Contrastive Learning with Adversarial Perturbations for Conditional Text Generation

Seanie Lee*, Dong Bok Lee*, Sung Ju Hwang

ICLR 2021 (4 6 5 6)

Reporter: Xiachong Feng

Authors

Seanie Lee

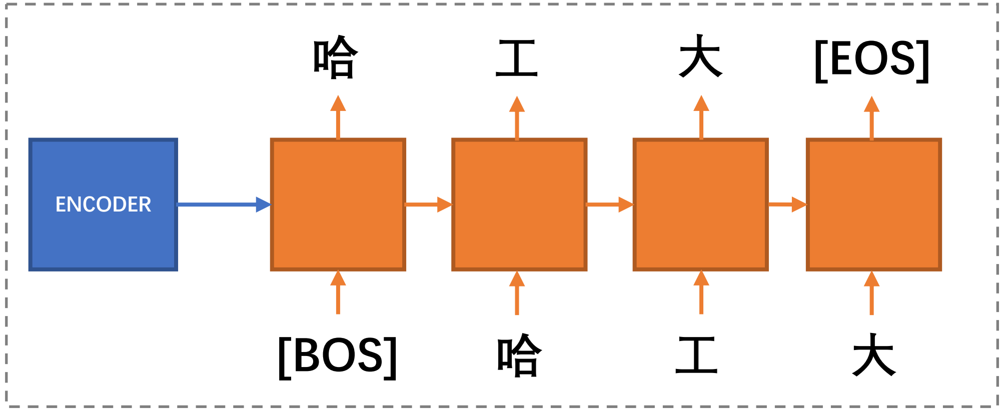
Dong Bok Lee

Sung Ju Hwang

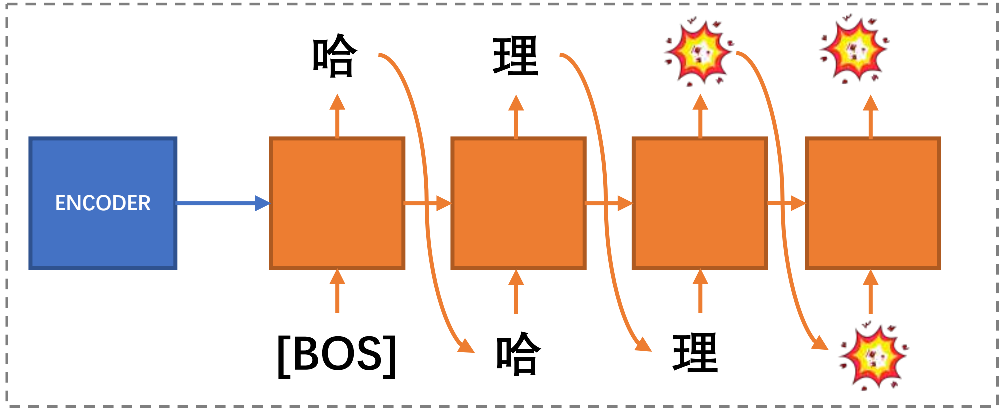
Background

Problem: Exposure Bias

- **Train: teacher forcing**
  - Input ground truth label.
- **Test:**
  - Input previous generated words.



Train

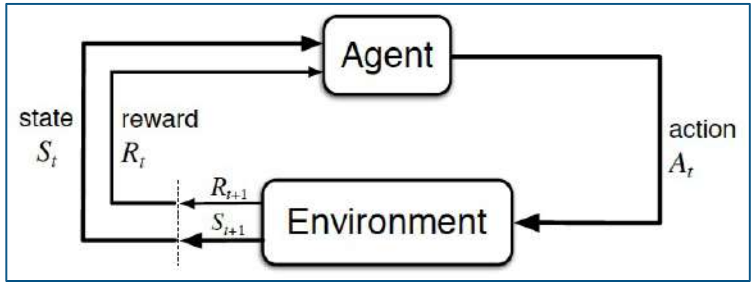


Test

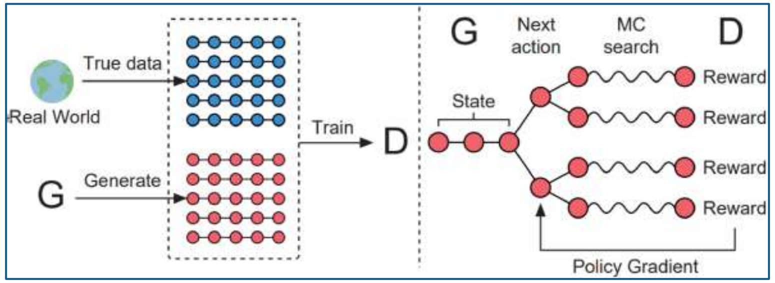
Prior Works to Tackle the Exposure Bias

```
# Probabilities indicating whether to use ground truth labels
instead of previous decoded tokens
use_ground_truth = get_cuda((torch.rand(len(enc_out)) >
0.25)).long()
# Select decoder input based on use_ground_truth probabilities
x_t = use_ground_truth * dec_batch[:, t] + (1 - use_ground_truth)
* x_t
# Decoder input
x_t = self.model.embeds(x_t)
```

