BERT-EMD: Many-to-Many Layer Mapping for BERT Compression with Earth Mover's Distance

Jianquan Li[1]*, Xiaokang Liu[1]*, Honghong Zhao[1], Ruifeng Xu[2], Min Yang[3]†, Yaohong Jin[1]

[1]Beijing Ultrapower Software Co.,Ltd., China

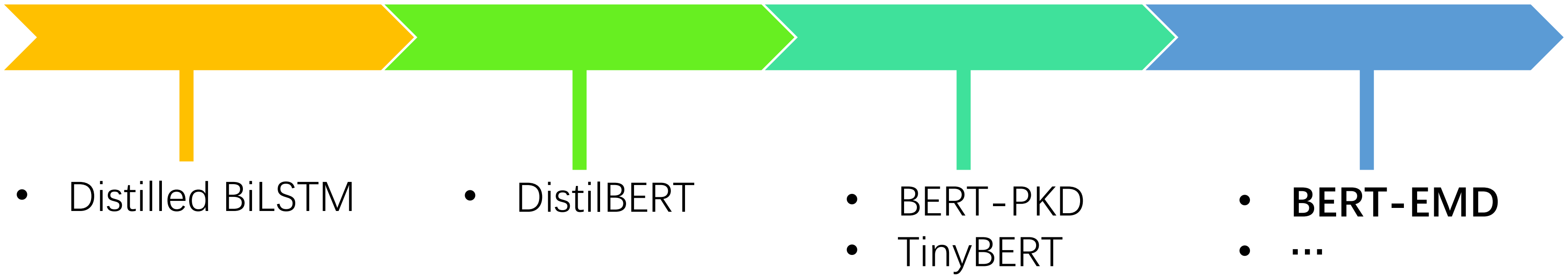
[2]Harbin Institute of Technology (Shenzhen), China

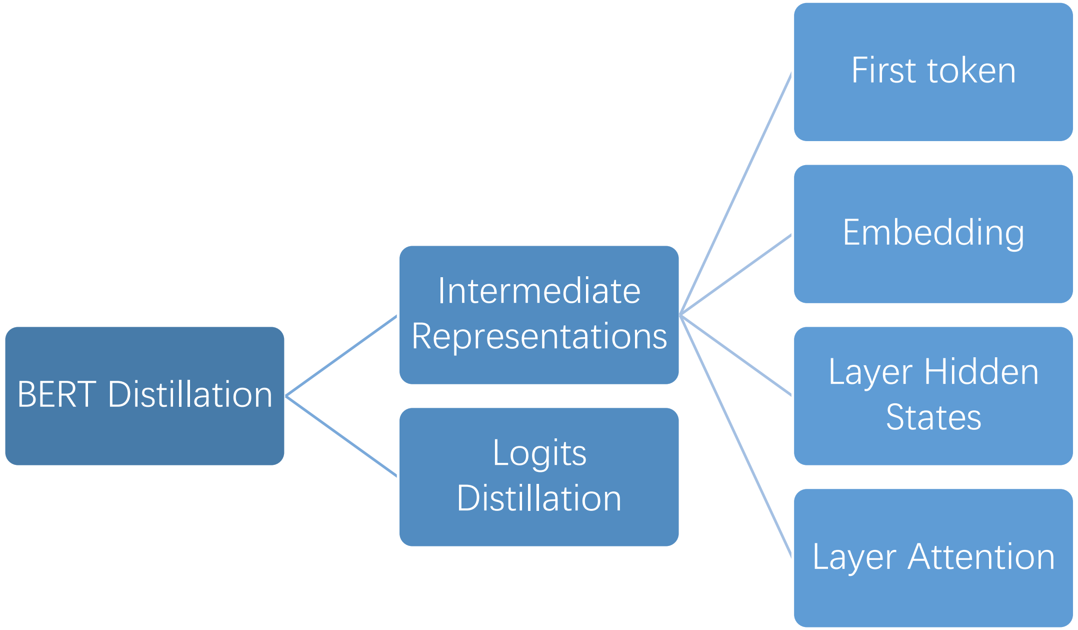
[3]Shenzhen Key Laboratory for High Performance Data Mining,
Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, China

{lijianquan2,liuxiaokang1,zhaohonghong1,jinyaohong}@ultrapower.com.cn

xuruifeng@hit.edu.cn, min.yang@siat.ac.cn

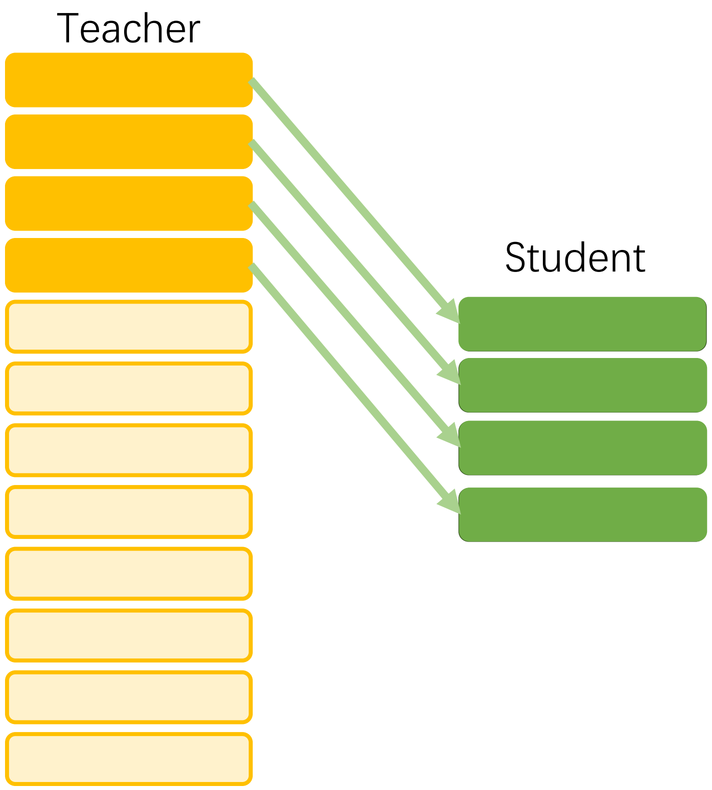
BERT Distillation Timeline





- Transformer to transformer
- KD on finetune and pretraining
- Intermediate representations

Layer mapping functions

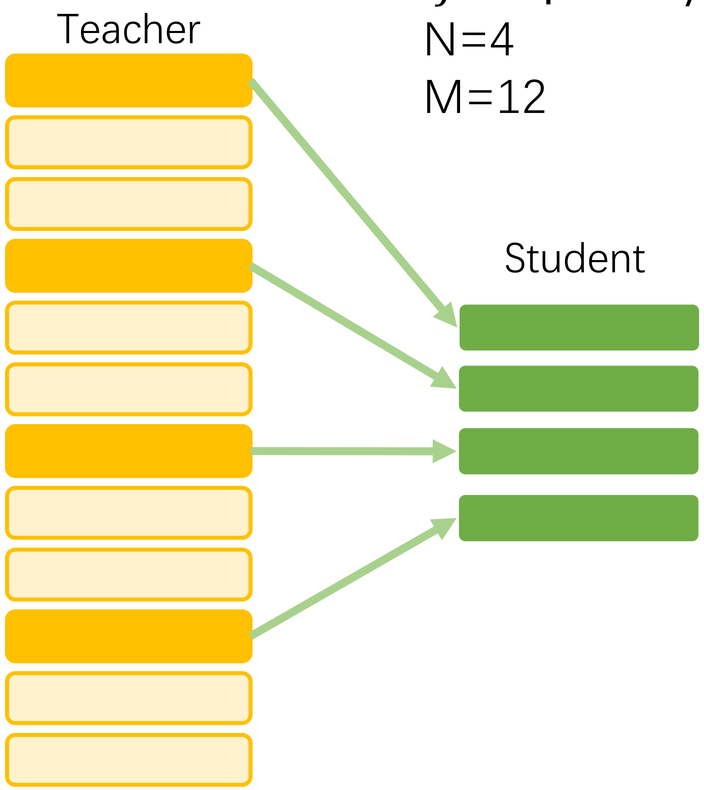


Last strategy

$j = \lceil i \times N/M \rceil$

N=4

M=12



Skip strategy

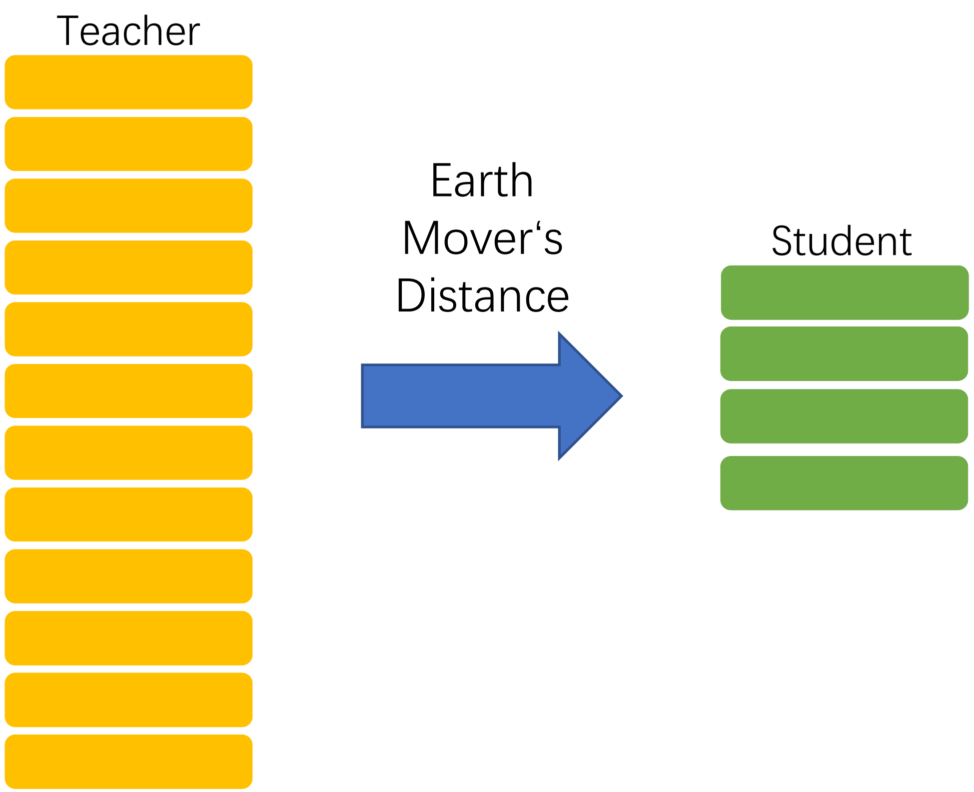
- cannot take full advantage of the teacher network
- Hard to find the best mapping function for different tasks
- Should the weight be the same?

$C_{12}^{4} = 495$ in Total

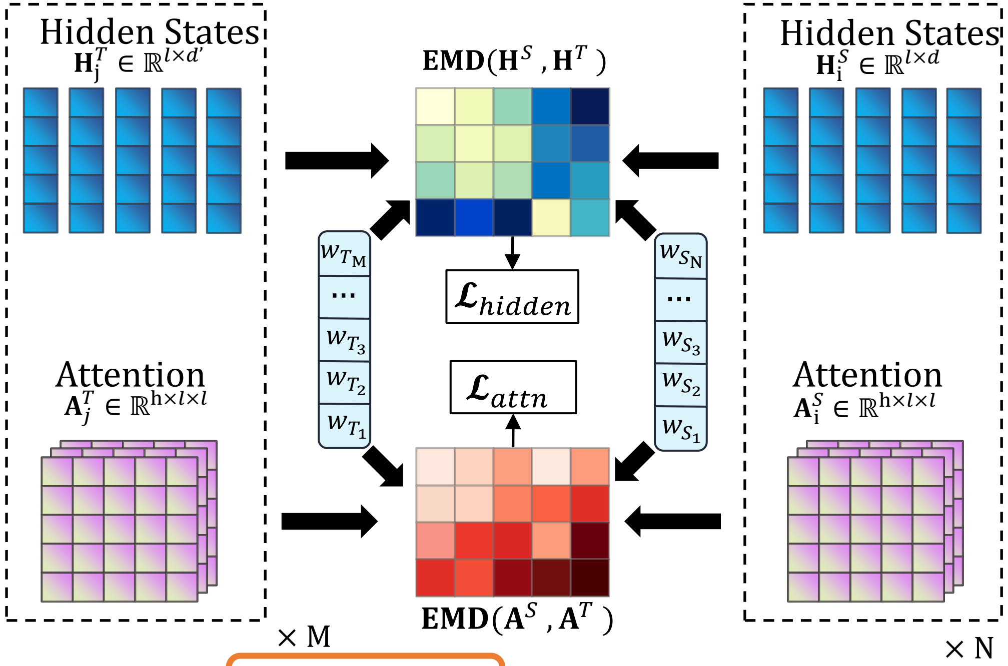
Model	SST-2	MRPC	QQP	MNLI-m	MNLI-mm	QNLI	RTE
BERT_6 (PKD-Last)	91.9	**85.1**/79.5	70.5/**88.9**	80.9	**81.0**	88.2	65.0
BERT_6 (PKD-Skip)	**92.0**	85.0/**79.9**	**70.7/88.9**	**81.5**	**81.0**	**89.0**	**65.5**

