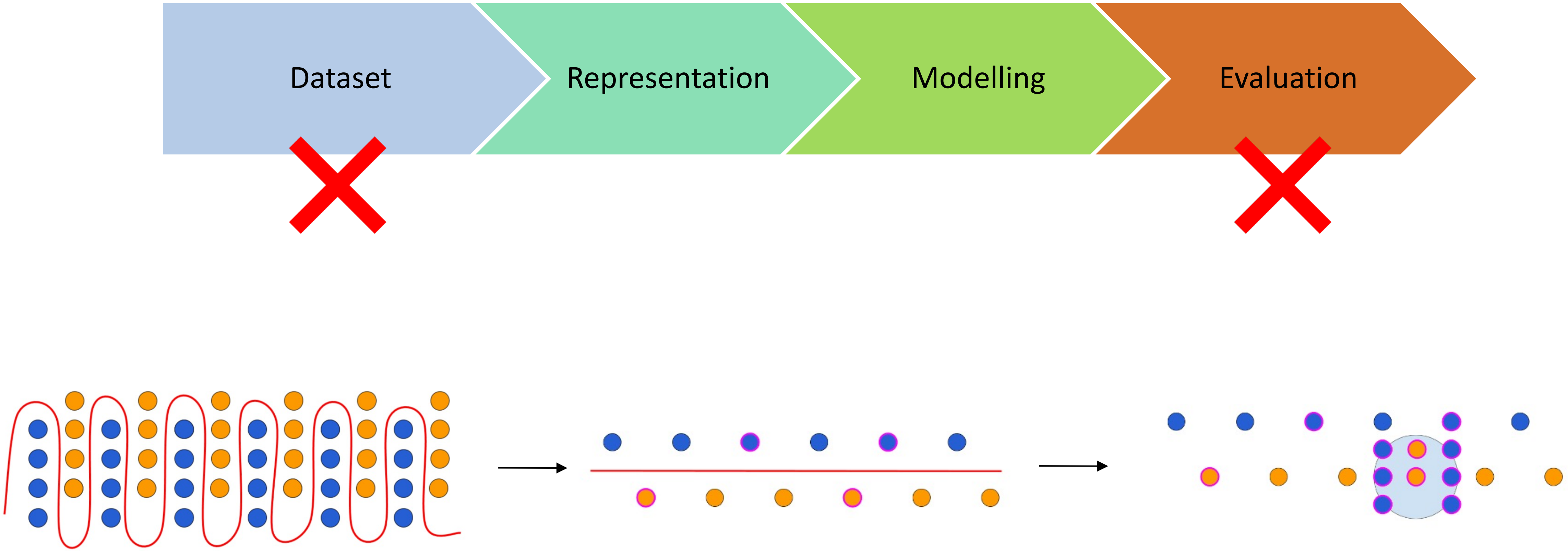
Robustness Improvement in Natural Language Processing

Tao Gui

Fudan University

0 General Framework of Natural Language Processing



0 General Framework of Natural Language Processing

ORIGINAL The government made a quick decision

BAE - R The **MASK** made a quick decision

judge , doctor , captain

BAE - I The **MASK** government made a quick decision

state , british , federal

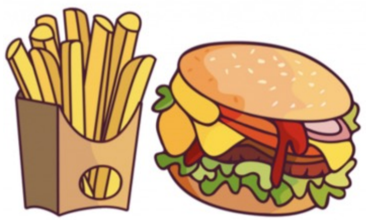
The government **MASK** made a quick decision

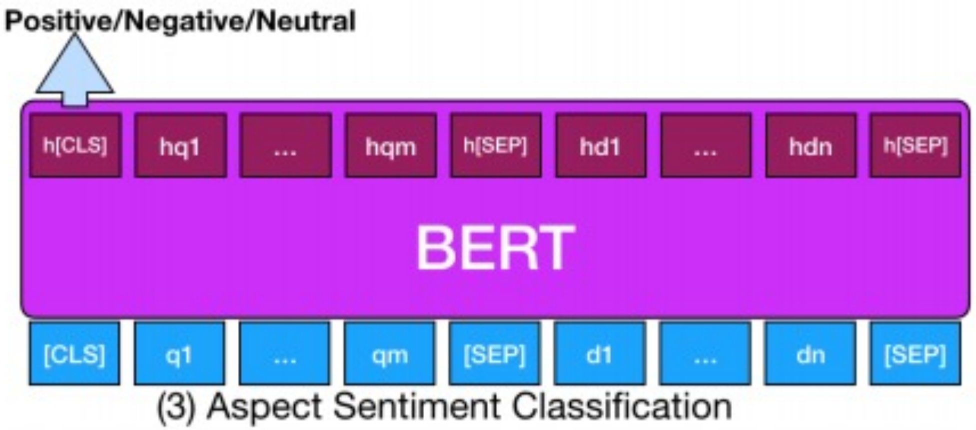
officials , then , immediately

FNLP

0 General Framework of Natural Language Processing

Dataset → Representation → Modelling → Evaluation

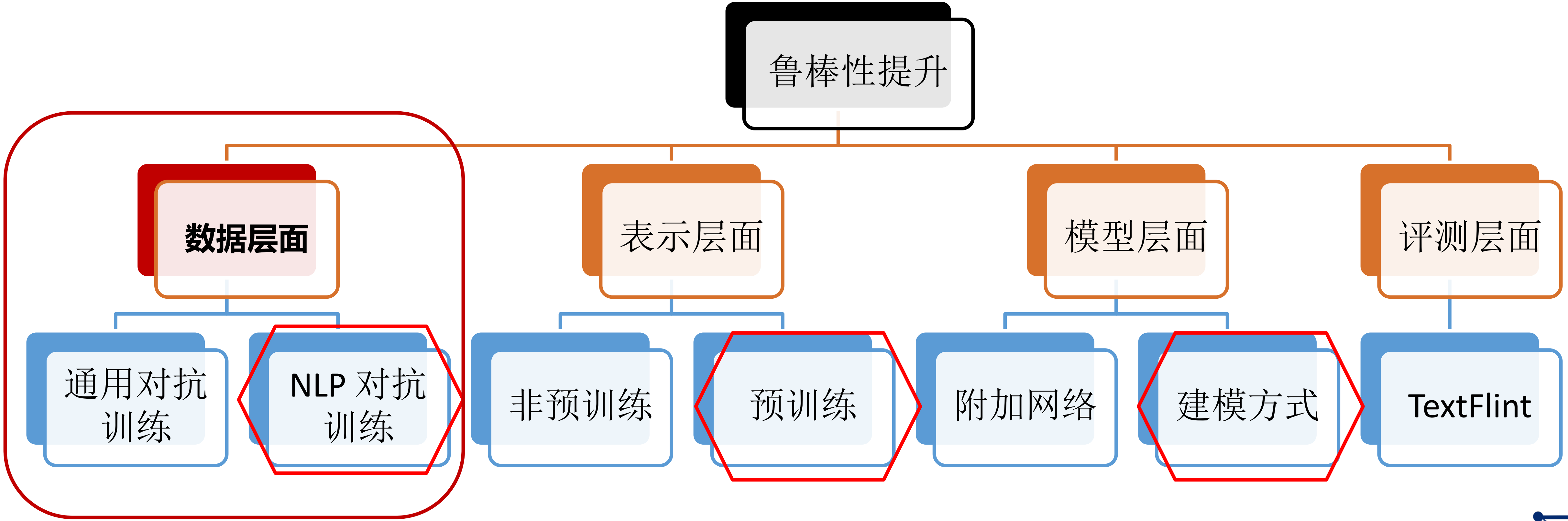




0 Outline in Robustness Improvement

- 鲁棒性提升
  - **数据层面**
    - 通用对抗训练
    - NLP 对抗训练
  - 表示层面
    - 非预训练
    - 预训练
  - 模型层面
    - 附加网络
    - 建模方式
  - 评测层面
    - TextFlint

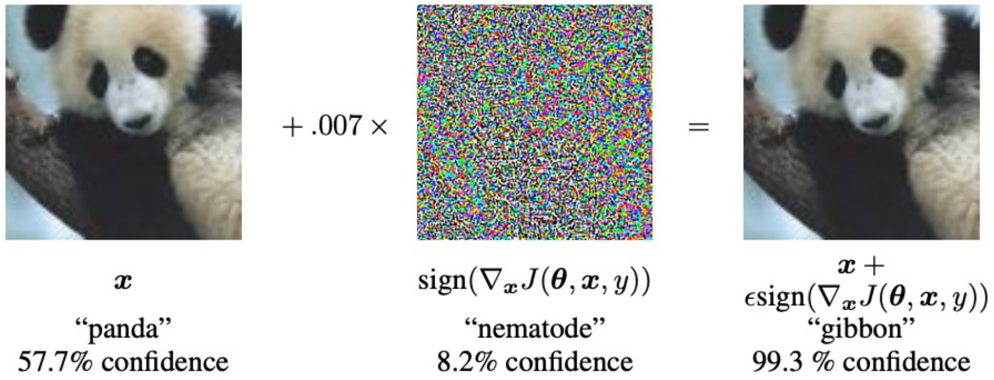
0 Outline in Robustness Improvement

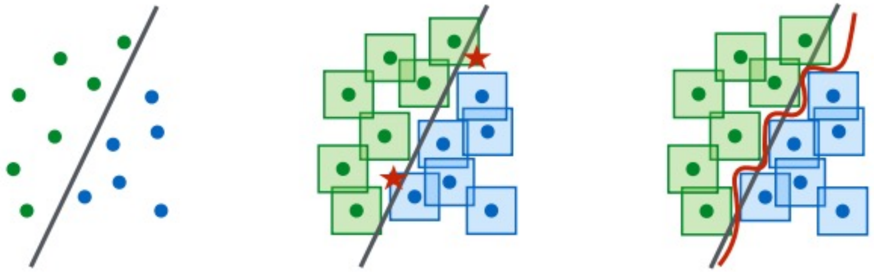


1 Robustness Improvement from Data Perspective---Universal AT

$$\mathbf{x}' = \mathbf{x} + \eta, f(\mathbf{x}) = \mathbf{y}, \mathbf{x} \in \mathbf{X}$$
$$f(\mathbf{x}') \neq \mathbf{y}$$
$$\text{or } f(\mathbf{x}') = \mathbf{y}', \mathbf{y}' \neq \mathbf{y}$$

$$\min_{\boldsymbol{\theta}} \mathbb{E}_{(\boldsymbol{Z},y)\sim\mathcal{D}} \left[\max_{\|\boldsymbol{\delta}\|\leq\epsilon} L(f_{\boldsymbol{\theta}}(\boldsymbol{X}+\boldsymbol{\delta}), y) \right],$$

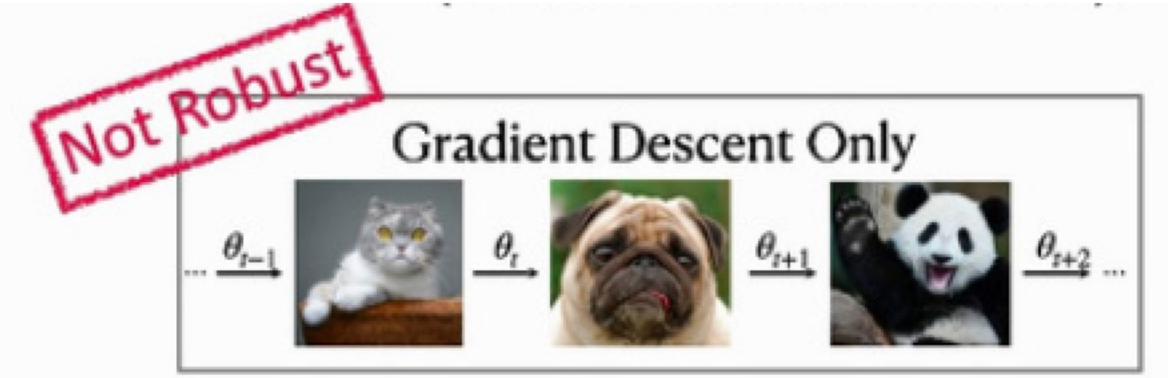




Adversarial Attack ⟷ **Adversarial Training**

Szegedy, Christian, et al. "Intriguing properties of neural networks." *arXiv preprint arXiv:1312.6199* (2013).

1 Robustness Improvement from Data Perspective---Universal AT



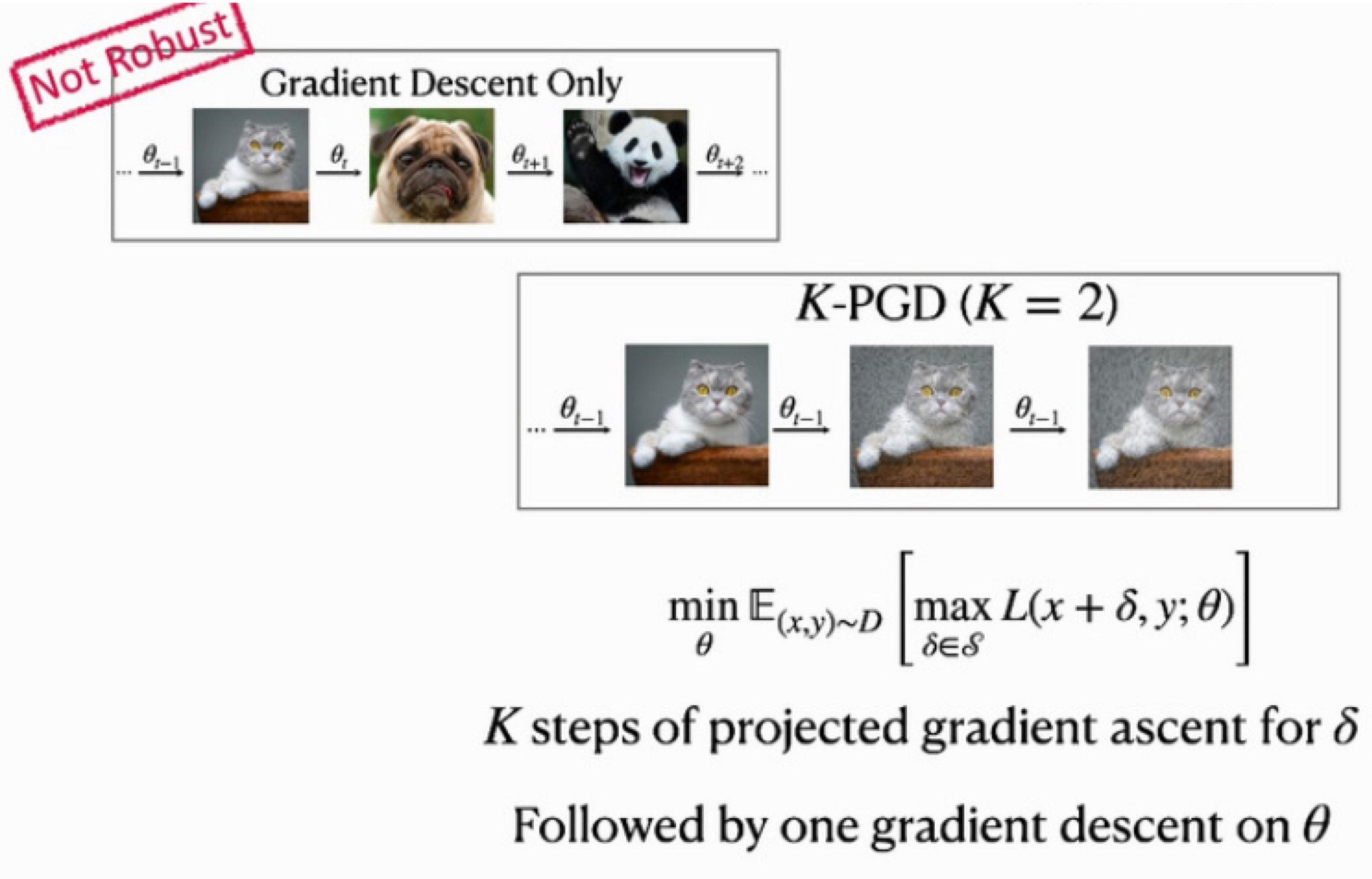
FGSM: $\delta = \epsilon \cdot Sign(g)$

FGM: $\delta = \epsilon \cdot (g/||g||_2)$

Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples." *arXiv preprint arXiv:1412.6572* (2014).
Miyato, Takeru, Andrew M. Dai, and Ian Goodfellow. "Adversarial training methods for semi-supervised text classification." *ICLR* 2016.
https://iclr.cc/virtual_2020/poster_BygzbyHFvB.html

FNLP

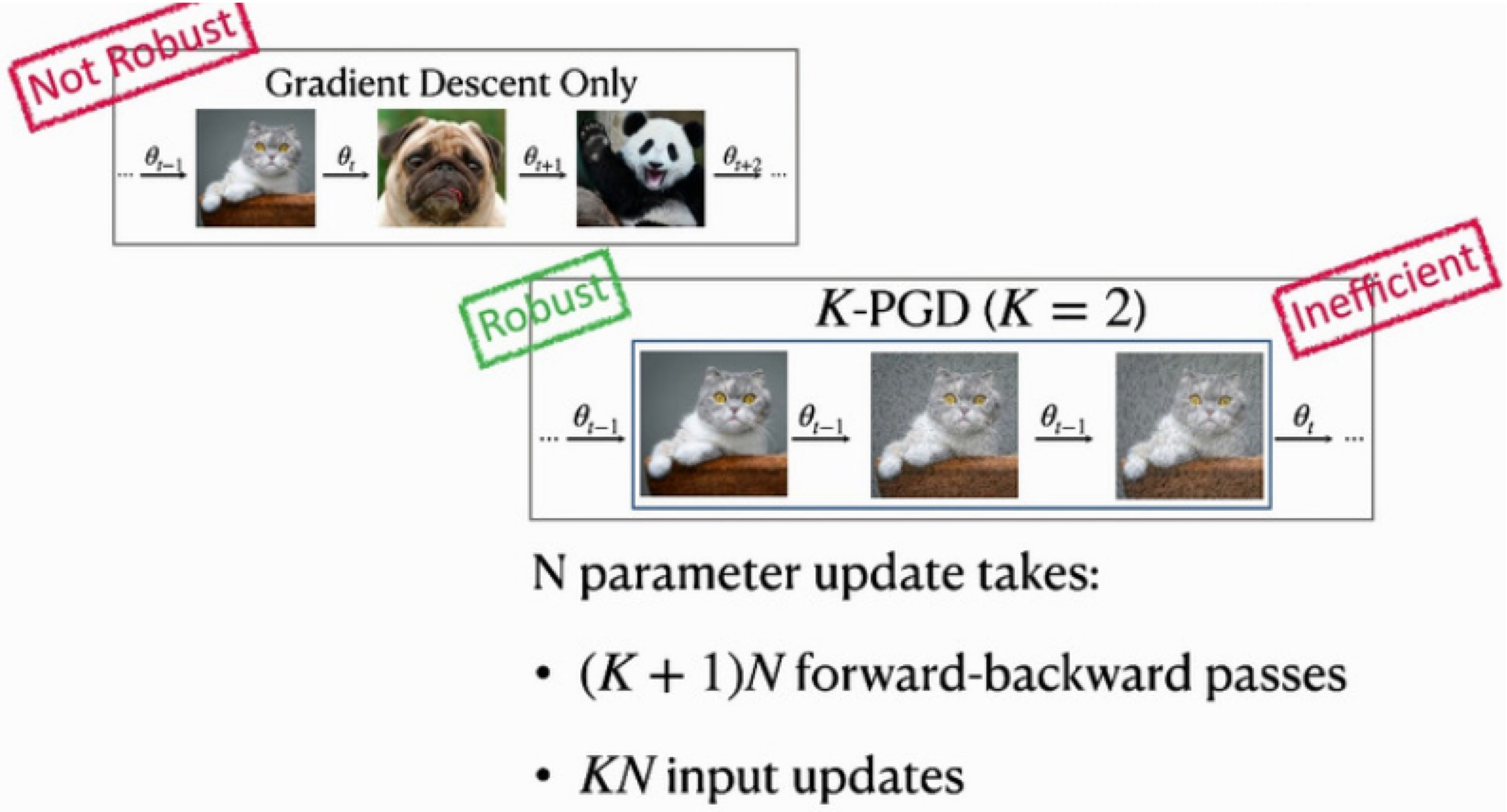
1 Robustness Improvement from Data Perspective---Universal AT



Madry, Aleksander, et al. "Towards deep learning models resistant to adversarial attacks." *arXiv preprint arXiv:1706.06083* (2017).

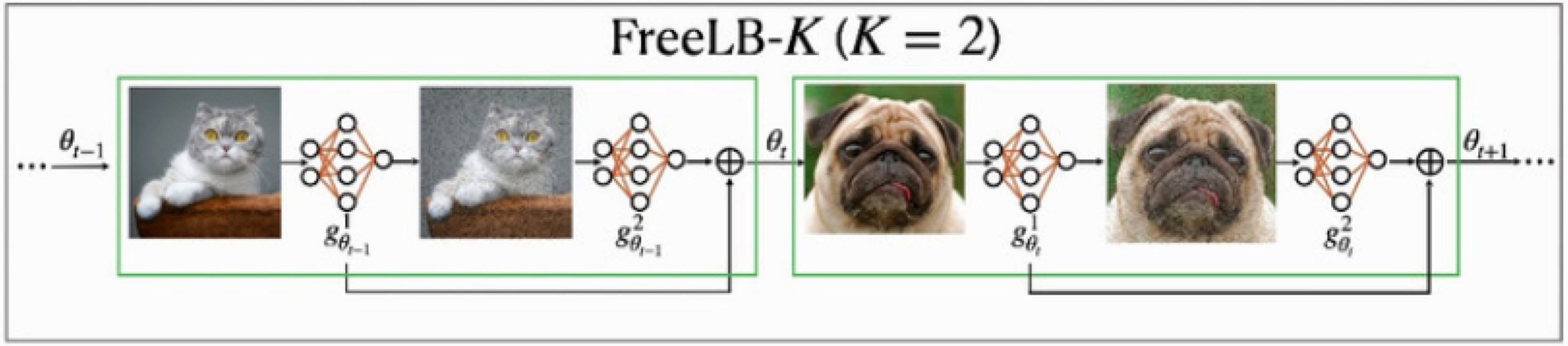
FNLP

1 Robustness Improvement from Data Perspective---Universal AT



Madry, Aleksander, et al. "Towards deep learning models resistant to adversarial attacks." *arXiv preprint arXiv:1706.06083* (2017).