Scheduled Sampling

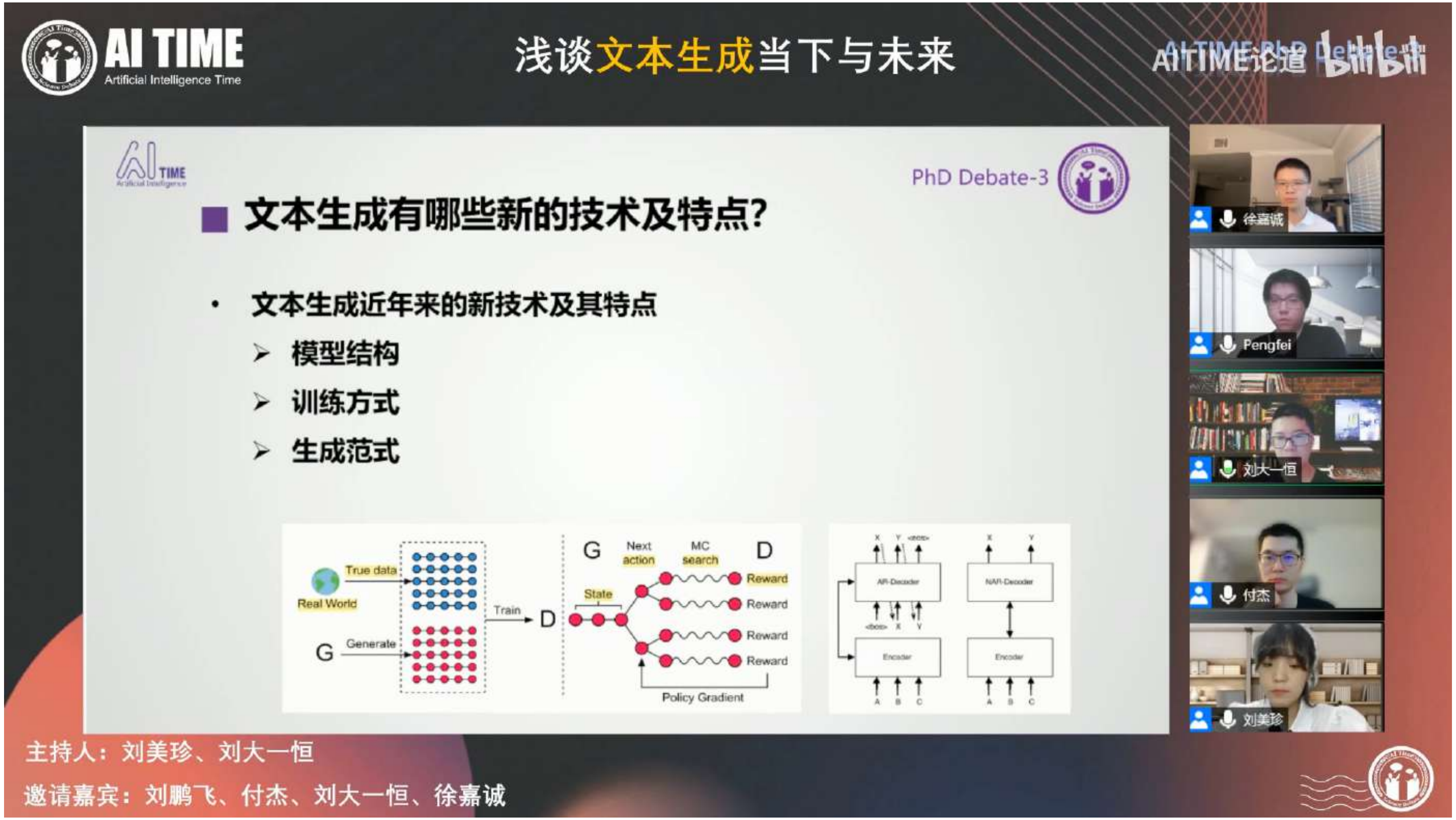


Reinforcement Learning



GAN

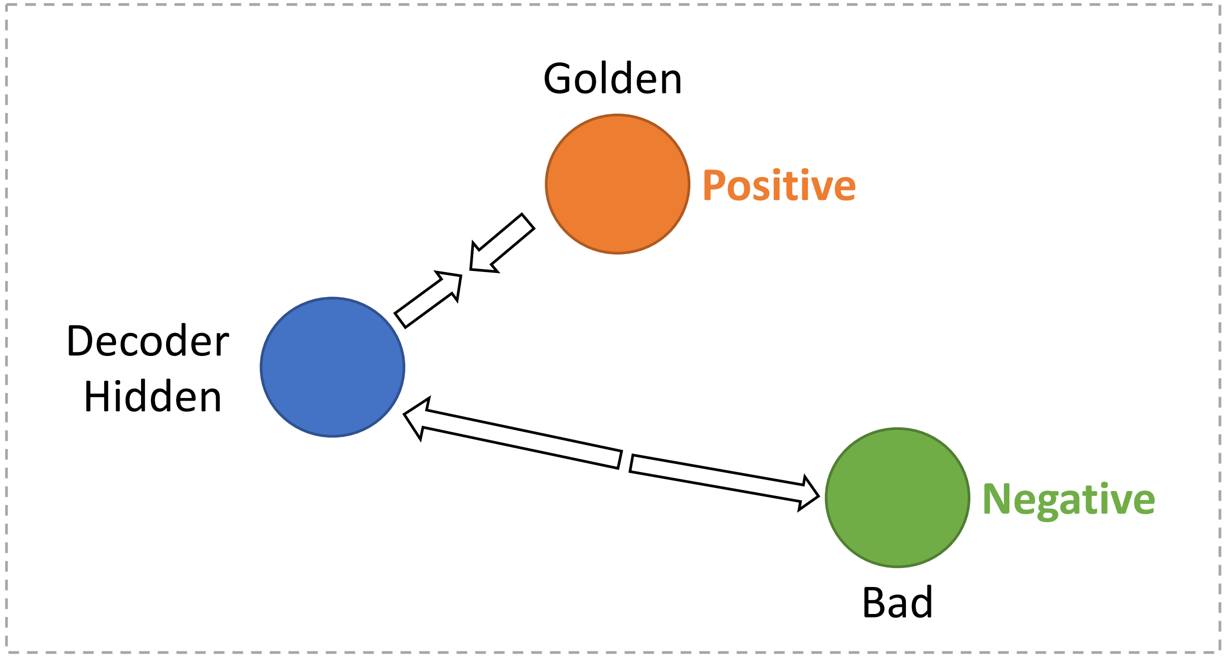
How about Reinforce and GAN?



