The AAAI Conference on Artificial Intelligence

Adaptive Prior-Dependent Correction Enhanced Reinforcement Learning for Natural Language Generation

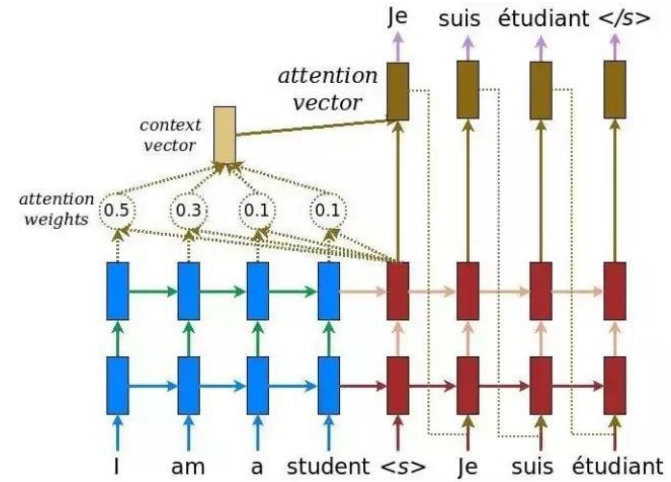
Wei Cheng, Ziyan Luo[#], Qiyue Yin[*]

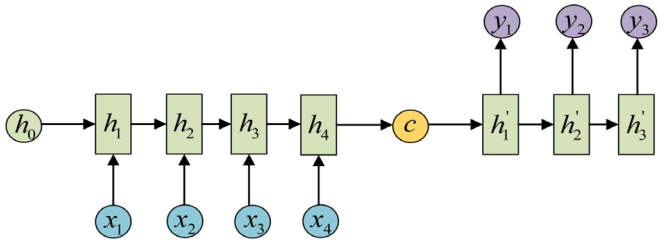
weicheng5993@foxmail.com, z5luo@ucsd.edu, qyyin@nlpr.ia.ac.cn

[#]Presenting author [*]Corresponding author

Natural Language Generation

- Neural natural language generation (NLG) aims to generate a piece of new text
  - neural machine translation (NMT)
  - image captioning
  - text summarization
  - …
- NLG models have recently shown remarkable progress in language fluency and coherence
- We focus the training of Seq2Seq NLG models





Motivation

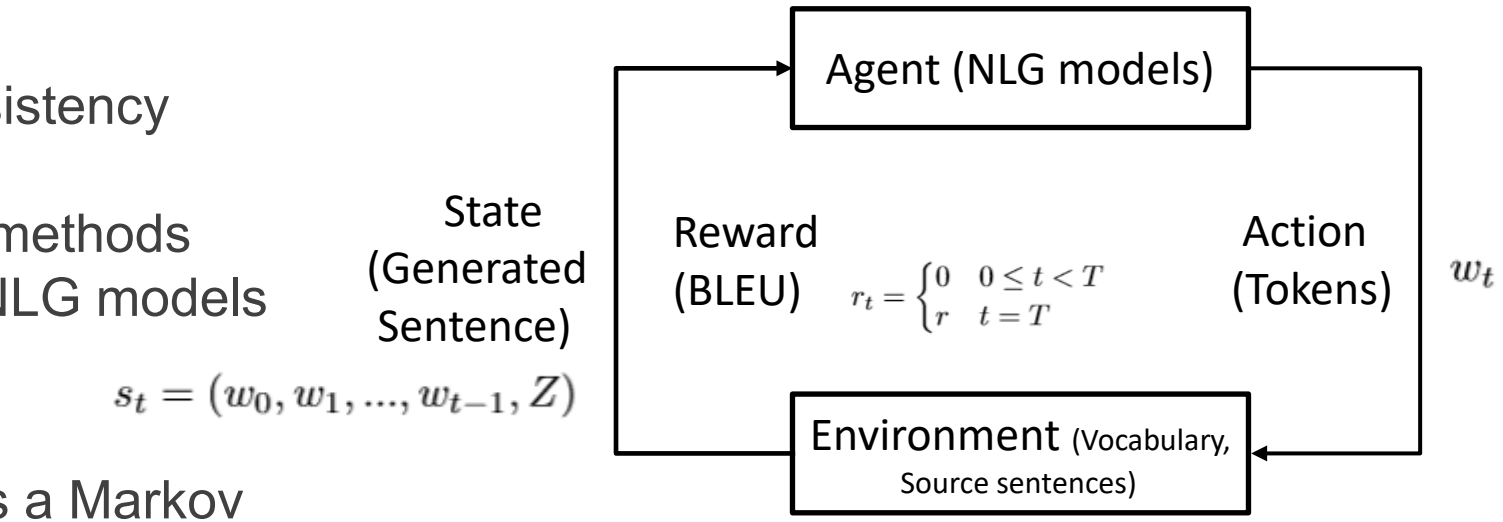
- Advantages of RL-based methods:
  - using the current output as the input of next step
  - directly optimizing the evaluation metric
  - avoiding the issues of loss inconsistency and exposure bias