1 Robustness Improvement from Data Perspective---Universal AT



Equivalent to enlarging the batch size with adversarial examples

$$\min_{\theta} \mathbb{E}_{(Z,y)\sim\mathcal{D}} \left[\frac{1}{K} \sum_{t=0}^{K-1} \max_{\delta_t \in \mathcal{I}_t} L(f_\theta(X+\delta_t), y; \theta) \right]$$

Improved invariance with a set of different transforms

Better generalization (Sokolic et al. 2017, Gong et al. 2020)

Zhu, Chen, et al. "FreeLB: Enhanced Adversarial Training for Natural Language Understanding." International Conference on Learning Representations. 2019.

1 Robustness Improvement from Data Perspective---Universal AT

AT in Image

Table 1: Validation accuracy and robustness of CIFAR-10 models trained with various methods.

Training	Evaluated Against: Nat. Images	PGD-20	PGD-100	CW-100	10 restart PGD-20	Train Time (min)
Natural	**95.01%**	0.00%	0.00%	0.00%	0.00%	**780**
Free $m=2$	91.45%	33.92%	33.20%	34.57%	33.41%	816
Free $m=4$	87.83%	41.15%	40.35%	41.96%	40.73%	800
Free $m=8$	85.96%	**46.82%**	**46.19%**	**46.60%**	**46.33%**	785
Free $m=10$	83.94%	46.31%	45.79%	45.86%	45.94%	785
7-PGD trained	87.25%	45.84%	45.29%	46.52%	45.53%	5418

Robustness improves
Accuracy decreases

AT in Text

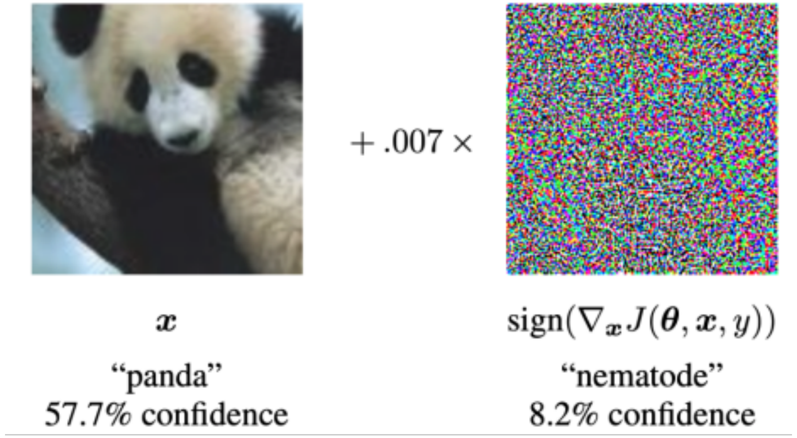
Experiment results of different defenders on AGNEWS

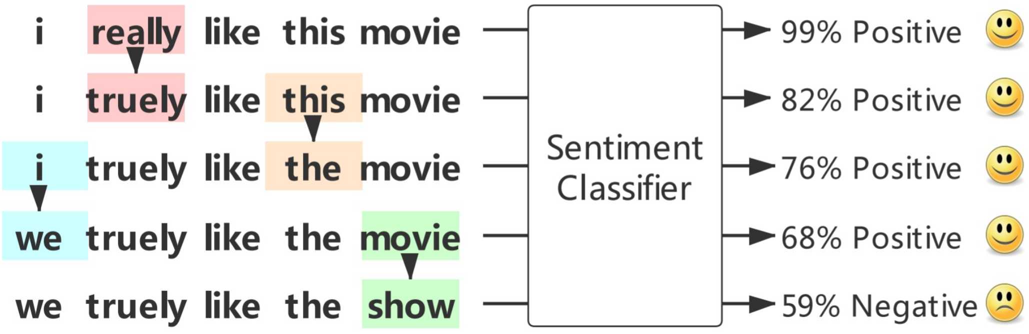
Method	Clean%	TextFooler: Aua%	Suc%	#Query
Baseline (BERT)	94.5	19.1	79.6	317.4
Adversarial Data Augmentation	94.4	38.6	58.9	404.6
MixADA (Si et al., 2020)	94.3	37.5	60.3	410.7
PGD-K (Madry et al., 2018)	94.7	24.8	73.9	353.5
FreeLB (Zhu et al., 2020)	94.7	31.6	66.7	382.1
TA-VAT (Li and Qiu, 2020)	94.8	31.0	67.3	382.5
InfoBERT (Wang et al., 2020)	**95.1**	31.8	66.5	369.9
DNE (Zhou et al., 2020)	93.9	28.7	69.8	367.9
ASCC (Dong et al., 2021)	92.3	28.2	69.6	326.5
SAFER (Ye et al., 2020)	94.3	31.8	66.1	350.1
RanMASK (Zeng et al., 2021)	91.7	37.9	58.7	**583.4**

Robustness improves
Accuracy improves

Shafahi, Ali, et al. "Adversarial training for free!." *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. 2019.

Problem in current Virtual Adversarial Train





Effectiveness of Adversarial Training

$$\min_{\theta} \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}} \left[\max_{\|\boldsymbol{\delta}\|\leq\epsilon} \mathcal{L}(f_{\theta}(\boldsymbol{X}+\boldsymbol{\delta}), y) \right], \quad (3)$$

$$\boldsymbol{\delta}_{t+1} = \prod_{\|\boldsymbol{\delta}\|_F \leq \epsilon} \left(\boldsymbol{\delta}_t + \alpha \frac{g(\boldsymbol{\delta}_t)}{\|g(\boldsymbol{\delta}_t)\|_F} \right), \quad (4)$$

$$\|\boldsymbol{\delta}_{t+1}\| \leq \|\boldsymbol{\delta}_t\| + \left\| \alpha \frac{g(\boldsymbol{\delta}_t)}{\|g(\boldsymbol{\delta}_t)\|_F} \right\| \leq \|\boldsymbol{\delta}_t\| + \alpha. \quad (5)$$

$$\begin{aligned} \|\boldsymbol{\delta}_t\| &\leq \|\boldsymbol{\delta}_{t-1}\| + \alpha \leq \|\boldsymbol{\delta}_{t-2}\| + 2\alpha \\ &\leq \cdots \leq \|\boldsymbol{\delta}_1\| + (t-1) * \alpha \leq t\alpha, \end{aligned} \quad (6)$$

Method	Norm ϵ	Clean%	TextFooler		BERT-Attack	
			Aua%	Suc%	Aua%	Suc%
PGD-K	0.01	94.9	21.8	77.0	31.5	66.8
	0.1	95.3	43.6	54.3	45.1	52.7
	1	**95.2**	**45.2**	**55.3**	**45.3**	**52.4**
	w/o	**95.2**	**45.2**	**55.3**	**45.3**	**52.4**
FreeLB	0.01	95.4	30.5	68.0	43.6	54.3
	0.1	95.5	36.1	62.2	40.0	58.1
	1	**94.9**	**45.8**	**51.7**	**42.5**	**55.2**
	w/o	**94.9**	**45.8**	**51.7**	**42.5**	**55.2**

Li, Zongyi, et al. "Searching for an Effective Defender: Benchmarking Defense against Adversarial Word Substitution." EMNLP 2021.

FNLP

Experiments

Method	Clean%	TextFooler			TextBugger			BERT-Attack		
		Aua%	Suc%	#Query	Aua%	Suc%	#Query	Aua%	Suc%	#Query
Baseline (BERT)	94.5	19.1	79.6	317.4	23.5	75.0	320.6	27.2	71.0	338.8
Adversarial Data Augmentation	94.4	38.6	58.9	404.6	43.3	53.9	418.3	42.9	54.5	407.0
MixADA (Si et al., 2020)	94.3	37.5	60.3	410.7	36.4	61.4	423.5	39.1	58.6	408.4
PGD-K (Madry et al., 2018)	94.7	24.8	73.9	353.5	26.7	71.9	367.1	39.4	58.5	399.3
FreeLB (Zhu et al., 2020)	94.7	31.6	66.7	382.1	32.9	65.4	390.6	43.9	53.8	417.1
TA-VAT (Li and Qiu, 2020)	94.8	31.0	67.3	382.5	34.2	63.9	415.2	45.0	52.5	436.9
InfoBERT (Wang et al., 2020)	**95.1**	31.8	66.5	369.9	36.3	61.8	391.6	42.4	55.3	392.8
DNE (Zhou et al., 2020)	93.9	28.7	69.8	367.9	28.2	70.3	377.6	42.4	55.5	470.1
ASCC (Dong et al., 2021)	92.3	28.2	69.6	326.5	37.0	60.1	307.4	32.7	64.7	337.1
SAFER (Ye et al., 2020)	94.3	31.8	66.1	350.1	41.2	56.1	398.8	39.3	58.2	373.5
RanMASK (Zeng et al., 2021)	91.7	37.9	58.7	**583.4**	45.0	50.9	**626.8**	**49.5**	**46.1**	**661.8**
FreeLB++	**95.1**	**51.5**	**46.0**	419.1	**55.9**	**41.4**	416.9	41.8	56.2	386.1

Table 2: The experiment results of different defenders on AGNEWS, where all models are trained on BERT. The best performance is marked in **bold**. FreeLB++ not only achieves best defense performance under both TextBugger and TextFooler, but also improves **Clean**%. Although RanMASK has also achieved significant defense performance, it drops a lot in **Clean**%.

Li, Zongyi, et al. "Searching for an Effective Defender: Benchmarking Defense against Adversarial Word Substitution." EMNLP 2021.

FNLP

Problem in current Virtual Adversarial Train

- 1. Randomization Problem
- random initialize at step zero

At step t:

generate perturbations

$$\boldsymbol{\delta}_{t+1} = \prod_{||\boldsymbol{\delta}_t||_F \leq \epsilon} \frac{(\boldsymbol{\delta}_t + \alpha g(\boldsymbol{\delta}_t))}{||g(\boldsymbol{\delta}_t)||_F} \quad (2)$$

constrain using L2norm

$$g(\boldsymbol{\delta}_t) = \nabla_{\boldsymbol{\delta}} L(f_\theta(\boldsymbol{X} + \boldsymbol{\delta}_t), y) \quad (3)$$

- 2. Constraint Problem
- Frobenius Norm on X
- X = [v0,v1,… vn …] is output of a sequence

FNLP

Method

- 1. Global Perturbation Vocabulary

- 2. Token-Level Constraint

Li, Linyang, and Xipeng Qiu. "Token-Aware Virtual Adversarial Training in Natural Language Understanding." *AAAI* 2021.

1. Perturbation Generation

Algorithm 1 Token-Aware Virtual Adversarial Training

Require: Training Samples $S = \{(X = [w_0, \cdots, w_i, \cdots], y)\}$, perturbation bound ϵ, initialize bound σ adversarial steps K, adversarial step size α, model parameter θ

$\boldsymbol{V} \in \mathbb{R}^{N \times D} \leftarrow \frac{1}{\sqrt{D}} U(-\sigma, \sigma)$ // Initialize perturbation vocabulary $\boldsymbol{V}$

for epoch = $1, \cdots,$ **do**

  for batch $B \subset S$ **do**

    $\boldsymbol{\delta}_0 \leftarrow \frac{1}{\sqrt{D}} U(-\sigma, \sigma)$, $\boldsymbol{\eta}_0^i \leftarrow \boldsymbol{V}[w_i]$, $\boldsymbol{g}_0 \leftarrow 0$ //Initialize perturbation and gradient of θ

    for t = $1, \cdots, K$ **do**

      $\boldsymbol{g}_t \leftarrow \boldsymbol{g}_{t-1} + \frac{1}{K} \mathbb{E}_{(X,y) \in B} [\nabla_\theta L(f_\theta(X + \boldsymbol{\delta}_{t-1} + \boldsymbol{\eta}_{t-1}), y)]$ //Accumulate gradients of θ

      Update token-level perturbation $\boldsymbol{\eta}$:

      $\boldsymbol{g}_\eta^i \leftarrow \nabla_{\eta^i} L(f_\theta((X + \boldsymbol{\delta}_{t-1} + \boldsymbol{\eta}_{t-1}), y)$

      $\boldsymbol{\eta}_t^i \leftarrow n^i * (\boldsymbol{\eta}_{t-1}^i + \alpha \cdot \boldsymbol{g}_\eta^i / ||\boldsymbol{g}_\eta^i||_F)$

      $\boldsymbol{\eta}_t \leftarrow \prod_{||\boldsymbol{\eta}||_F < \epsilon} (\boldsymbol{\eta}_t)$

      Update instance-level perturbation $\boldsymbol{\delta}$:

      $\boldsymbol{g}_\delta \leftarrow \nabla_\delta L(f_\theta((X + \boldsymbol{\delta}_{t-1} + \boldsymbol{\eta}_{t-1}), y)$

      $\boldsymbol{\delta}_t \leftarrow \prod_{||\boldsymbol{\delta}||_F < \epsilon} (\boldsymbol{\delta}_{t-1} + \alpha \cdot \boldsymbol{g}_\delta / ||\boldsymbol{g}_\delta||_F)$

    end for

    $\boldsymbol{V}[w_i] \leftarrow \boldsymbol{\eta}_K^i$ //Update perturbation vocabulary $\boldsymbol{V}$

    $\theta \leftarrow \theta - g_K$ //Update model parameter θ

  end for

end for

1 Robustness Improvement from Data Perspective---NLP AT

2. Global Perturbation Vocabulary

Algorithm 1 Token-Aware Virtual Adversarial Training

Require: Training Samples $S = \{(X = [w_0, \cdots, w_i, \cdots], y)\}$, perturbation bound ϵ, initialize bound σ adversarial steps K, adversarial step size α, model parameter θ

$\boldsymbol{V} \in \mathbb{R}^{N \times D} \leftarrow \frac{1}{\sqrt{D}} U(-\sigma, \sigma)$ // Initialize perturbation vocabulary $\boldsymbol{V}$

for epoch $= 1, \cdots,$ **do**

  for batch $B \subset S$ **do**

    $\boldsymbol{\delta}_0 \leftarrow \frac{1}{\sqrt{D}} U(-\sigma, \sigma)$, $\boldsymbol{\eta}_0^i \leftarrow \boldsymbol{V}[w_i]$, $\boldsymbol{g}_0 \leftarrow 0$ //Initialize perturbation and gradient of θ

    for $t = 1, \cdots, K$ **do**

      $\boldsymbol{g}_t \leftarrow \boldsymbol{g}_{t-1} + \frac{1}{K} \mathbb{E}_{(X,y) \in B} [\nabla_\theta L(f_\theta(X + \boldsymbol{\delta}_{t-1} + \boldsymbol{\eta}_{t-1}), y)]$ //Accumulate gradients of θ

      Update token-level perturbation $\boldsymbol{\eta}$:

      $\boldsymbol{g}_\eta^i \leftarrow \nabla_{\eta^i} L(f_\theta((X + \boldsymbol{\delta}_{t-1} + \boldsymbol{\eta}_{t-1}), y)$

      $\boldsymbol{\eta}_t^i \leftarrow n^i * (\boldsymbol{\eta}_{t-1}^i + \alpha \cdot \boldsymbol{g}_\eta^i / ||\boldsymbol{g}_\eta^i||_F)$

      $\boldsymbol{\eta}_t \leftarrow \prod_{||\boldsymbol{\eta}||_F < \epsilon} (\boldsymbol{\eta}_t)$

      Update instance-level perturbation $\boldsymbol{\delta}$:

      $\boldsymbol{g}_\delta \leftarrow \nabla_\delta L(f_\theta((X + \boldsymbol{\delta}_{t-1} + \boldsymbol{\eta}_{t-1}), y)$

      $\boldsymbol{\delta}_t \leftarrow \prod_{||\boldsymbol{\delta}||_F < \epsilon} (\boldsymbol{\delta}_{t-1} + \alpha \cdot \boldsymbol{g}_\delta / ||\boldsymbol{g}_\delta||_F)$

    end for

    $\boldsymbol{V}[w_i] \leftarrow \boldsymbol{\eta}_K^i$ //Update perturbation vocabulary $\boldsymbol{V}$

    $\theta \leftarrow \theta - g_K$ //Update model parameter θ

  end for

end for

3. Token-Level Constraint

Algorithm 1 Token-Aware Virtual Adversarial Training

Require: Training Samples $S = \{(X = [w_0, \cdots, w_i, \cdots], y)\}$, perturbation bound ϵ, initialize bound σ adversarial steps K, adversarial step size α, model parameter θ

$\boldsymbol{V} \in \mathbb{R}^{N \times D} \leftarrow \frac{1}{\sqrt{D}} U(-\sigma, \sigma)$ // Initialize perturbation vocabulary $\boldsymbol{V}$
for epoch = $1, \cdots,$ **do**
  for batch $B \subset S$ **do**
    $\boldsymbol{\delta}_0 \leftarrow \frac{1}{\sqrt{D}} U(-\sigma, \sigma)$, $\boldsymbol{\eta}_0^i \leftarrow \boldsymbol{V}[w_i]$, $\boldsymbol{g}_0 \leftarrow 0$ //Initialize perturbation and gradient of θ
    for t = $1, \cdots, K$ **do**
      $\boldsymbol{g}_t \leftarrow \boldsymbol{g}_{t-1} + \frac{1}{K} \mathbb{E}_{(X,y) \in B}[\nabla_\theta L(f_\theta(X + \boldsymbol{\delta}_{t-1} + \boldsymbol{\eta}_{t-1}), y)]$ //Accumul
      Update token-level perturbation $\boldsymbol{\eta}$:
      $\boldsymbol{g}_\eta^i \leftarrow \nabla_{\eta^i} L(f_\theta((X + \boldsymbol{\delta}_{t-1} + \boldsymbol{\eta}_{t-1}), y)$
      $\boldsymbol{\eta}_t^i \leftarrow n^i \cdot (\boldsymbol{\eta}_{t-1}^i + \alpha \cdot \boldsymbol{g}_\eta^i / ||\boldsymbol{g}_\eta^i||_F)$
      $\boldsymbol{\eta}_t \leftarrow \prod_{||\boldsymbol{\eta}||_F < \epsilon}(\boldsymbol{\eta}_t)$
      Update instance-level perturbation $\boldsymbol{\delta}$:
      $\boldsymbol{g}_\delta \leftarrow \nabla_\delta L(f_\theta((X + \boldsymbol{\delta}_{t-1} + \boldsymbol{\eta}_{t-1}), y)$
      $\boldsymbol{\delta}_t \leftarrow \prod_{||\boldsymbol{\delta}||_F < \epsilon}(\boldsymbol{\delta}_{t-1} + \alpha \cdot \boldsymbol{g}_\delta / ||\boldsymbol{g}_\delta||_F)$
    end for
    $\boldsymbol{V}[w_i] \leftarrow \boldsymbol{\eta}_K^i$ //Update perturbation vocabulary $\boldsymbol{V}$
    $\theta \leftarrow \theta - g_K$ //Update model parameter θ
  end for
end for

$$n^i = \frac{||\boldsymbol{\eta}_t^i||_F}{\max_j(||\boldsymbol{\eta}_t^j||_F)}$$

FNLP

Results

Model	RTE	QNLI	MRPC	CoLA	SST	STS-B	MNLI-m/mm	QQP
	Acc	Acc	Acc/f1	Mcc	Acc	P/S Corr	Acc	Acc/f1
BERT-BASE								
BERT (Devlin et al. 2018)	-	88.4	-/86.7	-	92.7	-	84.4/-	-
BERT-ReImp	63.5	91.1	84.1/89.0	54.7	92.9	89.2/88.8	84.5/84.4	90.9/88.3
FreeAT-ReImp	68.0	91.3	85.0/89.2	57.5	93.2	89.5/89.0	84.9 / 85.0	91.2/88.5
FreeLB-ReImp	70.0	91.5	86.0/90.0	58.9	93.4	89.7/89.2	85.3 / 85.5	91.4/88.6
TA-VAT(ours)	**74.0**	**92.4**	**88.0/91.6**	**62.0**	**93.7**	**90.0/89.6**	**85.7 / 85.8**	**91.6/88.9**
ALBERT-xxlarge-v2								
ALBERT-xxlarge-v2(Lan et al. 2019)	89.2	95.3	-/90.9	71.4	96.9(96.5)	93.0/-	90.8/-	92.2/-
FreeLB(Zhu et al. 2020)	89.9	95.6*	-/92.4	73.1	97.0	93.2/-	90.9/-	92.5/-
TA-VAT(ours)	**90.3**	**95.7**	-/ **93.4**	**74.1**	96.8	**93.4**/-	**91.1**/-	**92.6** /-

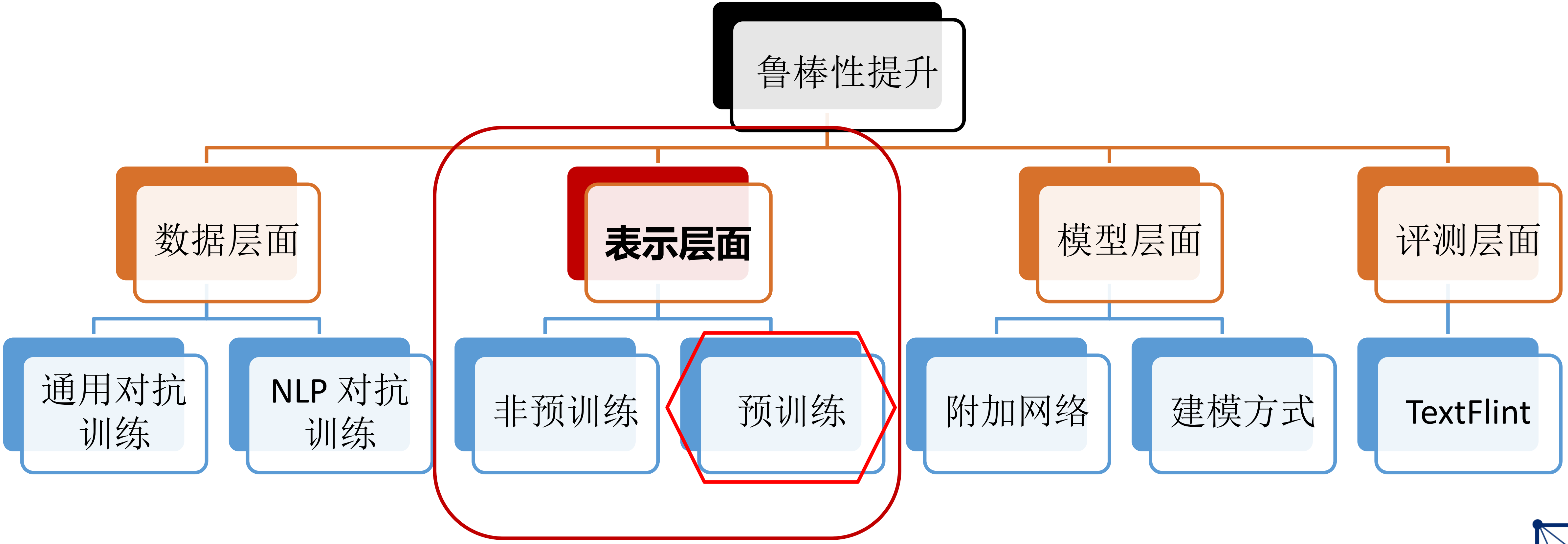
Table 1: Evaluation results on the development set of GLUE benchmark. QNLI* in FreeLB is formed as pairwise ranking task.