Result in BERT-PKD Paper

Single Layer distance to Model Distance



- many-to-many layer mapping
- leverage EMD to formulate the distance between the teacher and student networks



$$w_{S_i} = \frac{1}{N}$$
$$w_{T_j} = \frac{1}{M}$$

Attention

$$A^T = \{(A_1^T, w^{A_{T_1}}), \ldots, (A_M^T, w^{A_{T_M}})\}$$
$$A^S = \{(A_1^S, w^{A_{S_1}}), \ldots, (A_N^S, w^{A_{S_N}})\}$$
$$\text{Distance:}\quad d^{A_{ij}} = \mathrm{MSE}(A_i^S, A_j^T)$$

Hidden States

$$H^S = \{(H_1^S, w^{H_{S_1}}), \ldots, (H_N^S, w^{H_{S_N}})\}$$
$$H^T = \{(H_{1^T}, w^{H_{T_1}}), \ldots, (H_{M^T}, w^{H_{T_M}})\}$$
$$\text{Distance:}\quad d^{H_{ij}} = \mathrm{MSE}(H_i^S \boldsymbol{W}_h, H_j^{\mathrm{T}})$$

Flow

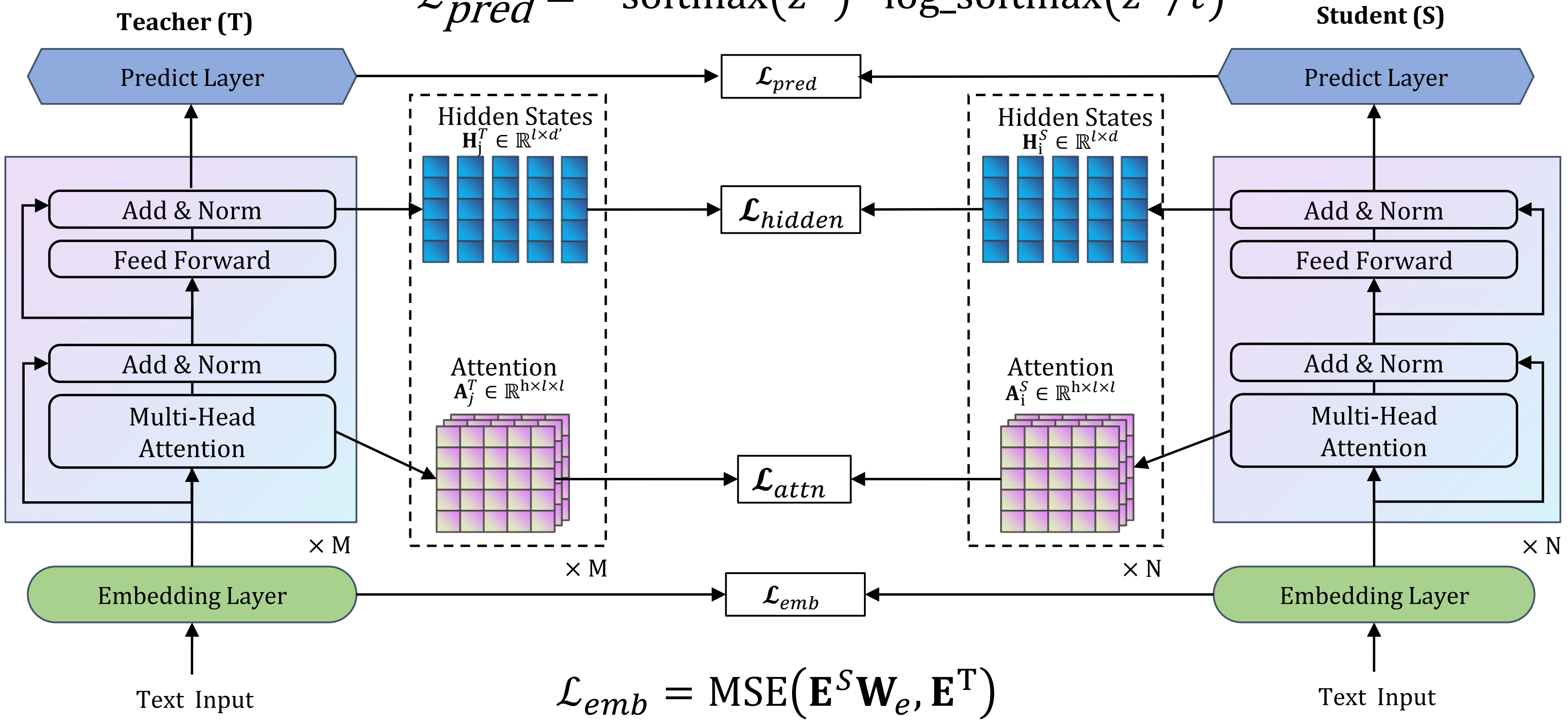
$$\boldsymbol{F}^{\mathrm{A}} = [f_{ij}^{\boldsymbol{A}}]$$
$$\boldsymbol{F}^{\mathrm{H}} = [f_{ij}^{\boldsymbol{H}}]$$
$$\mathrm{WORK}(H^T, H^S, F^H) = \sum_{i=1}^{M}\sum_{j=1}^{N} f^{H_{ij}} d^{H_{ij}}$$
$$s.t.\ f^{H_{ij}} \geq 0 \qquad 1 \leq i \leq M, \sim 1 \leq j \leq N$$
$$\sum_{j=1}^{N} f^{H_{ij}} \leq w^{H_{T_i}} \qquad 1 \leq i \leq M$$
$$\sum_{i=1}^{M} f^{H_{ij}} \leq w^{H_{S_j}} 1 \leq j \leq N$$
$$\sum_{i=1}^{M}\sum_{j=1}^{N} f^{H_{ij}} = \min\left(\sum_{i}^{M} w_{T_i}^{H}, \sum_{i}^{N} w_{S_i}^{H}\right)$$

Distance and Loss

$$\mathrm{EMD}(\boldsymbol{A^S}, \boldsymbol{A^T}) = \frac{\sum_{i=1}^{M}\sum_{j=1}^{N} f^{\boldsymbol{A}_{ij}} d^{\boldsymbol{A}_{ij}}}{\sum_{i=1}^{M}\sum_{j=1}^{N} f^{\boldsymbol{A}_{ij}}}$$
$$\mathrm{EMD}(H^S, H^T) = \frac{\sum_{i=1}^{M}\sum_{j=1}^{N} f^{H_{ij}} d^{H_{ij}}}{\sum_{i=1}^{M}\sum_{j=1}^{N} f^{H_{ij}}}$$
$$\mathcal{L}_{attn} = \mathrm{EMD}(A^S, A^T)$$
$$\mathcal{L}_{hidden} = \mathrm{EMD}(\boldsymbol{H^S}, \boldsymbol{H^T})$$

Network

$$\mathcal{L}_{pred} = -\mathrm{softmax}(z^T) \cdot \mathrm{log\_softmax}(z^S/t)$$



$$\mathcal{L}_{emb} = \mathrm{MSE}(\mathbf{E}^S \mathbf{W}_e, \mathbf{E}^T)$$

$$\mathcal{L}_{distill} = \beta(\mathcal{L}_{emb} + \mathcal{L}_{attn} + \mathcal{L}_{hidden}) + \mathcal{L}_{pred}$$

Cost attention weight update method

Target: Reducing transfer cost. Teacher model weight as example:

- Step 1 transferring cost between each teacher and student layers (unit transferring cost).
  - $\overline{C_{T_i}^{A}} = \frac{\sum_{j=1}^{N} d^{A_{ij}} f^{A_{ij}}}{w_{T_i}}$
  - $\overline{C_{T_i}^{H}} = \frac{\sum_{j=1}^{N} d^{H_{ij}} f^{H_{ij}}}{w_{T_i}}$
- Step 2. update weight based on the learned unit transferring cost:
  - $\overline{w_{T_i}^{A}} = \frac{\sum_{j=1}^{m} \overline{C_j^{A}}}{\overline{C_{T_i}^{A}}}$
  - $\overline{w_{T_i}^{H}} = \frac{\sum_{j=1}^{m} \overline{C_j^{H}}}{\overline{C_{T_i}^{H}}}$
- Step 3 softmax and average，get new weight:
  - $\overline{w_{T_i}} = \frac{1}{2}\left(\frac{e^\wedge(\overline{w_{T_i}^{A}}/\tau)}{\sum_{j\in M}\sum e^\wedge(\overline{w_{T_j}^{A}}/\tau)} + \frac{e^\wedge(\overline{w_{T_i}^{H}}/\tau)}{\sum_{j\in M}\sum e^\wedge(\overline{w_{T_j}^{H}}/\tau)}\right)$

Experiments

Model	Params Num	Inference Time	MNLI-m (393k)	MNLI-mm (393k)	QQP (364k)	SST-2 (67k)	CoLA (8.5k)	QNLI (108k)	MRPC (3.5k)	RTE (2.5k)	STS-b (5.7k)	AVE
$\text{BERT}_{\text{BASE12}}$-G	110M	×1	84.6	83.4	71.2	93.5	52.1	90.5	88.9	66.4	85.8	79.60
$\text{BERT}_{\text{BASE12}}$-T	110M	×1	84.4	83.3	71.6	93.4	52.8	90.5	88.1	66.9	85.2	79.58
$\text{BERT}_{\text{SMALL4}}$	14.5M	-	75.4	74.9	66.5	87.6	19.5	84.8	83.2	62.6	77.1	70.18
DistillBERT_4	52.2M	×3.0	78.9	78.0	68.5	**91.4**	**32.8**	85.2	82.4	54.1	76.1	71.93
BERT-PKD_4	52.2M	×3.0	79.9	79.3	**70.2**	89.4	24.8	85.1	82.6	62.3	79.8	72.60
TinyBERT_4	14.5M	×9.4	81.2	80.3	68.9	90.0	25.3	86.2	85.4	63.9	80.4	73.51
BERT-EMD_4	14.5M	×9.4	**82.1**	**80.6**	69.3	91.0	25.6	**87.2**	**87.6**	**66.2**	**82.3**	**74.66**
BERT-PKD_6	66.0M	×1.9	81.5	81.0	70.7	92.0	43.5	89.0	85.0	65.5	81.6	76.61
BERT-of-Theseus_6	66.0M	-	82.4	82.1	71.6	92.2	**47.8**	89.6	87.6	66.2	84.1	78.18
TinyBERT_6	66.0M	×1.9	84.4	83.1	71.3	92.6	46.1	89.8	88.0	69.7	83.9	78.77
BERT-EMD_6	66.0M	× 1.9	**84.7**	**83.5**	**72.0**	**93.3**	47.5	**90.7**	**89.8**	**71.7**	**86.8**	**80.00**

Table 1: Experimental results on the GLUE test set. The subscript within each model name represents the number of Transformer layers. AVE represents the average score over all tasks. $\text{BERT}_{\text{BASE12}}$-G and $\text{BERT}_{\text{BASE12}}$-T indicate the results of the fine-tuned BERT-base from (Devlin et al., 2018) and in our implementation, respectively.

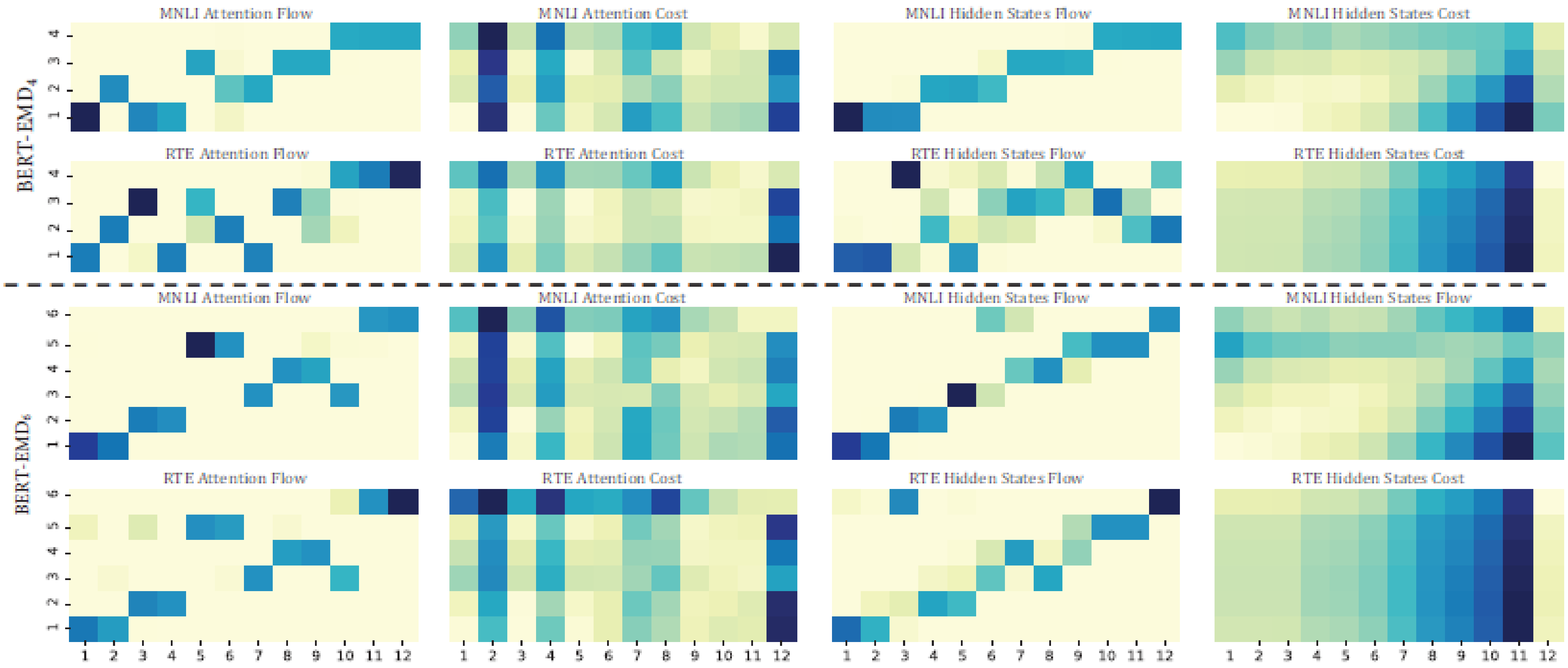


Figure 2: The visualization of flow matrices (**F**) and distance matrices (**D**) in developing BERT-EMD$_4$ (above) and BERT-EMD$_6$ (below) for two examples from MNLI and RTE tasks, respectively. The abscissa represents the Transformer layers of $\mathrm{BERT}_{\mathrm{BASE}_{12}}$, and the ordinate represents the Transformer layers of BERT-EMD$_4$/BERT-EMD$_6$. The color depth represents the values (weights) of the layers.

