any reference sentence

- Brevity Penalty

$$B = \begin{cases} e^{(1-|ref|/|hyp|)}, \text{if } |ref| > |hyp| \\ 1 \quad , \text{otherwise} \end{cases}$$

- BLEU
  - Often used in machine translation

$$BLEU = \mathrm{B} \cdot exp\left[\frac{1}{\mathrm{N}} \sum_{n=1}^{N} p_n\right]$$

ROUGE

- ROUGE (Recall-Oriented Understudy for Gisting Evaluation)
  - Often used in summarization tasks

$$\text{ROUGE-N} = \frac{\sum_{S \in \{ReferemceSummaries\}} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in \{ReferenceSummaries\}} \sum_{gram_n \in S} Count(gram_n)}$$

BLEU & ROUGE

- BLEU
  - Based on <u>n-gram overlap</u>
  - Consider precision
  - Reported as a single number
    - Combination of n = 1, 2, 3, 4 n-grams
- ROUGE
  - Based on <u>n-gram overlap</u>
  - Consider recall
  - Reported separately for each n-gram
    - ROUGE-1: unigram overlap
    - ROUGE-2: bigram overlap
    - ROUGE-L: LCS overlap

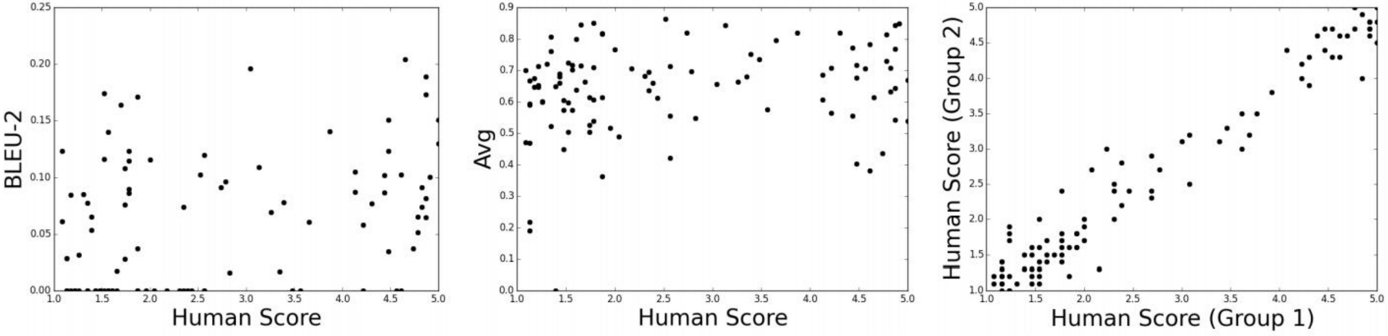
Automatic Evaluation Metrics

- Word overlap metrics: BLEU, ROUGE, METEOR, etc.
  - Not ideal for machine translation
  - Much worse for summarization
  - Even worse for dialogue, storytelling

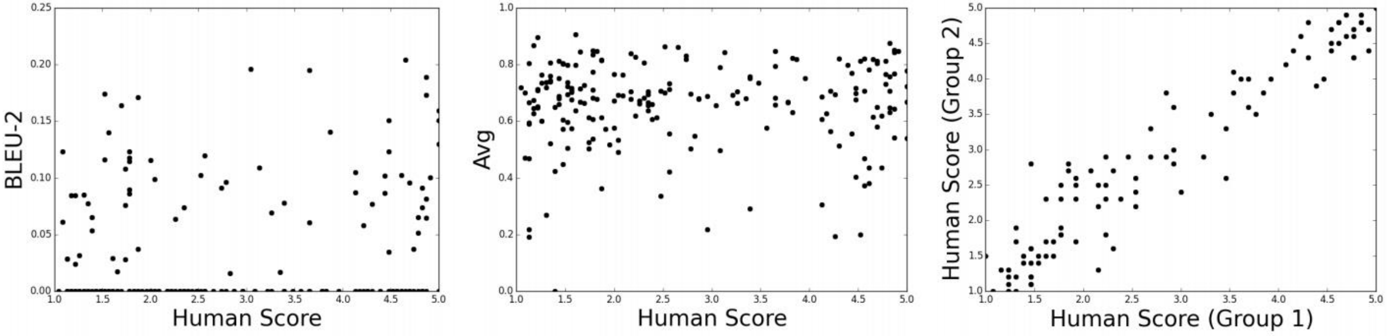
more open-ended

- Embedding metrics
  - Computing the similarity of word embeddings
  - Capturing semantics in a flexible way