Model	RTE	QNLI	MRPC	CoLA	SST	STS-B	MNLI-m/mm	QQP
	Acc	Acc	Acc/f1	Mcc	Acc	P/S Corr	Acc	Acc/f1
BERT-BASE(Devlin et al. 2018)	66.4	90.5	88.9/84.8	52.1	93.5	87.1/85.8	84.6/83.4	71.2/89.2
FreeLB(Zhu et al. 2020)	70.1	91.8*	88.1/83.5	54.5	93.6	87.7/86.7	85.7/84.6	72.7/89.6
TA-VAT(ours)	**71.0**	**91.7**	**88.9/84.5**	**55.9**	**94.5**	86.8/85.7	85.2/**84.7**	**72.8**/89.5

Table 2: Evaluation results on the test set of GLUE benchmark. Results use the evaluation server on GLUE website. QNLI* in FreeLB is formed as pairwise ranking task.

FNLP

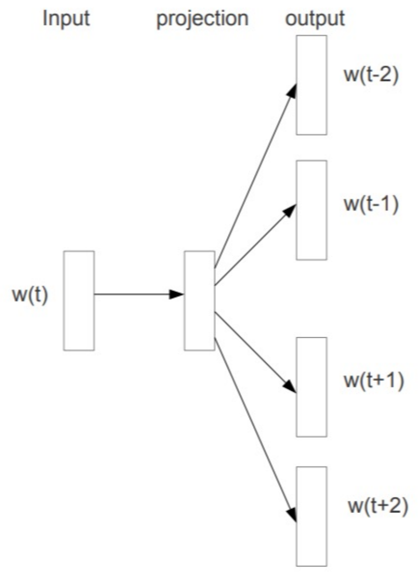
0 General Framework of Natural Language Processing



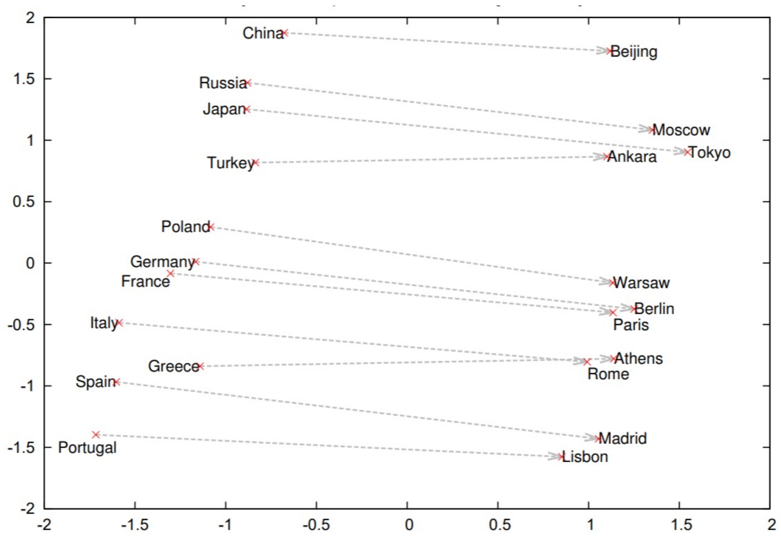
2 Robustness Improvement from Representation Perspective

- Word Embedding (unk 未见词)

  - Glove / Word2Vec / Random Init



Word2Vec (skip-gram)



Good Representation :
Country and Capital Vectors

Mikolov, Tomas, et al. "Distributed representations of words and phrases and their compositionality." Advances in neural information processing systems. 2013.

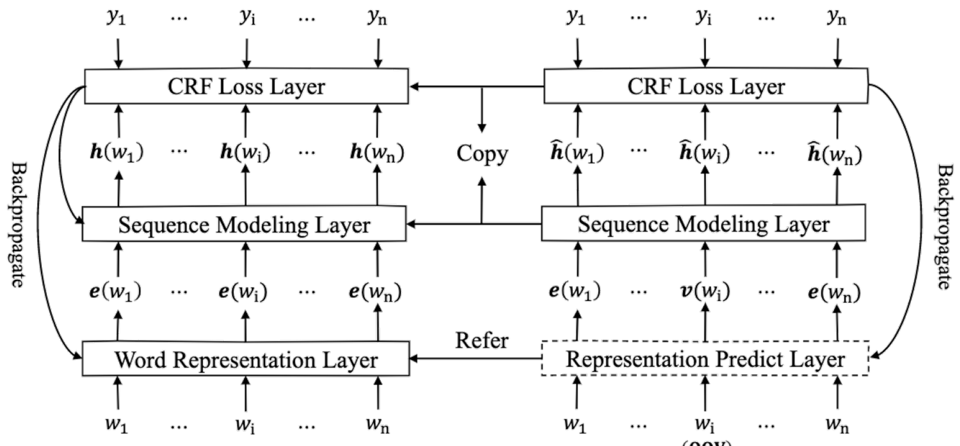
FNLP

2 Robustness Improvement from Representation Perspective---OOV

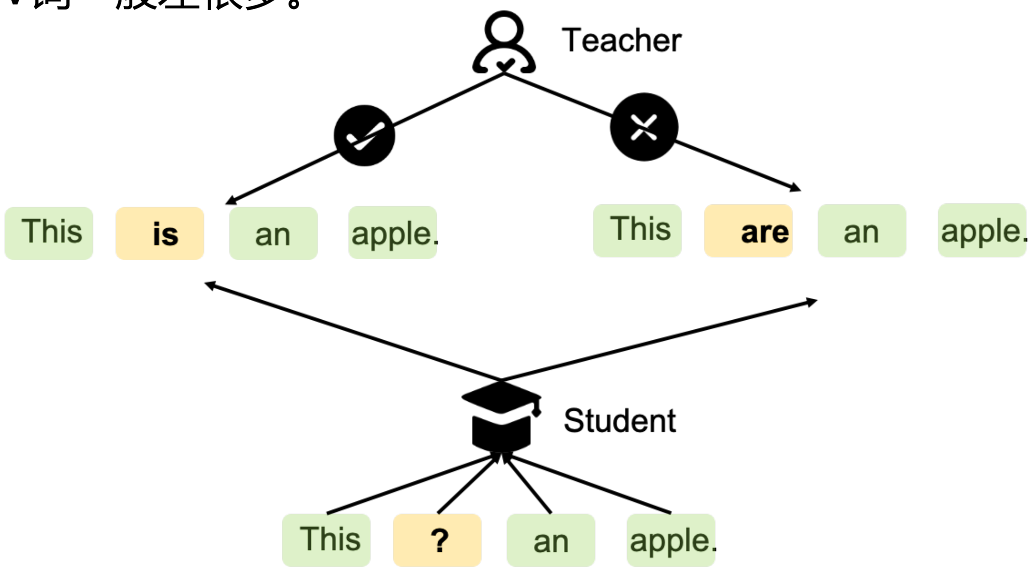
学习 OOV 词在序列标注任务中的表示

未 登 录 词

- 研究动机
  - ➢ 在测试集中经常会出现一些在训练集中没有出现过的OOV词。
  - ➢ 在序列标注任务中，模型在OOV词上的表现相较于非OOV词一般差很多。



Teacher Network　　Student Network



- 模型
  - ➢ 训练一个普通的序列标注模型，称为Teacher Network。
  - ➢ 训练一个OOV词表示模型(Student Network)：利用一个词的上下文和该词的字符序列预测改词的表示，使得预测的词表示在Teacher Network中有好的表现。

Peng, Minlong, et al., Learning Task-Specific Representation for Novel Words in Sequence Labeling, IJCAI 2019

FNLP

2 Robustness Improvement from Representation Perspective---OOV

Dataset	Dev #OOV	Dev OOV Rate	Test #OOV	Test OOV Rate
NER				
CoNLL02-Spanish	2,216	50.91%	1,544	43.38%
CoNLL02-Dutch	1,819	69.53%	2,564	65.05%
Twitter-English	1,266	79.15%	4,131	79.13%
CoNLL03-German	3,928	81.27%	2,685	73.10%

数据集 OOV 统计

Arch	Model	CoNLL02-Spanish		CoNLL02-Dutch		Twitter-English		CoNLL03-German	
		Dev	Test	Dev	Test	Dev	Test	Dev	Test
LSTM	RandomUNK	69.36	72.06	64.23	64.08	56.88	56.38	55.92	56.89
	SingleUNK	68.79	71.59	67.83	66.39	56.82	56.39	59.69	60.16
	[Lazaridou *et al.*, 2017]	68.61	69.08	65.99	65.43	47.72	47.20	47.87	49.17
	[Khodak *et al.*, 2018]	68.74	69.53	66.34	65.70	48.22	47.28	47.97	49.33
	[Schick and Schütze, 2018]	70.84	72.88	68.88	67.51	59.18	57.21	55.83	58.42
	[Akbik *et al.*, 2018]	61.78	64.06	60.49	62.09	49.68	50.22	55.06	53.01
	Proposed	**73.91**	**74.63**	**70.33**	**70.12**	**60.14**	**58.32**	**60.55**	**61.79**

面对 OOV 鲁棒性

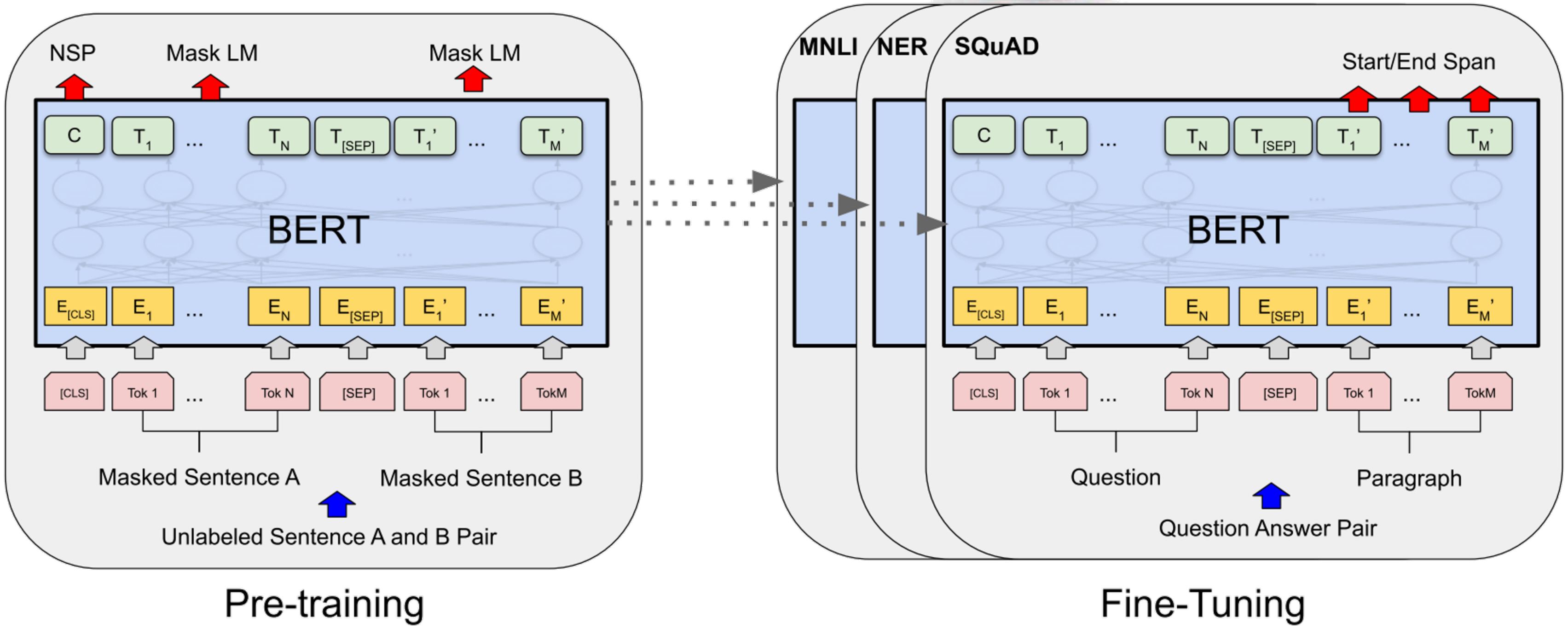
Peng, Minlong, et al., Learning Task-Specific Representation for Novel Words in Sequence Labeling, IJCAI 2019

FNLP

2 Robustness Improvement from Representation Perspective---Contextual Embedding

➢ Pretrained Contextual Embedding (contextual)

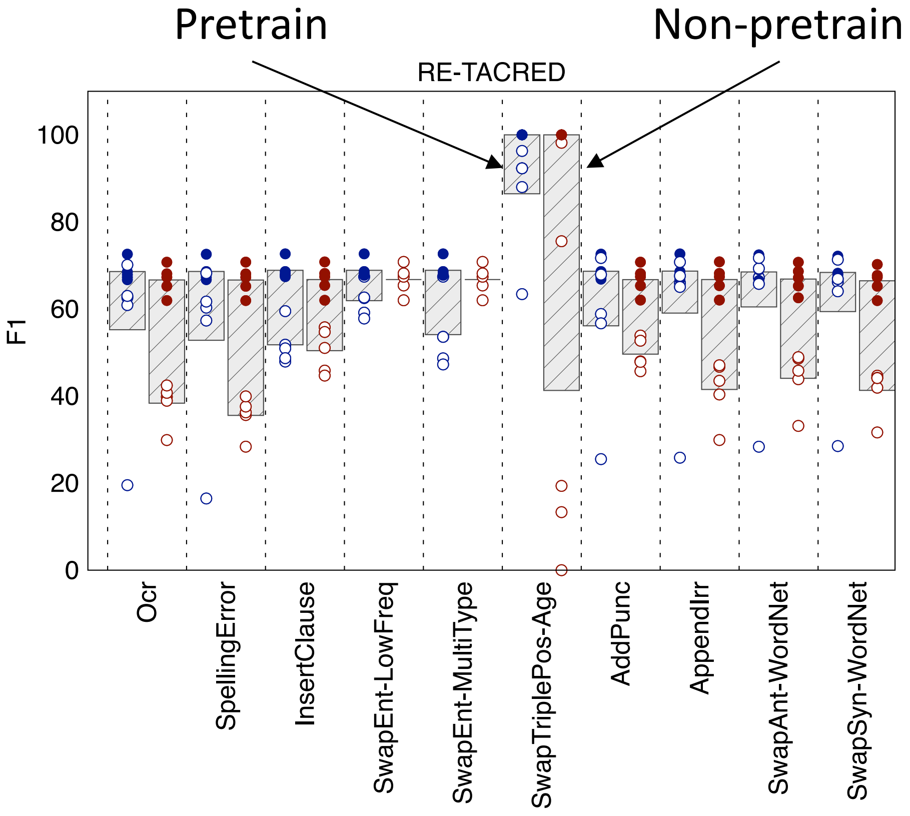
- ELMo / BERT (我在苹果公司工作/我吃了苹果)

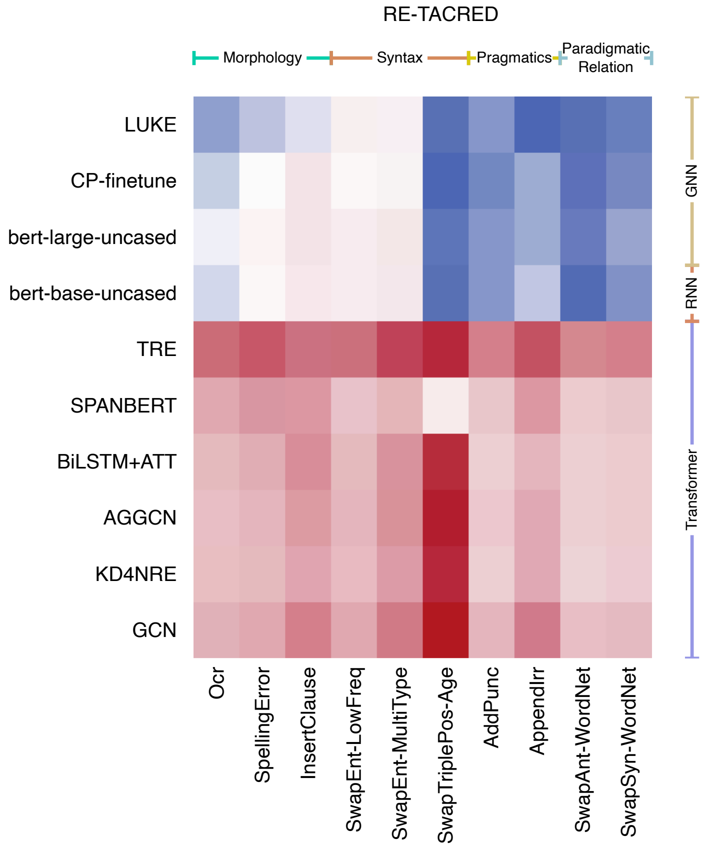


BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding
Deep contextualized word representations

FNLP

2 Robustness Improvement from Representation Perspective---Contextual Embedding





预训练与非预训练鲁棒性对比

FNLP

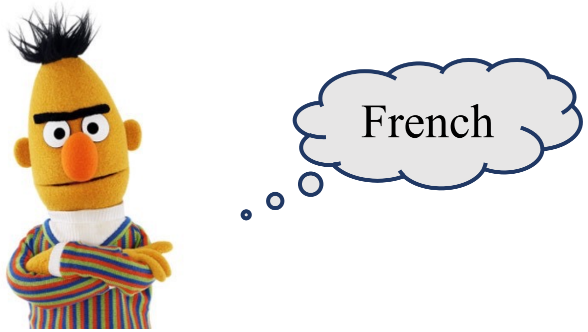
Language Models Need Knowledge

- **PLMs perform poorly on entity recognition**
  - Contextualized PLMs achieved small improvements on entity & semantic related tasks compared with non-contextualized methods. ([Tenney et al.](#))
  - BPE tokenization breaks entities

    `The native language of Jean Mara ##is is French.`
  - Surface form-based reasoning

	original BERT	E-BERT-replace	E-BERT-concat	ERNIE	Know-Bert
Jean Marais	French	French	French	french	french
Daniel Ceccaldi	Italian	French	French	french	italian
Orane Demazis	Albanian	French	French	french	french
Sylvia Lopez	Spanish	French	Spanish	spanish	spanish
Annick Alane	English	French	French	english	english

The native language of Jean Marais is `[MASK]`.



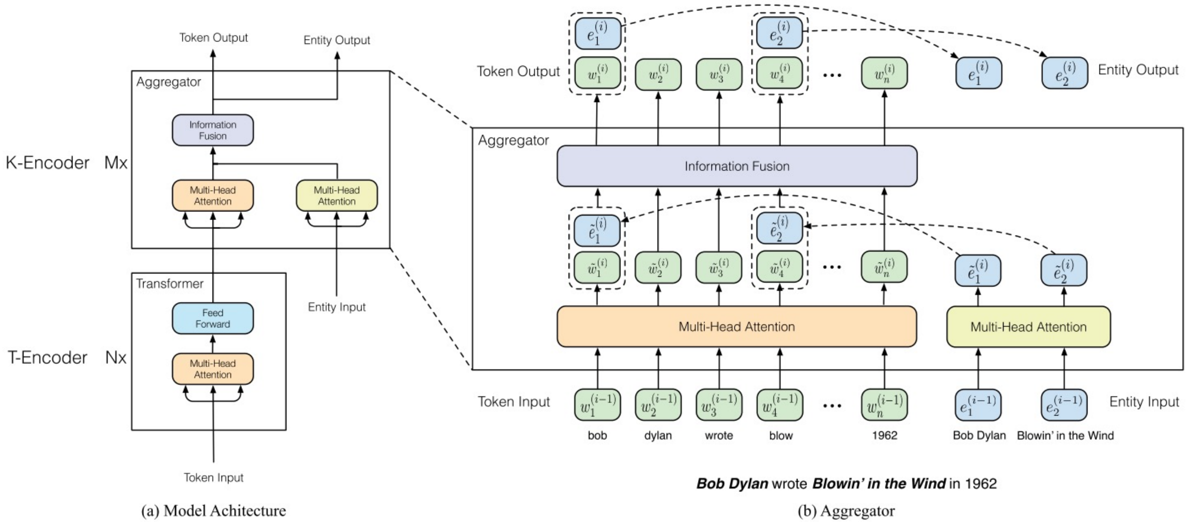
BERT does not know *Daniel Ceccaldi* (as an entity) at all.
It just think *Daniel Ceccaldi* looks like an Italian name.