Overview

Overview

- Contrastive Learning



Conditional Text Generation

$$\mathbf{x}^{(i)} = (x_1^{(i)}, \ldots, x_L^{(i)}) \rightarrow \mathbf{y}^{(i)} = (y_1^{(i)}, \ldots, y_T^{(i)})$$

$$\mathcal{L}_{MLE}(\theta) = \sum_{i=1}^{N} \log p_\theta(\mathbf{y}^{(i)} | \mathbf{x}^{(i)})$$

$$p_\theta(y_1^{(i)}, \ldots, y_T^{(i)} | \mathbf{x}^{(i)}) = \prod_{t=1}^{T} p_\theta(y_t^{(i)} | \boxed{\mathbf{y}_{<t}^{(i)},} \mathbf{x}^{(i)})$$

$$p_\theta(y_t^{(i)} | \mathbf{y}_{<t}^{(i)}, \mathbf{x}^{(i)}) = \mathrm{softmax}(\mathbf{W}\mathbf{h}_t^{(i)} + \mathbf{b})$$

$$\mathbf{h}_t^{(i)} = \underbrace{g}_{\text{Decoder}}(y_{t-1}^{(i)}, \mathbf{M}^{(i)}; \theta), \quad \mathbf{M}^{(i)} = \underbrace{f}_{\text{Encoder}}(\mathbf{x}^{(i)}; \theta)$$

$$\mathbf{M}^{(i)} = [\mathbf{m}_1^{(i)} \cdots \mathbf{m}_L^{(i)}] \in \mathbb{R}^{d \times L}$$

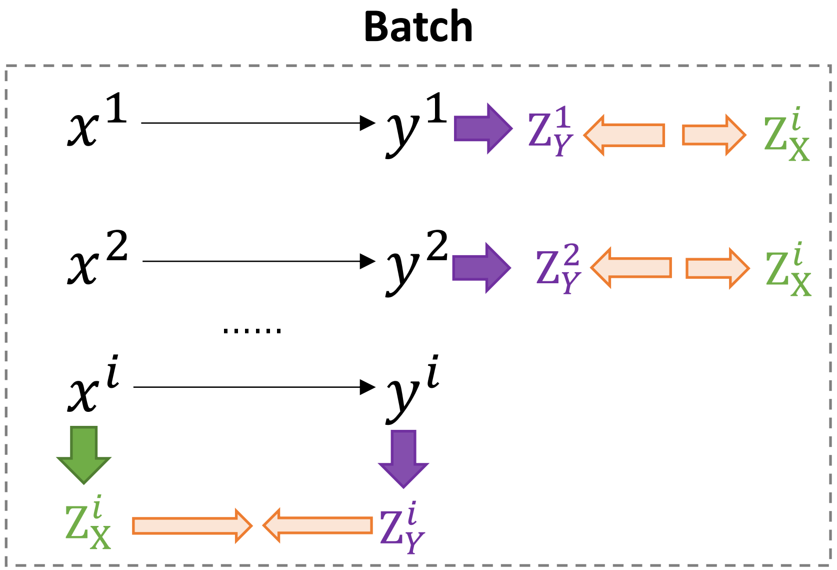
Simple Contrastive Learning Framework

- select the negative pairs as a random non-target output sequence from the same batch.
- maximize the similarity between the pair of source and target sequence, while minimizing the similarity between the negative pairs

$$\boxed{\mathcal{L}_{cont}(\theta) = \sum_{i=1}^{N} \log \frac{\exp(\mathrm{sim}(\mathbf{z}_{\mathbf{x}}^{(i)}, \mathbf{z}_{\mathbf{y}}^{(i)})/\tau)}{\sum_{\mathbf{z}_{\mathbf{y}}^{(j)} \in S} \exp(\mathrm{sim}(\mathbf{z}_{\mathbf{x}}^{(i)}, \mathbf{z}_{\mathbf{y}}^{(j)})/\tau)}}$$

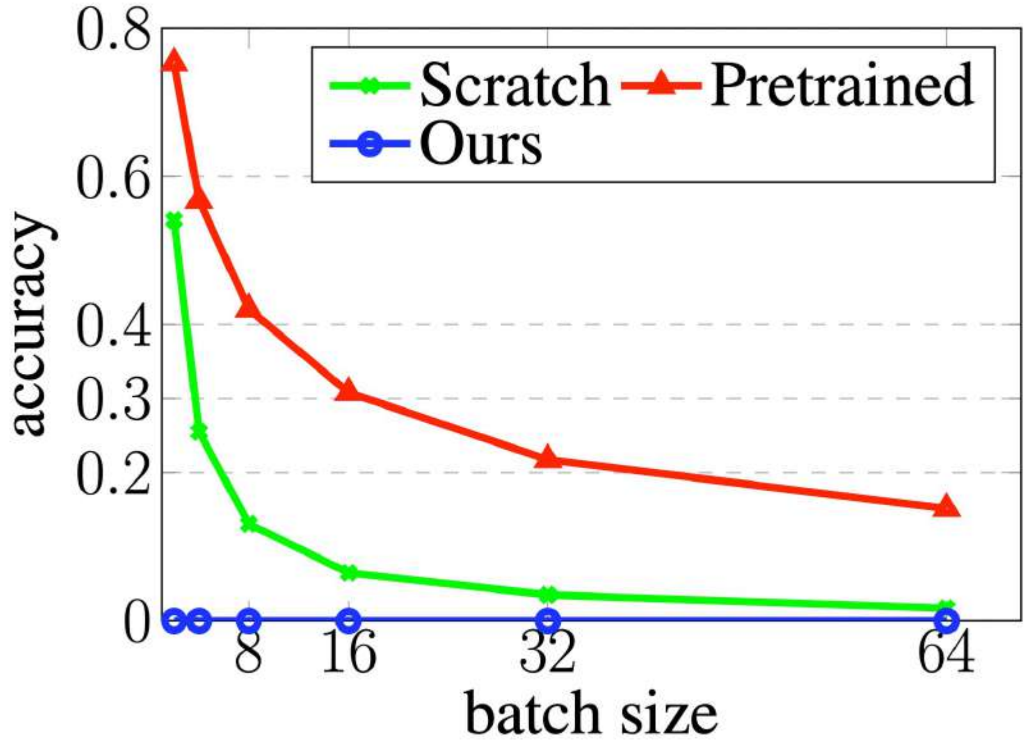
$$\underline{\mathbf{z}_{\mathbf{x}}^{(i)} = \xi(\mathbf{M}^{(i)}; \theta)}, \; \underline{\mathbf{z}_{\mathbf{y}}^{(i)} = \xi(\mathbf{H}^{(i)}; \theta)}$$