Automatic Metrics vs. Human Judgement



(a) Twitter



(b) Ubuntu

No agreement between automatic scores and human scores in dialogue quality

Focused Metrics for Particular Aspects

- Evaluating a single aspect instead of the overall quality
  - Fluency (compute probability w.r.t. well-trained LM)
  - Correct style (prob w.r.t. LM trained on target corpus)
  - Diversity (rare word usage, uniqueness of n-grams)
  - Relevance to input (semantic similarity measures)
  - Simple things like length and repetition
  - Task-specific metrics e.g. compression rate for summarization

Scores help us track some important qualities we care about

Perplexity

- Perplexity is a measurement of confusion degree when a language model predicts a sentence
  - A better LM predicts an unseen test set better → lower perplexity

$$PP(S) = p(w_1, w_2, \cdots, w_N)^{-1/N}$$
$$= \sqrt[N]{\frac{1}{p(w_1, w_2, \cdots, w_N)}}$$

$$PP(S) = 2^{-l}$$

where $l = \frac{1}{N} \log p(w_1, w_2, \cdots, w_N)$

inverse probability of the test set normalized by the number of words

Evaluating the trained generative (probabilistic) language model

Cross Entropy

LM ⟷ True

- Cross entropy is a distance between two distributions

$$H(\underset{\text{predicted}}{p}, \underset{\text{true}}{q}) = \mathbb{E}_q[-\log p(x)] = -\sum_x q(x) \log p(x)$$

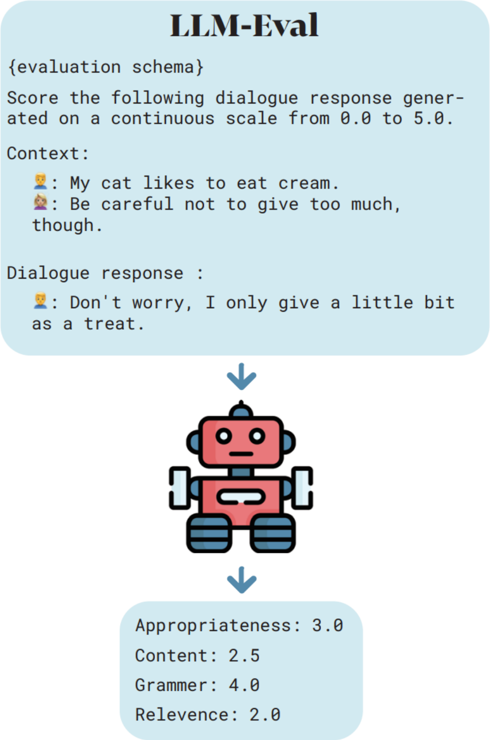
$$= -\frac{1}{N}\sum_{i=1}^{N}\left(\sum_x q(x \mid w_1, \cdots, w_{i-1}) \log p(x \mid w_1, \cdots, w_{i-1})\right)$$

the testing sentence is $w_1, w_2, \cdots, w_{i-1}, w_i$, so $q(w_i \mid w_1, \cdots, w_{i-1}) = 1$

$$= -\frac{1}{N}\sum_{i=1}^{N} \log p(w_i \mid w_1, \cdots, w_{i-1})$$

$$= -\frac{1}{N} \log p(w_1, w_2, \cdots, w_N) = \log PP(S)$$

LLM-Eval (Lin & Chen, 2023)