Future work

- Using more powerful pre-trained language model
- Other weight modeling method
- Pretrain model training with EMD
- Use EMD on the CV model

Thanks

利用自监督学习的开放端故事生成评价方法

关　健

个人主页：https://jianguanthu.github.io/

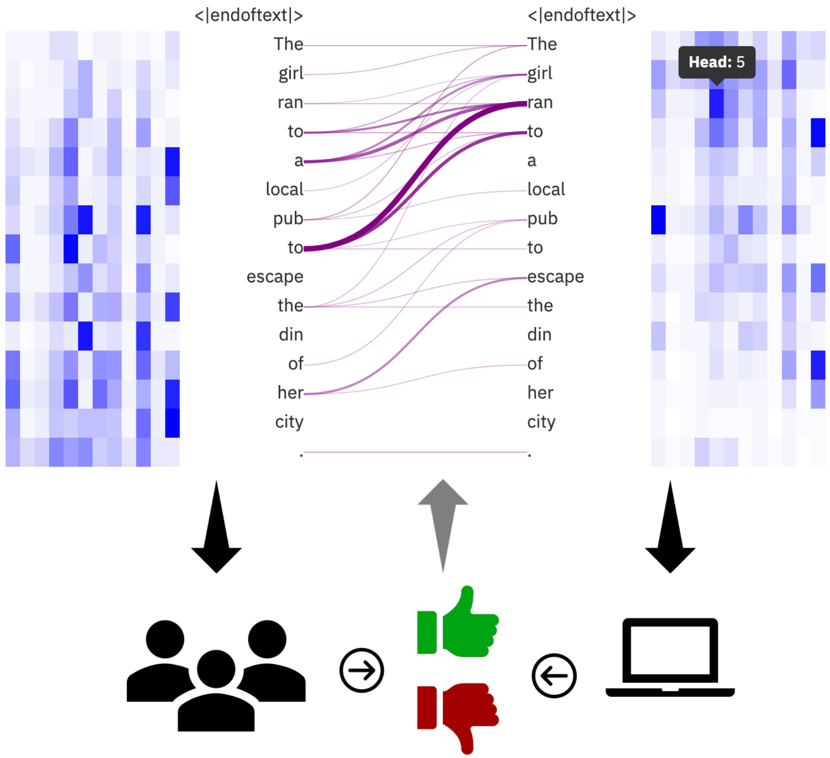
j-guan19@mails.tsinghua.edu.cn| 13051331318

2020年10月30日

介 绍

自然语言生成模型

- 模型框架：LM, Seq2Seq
- 模型结构：RNN, Transformer
- 预训练模型：GPT3, T5, BART



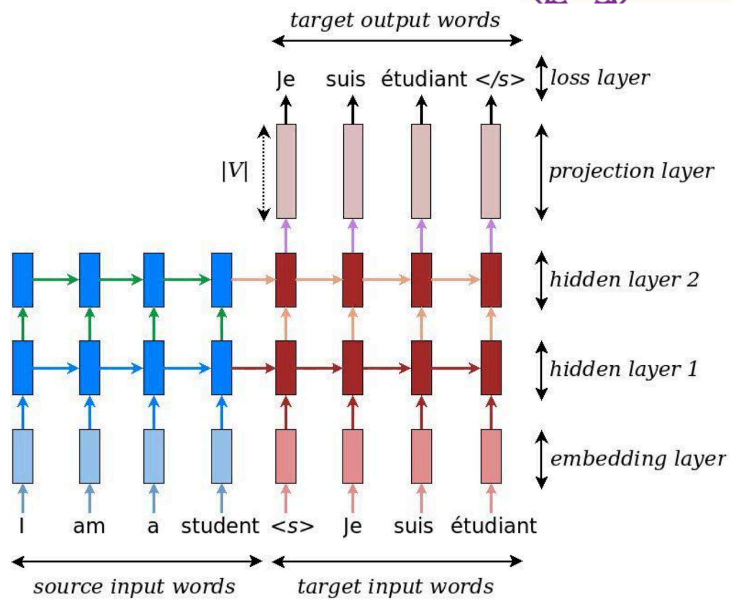
自然语言生成评价

- 意义：指导模型生成，提高生成质量
- 人工评价：耗时、昂贵、难以复现
- 自动评价：快速、低/零成本、易复现

介 绍

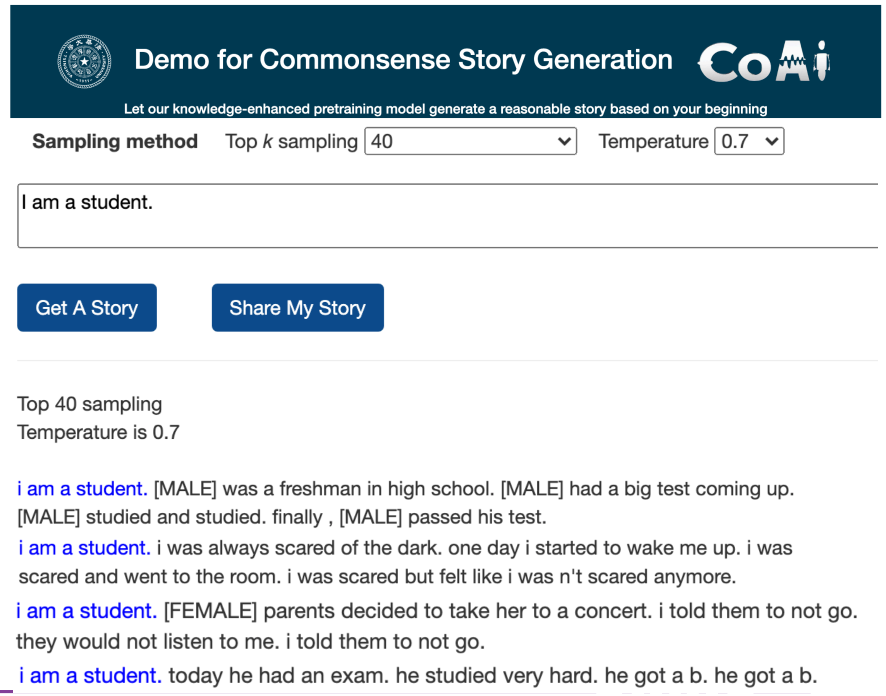
自然语言生成任务

- 受限语言生成 (Constrained NLG)
  - 机器翻译、自动摘要
  - 一对一：输入中包含生成所需的充分信息
  - 评价指标：**BLEU, MoverScore**



Luong et al. Effective Approaches to Attention-based Neural Machine Translation

- 开放端语言生成 (Open-Ended NLG)
  - 开放域对话、常识/科幻/寓言故事生成
  - 一对多：输入中仅仅包含非常有限的信息，同一个输入可能有许多合理的输出



http://coai.cs.tsinghua.edu.cn/static/CommonsenseStoryGen/

介绍

- 开放端语言生成的评价
  - 合理性与是否与参考文本在字面或语义上相似无关
  - 基于判别器的自动评价指标
    - 区分人撰写的文本和机器生成的文本
    - 容易过拟合到特定的数据或模型
  - 学习人工评价的自动评价指标
    - 从人工评价中学习人类偏好
    - 容易过拟合到训练数据上
  - 基于自监督学习的自动评价指标
    - 模仿生成模型自动构造大量的负样本
    - 不依赖于任何人工标注和生成模型
    - 具有与人工评价较高的相关性
    - 对质量/数据迁移具有好的泛化性

Leading Context Jack was at the bar.
Reference By Human He noticed a phone on the floor. He was going to take it to lost and found. But it started ringing on the way. Jack answered it and returned it to the owner's friends.
Sample 1 (Reasonable, B=0.29, M=0.49, U=1.00) On the way out he noticed a phone on the floor. He asked around if anybody owned it. Eventually he gave it to the bartender. They put it into their lost and found box. **Sample 2 (Reasonable, B=0.14, M=0.27, U=1.00)** He had a drinking problem. He kept having more beers. After a while he passed out. When he waked up, he was surprised to find that he lost over a hundred dollars. **Sample 3 (Unreasonable, B=0.20, M=0.35, U=0.00)** He was going to get drunk and get drunk. The bartender told him it was already time to leave. Jack started drinking. Jack wound up returning but cops came on the way home.

B: BLEU; **M**: MoverScore; **U**: UNION

UNION: An Unreferenced Metric for Evaluating Open-ended Story Generation

Jian Guan, Minlie Huang

CS Department, Tsinghua University

经 验 观 察

语言生成模型生成的故事为什么不合理？

- 基于ROCStories，分析381个NLG模型生成的不合理的故事

Mary noticed a bird's nest by her bedroom window.
She decided to climb the tree.
She climbed on the ladder and climbed on the ladder.
She climbed down the ladder and saw her step head.
She reached into her pocket and grabbed the bird's back.