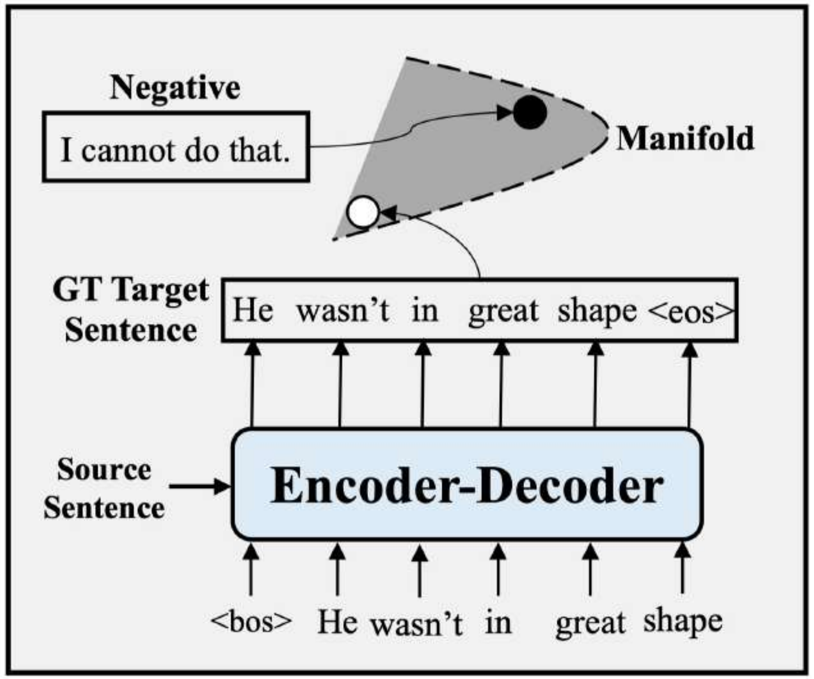
$$\xi([\mathrm{v}_1 \cdots \mathrm{v}_T]; \theta) := \mathrm{AvgPool}([\mathbf{u}_1 \cdots \mathbf{u}_T]), \text{ where } \mathrm{u}_t = \mathrm{ReLU}(\mathbf{W}^{(1)}\mathrm{v}_t + \mathbf{b}^{(1)})$$



Problem

- large portion of positive-negative pairs can be easily discriminated without any training