- LLM has a reasonable capability of evaluating dialogue responses

r / ρ (%)	TopicalChat	PersonaChat	ConvAI2	DD	ED	DSTC6	Average
BLEU-4	21.6 / 29.6	13.5 / 9.0	0.3 / 12.8	7.5 / 18.4	-5.1 / 0.2	13.1 / 29.8	8.5 / 16.6
ROUGE-L	27.5 / 28.7	6.6 / 3.8	13.6 / 14.0	15.4 / 14.7	2.9 / -1.3	33.2 / 32.6	16.5 / 15.4
BERTScore	29.8 / 32.5	15.2 / 12.2	22.5 / 22.4	12.9 / 10.0	4.6 / 3.3	36.9 / 33.7	20.3 / 19.0
DEB	18.0 / 11.6	29.1 / 37.3	42.6 / 50.4	33.7 / **36.3**	35.6 / 39.5	21.1 / 21.4	30.0 / 32.8
GRADE	20.0 / 21.7	35.8 / 35.2	56.6 / 57.1	27.8 / 25.3	33.0 / 29.7	11.9 / 12.2	30.9 / 30.2
USR	41.2 / 42.3	44.0 / 41.8	50.1 / 50.0	5.7 / 5.7	26.4 / 25.5	18.4 / 16.6	31.0 / 30.3
USL-H	32.2 / 34.0	49.5 / 52.3	44.3 / 45.7	10.8 / 9.3	29.3 / 23.5	21.7 / 17.9	31.3 / 30.5
without human reference							
LLM-EVAL 0-5	55.7 / 58.3	51.0 / 48.0	59.3 / 59.6	31.8 / 32.2	42.1 / 41.4	43.3 / 41.1	47.2 / 46.8
LLM-EVAL 0-100	49.0 / 49.9	53.3 / 51.5	**61.3** / **61.8**	**34.6** / 34.9	43.2 / 42.3	44.0 / 41.8	**47.6** / 47.0
with human reference							
LLM-EVAL 0-5	**56.5** / **59.4**	**55.4** / **53.1**	43.1 / 43.8	.320 / 32.2	40.0 / 40.1	47.0 / 45.5	45.7 / 45.7
LLM-EVAL 0-100	55.6 / 57.1	53.8 / 52.7	45.6 / 45.9	33.4 / 34.0	**43.5** / **43.2**	**49.8** / **49.9**	47.0 / **47.1**

LLM-Eval better correlates with human-judged scores than all existing metrics

LLM-Eval (Lin & Chen, 2023)

- LLM-Eval works good on not only **single-turn** but **multi-turn** evaluation

r / ρ (%)	DailyDialog-PE	FED		DSTC9	Average
	Turn-Level	Turn-Level	Dialog-Level	Dialog-Level	
DynaEval	16.7 / 16.0	31.9 / 32.3	50.3 / 54.7	9.3 / 10.1	27.1 / 28.3
USL-H	68.8 / 69.9	20.1 / 18.9	7.3 / 15.2	10.5 / 10.5	26.7 / 28.6
FlowScore	-	-6.5 / -5.5	-7.3 / -0.3	14.7 / 14.0	0.3 / 2.7
GPTScore	-	- / 38.3	- / 54.3	-	- / 46.3
LLM-EVAL 0-5	71.0 / **71.3**	**60.4** / **50.9**	**67.6** / **71.4**	15.9 / 16.5	**53.7** / **52.5**
LLM-EVAL 0-100	**71.4** / 71.0	59.7 / 49.9	64.4 / 70.4	**16.1** / **18.6**	52.9 / 52.5

Idea: LLM-Eval scores can be the proxy of human evaluation

Yen-Ting Lin and Yun-Nung Chen, "LLM-EVAL: Unified Multi-Dimensional Automatic Evaluation for Open-Domain Conversations with Large Language Models," in *Proceedings of NLP for Conversational AI Workshop (NLP4ConvAI)*, 2023.