情节重复

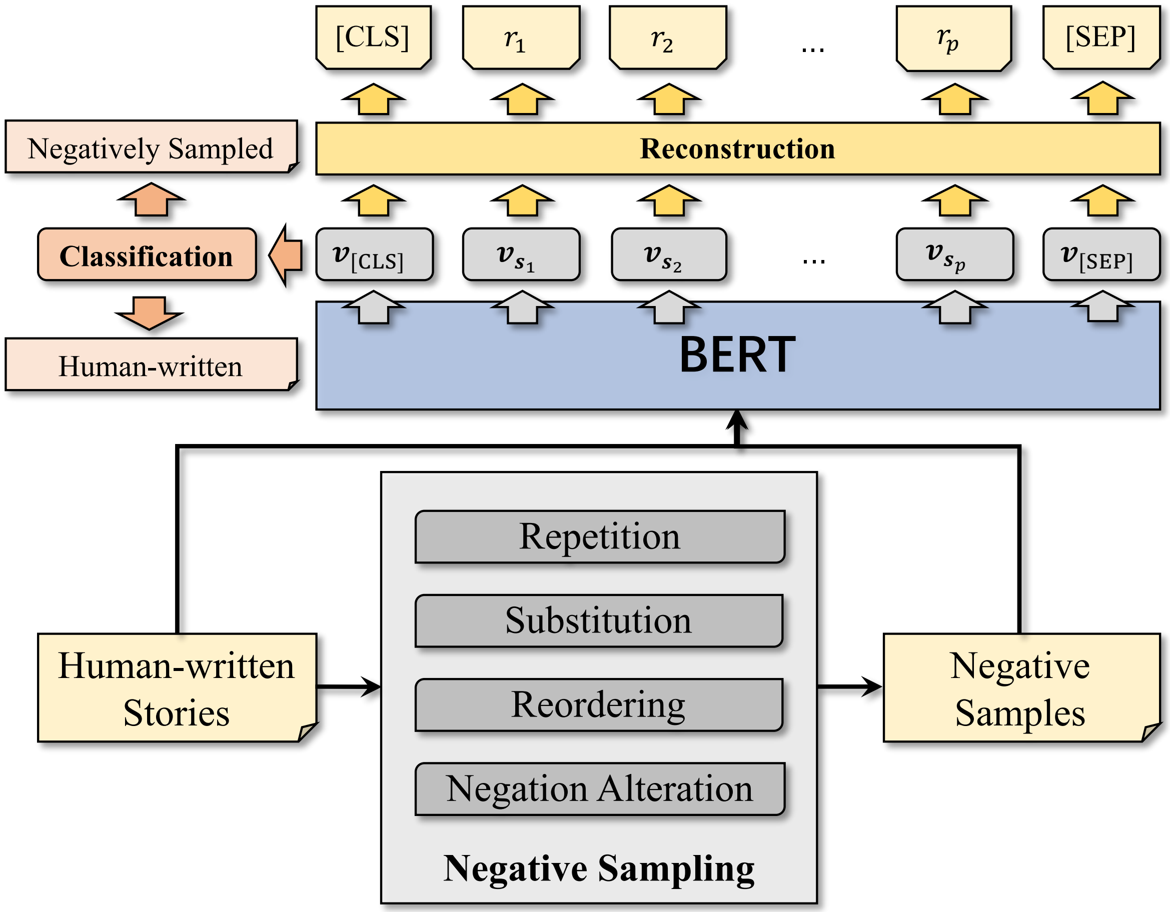
不连贯

逻辑冲突

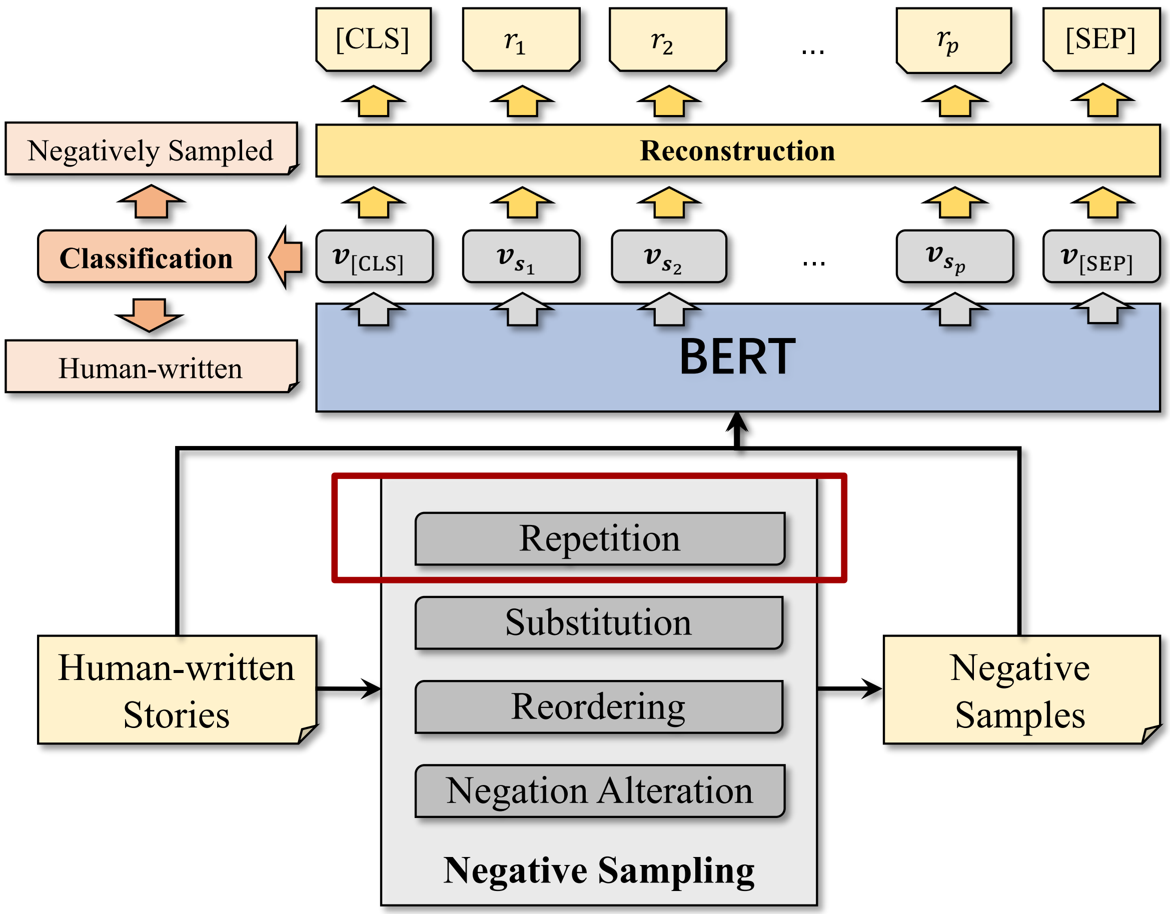
场景混乱

错误类型	情节重复	不连贯	逻辑冲突	场景混乱	其它
占比	44.1%	56.2%	67.5%	50.4%	12.9%

方 法


[CLS]
r_1
r_2
...
r_p
[SEP]
Reconstruction
Negatively Sampled
Classification
$v_{[CLS]}$
v_{s_1}
v_{s_2}
...
v_{s_p}
$v_{[SEP]}$
Human-written
BERT
Repetition
Substitution
Reordering
Negation Alteration
Negative Sampling
Human-written Stories
Negative Samples

方法：负采样



- Repetition
  - N-gram

    The weather was crisp and cool.

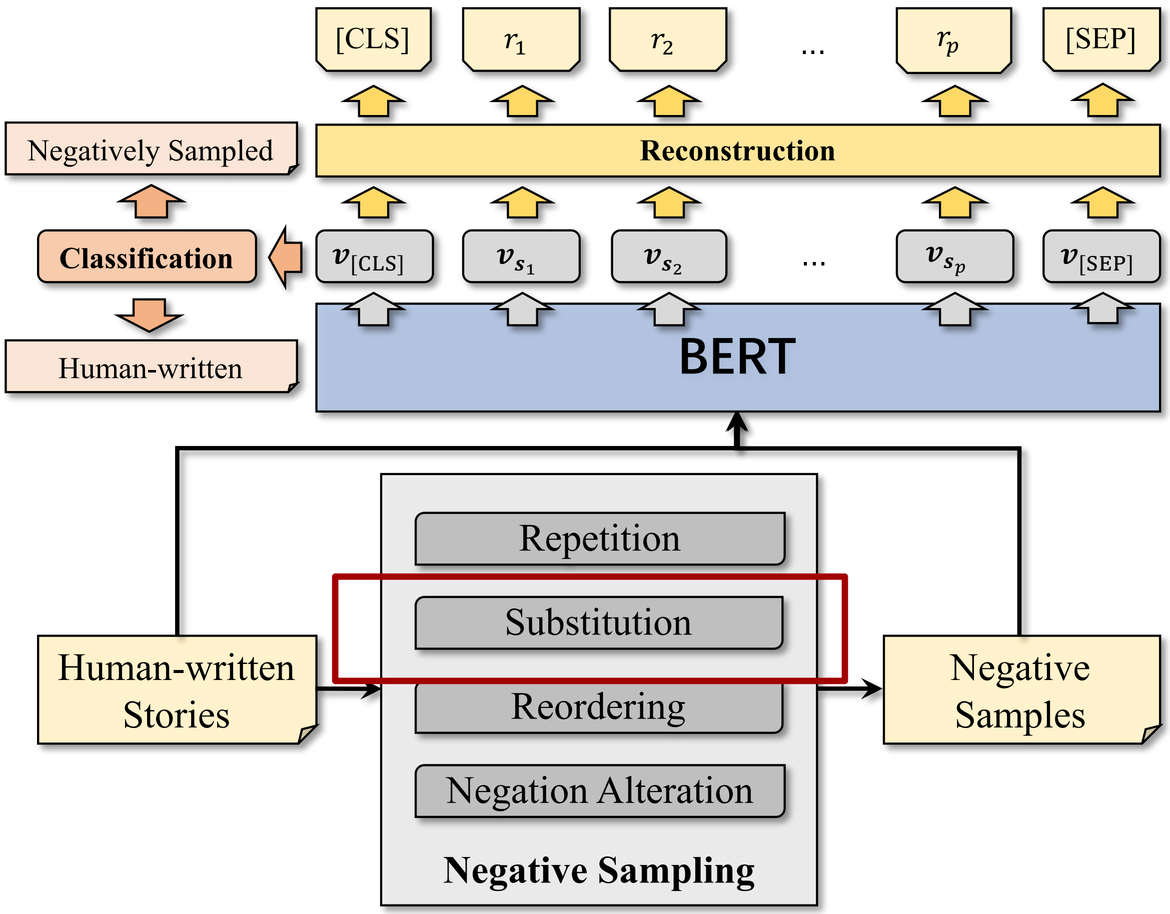
    The weather was crisp **and cool and cool**.

  - Sentence

    The weather was crisp and cool.

    The weather was crisp and cool.
    The weather was crisp and cool.

方法：负采样

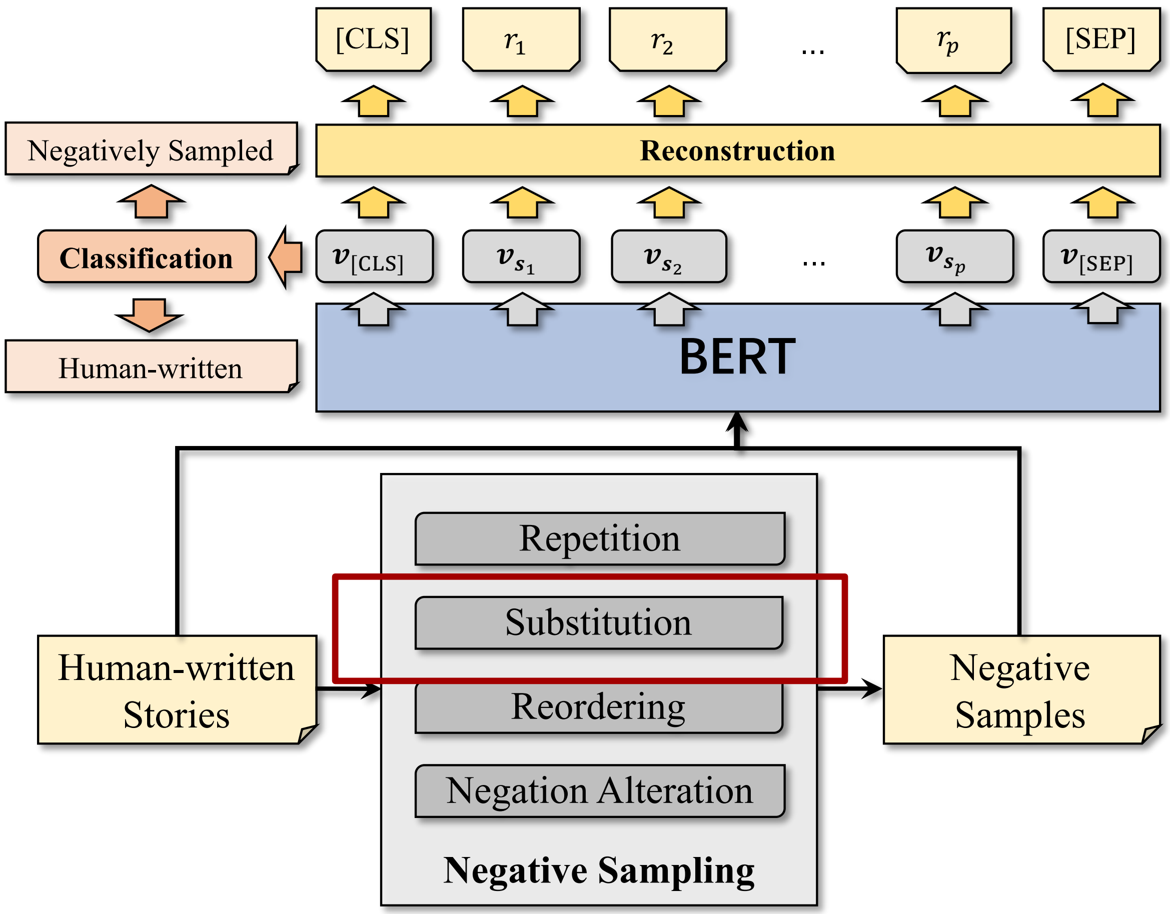


- Substitution
  - Keywords (head/tail in ConceptNet)
    - Antonym:
      - /r/Antonym
      - /r/NotDesires
      - /r/NotCapableOf
      - /r/NotHasProperty

Ken felt good and energetic.

Ken felt **bad** and energetic.

方法：负采样



- Substitution
  - Keywords (head/tail in ConceptNet)
    - Antonym:
      - /r/Antonym
      - /r/NotDesires
      - /r/NotCapableOf
      - /r/NotHasProperty

Ken felt good and energetic.

⇩

Ken felt **bad** and energetic.