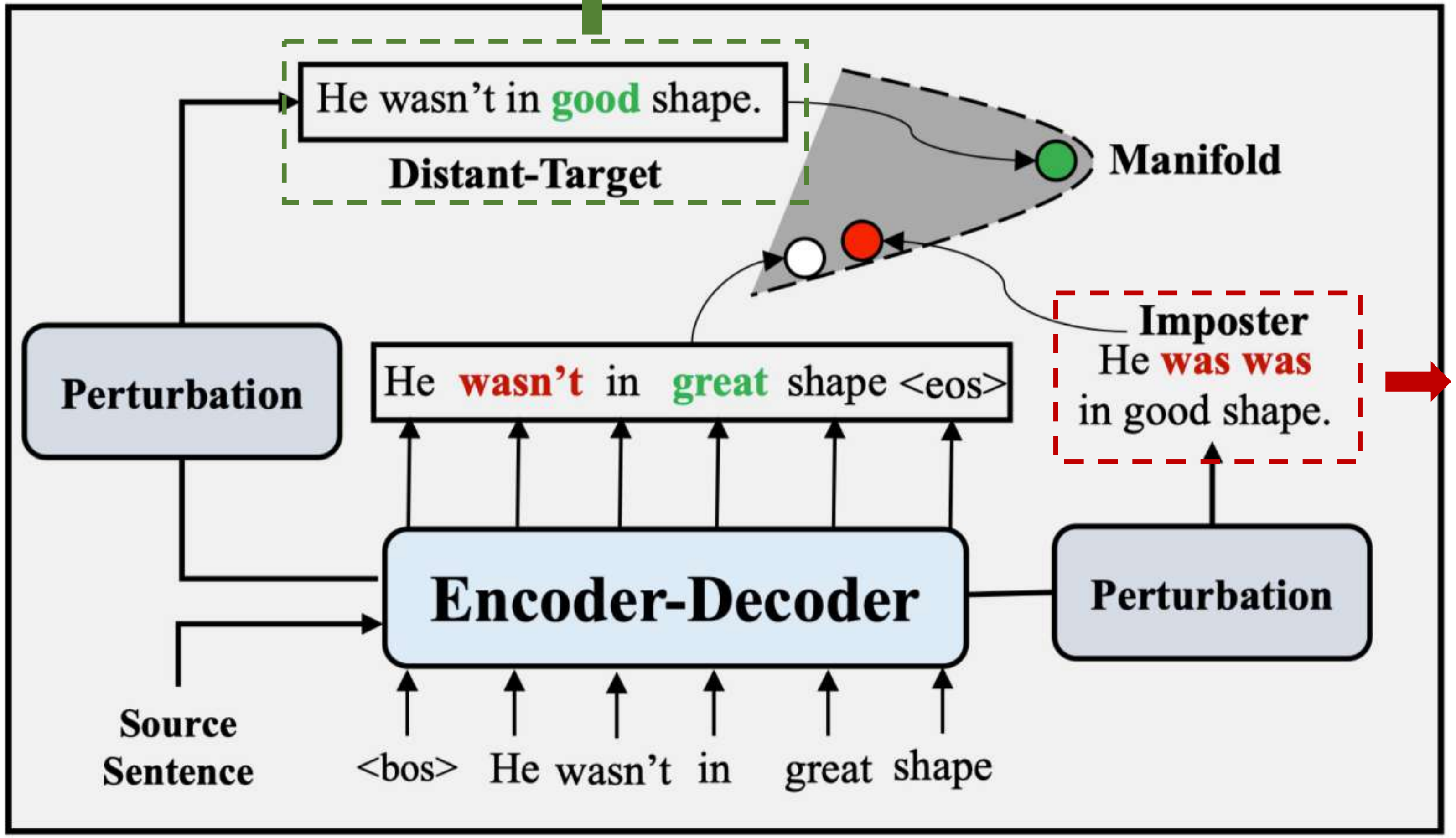


(b) Randomly Sampled Negative Example

Contrastive Learning with Adversarial Perturbations for Seq2Seq

Method

① Semantic space↓
② Embedding space↑



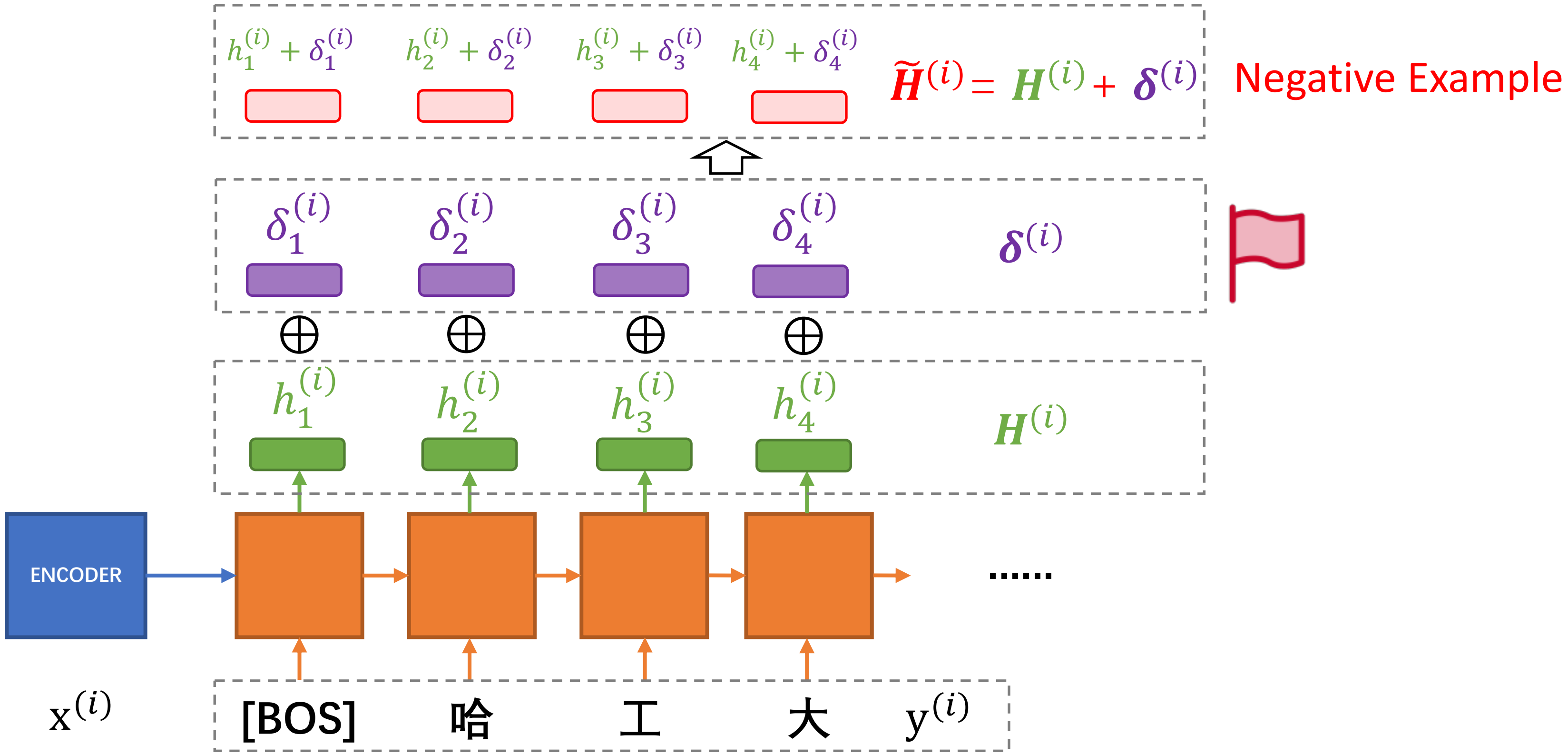
① Semantic space↑
② Embedding space↓

Imposter / Distant-Target Generation with perturbation

Generation of Imposters

$$p_\theta(y_t^{(i)}|\mathbf{y}_{<t}^{(i)}, \mathbf{x}^{(i)}; \mathbf{h}_t^{(i)} + \delta_t) = \text{softmax}\{\mathbf{W}(\underline{\mathbf{h}_t^{(i)}} + \underline{\delta_t}) + \mathbf{b}\}, \text{ where } \delta_t \in \mathbb{R}^d$$

$$p_\theta(\mathbf{y}^{(i)}|\mathbf{x}^{(i)}; \underline{\mathbf{H}^{(i)}} + \underline{\boldsymbol{\delta}}) = \prod_{t=1}^{T} p_\theta(y_t^{(i)}|\mathbf{y}_{<t}^{(i)}, \underline{\mathbf{x}^{(i)}}; \underline{\mathbf{h}_t^{(i)}} + \underline{\delta_t})$$



Generation of Imposters

① Semantic space↑
② Embedding space↓

$$\boldsymbol{\delta}^{(i)} = \underset{\boldsymbol{\delta}, ||\boldsymbol{\delta}||_2 \leq \epsilon}{\arg\min} \log p_\theta(\mathbf{y}^{(i)} | \mathbf{x}^{(i)}; \mathbf{H}^{(i)} + \boldsymbol{\delta})$$

Small perturbation

$$\tilde{\mathbf{H}}^{(i)} = \mathbf{H}^{(i)} - \epsilon \frac{\boldsymbol{g}}{||\boldsymbol{g}||_2}, \text{ where } \boldsymbol{g} = \nabla_{\mathbf{H}^{(i)}} \log p_\theta(\mathbf{y}^{(i)} | \mathbf{x}^{(i)})$$