13

Reinforcement Learning for NLG

Global Optimization

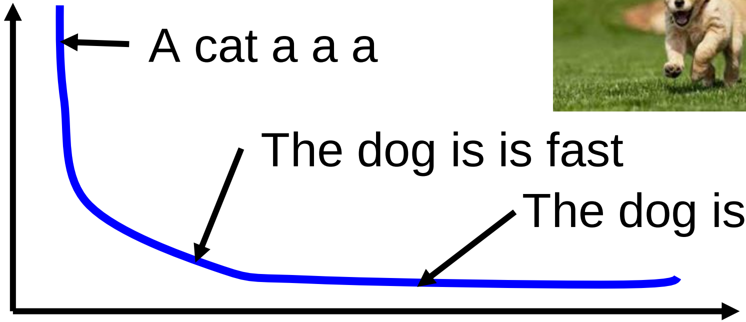
Global Optimization vs. Local Optimization

- Minimizing the error defined on component level (local) is not equivalent to improving the generated objects (global)

Reference: The dog is running fast

$$C = \sum_t C_t$$

Cross-entropy of each step

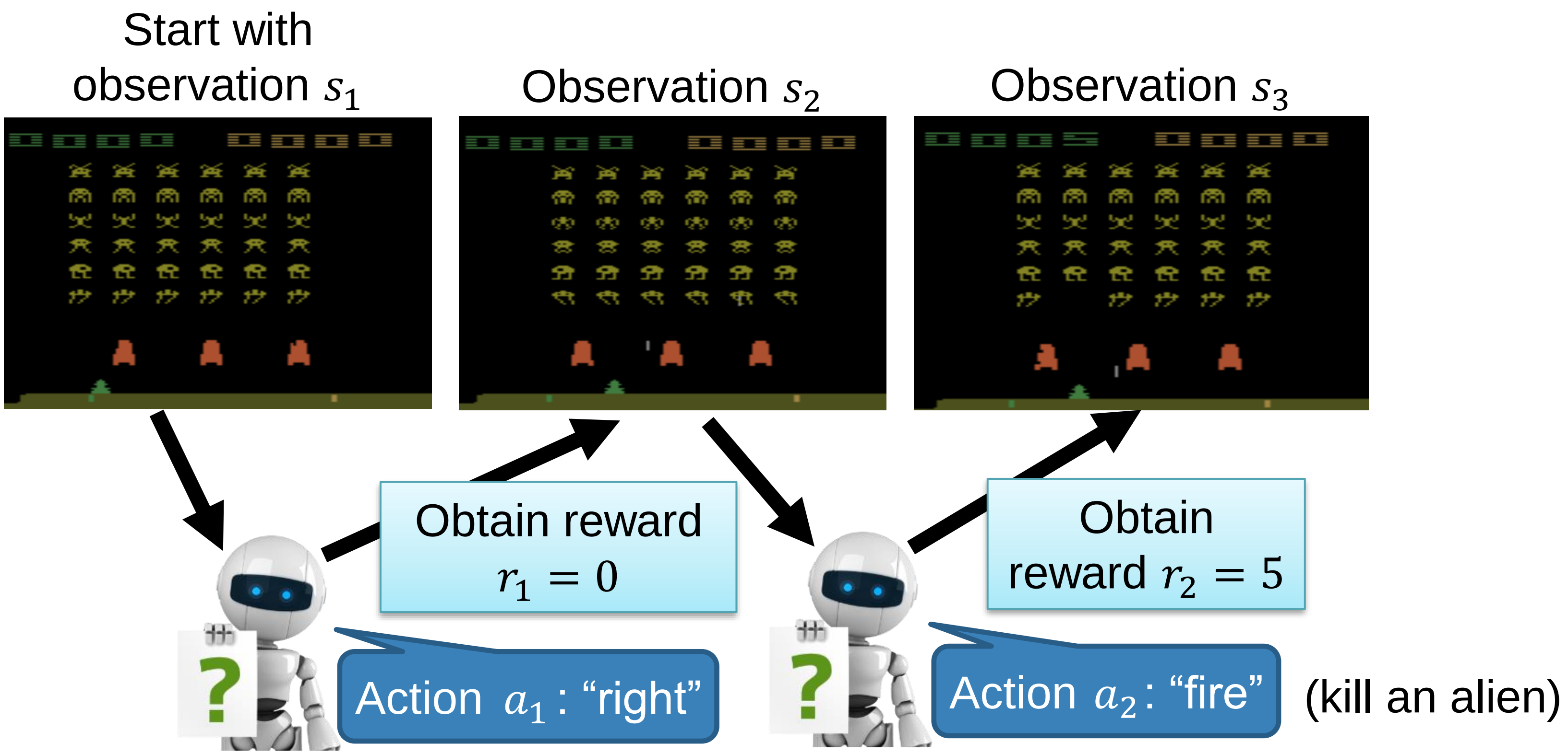


Optimize object-level criterion instead of component-level cross-entropy.

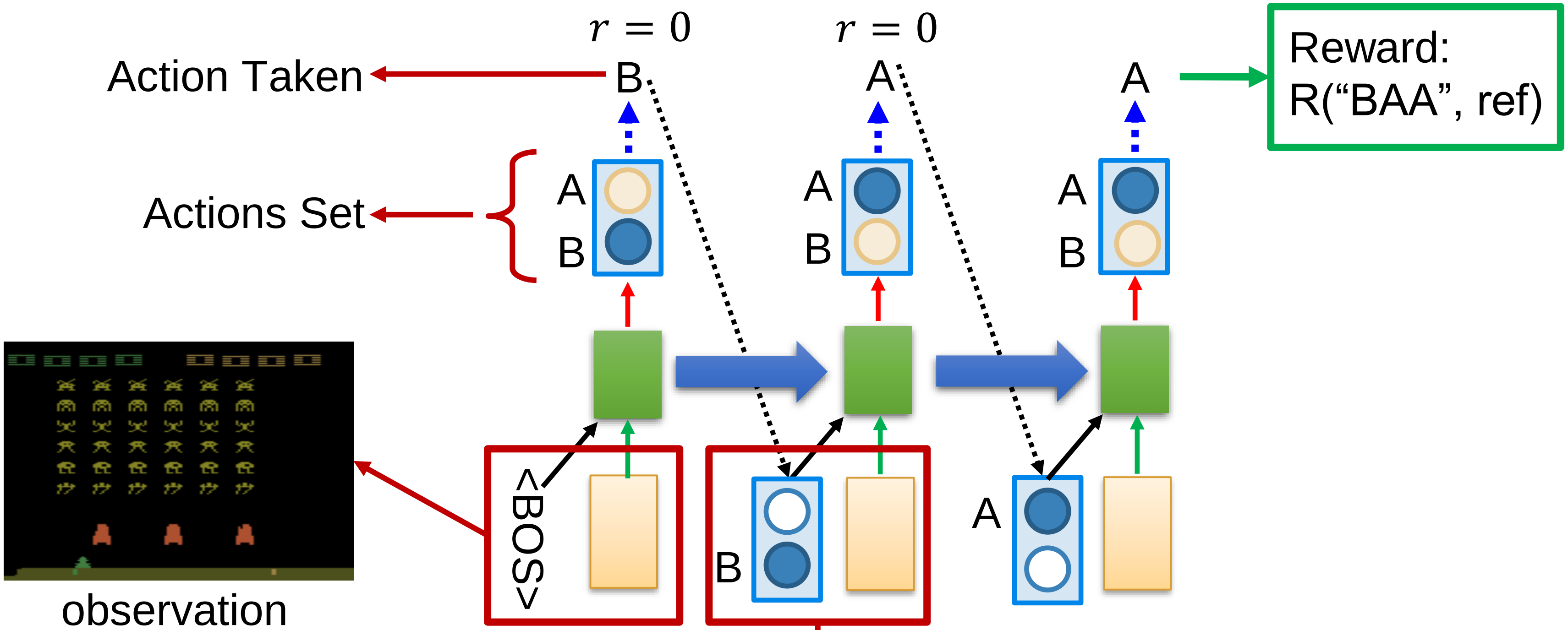
Object-level criterion: $R(y, \hat{y})$ y: ground truth, $\hat{y}$: generated sentence

Gradient Descent?