	MLE	Other RL methods	Ours
loss inconsistency	√	-	-
exposure bias	√	-	-

RL for NLG

- To address the loss inconsistency and exposure bias issues, reinforcement learning (RL) methods have been adopted to train NLG models

- Formulation the problem as a Markov Decision Process



MDP Formulation and trial-and-error learning

Motivation

- Advantages of RL-based methods:
  - using the current output as the input of next step
  - directly optimizing the evaluation metric
  - avoiding the issues of loss inconsistency and exposure bias
- However, the reward, such as BLEU/ROUGE, assign the same score to the different incorrect generated tokens, which is called *deviation ignorance*

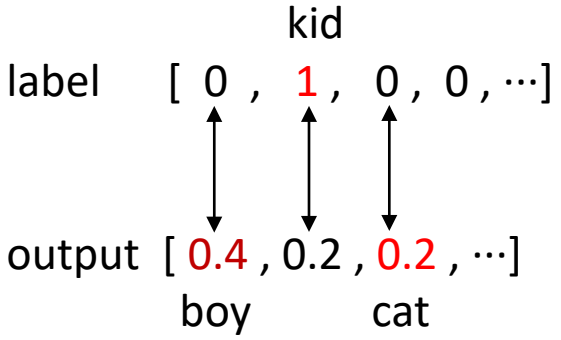
	MLE	**Other RL methods**	**Ours**
loss inconsistency	√	-	-
exposure bias	√	-	-
deviation ignorance	√	√	-

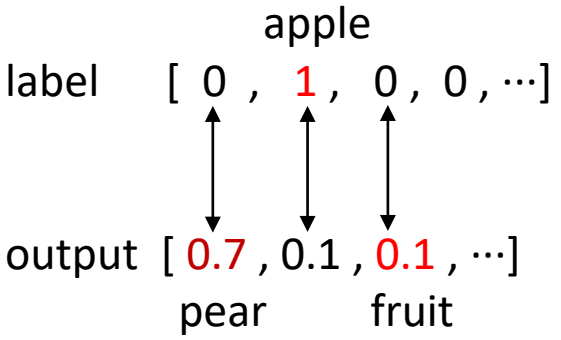
Deviation Ignorance

The models fail to understand how much the prediction distribution deviates from a prior distribution related to the ground-truth at token-level

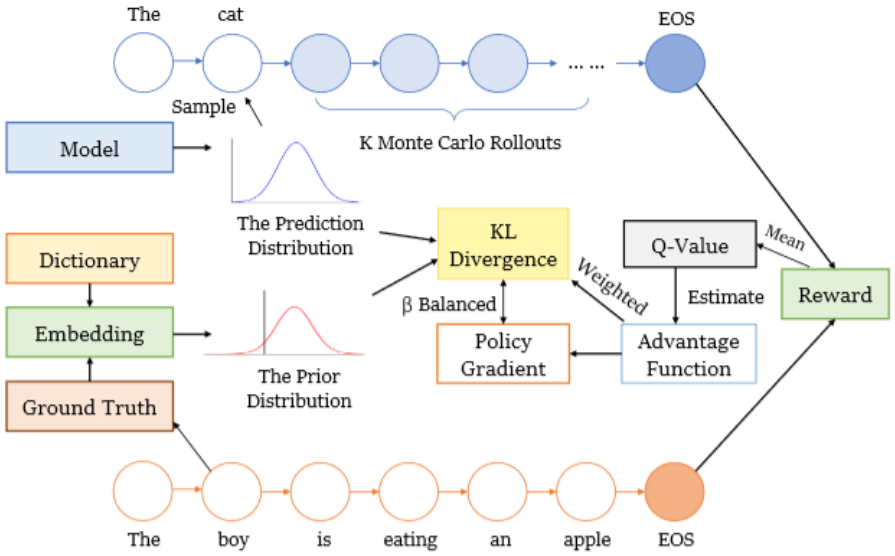
The metrics such as *BLEU* and *ROUGE* assign the same scores for the totally different predictions

Ground Truth: “the boy is eating an apple”
Predictions:
- “the *kid* is eating an apple” √
- “the boy is *having* an apple” √
- “the *cat* is eating an apple” ×
- “the boy is eating an *pear*” ×