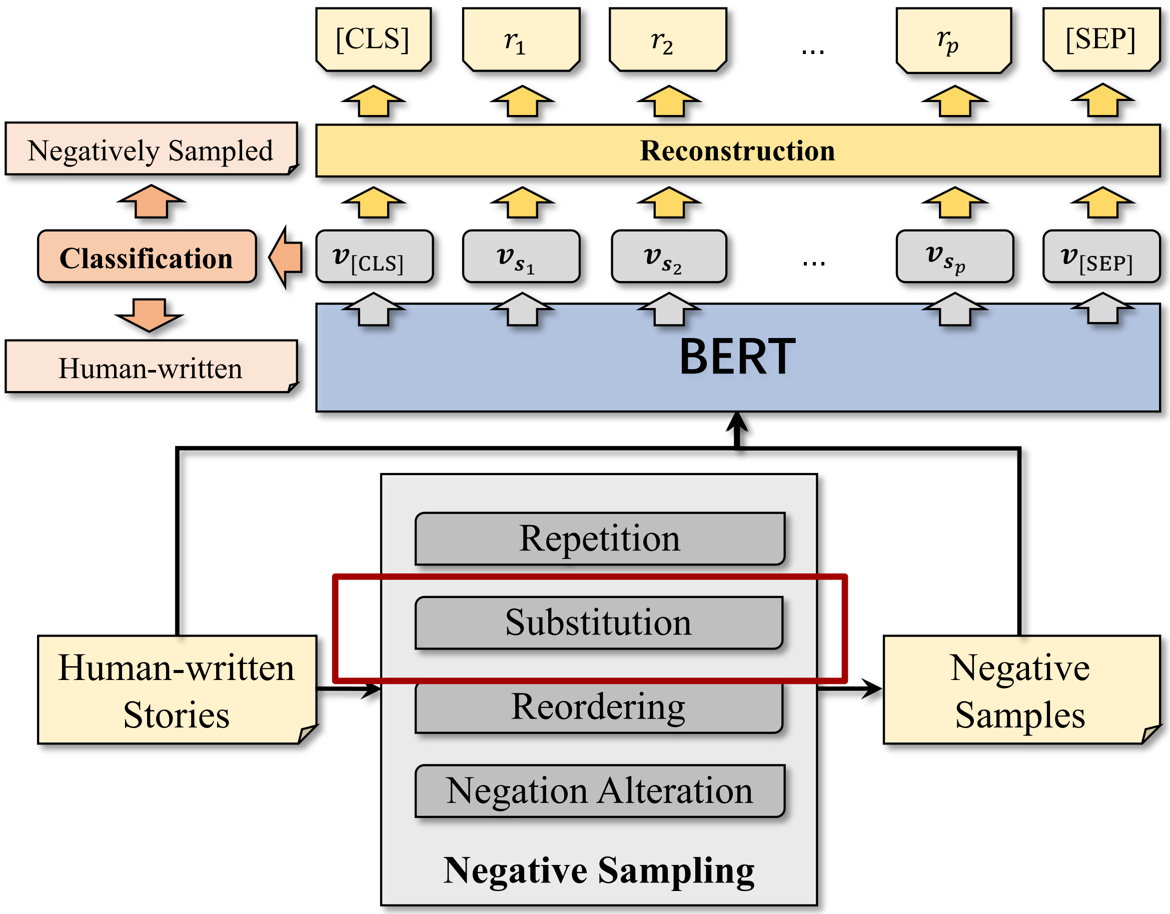
    - Random word with same POS

Ken felt good and energetic.

⇩

Ken **used** good and energetic.

方法：负采样



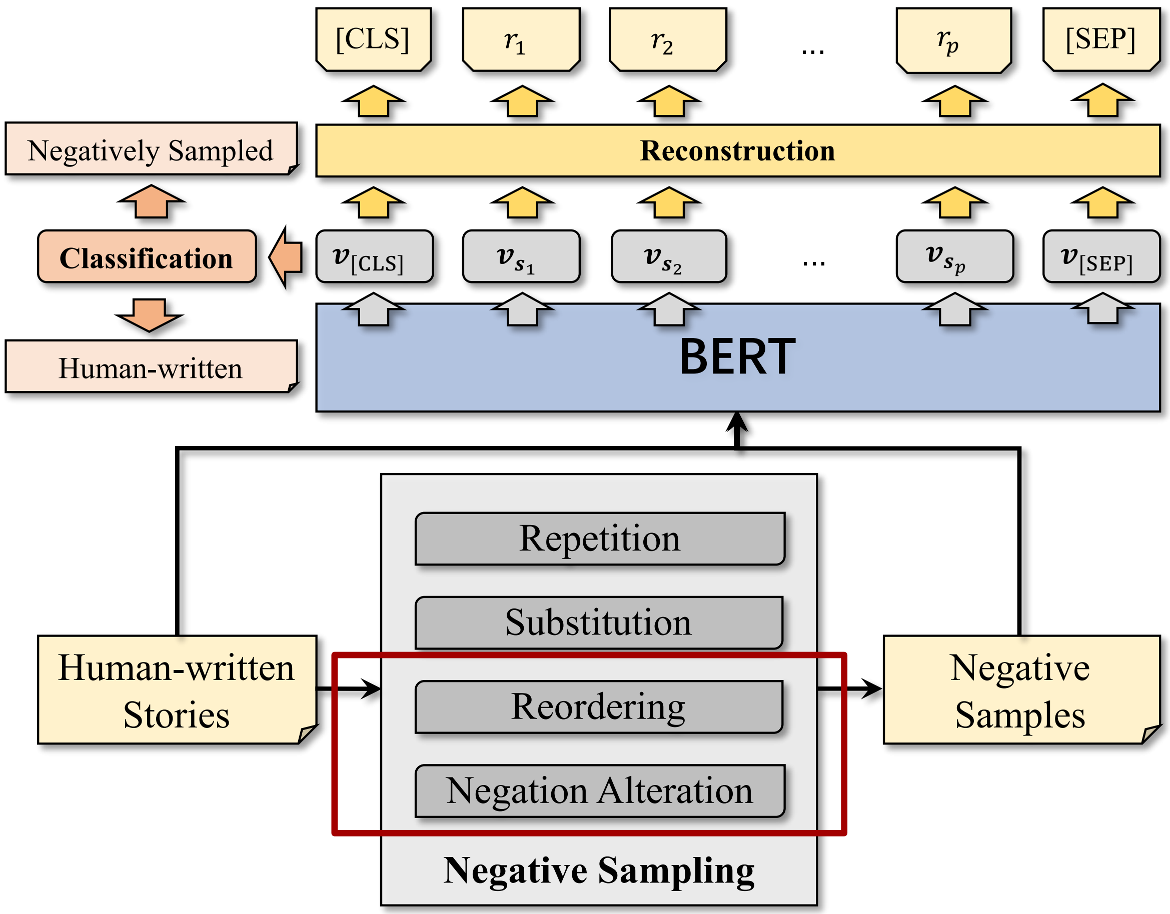
- Substitution
  - Keywords (head/tail in ConceptNet)
    - Antonym:
      - /r/Antonym
      - /r/NotDesires
      - /r/NotCapableOf
      - /r/NotHasProperty

      Ken felt good and energetic.
      ⇩
      Ken felt **bad** and energetic.
    - Random word with same POS

      Ken felt good and energetic.
      ⇩
      Ken **used** good and energetic.
  - Sentence

    Ken felt good and energetic.
    ⇩
    She decided to climb the tree.

方法：负采样

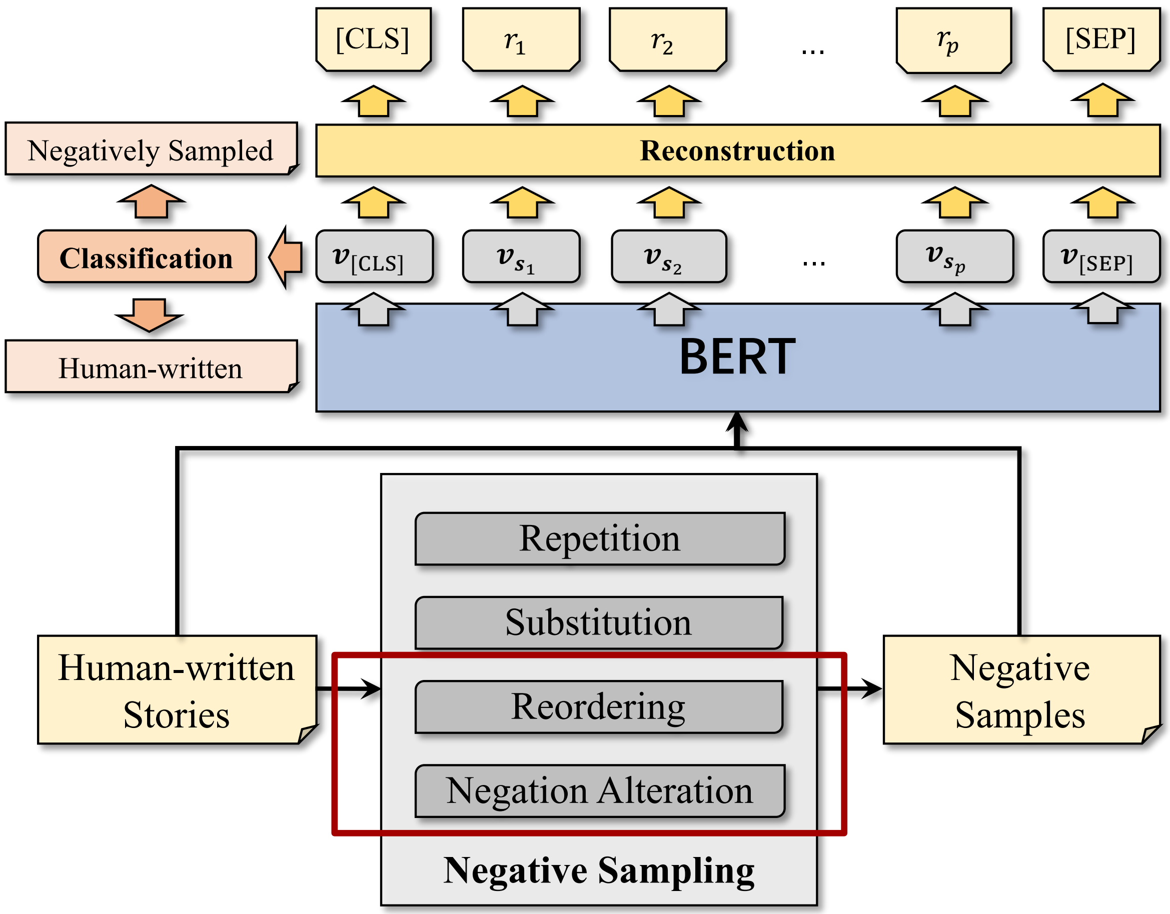


- Reorder
  - Sentences

He decided to keep jogging.
Ken then went several more miles.

Ken then went several more miles.
He decided to keep jogging.

方法：负采样



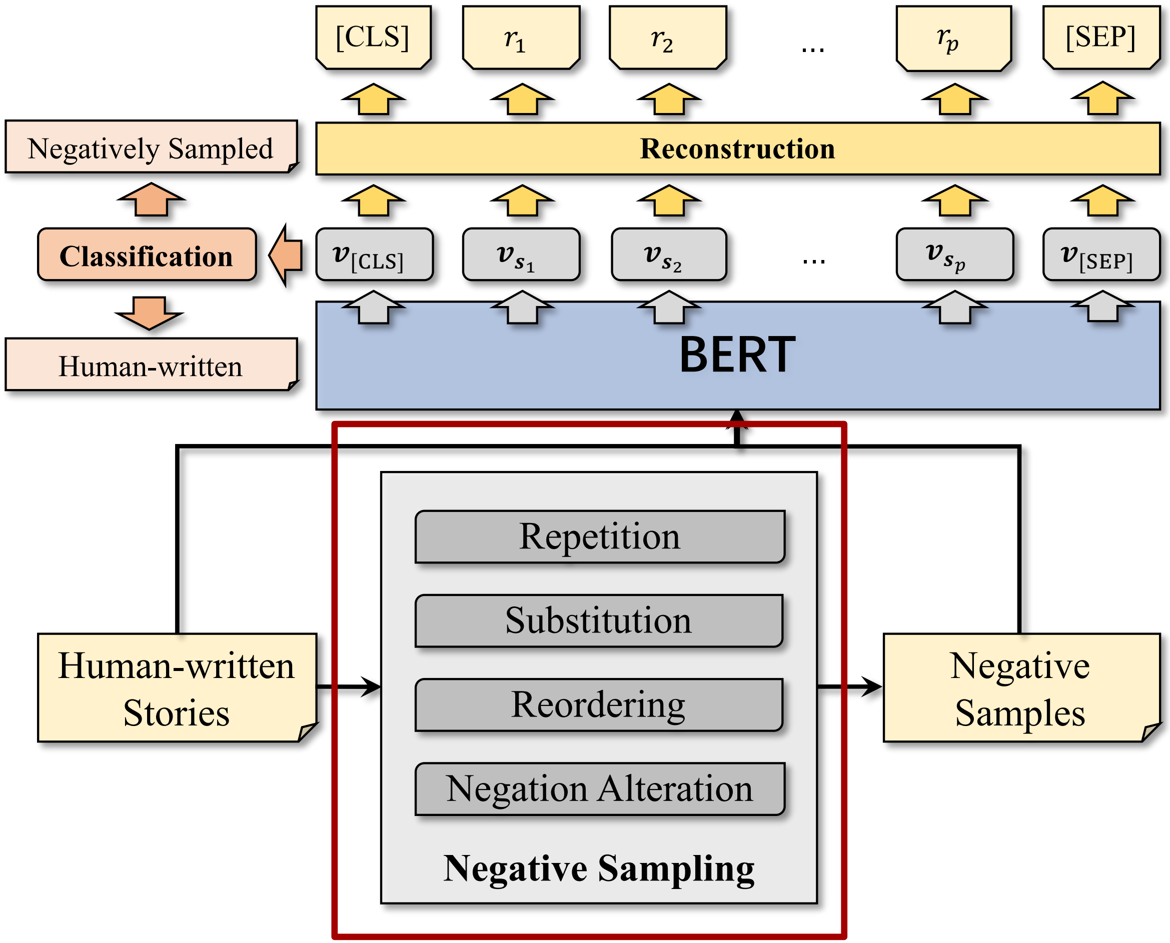
- Reorder
  - Sentences

He decided to keep jogging.
Ken then went several more miles.
⇩
Ken then went several more miles.
He decided to keep jogging.