Reinforcement Learning



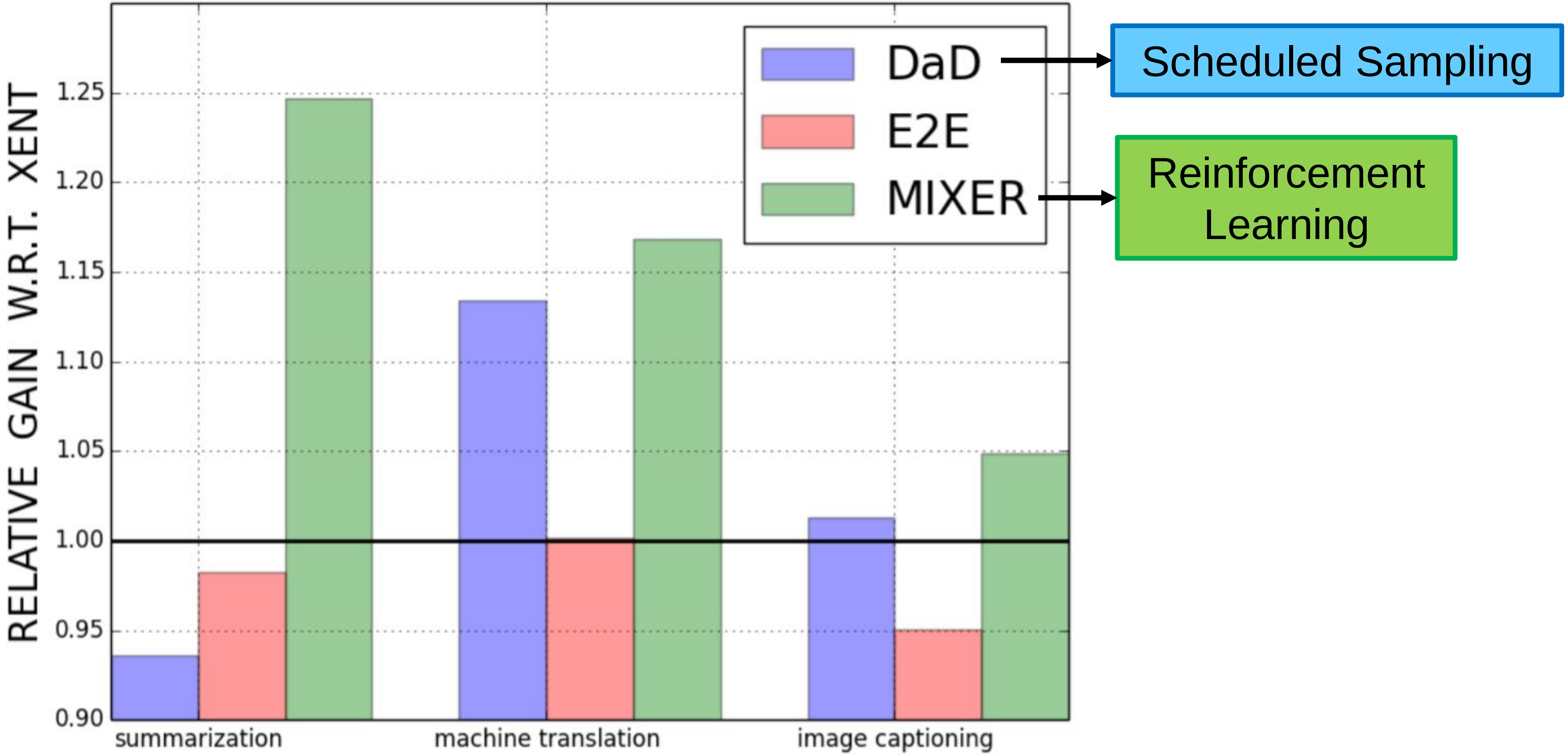
RL for NLG



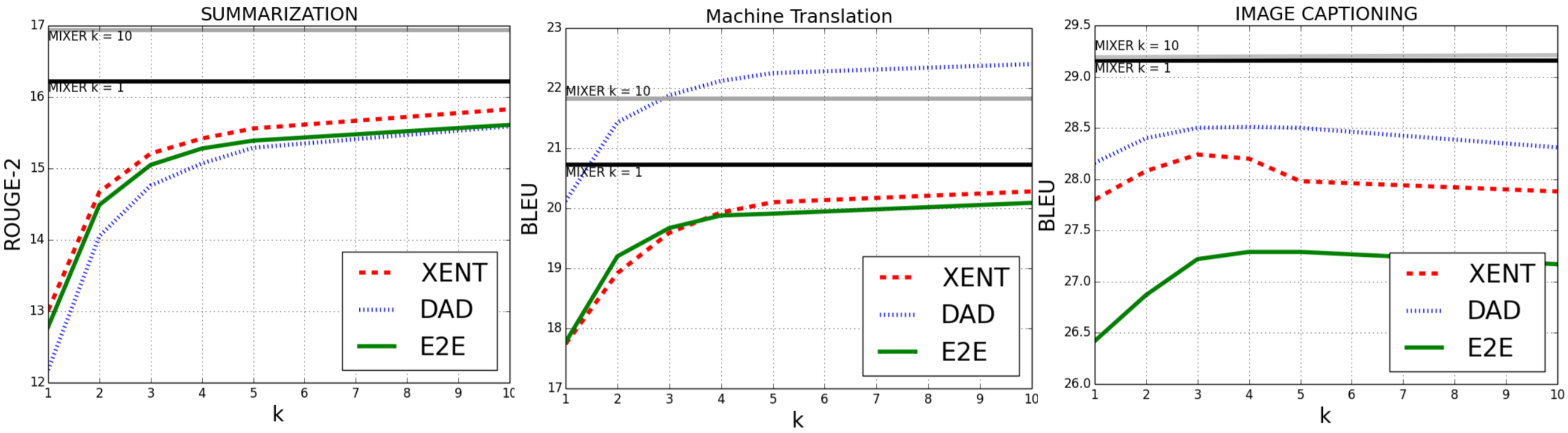
The action taken influences the next observation

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, Wojciech Zaremba, "Sequence Level Training with Recurrent Neural Networks", ICLR, 2016

RL for NLG


RELATIVE GAIN W.R.T. XENT
1.25
1.20
1.15
1.10
1.05
1.00
0.95
0.90
summarization
machine translation
image captioning
DaD
E2E
MIXER
Scheduled Sampling
Reinforcement Learning

RL for NLG


SUMMARIZATION
ROUGE-2
MIXER k = 10
MIXER k = 1
XENT
DAD
E2E
k
Machine Translation
BLEU
MIXER k = 10
MIXER k = 1
XENT
DAD
E2E
k
IMAGE CAPTIONING
BLEU
MIXER k = 10
MIXER k = 1
XENT
DAD
E2E
k

RL-Based Summarization

- RL: directly optimize ROUGE-L
- ML+RL: MLE + RL for optimizing ROUGE-L

Automatic

Model	ROUGE-1	ROUGE-2	ROUGE-L
ML, no intra-attention	44.26	27.43	40.41
ML, with intra-attention	43.86	27.10	40.11
RL, no intra-attention	**47.22**	30.51	**43.27**
ML+RL, no intra-attention	47.03	**30.72**	43.10

Human

Model	Readability	Relevance
ML	6.76	7.14
RL	4.18	6.32
ML+RL	**7.04**	**7.45**

Using RL instead of ML achieves higher ROUGE scores, but lower human scores.

Hybrid is the best.