Methodology



APDC: To alleviate *deviation ignorance* issue, we enhance the RL objective with a KL term with an **adaptive factor**:

$$L(\theta) = -L_{RL}(\theta) + \beta L_{KL}(\theta)$$

- Adaptive Factor: negatively correlated to advantage function

$$L_{KL}(\theta) = \sum_{t=0}^{T} e^{-\alpha \tilde{A}^{\pi}(s_t, a_t)} KL\left[p^*(w_t) || p_\theta(w_t)\right]$$

- The prior distribution:

$$p^*(w_t) = \sigma\left(cos\_sim\left(emb\left(w_t^*\right), emb\left(w_t\right)\right)\right)$$

Algorithm 1: APDC Enhanced Reinforcement Learning

Input: The model input S and the ground truth sentence $Y = (w_1^*, w_2^* ..., w_L^*)$;
Build the NLG model $M(\psi)$ with random initial weights ψ;
Pre-train the $M(\psi)$ with MLE and update $M(\psi)$ to $M(\theta)$;
Pre-train the word embedding and compute the prior distribution for all ground-truth tokens w_t^* with the word embedding
$p^*(w_t) = \sigma\left(cos\_sim\left(emb\left(w_t^*\right), emb\left(w_t\right)\right)\right)$;
while *not converged* **do**
  for *each time steps* $t = 0, ..., T$ **do**
    Sample token a_t from the policy π;
    Use K Monte Carlo rollouts inference algorithm to sample K $a_{t+1:T}$;
    Compute the estimated Q-value
    $\tilde{Q}^\pi(s_t, a_t) = \frac{1}{K}\sum_{k=1}^{K} R(a_{0:t-1}; a_t; a_{t+1:T}^k)$;
    Compute the estimated advantage value
    $\tilde{A}^\pi(s_t, a_t) = \tilde{Q}^\pi(s_t, a_t) - \tilde{Q}^\pi(s_{t-1}, a_{t-1})$;
    Compute the KL divergence
    $D_t(\theta) = e^{-\alpha \tilde{A}^\pi(s_t, a_t)} KL\left[p^*(w_t) || p_\theta(w_t)\right]$;
  end
  Compute the RL loss
  $L_{RL}(\theta) = \sum_{t=0}^{T} \tilde{A}^\pi(s_t, a_t)\pi(s_t, a_t)$;
  Compute the KL loss
  $L_{KL}(\theta) = \sum_{t=0}^{T} D_t(\theta)$;
  Update $M(\theta)$ by minimizing $(-L_{RL}(\theta) + \beta L_{KL}(\theta))$;
end

Methodology

- APG: to reduce variance, we estimate the advantage function per step using the prior knowledge that the transition **p(s'|s, a)** is determined (≡1 if s' is observed)

$$\tilde{A}^\pi(s_t, a_t) = \tilde{Q}^\pi(s_t, a_t) - \tilde{Q}^\pi(s_{t-1}, a_{t-1})$$

$$\nabla_\theta L_{RL}(\theta) = \sum_{t=0}^{T} \mathbb{E}_\pi[A^\pi(s_t, a_t)\nabla_\theta log(\pi_\theta(a_t|s_t))]$$

- Leverage K Monte-Carlo rollouts to estimate Q-values

$$\tilde{Q}^\pi(s_t, a_t) = \frac{1}{K}\sum_{k=1}^{K} R(a_{0:t-1}; a_t; a_{t+1:T}^k)$$

Experiments

We evaluate our algorithm in three tasks of NLG

Neural Machine Translation

Table 1: BLEU score of NMT for En-De, En-Zh, and Zh-En.

Method	En-De	En-Zh	Zh-En
Transformer+MLE (Vaswani et al. 2017)	27.30	34.12	24.29
MIXER (Ranzato et al. 2016)	27.43	34.38	24.59
RL4NMT (Wu et al. 2018)	27.52	34.46	24.70
APG	27.63	34.54	24.81
APG+PDC	27.70	34.56	24.90
RL4NMT+PDC	27.81	34.62	24.94
APG+APDC	**28.03**	**34.91**	**25.28**

Image Captioning

Table 2: Performance of image captioning on the MSCOCO Karpathy test split.