E-BERT: Efficient-Yet-Effective Entity Embeddings for BERT. https://arxiv.org/abs/1911.03681

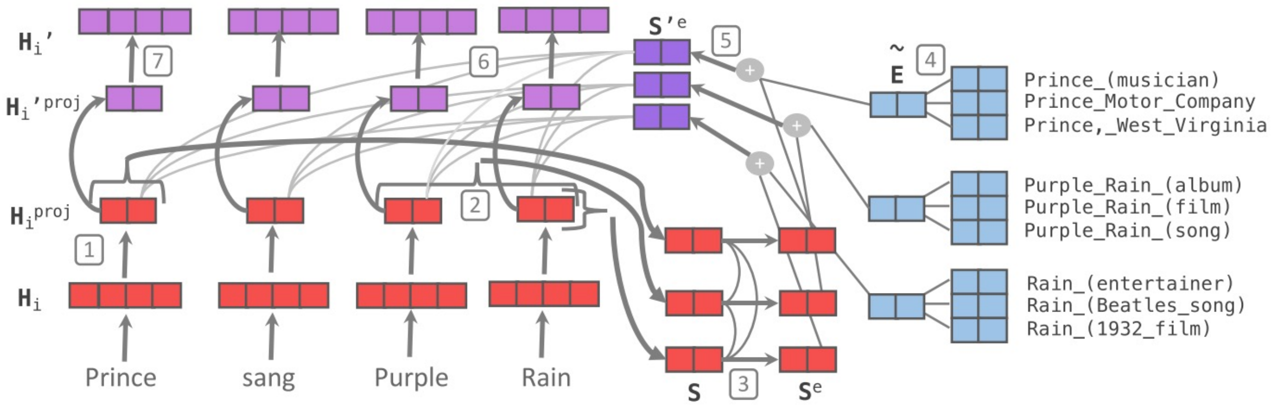
FNLP

Injecting Knowledge into PLMs

- **Injecting entity embeddings**
  - ERNIE, KnowBERT, K-BERT, etc.
  - The entity embeddings are NOT
    - Jointly learned along with PLM
    - Contextualized
- **Knowledge as supervision**
  - WKLM, etc.



ERNIE: Enhanced Language Representation with Informative Entities
https://arxiv.org/abs/1905.07129



Knowledge Enhanced Contextual Word Representations
https://arxiv.org/abs/1909.04164

Model	Acc.	Macro	Micro
NFGEC (Attentive)	54.53	74.76	71.58
NFGEC (LSTM)	55.60	75.15	71.73
BERT	52.04	75.16	71.63
ERNIE	**57.19**	**76.51**	**73.39**

Table 2: Results of various models on FIGER (%).

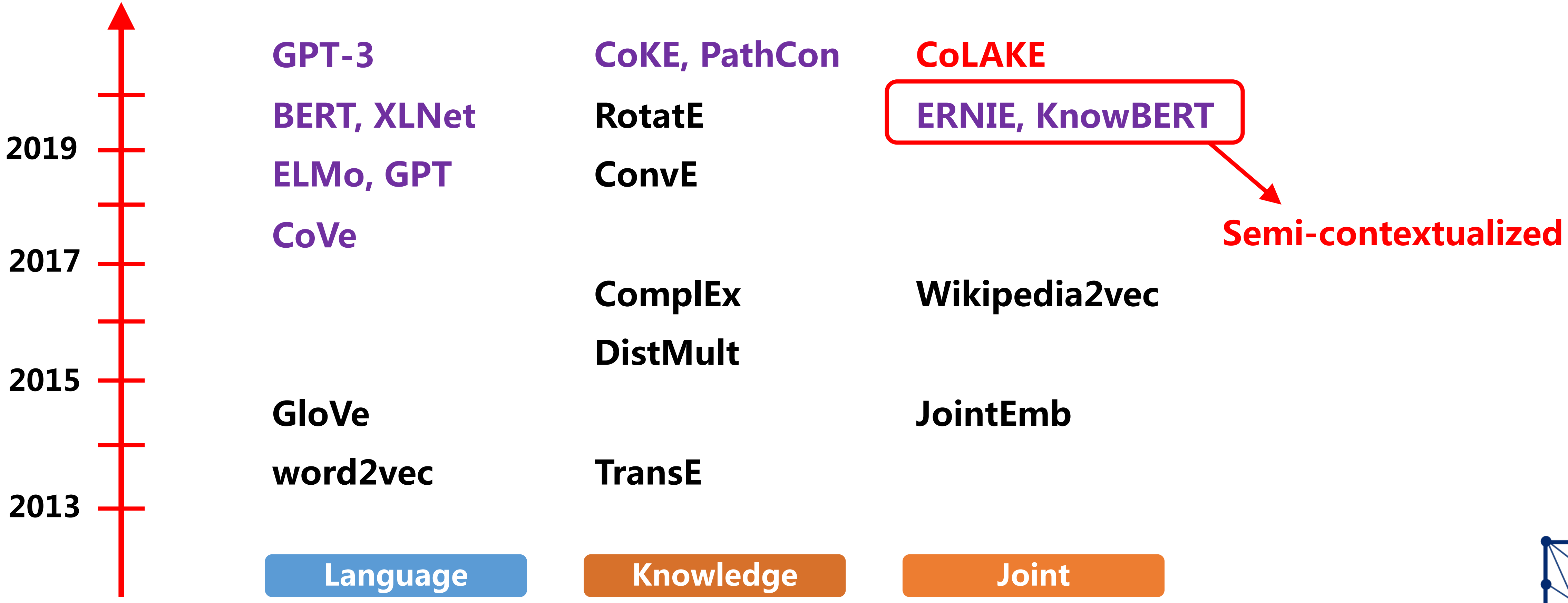
Model	FewRel			TACRED		
	P	R	F1	P	R	F1
CNN	69.51	69.64	69.35	70.30	54.20	61.20
PA-LSTM	-	-	-	65.70	64.50	65.10
C-GCN	-	-	-	69.90	63.30	66.40
BERT	85.05	85.11	84.89	67.23	64.81	66.00
ERNIE	88.49	88.44	**88.32**	69.97	66.08	**67.97**

Table 5: Results of various models on FewRel and TA-CRED (%).

ERNIE 在实体相关的任务上性能更好

Representation in Language and Knowledge

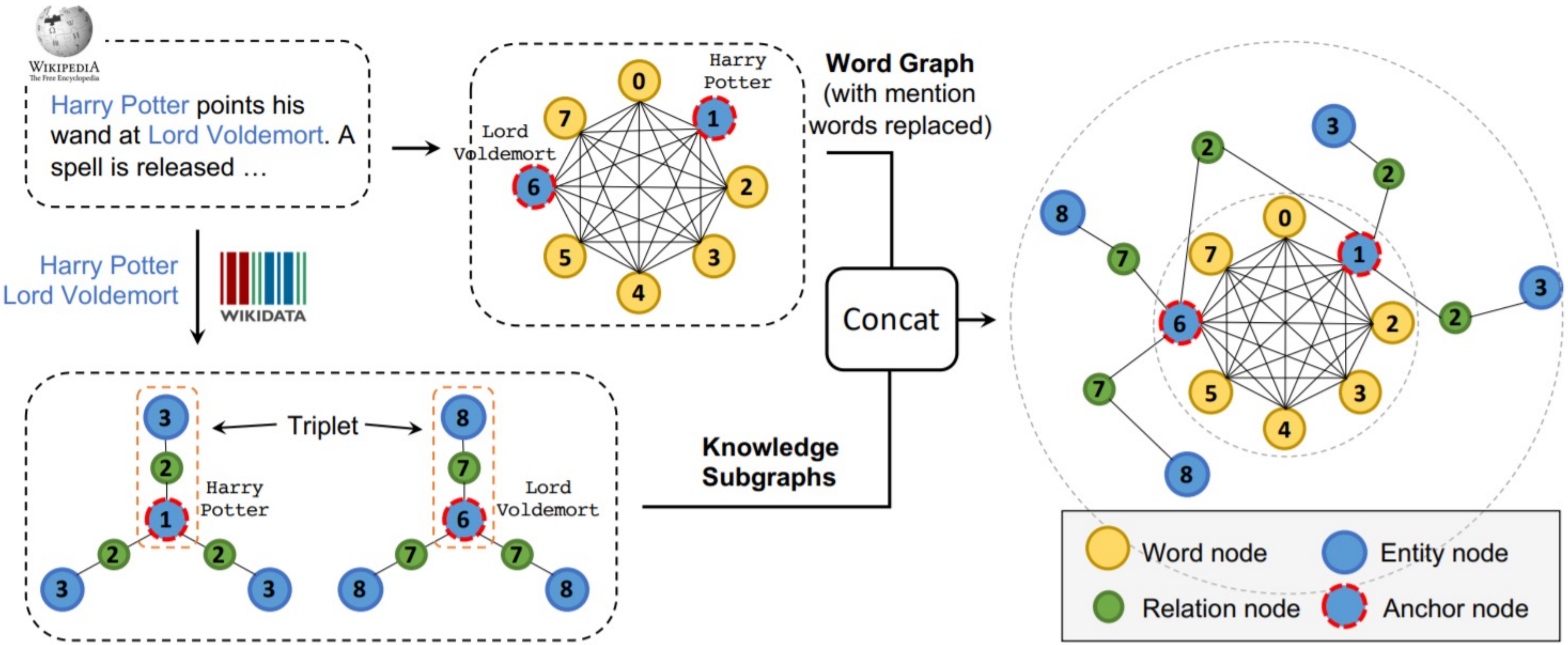
- **Combine the success of both sides -- CoLAKE**



FNLP

2 Representation – CoLAKE Word-knowledge Graph

- Word graph + Knowledge subgraph



Sun et al., CoLAKE: Contextualized Language and Knowledge Embedding, COLING 2020.

2 Representation – CoLAKE Word-knowledge Graph

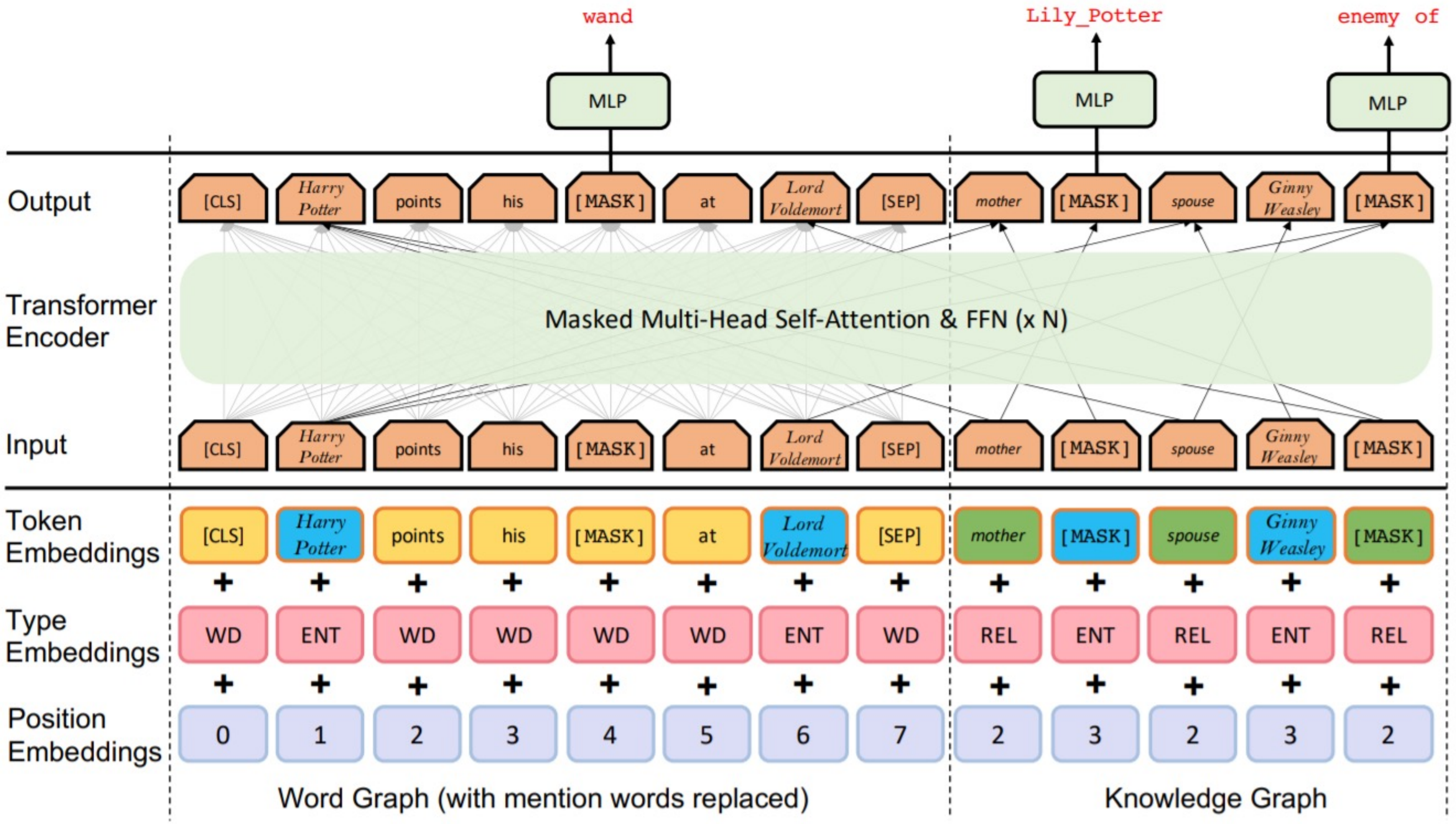
- **Modify Transformer for word-knowledge graph**

Word-knowledge graph is a positional heterogeneous graph.

Position Embedding

Type Embedding

Masked Self-Attention

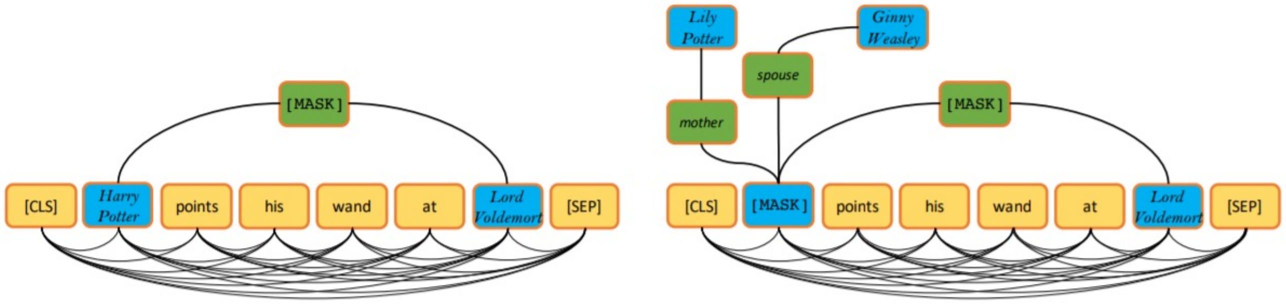


Sun et al., CoLAKE: Contextualized Language and Knowledge Embedding, COLING 2020.

FNLP

2 Representation – CoLAKE Word-knowledge Graph

- **Synthetic graph task**
  - Word-knowledge graph completion

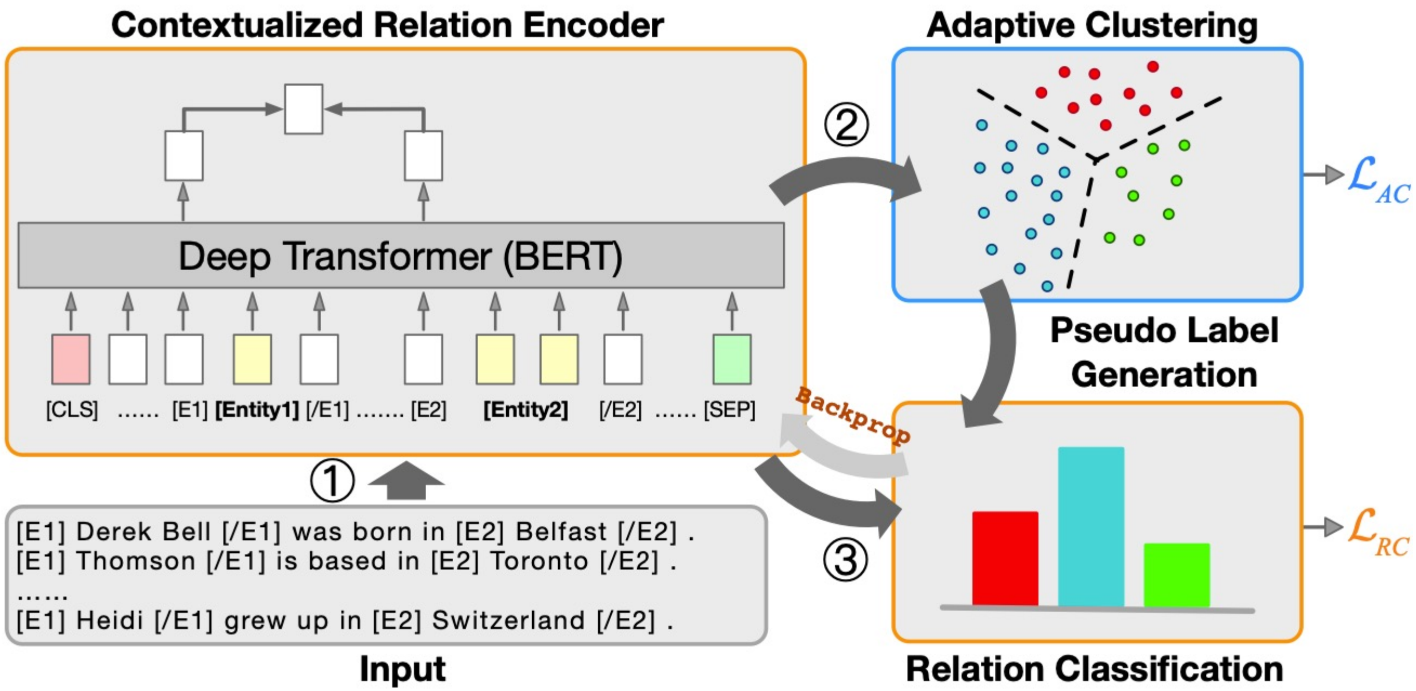


(a) Transductive setting. (b) Inductive setting.

Model	MR ↓	MRR	HITS@1	HITS@3	HITS@10
Transductive setting					
TransE (Bordes et al., 2013)	15.97	67.30	60.28	70.96	79.75
DistMult (Yang et al., 2015)	27.09	60.56	48.66	69.69	79.61
ComplEx (Trouillon et al., 2016)	26.73	61.09	49.80	70.64	79.78
RotatE (Sun et al., 2019)	30.36	70.90	64.74	74.89	81.05
CoLAKE	2.03	82.48	72.14	92.19	98.58
Inductive setting					
DKRL (Xie et al., 2016)	168.21	8.18	5.03	7.28	14.13
CoLAKE	31.01	28.10	15.69	30.28	58.05

Results on word-knowledge graph completion

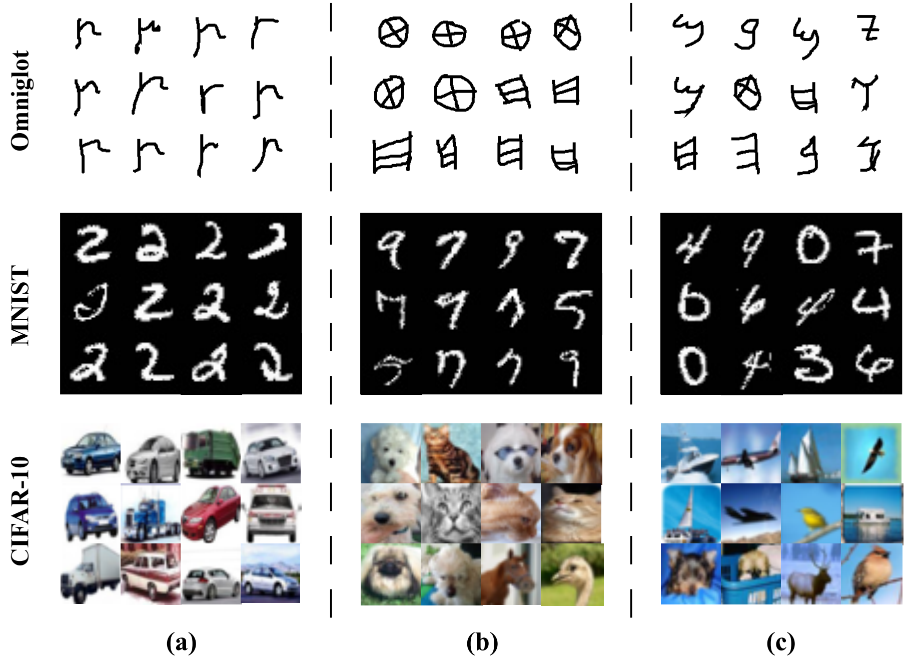
2 Representation---Target Oriented

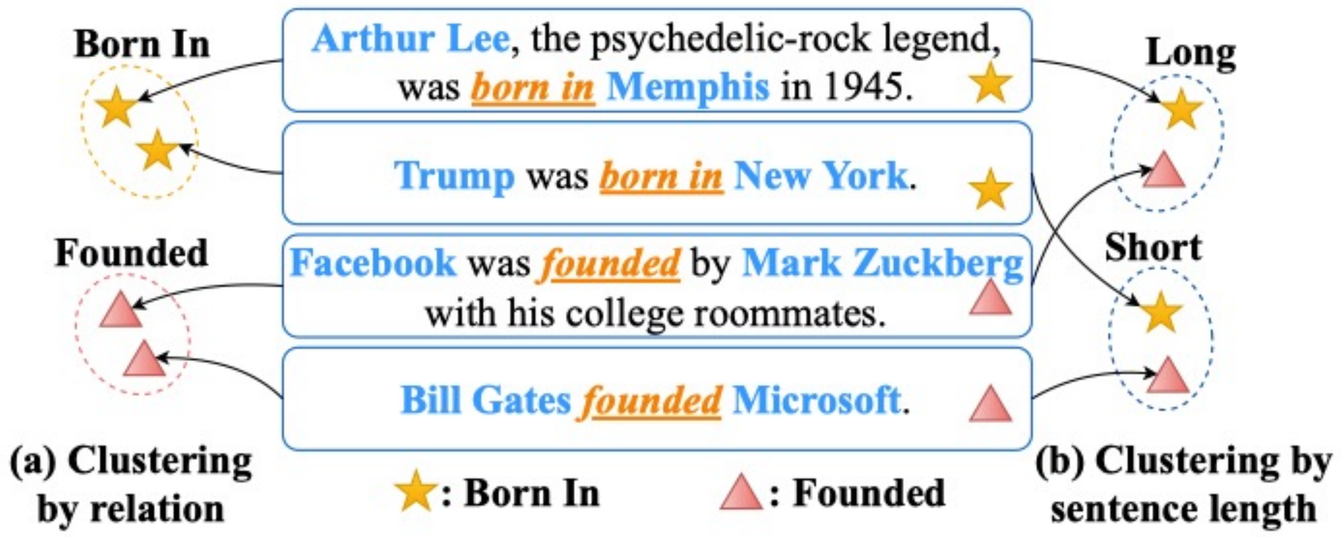


Open Relation Extraction via Self-supervised Learning.