- Negation Alteration
  - Add/Remove negation words

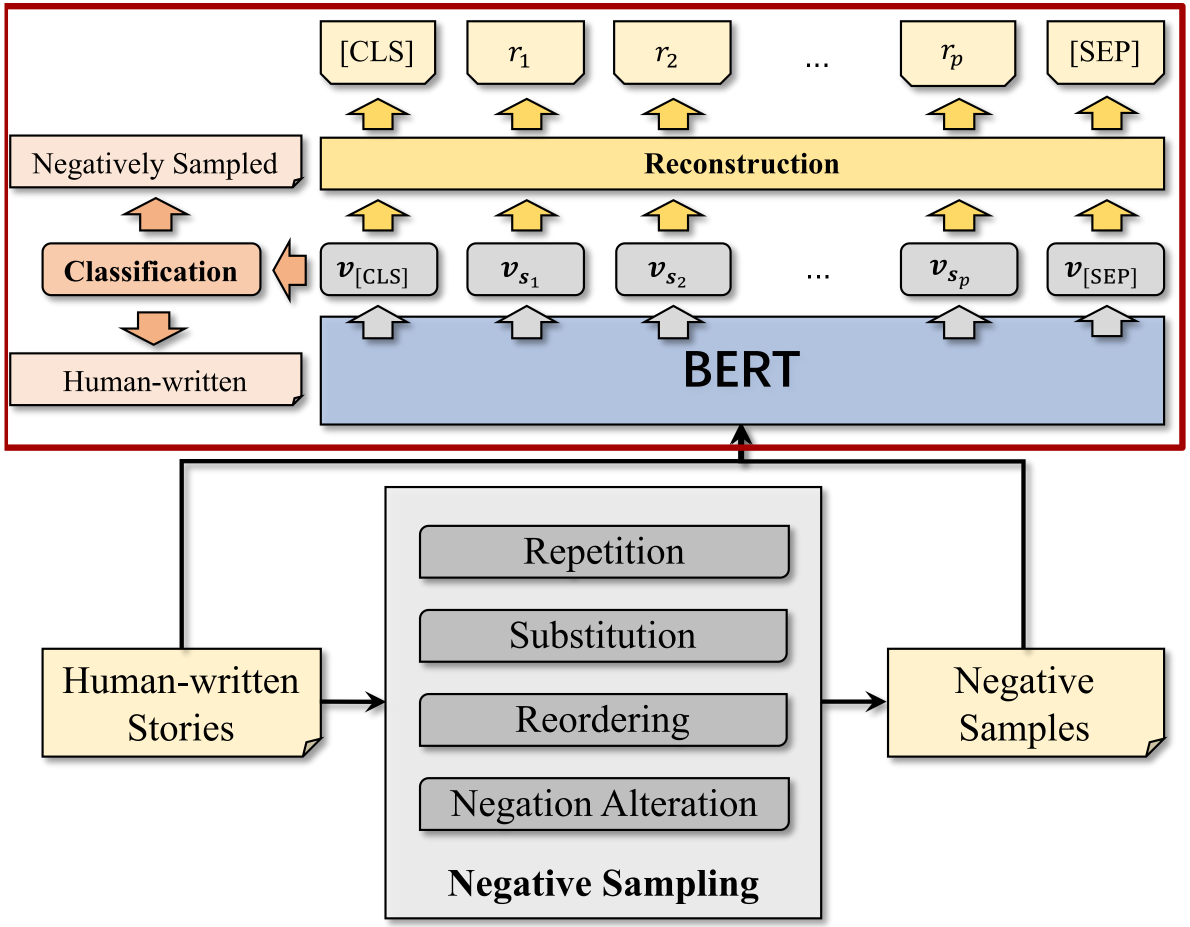
He **decided** to keep jogging.
⇳
He **did not decided** to keep jogging.

方 法：负 采 样



Leading Context Ken was out jogging one morning.
Reference By Human The weather was crisp and cool. Ken felt good and energetic. He decided to keep jogging longer than normal. Ken went several more miles out of his way.
Auto-Constructed Negative Sample The weather was crisp and **cool and cool**. Ken felt **bad** and energetic. Ken **DID NOT GO** several more miles out of his way. He decided to keep jogging longer than normal.

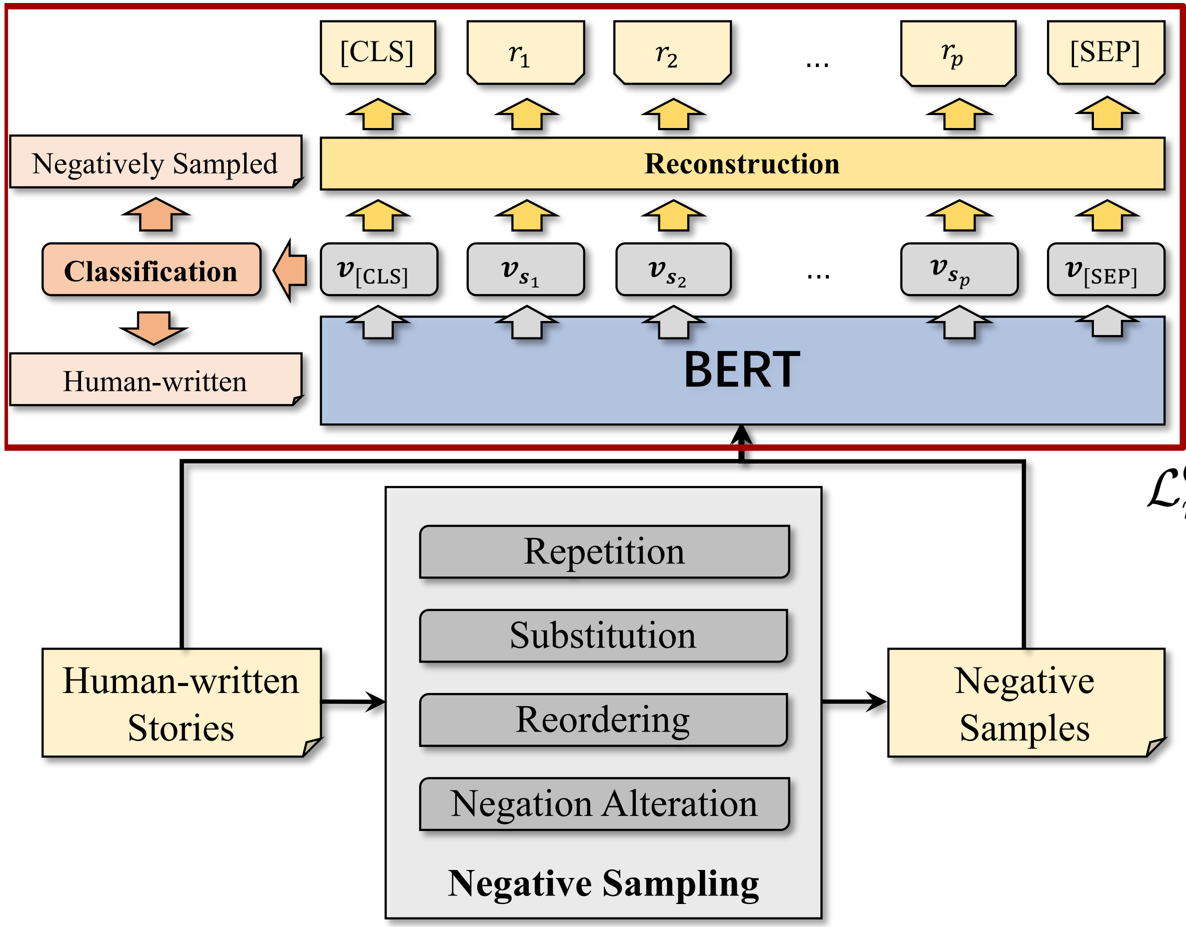
方 法：建 模



- **BERT 编码**

$$\boldsymbol{v}_{[\mathrm{CLS}]}, \boldsymbol{v}_{s_1}, \cdots, \boldsymbol{v}_{s_p}, \boldsymbol{v}_{[\mathrm{SEP}]} = \mathrm{BERT}(\boldsymbol{s}_n)$$

方 法：建 模



- **BERT 编码**

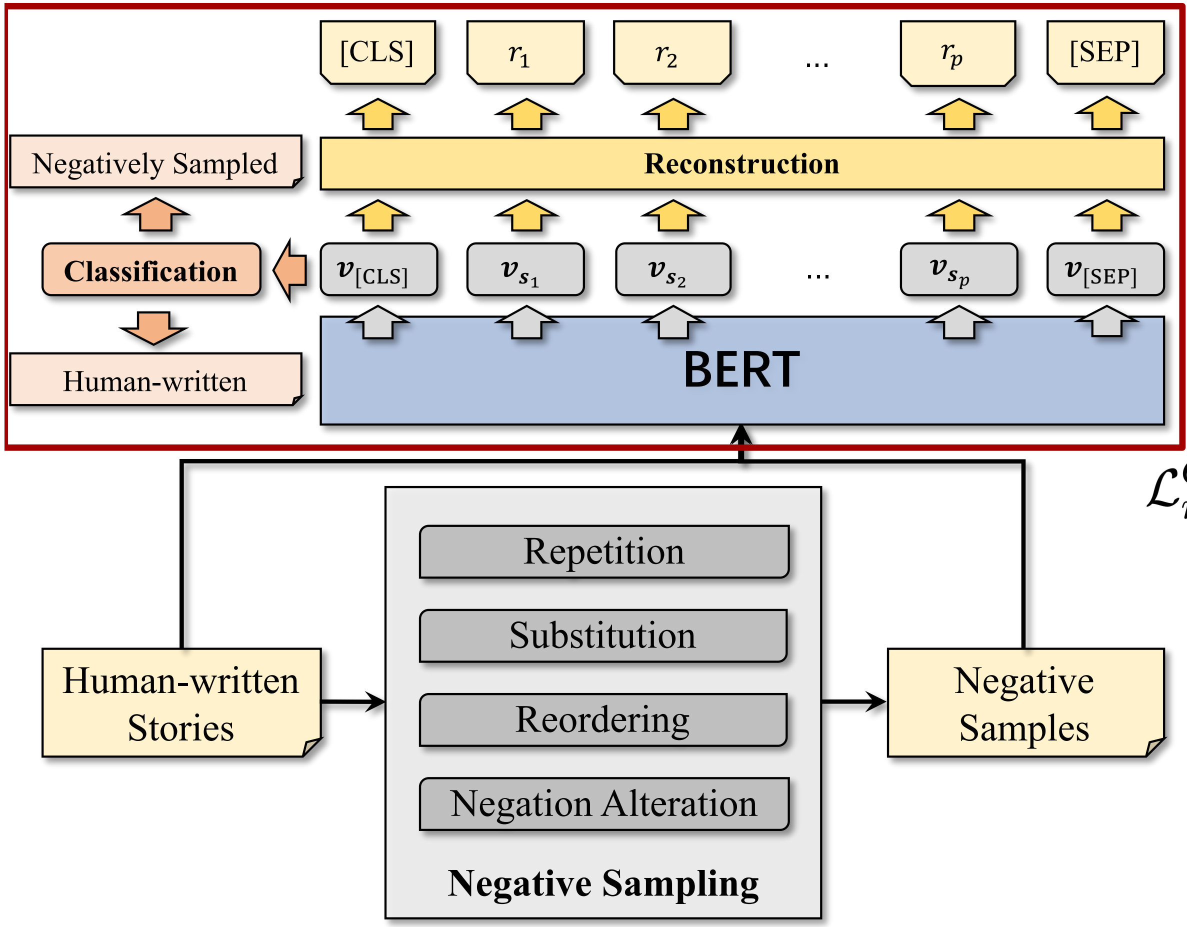
$$\boldsymbol{v}_{[\mathrm{CLS}]}, \boldsymbol{v}_{s_1}, \cdots, \boldsymbol{v}_{s_p}, \boldsymbol{v}_{[\mathrm{SEP}]} = \mathrm{BERT}(\boldsymbol{s}_n)$$

- **UNION分数预测**

$$\hat{y}_n = \mathrm{sigmoid}(\mathbf{W}_c \boldsymbol{v}_{[\mathrm{CLS}]} + \mathbf{b}_c)$$

$$\mathcal{L}_n^C = -y_n \log \hat{y}_n - (1 - y_n) \log (1 - \hat{y}_n)$$

方 法：建 模



- **BERT 编码**

$$\boldsymbol{v}_{[\mathrm{CLS}]}, \boldsymbol{v}_{s_1}, \cdots, \boldsymbol{v}_{s_p}, \boldsymbol{v}_{[\mathrm{SEP}]} = \mathrm{BERT}(\boldsymbol{s}_n)$$

- **UNION分数预测**

$$\hat{y}_n = \mathrm{sigmoid}(\mathbf{W}_c \boldsymbol{v}_{[\mathrm{CLS}]} + \mathbf{b}_c)$$