Method	BLEU-1	BLEU-4	METEOR	ROUGE-L	CIDEr
SCST (Att2all) (Rennie et al. 2017)	-	34.2	26.7	55.7	114.0
OTRL (Chen et al. 2020a)	79.3	34.4	26.8	56.2	111.8
Top-Down+MLE (Anderson et al. 2018)	77.2	36.2	27.0	56.4	113.5
Top-Down+SCST (Anderson et al. 2018)	79.8	36.3	27.7	56.9	120.1
Top-Down+MIXER (Ranzato et al. 2016)	78.4	36.2	27.4	56.5	115.6
Top-Down+SCST+APDC	79.9	36.6	27.8	56.9	120.2
APG	80.1	36.4	28.0	56.9	120.2
APG+PDC	80.3	36.7	28.4	57.3	121.2
APG+APDC (METEOR)	80.1	36.5	29.2	57.1	119.7
APG+APDC	**80.8**	**37.9**	**28.9**	**58.1**	**123.6**

Abstractive Text Summarization

Table 3: Results of abstractive text summarization on CNN/Daily Mail dataset. "Pointer" means the method Pointer-Generator+Coverage.

Method	ROUGE-1	ROUGE-2	ROUGE-L
Pointer (See, Liu, and Manning 2017)	39.53	17.28	36.38
MIXER (Ranzato et al. 2016)	39.78	17.91	37.15
OTRL (Chen et al. 2020a)	41.40	18.22	38.86
APG	41.51	18.34	38.93
APG+PDC	41.60	18.48	39.02
APG+APDC	**42.73**	**18.81**	**39.85**

Experiments

Comparison Approaches

APG performs better than other RL-based methods

Table 1: BLEU score of NMT for En-De, En-Zh, and Zh-En.

Method	En-De	En-Zh	Zh-En
Transformer+MLE (Vaswani et al. 2017)	27.30	34.12	24.29
MIXER (Ranzato et al. 2016)	27.43	34.38	24.59
RL4NMT (Wu et al. 2018)	27.52	34.46	24.70
APG	27.63	34.54	24.81
APG+PDC	27.70	34.56	24.90
RL4NMT+PDC	27.81	34.62	24.94
APG+APDC	**28.03**	**34.91**	**25.28**

Table 2: Performance of image captioning on the MSCOCO Karpathy test split.

Method	BLEU-1	BLEU-4	METEOR	ROUGE-L	CIDEr
SCST (Att2all) (Rennie et al. 2017)	-	34.2	26.7	55.7	114.0
OTRL (Chen et al. 2020a)	79.3	34.4	26.8	56.2	111.8
Top-Down+MLE (Anderson et al. 2018)	77.2	36.2	27.0	56.4	113.5
Top-Down+SCST (Anderson et al. 2018)	79.8	36.3	27.7	56.9	120.1
Top-Down+MIXER (Ranzato et al. 2016)	78.4	36.2	27.4	56.5	115.6
Top-Down+SCST+APDC	79.9	36.6	27.8	56.9	120.2
APG	80.1	36.4	28.0	56.9	120.2
APG+PDC	80.3	36.7	28.4	57.3	121.2
APG+APDC (METEOR)	80.1	36.5	29.2	57.1	119.7
APG+APDC	**80.8**	**37.9**	**28.9**	**58.1**	**123.6**

Table 3: Results of abstractive text summarization on CNN/Daily Mail dataset. "Pointer" means the method Pointer-Generator+Coverage.

Method	ROUGE-1	ROUGE-2	ROUGE-L
Pointer (See, Liu, and Manning 2017)	39.53	17.28	36.38
MIXER (Ranzato et al. 2016)	39.78	17.91	37.15
OTRL (Chen et al. 2020a)	41.40	18.22	38.86
APG	41.51	18.34	38.93
APG+PDC	41.60	18.48	39.02
APG+APDC	**42.73**	**18.81**	**39.85**

Experiments

APDC can also enhance other RL-based methods

APDC performs better than PDC (APDC without adaptive mechanism)

Table 1: BLEU score of NMT for En-De, En-Zh, and Zh-En.

Method	En-De	En-Zh	Zh-En
Transformer+MLE (Vaswani et al. 2017)	27.30	34.12	24.29
MIXER (Ranzato et al. 2016)	27.43	34.38	24.59
RL4NMT (Wu et al. 2018)	27.52	34.46	24.70
APG	27.63	34.54	24.81
APG+PDC	27.70	34.56	24.90
RL4NMT+PDC	27.81	34.62	24.94
APG+APDC	**28.03**	**34.91**	**25.28**