Hu, Xuming, et al. “SelfORE: Self-supervised Relational Feature Learning for Open Relation Extraction.” *EMNLP* 2020.

2 Representation---Target Oriented

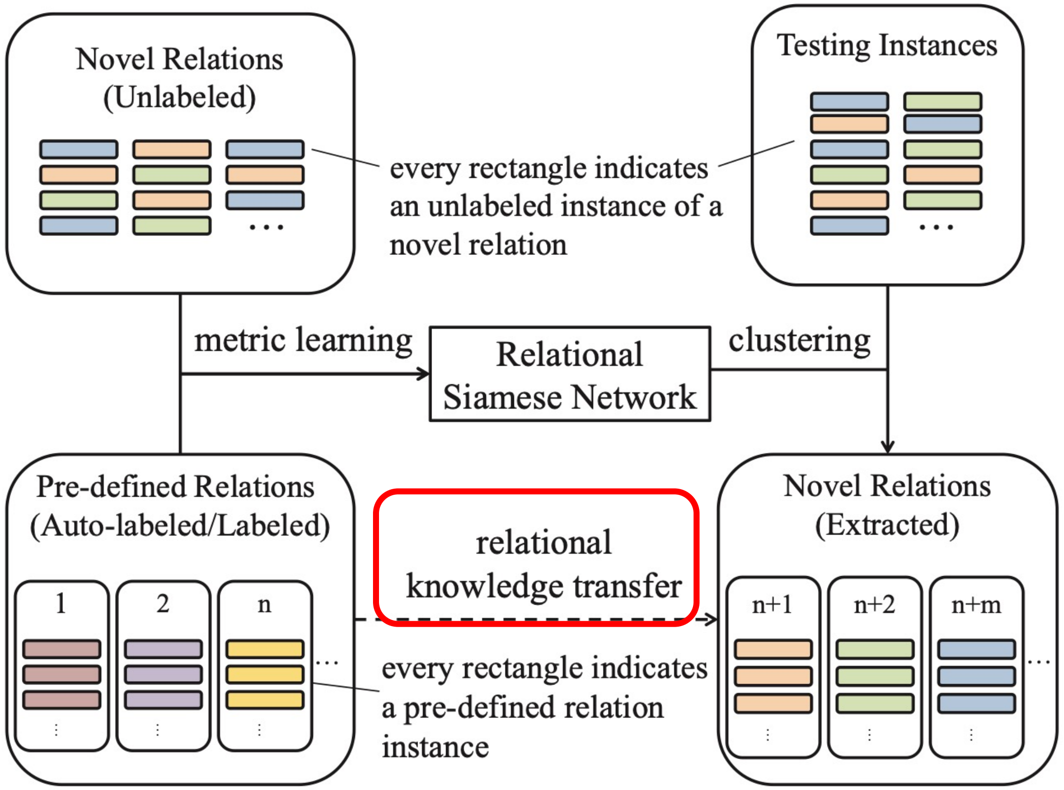




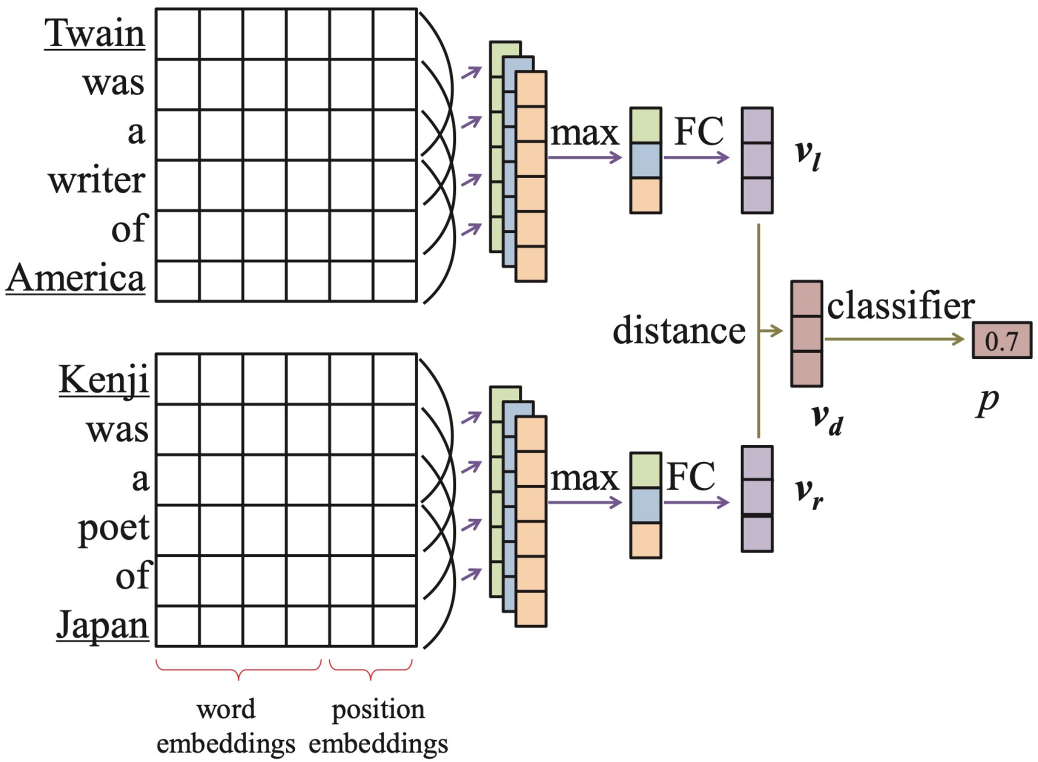
Problem in Unsupervised Clustering

Gui et al., Constructing Multiple Tasks for Augmentation: Improving Neural Image Classification With K-means Features, AAAI 2020.

2 Representation---Target Oriented



Transfer knowledge from Labeled Data



Relational Siamese Networks

Wu et al., Open Relation Extraction: Relational Knowledge Transfer from Supervised Data to Unsupervised Data, EMNLP 2019.

2 Representation---Target Oriented

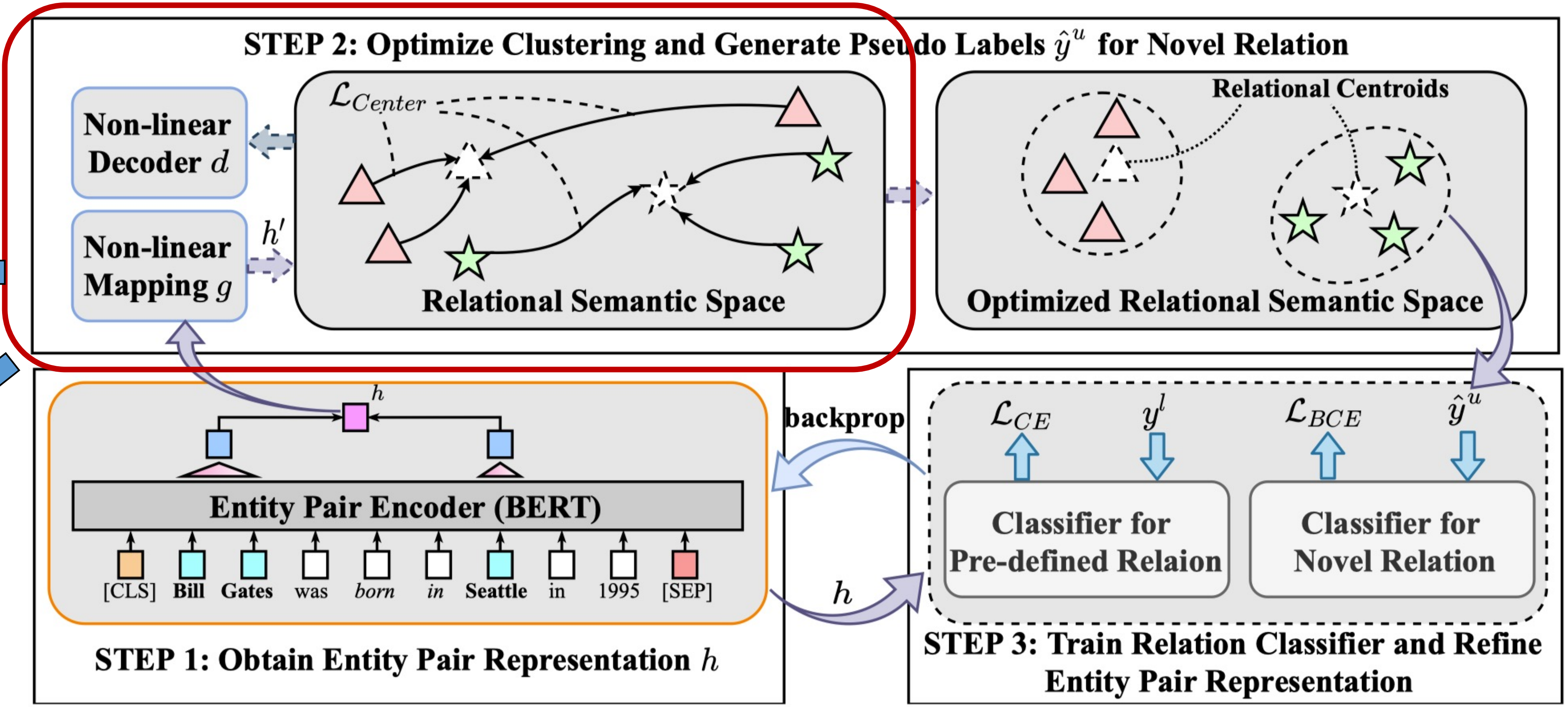
Center Loss for Learning Clustering:

$$\mathcal{L}_{center} = \frac{1}{2}\sum_{i=1}^{N}\left\|\boldsymbol{h}_i^{\boldsymbol{\ell}\prime} - \boldsymbol{c}_{y_i}\right\|_2^2$$

$$\boldsymbol{c}_r = \frac{1}{|\mathcal{D}_r|}\sum_{i\in\mathcal{D}_r}\boldsymbol{h}_i^{\ell\prime},$$

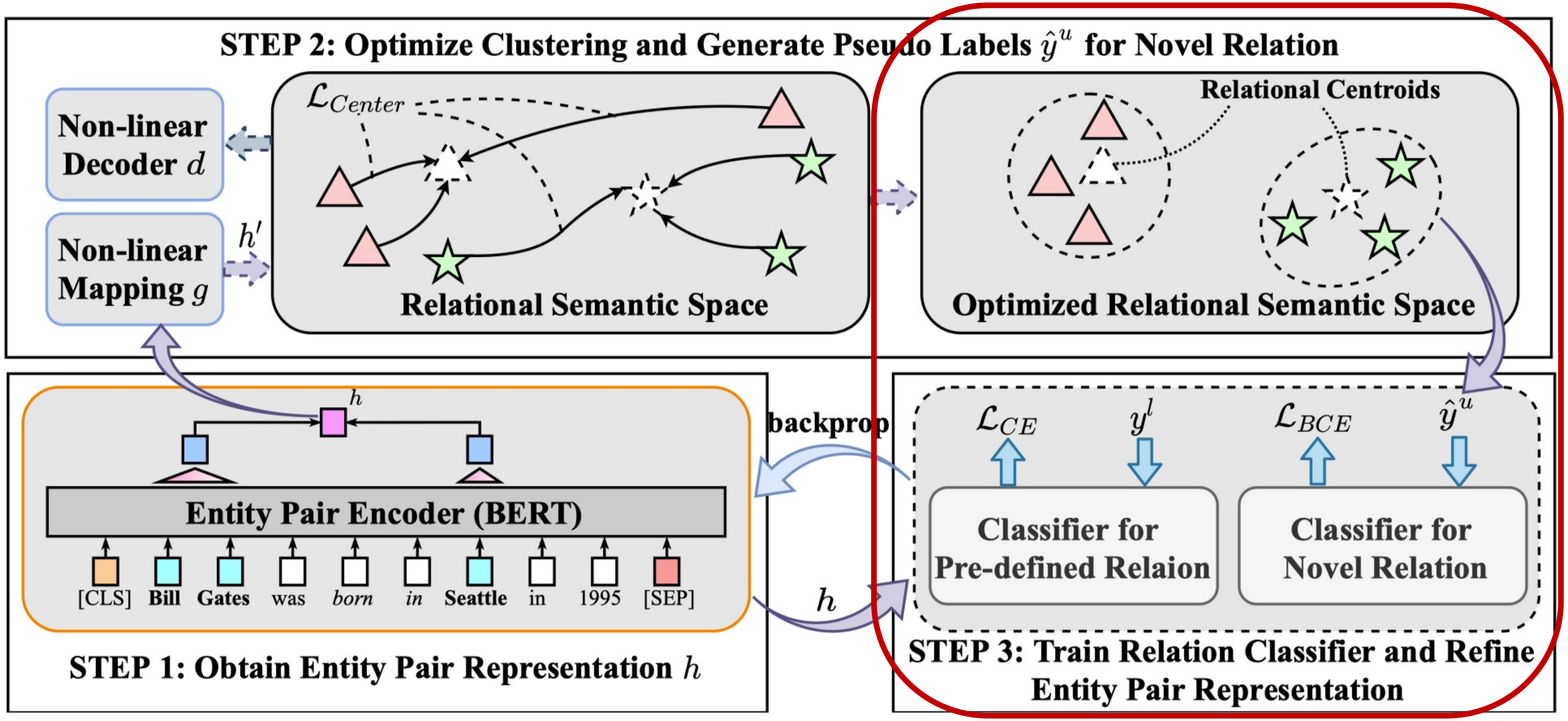
Reconstruction Loss Avoids Trivil Solution:

$$\mathcal{L}_C = \sum_{i=1}^{N}\ell(\boldsymbol{g}(\boldsymbol{h}_i'), \boldsymbol{h}_i)$$



Overview of our proposed RoCORE method.

Zhao, Jun, et al. "A Relation-Oriented Clustering Method for Open Relation Extraction." *EMNLP* 2021.



Kmeans Clustering for Pseudo Lables

$$\hat{y}^u = \text{k-means}(\boldsymbol{h}^{u\prime})$$

BCE Loss for New Classifier

$$\mathcal{L}_{BCE}(s_i^u, s_j^u) = q_{i,j}\mathcal{L}^+(s_i^u, s_j^u) + (1 - q_{ij})\mathcal{L}^-(s_i^u, s_j^u).$$

Overview of our proposed RoCORE method.

Zhao, Jun, et al. "A Relation-Oriented Clustering Method for Open Relation Extraction." *EMNLP* 2021.

2 Representation---Target Oriented

0-63 Labeled
64-79 Unlabeled

0-30 Labeled
31-40 Unlabeled

Method	FewRel			TACRED			FewRel→TACRED			TACRED→FewRel		
	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
VAE	0.179	0.697	0.285	0.234	0.606	0.338	0.234	0.606	0.338	0.179	0.697	0.285
RW-HAC	0.318	0.460	0.376	0.534	0.568	0.551	0.534	0.568	0.551	0.318	0.460	0.376
RSN	0.496	0.734	0.592	0.628	0.634	0.631	0.398	0.617	0.484	0.233	0.567	0.331
RSN+BERT	0.585	**0.899**	0.709	0.795	**0.878**	0.834	0.379	**0.843**	0.523	0.262	**0.898**	0.406
RoCORE	**0.749**$_{12}$	0.846$_{09}$	**0.794**$_{09}$	**0.871**$_{42}$	0.849$_{35}$	**0.860**$_{34}$	**0.616**$_{33}$	0.599$_{57}$	**0.607**$_{41}$	**0.687**$_{35}$	0.766$_{46}$	**0.724**$_{25}$

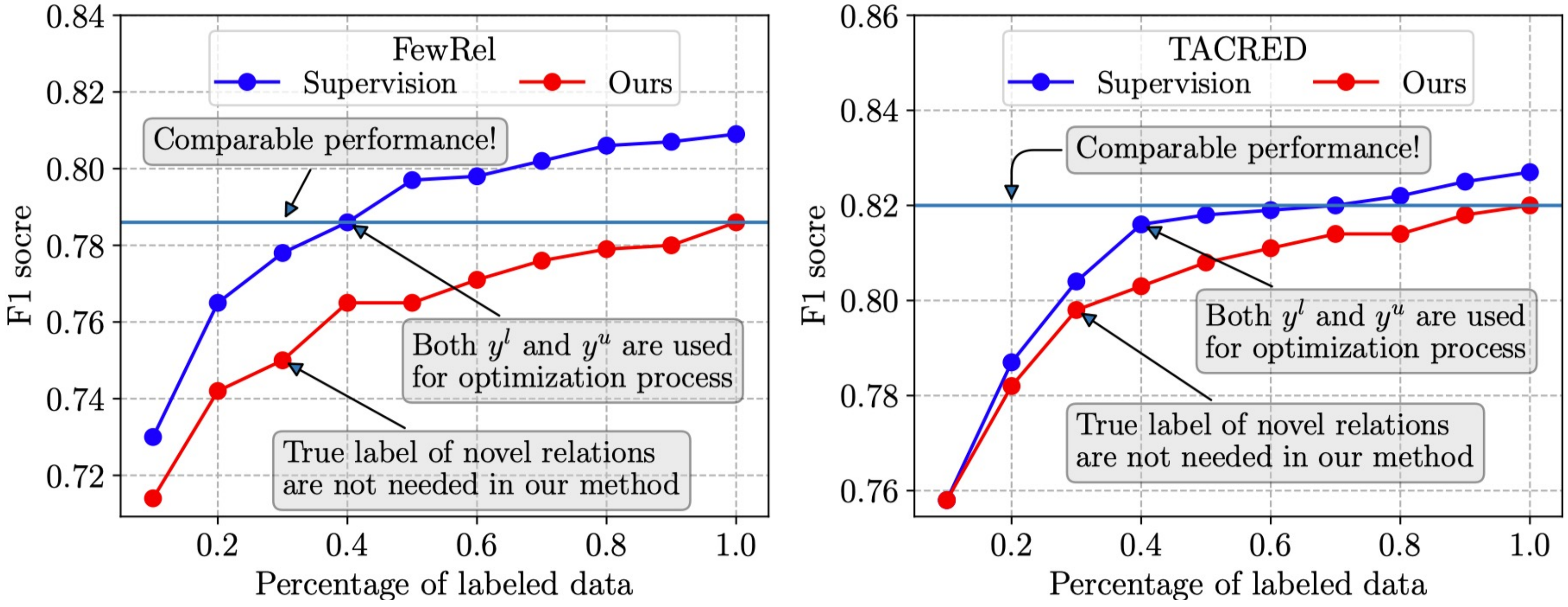
In Domain

Cross Domain

Zhao, Jun, et al. "A Relation-Oriented Clustering Method for Open Relation Extraction." *EMNLP* 2021.