Conditional likelihood is minimized

$$\boxed{\mathcal{L}_{cont-neg}(\theta)} = \sum_{i=1}^{N} \log \frac{\exp(\text{sim}(\mathbf{z}_{\mathbf{x}}^{(i)}, \mathbf{z}_{\mathbf{y}}^{(i)})/\tau)}{\sum_{\mathbf{z}_{\mathbf{y}}^{(k)} \in \boxed{S \cup \{\tilde{\mathbf{z}}_{\mathbf{y}}^{(i)}\}}} \exp(\text{sim}(\mathbf{z}_{\mathbf{x}}^{(i)}, \mathbf{z}_{\mathbf{y}}^{(k)})/\tau)}, \text{ where } \tilde{\mathbf{z}}_{\mathbf{y}}^{(i)} = \xi(\tilde{\mathbf{H}}^{(i)}; \theta)$$

Generation of Distant-Targets

① Semantic space↓
② Embedding space↑

①

$$\overline{\mathbf{H}}^{(i)} = \mathbf{H}^{(i)} - \eta \frac{\mathbf{g}}{||\mathbf{g}||_2} \text{ where } \mathbf{g} = \nabla_{\mathbf{H}^{(i)}} \mathcal{L}_{cont}(\theta)$$

$$p_\theta(\hat{y}_t^{(i)} | \hat{\mathbf{y}}_{<t}^{(i)}, \mathbf{x}^{(i)}) = \text{softmax}(\mathbf{W}\overline{\mathbf{h}}_t^{(i)} + \mathbf{b})$$

② Embedding space↑

②

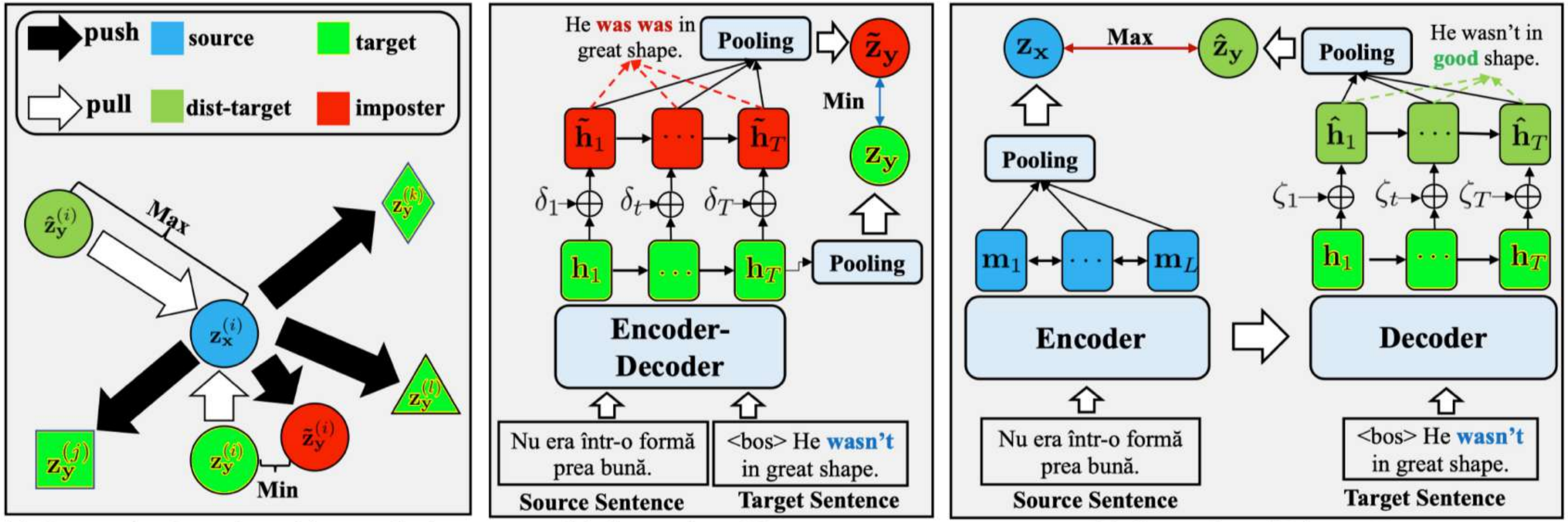
$$\mathcal{L}_{KL}(\theta) = \sum_{i=1}^{N} \sum_{t=1}^{T} D_{KL}(p_{\theta^*}(y_t^{(i)} | \mathbf{y}_{<t}^{(i)}, \mathbf{x}^{(i)}) || p_\theta(\hat{y}_t^{(i)} | \hat{\mathbf{y}}_{<t}^{(i)}, \mathbf{x}^{(i)})$$

$$\hat{\mathbf{H}}^{(i)} = \overline{\mathbf{H}}^{(i)} - \eta \frac{\mathbf{f}}{||\mathbf{f}||_2}, \text{ where } \mathbf{f} = \nabla_{\overline{\mathbf{H}}_1^{(i)}} \mathcal{L}_{KL}(\theta)$$

① Semantic space↓

$$\boxed{\mathcal{L}_{cont-pos}(\theta)} = \sum_{i=1}^{N} \log \frac{\exp(\text{sim}(\mathbf{z}_\mathbf{x}^{(i)}, \boxed{\hat{\mathbf{z}}_\mathbf{y}^{(i)}})/\tau)}{\sum_{\mathbf{z}_\mathbf{y}^{(k)} \in S \cup \{\tilde{\mathbf{z}}_\mathbf{y}^{(i)}\}} \exp(\text{sim}(\mathbf{z}_\mathbf{x}^{(i)}, \mathbf{z}_\mathbf{y}^{(k)})/\tau)}, \text{ where } \hat{\mathbf{z}}_\mathbf{y}^{(i)} = \xi(\hat{\mathbf{H}}^{(i)}; \theta)$$

Imposter and Distant-Target



(a) Contrastive Learning with perturbation (b) Generation of Imposters (c) Generation of Distant-Targets

Figure 3: **Generation of imposters and distant-targets with perturbation.** (a) We add small perturbation δ_t to $\mathbf{h}_t$ for $\tilde{\mathbf{z}}_{\mathbf{y}}$ so that its conditional likelihood is minimized to generate an invalid sentence. (b) We add large perturbation ζ_t to $\mathbf{h}_t$ for $\hat{\mathbf{z}}_{\mathbf{y}}$ by maximizing the distance from $\mathbf{z}_{\mathbf{x}}$, the representation of source sentence but enforcing its likelihood high to preserve the original semantics.