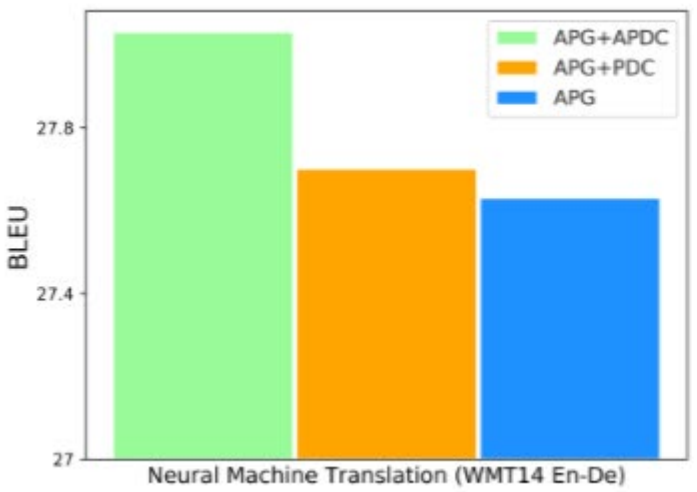
Table 2: Performance of image captioning on the MSCOCO Karpathy test split.

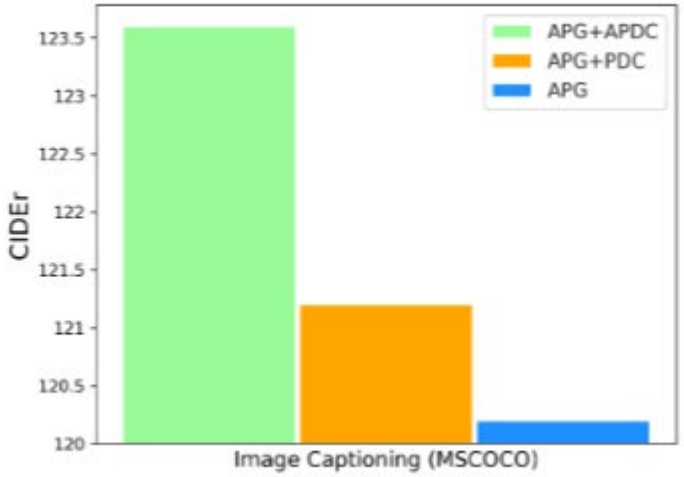
Method	BLEU-1	BLEU-4	METEOR	ROUGE-L	CIDEr
SCST (Att2all) (Rennie et al. 2017)	-	34.2	26.7	55.7	114.0
OTRL (Chen et al. 2020a)	79.3	34.4	26.8	56.2	111.8
Top-Down+MLE (Anderson et al. 2018)	77.2	36.2	27.0	56.4	113.5
Top-Down+SCST (Anderson et al. 2018)	79.8	36.3	27.7	56.9	120.1
Top-Down+MIXER (Ranzato et al. 2016)	78.4	36.2	27.4	56.5	115.6
Top-Down+SCST+APDC	79.9	36.6	27.8	56.9	120.2
APG	80.1	36.4	28.0	56.9	120.2
APG+PDC	80.3	36.7	28.4	57.3	121.2
APG+APDC (METEOR)	80.1	36.5	29.2	57.1	119.7
APG+APDC	**80.8**	**37.9**	**28.9**	**58.1**	**123.6**

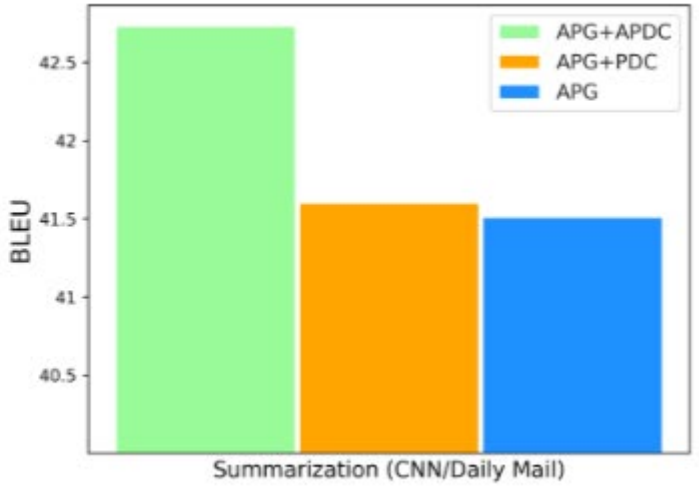
Experiments

Ablation Study illustration

APG+APDC > APG+PDC > APG







Summary

- Formulate the NLG problem as a Markov Decision Process and use an RL to solve the exposure bias and loss inconsistency issues
- Propose a novel technique: adaptive prior-dependent correction to further address the deviation ignorance issue
- Combine some advantage function estimation techniques
- Enhancing APG with APDC can strike a balance between token-level and sequence-level optimization
- Extensive experiments show that, on three tasks, our method consistently outperforms the state-of-the-art approaches

Thank you
for your careful listening!