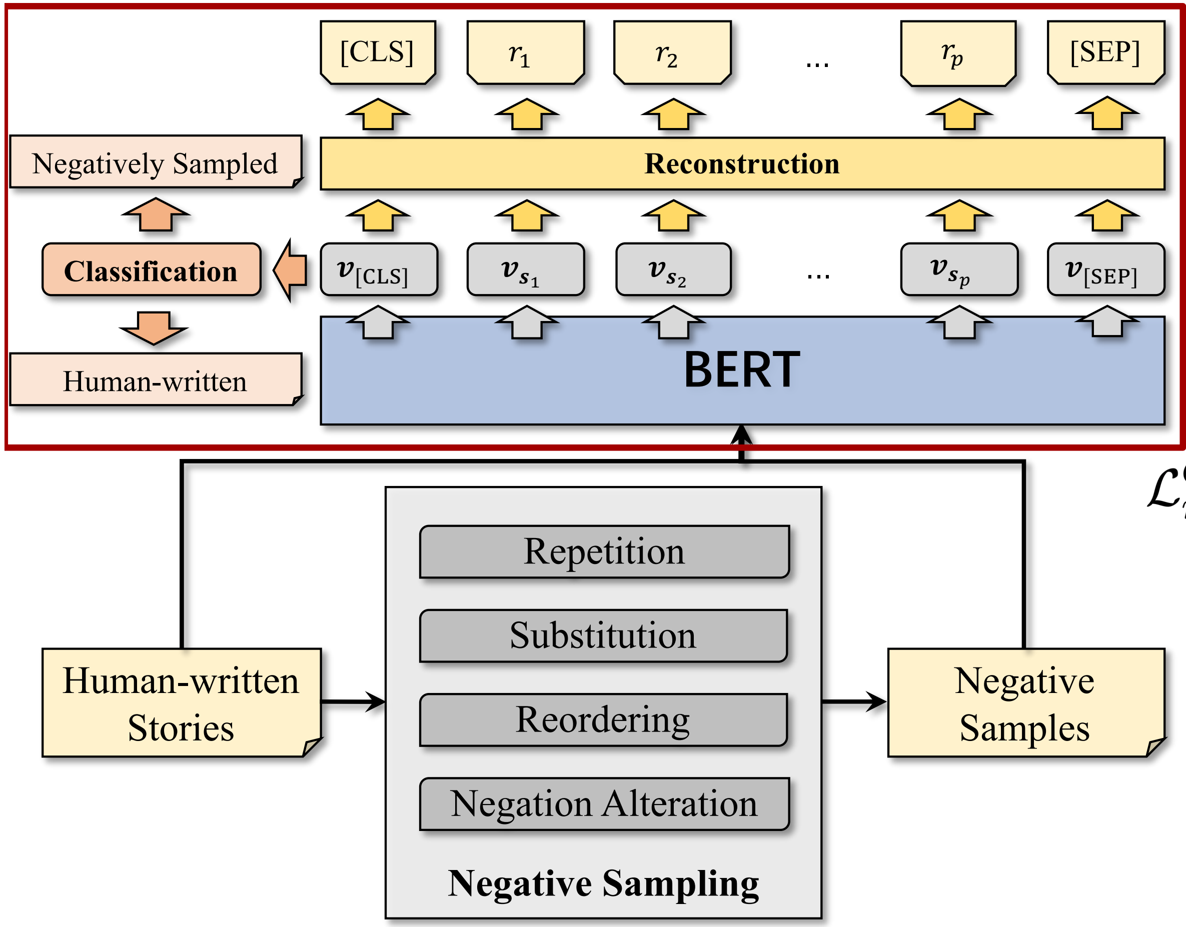
$$\mathcal{L}_n^C = -y_n \log \hat{y}_n - (1 - y_n) \log (1 - \hat{y}_n)$$

- **负样本重构**

$$P(\hat{r}_i | \boldsymbol{s}_n) = \mathrm{softmax}(\mathbf{W}_r \boldsymbol{v}_{s_i} + \mathbf{b}_r)$$

$$\mathcal{L}_n^R = -\frac{1}{p} \sum_{i=1}^{p} \log P(\hat{r}_i = r_i | \boldsymbol{s}_n)$$

方 法：建 模



- **BERT 编码**

$$\boldsymbol{v}_{[\mathrm{CLS}]}, \boldsymbol{v}_{s_1}, \cdots, \boldsymbol{v}_{s_p}, \boldsymbol{v}_{[\mathrm{SEP}]} = \mathrm{BERT}(\boldsymbol{s}_n)$$

- **UNION分数预测**

$$\hat{y}_n = \mathrm{sigmoid}(\mathbf{W}_c \boldsymbol{v}_{[\mathrm{CLS}]} + \mathbf{b}_c)$$

$$\mathcal{L}_n^C = -y_n \log \hat{y}_n - (1 - y_n) \log (1 - \hat{y}_n)$$

- **负样本重构**

$$P(\hat{r}_i | \boldsymbol{s}_n) = \mathrm{softmax}(\mathbf{W}_r \boldsymbol{v}_{s_i} + \mathbf{b}_r)$$

$$\mathcal{L}_n^R = -\frac{1}{p} \sum_{i=1}^{p} \log P(\hat{r}_i = r_i | \boldsymbol{s}_n)$$

- **整体损失**

$$\mathcal{L} = \frac{1}{N} \sum_{n=1}^{N} (\mathcal{L}_n^C + \lambda \mathcal{L}_n^R)$$

实 验

- 基线指标
  - 有参考指标：BLEU, MoverScore, RUBER$_r$-BERT
  - 无参考指标：Perplexity (of GPT2), DisScore, RUBER$_u$-BERT
  - 混合指标：RUBER-BERT, BLEURT
- 模型设置
  - UNION和所有的基线模型均基于BERT/GPT的base版本
- 数据: ROCStories (ROC), WritingPrompts (WP)
- NLG模型:
  - Fusion
  - Plan&Write
  - Fine-tuned GPT2
  - KG-enhanced GPT2

Split	Metrics	ROC	WP	NS
Train/ Validate	Perplexity	88,344/ 4,908	272,600/ 15,620	✗
	DisScore			✓
	RUBER$_u$			✓
	UNION			✓
	BLEURT	360†/40†	180†/20†	✗
Test	All metrics	400†	200†	N/A

实 验

- 和人工评价的相关性
  - r/ρ/τ : Pearson/Spearman/Kendall 相关系数

Metrics		ROC			WP		
		r	ρ	τ	r	ρ	τ
Referenced	**BLEU**	0.0299	0.0320	0.0231	0.1213	0.0941	0.0704
	MoverScore	0.1538*	0.1535*	0.1093*	0.1613	0.1450	0.1031
	RUBER$_r$-BERT	0.0448	0.0517	0.0380	0.1502	0.1357	0.0986
Unreferenced	**Perplexity**	0.2464*	0.2295*	0.1650*	-0.0705	-0.0479	-0.0345
	RUBER$_u$-BERT	0.1477*	0.1434*	0.1018*	0.1613	0.1605	0.1157
	DisScore	0.0406	0.0633	0.0456	0.0627	-0.0234	-0.0180
	UNION	**0.3687***	**0.4599***	**0.3386***	**0.3663***	**0.4493***	**0.3293***
	-Recon	0.3101*	0.4027*	0.2927*	0.3292*	0.3786*	0.2836*
Hybrid	**RUBER-BERT**	0.1412*	0.1395*	0.1015*	0.1676	0.1664	0.1194
	BLEURT	0.2310*	0.2353*	0.1679*	0.2229*	0.1602	0.1180

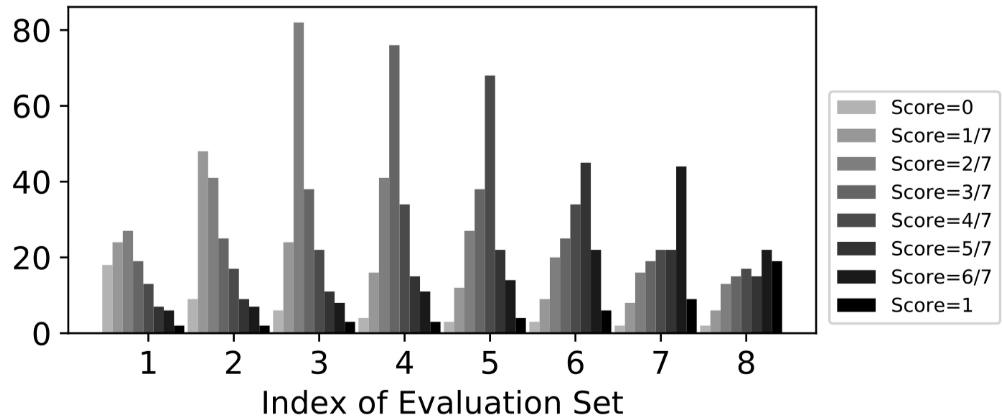
实 验

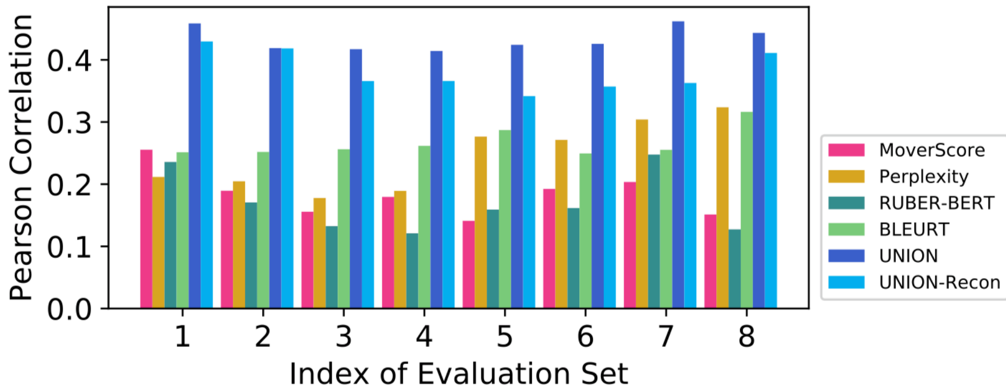
- 对数据迁移的泛化性

Metrics	r	ρ	τ
Training: WP **Test**: ROC			
Perplexity	-0.0015	0.0149	0.0101
RUBER$_u$-BERT	-0.0099	-0.0162	-0.0110
BLEURT	0.1326*	0.1137*	0.0828*
UNION	**0.1986***	**0.2501***	**0.1755***
-Recon	0.1704*	0.2158*	0.1523*

Metrics	r	ρ	τ
Training: ROC **Test**: WP			
Perplexity	0.0366	0.0198	0.0150
RUBER$_u$-BERT	0.1392	0.1276	0.0912
BLEURT	0.1560	0.1305	0.0941
UNION	**0.2872***	**0.2935***	**0.2142***
-Recon	0.2397*	0.2712*	0.1971*

- 对质量迁移的泛化性





实 验

◉ 消融实验

Evaluation Set	All Samples (400)	Reasonable Samples (19) + Unreasonable Samples with Repe (24)	Cohe (38)	Conf (61)	Chao (23)
UNION	0.3687	0.6943	0.5144	0.4571	0.6744
-Repetition	0.3167 (↓**14%**)	0.4743 (↓**32%**)	0.5308 (↑3%)	0.4316 (↓6%)	0.6561 (↓3%)
-Substitution	0.3118 (↓**15%**)	0.7034 (↑1%)	0.4185 (↓**19%**)	0.4468 (↓2%)	0.5850 (↓**13%**)
-Reordering	0.2302 (↓**38%**)	0.6546 (↓6%)	0.5077 (↓1%)	0.3507 (↓**23%**)	0.5393 (↓**20%**)
-Negation Alteration	0.3304 (↓**10%**)	0.6665 (↓4%)	0.4987 (↓3%)	0.3946 (↓**14%**)	0.5176 (↓**23%**)

◆ 四个负采样技巧对于评价故事生成都是必需的

◆ Reordering 可能是最重要的技巧

◆ 对UNION 来说，评价情节重复和场景混乱的故事可能是更容易的

案例研究

ID	Leading Context	Reference	Generated Samples	<u>H</u>	M	B	U
S1	[MALE] had joined the volunteer fire department.	He had to go through a lot of training. He took a first responder's course. [MALE] was first to respond on a scene one time. He saved a man 's life.	His first day there he saw a homeless man. He gave the man some water because he was thirsty. The man told [MALE] it was the most delicious water he ever tasted. [MALE] gave the man a small bucket of water.	<u>1.00</u>	0.34	0.43	0.99
S2	We were looking for something fun to do on a Tuesday night.	We decided to see a new movie that was out. When we got there we found out the tickets were half price on Tuesdays. We decided Tuesdays will now be our standing date night. It is such a nice, fun, cheap night that we can look forward to.	My wife and I were so excited. We went to the mall. *We had a great time. We had a great time. (Repe)*	<u>0.00</u>	0.44	0.49	0.00
S3	[NEUTRAL] had a new baby brother.	The baby would cry all night. [NEUTRAL] wasn't able to sleep. [NEUTRAL] started to despise his brother. He asked his mom if he could move to his grandmother's.	He wanted to do something new. He was sad to see other kids and play his own. [NEUTRAL] had a great time. [NEUTRAL]'s dad decided to *go shopping*. *(Cohe)*	<u>0.00</u>	0.48	0.54	0.00