ChatGPT: Reinforcement Learning from Human Feedback

- **Improving GPT via teacher's feedback**

a conversation history

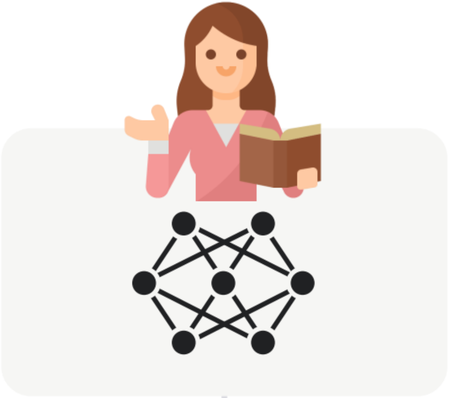
: 我想了解台灣的知名女歌手。

: 台灣知名女歌手包含蔡依林、…

: 可以替我介紹一下蔡依林嗎？

a model-generated output

當然沒問題！蔡依林曾獲x次金曲獎…



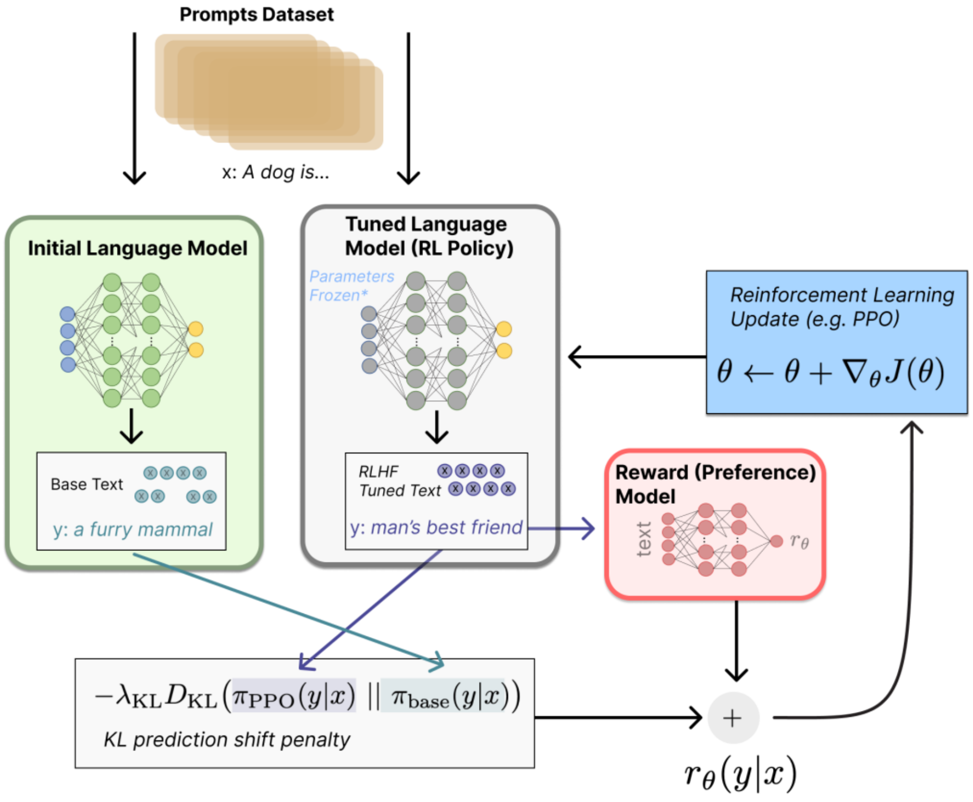
Teacher-given scores

Human Feedback

generation update via reinforcement learning

Idea: optimize abstract indicators (e.g. human's satisfaction)

RLHF: RL from Human Feedback



Concluding Remarks

- Automatic evaluation
  - Output evaluation
  - Model evaluation
- Perplexity
  - Confusion degree when a language model predicts a sentence
  - Cross entropy between true and predicted distributions
  - Lower is better
- RL for NLG
  - Hybrid is better (MLE first, RL later)
  - RL enables models to improve abstract indicators