FNLP

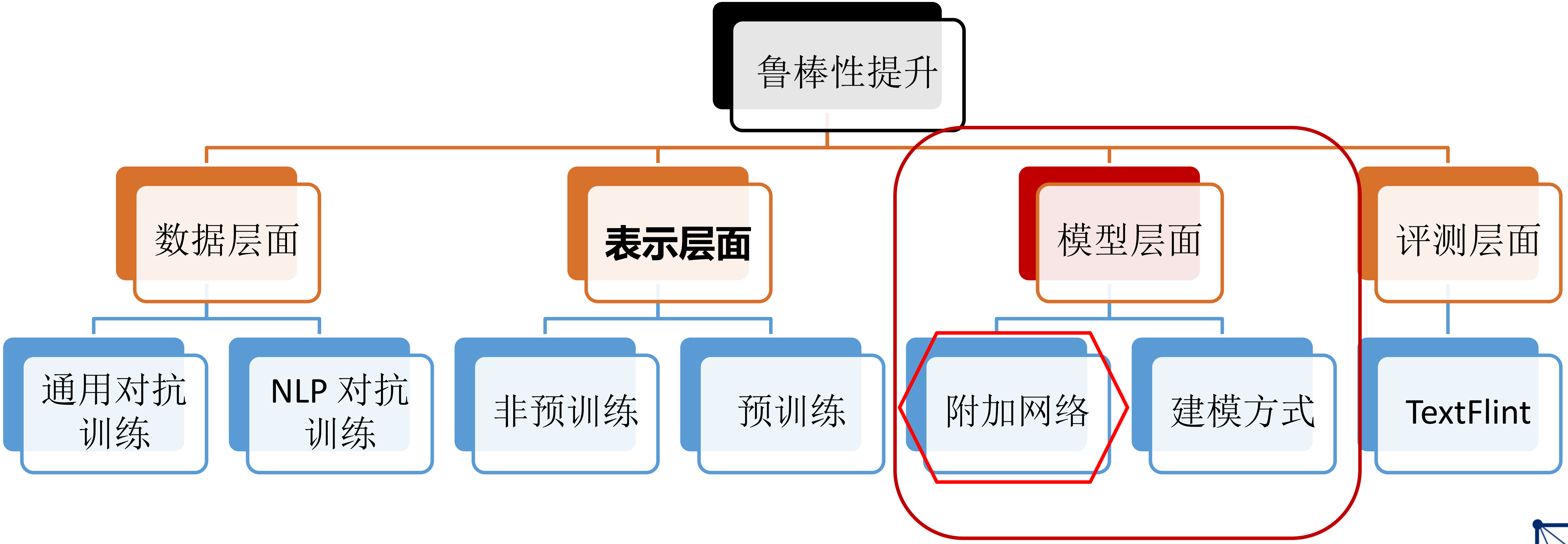
2 Representation---Target Oriented



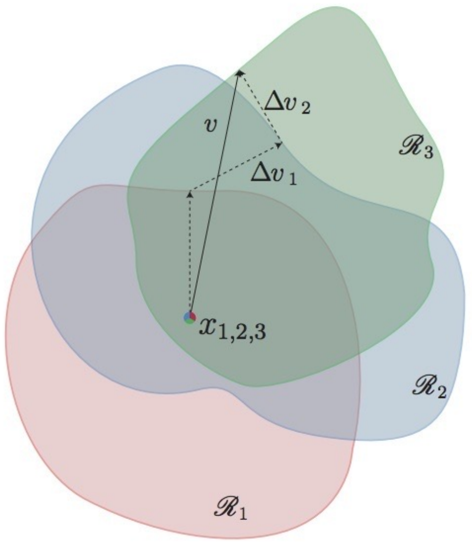
Model performance with different amounts of labeled data.

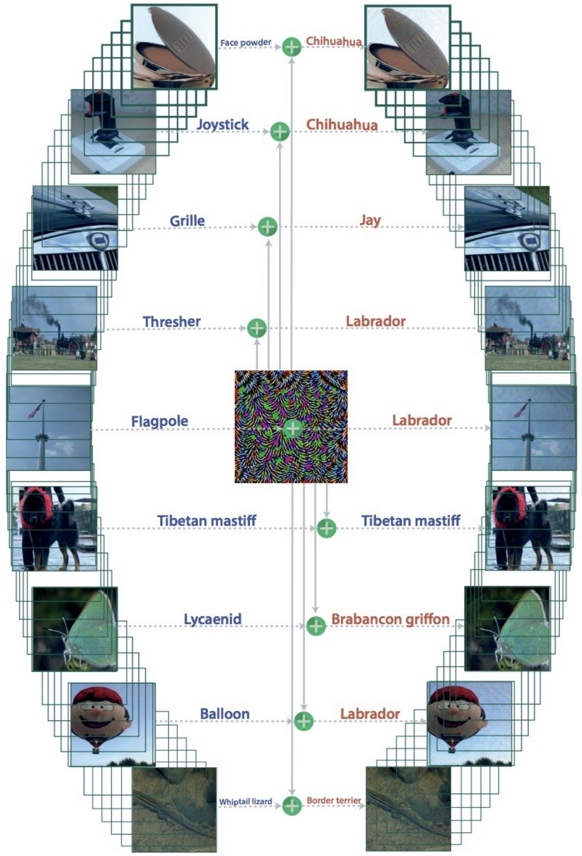
Zhao, Jun, et al. “A Relation-Oriented Clustering Method for Open Relation Extraction.” *EMNLP* 2021.

0 General Framework of Natural Language Processing



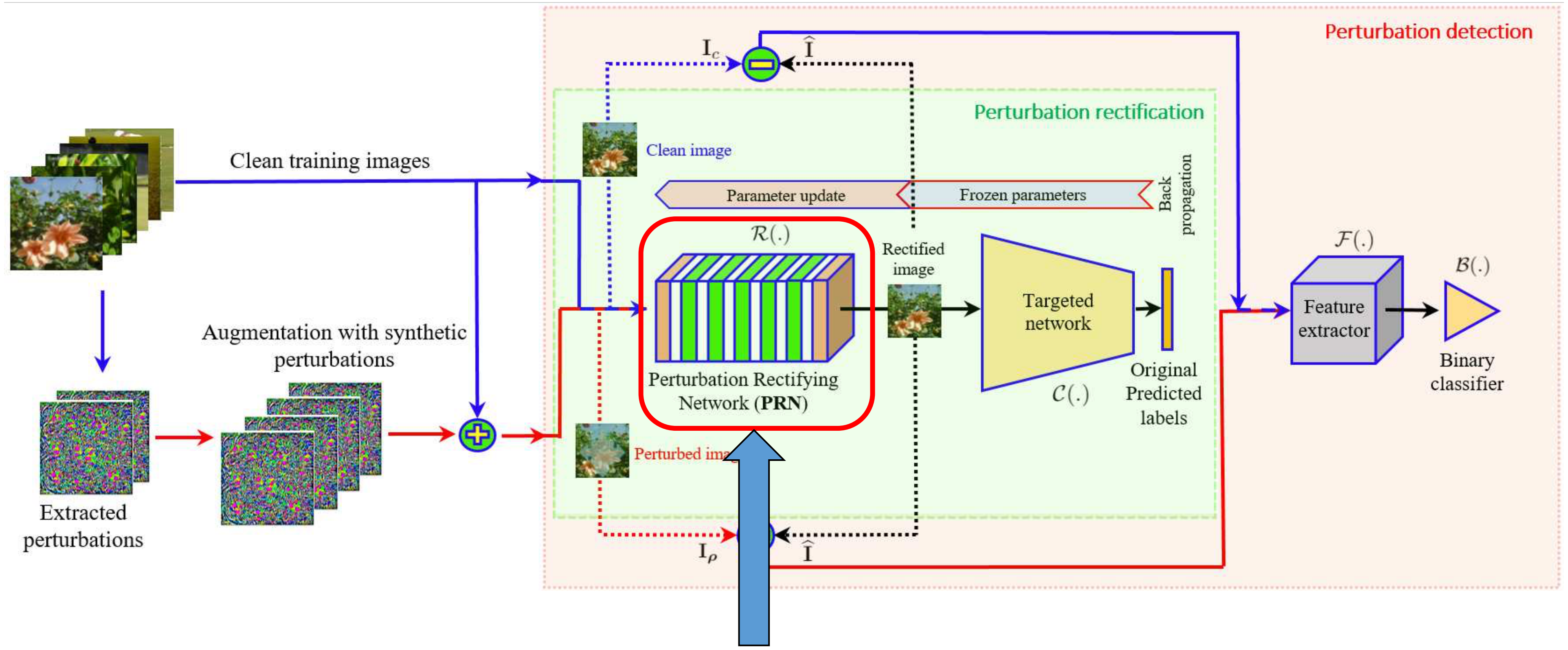
3 Robustness Improvement form Model Perspective---Affiliate network





Moosavi-Dezfooli, Seyed-Mohsen, et al. "Universal adversarial perturbations." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.

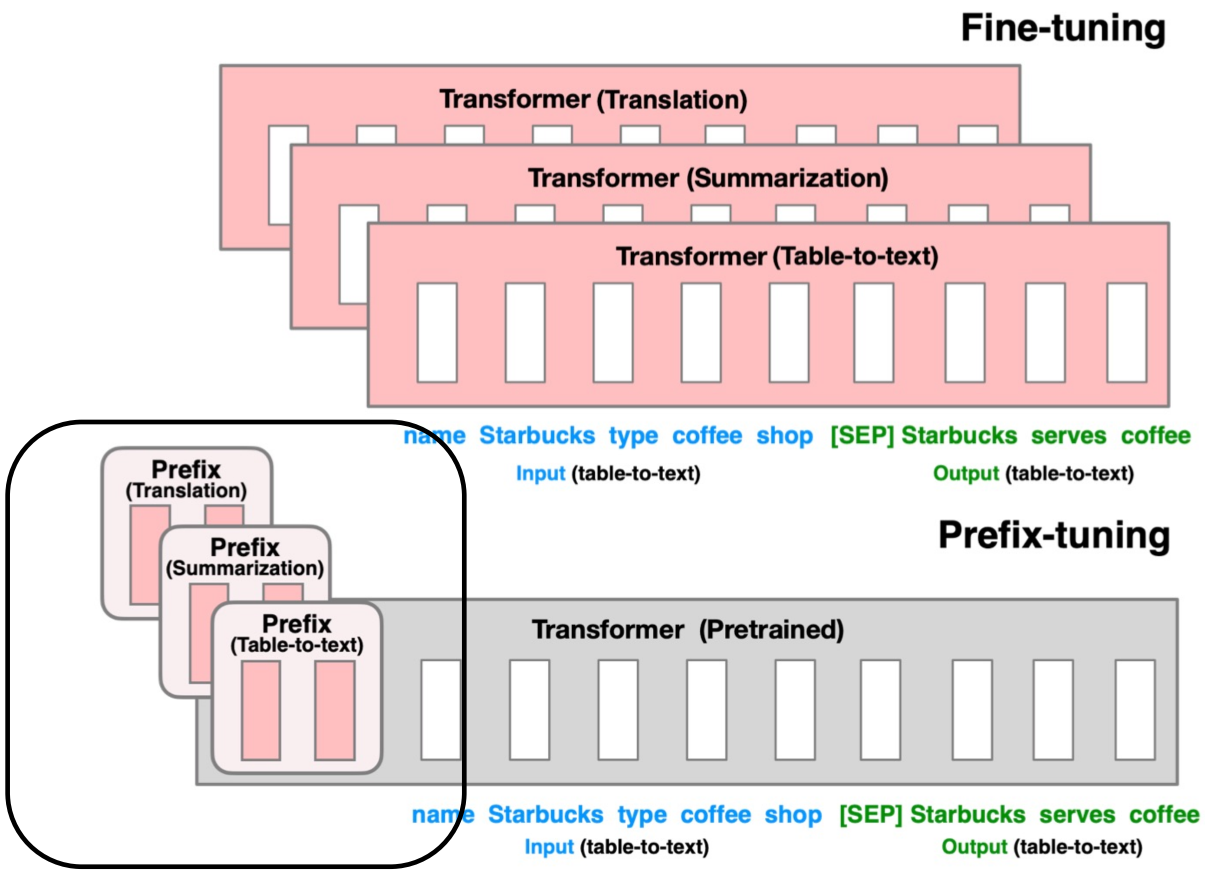
3 Robustness Improvement form Model Perspective---Affiliate network



插件模块

Akhtar, Naveed, Jian Liu, and Ajmal Mian. "Defense against universal adversarial perturbations." *CVPR* 2018.

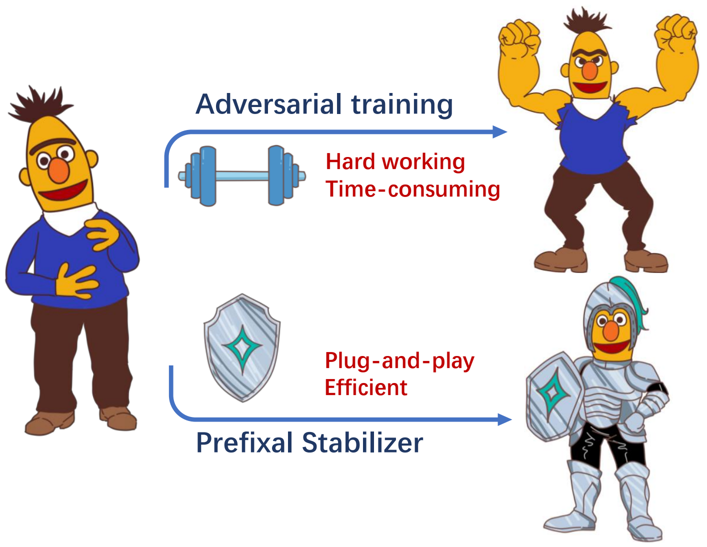
3 Robustness Improvement form Model Perspective---Affiliate network



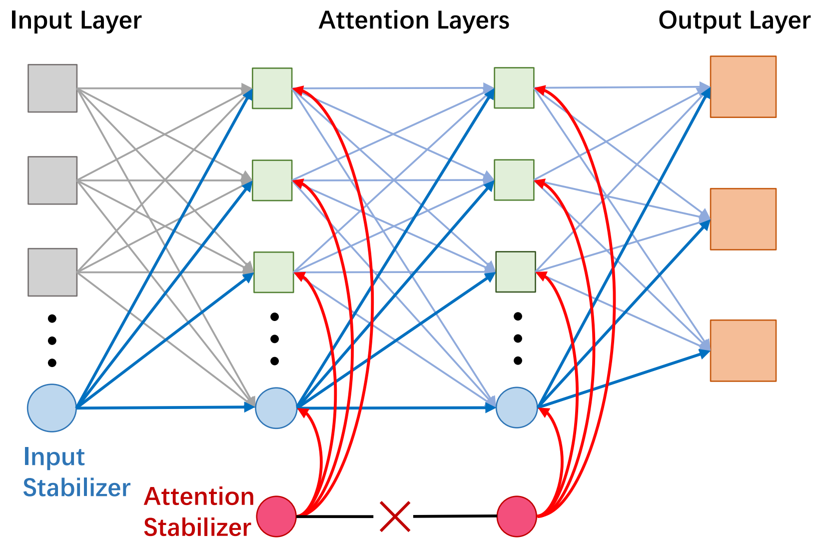
Prefix Tuning

Li, Xiang Lisa, and Percy Liang. "Prefix-tuning: Optimizing continuous prompts for generation." *arXiv preprint arXiv:2101.00190* (2021).

3 Robustness Improvement form Model Perspective---Affiliate network



Adversarial training VS Prefixal Stabilizer



Architecture of Prefixal Stabilizer

Rui Zheng et al., Prefixal Stabilizer: A Plug and Play Defense Moduleto Defend Textual Adversarial Attack

FNLP

Robustness Improvement form Model Perspective---Affiliate network

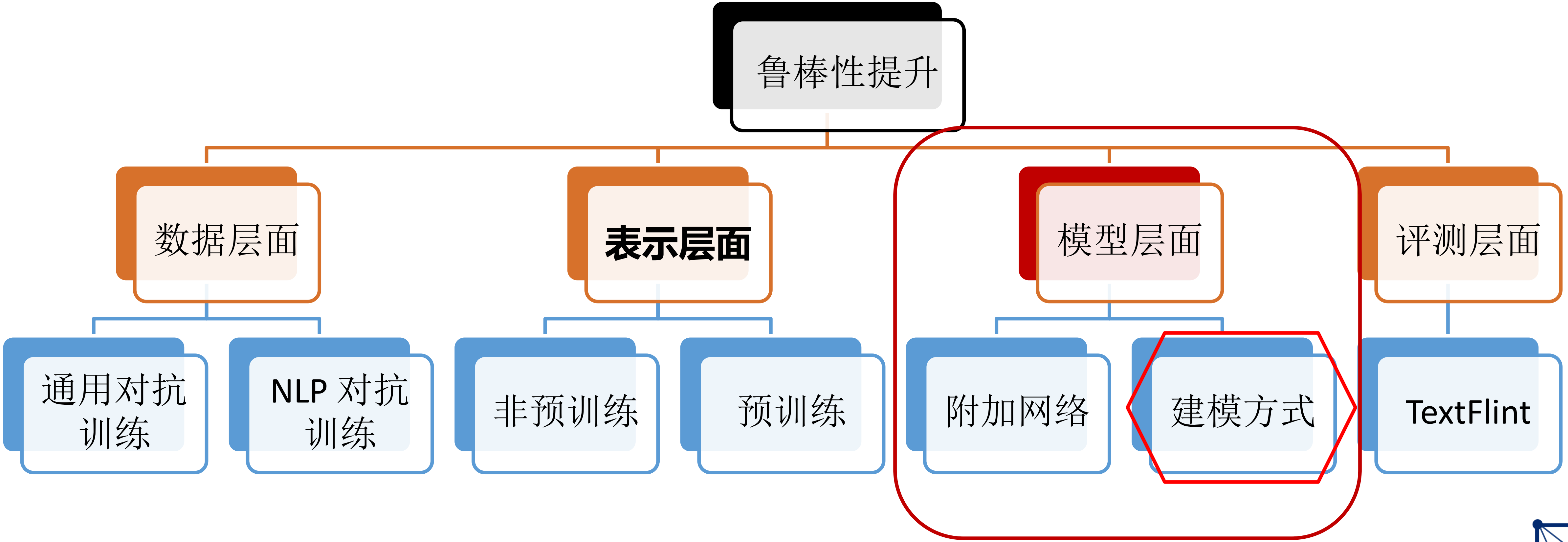
Methods	*Clean%*	TextFooler			PWWS			TextBugger		
		Aua%	*Suc%*	*#Query*	*Aua%*	*Suc%*	*#Query*	*Aua%*	*Suc%*	*#Query*
Baseline (BERT)	94.2	9.5	89.9	1007.8	20.0	78.8	2654.2	6.0	93.6	1038.7
PGD (Madry et al. 2017)	94.6	25.5	73.0	1416.0	35.0	68.3	2741.5	23.0	75.9	1625.5
FreeLB (Zhu et al. 2019)	**94.7**	18.0	81.0	1160.7	25.5	73.1	2699.4	19.0	79.9	1445.0
TAVAT (Li and Qiu 2020)	94.3	14.0	85.2	1117.7	14.5	84.6	2625.4	12.0	87.3	1122.6
InfoBERT (Wang et al. 2020)	94.5	23.5	75.1	1374.7	27.5	70.9	2655.9	21.0	77.8	1420.6
MixADA (Si et al. 2021)	93.9	10.5	88.8	1023.4	23.5	75.0	2656.6	6.5	93.1	1077.3
Stabilizer-PGD	93.9	**33.5**	**64.3**	**1626.6**	**39.0**	**58.5**	**2777.1**	**32.5**	**65.4**	**1845.3**
Stabilizer-FreeLB	94.1	24.0	74.5	1300.3	32.5	65.5	2772.6	23.0	75.6	1472.7

Table 2: Experiment results of different defenders on IMDB, where all models are trained on BERT. The best performance is marked in bold.

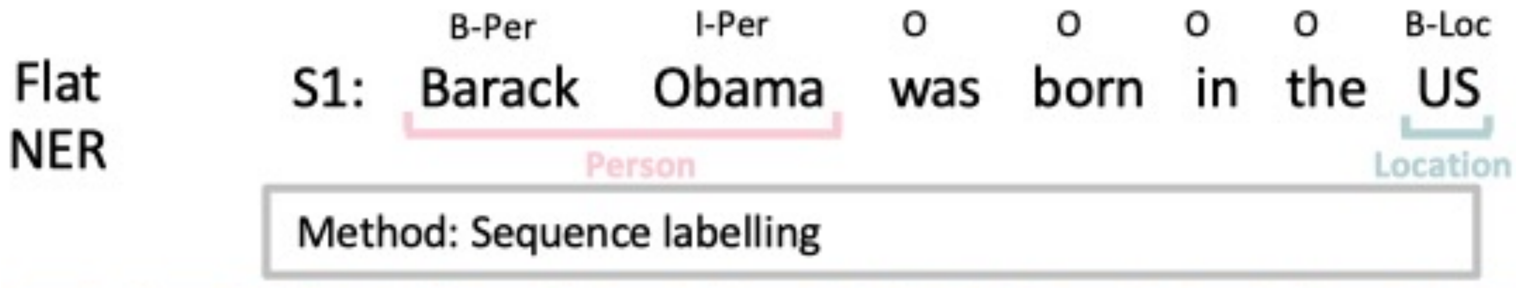
Methods	**SST-2** to IMDB			
	Clean%	*Aua%*	*Suc%*	*#Query*
BERT (SST-2)	90.8	2.0	97.8	564.1
PGD (SST-2)	89.9	20.0	77.8	827.8
FreeLB (SST-2)	89.2	22.0	75.3	874.2
Stabilizer-PGD (SST-2) +BERT (IMDB)	93.7	**36.0**	**61.6**	**1664.5**
Stabilizer-FreeLB (SST-2) +BERT (IMDB)	**93.8**	34.0	63.7	1536.0

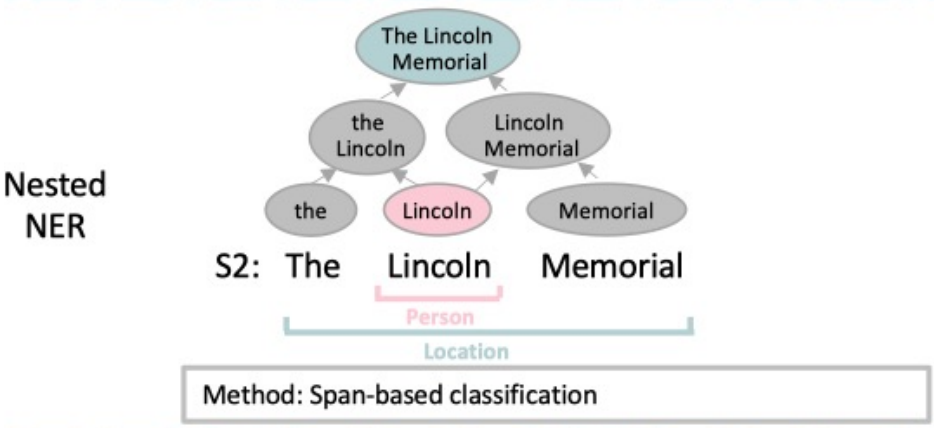
Table 3: Experimental results when the models **trained on SST-2** is transferred to **IMDB for testing**. BERT (fine-tuned on IMDB) equipped with our module (trained on SST-2) has improved robustness to a large extent and has little damage to cleaning accuracy.