CLAPS Objective

$$\max_{\theta} \underline{\mathcal{L}_{MLE}(\theta)} - \underline{\alpha\mathcal{L}_{KL}(\theta)} + \underline{\beta\{\mathcal{L}_{cont-neg}(\theta) + \mathcal{L}_{cont-pos}(\theta)\}}$$

Experiment

Experiment

- **Machine Translation**
  - WMT16 Romanian-English parallel corpus (WMT’16 RO-EN)
  - T5-small model
- **Text Summarization**
  - XSum dataset
  - T5-small model
- **Question Generation**
  - SQuAD dataset
  - T5-small model

Table 4: The statistics and the data source of WMT’16 RO-EN, Xsum, and SQuAD.

Datasets	**Train (#)**	**Valid (#)**	**Test (#)**	**Source**
WMT’16 RO-EN	610,320	1,999	1,999	Romanian-English Parallel corpus.
Xsum	204,045	11,332	11,334	One-sentence summary of BBC news articles.
SQuAD	86,588	5,192	5,378	Crowd-sourced questions from Wikipedia paragraph

Results

Method	Aug.	BLEU-1	BLEU-2	BLEU-3	BLEU-4	BLEU	F1/EM
			Question Generation - SQuAD				
Harvesting-QG	-	-	-	20.90	15.16	-	66.05/54.62
T5-MLE	-	41.26	30.30	23.38	18.54	21.00	67.64/55.91
α-T5-MLE ($\alpha = 0.7$)	-	40.82	29.79	22.84	17.99	20.50	68.04/56.30
α-T5-MLE ($\alpha = 2.0$)	-	37.35	27.20	20.79	16.36	18.41	65.74/54.76
T5-SSMBA	Pos.	41.67	30.59	23.53	18.57	21.07	68.47/56.37
T5-WordDropout Contrastive	Neg.	41.37	30.50	23.58	18.71	21.19	68.16/56.41
R3F	-	41.00	30.15	23.26	18.44	20.97	65.84/54.10
T5-MLE-contrastive	-	41.23	30.28	23.33	18.45	20.91	67.32/55.25
T5-CLAPS w/o negative	Pos.	41.87	30.93	23.90	18.92	21.38	-
T5-CLAPS w/o positive	Neg.	41.65	30.69	23.71	18.81	21.25	68.26/56.41
T5-CLAPS	Pos.+Neg.	**42.33**	**31.29**	**24.22**	**19.19**	**21.55**	**69.01/57.06**
ERNIE-GEN (Xiao et al., 2020)	-	-	-	-	**26.95**	-	-
Info-HCVAE (Lee et al., 2020)	-	-	-	-	-	-	**81.51/71.18**
			Machine Translation - WMT'16 RO-EN				
Transformer	-	50.36	37.18	28.42	22.21	26.17	
Scratch-T5-MLE	-	51.62	37.22	27.26	21.13	25.34	
Scratch-CLAPS	Pos.+Neg.	53.42	39.57	30.24	23.59	27.61	
T5-MLE	-	57.76	44.45	35.12	28.21	32.43	
α-T5-MLE ($\alpha = 0.7$)	-	57.63	44.23	33.84	27.90	32.14	
α-T5-MLE ($\alpha = 2.0$)	-	56.03	42.59	33.29	26.45	30.72	
T5-SSMBA	Pos.	58.23	44.87	35.50	28.48	32.81	
T5-WordDropout Contrastive	Neg.	57.77	44.45	35.12	28.21	32.44	
R3F	-	58.07	44.86	35.57	28.66	32.99	
T5-MLE-contrastive	-	57.64	44.12	34.74	27.79	32.03	
T5-CLAPS w/o negative	Pos.	58.81	45.52	36.20	29.23	33.50	67.58/55.91
T5-CLAPS w/o positive	Neg.	57.90	44.60	35.27	28.34	32.55	
T5-CLAPS	Pos.+Neg.	**58.98**	**45.72**	**36.39**	**29.41**	**33.96**	
Conneau & Lample (2019)	-	-	-	-	-	**38.5**	

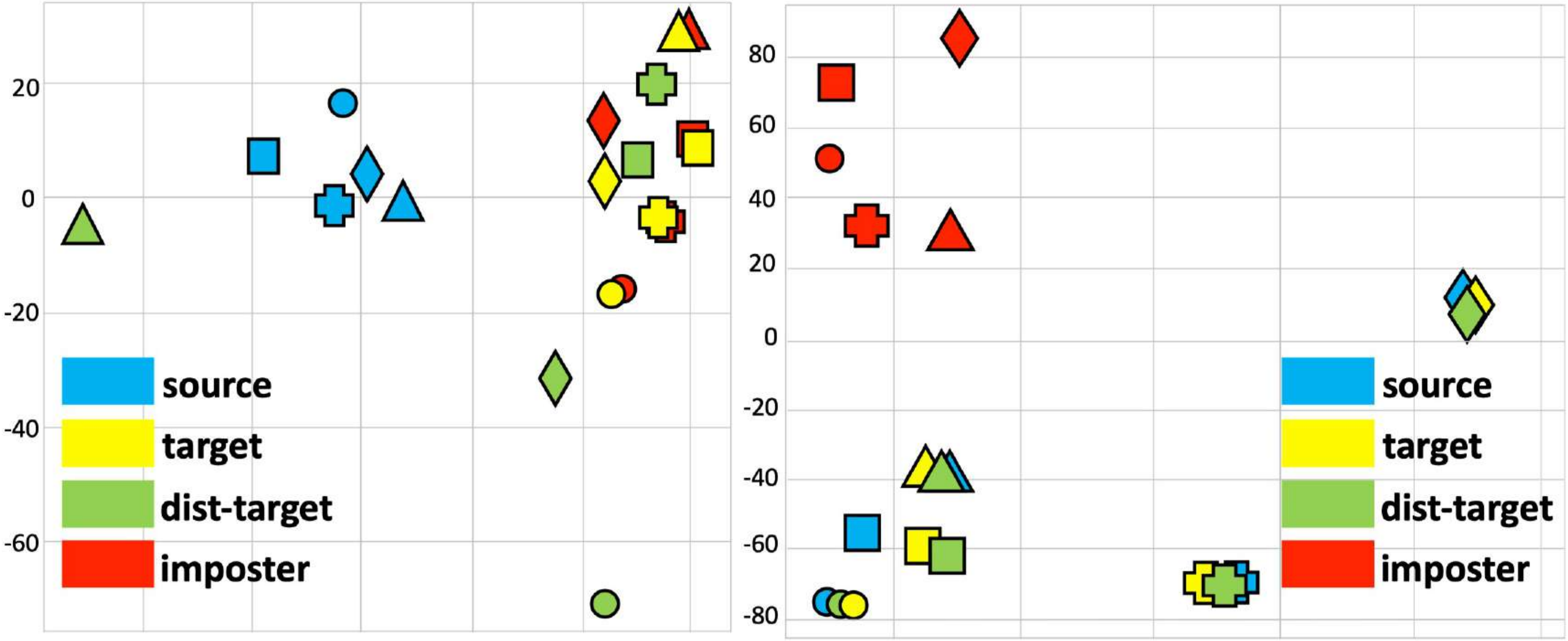
Table 1: BLEU scores on WMT'16 RO-EN and SQuAD for machine translation and question generation. EM/F1 scores with BERT-base QA model for question generation.