案例研究

ID	Leading Context	Reference	Generated Samples	H	M	B	U
S4	[MALE] went to work for his father's business.	His father was the boss. [MALE] was lazy at work. Everyone was scared to tell his father. [MALE] continued to do a bad job.	He was very careful with his business. He *didn't get into trouble* for his mistakes. His father found out and *fired him*. He *was a bit sad* but *never did. (Conf)*	0.14	0.62	0.69	0.00
S5	[FEMALE]'s mom married [FEMALE]'s dad, and the two girls became stepsisters.	[FEMALE], 12, had grown up in a low-income single-parent household. But ani, 7, was wealthy and spoiled, so she was very bratty. At first she hated [FEMALE] and was always mean to her! But then, finally, the two girls began to become friends.	When their dad *left the house*, *he went to their room*. When he came back, he found them in the closet. He scolded them and grounded them for a year. The girls *weren't happy with their new stepmother. (Chao)*	0.00	0.45	0.52	0.00

总 结

- **UNION：评价开放端故事生成的自动评价指标**
  - 基于自监督学习框架
  - 不依赖于任何NLG模型或者人工标注
  - 达到和人工评价更好的相关性
  - 达到对数据和质量迁移更好的泛化性
  - 相似的想法可以被迁移到其他领域（例如，开放域闲聊对话生成的评价）

感 谢 聆 听 欢 迎 提 问

代码&数据： https://github.com/thu-coai/UNION

ONLG论文列表： https://github.com/thu-coai/PaperForONLG

个人主页： https://jianguanthu.github.io/

An information theoretic view on selecting linguistic probes

Zining Zhu, Frank Rudzicz

VECTOR INSTITUTE

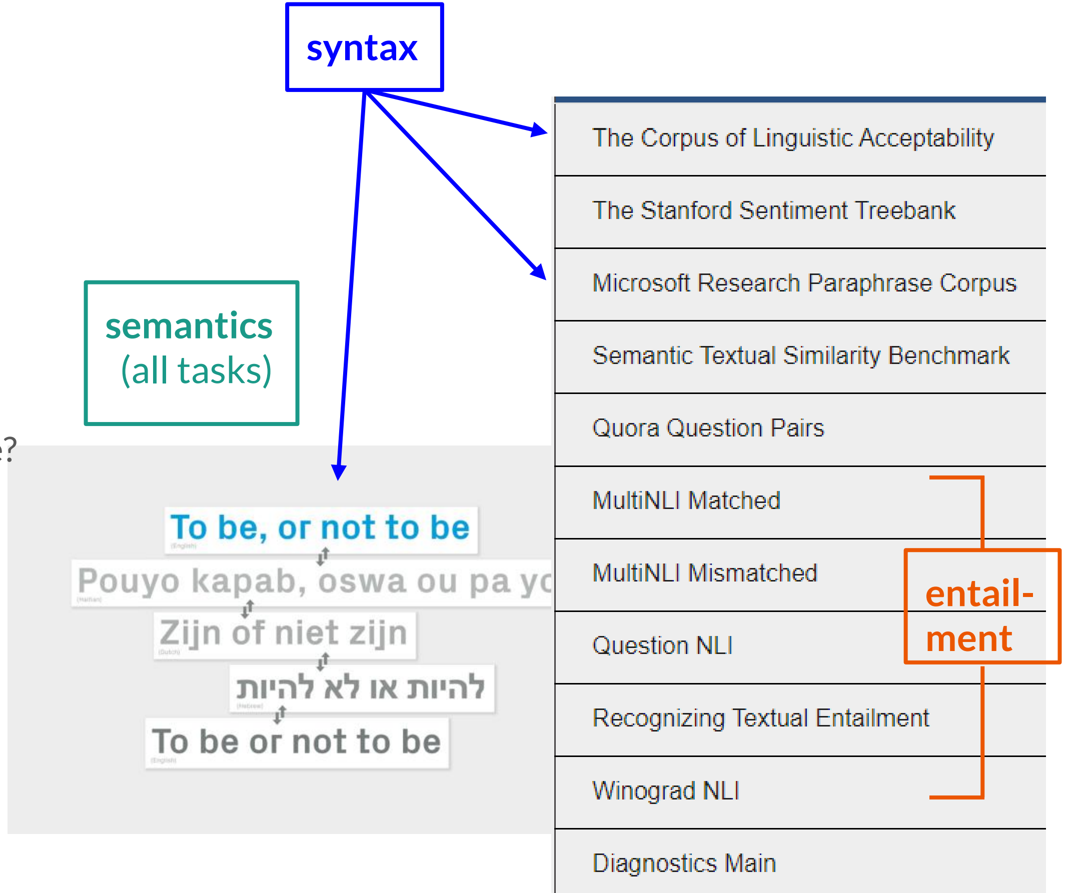
UNIVERSITY OF TORONTO

SURGICAL SAFETY TECHNOLOGIES

St. Michael's
Inspired Care.
Inspiring Science.

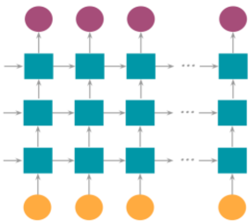
Introduction

- Neural NLP models takes over SOTA
- Why can neural NLP models do so well?
- They have linguistic knowledge.
- If yes, how much knowledge do they have?
    - Where do they encode the knowledge?
    - Do they encode only in some neurons?
    - Patterns of encoding knowledge?



Diagnostic classifiers

- (Ettinger et al., 2016): Diagnostic classifier task
    - Probe the sentence representations.
- (Alain & Bengio, 2017):
    - Diagnostic classifier essentially probes: *"Is there information about factor __ in this part of the model?"*



Probing for semantic evidence of composition by means of simple classification tasks

Allyson Ettinger[1], Ahmed Elgohary[2], Philip Resnik[1,3]
[1]Linguistics, [2]Computer Science, [3]Institute for Advanced Computer Studies
University of Maryland, College Park, MD
{aetting,resnik}@umd.edu,elgohary@cs.umd.edu

Understanding intermediate layers using linear classifier probes

Guillaume Alain
Mila, University of Montreal
guillaume.alain.umontreal@gmail.com

Yoshua Bengio
Mila, University of Montreal

What knowledge to probe?

- Many types of probing tasks:
    - Syntax distance (Hewitt and Manning, 2019)
    - Syntax & semantic tasks (Tenney et al., 2019)
    - Rhetorical discourse features (Zhu et al., 2020)
    - Many other knowledge.
- Diagnostic classifier:
    - Simple set-up.
    - Good performance -> rich knowledge.

A Structural Probe for Finding Syntax in Word Representations

John Hewitt
Stanford University
johnhew@stanford.edu

Christopher D. Manning
Stanford University
manning@stanford.edu

BERT Rediscovers the Classical NLP Pipeline

Ian Tenney[1] **Dipanjan Das**[1] **Ellie Pavlick**[1,2]
[1]Google Research [2]Brown University
{iftenney,dipanjand,epavlick}@google.com

Examining the rhetorical capacities of neural language models

Zining Zhu[1,2], **Chuer Pan**[1], **Mohamed Abdalla**[1,2], **Frank Rudzicz**[1,2,3,4]
1: University of Toronto; 2: Vector Institute
3: Li Ka Shing Knowledge Institute, St Michael's Hospital
4: Surgical Safety Technologies
zining@cs.toronto.edu, chuer.pan@mail.utoronto.ca
{msa, frank}@cs.toronto.edu

A dichotomy about probe

- (Hewitt and Liang, EMNLP 2019)
- When we observe good performance:
  - Do the representations contain rich knowledge?
  - Or, do the probe learns the task?
  - (Zhang et al., ICLR 2017): NNs can learn even from random vectors!
- Propose: select probes using "selectivity" criterion.
  - Selectivity: how much the probing accuracy *improves* compared to the control task (random labels).
- Propose to use the probes with as few parameters as possible.

UNDERSTANDING DEEP LEARNING REQUIRES RETHINKING GENERALIZATION

Chiyuan Zhang[*]
Massachusetts Institute of Technology
chiyuan@mit.edu

Samy Bengio
Google Brain
bengio@google.com

Moritz Hardt
Google Brain
mrtz@google.com

Benjamin Recht[†]
University of California, Berkeley
brecht@berkeley.edu

Oriol Vinyals
Google DeepMind
vinyals@google.com

Designing and Interpreting Probes with Control Tasks

John Hewitt
Stanford University
johnhew@stanford.edu

Percy Liang
Stanford University
pliang@cs.stanford.edu

An info-theoretic formulation

- The purpose of a diagnostic classifier is to approximate I(T;R)
  - T: the label
  - R: the representation
- Reject the “good representation or good probe” dichotomy
- Reject the *control tasks*
  - Propose to use *control function* (randomize R) as a control setting.
  - With control function, compute the “information gain” criterion to select probes.
  - Info gain: the difference of cross-entropy losses between control task and probing task:

$$\tilde{\mathcal{G}}(T, R, \mathbf{c}) = H(p_c, q_{\phi_c}) - H(p, q_\phi)$$

Information-Theoretic Probing for Linguistic Structure

Tiago Pimentel[∂] Josef Valvoda[∂] Rowan Hall Maudslay[∂] Ran Zmigrod[∂]
Adina Williams[ə] Ryan Cotterell[∂,ς]
[∂]University of Cambridge [ə]Facebook AI Research [ς]ETH Zürich
tp472@cam.ac.uk, jv406@cam.ac.uk, rh635@cam.ac.uk,
rz279@cam.ac.uk, adinawilliams@fb.com, rdc42@cl.cam.ac.uk

An info-theoretic view on the dichotomy

- The dichotomy is valid, info-theoretically.
- Decompose the good probing performance:
    - T: target. R: representation.
    - p: unknown true distribution; q: the probe model

$$H(p, q_\theta) = H(T) - I(T; R) + \mathrm{KL}(p \,\|\, q_\theta)$$

A low cross-entropy "probing loss" could be the results of:

(a) High code-target mutual information ("good representation")

Or (b) A low KL ("probe learns the task")

An information theoretic view on selecting linguistic probes

Zining Zhu[1,2], **Frank Rudzicz**[3,4,1,2]
[1] University of Toronto, [2] Vector Institute, [3] Surgical Safety Technologies
[4] Li Ka Shing Knowledge Institute, St Michael's Hospital
{zining, frank}@cs.toronto.edu

An info-theoretic view on selecting probes

- Selecting probes with the two controlling mechanisms:
  - *Control task* (Hewitt and Liang, 2019)
  - *Control function* (Pimentel et al., 2020)
- They still contain errors, but much smaller:
  - The error terms are both the diff of a pair of KL divergences
- They differ by only irrelevant terms!
  - … as long as the randomization is done well
  - How about empirically?

$$H(p_c, q_{\theta_c}) - H(p, q_\theta) = I(T; R) - \Delta$$

$$\Delta_h = \mathrm{KL}(p \,\|\, q_\theta) - \mathrm{KL}(p_c \,\|\, q_{\theta_c}) + \mathrm{Const}$$

$$\Delta_p = \mathrm{KL}(p \,\|\, q_\phi) - \mathrm{KL}(p_c \,\|\, q_{\phi_c})$$

An information theoretic view on selecting linguistic probes

Zining Zhu[1,2], **Frank Rudzicz**[3,4,1,2]
[1] University of Toronto, [2] Vector Institute, [3] Surgical Safety Technologies
[4] Li Ka Shing Knowledge Institute, St Michael's Hospital
{zining,frank}@cs.toronto.edu

VECTOR INSTITUTE

The two control mechanisms agree well

- Ran 10,000+ POS probing experiments, sweeping different param configs.
- Empirically: the criteria of Hewitt and Liang (2019) and Pimentel et al., (2020) agree,
  - To the extent similar to the "accuracy vs. cross entropy loss" agreement

Language	# POS	# Tokens train / dev / test	Correlations (t_acc,f_ent)	(t_acc,t_ent)	(f_acc,f_ent)
English	17	177k / 22k / 22k	0.1615	0.1334	0.1763
French	15	303k / 31k / 8k	0.0906	0.0606	0.1295
Spanish	16	341k / 33k / 11k	0.1360	0.0560	0.1254

Table 1: Spearman correlations between t_acc (the "selectivity" criterion (Hewitt and Liang, 2019)) and f_ent (the "gain" criterion (Pimentel et al., 2020)) are on par with two "accuracy vs. cross entropy" correlations.

An information theoretic view on selecting linguistic probes

Zining Zhu[1,2], Frank Rudzicz[3,4,1,2]
[1] University of Toronto, [2] Vector Institute, [3] Surgical Safety Technologies
[4] Li Ka Shing Knowledge Institute, St Michael's Hospital
{zining,frank}@cs.toronto.edu

Takeaways

- “Diagnostic classifier” probes can be formulated better with information theory.
    - We analyzed the sources of error of (1) single loss, (2) control mechanisms.
    - We showed the two control mechanisms are equivalent.

Special thanks to:

Mohamed, Amanjit, Bai, Yuchen, Chuer, Aparna, Frank -- for discussions about the ideas + relevant papers