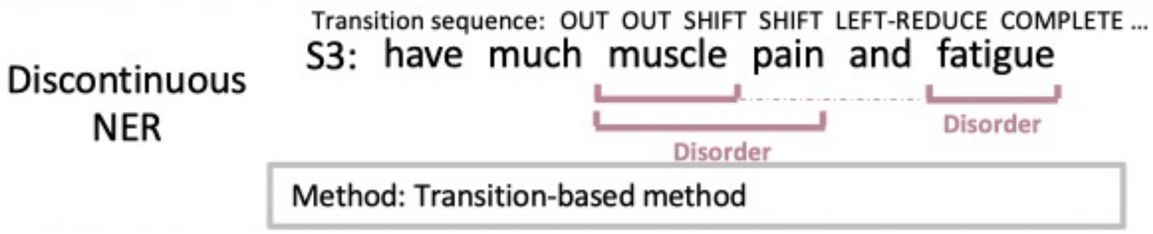
Rui Zheng et al., Prefixal Stabilizer: A Plug and Play Defense Moduleto Defend Textual Adversarial Attack


鲁棒性提升
数据层面
表示层面
模型层面
评测层面
通用对抗训练
NLP 对抗训练
非预训练
预训练
附加网络
建模方式
TextFlint

3 Robustness Improvement form Model Perspective---Modeling Method



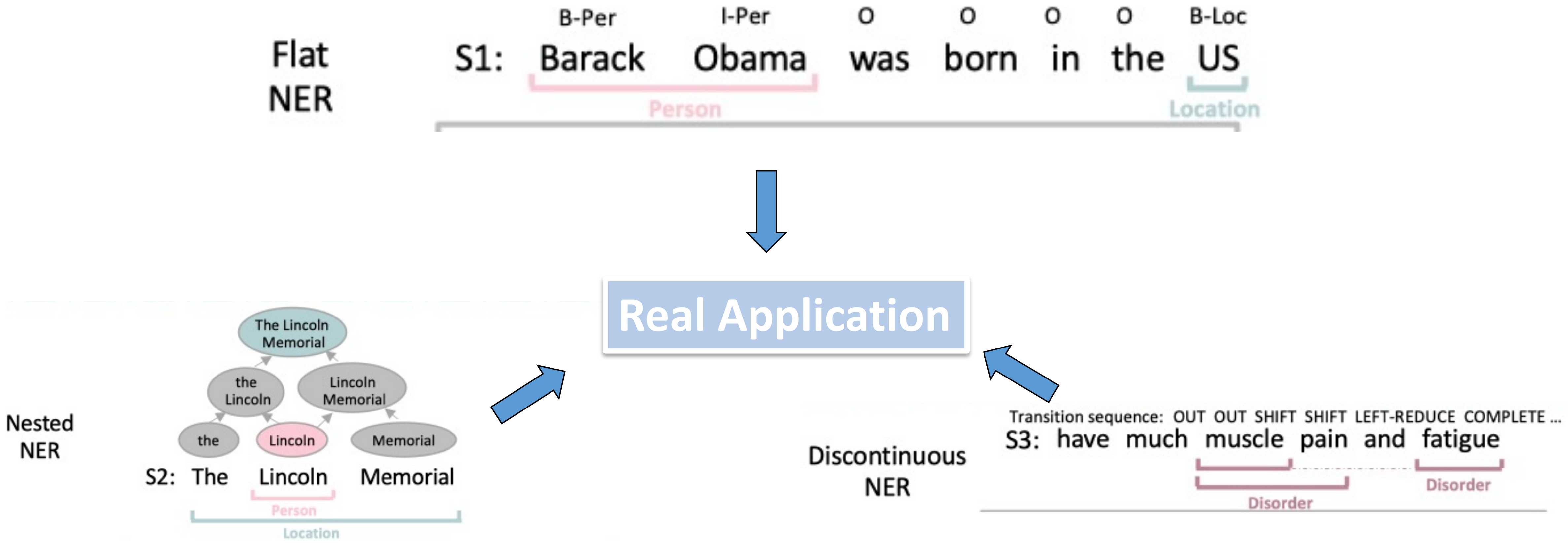




Three kinds of named entities

Yan et al., A Unified Generative Framework for Various NER tasks, ACL 2021

3 Robustness Improvement form Model Perspective---Modeling Method

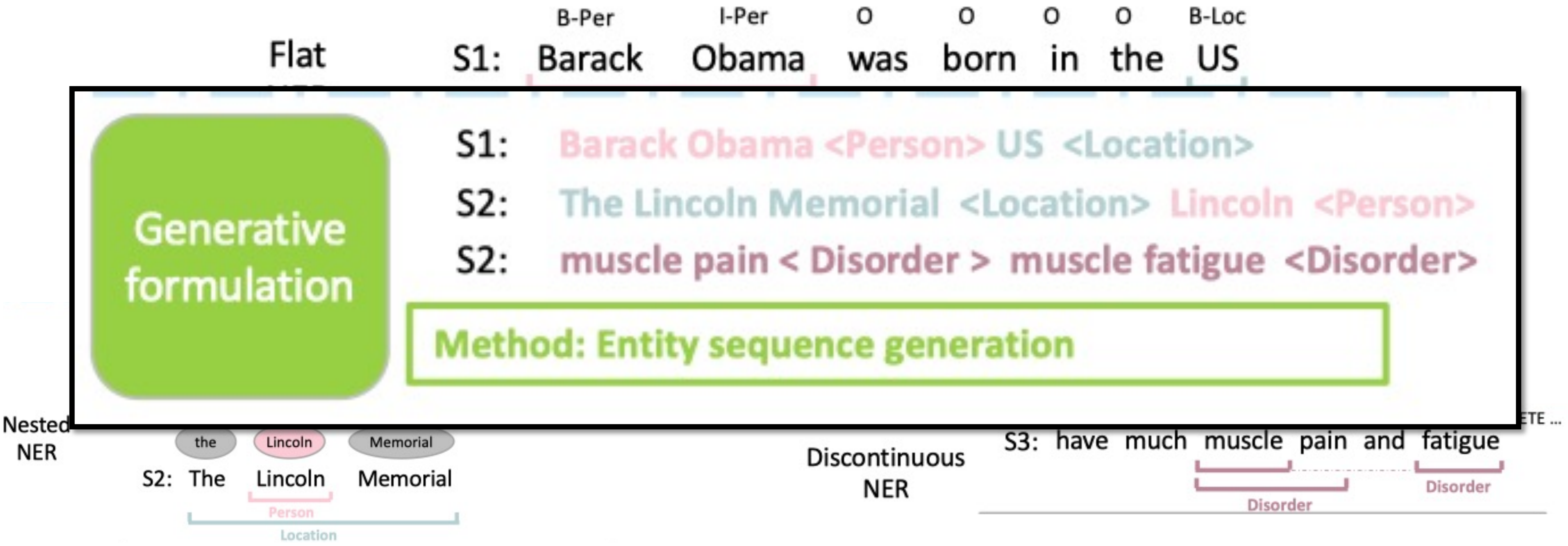


The different formulations make it hard to solve all NER tasks in a unified method

Yan et al., A Unified Generative Framework for Various NER tasks, ACL 2021

FNLP

3 Robustness Improvement form Model Perspective---Modeling Method



The different formulations make it hard to solve all NER tasks in a unified method

Yan et al., A Unified Generative Framework for Various NER tasks, ACL 2021

3 Robustness Improvement form Model Perspective---Modeling Method

Input: <s> have muscle pain and fatigue </s>
Output: 2 3 6 2 5 </s>

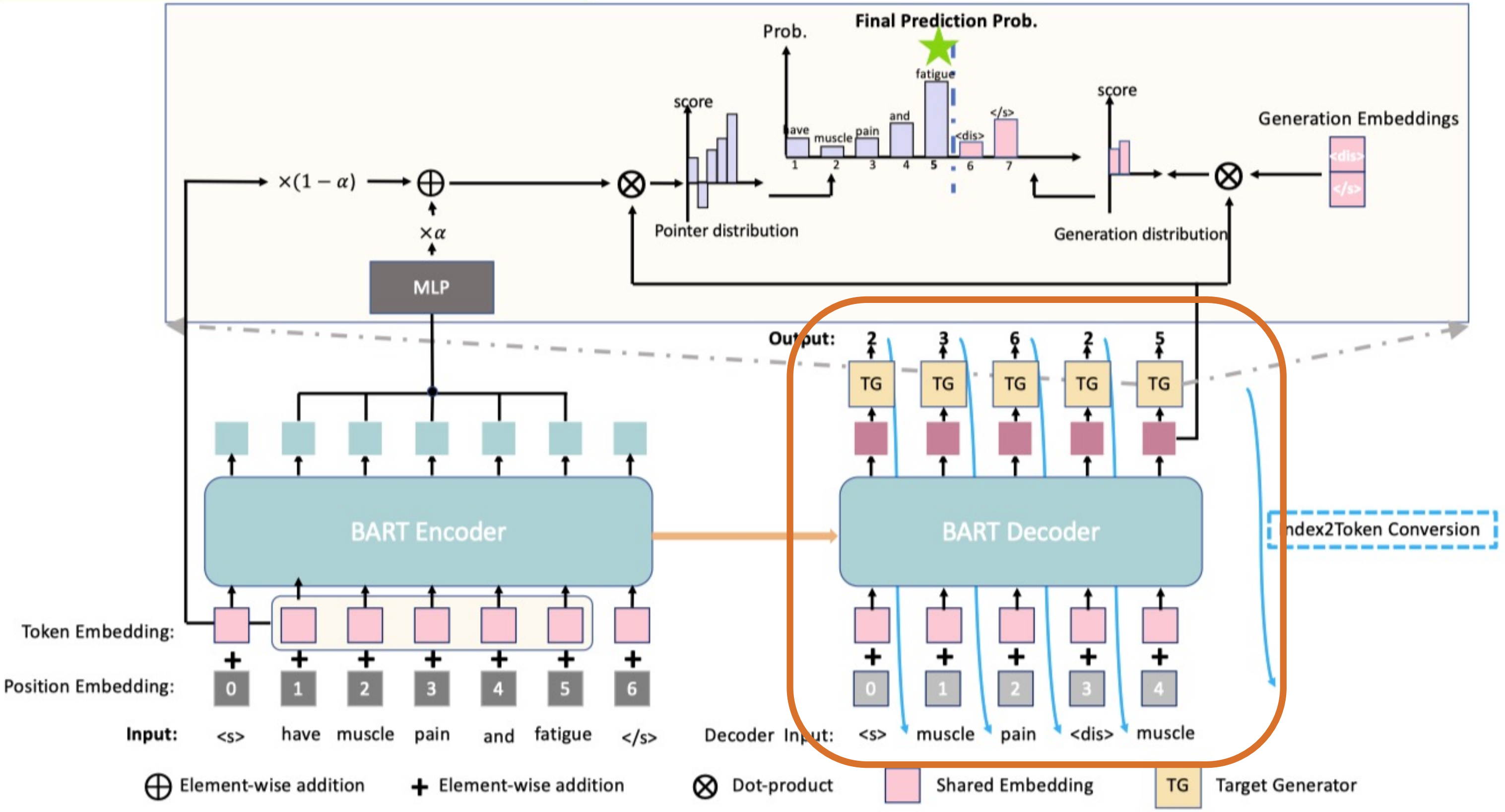
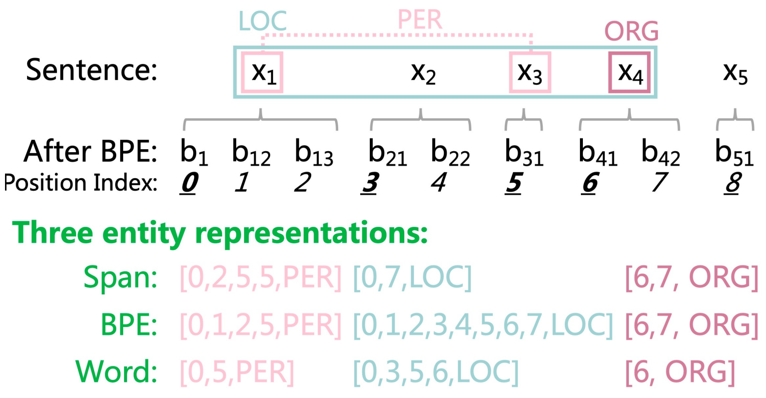


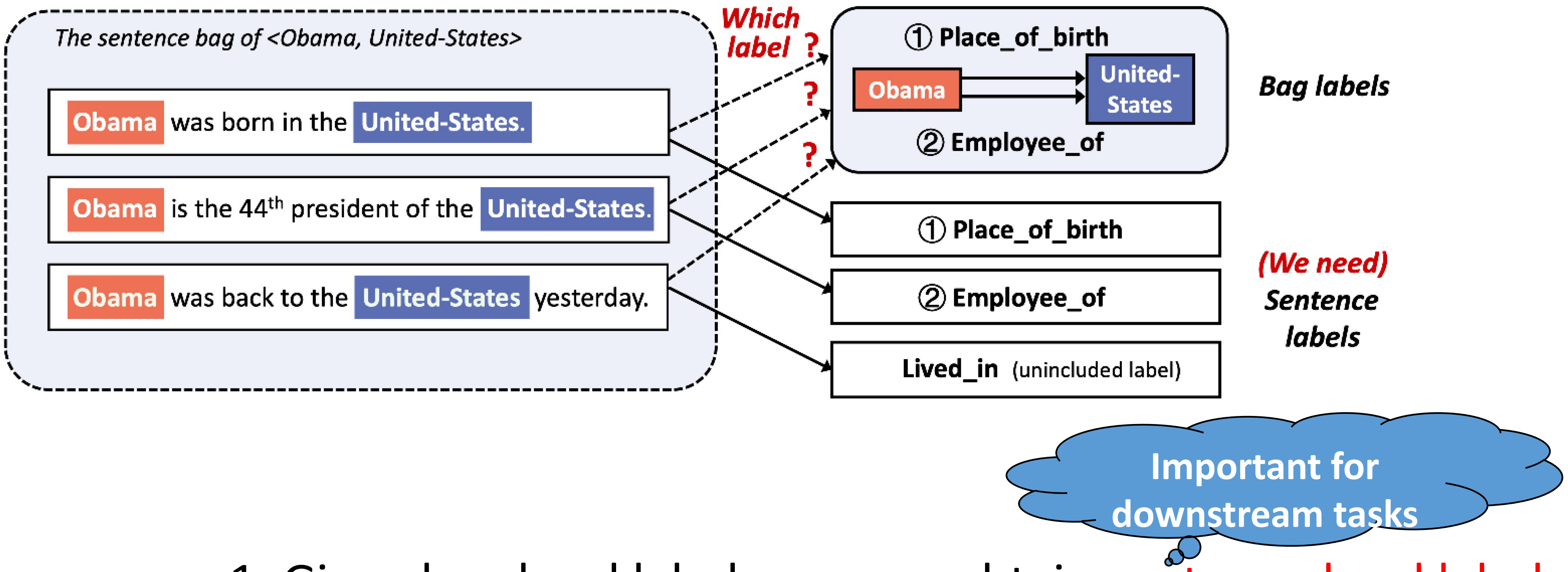
Figure 2: Model structure used in our method. “<s>” and “</s>” are the predefined start-of-sentence and end-of-sentence tokens in BART. We assume the there is only one entity tag “<dis>”, therefore, only this token and the “</s>” token need to be generated, other tokens can be generated by the pointer.



Three kinds of entity representations

Yan et al., A Unified Generative Framework for Various NER tasks, ACL 2021

Robustness Improvement form Model Perspective---Modeling Method



1. Given bag-level labels, can we obtain sentence-level labels?
2. Sentence bag contains correct labels, incorrect labels, and unincluded labels.
3. Previous positive learning framework cannot distinguish noisy data.

SENT: Sentence-level Distant Relation Extraction via Negative Training

FNLP

Robustness Improvement form Model Perspective---Modeling Method

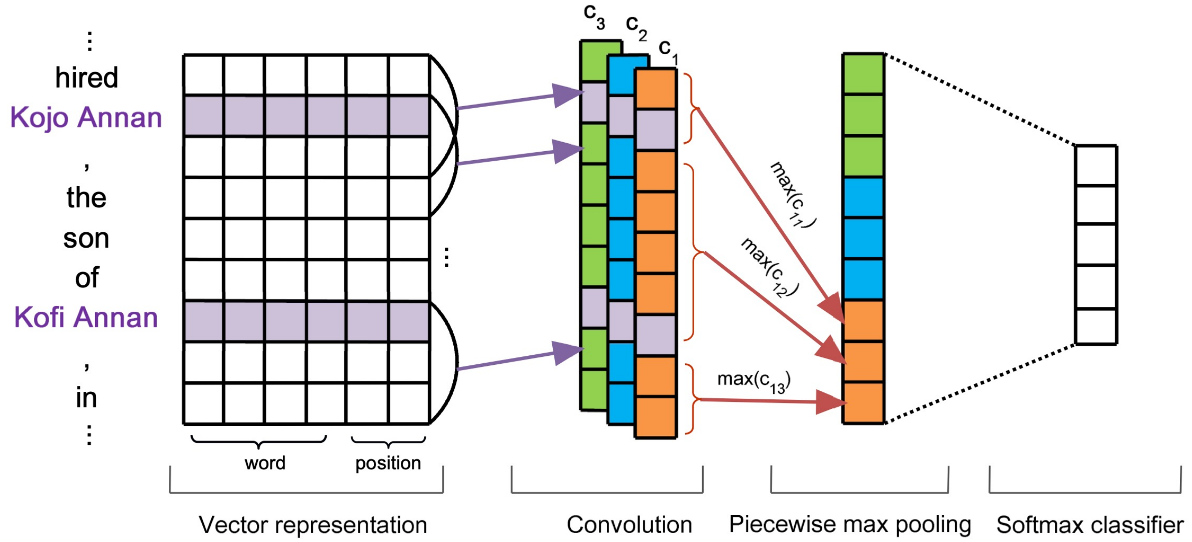


Figure 3: The architecture of PCNNs (better viewed in color) used for distant supervised relation extraction, illustrating the procedure for handling one instance of a bag and predicting the relation between *Kojo Annan* and *Kofi Annan*.

Algorithm 1 Multi-instance learning

1: Initialize θ. Partition the bags into mini-batches of size b_s.
2: Randomly choose a mini-batch, and feed the bags into the network one by one.
3: Find the j-th instance m_i^j $(1 \leq i \leq b_s)$ in each bag according to Eq. (9).
4: Update θ based on the gradients of m_i^j $(1 \leq i \leq b_s)$ via Adadelta.
5: Repeat steps 2-4 until either convergence or the maximum number of epochs is reached.

Zeng et al., Distant Supervision for Relation Extraction via Piecewise Convolutional Neural Networks, EMNLP 2015.

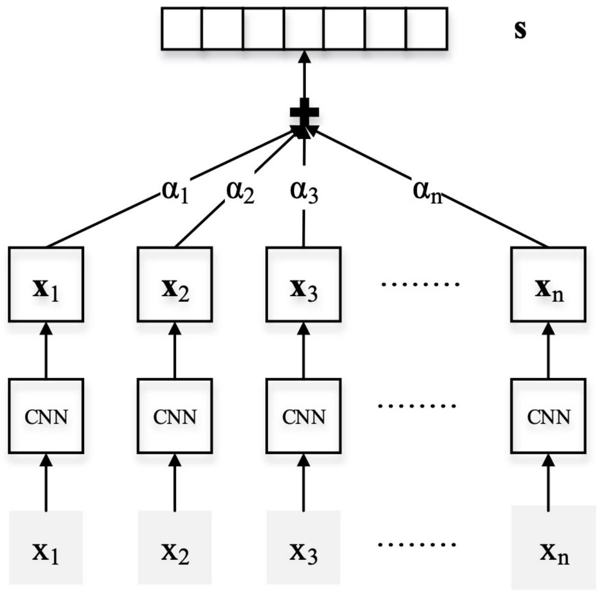


Figure 1: The architecture of sentence-level attention-based CNN, where x_i and $\mathbf{x}_i$ indicate the original sentence for an entity pair and its corresponding sentence representation, α_i is the weight given by sentence-level attention, and $\mathbf{s}$ indicates the representation of the sentence set.