Results

Table 2: Rouge and Meteor on Xsum test set for text summarization.

Method	Aug.	Rouge-1	Rouge-2	Rouge-L	METEOR
Text Summarization - XSum					
PTGEN-COVG	-	28.10	8.02	21.72	12.46
CONVS2S	-	31.89	11.54	25.75	13.20
Scratch-T5-MLE	-	31.44	11.07	25.18	13.01
Stcratch-CLAPS	Pos.+Neg.	33.52	12.59	26.91	14.18
T5-MLE	-	36.10	14.72	29.16	15.78
α-T5-MLE ($\alpha = 0.7$)	-	36.68	15.10	29.72	15.78
α-T5-MLE ($\alpha = 2.0$)	-	34.18	13.53	27.35	14.51
T5-SSMBA	Pos.	36.58	14.81	29.68	15.38
T5-WordDropout Contrastive	Neg.	36.88	15.11	29.79	15.77
R3F	-	36.96	15.12	29.76	15.68
T5-MLE-contrastive	-	36.34	14.81	29.41	15.85
T5-CLAPS w/o negative	Pos.	37.49	15.31	30.42	16.36
T5-CLAPS w/o positive	Neg.	37.72	15.49	**30.74**	16.06
T5-CLAPS	Pos.+Neg.	**37.89**	**15.78**	30.59	**16.38**
PEGASUS (Zhang et al., 2020)	-	**47.21**	**24.56**	**39.25**	-

Visualization



(a) Finetune without contrastive learning (b) Finetune with contrastive learning

Qualitative Examples

(MT) Lupta lui Hilary a fost mai atractivă.
=>(**GT**): Hillary's **struggle** was more attractive
=>(**Dist.**): Hilary's **fight** was more attractive
=>(**Imp.**): **Thearies'** battle fight has attractive appealing

(QG) … Von Miller … recording **five** solo tackles, …
=>(**GT**): How many solo tackles did Von Miller **make** at Super Bowl 50?
=>(**Dist.**): How many solo tackles did Von Miller **record** at Super Bowl 50?
=>(**Imp.**): What much tackle **did was** Miller record at Super Bowl 50?

(Sum.) Pieces from the board game … have been found in … China. …
=>(**GT**): An ancient board game has been **found** in a Chinese Tomb.
=>(**Dist.**): An ancient board game has been **discovered** in a Chinese Tomb.
=>(**Imp.**): America's gained vast Africa **most well geographical** countries, 22

Table 3: Greedy decoding from hidden representation of imposters and distant-targets. The answer span is highlighted for QG.

Human Evaluation

- Conduct a human evaluation of the 20 summaries and 20 questions generated by our CLAPS and T5-MLE trained for text summarization and QG task.
- 20 human judges perform blind quality assessment

- For text summarization, 70% of the human annotators chose the sentences generated by our model as better than the baseline, and
- For QG, 85% favored the sentences generated by our model over that of the baseline.

Conclusion

Conclusion

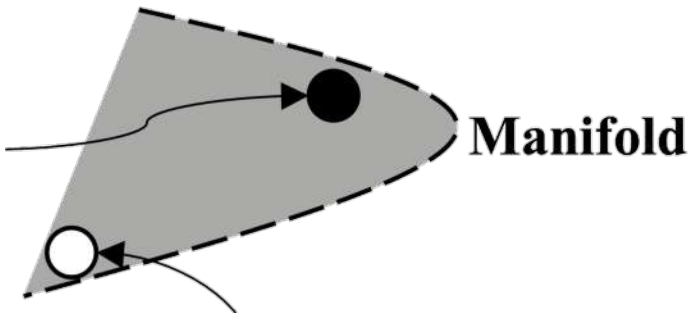
- Contrastive learning framework to mitigate the exposure bias problem.
- New principled approach to automatically construct “hard” negative and positive examples.
- Method improved the performance of seq2seq model on machine translation, question generation, and text summarization tasks.

Last of the Last

Manifold?

- 黎曼开始了关于延展性、维数、延展性数量化的讨论，他给了这些多度延展的量一个名称，德文写作mannigfaltigkeit，英文翻译为Manifold，英文字面意思可以理解为“多种多样”。

- 中国第一个拓扑学家江泽涵（北大教授）把这个词翻译为“流形”，取自
  - 文天祥《正气歌》，“天地有正气，杂然赋流形”，
  - 而其原始出处为《易经》，“大哉乾元，万物资始，乃统天。云行雨施，品物流形。”



https://www.zhihu.com/question/30553302/answer/311788410

Thanks~