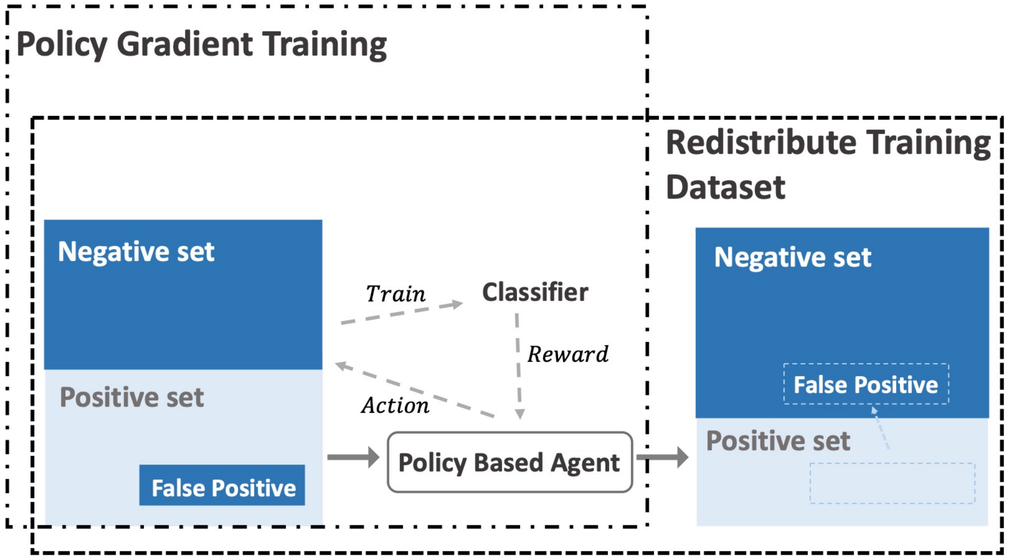


Figure 1: Our deep reinforcement learning framework aims at dynamically recognizing false positive samples, and moving them from the positive set to the negative set during distant supervision.

Attention

Reinforcement Learning

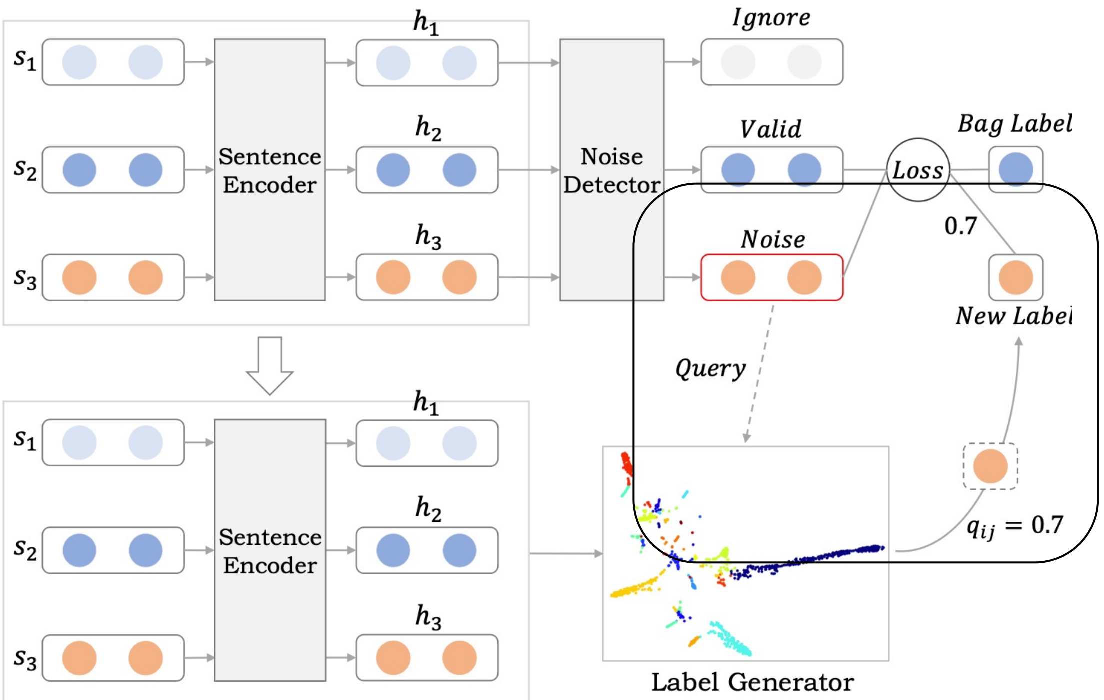
Lin et al., Neural relation extraction with selective attention over instances, ACL 2016.
Qin et al., Robust Distant Supervision Relation Extraction via Deep Reinforcement Learning, ACL 2018.

FNLP

3 Robustness Improvement form Model Perspective---Modeling Method

	Sentence	Bag Label	Noise?	Correct Label
Bag	#1: **Barack Obama** was born in the **United States**.	president of	Yes	born in
	#2: **Barack Obama** was the first African American to be elected to the president of the **United States**.		No	president of
	#3: **Barack Obama** served as the 44th president of the **United States** from 2009 to 2017.		No	president of

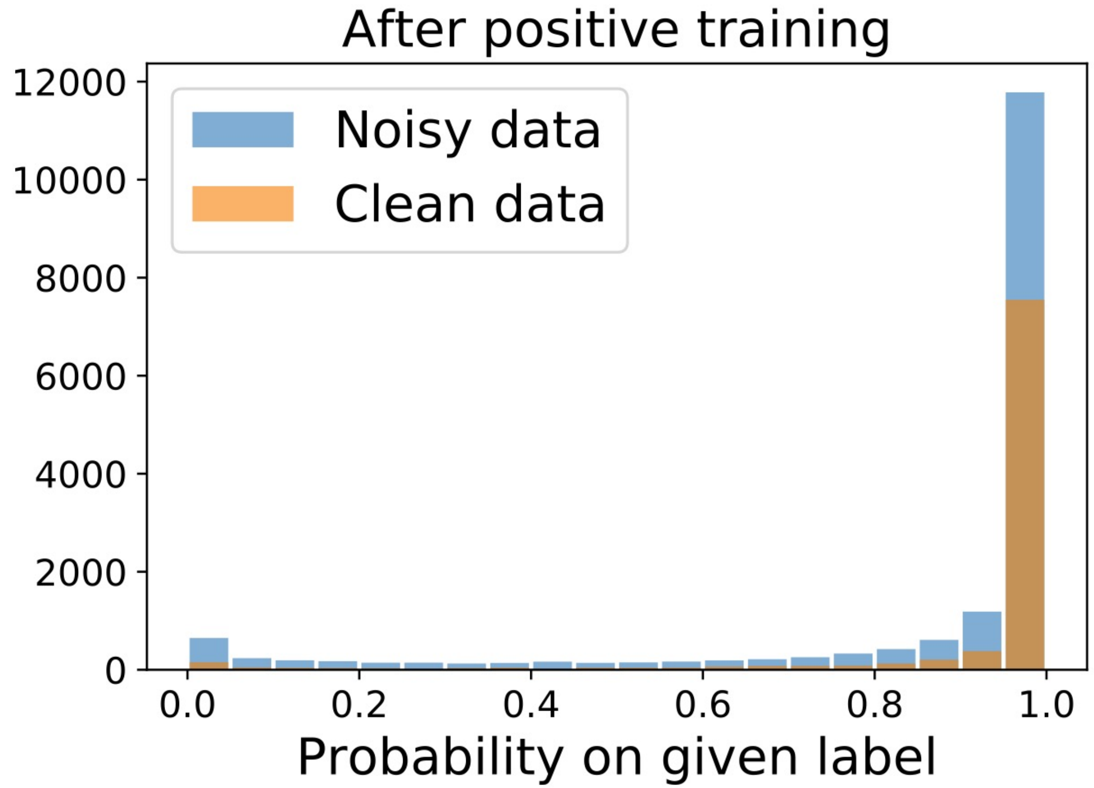
Table 1: An example of sentence-bag annotated by distant supervision. "Yes" and "No" indicate whether or not each sentence is a noisy sentence. "Correct Label" means the true relationship between the entity pair expressed in each sentence.



Shang, et al., Are Noisy Sentences Useless for Distant Supervised Relation Extraction? AAAI 2020.

3 Robustness Improvement form Model Perspective---Modeling Method

$$\mathcal{L}_{PT}(f, y^*) = -\sum_{k=1}^{C} y_k \log p_k$$



Positive Training 范式很难区分出干净数据与噪音数据

FNLP

2 Noisy Label in Distant Supervision

Positive Training

"is a **dog**"

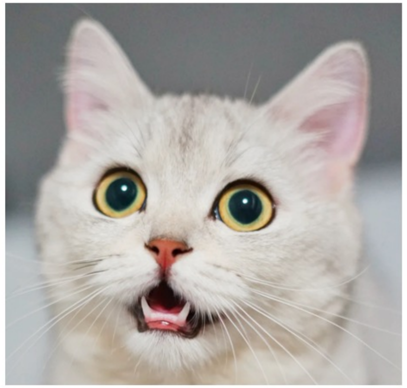
Correct Labels

"is a **cat**"
"is a **bird**"
"is an **elephant**"
"is a **chicken**"
"is a **pig**"

Incorrect labels

$$\mathcal{L}_{PT}(f, y^*) = -\sum_{k=1}^{C} y_k \log p_k$$

Negative Training



Correct Labels

"is not a **dog**"
"is not a **bird**"
"is not an **elephant**"
"is not a **chicken**"
"is not a **pig**"

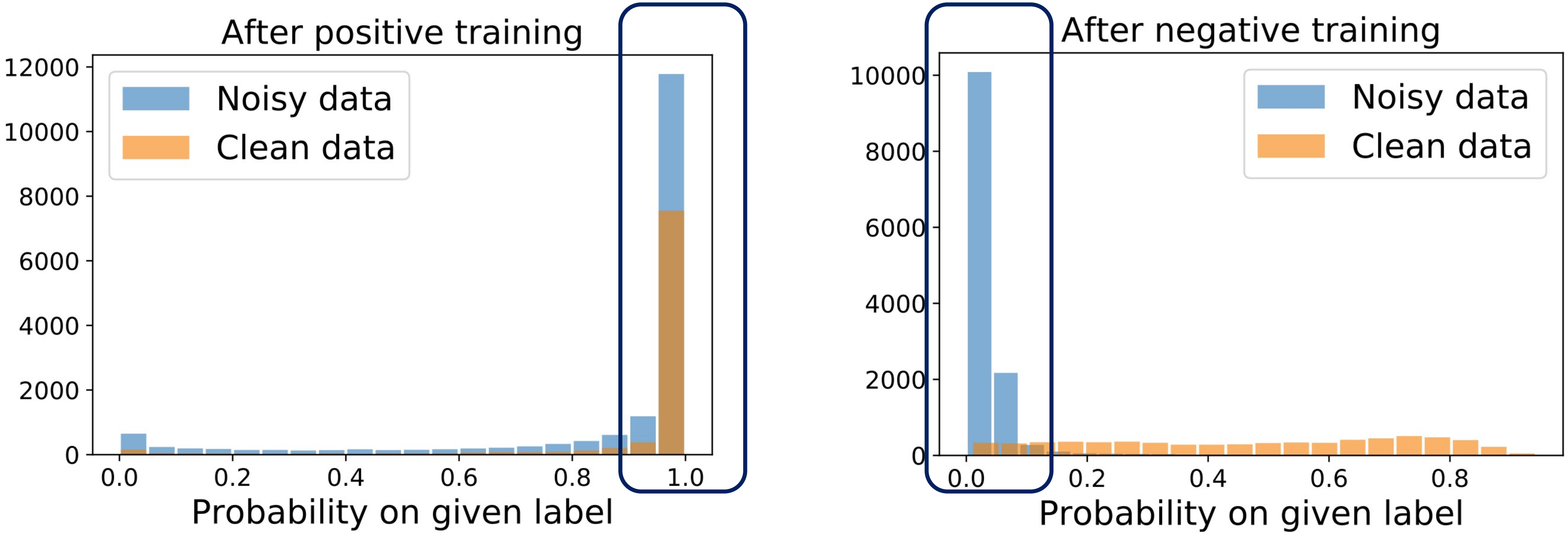
Incorrect labels

"is not a **cat**"

$$\mathcal{L}_{NT}(f, y^*) = -\sum_{k=1}^{C} \overline{y_k} \log(1 - p_k)$$

FNLP

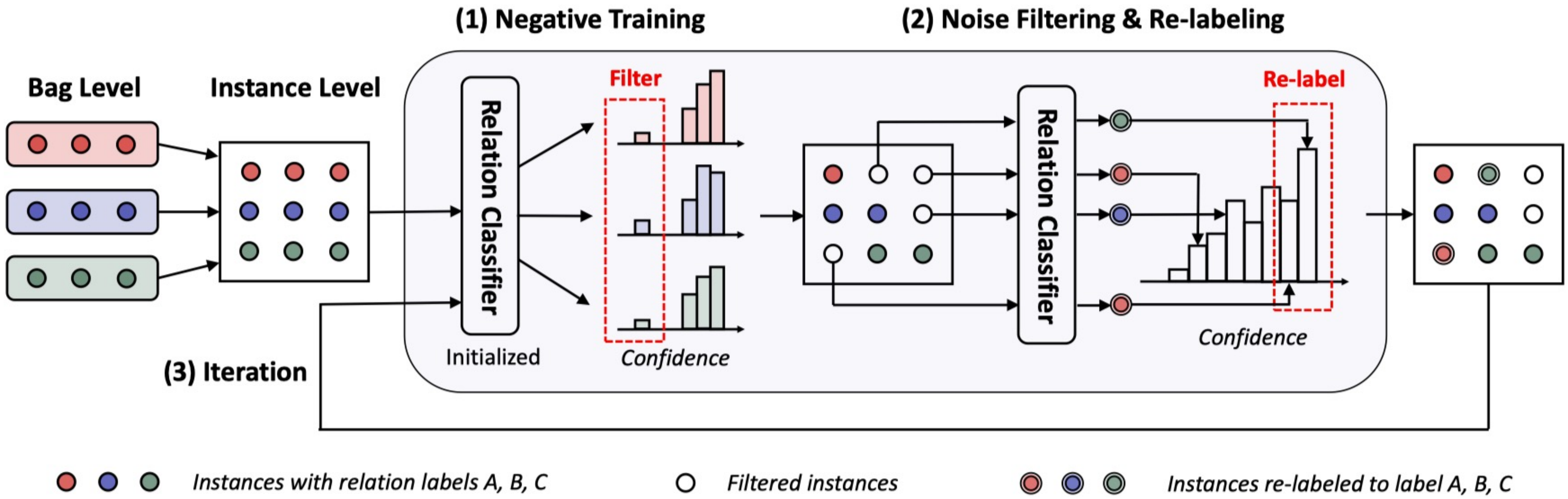
2 Noisy Label in Distant Supervision



Comparison between positive and negative training

FNLP

2 Noisy Label in Distant Supervision



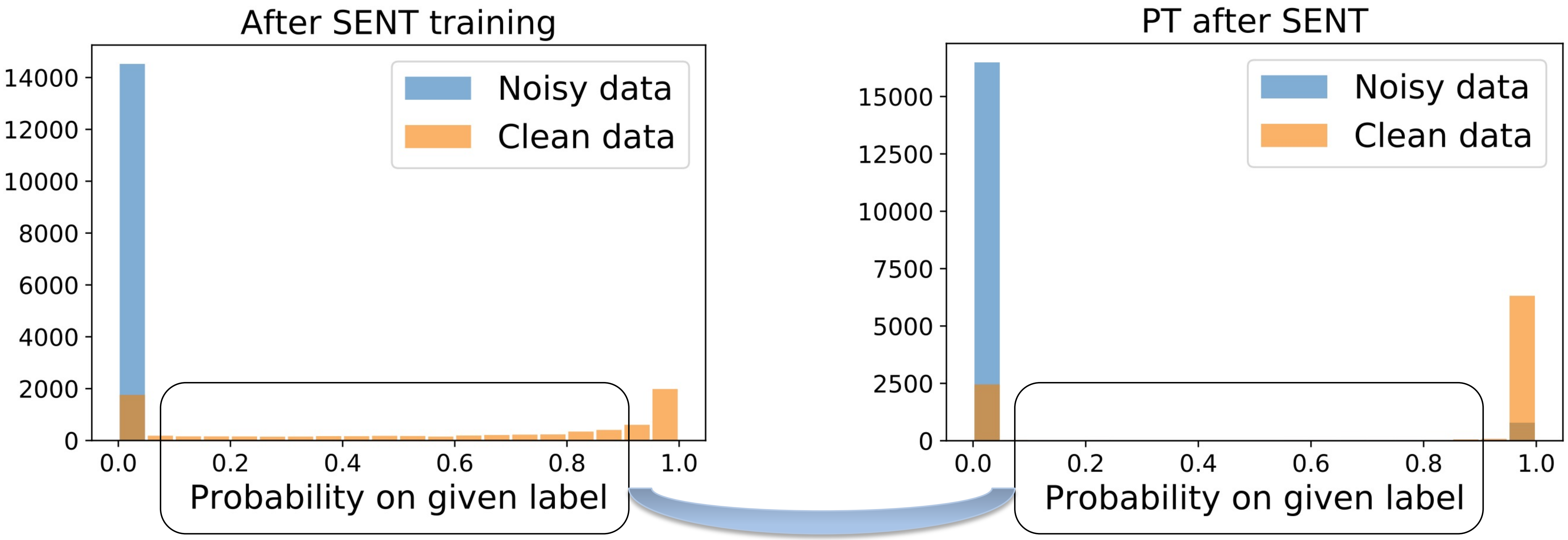
An overview of the proposed framework, SENT, for sentence-level distant RE.

(1) Negative training for separating the noisy data from the training data
(2) Noise-filtering and re-labeling
(3) Iterative training to further boost the performance.

Ma et al., SENT: Sentence-level Distant Relation Extraction via Negative Training, ACL 2021

FNLP

2 Noisy Label in Distant Supervision



(1) After SENT training, the clean and noisy data are further separated

(2) PT after SENT helps improve the convergence of the clean data

Ma et al., SENT: Sentence-level Distant Relation Extraction via Negative Training, ACL 2021

FNLP

2 Noisy Label in Distant Supervision

<table>
<tr><td colspan="2">Datasets</td><td>NYT-10</td><td>noisy-TACRED</td></tr>
<tr><td colspan="2">Label</td><td>24</td><td>41</td></tr>
<tr><td rowspan="3">Train</td><td>inst.</td><td>371461</td><td>68124</td></tr>
<tr><td>posi.</td><td>110518</td><td>13012</td></tr>
<tr><td>noise</td><td>unknown</td><td>20586</td></tr>
<tr><td rowspan="2">Dev</td><td>inst.</td><td>2379</td><td>22631</td></tr>
<tr><td>posi.</td><td>337</td><td>5436</td></tr>
<tr><td rowspan="2">Test</td><td>inst.</td><td>2164</td><td>15509</td></tr>
<tr><td>posi.</td><td>323</td><td>3325</td></tr>
</table>

Table 1: Statistics of datasets.

Method	**Dev**			**Test**		
	Prec.	**Rec.**	**F1**	**Prec.**	**Rec.**	**F1**
CNN(Zeng et al., 2014)	38.32	65.22	48.28	35.75	64.54	46.01
PCNN(Zeng et al., 2015)	36.09	63.66	46.07	36.06	64.86	46.35
BiLSTM(Zhang et al., 2015)	36.71	66.46	47.29	35.52	67.41	46.53
BiLSTM+ATT(Zhang et al., 2017)	37.59	64.91	47.61	34.93	65.18	45.48
BERT(Devlin et al., 2019)	34.78	65.17	45.35	36.19	70.44	47.81
BiLSTM+BERT(Devlin et al., 2019)	36.09	73.17	48.34	33.23	72.70	45.61
PCNN+SelATT(Lin et al., 2016)	46.01	30.43	36.64	45.41	30.03	36.15
PCNN+RA_BAG_ATT(Ye and Ling, 2019)	49.84	46.90	48.33	56.76	50.60	53.50
CNN+RL$_1$ (Qin et al., 2018)	37.71	52.66	43.95	39.41	61.61	48.07
CNN+RL$_2$ (Feng et al., 2018)	40.00	59.17	47.73	40.23	63.78	49.34
ARNOR(Jia et al., 2019)	62.45	58.51	60.36	65.23	56.79	60.90
SENT (BiLSTM)	66.44	56.97	61.34	71.80	59.13	64.86
SENT (BiLSTM+BERT)	**69.43**	**63.72**	**66.45**	**75.78**	**63.82**	**69.29**

Table 2: Main results on sentence-level evaluation. Compared baselines include normal RE model (the first part of the table), and models for distant RE (the second part of the table)

Ma et al., SENT: Sentence-level Distant Relation Extraction via Negative Training, ACL 2021

FNLP

2 Noisy Label in Distant Supervision

323 Samples including 200 incorrect samples

Noise Reduction	Prec.	Rec.	F1
CNN+RL$_2$	40.58	96.31	57.10
ARNOR	76.37	68.13	72.02
SENT (biLSTM)	80.00	**88.46**	84.02
SENT (biLSTM+BERT)	**84.33**	85.67	**84.99**

Table 3: The noise-filtering effect evaluated on a noise-annotated test set of NYT-10.

13012 correct samples
20586 incorrect samples

	Method	Prec.	Rec.	F1
Clean Data	BiLSTM+ATT	67.7	63.2	65.4
	BiLSTM	61.4	61.7	61.5
Noisy Data	BiLSTM+ATT	32.8	43.8	37.5
	BiLSTM	37.8	45.5	41.3
	SENT (biLSTM)	**66.0**	**52.9**	**58.7**

Table 4: Model performance on clean and noisy-TACRED. When trained on noisy data, the performance of base models degrade dramatically while SENT achieves comparable results with the models trained on clean data.

Ma et al., SENT: Sentence-level Distant Relation Extraction via Negative Training, ACL 2021

FNLP

0 Conclusion

- 鲁棒性提升
  - **数据层面**
    - 通用对抗训练
    - NLP 对抗训练
      - **Discrete Nature**
  - 表示层面
    - 非预训练
    - 预训练
      - **Robust Compression**
      - **Robust Prompt**
  - 模型层面
    - 附加网络
    - 建模方式
      - **Various Tasks**
      - **Redundant Features**
      - **Causality**
  - 评测层面
    - TextFlint
      - **Guidance**

FNLP

Thanks for your attention!